Penalized Generalized Linear Mixed Models for Variable Selection in Genetic Association Studies

Julien St-Pierre

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec
January 2025

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Dedication

I dedicate this thesis to my brother, in loving memory.

Acknowledgements

First and foremost, I would like to express my most sincere gratitude to my PhD co-supervisors, Dr. Karim Oualkacha, Dr. Sahir Rai Bhatnagar and Dr. Josée Dupuis, for their help and support over these past years. I wish to thank Karim for sharing his passion for research, for introducing me to the area of statistical genetics during my master studies, and for the continuous support, encouragements, generosity, and for the many hours we spent discussing where he shared valuable knowledge with me. I would like to thank Sahir for advising me, inviting me to various conferences and introducing me to many collaborators. I wish to thank Josée for advising me, for sharing her valuable knowledge and experience, and for always being available when I needed her guidance. I am grateful to all of my co-supervisor for supporting me during my PhD, and for making these past few years so enjoyable.

At the Department of Epidemiology, Biostatistics and Occupational Health, I am indebted to the several Professors who generously shared their knowledge and excitement for biostatistics with me; in particular, I wish to thank Dr. Alexandra Schmidt, Dr. Erica E. M. Moodie, Dr. James Hanley, Dr. Andrea Benedetti, Dr. Celia M. T. Greenwood and Dr. Qihuang Zhang for their teaching and mentoring. I want to thank my friends and colleagues at McGill University, and in particular, Victoire, Renaud, Vanessa, Marc, Niki, Janie, Daniel, Armando and Steve for the great times we had, for the laughters, and for your support all throughout this journey. I am also very grateful to Andre Yves Gagnon, Katherine Hayden, Dolores Coleto, and Nathalie Theoret, for their support and for facilitating my studies at McGill University. I also wish to thank the Professors of the Department of Statistics at McGill University for their outstanding teaching.

I wish to acknowledge the financial support I received from the *Réseau Québécois sur le Suicide, Troubles de l'Humeur, et Troubles associés*, from my PhD co-supervisors, and from

the Faculty of Medicine and Health Sciences at McGill University. I would also like to thank Compute Canada for the computing resources they provided, which were crucial to this work, and for assisting whenever I needed technical assistance. I am also indebted to Dr. Massimiliano Orri from the Douglas Hospital Research Centre for his financial support, collaboration and guidance during the projects we undertook together. I would also like to thank Léa Perret and the other members of the Life-course Suicide Epidemiology Research lab who welcomed me into their research group and made me feel like a part of the team.

I am indebted to a few other mentors whom I would like to acknowledge here, who influenced me in pursuing doctoral studies and taught me a lot along the way. First, I would like to thank Dr. Sorana Froda who always believed in me and who transmitted her passion for statistics as I was starting my journey from physics engineering to statistics. She was the first one who recognized my potential and I wouldn't have pursue graduate studies without her support and encouragements. I would also tank Dr. Fabrice Larribe and Dr. Geneviève Lefebvre for their support and collegiality during my master studies. I am also indebted to Dr. Simon Drouin who supervised me during my internship at the CHU Sainte-Justine Research Center.

I would also like to thank my close friends who always supported me and been there during the ups and downs: Eve, Catherine, Jérémy, Éliote, Jeanne, Dominique, Renaud, Patrick, Victoire, Vanessa, Niki and Marc. I am forever grateful for your friendship and for always believing in me. This thesis belongs in part to my father who has been supporting and encouraging me throughout this whole journey. I am grateful for his wisdom and for reminding me to always pursue my dreams when I was doubting. I hope he sees himself reflected in this accomplishment. J'aimerais aussi dédier cette thèse à mon frère Léandre, dont le courage et la force m'inspirent tous les jours à me surpasser. Finalement, j'aimerais remercier Félix pour son amour et son support indéfectible durant ces nombreuses années.

Preface

This manuscript-based thesis contains six chapters: an introduction, an original literature review, three chapters that correspond to three stand alone manuscripts, and a conclusion. A complete bibliography is presented after the appendices, at the end of this thesis. Chapters 3, 4 and 5 are linked by the main research topic of this thesis, and each adds novel methodological developments and new insights to the current statistical literature in the area of penalized mixed models for genetic association studies. Each of these three chapters begins with a short preamble that introduces the topic of the chapter and that briefly describes the gap in literature that I seek to fill with the proposed methodology. The proposed methodologies are all demonstrated using real genetic cohort studies.

The introduction and the literature review (Chapters 1 and 2) of this thesis were conceived and written by Julien St-Pierre (JStP), and both were further edited by Karim Oualkacha (KO) and Josée Dupuis (JD). The work in Chapter 3 was conceptualized in a series of discussion between JStP, Sahir Rai Bhatnagar (SB) and KO. JStP conducted the methodological derivations, designed and conducted the simulation study, performed the data analysis and wrote the manuscript draft. SB and KO provided substantial help and guidance with the methodological derivations, simulation studies and data analysis. SB and KO further corrected and edited the chapter. The methodological work in Chapter 4 was conceptualized by JStP with help and guidance by SB and KO. JStP conducted the methodological derivations, designed and conducted the simulation study, performed the data analysis and wrote the manuscript draft. The work was advised by SB and KO, which also corrected and edited the chapter. The work in Chapter 5 was conceptualized in a series of discussion between JStP and KO. JStP conducted the methodological derivations, designed and conducted the simulation study, performed the data analysis and wrote the manuscript draft. KO provided substantial help and guidance with the methodological derivations and simulation studies.

Massimiliano Orri (MO), KO and JD provided guidance and support for the real data analysis. KO, JD and MO advised the work and edited the chapter. The conclusion was conceived and written by JStP and edited by JD and KO.

Preface

The case study in Chapter 3 of this thesis uses the data from the UK Biobank. UK Biobank is a large-scale biomedical database and research resource containing de-identified genetic, lifestyle and health information and biological samples from half a million UK participants (Sudlow et al. [2015]). UK Biobank is generously supported by its founding funders the Wellcome Trust and UK Medical Research Council, as well as the Department of Health, Scottish Government, the Northwest Regional Development Agency, British Heart Foundation and Cancer Research UK. The organisation has over 150 dedicated members of staff, based in multiple locations across the UK. More information on obtaining access to the UK Biobank data is available on the project website (http://www.ukbiobank.ac.uk).

The case study in Chapter 4 of this thesis uses data from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA), OPPERA II and Complex Persistent Pain Conditions (CPPC): Unique and Shared Pathways of Vulnerability 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. 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).

The case study in Chapter 5 of this thesis uses data from the Quebec Longitudinal Study of Child Development (QLSCD) and the Quebec Newborn Twin Study (QNTS). QLSCD and QNTS data are accessible to researchers on the premises of the Centre d'accès aux données de recherche de l'Institut statistique du Québec (CADRISQ) located in Montreal and Quebec City. More information on obtaining access to the QLSCD data can be found on Research Data Access Point website (https://www.stat.gouv.qc.ca/research/#/accueil). More information on obtaining access to the QNTS data can be found on the Groupe de recherche sur l'inadaptation psychosociale chez l'enfant (GRIP) website (http://www.gripinfo.ca/grip/public/www/etudes/fr/dadprocedures.asp).

Abstract

Genome-wide association studies (GWAS) aim to identify important genetic predictors associated with measurable traits, i.e, phenotypes. GWAS are typically conducted by testing association on each genetic variant independently, requiring a stringent multiple-testing threshold to avoid false positives. Penalized regression methods have been proposed as an alternative method to increase the power for identifying weaker genome-wide associations. Population-based cohorts often include diverse and admixed individuals, as well as individuals with known or unknown familial relatedness, and failing to account for the aforementioned can decrease power and lead to spurious associations. Thus, including the top principal components (PCs) of a genetic relatedness matrix (GRM) and/or a random effect with variance-covariance structure proportional to a GRM is warranted. While mixed models are now widely employed in GWAS and the literature is expanding, existing methods are mostly focusing on univariate association models. In my doctoral thesis, I focus on developing penalized generalized linear mixed models (GLMMs) for identifying and estimating important genetic predictors effects, while accounting for non-normality of the traits and correlation between different observations.

In a first manuscript, I introduce a new method that allows to simultaneously select genetic markers and estimate their effects, accounting for between-individual correlations and binary nature of the trait. I develop a computationally efficient algorithm based on penalized quasi-likelihood (PQL) estimation that allows to scale regularized mixed models on high-dimensional binary trait GWAS. I show through simulations that when the dimensionality of the relatedness matrix is high, a penalized LMM or logistic regression model with PC adjustment fail to select important predictors, and have inferior prediction accuracy compared to the proposed model. Further, I demonstrate through the analysis of two polygenic binary traits that the proposed method can achieve higher predictive performance, while also selecting fewer predictors than a sparse regularized logistic lasso with PC adjustment.

In a second manuscript, I extend the model proposed in the first manuscript to perform hierarchical selection of gene-environment interaction (GEI) effects in sparse regularized GLMMs, accounting for population structure, close relatedness, shared environmental exposure and binary nature of the trait. I propose to combine PQL estimation with a sparse group lasso penalty and derive a proximal Newton-type algorithm with block coordinate descent. I show that for all simulation scenarios, the proposed method always select the lowest number of predictors in the model, while maintaining low false positive rates (FPR) and false discovery rates (FDR). Moreover, using real data from the OPPERA study to explore the comparative performance of the model in selecting important predictors of temporomandibular disorder (TMD), I show that the proposed method is able to identify a previously reported significant SNP in a combined or sex-segregated GWAS.

The third manuscript 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 behavioral difficulties in childhood and adolescence. The methodology proposed in the previous two chapters is extended to allow multiple measurements per subject through the use of multiple random effects. Through simulation studies, I first show that using a GRM to account for both population structure and closer relatedness is not enough as it inflates the relative bias of variance components estimates, and adjusting for population structure by adding the 10 PCs is warranted. Moreover, I show that even though using a sparse GRM in place of the full GRM does introduce bias in the variance components estimates, the computational gain is major while the impact on the performance of the penalized model to identify important predictors is negligible. I 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, I show an application of the proposed model to predict externalizing scores in the combined QLSCD and QNTS longitudinal cohorts.

Abrégé

Les études d'association pangénomique (GWAS) visent à identifier d'importants prédicteurs génétiques associés à des traits mesurables, c'est-à-dire des phénotypes. Les GWAS sont généralement menées en testant l'association entre un phénotype et chaque variant génétique de manière indépendante, nécessitant un seuil de test multiple strict pour éviter les faux positifs. Des méthodes de régression pénalisée ont été proposées comme une alternative pour augmenter la puissance d'identification des associations pangénomiques plus faibles. Les cohortes populationnelles incluent souvent des individus provenant de diverses populations, ainsi que des individus avec des liens familiaux connus ou inconnus, et ne pas prendre en compte ces facteurs peut réduire la puissance et conduire à des fausses associations. Ainsi, inclure les premières composantes principales (PCs) d'une matrice de similarité génétique (GRM) et/ou un effet aléatoire avec une structure de variance-covariance proportionnelle à une GRM est nécessaire. Bien que les modèles mixtes soient désormais largement utilisés dans les GWAS et que la littérature s'agrandisse, les méthodes existantes se concentrent principalement sur les modèles d'association univariés. Dans ma thèse de doctorat, je me concentre sur le développement de modèles linéaires généralisés mixtes (GLMMs) pénalisés pour identifier et estimer les effets des prédicteurs génétiques importants, tout en tenant compte de la non-normalité des traits et de la corrélation entre différentes observations.

Dans un premier article, je présente une nouvelle méthode qui permet de sélectionner simultanément les marqueurs génétiques et d'estimer leurs effets, en tenant compte des corrélations entre les individus et de la nature binaire du trait. Je développe un algorithme efficace sur le plan computationnel basé sur l'estimation par quasi-vraisemblance pénalisée (PQL) qui permet d'adapter les modèles mixtes régularisés aux GWAS de traits binaires. Je montre à travers des simulations que lorsque la dimensionnalité de la matrice de similarité est élevée, les modèles linéaires mixtes pénalisés et la régression logistique avec ajustement par PC échouent à sélectionner les prédicteurs importants et ont une précision de prédiction

inférieure par rapport au modèle proposé. De plus, je démontre, à travers l'analyse de deux traits polygéniques binaires, que la méthode proposée peut atteindre une performance prédictive plus élevée, tout en sélectionnant moins de prédicteurs qu'une régression logistique régularisée par lasso avec ajustement par PC.

Dans un second article, j'étends le modèle proposé dans le premier article pour effectuer une sélection hiérarchique des effets d'interaction gène-environnement (GEI) dans les GLMMs régularisés sparses, en tenant compte de la structure de la population, des relations familiales, de l'exposition environnementale partagée et de la nature binaire du trait. Je propose de combiner l'estimation par PQL avec une pénalité de groupe lasso sparse et je dérive un algorithme de Newton proximal avec descente par coordonnées en bloc. Je montre que pour tous les scénarios de simulation, la méthode proposée sélectionne toujours le plus petit nombre de prédicteurs dans le modèle, tout en maintenant de faibles taux de faux positifs (FPR) et de fausses découvertes (FDR). De plus, en utilisant des données réelles de l'étude OPPERA pour explorer la performance comparative du modèle dans la sélection des prédicteurs importants du trouble temporomandibulaire (TMD), je montre que le modèle proposé est capable d'identifier un SNP significatif précédemment rapporté dans une GWAS stratifiée ou non par sexe.

Le troisième article est motivé par des analyses de données longitudinales collectées auprès de participants de l'Étude longitudinale du développement des enfants du Québec (QLSCD) et de l'Étude des jumeaux nouveau-nés du Québec (QNTS) pour identifier d'importants prédicteurs génétiques des difficultés émotionnelles et comportementales durant l'enfance et l'adolescence. À travers des études de simulation, je montre d'abord que l'utilisation d'une GRM pour tenir compte à la fois de la structure de la population et des liens familiaux n'est pas suffisante, car cela augmente le biais relatif des composantes de variance, et que l'ajustement pour la structure de la population en ajoutant les 10 premiers PCs est nécessaire. De plus, je montre que bien que l'utilisation d'une GRM sparse à la place de la GRM complète

introduise un biais dans l'estimation des composantes de variance, le gain computationnel est majeur tandis que l'impact sur la performance du modèle pénalisé pour identifier les prédicteurs importants est négligeable. Je compare la performance du modèle mixte pénalisé proposé à un lasso standard et à un test d'association mixte univarié et montre que le modèle proposé identifie toujours les prédicteurs causaux avec une plus grande précision. Enfin, je montre une application du modèle proposé pour prédire les scores externalisés dans les cohortes longitudinales combinées QLSCD et QNTS.

Table of contents

List of Tables

List of Figures

Abbreviations

AIREML average information restricted maximum likelihood

BCD blockwise coordinate descent

CAP composite absolute penalties

CGD coordinate gradient descent

DNA deoxyribonucleic acid

FDR false discovery rate

FPR false positive rate

GEI gene-environment interaction

GLM generalized linear model

GLMM generalized linear mixed model

GMMAT generalized linear mixed model association test

GRM genetic relatedness matrix

GSM genetic similarity matrix

GWAS genome-wide association studies

LD linkage disequilibrium

LMM linear mixed model

MCP minimax concave penalty

ML maximum likelihood

MLE maximum-likelihood estimation

MM majorization-minimization

MM mixed model

OLS ordinary least squares

PC principal component

PCA principal component analysis

PQL penalized quasi-likelihood

PRS polygenic risk score

QR quantile regression

REML restricted maximum likelihood

SCAD smoothly clipped absolute deviation

SNP single nucleotide polymorphism

TMD temporomandibular disorder

TPR true positive rate

WLS weighted least squares

Chapter 1

Introduction

This thesis focuses on variable selection models in the context of genetic association studies. I address two important challenges in those studies, which are i) the spurious selection of variables due to confounding by population stratification and close relatedness of the relationship between genetic predictors and the outcome, and ii) the computational complexity associated with fitting mixed models when the number of predictors is significantly larger than the sample size.

The observed variation for certain measurable traits in a population, also called phenotypes, are influenced by the environment and variations in the genes that code for these phenotypes, to varying degrees depending on the trait under study. The deoxyribonucleic acid (DNA) located on the gene locus differs from one individual to another. These variations in nucleotides for a specific locus are called alleles. Since each individual inherits pairs of homologous chromosomes at birth, one chromosome from the father and one from the mother, there are precisely two alleles at each genome location for a trait. The majority of observed genetic diversity is due to genetic variations affecting only a single pair of nucleotide bases, called single nucleotide polymorphisms (SNPs). A biallelic site is a specific locus in a genome that contains two observed alleles. In this case, the allele most prevalent in a population for

a SNP is referred to as the major allele, contrasting with the minor allele, which is less frequent. Multiallelic sites refer to specific loci in a genome that contain three or more observed alleles. Confounding due to population structure comes from the fact that allele frequencies can differ greatly between individuals who do not share similar ancestry. Moreover, genetic similarity between close individuals, referred to as genetic kinship, can also confound the association between genetic predictors and phenotypes.

Genome-wide association studies (GWAS) have led to the identification of hundreds of common genetic variants, or SNPs, associated with complex traits (Visscher et al. [2017]). This thesis focuses on regularized regression models for GWAS, in which the effects of a large number of genetic predictors are jointly estimated in a penalized model in order to induce sparsity among the predictors fixed effects. This is in contrast to univariate methods which typically test the statistical association between one or many phenotypes and each genetic variant, or SNP, independently. Although these two approaches are fundamentally different in the way they address having significantly more predictors than observations, they are prone to the same biases that arise from population structure and genetic kinship.

Several authors have proposed in the statistical literature methods to adjust for population structure and genetic similarity between individuals in univariate association tests (see e.g., Yu et al. [2005], Price et al. [2006], Price et al. [2010], Yang et al. [2011], Conomos et al. [2015], Sul et al. [2016], Chen et al. [2016]). Sparse regularized linear mixed models (LMMs) have also been proposed in the statistical literature to perform variable selection in GWAS of continuous phenotypes while adjusting for population structure and genetic kinship (see e.g., Rakitsch et al. [2012], Bhatnagar et al. [2020b]). Nevertheless, sparse regularized mixed models were not yet proposed in the literature for GWAS of binary traits, due to the computational complexity associated with fitting high dimensional generalized linear mixed models (GLMMs).

In this thesis, I therefore build on the previous statistical literature on sparse regularized

GLMMs to propose and demonstrate a methodology for the selection and estimation of genetic effects for GWAS of binary traits. In Chapter 2, I provide a review of the fundamental concepts to consider in the development of that methodology. More explicitly, I discuss the important aspects of GWAS, variance components estimation in mixed models, and algorithms to estimate fixed effects coefficients in sparse regularized regression models.

In Chapter 3, a method is proposed to select important genetic predictors and estimate their effects in GWAS of binary traits, accounting for between-individual correlations and binary nature of the outcome. An algorithm based on regularized penalized quasi-likelihood (PQL) estimation is presented. The method is demonstrated in extensive simulation studies and is used to assess the predictive performance of a polygenic model for asthma and high cholesterol, in an analysis of data from the UK Biobank.

In Chapter 4, hierarchical selection of gene-environment interaction (GEI) effects in sparse regularized GLMMs is considered. A proximal Newton-type algorithm with blockwise coordinate descent (BCD) combining regularized PQL estimation with a sparse group lasso penalty is derived. The use of a second random effect to account for shared environmental exposure is demonstrated using different simulation scenarios. Finally, the proposed methodology is used to identify important predictors of temporomandibular disorder (TMD) that may interact with sex, in a GWAS using individuals from the OPPERA study.

Chapter 5 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 behavioral difficulties in childhood and adolescence. The methodology proposed in the previous two chapters is extended to allow multiple measurements per participant through the use of multiple random effects. Using simulation studies, it is demonstrated that using a genetic relatedness matrix (GRM) to account for both population structure and closer relatedness may not be sufficient as it inflates the relative bias of variance components estimates, and adjusting for

population structure by adding the top principal components (PCs) to the model is warranted. Moreover, even though the use of a sparse GRM introduces bias in the variance components estimates, it is demonstrated that the computational gain is major while the impact on the performance of the penalized model to identify important predictors is negligible. Using simulation studies of both continuous and binary traits, the proposed model is shown to always identify causal predictors with greater precision than other methods. Finally, an application of the methodology to build a prediction model for externalizing scores in the combined QLSCD and QNTS longitudinal cohorts is demonstrated.

Chapters 3 to 5 were written as stand-alone manuscripts. Chapter 3 is published in *Bioinformatics*. Chapter 4 is published in *Statistical Methods in Medical Research*. Chapter 5 is ready to be submitted for publication in a statistical journal. This thesis ends with a conclusion in which I review the notable contributions of this thesis, discuss important limitations of this work, and mention potential ideas for future work.

Chapter 2

Literature review

This chapter introduces and reviews the theory that we expanded upon in all three manuscripts comprising this thesis.

In this chapter, vector and matrices are denoted in bold. I use the notation y_i to denote the phenotypic outcome of individual $i, i = 1, ..., n$, $\mathbf{G}_i$ to refer to a vector of genetic predictors for individual i, taking values $\{0, 1, 2\}$ as the number of copies of the minor allele, with corresponding coefficients $\boldsymbol{\beta}$ denoting the effect of $\mathbf{G}_i$ on the outcome mean. In the univariate models, G_i and β represent the number of copies of the minor allele and regression coefficient for a single genetic predictor. Important covariates that are generally collected in GWAS, such as age, sex and ancestry, are included in the vector $\mathbf{X}_i$, with their corresponding fixed effects $\boldsymbol{\alpha}$.

2.1 Genome wide association studies

In this section, I first present the single-SNP linear regression model that is widely used in GWAS for continuous phenotypes, and I explain the rationale behind the use of a polygenic random effect to account for unexplained heritability. Second, I give an overview of the

average information restricted maximum likelihood (AIREML) algorithm that is a popular method to estimate variance components in mixed models. Third, I review the literature on population structure and participant relatedness, and the different mixed models that have been proposed in the recent years to address these two confounding sources. Fourth, I present the generalized linear mixed model association test (GMMAT) proposed by Chen et al. [2016] which is an extension of the mixed models literature for non-normally distributed phenotypes, and I review PQL estimation. I finish this section by introducing the concept of GEIs.

2.1.1 Single-SNP regression model

The single-SNP linear regression model can be written as

$$y_i = \boldsymbol{X}_i^{\intercal}\boldsymbol{\alpha} + G_i\beta + e_i,$$

where G_i is the number of copies of the minor allele for the SNP under study, and $e_i \sim N(0, \sigma_e^2)$ is the residual error. Standard maximum-likelihood estimation (MLE) procedures, such as the Wald test can be conducted to perform inference on β and draw evidence on the relative importance of a genetic predictor in explaining the observed variation in the phenotype $\mathbf{y}$.

Given the very large number of independent SNPs being tested for association, typically in the order of one million, a stringent multiple-testing threshold is required to avoid false positives. The Bonferonni genome-wide significance p-value threshold of 5×10^{-8} has become the standard, as it represents a false discovery rate of $0.05/10^6$ (Pe'er et al. [2008]). Moreover, GWAS have brought to light the problem of unexplained heritability, that is, identified variants only explain a low fraction of the total observed variability for traits under study (Manolio et al. [2009]). Yang et al. [2010] hypothesized that most causal variants explain such a small amount of variation that their effects do not reach stringent signifi-

cance thresholds, and they proposed modelling the additive genetic effects of all the SNPs as random effects in a LMM to estimate the heritability, i.e. the variance explained by all the causal genetic predictors. They proposed the following model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{W}\mathbf{u} + \mathbf{e},$$

where $\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \mathbf{I_M}\boldsymbol{\sigma^2_u})$ is a vector of random SNP effects, $\mathbf{I}_M$ is an $M \times M$ identity matrix with M the total number of SNPs, and $\mathbf{e} \sim N(0, \mathbf{I}_n\sigma^2_e)$ is the vector of residual (random) effects. If we define $\boldsymbol{\Phi} = \mathbf{W}\mathbf{W}'/M$, where $\mathbf{W}$ is the standardized design matrix for the M genetic predictors, and let $\sigma^2_g = M\sigma^2_u$ be the additive genetic variance, then we can rewrite the previous LMM as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{b} + \mathbf{e},$$

where $\mathbf{b} \sim N(0, \sigma^2_g\boldsymbol{\Phi})$ is an $n \times 1$ vector of polygenic random effects, and $\boldsymbol{\Phi}$ is the GRM between individuals. Yang et al. [2011] proposed using an AIREML approach (Harville [1977], Gilmour et al. [1995]) to estimate the polygenic variance component σ^2_g. I briefly review in the next section this estimation method in the context of LMMs.

2.1.2 AIREML estimation in LMMs

The previous LMM can be extended in a more general form to incorporate more than a single random effect, such as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{Z} = [\mathbf{Z}_1, ..., \mathbf{Z}_r]$ is a $n \times nr$ design matrix with $\mathbf{Z}_j$ the $n \times n$ design matrix for the j^{th} random effect, and $\mathbf{b} = (\mathbf{b}_1^\intercal, ..., \mathbf{b}_r^\intercal)^\intercal$ is a stacked $nr \times 1$ vector of random effects. For many genetic applications, it is convenient to assume that all random effects are mutually

independent, such that the total variance can be written as

$$\boldsymbol{\Sigma} = \sum_{j=1}^{r} \sigma_j^2 \mathbf{Z}_j \mathbf{V}_j \mathbf{Z}_j^\intercal + \sigma_e^2 \boldsymbol{I}_n,$$

where $\mathbf{V}_j$ is a $n \times n$ similarity matrix that accounts for between-subject relatedness, and σ_j^2 are variance components. The restricted log-likelihood function (Harville [1977]) for the model in (2.1) is equal to

$$\ell_R(\boldsymbol{\alpha}, \sigma_e^2, \sigma_1^2, ..., \sigma_r^2) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log\left|\boldsymbol{X}^{*\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^*\right| - \frac{1}{2}\mathbf{y}^\intercal\mathbf{P}\mathbf{y},$$

where $\boldsymbol{X}^*$ is a full-rank submatrix of $\boldsymbol{X}$ and $\mathbf{P}$ is a projection matrix defined by $\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}^*\left(\boldsymbol{X}^{*\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*\intercal}\boldsymbol{\Sigma}^{-1}$. In the restricted maximum likelihood (REML) method, estimates for the variance components and fixed effects are iteratively updated using Newton's method

$$\begin{aligned}
\boldsymbol{\alpha}^{(t+1)} &= \boldsymbol{\alpha}^{(t)} - [\nabla^2_{\boldsymbol{\alpha}}\ell_R]^{-1}\frac{\partial \ell_R}{\partial \boldsymbol{\alpha}}\,|_{(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\vartheta}^{(t)})}, \\
\boldsymbol{\vartheta}^{(t+1)} &= \boldsymbol{\vartheta}^{(t)} - [\nabla^2_{\boldsymbol{\vartheta}}\ell_R]^{-1}\frac{\partial \ell_R}{\partial \boldsymbol{\vartheta}}\,|_{(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\vartheta}^{(t)})},
\end{aligned}$$

where $-\nabla^2_{\boldsymbol{\alpha}}\ell_R$, $-\nabla^2_{\boldsymbol{\vartheta}}\ell_R$ are the observed information matrices with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\vartheta} = (\sigma_e^2, \sigma_1^2, ..., \sigma_r^2)$. The first and second derivative of ℓ_R with respect to $\boldsymbol{\vartheta}$ are

$$\begin{aligned}
\frac{\partial \ell_R}{\partial \vartheta_j} &= \frac{1}{2}\left\{\mathbf{y}^\intercal\mathbf{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_j}\mathbf{P}\mathbf{y} - tr\left(\boldsymbol{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_j}\right)\right\}, \\
\frac{\partial^2 \ell_R}{\partial \vartheta_l \partial \vartheta_j} &= \frac{1}{2}\left\{-2\mathbf{y}^\intercal\mathbf{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_l}\boldsymbol{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_j}\mathbf{P}\mathbf{y} - tr\left(\boldsymbol{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_l}\boldsymbol{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_j}\right)\right\},
\end{aligned}$$

with

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_j} = \begin{cases} \mathbf{I}_n & \text{if } j = 1, \\ \mathbf{Z}_j\mathbf{V}_j\mathbf{Z}_j^\intercal & \text{otherwise.} \end{cases}$$

The elements of the expected information matrix are

$$E\left(-\frac{\partial^2 \ell_R}{\partial \vartheta_l \partial \vartheta_j}\right) = \frac{1}{2} tr\left(\boldsymbol{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_l}\boldsymbol{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_j}\right).$$

Evaluating the trace in either the observed or expected information matrix is computationally expensive. Thus, Gilmour et al. [1995] proposed replacing the observed information matrix by the average of the observed and expected information matrices, which yields the updates

$$\boldsymbol{\vartheta}^{(t+1)} = \boldsymbol{\vartheta}^{(t)} + [\mathbf{AI}^{(t)}]^{-1}\frac{\partial \ell_R}{\partial \boldsymbol{\vartheta}} \mid_{(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\vartheta}^{(t)})},$$

with

$$\mathbf{AI}_{lj} = \frac{1}{2}\mathbf{y}^\intercal \mathbf{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_l}\boldsymbol{P}\frac{\partial \boldsymbol{\Sigma}}{\partial \vartheta_j}\mathbf{P}\mathbf{y}.$$

2.1.3 Population structure and subject relatedness

Confounding due to population structure or participants relatedness is a major issue in genetic association studies. Confounding comes from the fact that allele frequencies can differ greatly between individuals who do not share similar ancestry. Modern large scale cohorts will often include participants from different ethnic groups as well as admixed individuals, that is, subjects with individual-specific proportions of ancestries, or individuals with known or unknown familial relatedness, defined as cryptic relatedness (Sul et al. [2018]). When ignored, population structure and subject relatedness can decrease power and lead to spurious associations (Price et al. [2010]).

Principal component analysis (PCA) can control for the confounding effect due to population structure by including the top eigenvectors of the GRM as fixed effects in the regression model (Price et al. [2006]). With admixture and population structure being low dimensional fixed-effects processes, they can correctly be accounted for by using a relatively small number

of PCs (e.g. 10) (Novembre and Stephens [2008]). However, using too few PCs can result in residual bias leading to false positives, while adding too many PCs as covariates can lead to a loss of efficiency (Zhao et al. [2018]). More importantly, genetic studies frequently contain sample structure influenced by both population stratification and closer relatedness. Existing methods for inferring population structure may struggle when applied to samples with related individuals (Price et al. [2010], Thornton and Bermejo [2014]). Conomos et al. [2015] proposed a method called PC-AiR (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.

To adjust for both population structure and closer relatedness in a sample, including the top PCs and a random effect with variance-covariance structure proportional to a grm is warranted (Yu et al. [2005], Price et al. [2010]). Indeed, kinship is a high dimensional process, such that it cannot be fully captured by a few PCs, and it would require the inclusion of too many PCs as fixed effects covariates relative to the dimension of the sample size. An important limitation of mixed model association tests is the need to fit a large number of mixed-effects regression models, one per variant, across the genome in order to estimate the variance components. A common approach to reduce this computational burden is to fit a regression mixed model under the null hypothesis only once per GWAS, assuming that the variance components are the same for all variants. This is is known as the P3D (population parameters previously determined) method (Zhang et al. [2010]), and it has been shown to outperform both PCA and genomic control in correcting for sample structure (Kang et al. [2010]) in mixed model association tests. Thus, LMMs are now widely used in GWAS to test for association between individual genetic variants and a phenotype of interest while adjusting for all sources of confounding (Kang et al. [2008], Zhang et al. [2010], Kang et al. [2010], Yang et al. [2011], Lippert et al. [2011], Zhou and Stephens [2012], Loh et al. [2015]). By possibly incorporating more than a single random effect, LMMs allow to account for

other sources of confounding that may arise from more complex designs.

2.1.4 GMMAT

For binary traits, Chen et al. [2016] have shown that LMMs are generally inappropriate when population stratification leads to violation of the LMM's constant-residual variance assumption. They proposed the following GLMM for a continuous or binary phenotype y_i, $i = 1, ..., n$,

$$\eta_i = g(\mu_i) = \boldsymbol{X}_i^\intercal \boldsymbol{\alpha} + G_i\beta + b_i, \tag{2.2}$$

where $\mathbb{E}[y_i|\mathbf{b}] = \mu_\mathbf{i}$, $\boldsymbol{b} = (b_1, ..., b_n)^\intercal \sim \mathcal{N}(\mathbf{0}, \sum_{k=1}^{K} \tau_k \boldsymbol{V}_k)$ is an $n \times 1$ column vector of random intercepts, $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)^\intercal$ are the variance component parameters that account for the relatedness between individuals, and $\boldsymbol{V}_1, ..., \boldsymbol{V}_K$ are known relatedness matrices. Typically, $\boldsymbol{V}_1$ is the GRM $\boldsymbol{\Phi}$ between individuals to account for closer relatedness that cannot be captured by the leading PCs.

Chen et al. [2016] proposed estimating the variance component parameters $\boldsymbol{\tau}$ only once, under the null model of no genetic association, that is assuming that $\beta = 0$. GMMAT relies on PQL to estimate the fixed effects parameters $\boldsymbol{\alpha}$ and the average information REML algorithm (Harville [1977], Gilmour et al. [1995]) to estimate the variance components. Similarly to other LMMs, the fitted null model is the same for all genetic predictors in a GWAS. To test for association between each genetic variant and a binary trait, GMMAT applies a score test which is computationally fast and scalable for large-size GWASs. However, obtaining effect size estimates $\hat{\beta}$ for a large number of predictors is computationally demanding as it requires fitting one mixed model per tested predictor. Thus, the authors suggested a two-steps method where only the important predictors that passed a user-defined stringent significance threshold are estimated using a Wald test. Of note, the GMMAT assumes that score test statistics asymptotically follow a Gaussian distribution to estimate asymptotic

p-values. Zhou et al. [2018] argued that this can potentially lead to type I error rate inflation when case-control ratios are unbalanced, and proposed a Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) that relies on the saddlepoint approximation to calibrate unbalanced case-control ratios in score tests. They further proposed to estimate $\hat{\beta}$ using the variance components estimated under the null model of no association to avoid fitting a large number of mixed models. Their estimate for $\hat{\beta}$ is given by

$$\hat{\beta} = \left(\mathbf{G}^\intercal \hat{\mathbf{P}} \mathbf{G}\right)^{-1} \mathbf{G}^\intercal (\mathbf{y} - \hat{\mu}),$$

where $\hat{\mathbf{P}} = \hat{\mathbf{\Sigma}}^{-1} - \hat{\mathbf{\Sigma}}^{-1}\mathbf{X}(\mathbf{X}^\intercal \hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal \hat{\mathbf{\Sigma}}^{-1}$ is a projection matrix, $\hat{\mathbf{\Sigma}}^{-1} = \hat{\mathbf{W}}^{-1} + \sum_{k=1}^{K} \hat{\tau}_k \mathbf{V}_k$ is the variance-covariance matrix, and $\mathbf{W}$ is the diagonal matrix of GLM weights.

In the next section, I briefly review the PQL estimation method given it is an important component of the models that I propose in this thesis.

2.1.5 PQL estimation

Let $f(\cdot)$ be a probability density function, such that the marginal log-likelihood for the previous GLMM in (2.2) is

$$\begin{aligned}
\ell(\boldsymbol{\alpha}, \beta, \boldsymbol{\tau}) &= \log \int \prod_{i}^{n} f(y_i|\boldsymbol{\alpha}, \beta, \boldsymbol{\tau}, \boldsymbol{b}) f(\boldsymbol{b}|\boldsymbol{\tau}) d\boldsymbol{b} \\
&= \log \int \exp\left\{ \sum_{i=1}^{n} \log\, f(y_i|\boldsymbol{\alpha}, \beta, \boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^\intercal \left(\sum_{k=1}^{K} \tau_k \boldsymbol{V}_k \right)^{-1} \boldsymbol{b} \right\} \times \left| \sum_{k=1}^{K} \tau_k \boldsymbol{V}_k \right|^{-1/2} d\boldsymbol{b}.
\end{aligned}$$

MLE requires exact information about the data distribution through specification of the conditional densities $f(y_i|\boldsymbol{\alpha}, \beta, \mathbf{b})$. Moreover, except for normal responses with an identity link function, the marginal likelihood $\ell(\boldsymbol{\alpha}, \beta, \boldsymbol{\tau})$ has no analytic form. Let

$$h(\boldsymbol{b}) = \sum_{i=1}^{n} \log\, f(y_i|\boldsymbol{\alpha}, \beta, \boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^\intercal \left(\sum_{k=1}^{K} \tau_k \boldsymbol{V}_k \right)^{-1} \boldsymbol{b},$$

we can use Laplace method (Tierney and Kadane [1986]) to approximate the integral

$$\int \exp\{h(\boldsymbol{b})\}d\boldsymbol{b} \approx (2\pi)^{\frac{n}{2}} \left|-h''(\tilde{\boldsymbol{b}})\right|^{-\frac{1}{2}} \exp\{h(\tilde{\boldsymbol{b}})\},$$

such that the Laplace approximated log-likelihood is defined as

$$\ell_{LA}(\boldsymbol{\alpha}, \beta, \boldsymbol{\tau}) = \sum_{i=1}^{n} \log\, f(y_i|\boldsymbol{\alpha}, \beta, \tilde{\boldsymbol{b}}) - \frac{1}{2}\tilde{\boldsymbol{b}}^{\intercal}\left(\sum_{k=1}^{K}\tau_k \boldsymbol{V}_k\right)^{-1}\tilde{\boldsymbol{b}} - \frac{1}{2}\log\left|\sum_{k=1}^{K}\tau_k \boldsymbol{V}_k \boldsymbol{W} + \boldsymbol{I}\right|,$$

where $\tilde{\boldsymbol{b}}$ is the solution of $h'(\boldsymbol{b}) = 0$, $\boldsymbol{W}$ is a diagonal matrix with weights for each observation $w_i = \phi^{-1}\text{diag}\left\{\frac{a_i}{\nu(\mu_i)[g'(\mu_i)^2]}\right\}$, ϕ is the dispersion parameter, a_i are known weights, and $\nu(\mu_i)$ is the variance function. Assuming that the weights in $\boldsymbol{W}$ varies slowly with the conditional mean $\boldsymbol{\mu}$, Breslow and Clayton [1993] proposed maximizing instead a PQL function

$$\ell_{PQL}(\boldsymbol{\alpha}, \beta, \boldsymbol{\tau}) = \sum_{i=1}^{n} ql_i(\boldsymbol{\alpha}, \beta, \tilde{\boldsymbol{b}}) - \frac{1}{2}\tilde{\boldsymbol{b}}^{\intercal}\left(\sum_{k=1}^{K}\tau_k \boldsymbol{V}_k\right)^{-1}\tilde{\boldsymbol{b}},$$

where $ql_i(\boldsymbol{\alpha}, \beta, \tilde{\boldsymbol{b}}) = \int_{y_i}^{\mu_i} \frac{a_i(y_i-\mu)}{\phi\nu(\mu)}d\mu$ is the quasi-likelihood for the ith subject given the random intercept $\boldsymbol{b}$.

PQL is formally equivalent to replacing the observation vector $\boldsymbol{y}$ by a working vector $\tilde{\mathbf{Y}} = \boldsymbol{\eta} + \boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu})$ with $\boldsymbol{\Delta} = \text{diag}(g'(\mu_1), ..., g'(\mu_n))$ and assuming under the normal theory linear model that $\tilde{\mathbf{Y}} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N(0, \mathbf{W^{-1}})$. Moreover, PQL does not require specifying the exact distribution of the data, but only some information about the mean and variance relationship. PQL estimation can be justified by using a Taylor series expansion of the linear predictor $g(y_i)$ around the conditional mean of y_i (Schall [1991])

$$g(y_i) \approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i) = \eta_i + g'(\mu_i)(y_i - \mu_i) = \tilde{Y}_i.$$

2.1.6 GEI

Aside from population structure and closer relatedness, genetic association studies are also subject to unobserved confounding factors arising from heterogeneous environmental exposure. Moreover, interactions between genes and environmental factors may play a key role in the etiology of many common disorders that have both genetic and environmental risk factors. Ottman [1996] defined GEI as "a different effect of an environmental exposure on disease risk in persons with different genotypes", or alternatively, "a different effect of a genotype on disease risk in persons with different environmental exposures". Thus, we may hypothesize that the average effect of a treatment or any environmental risk factor on an individual may be modified by its genotype, or alternatively that the effect of a gene on an individual's phenotype may be altered by a treatment or environmental exposure.

Sul et al. [2016] studied the impact of population structure on GEI statistics in GWAS and proposed a statistical approach based on mixed models with an additional random effect that captures the similarity of individuals due to random polygenic GEI effects. Given the kinship matrix $\boldsymbol{K}$ and binary exposure D, they define the matrix $\boldsymbol{K}^D$ where each entry $K^D_{ij} = K_{ij}$ if $D_i = D_j$ and $K^D_{ij} = 0$ otherwise (Yang et al. [2011]). This derived kinship matrix $\boldsymbol{K}^D$ 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.

2.2 Regularized regression models

Multivariable regression methods, as opposed to single-SNP models, simultaneously fit many genetic variants as fixed effects in a single regression model. They have been proposed as an alternative method to increase the power for identifying weaker genome-wide associations compared to univariable methods (Li et al. [2010]). Moreover, contrary to single-SNP methods, a multivariable model can incorporate information about the dependence structure

between different variants, i.e. the nonrandom association of alleles of different loci referred to as linkage disequilibrium (LD) (Slatkin [2008]), potentially differentiating causative from spurious associations (Malo et al. [2008]). In both simulations and analysis of high dimensional data, that is where the number of predictors p is much greater than the sample size n, multivariable logistic models have shown to achieve lower false-positive rates and higher precision than methods based on univariable GWAS summary statistics in case-control studies (Qian et al. [2020], Privé et al. [2019]).

In high dimensional data, because the number of predictors exceeds the number of observations, regularization that imposes a penalty on the size of the predictors coefficients is required to avoid over-fitting and numerical instability of the standard regression solution. A common regularization penalty used in genetic studies is the lasso, for least absolute shrinkage and selection operator, which imposes a ℓ_1 penalty on the regression coefficients of the genetic predictors (Tibshirani [1996]). The popularity of the lasso is due to the fact that the constraint it imposes results in many coefficients being exactly equal to zero, leading to more interpretable models, especially in high dimensional settings. A very efficient implementation of the lasso can be found in the `glmnet` package in R (Friedman et al. [2010b]). The great computational efficiency of the lasso relies on three key concepts, cyclic coordinate gradient descent (CGD) to minimize the loss function with respect to one predictor at a time (Wu and Lange [2008]), pathwise solutions and strong rules for discarding the vast majority of predictors from each iteration (Tibshirani et al. [2012]).

In this section, I first present the lasso regularized multivariable genetic linear model for continuous phenotypes. Second, I detail how pathwise solutions and strong rules help reduce the computational burden associated with fitting high dimensional regularized models. Third, I present the proximal Newton algorithm (Lee et al. [2014]) that was proposed by Friedman et al. [2010b] to obtain estimates of the fixed effects predictors for the logistic lasso model. Fourth, the adaptive lasso (Zou [2006]) and elastic-net (Zou and Hastie [2005]) penalties are

presented. I finish this section by reviewing the literature behind the group lasso and sparse group lasso models.

2.2.1 Lasso regularization

Consider the following multivariable genetic linear model, where for simplicity we include no covariate except for an intercept term α,

$$y_i = \alpha + \mathbf{G}_i^\intercal \boldsymbol{\beta} + e_i. \tag{2.3}$$

The lasso solves the minimization problem

$$\min_{\alpha,\boldsymbol{\beta}} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \alpha - \mathbf{G}_i^\intercal\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where λ is a tuning parameter that controls the sparsity of the solution. Assume we have initial estimates for the intercept $\tilde{\alpha}$ and for genetic predictors fixed effects $\tilde{\beta}_l$ for $l \neq j$ and $j = 1, ..., p$. A coordinate descent step involves computing the subgradient at $\beta_j = \tilde{\beta}_j$ such that the stationary condition is given by

$$0 = -\frac{1}{n}\sum_{i=1}^{n} G_{ij}(y_i - \tilde{\alpha} - \mathbf{G}_i^\intercal\tilde{\boldsymbol{\beta}}) + \lambda u,$$

where we define the subgradient $u \in \partial|\tilde{\beta}_j| = \begin{cases} [-1, 1] & \text{if } \tilde{\beta}_j = 0, \\ \text{sign}(\tilde{\beta}_j) & \text{if } \tilde{\beta}_j \neq 0. \end{cases}$

Let $r_i^{(j)} = y_i - \tilde{\alpha} - \sum_{l\neq j} G_{il}\tilde{\beta}_l$ be the partial residual when solving for β_j, then it is simple to show that the coordinate step solution is

$$\tilde{\beta}_j = S_\lambda\left(\frac{1}{n}\sum_{i=1}^{n} G_{ij}r_i^{(j)}\right),$$

where $S_\lambda(\cdot)$ is the soft-tresholding function defined as

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

Let $\hat{y}_i = \tilde{\alpha} + \sum_{l=1}^{p} G_{il}\tilde{\beta}_l$ be the current fit of the model for the i^{th} observation, and $r_i = y_i - \hat{y}_i$ the current residual. Assuming $\mathbf{G}_j$ is standardized such that $\sum_{i=1}^{n} G_{ij}^2 = n$, we can rewrite the coordinate descent update as

$$\tilde{\beta}_j^{\,(t+1)} = S_\lambda\left(\frac{1}{n}\sum_{i=1}^{n} G_{ij} r_i^{(t)} + \tilde{\beta}_j^{\,(t)}\right). \tag{2.4}$$

2.2.2 Pathwise solution and strong rules

Suppose that the current estimates $\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_p$ are all equal to 0 for $\lambda = \lambda_1$, and that we have an initial estimate $\tilde{\alpha}_0$ of the intercept, then from (2.4) we must have that for all j

$$\frac{1}{n}\left|\sum_{i=1}^{n} G_{ij}(y_i - \tilde{\alpha}_0)\right| \leq \lambda_1.$$

Thus, we can create a path of decreasing values of the tuning parameter $\lambda_1 > \lambda_2 > ... > \lambda_m$ such that at $\lambda_1 = \max_j \frac{1}{n}\left|\sum_{i=1}^{n} G_{ij}(y_i - \tilde{\alpha}_0)\right|$ all coefficients are equal to 0. This procedure also exploits *warm starts*, that is the solution $\tilde{\boldsymbol{\beta}}(\lambda_{k-1})$ at λ_{k-1} is used as the initial estimate for minimizing the loss function at λ_k.

When the number of genetic predictors is very large, assuming that most of their effects are equal to 0, it is desirable to discard them from the coordinate descent steps to speedup the optimization procedure. Tibshirani et al. [2012] derived sequential strong rules that can be used when solving lasso-type problems over a grid of tuning parameter values. There-

fore, having already computed the solutions $\tilde{\alpha}(\lambda_{k-1})$ and $\tilde{\boldsymbol{\beta}}(\lambda_{k-1})$, the sequential strong rule discards the j^{th} genetic predictor from the optimization problem at λ_k if

$$|\mathbf{G}_j^\intercal(\mathbf{y} - \mathbf{1}\tilde{\alpha}(\lambda_{k-1}) - \mathbf{G}\tilde{\boldsymbol{\beta}}(\lambda_{k-1}))| < 2\lambda_k - \lambda_{k-1}. \tag{2.5}$$

Thus, we define the strong set $\mathcal{S}(\lambda_k)$ as the indices of the predictors that survive the screening rule (2.5). Once all coefficients in the strong set has converged, we check the stationary condition $|\frac{1}{n}\sum_{i=1}^n G_{ij}r_i| < \lambda_k$ for all remaining predictors. If any predictor violates the stationary condition, then we add them to $\mathcal{S}(\lambda_k)$ and rerun the coordinate descent algorithm on $\mathcal{S}(\lambda_k)$ only.

2.2.3 Logistic lasso

Consider the following multivariable genetic logistic model, where for simplicity we again include no covariate except for an intercept term α,

$$\text{logit}(\mu_i) = \alpha + \sum_{j=1}^{p} G_{ij}\beta_j,$$

where $\mu_i = P(y_i = 1|\mathbf{G}_i)$. The lasso solves the minimization problem

$$\min_{\alpha,\boldsymbol{\beta}} \left\{ -\left[\frac{1}{n}\sum_{i=1}^{n} y_i(\alpha + \mathbf{G}_i^\intercal\boldsymbol{\beta}) - \log(1 + e^{\alpha+\mathbf{G}_i^\intercal\boldsymbol{\beta}})\right] + \lambda\sum_{j=1}^{p}|\beta_j| \right\}. \tag{2.6}$$

Clearly, taking the derivative of the regularized loss function (2.6) and setting it to zero does not provide any closed form solution for $(\tilde{\alpha}, \tilde{\boldsymbol{\beta}})$. Since the objective function in (2.6) consists of a smooth convex function $f(\alpha, \boldsymbol{\beta}) := -\left[\frac{1}{n}\sum_{i=1}^n y_i(\alpha + \mathbf{G}_i^\intercal\boldsymbol{\beta}) - \log(1 + e^{\alpha+\mathbf{G}_i^\intercal\boldsymbol{\beta}})\right]$ and a non-smooth convex regularizer $g(\boldsymbol{\beta}) := \lambda\|\boldsymbol{\beta}\|_1$, Friedman et al. [2010b] proposed a proximal Newton algorithm (Lee et al. [2014]) with cyclic coordinate descent to find solutions to the minimization problem (2.6). This consists in replacing $f(\alpha, \boldsymbol{\beta})$ in (2.6) by its second

order Taylor expansion about the current iterate $\boldsymbol{\Theta}^{(t)} = (\alpha^{(t)}, \boldsymbol{\beta}^{(t)})$ while maintaining the regularization function $g(\boldsymbol{\beta})$ unchanged. Let $\boldsymbol{X\Theta} = \mathbf{1}\alpha + \mathbf{G}\boldsymbol{\beta}$, the iterative step reduces to

$$\begin{aligned}\boldsymbol{\Theta}^{(t+1)} &= \underset{\boldsymbol{\Theta}}{\text{argmin}} \left\{ \frac{1}{2s_t} \left\| \boldsymbol{\Theta} - \left(\boldsymbol{\Theta}^{(t)} - s_t \left[\nabla^2_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}^{(t)}) \right]^{-1} \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}^{(t)}) \right) \right\|_2^2 + g(\boldsymbol{\Theta}) \right\} \\ &= \underset{\boldsymbol{\Theta}}{\text{argmin}} \left\{ \frac{1}{2s_t} \left\| \boldsymbol{\Theta} - \left[\boldsymbol{X}^\intercal \mathbf{W}^{(t)} \boldsymbol{X} \right]^{-1} \boldsymbol{X}^\intercal \mathbf{W}^{(t)} \left(\boldsymbol{X}\boldsymbol{\Theta}^{(t)} + s_t \mathbf{W}^{-(t)} (\boldsymbol{y} - \boldsymbol{\mu}^{(t)}) \right) \right\|_2^2 + g(\boldsymbol{\Theta}) \right\},\end{aligned}$$

where s_t is a suitable step size and $\mathbf{W}^{(t)}$ is a diagonal matrix of weights, with the i^{th} element being equal to $\mu_i^{(t)}(1-\mu_i^{(t)})$. Defining the working response vector $\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\Theta}^{(t)} + s_t \mathbf{W}^{-(t)}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)})$, we can rewrite the minimization problem as a weighted least squares (WLS) problem where

$$\begin{aligned}\boldsymbol{\Theta}^{(t+1)} &= \underset{\boldsymbol{\Theta}}{\text{argmin}} \left\{ \frac{1}{2s_t} \left\| \boldsymbol{\Theta} - \left[\boldsymbol{X}^\intercal \mathbf{W}^{(t)} \boldsymbol{X} \right]^{-1} \boldsymbol{X}^\intercal \mathbf{W}^{(t)} \tilde{\boldsymbol{Y}} \right\|_2^2 + g(\boldsymbol{\Theta}) \right\} \\ &= \underset{\boldsymbol{\Theta}}{\text{argmin}} \left\{ \frac{1}{2s_t} \sum_{i=1}^{n} w_i \left(\tilde{Y}_i - \boldsymbol{X}_i \boldsymbol{\Theta} \right)^2 + \lambda \sum_j |\beta_j| \right\},\end{aligned}$$

where $w_i = \mu_i^{(t)}(1-\mu_i^{(t)})$. Thus, for a fixed vector of weighs $(w_1, ..., w_n)$, the update step is simply given by

$$\tilde{\beta}_j = \frac{S_\lambda(\sum_{i=1}^n w_i G_{ij} r_i^{(j)})}{\sum_{i=1}^n w_i G_{ij}^2}$$

with $r_i^{(j)} = \tilde{Y}_i - \tilde{\alpha} - \sum_{l \neq j} G_{il}\tilde{\beta}_l$ the working partial residual when solving for β_j. Compared to the lasso for the linear model, this algorithm requires iteratively fitting WLS as an inner loop combined with an outer quadratic approximation around the current iterates. Although the Newton algorithm is not guaranteed to converge with a fixed step size $s_t = 1$, the fact that the quadratic approximation is always *warm-started* from the previous iterative solution makes it very accurate in practice (Friedman et al. [2010b]).

2.2.4 Adaptative lasso and elastic-net

The ℓ_1 penalty imposed by the lasso produces biased estimates for large coefficients. Indeed, in the orthonormal case, one can show that

$$\hat{\beta}_j^{lasso} = S_\lambda(\hat{\beta}_j^{OLS}) = \text{sign}(\hat{\beta}_j^{OLS})(|\hat{\beta}_j^{OLS}| - \lambda)_+,$$

where $\hat{\beta}_j^{OLS}$ is the ordinary least squares (OLS) estimator and $(\cdot)_+$ is equal to zero if not positive. Fan and Li [2001] conjectured that because of this resulting bias the oracle properties do not hold for the lasso. They proposed a smoothly clipped absolute deviation (SCAD) penalty for variable selection and proved its oracle properties. Alternatively, Zou [2006] proposed assigning different weights to different coefficients through a weighted lasso, and showed that if the choice of the weights is informed by the data, then the weighted lasso can achieve the oracle properties. They proposed the adaptive lasso methodology which solves the minimization problem

$$\min_{\alpha,\boldsymbol{\beta}} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \alpha - \mathbf{G}_i^\intercal\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j|,$$

where the weights w_j for $j = 1, ..., p$ are defined by $\hat{w}_j = 1/|\hat{\beta}_j|^\gamma$ and $\hat{\beta}_j$ is a root-n-consistent estimator of β_j, for example the OLS estimator, and $\gamma > 0$. In the high-dimensional setting, where the number of predictors diverges with the sample size, as it is the case with genetic association studies, the OLS estimator is no longer applicable. A practical solution is to use the ℓ_2 penalized estimator for constructing the weights in the adaptive lasso (Zou [2006]). Moreover, in the case where there is strong collinearity between the predictors, as with SNPs in LD blocks, the performance of the lasso in retrieving important predictors breaks down (Zou and Hastie [2005]). Indeed, when there is strong pairwise correlations among a group of predictors, the lasso tends to select only one variable from the group randomly. To address this challenge, Zou and Hastie [2005] proposed a new regularization and variable

selection technique called the elastic-net which is a convex combination of the lasso and ridge (Hoerl and Kennard [1970]) penalty on the regression coefficients. For the model in (2.3), the elastic-net estimator is defined as

$$\hat{\boldsymbol{\beta}}^{enet} = (1 + \frac{\lambda_2}{n}) \left\{ \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \alpha - \mathbf{G}_i^\intercal \boldsymbol{\beta})^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2.7}$$

The elastic-net method produces a sparse model with good prediction accuracy, while encouraging a grouping effect. However, it lacks the oracle property of the adaptative lasso. Thus, similarly to the adaptive lasso methodology, the adaptive elastic-net (Zou and Zhang [2009]) incorporates cleverly chosen penalization weights to achieve both the oracle property and good selection performance when there is collinearity between predictors. Zou and Zhang [2009] proposed to solve the following minimization problem

$$(1 + \frac{\lambda_2}{n}) \left\{ \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \alpha - \mathbf{G}_i^\intercal \boldsymbol{\beta})^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1^* \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right\},$$

where the adaptive weights w_j for $j = 1, ..., p$ are equal to $\hat{w}_j = 1/|\hat{\beta}_j^{enet}|^\gamma$. For $\lambda_2 = 0$, the adaptive elastic-net reduces to the adaptive lasso. Moreover, in the orthogonal case, the adaptive elastic-net also reduces to the adaptive lasso for any value of λ_2.

2.2.5 Group lasso

It is often the case that predictors are organized into natural groups, for example SNPs located on the same genes, genes belonging to the same pathways, or different levels from a categorical predictor. Thus, instead of focusing on selection of individual predictors, we may be interested in selecting groups of predictors all together. The group lasso (Yuan and Lin [2005], Simon and Tibshirani [2012]) applied to model (2.3) minimizes the convex loss

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \alpha - \mathbf{G}_i^\intercal \boldsymbol{\beta})^2 + \lambda \sum_{k=1}^{L} \sqrt{p_k} \|(\beta_{k1}, ..., \beta_{kp_k})\|_2,$$

where p_k is the size of the k^{th} group of predictors. The non-differentiability of the ℓ_2 norm at 0 ensures that some groups will have all their coefficients to be exactly 0. Indeed, for the ℓ^{th} group of predictors $\mathbf{G}^{(\ell)}$, $\tilde{\boldsymbol{\beta}}_\ell = (\tilde{\beta}_{\ell 1}, ..., \tilde{\beta}_{\ell p_\ell})$ must satisfy the stationary condition

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{G}_i^{(\ell)}(y_i - \alpha - \sum_{k=1}^{L} \mathbf{G}_i^{(k)\intercal}\boldsymbol{\beta}_k) = \lambda\sqrt{p_\ell}\mathbf{v},$$

where $\mathbf{v}$ is the subgradient of $\|\boldsymbol{\beta}_\ell\|_2$ evaluated at $\tilde{\boldsymbol{\beta}}_\ell$ which is given by

$$\mathbf{v} = \begin{cases} \{\mathbf{v} |\ \|\mathbf{v}\|_2 \leq 1\} & \text{if } \tilde{\boldsymbol{\beta}}_\ell = 0, \\ \tilde{\boldsymbol{\beta}}_\ell / \|\tilde{\boldsymbol{\beta}}_\ell\|_2 & \text{if } \tilde{\boldsymbol{\beta}}_\ell \neq 0. \end{cases}$$

For $\tilde{\boldsymbol{\beta}}_\ell = 0$ to be a solution, the constraint $\|\mathbf{v}\|_2 \leq 1$ implies that

$$\|\mathbf{G}^{(\ell)\intercal}(\mathbf{y} - \mathbf{1}\alpha - \sum_{k\neq\ell}^{L} \mathbf{G}^{(k)}\tilde{\boldsymbol{\beta}}_\ell)\|_2 \leq \lambda\sqrt{p_\ell}.$$

For $\tilde{\boldsymbol{\beta}}_\ell \neq 0$, solving the stationary condition yields

$$\tilde{\boldsymbol{\beta}}_\ell = \left(\sum_{i=1}^{n} \mathbf{G}_i^{(\ell)}\mathbf{G}_i^{(\ell)\intercal}/n + \lambda\sqrt{p_\ell}/\|\tilde{\boldsymbol{\beta}}_\ell\|_2 \mathbf{I}_{p_\ell}\right)^{-1} \sum_{i=1}^{n} \mathbf{G}_i^{(\ell)} r_i^{(\ell)}/n$$

which is a function of the optimal solution $\lambda\sqrt{p_\ell}/\|\tilde{\boldsymbol{\beta}}_\ell\|_2$, with $r_i^{(\ell)} = y_i - \tilde{\alpha} - \sum_{k\neq\ell}^{L} \mathbf{G}_i^{(k)\intercal}\tilde{\boldsymbol{\beta}}_k$ the partial residual of y_i when subtracting all group fits other than the ℓ^{th} group of predictors.

Assuming that $\sum_{i=1}^{n} \mathbf{G}_i^{(\ell)}\mathbf{G}_i^{(\ell)\intercal}/n = \mathbf{I}_{p_\ell}$, the solution simplifies to (Yuan and Lin [2005])

$$\tilde{\boldsymbol{\beta}}_\ell = \left(1 - \frac{\lambda\sqrt{p_\ell}}{\|\sum_{i=1}^{n} \mathbf{G}_i^{(\ell)} r_i^{(\ell)}/n\|_2}\right) \sum_{i=1}^{n} \mathbf{G}_i^{(\ell)} r_i^{(\ell)}/n.$$

In the case where the columns of $\mathbf{G}^{(\ell)}$ are not orthogonal, one may be tempted to orthogonal-

ize them before applying the group lasso. However, Friedman et al. [2010a] showed that the resulting solution, when transformed back to the original basis for each group, will not generally provide a solution to the group lasso problem with the original covariates. Meier et al. [2008] implemented a block coordinate gradient descent (BCGD) that combines a quadratic approximation of the log-likelihood via a quasi-Newton method with an additional Armijo line search. Alternatively, Kim et al. [2006] proposed updating the entire coefficient vector simultaneously at each step by bounding the sum of the groups' norms instead of using a penalty. Their method relies on projecting the solution vector of the unpenalized loss function to the closest vector satisfying the bound condition on the sum of group norms. Finally, Foygel and Drton [2010] fit the group lasso by finding the exact optimal value for each group block-wise with one univariate line search using the spectral decomposition of $\mathbf{G}^{(\ell)\intercal}\mathbf{G}^{(\ell)}$.

2.2.6 Sparse group lasso

While the group lasso effectively allows to perform selection of groups of predictors, it may select too many false positives in the final model due to the fact that once a predictor from a group enters the model, it favours other predictors from the same group to also be selected even though they may not have any direct effect on the response. The sparse group lasso is a natural extension of the group lasso formulation that allows both sparsity between groups and within each group of predictors by adding a ℓ_1 penalty on the coefficients of the predictors (Wu and Lange [2008], Foygel and Drton [2010], Friedman et al. [2010a], Zhou et al. [2010], Simon and Tibshirani [2012]). The sparse group lasso for model (2.3) minimizes the convex loss function given by

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \alpha - \mathbf{G}_i^{\intercal}\boldsymbol{\beta})^2 + (1-\rho)\lambda\sum_{k=1}^{L}\sqrt{p_k}\|(\beta_{k1},...,\beta_{kp_k})\|_2 + \rho\lambda\|\boldsymbol{\beta}\|_1$$

with $\rho \in [0, 1]$ an additional tuning parameter that controls the relative strength of the ℓ_2 and ℓ_1 penalties. Setting $\rho = 0$ is equivalent to a group lasso problem, while $\rho = 1$ gives a lasso problem. Thus, the sparse group lasso can be seen as a linear combination of the two penalties.

The stationary condition for the ℓ^{th} group of predictors $\mathbf{G}^{(\ell)}$ must satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{G}_i^{(\ell)}(y_i - \alpha - \sum_{k=1}^{L}\mathbf{G}_i^{(k)\mathsf{T}}\boldsymbol{\beta}_k) = (1-\rho)\lambda\sqrt{p_\ell}\mathbf{v} + \rho\lambda\mathbf{u},$$

where again $\mathbf{u}$ and and $\mathbf{v}$ are respectively the subgradients of $\|\boldsymbol{\beta}_\ell\|_1$ and $\|\boldsymbol{\beta}_\ell\|_2$ evaluated at $\tilde{\boldsymbol{\beta}}_\ell$. It is easy to show that $\tilde{\boldsymbol{\beta}}_\ell = 0$ satisfies the subgradient equations if

$$\|S_{\rho\lambda}(\sum_{i=1}^{n}\mathbf{G}_i^{(\ell)}r_i^{(\ell)}/n)\|_2 \leq (1-\rho)\lambda\sqrt{p_\ell},$$

where $r_i^{(\ell)}$ is the partial residual of y_i, and the soft-thresholding operator is applied coordinate-wise. Hence, the sparse group lasso applies univariate soft-thresholding before applying shrinkage to each group.

If $\boldsymbol{\beta}_\ell \neq 0$, then the subgradient equations for the j^{th} predictor $\beta_{\ell j}$ or group ℓ is given by

$$\frac{1}{n}\sum_{i=1}^{n}G_i^{(\ell j)}(y_i - \alpha - \sum_{k=1}^{L}\mathbf{G}_i^{(k)\mathsf{T}}\boldsymbol{\beta}_k) = (1-\rho)\lambda\sqrt{p_\ell}\frac{\beta_{\ell j}}{\|\boldsymbol{\beta}_\ell\|_2} + \rho\lambda u_j.$$

Again, since $\beta_{\ell j} = 0$ implies $|u_j| < 1$, we have that the stationary condition is satisfied for $\beta_{\ell j} = 0$ whenever

$$\left|\sum_{i=1}^{n}G_i^{(\ell j)}r_i^{(\ell j)}\right| \leq n\rho\lambda$$

with $r_i^{(\ell j)} = r_i^{(\ell)} - \sum_{k\neq j}G_i^{(\ell k)\mathsf{T}}\beta_{\ell k}$ the partial residual when fitting all other predictors than $\beta_{\ell j}$.

For $\beta_{\ell j}$ different from 0, the solution is

$$\tilde{\beta}_{\ell j} = \frac{S_{\rho\lambda}(\sum_{i=1}^{n} G_i^{(\ell j)} r_i^{(\ell j)}/n)}{\sum_{i=1}^{n} {G_i^{(\ell j)}}^2/n + (1-\rho)\lambda\sqrt{p_\ell}/\|\tilde{\boldsymbol{\beta}}_\ell\|_2}$$

which depends on the optimal solution $\|\tilde{\boldsymbol{\beta}}_\ell\|_2$. Friedman et al. [2010a] proposed to fit the sparse-group lasso using blockwise descent and an accelerated generalized gradient algorithm with backtracking line search to solve within each group. Alternatively, Wu and Lange [2008] proposed majorizing the group ℓ_2 penalty of the loss function using the concavity of the square root function. This approach yields an elastic-net solution to the problem, that is a combination of a ridge (Hoerl and Kennard [1970]) and a lasso penalty as first introduced by Zou and Hastie [2005]. The authors note that although convergence may be slowed by majorization, the descent property of the majorization-minimization (MM) algorithm (Lange et al. [2000]) ensures that minimizing the surrogate function minimizes in turn the loss function.

2.3 Sparse regularized mixed models

To account for the correlation between observations that may arise from the sampling design or longitudinal nature of the data, a common and flexible approach is to fit sparse regularized multivariable mixed models. Schelldorfer et al. [2011] first proposed a CGD algorithm for ℓ_1 penalized high dimensional LMMs which they proved to converge numerically to a stationary point of the loss function, albeit with no guarantee to converge to the global optimum due to non-convexity of the negative log-likelihood function. Alternatively, Bondell et al. [2010] considered the simultaneous selection of fixed and random effects for low dimensional LMMs. Ibrahim et al. [2011] and Ghosh and Thoresen [2017] further extended the framework to other penalties than the lasso, such as the SCAD and the adaptive least absolute shrinkage and selection operator (ALASSO) penalty functions. For GLMMs, Schelldorfer et al. [2014] considered joint estimation of random effects and sparse fixed effects using a Laplace

approximation of the log-likelihood. They proposed a two steps method in which they first perform variable screening to reduce the dimensionality of the model, and in a second step refit the model by maximum likelihood (ML) estimation to get accurate parameter estimates. On the other hand, Groll and Tutz [2012] and Hui et al. [2017] considered joint selection of mixed effects using regularized PQL. More recently, M. Heiling et al. [2024] proposed an R package called `glmmPen` to jointly select fixed and random effects in GLMMs using a Monte Carlo expectation conditional minimization (MCECM) algorithm.

In genetic association studies, the number of important random effects is often low dimensional and pre-specified. Indeed, it is common to only include a random polygenic effect with variance-covariance structure proportional to a known GRM. Thus, we are mainly interested in performing selection of important predictors while jointly estimating the variance components. However, it is computationally challenging to jointly estimating the variance components with the regression fixed effects vector since the penalized negative log-likelihood function is non-convex with respect to the variance components. For this reason, to obtain a scalable algorithm, Rakitsch et al. [2012] proposed estimating first the variance components assuming no SNP main effects in a null model, and in the second step, using the residuals from the null model as the response in a high dimensional linear model that assumes uncorrelated errors. As an alternative to the two-steps approach, Bhatnagar et al. [2020b] developed a BCD algorithm to simultaneously select predictors and estimate their effects in LMMs, accounting for between-individual correlations, while jointly estimating the variance components parameters. I briefly review below the model from Bhatnagar et al. [2020b].

Consider the following LMM

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{G}\boldsymbol{\beta} + \mathbf{b} + \mathbf{e},$$

with $\mathbf{b} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi})$ and $\mathbf{e} \sim \mathcal{N}(0, (1-\eta)\sigma^2\mathbf{I})$. The scalar parameter η represents the

narrow-sense heritability, that is the proportion of phenotypic variance explained by additive genetic effects (Manolio et al. [2009]). Defining the complete parameter vector as $\boldsymbol{\Theta} := (\boldsymbol{\alpha}, \boldsymbol{\beta}, \eta, \sigma^2)$, the negative log-likelihood for the previous model is given by

$$-\ell(\boldsymbol{\Theta}) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log(\det(\mathbf{V})) + \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{G}\boldsymbol{\beta})^\intercal \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{G}\boldsymbol{\beta}),$$

where $\mathbf{V} = \eta\boldsymbol{\Phi} + (1-\eta)\mathbf{I}$.

Let $\boldsymbol{\Phi} = \mathbf{U}\mathbf{D}\mathbf{U}^\intercal$ be the spectral decomposition of the kinship or GRM matrix, then we have

$$\log(\det(\mathbf{V})) = \sum_{i=1}^{n} \log(1 + \eta(\Lambda_i - 1)),$$

with $\Lambda_1 \geq \Lambda_2 \geq ... \geq 0$ the (positive) eigenvalues of $\boldsymbol{\Phi}$. Further, the inverse of $\mathbf{V}$ can be expressed as

$$\mathbf{V}^{-1} = \mathbf{U}\tilde{\mathbf{D}}^{-\mathbf{1}}\mathbf{U}^\intercal,$$

where $\tilde{\mathbf{D}} = \eta\mathbf{D} + (1-\eta)\mathbf{I}$ is diagonal. Using the previous two equations, the negative log-likelihood becomes

$$-\ell(\boldsymbol{\Theta}) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{n} \log(1 + \eta(\Lambda_i - 1)) + \frac{1}{2\sigma^2}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\alpha} - \tilde{\mathbf{G}}\boldsymbol{\beta})^\intercal \tilde{\mathbf{D}}^{-\mathbf{1}}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\alpha} - \tilde{\mathbf{G}}\boldsymbol{\beta}),$$

where $\tilde{\mathbf{y}} = \mathbf{U}^\intercal\mathbf{y}$ is the rotated response vector, and $\tilde{\mathbf{X}} = \mathbf{U}^\intercal\mathbf{X}$, $\tilde{\mathbf{G}} = \mathbf{U}^\intercal\mathbf{G}$ are respectively the rotated covariates and genetic predictors matrices. Thus, the spectral decomposition of the matrix $\boldsymbol{\Phi}$ results in a diagonal covariance matrix making the negative log-likehood a function of the WLS function

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n} \frac{(\tilde{y}_i - \tilde{\mathbf{X}}_i^\intercal\boldsymbol{\alpha} - \tilde{\mathbf{G}}_i^\intercal\boldsymbol{\beta})^2}{1 + \eta(\Lambda_i - 1)}.$$

Adding a lasso regularization term to the negative log-likelihood function, the objective function to minimize is defined as

$$Q_\lambda(\boldsymbol{\Theta}) :=$$
$$\frac{n}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{n}\log(1+\eta(\Lambda_i-1)) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(\tilde{y}_i - \tilde{\mathbf{X}}_i^\intercal\boldsymbol{\alpha} - \tilde{\mathbf{G}}_i^\intercal\boldsymbol{\beta})^2}{1+\eta(\Lambda_i-1)} + \lambda\sum_{j=1}^{k}\nu_j^\alpha|\alpha_j| + \lambda\sum_{j=1}^{p}\nu_j^\beta|\beta_j|,$$

where ν_j^α and ν_j^β are regularization weights to achieve different penalties for each parameter. Bhatnagar et al. [2020b] proposed a general purpose block CGD algorithm to estimate $\boldsymbol{\Theta}$ by minimizing $Q_\lambda(\boldsymbol{\Theta})$ with respect to one parameter at a time while holding all others fixed. Through simulation studies, Reisetter and Breheny [2021] showed that estimating the variance components only once under the null as proposed by Rakitsch et al. [2012] performed similarly in terms of estimating the SNPs coefficients than by including the variance components in the iterative procedure as in Bhatnagar et al. [2020b], while showing much greater computational efficiency and numerical stability.

Chapter 3

Efficient Penalized Generalized Linear Mixed Models for Variable Selection and Genetic Risk Prediction in High-Dimensional Data

Preamble to Manuscript 1.

Rakitsch et al. [2012] and Bhatnagar et al. [2020b] have proposed efficient algorithms to fit penalized LMMs to high-dimensional genetic data. These methods are typically restricted to linear models since in GLMMs, it is no longer possible to perform a single spectral decomposition to rotate the phenotype and design matrix, as the covariance matrix depends on the sample weights which in turn depend on the estimated regression coefficients that are being iteratively updated. This limits the application of high-dimensional MMs to analysis of binary traits in genetic association studies. This gap motivated the methodological development proposed in this manuscript.

The original contributions of this chapter are i) developing a scalable algorithm based

on regularized PQL estimation which makes it possible to fit penalized GLMMs on high-dimensional GWAS of binary traits, ii) developing an open source `Julia` programming language package called `PenalizedGLMM.jl`, and iii) showing through simulation studies that our proposed method has higher precision and better prediction accuracy than a penalized LMM and logistic lasso with PC adjustment when the number of subpopulations is greater than the number of PCs included in the model, or when there is genetic relatedness between individuals.

The corresponding manuscript was published in *Bioinformatics* (St-Pierre et al. [2023]).

Efficient Penalized Generalized Linear Mixed Models for Variable Selection and Genetic Risk Prediction in High-Dimensional Data

Julien St-Pierre[1], Karim Oualkacha[2], Sahir Rai Bhatnagar[1].

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

[2]*Département de Mathématiques, Université du Québec á Montréal*

This thesis contains the accepted version of the corresponding paper published in *Bioinformatics* (St-Pierre et al. [2023]).

Abstract

Motivation: Sparse regularized regression methods are now widely used in genome-wide association studies (GWAS) to address the multiple testing burden that limits discovery of potentially important predictors. Linear mixed models (LMMs) have become an attractive alternative to principal components (PC) adjustment to account for population structure and relatedness in high-dimensional penalized models. However, their use in binary trait GWAS rely on the invalid assumption that the residual variance does not depend on the estimated regression coefficients. Moreover, LMMs use a single spectral decomposition of the covariance matrix of the responses, which is no longer possible in generalized linear mixed models (GLMMs).

Results: We introduce a new method called `pglmm`, a penalized GLMM that allows to simultaneously select genetic markers and estimate their effects, accounting for between-individual correlations and binary nature of the trait. We develop a computationally efficient algorithm based on penalized quasi-likelihood (PQL) estimation that allows to scale regularized mixed models on high-dimensional binary trait GWAS. We show through simulations that when the dimensionality of the relatedness matrix is high, penalized LMM and logistic regression with PC adjustment fail to select important predictors, and have inferior prediction accuracy compared to `pglmm`. Further, we demonstrate through the analysis of two polygenic binary traits in a subset of 6731 related individuals from the UK Biobank data with 320K SNPs that our method can achieve higher predictive performance, while also selecting fewer predictors than a sparse regularized logistic lasso with PC adjustment.

Availability and implementation: Our Julia package `PenalizedGLMM.jl` is publicly available on github : https://github.com/julstpierre/PenalizedGLMM.

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

Supplementary information: Supplementary data are available at *Bioinformatics* online.

3.1 Introduction

Genome-wide association studies (GWAS) have led to the identification of hundreds of common genetic variants, or single nucleotide polymorphisms (SNPs), associated with complex traits (Visscher et al. [2017]) 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. Moreover, GWAS 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 (Manolio et al. [2009]). Multivariable regression methods, on the other hand, simultaneously fit many SNPs in a single model and have been proposed to increase the power for identifying weaker associations compared to univariable methods (Li et al. [2010]). Moreover, sparse regularized multivariable regression models, which can perform variable selection, are exempt from the multiple testing burden.

Principal component analysis (PCA) can control for the confounding effect due to population structure by including the top eigenvectors of a genetic similarity matrix (GSM) as fixed effects in the regression model (Price et al. [2006]). Alternatively, using mixed models (MMs), one can model population structure and/or closer relatedness by including a polygenic random effect with variance-covariance structure proportional to the GSM (Yu et al. [2005]). Hence, while both PCA and 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 (Price et al. [2010]). Moreover, MMs alleviate the need to evaluate the optimal number of PCs to retain in the model as fixed effects.

Several authors have proposed to combine penalized quasi-likelihood (PQL) estimation with sparsity inducing regularization to perform selection of fixed and/or random effects in generalized linear mixed model (GLMMs) (Groll and Tutz [2012], Hui et al. [2017]). However, none of these methods are currently scalable for modern large-scale genome-wide data, nor can

they directly incorporate relatedness structure through the use of a kinship matrix. Indeed, the computational efficiency of recent multivariable methods for high-dimensional MMs rely on performing a spectral decomposition of the covariance matrix to rotate the phenotype and design matrix such that the transformed data become uncorrelated (Bhatnagar et al. [2020b], Rakitsch et al. [2012]). These methods are typically restricted to linear models since in GLMMs, it is no longer possible to perform a single spectral decomposition to rotate the phenotype and design matrix, as the covariance matrix depends on the sample weights which in turn depend on the estimated regression coefficients that are being iteratively updated. This limits the application of high-dimensional MMs to analysis of binary traits in genetic association studies.

In this paper, we introduce a new method called `pglmm` that allows to simultaneously select variables and estimate their effects, accounting for between-individual correlations and binary nature of the trait. We develop a scalable algorithm based on PQL estimation which makes it possible to fit penalized GLMMs on high-dimensional GWAS of binary traits. To speedup computation, we estimate the variance components and dispersion parameter of the model under the null hypothesis of no genetic effect. Secondly, we use an upper-bound for the inverse variance-covariance matrix in order to perform a single spectral decomposition of the GSM and greatly reduce memory usage. Finally, we implement an efficient cyclic coordinate descent algorithm in order to find the optimal estimates for the fixed and random effects parameters. Our method is implemented in an open source `Julia` programming language (Bezanson et al. [2017]) package called `PenalizedGLMM.jl` and freely available at `https://github.com/julstpierre/PenalizedGLMM`.

The rest of this paper is structured as follows. In Section 3.2 we present our model, describe the cyclic coordinate descent algorithm that is used to estimate the parameters and detail how predictions are obtained in GLMs with PC adjustment versus our proposed mixed model. In Section 3.3, we show through simulations that both LMM and logistic model

with PC adjustment fail to correctly select important predictors and estimate their effects when the dimensionality of the kinship matrix is high. Further, we demonstrate through the analysis of two polygenic binary traits in a subset of 6731 related individuals from the UK Biobank data that our method achieves higher predictive performance, while also selecting consistently fewer predictors than a logistic lasso with PC adjustment. We finish with a discussion of our results, some limitations and future directions in Section 3.4.

3.2 Methods

3.2.1 Model

We consider the following GLMM

$$g(\mu_i) = \eta_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{G}_i\boldsymbol{\gamma} + b_i, \tag{3.1}$$

for $i = 1, .., n$, where $\mu_i = \mathbb{E}(y_i|\boldsymbol{X}_i, \boldsymbol{G}_i, b_i)$, $\boldsymbol{X}_i$ is a $1 \times m$ row vector of covariates for subject i, $\boldsymbol{\alpha}$ is a $m \times 1$ column vector of fixed covariate effects including the intercept, $\boldsymbol{G}_i$ 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, and $\boldsymbol{\gamma}$ is a $p \times 1$ column vector of fixed additive genotype effects. We assume that $\boldsymbol{b} = (b_1, ..., b_n)^\intercal \sim \mathcal{N}(0, \sum_{s=1}^{S} \tau_s \boldsymbol{V}_s)$ is an $n \times 1$ column vector of random effects, $\boldsymbol{\tau} = (\tau_1, ..., \tau_S)^\intercal$ are variance component parameters, $\boldsymbol{V}_1$ is a known kinship matrix or GSM typically estimated from high-quality common genotype markers (MAF ≥ 0.01) (Yang et al. [2011]) and $\boldsymbol{V}_2, ..., \boldsymbol{V}_S$ are any known $n \times n$ positive semi-definite matrices to account for shared environmental effects or complex sampling designs. The phenotypes y_i are assumed to be conditionally independent and identically distributed given $(\boldsymbol{X}_i, \boldsymbol{G}_i, \boldsymbol{b})$ and follow any exponential family distribution with canonical link function $g(\cdot)$, mean $\mathbb{E}(y_i|\boldsymbol{b}) = \mu_i$ and variance $\text{Var}(y_i|\boldsymbol{b}) = \phi a_i^{-1} \nu(\mu_i)$, where ϕ is a dispersion parameter, a_i are known weights and $\nu(\cdot)$ is the variance function. In order to estimate the parameters of interest and perform

variable selection, we need to use an approximation method to obtain a closed analytical form for the marginal likelihood of model (3.1). Following the derivation of (Chen et al. [2016]), we propose to fit (3.1) using a PQL method, from where the log integrated quasi-likelihood function is equal to

$$ql(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \phi, \boldsymbol{\tau}) = -\frac{1}{2}\log\left|\sum_{s=1}^{S}\tau_s \boldsymbol{V}_s \boldsymbol{W} + \boldsymbol{I}_n\right| + \sum_{i=1}^{n} ql_i(\boldsymbol{\alpha}, \boldsymbol{\gamma}|\tilde{\boldsymbol{b}}) - \frac{1}{2}\tilde{\boldsymbol{b}}^{\intercal}\left(\sum_{s=1}^{S}\tau_s \boldsymbol{V}_s\right)^{-1}\tilde{\boldsymbol{b}}, \tag{3.2}$$

where $\boldsymbol{W} = \text{diag}\left\{\frac{a_i}{\phi\nu(\mu_i)[g'(\mu_i)^2]}\right\}$ is a diagonal matrix containing weights for each observation, $ql_i(\boldsymbol{\alpha}, \boldsymbol{\gamma}|\boldsymbol{b}) = \int_{y_i}^{\mu_i}\frac{a_i(y_i-\mu)}{\phi\nu(\mu)}d\mu$ is the quasi-likelihood for the ith individual given the random effects $\boldsymbol{b}$, and $\tilde{\boldsymbol{b}}$ is the solution which maximizes (3.2).

In typical genome-wide studies, the number of predictors is much greater than the number of observations ($p > n$), and the parameter vector $\boldsymbol{\gamma}$ becomes underdetermined when modelling SNPs jointly. Thus, we propose to add a lasso regularization term (Tibshirani [1996]) to the negative quasi-likelihood function in (3.2) to seek a sparse subset of $\boldsymbol{\gamma}$ that gives an adequate fit to the data. Because $ql(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \phi, \boldsymbol{\tau})$ is a non-convex loss function, we propose a two-step estimation method to reduce the computational complexity. First, we obtain the variance component estimates $\hat{\phi}$ and $\hat{\boldsymbol{\tau}}$ under the null hypothesis of no genetic effect ($\boldsymbol{\gamma} = \boldsymbol{0}$) using the AI-REML algorithm (Gilmour et al. [1995]) detailed in Appendix A.1 of the Supplementary Material. Second, assuming that the weights in $\boldsymbol{W}$ vary slowly with the conditional mean, we drop the first term in (3.2) (Breslow and Clayton [1993]) and define the following objective

function which we seek to minimize with respect to $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \tilde{\boldsymbol{b}})$:

$$
\begin{aligned}
(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{b}}) &= \underset{\boldsymbol{\alpha},\boldsymbol{\gamma},\tilde{\boldsymbol{b}}}{\operatorname{argmin}}\, Q_\lambda(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \tilde{\boldsymbol{b}}), \\
Q_\lambda(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \tilde{\boldsymbol{b}}) &= -\sum_{i=1}^{n} ql_i(\boldsymbol{\alpha}, \boldsymbol{\gamma} | \tilde{\boldsymbol{b}}) + \frac{1}{2}\tilde{\boldsymbol{b}}^\intercal \left(\sum_{s=1}^{S} \hat{\tau}_s \boldsymbol{V}_s \right)^{-1} \tilde{\boldsymbol{b}} + \lambda \sum_j v_j |\gamma_j| \\
&:= -\ell_{PQL}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \hat{\phi}, \hat{\boldsymbol{\tau}} | \tilde{\boldsymbol{b}}) + \lambda \sum_j v_j |\gamma_j|, \qquad (3.3)
\end{aligned}
$$

where λ is a nonnegative regularization parameter, and v_j is a penalty factor for the j^{th} predictor. This two-step approach is known as the P3D (population parameters previously determined) method (Zhang et al. [2010]), which is a common approach in mixed-model association tests, and it has been shown to outperform both PCA and genomic control in correcting for sample structure (Kang et al. [2010]). Moreover, Reisetter and Breheny [2021] showed through simulation studies that in the case of penalized LMMs, estimating the variance components once performed similarly in terms of estimating the SNP coefficients than by including the variance components in the iterative procedure, while showing much greater computational efficiency and numerical stability. By default, we standardize the genotype counts and assign $v_j = 1$ for all genetic predictors in (3.3), which is equivalent to using unscaled genotypes with $v_j^{-1} = \sqrt{2MAF_j(1 - MAF_j)}$ where the MAFs are estimated from the data. Alternatively, it is possible to use an adaptive lasso penalty with weights $v_j = |\hat{\beta}_j|^{-\kappa}$, where κ is a common power parameter and $\hat{\beta}_j$ is the coefficient estimate obtained by univariable marginal regression (Waldmann et al. [2019]).

In Appendix A.2 of the Supplementary Material, we detail our proposed cyclic coordinate gradient descent algorithm to solve (3.3) and obtain regularized PQL estimates for $\boldsymbol{\beta} = (\boldsymbol{\alpha}^\intercal, \boldsymbol{\gamma}^\intercal)^\intercal$ and $\tilde{\boldsymbol{b}}$. Briefly, our algorithm is equivalent to iteratively solving the two penalized

weighted least squares (WLS)

$$\underset{\tilde{\boldsymbol{b}}}{\operatorname{argmin}} \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta} - \tilde{\boldsymbol{b}}\right)^{\intercal} \boldsymbol{W} \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta} - \tilde{\boldsymbol{b}}\right) + \tilde{\boldsymbol{b}}^{\intercal} \left(\sum_{s=1}^{S} \hat{\tau}_s \boldsymbol{V}_s\right)^{-1} \tilde{\boldsymbol{b}},$$

and

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right)^{\intercal} \boldsymbol{\Sigma}^{-1} \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right) + \lambda \sum_{j} v_j |\beta_j|, \tag{3.4}$$

where $\boldsymbol{\Sigma} = \boldsymbol{W}^{-1} + \sum_{s=1}^{S} \hat{\tau}_s \boldsymbol{V}_s$ is the covariance matrix of the working response vector $\tilde{\boldsymbol{Y}}$, and $\tilde{\boldsymbol{X}} = [\boldsymbol{X};\ \boldsymbol{G}]$. We use the spectral decomposition of $\boldsymbol{\Sigma}$ to rotate $\tilde{\boldsymbol{Y}}$, $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{b}}$ in (3.4) such that the transformed data is uncorrelated. Given the current estimate for $\boldsymbol{\beta}$, $\tilde{\boldsymbol{b}}$ can be shown to be equal to a generalized ridge-like WLS estimator with $\tilde{\boldsymbol{X}}\boldsymbol{\beta}$ as an offset. Hence, by profiling out $\tilde{\boldsymbol{b}}$ from the objective function and replacing it by its closed-form estimate, we estimate $\boldsymbol{\beta}$ by cycling through its coordinates and minimizing the objective function with respect to one coordinate at a time. In this work, we focus on penalized GLMMs for high-dimensional ($p > n$) GWAS data of binary traits, for which we can use a lower bound on $\boldsymbol{\Sigma}$ so that a single spectral decomposition is performed (Böhning and Lindsay [1988]). Although the methods apply to GLMMs for any exponential family, e.g. counts following a Poisson distribution, we need to perform a spectral decomposition of $\boldsymbol{\Sigma}$ each time we update the weight matrix $\boldsymbol{W}$. Hence, for other distributions, further work is needed to address these computational limitations for application to high-dimensional GWAS data. All calculations and algorithmic steps are detailed in Appendix A.2 of the Supplementary Material.

3.2.2 Prediction

It is often of interest in genetic association studies to make predictions on a new set of individuals, e.g., the genetic risk of developing a disease for a binary response or the expected outcome in the case of a continuous response. In what follows, we compare how predictions

are obtained using `pglmm` versus a GLM with PC adjustment.

`pglmm`

Suppose a single variance component is needed such that $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \tau_1 \boldsymbol{V_1})$ where $\boldsymbol{V_1}$ is the GSM between n subjects that are used to fit the GLMM (3.1). We iteratively fit on a training set of size n the working linear mixed model

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = g'(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \boldsymbol{W}^{-1})$. Let $\tilde{\boldsymbol{Y}}_s$ be the latent working vector in a testing set of n_s individuals with predictor set $\tilde{\boldsymbol{X}}_s$. Similar to Bhatnagar et al. [2020b], we assume that the marginal joint distribution of $\tilde{\boldsymbol{Y}}_s$ and $\tilde{\boldsymbol{Y}}$ is multivariate Normal :

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

where $\boldsymbol{\Sigma}_{12} = \tau_1 \boldsymbol{V}_{12}$ and $\boldsymbol{V}_{12}$ is the $n_s \times n$ GSM between the testing and training individuals. It follows from standard normal theory that

$$\begin{aligned} &\tilde{\boldsymbol{Y}}_s | \tilde{\boldsymbol{Y}}, \phi, \tau_1, \boldsymbol{\beta}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}_s \sim \\ &\quad \mathcal{N}\left(\tilde{\boldsymbol{X}}_s\boldsymbol{\beta} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \right). \end{aligned}$$

The predictions are based on the conditional expectation $\mathbb{E}[\tilde{\boldsymbol{Y}}_s|\tilde{\boldsymbol{Y}}, \phi, \tau_1, \boldsymbol{\beta}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}_s]$, that is

$$\begin{aligned}\hat{\boldsymbol{\mu}}_s &= g^{-1}\left(\mathbb{E}[\tilde{\boldsymbol{Y}}_s|\tilde{\boldsymbol{Y}}, \hat{\phi}, \hat{\tau}_1, \hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}_s]\right) \\ &= g^{-1}\left(\tilde{\boldsymbol{X}}_s\hat{\boldsymbol{\beta}} + \hat{\tau}_1\boldsymbol{V}_{12}\left(\boldsymbol{W}^{-1} + \hat{\tau}_1\boldsymbol{V}_1\right)^{-1}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right) \\ &= g^{-1}\left(\tilde{\boldsymbol{X}}_s\hat{\boldsymbol{\beta}} + \boldsymbol{V}_{12}\boldsymbol{U}\left(\frac{1}{\hat{\tau}_1}\boldsymbol{D}^{-1} + \boldsymbol{U}^\intercal\boldsymbol{W}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^\intercal\boldsymbol{W}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right), \end{aligned} \tag{3.5}$$

where $g(\cdot)$ is the link function and $\boldsymbol{U}$ is the $n \times n$ matrix of PCs obtained from the spectral decomposition of the GSM for training subjects.

GLM with PC adjustment

Another approach to control for population structure and/or subjects' relatedness is to use the first r columns of $\boldsymbol{U}$ as unpenalized fixed effects covariates (Privé et al. [2020]). This leads to the following GLM

$$g(\boldsymbol{\mu}) = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{U}_r\boldsymbol{\delta},$$

where $\boldsymbol{U}_r$ is the $n \times r$ design matrix for the first r PCs and $\delta \in \mathbb{R}^r$ is the corresponding vector of fixed effects. Letting $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{U}_r\boldsymbol{\delta} + g'(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})$ be the working response vector, one can show that

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{U}_r^\intercal\boldsymbol{W}\boldsymbol{U}_r)^{-1}\boldsymbol{U}_r^\intercal\boldsymbol{W}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}\right), \tag{3.6}$$

where $\boldsymbol{W}$ is the diagonal matrix of GLM weights. Recall that $\boldsymbol{V}_{12}$ is the $n_s \times n$ GSM between the test and training sets subjects such that the projected PCs on the testing subjects are

equal to $\boldsymbol{V}_{12}\boldsymbol{U}_r$. Then, the estimated mean response $\hat{\boldsymbol{\mu}}_s$ for the testing set is given by

$$\begin{aligned}\hat{\boldsymbol{\mu}}_s &= g^{-1}\left(\tilde{\boldsymbol{X}}_s\hat{\boldsymbol{\beta}} + \boldsymbol{V}_{12}\boldsymbol{U}_r\hat{\boldsymbol{\delta}}\right) \\ &= g^{-1}\left(\tilde{\boldsymbol{X}}_s\hat{\boldsymbol{\beta}} + \boldsymbol{V}_{12}\boldsymbol{U}_r\left(\boldsymbol{U}_r^\intercal\boldsymbol{W}\boldsymbol{U}_r\right)^{-1}\boldsymbol{U}_r^\intercal\boldsymbol{W}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}\right)\right). \end{aligned} \tag{3.7}$$

By comparing (3.5) and (3.7), we see that both GLM with PC adjustment and `pglmm` use a projection of the training PCs on the testing set to predict new responses, but with different coefficients for the projected PCs. For the former, the estimated coefficients for the first r projected PCs in (3.6) are obtained by iteratively solving generalized least squares (GLS) on the partial working residuals $\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$. For `pglmm`, the estimated coefficients for all projected PCs are also obtained by iteratively solving GLS on the partial working residuals $\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$, with an extra ridge penalty for each coefficient that is equal to ${\hat{\tau}_1}^{-1}\Lambda_i^{-1}$ with Λ_i the i^{th} eigenvalue of $\boldsymbol{V}$ that is associated with the i^{th} PC.

Hence, `pglmm` shrinks PCs coefficients proportionally to their corresponding eigenvalues in a smooth way, while the fixed effect GLM uses a thresholding approach; the first r predictors with larger eigenvalues are kept intact, and the others are completely removed. This implies that the confounding effect from population structure and/or relatedness on the phenotype is fully captured by the first r PCs. As we show in simulations, departure from this assumption may lead to higher false-positive rates and decrease prediction accuracy.

3.2.3 Simulation design

We evaluated the performance of our proposed method against that of a lasso LMM, using the `R` package `ggmix` (Bhatnagar et al. [2020a]), and a logistic lasso, using the `Julia` package `GLMNet` which wraps the `Fortran` code from the original `R` package `glmnet` (Friedman et al. [2010b]). We compared `glmnet` when we included or not the first 10 PCs in the model (`glmnetPC`). We performed a total of 50 replications for two simulation scenarios, drawing

Table 3.1: Values for all simulation parameters. In the first scenario, we simulated binary phenotypes and random genotypes from the BN-PSD admixture model using the `bnpsd` package in `R`. In the second scenario, we simulated binary phenotypes using a total of 6731 subjects of White British ancestry from the UK Biobank data having estimated 1st, 2nd or 3rd degree relationships with at least one other individual.

Parameter	Definition	Scenario 1 BN-PSD model	Scenario 2 Real genotype
M	Number of replications	50	50
h_g^2	Fraction of variance due to fixed genetic effects	0.5	0.17
h_b^2	Fraction of variance due to random genetic effects	0.4	0.4
π_0	Prevalence under the null	0.1	0.1
n	Sample size	2500	6731
p	Number of SNPs	5000	15 000
c	Fraction of causal SNPs	1%	1%

anew genotypes and simulated traits. Values for all simulation parameters are presented in Table 3.1.

Simulated genotype from the admixture model

In the first scenario, we studied the performance of all methods for different population structures by simulating random genotypes from the BN-PSD admixture model for 10 or 20 subpopulations with 1d geography or independent subpopulations using the `bnpsd` package in `R` (Ochoa and Storey [2021]). Sample size was set to $n = 2500$. We simulated $p = 5000$ candidate SNPs and randomly selected c=1% to be causal. The kinship matrix $\boldsymbol{V}$ and PCs were calculated using a set of $50,000$ additional simulated SNPs. We simulated covariates for age and sex using Normal and Binomial distributions, respectively.

For each replication, subjects were partitioned into training and test sets using an 80/20 ratio. Variable selection and coefficient estimation were performed on training subjects for all methods. We compared each method at a fixed number of active predictors, ranging from 5 to either 50 which corresponds to the number of true causal SNPs. Comparisons were

based on three criteria: the ability to retrieve the causal predictors, measured by the true positive rate

$$\text{TPR} = \frac{|\{1 \leq k \leq p : \hat{\beta}_k \neq 0 \cap \beta_k \neq 0\}|}{|\{1 \leq k \leq p : \beta_k \neq 0\}|},$$

the ability to accurately estimate coefficients, measured by the root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}(\hat{\beta}_k - \beta_k)^2},$$

and the ability to predict outcomes in the test sets, measured by the area under the roc curve (AUC).

Real genotypes from the UK Biobank data

In the second scenario, we compared the performance of all methods when a high proportion of related individuals are present, using real genotype data from the UK Biobank. We retained a total of 6731 subjects of White British ancestry having estimated 1st, 2nd or 3rd degree relationships with at least one other individual. We compared methods in a more realistic setting with weaker effect sizes and more causal variants than the first scenario. We sampled $p = 15\,000$ candidate SNPs among all chromosomes and randomly selected c=1% to be causal. We used PCs as provided with the data set. These were computed using a set of unrelated samples and high quality markers pruned to minimise LD (Bycroft et al. [2018]). Then, all subjects were projected onto the principal components using the corresponding loadings. Since the markers that were used to compute the PCs were potentially sampled as candidate causal markers in our simulations, we included all candidate SNPs in the set of markers used for calculating the kinship matrix $\boldsymbol{V}$. We simulated age using a Normal distribution and used the sex covariate provided with the data.

For this simulation scenario, we evaluated the performance of all methods when using cross-validation as a model selection criteria, rather than fixing the number of active predictors in

the model. For this, the 6731 subjects from the UK Biobank data were randomly split into training (40%), validation (30%) and test (30%) sets, ensuring all related individuals were assigned into the same set. For cross-validation, the full lasso solution path was fitted on the training set, and the regularization parameter was obtained on the model which maximized AUC on the validation set. We also evaluated the performance of our proposed method when using AIC as a model selection criterion. Again, we compared methods performance on the basis of TPR, AUC on the test sets and RMSE. Additionally, we compared each model selection approach on the total number of predictors selected and on the model precision, which is defined as the proportion of selected predictors that are true positives.

Simulation model

Let S be the set of candidate causal SNPs, with $|S| = p \times c$, then the causal SNPs fixed effects β_j were generated from a Gaussian distribution $\mathcal{N}(0, h_g^2\sigma^2/|S|)$, where h_g^2 is the fraction of variance on the logit scale that is due to total additive genetic fixed effects. That is, we assumed the candidate causal markers explained a fraction of the total polygenic heritability, and the rest was explained by a random polygenic effect $b \sim \mathcal{N}(0, h_b^2\sigma^2\boldsymbol{V})$. For the first scenario, we simulated a signal-to-noise ratio (SNR) equal to 1 for the fixed genetic effects ($h_g^2 = 50\%$) under strong random polygenic effects ($h_b^2 = 40\%$). For the second scenario, we simulated fixed effects using $h_g^2 = 17\%$, which corresponds to the estimated SNP heritability for asthma on the liability scale (see https://nealelab.github.io/UKBB_ldsc/h2_summary_20002_1111.html), again under strong random polygenic effects ($h_b^2 = 40\%$). We then simulated a binary phenotype using a logistic link function

$$\begin{aligned}\text{logit}(\pi) = \text{logit}(\pi_0) - \log(1.3) \times Sex + \log(1.05)Age/10 \\ + \sum_{j \in S} \beta_j \cdot \widetilde{G}_j + b,\end{aligned} \tag{3.8}$$

Table 3.2: Demographics for the real data application. We retained a total of 6731 subjects of White British ancestry from the UK Biobank data having estimated 1st, 2nd or 3rd degree relationships with at least one other individual.

	Asthma		High Cholesterol	
	Cases	Controls	Cases	Controls
N (%)	819 (12.2)	5912 (87.8)	883 (13.1)	5848 (86.7)
Age Median (IQR)	58 (16)	59 (15)	64 (7)	57 (16)
Male (%)	306 (37.4)	2571 (43.5)	467 (52.9)	2410 (41.2)

where the parameter π_0 was chosen to specify the prevalence under the null, and $\widetilde{G}_j$ 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_i is the MAF. By using the spectral decomposition of the kinship matrix $\boldsymbol{V}$, we can show that $\boldsymbol{b} = \boldsymbol{U}\delta$, where $\delta \sim \mathcal{N}(0, h_b^2\sigma^2\boldsymbol{D})$, $\boldsymbol{U}$ is the $n \times n$ matrix of PCs, and $\boldsymbol{D}$ is a diagonal matrix of corresponding eigenvalues. Thus, the unmeasured confounding effect $\boldsymbol{\delta}$ is correlated with the population structure through the design matrix of PCs $\boldsymbol{U}$.

3.2.4 Real data application

We used the same set of 6731 related subjects from the UK Biobank data set presented in Section 2.3.2 to construct a polygenic risk score (PRS) on two highly heritable binary traits, asthma (self-reported, UK Biobank code: 20002_1111) and high cholesterol (self-reported, UK Biobank code: 20002_1473). We present demographics and number of cases for both analyses in Table 3.2. After filtering for SNPs with missing rate smaller than 0.01, MAF above 0.05 and a p-value for the Hardy–Weinberg exact test above 10^{-6}, a total of 320K genotype SNPs were remaining.

To better understand the contribution of the PRS for predicting asthma and high cholesterol, we fitted for each trait a null model with only age, sex, genotyping platform and the first 10 PCs as fixed effects. Since for highly polygenic traits, it is generally considered that there are a large number of predictors with small to moderate effects (O'Connor et al. [2019]), we also fitted a standard genomic best linear unbiased prediction (gBLUP) model (Ødegård

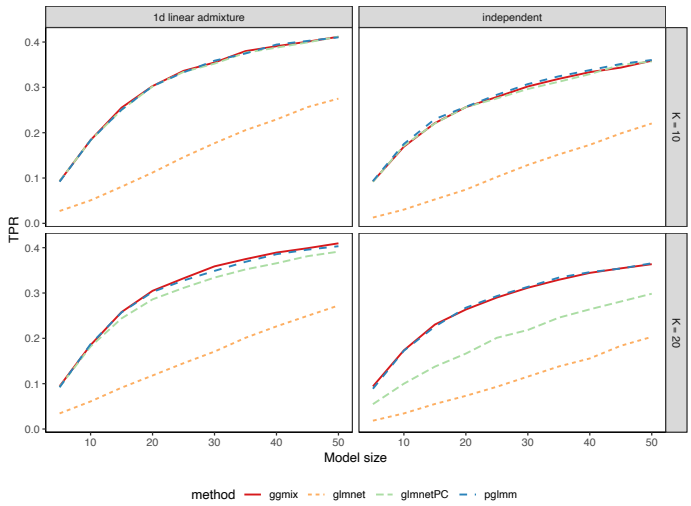
et al. [2018]). The gBLUP model corresponds to the null model of `pglmm`, i.e, a model where we include age, sex and genotyping platform as fixed effects, and one random effect with variance-covariance proportionnal to the GSM. For both gBLUP and `pglmm`, we did not include any PC since kinship is accounted for by the random effect. Finally, we also fitted a logistic lasso in which the top 10 PCs were included as unpenalized covariates in addition to age, sex and genotyping platform (`glmnetPC`). To evaluate the predictive performance of the compared methods in independent subjects, we randomly split the subjects in training (80%) and test (20%) sets for a total of 40 times. For each of the 40 replications, the full lasso solution path was fitted on the training set only. For `pglmm`, the regularization parameter (λ) was selected to minimize the AIC on the training data. For `glmnetPC`, the regularization parameter was obtained by minimizing the deviance using 10-fold cross-validation on the training data. We compared mean prediction accuracy on the test sets as well as the median number of predictors included in all models.

3.3 Results

3.3.1 Simulation results for the first scenario

Results for selection of important predictors, as measured by the mean TPR in 50 replications, are presented in Figure 3.1. For both 1d linear admixture and independent subpopulations, `glmnet` without PC adjustment failed to retrieve causal markers compared to all other methods. This is expected under population stratification; SNPs that differ in frequency between subpopulations are identified as important predictors because prevalence is not constant across each group. When the first 10 PCs were added as unpenalized covariates, `glmnetPC`'s ability to select causal predictors was lesser to that of `pglmm` and `ggmix` for the 20 independent subpopulations. Since in the independent subpopulations simulated data, each subpopulation indicator function is strongly associated with only a few PCs, as shown in Appendix A.5 of the Supplementary Material, omitting to include all important PCs in

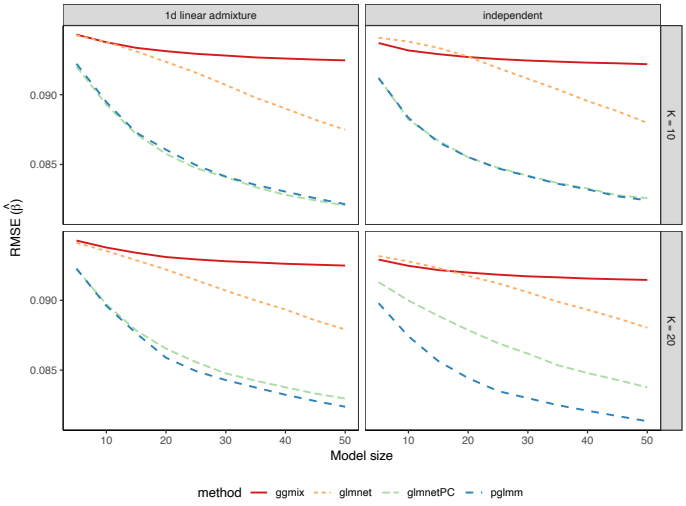
Figure 3.1: Mean of 50 TPRs for the first simulation scenario where we simulated random genotypes from the BN-PSD admixture model. K represents the number of intermediate subpopulations in the 1d linear admixture data (left panel), and the number of independent subpopulations in the independent data (right panel).



the model leads to incorrectly capturing the confounding structure. On the other hand, because there is more overlap between subpopulations in the admixture data compared to the independent subpopulations (Reisetter and Breheny [2021]), each subpopulation indicator function is moderately correlated with many PCs. Thus, including only the first 10 PCs in the model is enough to correct for confounding even when K = 20. (bottom-left panel of Figure 3.1). Alternatively, including a random effect with variance-covariance structure proportional to the GSM correctly adjusts for population structure in all scenarios while alleviating the burden of choosing the right number of fixed predictors to include in the model. Even though `ggmix` assumes a standard LMM for the binary trait, it was able to identify causal markers at the same rate as `pglmm`.

Results for estimation of SNP effects as measured by the mean RMSE in 50 replications are presented in Figure 3.2. Results are consistent with TPR results in that `glmnet` without PC adjustment performed poorly in all scenarios, while `pglmm` outperformed all other methods for the 20 independent subpopulations and performed comparably with `glmnetPC` for all other settings. As expected, `ggmix` had higher RMSE compared to `pglmm` and `glmnetPC`.

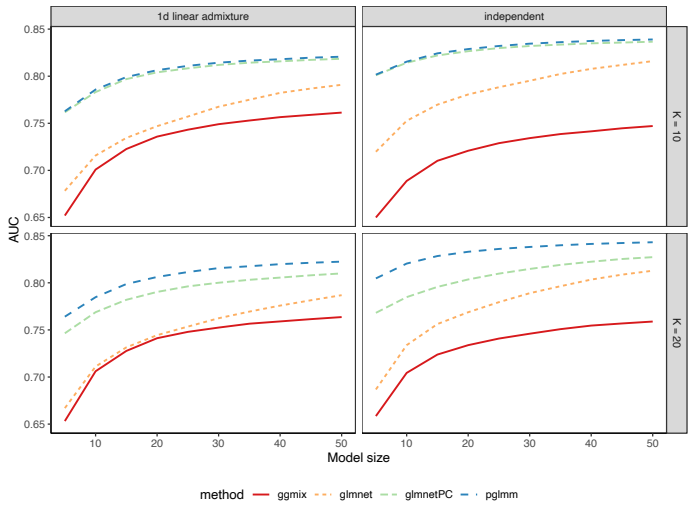
Figure 3.2: Mean of 50 RMSEs for the first simulation scenario where we simulated random genotypes from the BN-PSD admixture model. K represents the number of intermediate subpopulations in the 1d linear admixture data (left panel), and the number of independent subpopulations in the independent data (right panel).



Thus, even though `ggmix` was able to identify causal markers at the same rate as other methods that accounted for the binary nature of the response, resulting estimates for the SNP effects were not accurate.

For both 1d linear admixture and independent subpopulations, `ggmix` and `glmnet` had poor predictive performance for $K = 10$ and $K = 20$, as reported in Figure 3.3. Also, the predictive performance of `glmnetPC` was greatly reduced when $K = 20$ for both admixture and independent populations, even if in the case of the admixture data, the RMSE for estimation of SNP effects was comparable for `glmnetPC` and `pglmm`. This suggests that the observed discrepancy in predictive accuracy might be caused by how each method handle the confounding effects. Using only 10 PCs as fixed effects when $K = 20$ may result in overfitted coefficients for `glmnetPC`, which may in turn potentially decrease prediction accuracy and increase variance of predictions in independent subjects. By using a ridge-like estimator for the random effects, `pglmm` is less likely to overfit the confounding effects compared to `glmnetPC`.

Figure 3.3: Mean of 50 AUCs in test sets for the first simulation scenario where we simulated random genotypes from the BN-PSD admixture model. K represents the number of intermediate subpopulations in the 1d linear admixture data (left panel), and the number of independent subpopulations in the independent data (right panel).



3.3.2 Simulation results for the second scenario

In the second simulation scenario, we evaluated the performance of our method when using AIC or cross-validation as a model selection strategy, i.e, for selecting the optimal value of the regularization parameter, rather than fixing the number of active predictors in the model. For all other methods, we used cross-validation to perform model selection and compared the ability of all methods to adjust for potential confounding stemming from subjects' relatedness. We present median and interquartile values for AUC, model size, RMSE, TPR and precision in Table 3.3. In addition to the penalized methods, we reported prediction accuracy for the standard gBLUP model where only non-genetic covariates and a polygenic random effect were included.

Contrarily to the previous simulations under the admixture and independent populations models, `glmnetPC` had lower prediction accuracy compared to `glmnet`. This highlights the fact that using a fixed number of PCs to control for sample relatedness is not robust compared to using a random effect. In comparison to the first simulation scenario, where the TPR was between 30% and 40% when the number of active predictors in the model was equal to the

Table 3.3: Results of the model selection simulations for the second scenario. For each replication, the best model for `pglmm` was chosen using either AIC, or CV. For all other methods, the best model was chosen using CV. For all metrics, we report median and interquartile range. Since the gBLUP model makes prediction using only non-genetic covariates and a polygenic random effect, we only report median AUC values.

	`ggmix`	`glmnet`	`glmnetPC`	`pglmm` (AIC)	`pglmm` (CV)	gBLUP
Model size	341 (1598)	226 (1050)	378 (1021)	50.5 (54.2)	102 (487)	0 (0)
AUC	0.558 (0.023)	0.561 (0.030)	0.552 (0.022)	0.569 (0.021)	0.568 (0.028)	0.549 (0.020)
RMSE	0.0334 (0.0038)	0.0328 (0.0117)	0.0336 (0.0127)	0.0318 (0.0036)	0.0324 (0.0051)	-
TPR	0.167 (0.248)	0.133 (0.192)	0.17 (0.165)	0.06 (0.0517)	0.08 (0.155)	-
Precision	0.0576 (0.141)	0.0854 (0.117)	0.0584 (0.0653)	0.203 (0.156)	0.107 (0.200)	-

number of causal markers, the maximum value for the TPR for all methods was equal to 17% in the second simulation scenario. This is because although in both scenarios the proportion of causal markers was the same ($c = 1\%$), we simulated more causal predictors with weaker effects size in the second scenario. Indeed, the number of causal markers and simulated heritability in the second scenario were equal to $c * p = 150$ and $h_g^2 = 17\%$ respectively, compared to $c * p = 50$ and $h_g^2 = 50\%$ in the first scenario.

In term of prediction accuracy and estimation of predictor coefficients, `pglmm` performed comparably using either cross-validation or AIC, while achieving better performance than all other methods. Moreover, our method led to sparser models with higher precision than all other methods, especially when using AIC as a model selection criteria. Thus, using a logistic lasso model with 10 PCs to control for relatedness led to models with more false positives and worse prediction accuracy than all other methods, including the logistic lasso with no PC adjustment. These results highlight once again the robustness of using a random effect rather than PCs to account for relatedness between subjects. In summary, by explicitly modeling the correlation between subjects and binary nature of the trait, our method led to sparser models with higher precision and prediction accuracy than all other methods.

3.3.3 PRS for the UK Biobank real data application

Results for asthma and high cholesterol PRSs are summarized in Table 3.4. For asthma, `pglmm` performed better than all other methods when comparing AUC on the test sets. In addition, the median number of predictors selected by `pglmm` was 2.5 times smaller than for `glmnetPC`, and the variability in predictors selected was more important for `glmnetPC`, as reported by an IQR value equal to 113.25, compared to 43.25 for `pglmm`. This is consistent with our simulation results showing that `pglmm` leads to sparser models with higher predictive power than logistic lasso. For high cholesterol, the median number of predictors selected by both penalized models was equal or close to 0, which suggests that SNP effects may be too small to detect. Indeed, both methods based on sparse regression do not perform as well as either the gBLUP or null model with non-genetic covariates and 10 PCs. For both asthma and high cholesterol, fitting the null model for `pglmm` took a median time of approximately 1.7 minutes, while fitting the full lasso path for 100 values of the tuning parameter λ took a median time of 51 and 55 minutes respectively. Analyses were performed using 2 cores of an AMD Rome 7532 (2.40 GHz), each with 64GB of RAM. As implemented in the `glmnet` package and other high-dimensional sparse regression methods, we use sequential strong rules for solving the lasso problem such that most of the predictors are discarded from the optimization problem at each iteration (Tibshirani et al. [2012]). This allows our sparse regularized mixed regression method to remain computationally efficient when the number of genetic variants is very large.

3.3.4 Computational efficiency

In this additional simulation scenario, we compared the computational efficiency of all methods. We considered a grid of values for the sample size n and number of predictors p, and we simulated a total of 10 replications of the 1d linear admixture model with 20 populations, for each of the nine combinations of (n, p). For each replication, we randomly selected 1% of the predictors to be causal. Simulations were performed on a single core of an AMD Rome

Table 3.4: PRS results for asthma and high cholesterol using a total of 6731 subjects of White British ancestry from the UK Biobank data having estimated 1st, 2nd or 3rd degree relationships with at least one other individual. To find the optimal regularization parameter for both penalized methods, we split the subjects in training (80%) and test (20%) sets for a total of 40 times.

Model	AUC_{test}	Model size
Asthma	Mean (SD)	Median (IQR)
Covariates + 10PCs	0.5227 (0.021)	-
gBLUP	0.5447 (0.017)	-
`glmnetPC`	0.5258 (0.020)	42 (113.25)
`pglmm`	0.5484 (0.017)	16.5 (43.25)
High cholesterol		
Covariates + 10PCs	0.7126 (0.020)	-
gBLUP	0.7142 (0.020)	-
`glmnetPC`	0.7106 (0.020)	0 (7.25)
`pglmm`	0.7118 (0.020)	0.5 (16)

7532 (2.40 GHz) with 64 GB of RAM. Results for the computational efficiency of all methods for different sample sizes and number of predictors are reported in Table 3.5. The median computational time of `pglmm` ranged between 5.3 minutes ($n = 2500$, $p = 10\ 000$) and 48.6 minutes ($n = 7500$, $p = 30\ 000$), while `ggmix` running time varied between 13.7 and 93.3 minutes respectively. Thus, `pglmm` was considerably faster than `ggmix` because contrarily to the latter, we estimate the variance components only once under the null model, which dramatically decreases the computational complexity of the regularized minimization problem. The maximum running time for `glmnetPC` was equal to 2 minutes, for the simulations with $n = 7500$ and $p = 30\ 000$. The large difference in computation time between `pglmm` and `glmnetPC` is explained by the dimension of the parameter space that each method is estimating. Indeed, to account for population structure, `glmnetPC` only needs to fit the first 10 PCs, while under the mixed model approach, `pglmm` needs to estimate the random effects vector of dimension equal to the sample size. As we show in Appendix A.2 of the Supplementary Material, by profiling out the random effects vector from the regularized minimization problem in our proposed algorithm, we need to rotate the response vector using the eigenvectors of the variance-covariance matrix after each WLS iteration such that the transformed data

Table 3.5: Median computation time in minutes of `pglmm`, `glmnetPC` and `ggmix` for fitting a sequence of 100 regression models for different sample sizes and number of predictors. For `pglmm`, we also present the median computation time for fitting the null model. Simulations were performed on a single core of an AMD Rome 7532 (2.40 GHz) with 64 GB of RAM. We simulated a total of 10 replications of the 1d linear admixture model with 20 populations.

		`pglmm`		`glmnetPC`	`ggmix`
n	p	Null model	Full model	Full model	Full model
2500	10 000	0.6	5.3	0.2	13.7
	20 000	-	11.0	0.4	26.6
	30 000	-	13.8	0.6	34.0
5000	10 000	2.1	11.9	0.6	25.0
	20 000	-	24.4	1.0	42.0
	30 000	-	34.9	1.2	55.8
7500	10 000	4.8	24.0	1.3	51.3
	20 000	-	33.5	1.7	78.2
	30 000	-	48.6	2.0	93.3

is uncorrelated. This requires performing multiple matrix-vector multiplications, with complexity $O(n^2)$, while `glmnetPC` only needs performing vector multiplications with complexity $O(n)$.

3.4 Discussion

We have introduced a new method called `pglmm` based on regularized PQL estimation, for selecting important predictors and estimating their effects in high-dimensional GWAS data, accounting for population structure, close relatedness and binary nature of the trait. By simulating random genotypes from the BN-PSD admixture model for 10 or 20 subpopulations with 1d geography or independent subpopulations, we showed that `pglmm` was markedly better than a logistic lasso with PC adjustment when the number of subpopulations was greater than the number of PCs included. We also showed that a lasso LMM was unable to estimate predictor effects with accuracy for binary responses, which greatly decreased its predictive performance. Performance assessment was based on TPR of selected predictors, RMSE of estimated effects and AUC of predictions. These results strongly advocate for

using methods that explicitly account for the binary nature of the trait while effectively controlling for population structure and relatedness in genetic studies.

In the second simulation scenario, we used real genotype data from a subset of related individuals from the UK Biobank data to simulate binary responses, and showed that `pglmm` effectively led to sparser models with higher precision and prediction accuracy than a lasso LMM and a logistic lasso model with or without PC adjustment. We also demonstrated that using AIC as a model selection strategy led to similar prediction performance than cross-validation, with even sparser models. Using the same data set, we illustrated the potential advantages of `pglmm` over a logistic lasso with PC adjustment in a real data application for constructing a PRS on two highly heritable binary traits. Although these analyses have limited power compared to using all UK Biobank subjects of White British ancestry, it is often the case that researchers may be limited to relatively smaller data sets. For these cases, it is of primary importance to avoid discarding samples based on relatedness and properly account for the possible correlation between observations. Thus, we used this reduced sample from the UK Biobank to demonstrate the potential advantages of using penalized multivariable GLMMs in smaller data sets where subjects' relatedness might be an important confounder.

A limitation of `pglmm` compared to a logistic 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 are clearly too prohibitive for application to large cohorts such as the full UK Biobank with a total of $500K$ samples. 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. [2018]), and to use a sparse GSM to adjust for the sample relatedness (Jiang et al. [2019]).

In this study, we focused solely on the lasso as a regularization penalty for the genetic markers

effects. However, 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, although we note that both SCAD and MCP require tuning an additional parameter which controls the relaxation rate of the regularization. Another alternative includes implementation of 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 (Meinshausen [2007]). Finally, it would also be of interest to explore if tuning the generalized ridge regularization on the random effects, or replacing it by a lasso regularization to perform selection of individual random effects, could result in better predictive performance.

Acknowledgments

We thank the UK Biobank and all participants for providing information. 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 also like to thank the three anonymous reviewers for their valuable suggestions.

Funding

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.].

Conflict of Interest: None declared.

Data availability statement

Our Julia package `PenalizedGLMM` and simulated data are available on github https://github.com/julstpierre/PenalizedGLMM. UK Biobank data are available via application directly to UK Biobank (https://www.ukbiobank.ac.uk/enable-your-research). The current study was conducted under UK Biobank application number 20802.

References

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. URL https://doi.org/10.1137/141000671.

Sahir R. Bhatnagar, Yi Yang, and Celia M. T. Greenwood. ggmix: Variable selection in linear mixed models for snp data. R package version 0.0.1. 2020a.

Sahir R. Bhatnagar, Yi Yang, Tianyuan Lu, Erwin Schurr, JC Loredo-Osti, Marie Forest, Karim Oualkacha, and Celia M. T. Greenwood. Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. *PLOS Genetics*, 16(5):e1008766, May 2020b. 10.1371/journal.pgen.1008766. URL https://doi.org/10.1371/journal.pgen.1008766.

Dankmar Böhning and Bruce G. Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, December 1988. ISSN 0020-3157, 1572-9052. 10.1007/BF00049423.

Norman E. Breslow and David G. Clayton. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421):9–25, March 1993. ISSN 0162-1459, 1537-274X. 10.1080/01621459.1993.10594284.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes,

Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018. 10.1038/s41586-018-0579-z. URL https://doi.org/10.1038/s41586-018-0579-z.

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. 10.1016/j.ajhg.2016.02.012. URL https://doi.org/10.1016/j.ajhg.2016.02.012.

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.

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.

Andreas Groll and Gerhard Tutz. Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing*, 24(2):137–154, October 2012. ISSN 1573-1375. 10.1007/s11222-012-9359-z. URL http://dx.doi.org/10.1007/s11222-012-9359-z.

Francis K. C. Hui, Samuel Müller, and A. H. Welsh. Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association*, 112(519):1323–1333, June 2017. ISSN 1537-274X. 10.1080/01621459.2016.1215989. URL http://dx.doi.org/10.1080/01621459.2016.1215989.

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.

Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, March 2008. 10.1534/genetics.107.080101. URL https://doi.org/10.1534/genetics.107.080101.

Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, March 2010. 10.1038/ng.548. URL https://doi.org/10.1038/ng.548.

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.

Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the

missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct 2009. ISSN 1476-4687. 10.1038/nature08494. URL https://doi.org/10.1038/nature08494.

Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, September 2007. 10.1016/j.csda.2006.12.019. URL https://doi.org/10.1016/j.csda.2006.12.019.

Alejandro Ochoa and John D. Storey. Estimating FST and kinship for arbitrary population structures. *PLOS Genetics*, 17(1):e1009241, January 2021. 10.1371/journal.pgen.1009241. URL https://doi.org/10.1371/journal.pgen.1009241.

Luke J. O'Connor, Armin P. Schoech, Farhad Hormozdiari, Steven Gazal, Nick Patterson, and Alkes L. Price. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476, September 2019. 10.1016/j.ajhg.2019.07.003. URL https://doi.org/10.1016/j.ajhg.2019.07.003.

Jørgen Ødegård, Ulf Indahl, Ismo Strandén, and Theo H. E. Meuwissen. Large-scale genomic prediction using singular value decomposition of the genotype matrix. *Genetics Selection Evolution*, 50(1), February 2018. 10.1186/s12711-018-0373-2. URL https://doi.org/10.1186/s12711-018-0373-2.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, July 2006. 10.1038/ng1847. URL https://doi.org/10.1038/ng1847.

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.

Florian Privé, Bjarni J. Vilhjálmsson, and Hugues Aschard. Fitting penalized regressions

on very large genetic data using snpnet and bigstatsr. October 2020. 10.1101/2020.10.30.362079. URL https://doi.org/10.1101/2020.10.30.362079.

Barbara Rakitsch, Christoph Lippert, Oliver Stegle, and Karsten Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, November 2012. 10.1093/bioinformatics/bts669. URL https://doi.org/10.1093/bioinformatics/bts669.

Anna C. Reisetter and Patrick Breheny. Penalized linear mixed models for structured genetic data. *Genetic Epidemiology*, May 2021. 10.1002/gepi.22384. URL https://doi.org/10.1002/gepi.22384.

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.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 74(2):245–266, 2012. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/41430939.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, July 2017. 10.1016/j.ajhg.2017.06.005. URL https://doi.org/10.1016/j.ajhg.2017.06.005.

Patrik Waldmann, Maja Ferenčaković, Gábor Mészáros, Negar Khayatzadeh, Ino Curik, and Johann Sölkner. AUTALASSO: an automatic adaptive LASSO for genome-wide prediction. *BMC Bioinformatics*, 20(1), April 2019. 10.1186/s12859-019-2743-3. URL https://doi.org/10.1186/s12859-019-2743-3.

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.

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.

Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, and Edward S Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360, March 2010. 10.1038/ng.546. URL https://doi.org/10.1038/ng.546.

Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A. Gagliano, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, August 2018. ISSN 1546-1718. 10.1038/s41588-018-0184-y. URL http://dx.doi.org/10.1038/s41588-018-0184-y.

Chapter 4

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

Preamble to Manuscript 2.

The etiology of most complex diseases involves genetic variants, environmental factors, and gene-environment interaction (GEI) effects. In high-dimensional regularized regression models, hierarchical selection of GEI effects is desired, where a GEI effect is selected in the model only if the corresponding genetic main effect is also selected. Hierarchy can be induced through the use of group lasso penalties (Yuan and Lin [2005], Lim and Hastie [2015]) and generalizations thereof such as the sparse group lasso (Friedman et al. [2010a], Liang et al. [2024]) or group L_∞ penalty (Zemlianskaia et al. [2022]), or by adding a set of convex constraints to the lasso (Bien et al. [2013]).

Dependence between gene and environment can be induced by population structure and closer relatedness (Dudbridge and Fletcher [2014]). Thus, spurious selection of GEI effects is also of concern in genetic association studies. Sul et al. [2016] showed that under the

polygenic model, population structure and closer relatedness 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.

The original contributions of this chapter are i) developing a unified approach based on regularized penalized quasi-likelihood (PQL) estimation to perform hierarchical selection of GEI effects in sparse regularized logistic mixed models, ii) deriving a proximal Newton-type algorithm with block coordinate descent for PQL estimation with mixed lasso and group lasso penalties, and iii) showing through simulation studies that including and additional random effect to account for the shared environmental exposure reduced the false positive rate for the selection of both GEI and main genetic effects in sparse regularized mixed models.

The corresponding manuscript has been accepted for publication in *Statistical Methods in Medical Research* and is currently available online (St-Pierre et al. [2024]).

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

Julien St-Pierre[1], Karim Oualkacha[2], Sahir Rai Bhatnagar[1].

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

[2]*Département de Mathématiques, Université du Québec à Montréal*

This thesis contains the accepted version of the corresponding paper published in *Statistical Methods in Medical Research* (St-Pierre et al. [2024]).

Abstract

Interactions between genes and environmental factors may play a key role in the etiology of many common disorders. Several regularized generalized linear models (GLMs) have been proposed for hierarchical selection of gene by environment interaction (GEI) effects, where a GEI 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 paper, we develop a unified approach based on regularized penalized quasi-likelihood (PQL) estimation to perform hierarchical selection of GEI 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 (FPR) and false discovery rate (FDR) of our proposed method for selection of both GEI and main effects. Using the F_1 score as a balanced measure of the FDR and true positive rate (TPR), we further show that in the hierarchical simulation scenarios, our method outperforms other methods for retrieving important GEI effects. Thus, compared to other penalized methods, our proposed method enforces sparsity by controlling the number of false positives in the model while having the best predictive performance. 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.

4.1 Introduction

Genome-wide association studies (GWAS) have led to the identification of hundreds of common genetic variants, or single nucleotide polymorphisms (SNPs), associated with complex traits (Visscher et al. [2017]) 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, GWAS 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 (Manolio et al. [2009]). Beyond the identified genetic variants, interactions between 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. [2020] 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 (Mukherjee et al. [2009]).

Several regularized generalized linear models (GLMs) have been proposed for selection of both genetic and GEI effects in genetic association studies (Fang et al. [2023], Zemlianskaia et al. [2022], Lim and Hastie [2015]), 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) (Yu et al. [2005]). 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 (Hoffman [2013]). Hence, while both Principal Component 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 (Price et al. [2010]). 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 closed 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 matrix respectively, where p is the number of genetic predictors. Even for moderately small cohorts, the number of predictors in GWAS is often greater than one 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 (Qian et al. [2020]). 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 (St-Pierre et al. [2023], Hu et al. [2019], Hui et al. [2017]). 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 (Bien et al. [2013]). Hierarchical variable selection of GEI effects is appealing both for increasing statistical power (Cox [1984]) 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 (Dudbridge and Fletcher [2014]). In the context of GWAS, Sul et al. [2016] 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 paper, 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. [2016], 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 selection (Zhao et al. [2009]) 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 (St-Pierre et al. [2023]). 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 cohort (Maixner et al. [2011]) to study the sex-specific association between temporomandibular disorder (TMD) and genetic predictors.

4.2 Methodology

4.2.1 Model

We have the following generalized linear mixed model (GLMM)

$$g(\mu_i) = \eta_i = \boldsymbol{Z_i\theta} + D_i\alpha + \boldsymbol{G_i\beta} + (D_i\boldsymbol{G}_i)\boldsymbol{\gamma} + b_i \tag{4.1}$$

for $i = 1, .., n$, where $\mu_i = \mathbb{E}(y_i|\boldsymbol{Z}_i, \boldsymbol{G}_i, D_i, b_i)$, $\boldsymbol{Z}_i$ is a $1 \times m$ row vector of covariates for subject i, $\boldsymbol{G}_i$ 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, $(\boldsymbol{\theta}^\intercal, \boldsymbol{\beta}^\intercal)^\intercal$ 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 continous environmental factor D with fixed effect α, and $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, ..., \gamma_p]^\intercal \in \mathbb{R}^p$ is the vector of fixed GEI effects. Thus, we have a total of $2p + m + 1$ coefficients. We assume that $\boldsymbol{b} = (b_1, ..., b_n)^\intercal \sim \mathcal{N}(0, \tau_g\boldsymbol{K} + \tau_d\boldsymbol{K}^D)$ is an $n \times 1$ column vector of random

effects, with $\boldsymbol{\tau} = (\tau_g, \tau_d)^\intercal$ the variance components that account for the relatedness between individuals. $\boldsymbol{K}$ is a known GSM or kinship matrix and $\boldsymbol{K}^D$ 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 $\boldsymbol{K}^D$ corrects for the spurious association of GEI effects due to population structure and subjects relatedness, in the same way that the kinship matrix $\boldsymbol{K}$ corrects for population structure and subjects relatedness on the main effects. Thus, the matrix $\boldsymbol{K}^D$ 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 $K^D_{ij} = K_{ij}$ if $D_i = D_j$, and $K^D_{ij} = 0$ otherwise. For a continuous exposure, one possibility is to set $K^D_{ij} = K_{ij}(1 - d(D_i, D_j))$, where d is a metric with range $[0, 1]$. The phenotypes y_i's are assumed to be conditionally independent and identically distributed given $(\boldsymbol{Z}_i, \boldsymbol{G}_i, D_i, \boldsymbol{b})$ and follow any exponential family distribution with canonical link function $g(\cdot)$, mean $\mathbb{E}(y_i|\boldsymbol{Z}_i, \boldsymbol{G}_i, D_i, \boldsymbol{b}) = \mu_i$ and variance $\text{Var}(y_i|\boldsymbol{Z}_i, \boldsymbol{G}_i, D_i, \boldsymbol{b}) = \phi a_i^{-1}\nu(\mu_i)$, where ϕ is a dispersion parameter, a_i are known weights and $\nu(\cdot)$ is the variance function.

4.2.2 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 (4.1). We propose to fit (4.1) using a PQL method (St-Pierre et al. [2023], Chen et al. [2016]), from where the log integrated quasi-likelihood function is equal to

$$\ell_{PQL}(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}; \tilde{\boldsymbol{b}}) = -\frac{1}{2}\log\left|\left(\tau_g\boldsymbol{K} + \tau_d\boldsymbol{K}^D\right)\boldsymbol{W} + \boldsymbol{I}_n\right| + \sum_{i=1}^{n} ql_i(\boldsymbol{\Theta}; \tilde{\boldsymbol{b}}) - \frac{1}{2}\tilde{\boldsymbol{b}}^\intercal\left(\tau_g\boldsymbol{K} + \tau_d\boldsymbol{K}^D\right)^{-1}\tilde{\boldsymbol{b}}, \tag{4.2}$$

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

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 modelling 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 penalty (Simon et al. [2013]) to the negative quasi-likelihood function in (4.2) 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 composite absolute penalties (CAP) that can induce hierarchical variable selection (Zhao et al. [2009]). We define the following objective function Q_λ which we seek to minimize with respect to $(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau})$:

$$Q_\lambda(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}; \tilde{\boldsymbol{b}}) := -\ell_{PQL}(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}; \tilde{\boldsymbol{b}}) + (1-\rho)\lambda\sum_j \|(\beta_j, \gamma_j)\|_2 + \rho\lambda\sum_j |\gamma_j|, \tag{4.3}$$

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 modelling 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.

4.2.3 Estimation of variance components

Jointly estimating the variance components τ_g, τ_d and scale parameter ϕ with the regression effects vector $\boldsymbol{\Theta}$ and random effects vector $\boldsymbol{b}$ is a computationally challenging non-convex optimization problem. Updates for τ_g, τ_d and ϕ based on a majorization-minimization (MM) algorithm (Zhou et al. [2019a]) 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 $\boldsymbol{\beta} = \boldsymbol{\gamma} = 0$, using the AI-REML algorithm (St-Pierre et al. [2023], Gilmour et al. [1995]).

4.2.4 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

$$\begin{aligned}\left(\hat{\tau}_g \boldsymbol{K} + \hat{\tau}_d \boldsymbol{K}^D\right)^{-1} &= \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\intercal\right)^{-1} \\ &= \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\intercal, \end{aligned} \tag{4.4}$$

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

Using (4.4) and assuming that the weights in $\boldsymbol{W}$ vary slowly with the conditional mean (Bres-

low and Clayton [1993]), minimizing (4.3) is now equivalent to

$$\begin{aligned}\hat{\boldsymbol{\Theta}} &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \; -\sum_{i=1}^{n} ql_i(\boldsymbol{\Theta}; \tilde{\boldsymbol{\delta}}) + \frac{1}{2}\tilde{\boldsymbol{\delta}}^{\intercal}\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{\delta}} + (1-\rho)\lambda\sum_{j}\|(\beta_j, \gamma_j)\|_2 + \rho\lambda\sum_{j}|\gamma_j| \\ &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \; f(\boldsymbol{\Theta}; \tilde{\boldsymbol{\delta}}) + g(\boldsymbol{\Theta}), \end{aligned} \tag{4.5}$$

where $\tilde{\boldsymbol{\delta}} = \boldsymbol{U}^{\intercal}\tilde{\boldsymbol{b}}$ is the minimizer of $f(\boldsymbol{\Theta}; \boldsymbol{\delta}) := -\sum_{i=1}^{n} ql_i(\boldsymbol{\Theta}; \boldsymbol{\delta}) + \frac{1}{2}\boldsymbol{\delta}^{\intercal}\boldsymbol{\Lambda}^{-1}\boldsymbol{\delta}$. Thus, iteratively solving (4.5) also requires updating the solution $\tilde{\boldsymbol{\delta}}$ at each step until convergence. Conditioning on the previous solution for $\boldsymbol{\Theta}$, $\tilde{\boldsymbol{\delta}}$ is obtained by minimizing a generalized ridge weighted least-squares (WLS) problem with $\boldsymbol{\Lambda}^{-1}$ as the regularization matrix. Then, conditioning on $\tilde{\boldsymbol{\delta}}$, $\hat{\boldsymbol{\Theta}}$ is found by minimizing a WLS problem with a sparse group lasso penalty. We present in Appendix B our proposed proximal Newton-type algorithm that cycles through updates of $\tilde{\boldsymbol{\delta}}$ and $\boldsymbol{\Theta}$.

4.3 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` (Friedman et al. [2010b]). Then, among logistic models that impose hierarchical interactions, we compared our method with the `glinternet` (Lim and Hastie [2015]) and `gesso` (Zemlianskaia et al. [2022]) 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, fitting a total of 500 models. The default implementation for `gesso` is to solve the minimization problem over a 20 by 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.

Table 4.1: 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

4.3.1 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 4,097 whole genomes from the Human Genome Diversity Project (HGDP) and the 1000 Genomes Project (1000G) (Koenig et al. [2023]). 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 $\epsilon \sim \mathcal{N}(0, h_g^2\sigma^2\boldsymbol{K} + h_d^2\sigma^2\boldsymbol{K^D})$, 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 $\boldsymbol{K}$ and $\boldsymbol{K^D}$ 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 $\boldsymbol{K}$. 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 $i = 1, ..., 4097$, binary phenotypes were generated using the following model

$$\text{logit}(\pi) = \text{logit}(\pi_{0k}) - \log(1.3) \times Sex + \log(1.05)Age/10 + \sum_{j \in S} \beta_j \widetilde{G}_j + \sum_{j \in S'} \gamma_j \cdot (Sex \times \widetilde{G}_j) + \epsilon, \tag{4.6}$$

where π_{0k}, for $k = 1, ..., 7$, was simulated using a $U(0.1, 0.9)$ distribution to specify a different prevalence for each population in Table 4.1 under the null, and $\widetilde{G}_j$ 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 $j = 1, ..., p$, 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.

4.3.2 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 $\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0)$ for the main effects, and $\sum_{j=1}^{p} \mathbb{I}(\hat{\gamma}_j \neq 0)$ 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.

4.3.3 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, i.e. the proportion of selected predictors that are causal, over 100 replications for the selection of GEI effects (Figure 4.1) and main genetic effects (Figure 4.2) respectively. We see from Figure 4.1 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 (Zhao et al. [2009]). On the other hand, we see from Figure 4.2 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.

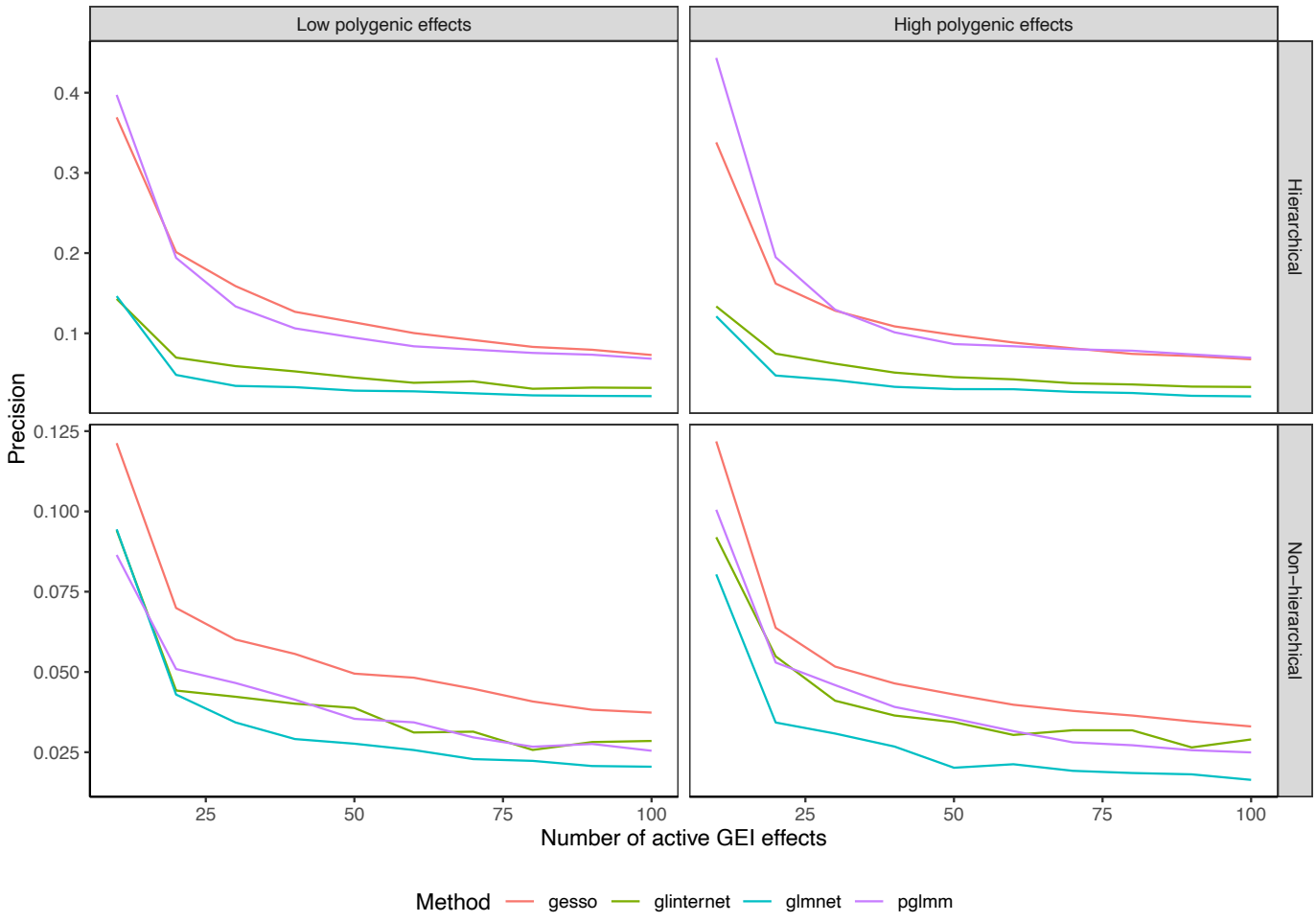
In practice, we often do not have any a priori knowledge for the number of main effects

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, false positive rate (FPR), true positive rate (TPR), false discovery rate (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 K (`plmm (1 Random effect (RE))`) and when including both K and K^D matrices (`pglmm (2 REs)`).

With respect to selection of the GEI effects (Table 4.2), 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.

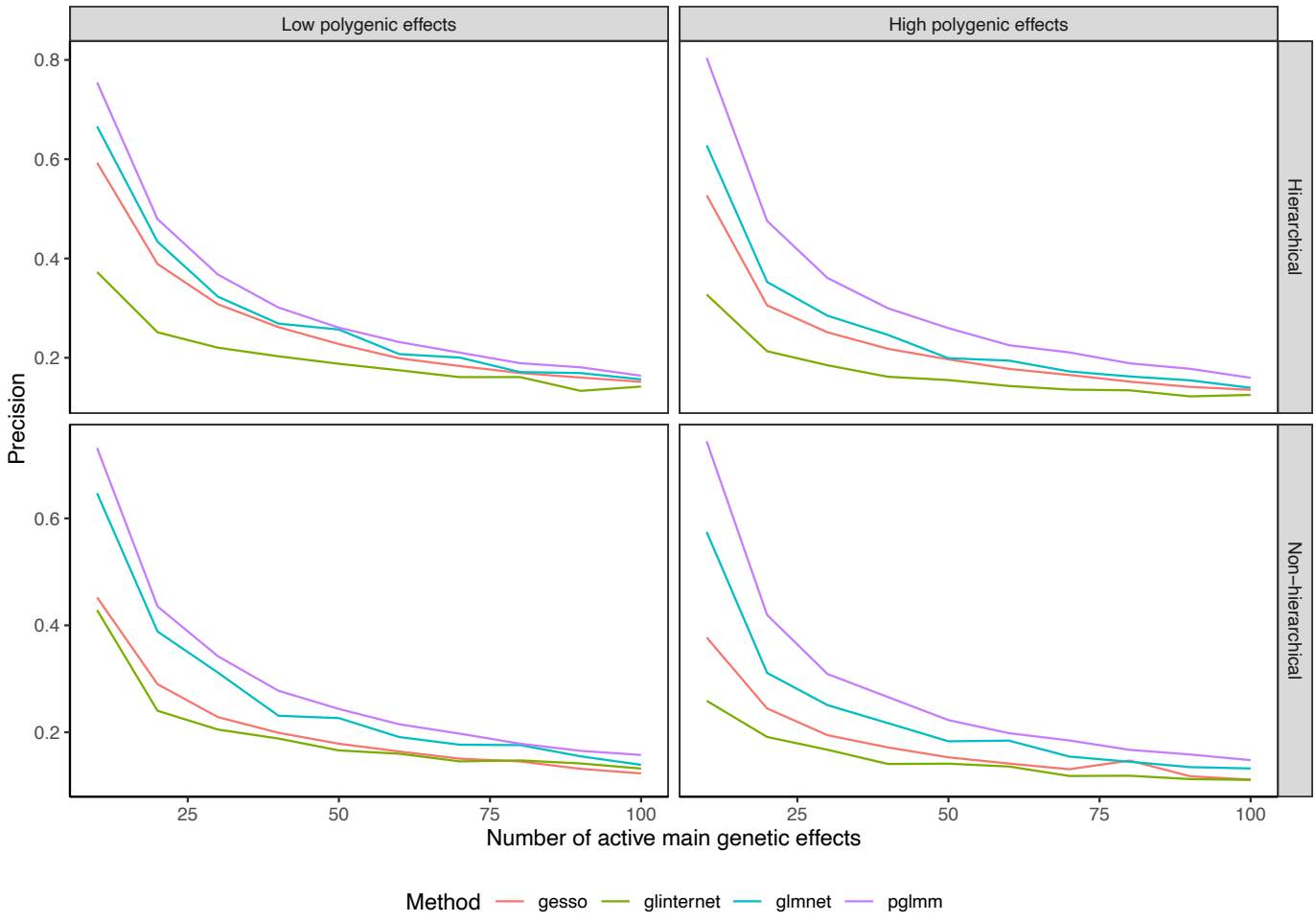
With respect to the genetic main effects (Table 4.3), `pglmm` selects the lowest number of predictors in the model, and thus has the lowest FPR and FDR in all simulation scenarios. Again, adding an additional random effect reduces the FPR for `pglmm`, but to a lower extent than for selection of GEI effects. 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

Figure 4.1: Precision of compared methods averaged over 100 replications as a function of the number of active GEI effects.



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 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 4.4. We see that `pglmm` with two random effects outperforms all other methods for all simulation scenarios.

Figure 4.2: Precision of compared methods averaged over 100 replications as a function of the number of active main effects in the model.



4.4 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 Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study cohort (Smith et al. [2018]). Moreover, TMD has much greater prevalence in females than in males and is believed to have some sex-specific pathophysiologic mechanisms (Bueno et al. [2018]). In this analysis, we wanted to explore the com-

parative performance of our 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 4.5, and further details on study design, recruitment, subject characteristics, and phenotyping for each study are provided in the Supplementary Materials of Smith et al. [2018] (available at `http://links.lww.com/PAIN/A688`).

We used the imputed data described in Smith et al. [2018]. Genotypes were imputed to the 1000 Genomes Project phase 3 reference panel using the software packages SHAPEIT (Delaneau et al. [2011]) for prephasing and IMPUTE version 2 (Howie et al. [2009]). 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 selection (Marees et al. [2018]) ($< 10^{-6}$ in controls, $< 10^{-11}$ in cases). We filtered using a SNP call rate greater than 95% on the combined dataset to retain imputed variants present in all cohorts, which resulted in a total of $4.8M$ imputed SNPs. PCs and kinship matrices were calculated on the merged genotype data using the `-pca` and `-make-rel` flags in PLINK (Chang et al. [2015]), 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 243K predictors.

We present in Table 4.6 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 `glinternet` package due to computational considerations, its memory requirement being too large for the joint analysis of the 243K 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): $[1.26; 1.61]$, $P = 2.82 \times 10^{-8}$) (Smith et al. [2018]). 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 4.7 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 7 SNPs, among which 3 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. 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, `pglmm` 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, `pglmm` 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`.

Table 4.2: Results for the GEI effects γ. 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.

		Non-hierarchical model		Hierarchical model	
Metric	Method	Low ϵ	High ϵ	Low ϵ	High ϵ
Model size	`pglmm (1 RE)`	84.6	99.2	95.5	103
	`pglmm (2 REs)`	58.4	57.8	63.9	59.8
	`glmnet`	21.6	38.8	22.7	38.6
	`glinternet`	17.5	39.2	19.5	39.6
	`gesso`	64.3	102	78.6	110
FPR	`pglmm (1 RE)`	8.31×10^{-3}	9.76×10^{-3}	8.96×10^{-3}	9.69×10^{-3}
	`pglmm (2 REs)`	5.72×10^{-3}	5.65×10^{-3}	6.00×10^{-3}	5.55×10^{-3}
	`glmnet`	2.09×10^{-3}	3.80×10^{-3}	2.20×10^{-3}	$\mathbf{3.76\times10^{-3}}$
	`glinternet`	$\mathbf{1.68\times10^{-3}}$	$\mathbf{3.79\times10^{-3}}$	$\mathbf{1.84\times10^{-3}}$	3.79×10^{-3}
	`gesso`	6.20×10^{-3}	9.94×10^{-3}	7.38×10^{-3}	1.06×10^{-2}
TPR	`pglmm (1 RE)`	0.039	0.043	**0.126**	**0.124**
	`pglmm (2 REs)`	0.030	0.031	0.084	0.091
	`glmnet`	0.016	0.020	0.018	0.025
	`glinternet`	0.016	0.030	0.024	0.038
	`gesso`	**0.052**	**0.068**	0.104	0.103
FDR	`pglmm (1 RE)`	0.966	0.965	0.926	**0.908**
	`pglmm (2 REs)`	**0.936**	0.968	0.921	**0.889**
	`glmnet`	0.962	0.974	0.955	0.967
	`glinternet`	0.948	0.956	0.923	0.949
	`gesso`	0.945	**0.948**	**0.912**	0.930
F_1	`pglmm (1 RE)`	0.035	0.033	**0.091**	**0.084**
	`pglmm (2 REs)`	0.039	0.036	0.082	**0.090**
	`glmnet`	0.036	0.033	0.035	0.036
	`glinternet`	0.040	0.038	0.045	0.048
	`gesso`	**0.052**	**0.048**	0.083	0.067

Model size is defined as $\sum_{j=1}^{p} \mathbb{I}(\hat{\gamma}_j \neq 0)$.
FPR is defined as $\sum_{j=1}^{p} \mathbb{I}(\hat{\gamma}_j \neq 0 \ \cap \gamma_j = 0) / \sum_{j=1}^{p} \mathbb{I}(\gamma_j = 0)$.
TPR is defined as $\sum_{j=1}^{p} \mathbb{I}(\hat{\gamma}_j \neq 0 \ \cap \gamma_j \neq 0) / \sum_{j=1}^{p} \mathbb{I}(\gamma_j \neq 0)$.
FDR is defined as $\sum_{j=1}^{p} \mathbb{I}(\hat{\gamma}_j \neq 0 \ \cap \gamma_j = 0) / \sum_{j=1}^{p} \mathbb{I}(\hat{\gamma}_j \neq 0)$.
F_1 is defined as $2 \times \left(\frac{1}{1-FDR} + \frac{1}{TPR}\right)^{-1}$.

Table 4.3: Results for the genetic predictors main effects β. 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.

		Non-hierarchical model		Hierarchical model	
Metric	Method	Low ϵ	High ϵ	Low ϵ	High ϵ
Model size	`pglmm (1 RE)`	227	212	220	204
	`pglmm (2 REs)`	206	190	214	179
	`glmnet`	278	444	286	450
	`glinternet`	299	481	312	480
	`gesso`	212	361	224	367
FPR	`pglmm (1 RE)`	2.09×10^{-2}	$\mathbf{1.96\times10^{-2}}$	$\mathbf{2.01\times10^{-2}}$	$\mathbf{1.87\times10^{-2}}$
	`pglmm (2 REs)`	$\mathbf{1.89\times10^{-2}}$	$\mathbf{1.74\times10^{-2}}$	$\mathbf{1.95\times10^{-2}}$	$\mathbf{1.62\times10^{-2}}$
	`glmnet`	2.59×10^{-2}	4.24×10^{-2}	2.66×10^{-2}	4.29×10^{-2}
	`glinternet`	2.80×10^{-2}	4.61×10^{-2}	2.91×10^{-2}	4.57×10^{-2}
	`gesso`	1.95×10^{-2}	3.42×10^{-2}	2.05×10^{-2}	3.47×10^{-2}
TPR	`pglmm (1 RE)`	0.195	0.181	0.208	0.196
	`pglmm (2 REs)`	0.188	0.177	0.206	0.188
	`glmnet`	0.215	0.244	0.226	0.257
	`glinternet`	**0.216**	**0.246**	**0.237**	**0.271**
	`gesso`	0.190	0.220	0.210	0.238
FDR	`pglmm (1 RE)`	**0.895**	**0.895**	**0.891**	**0.883**
	`pglmm (2 REs)`	**0.888**	**0.885**	**0.885**	**0.870**
	`glmnet`	0.920	0.944	0.917	0.942
	`glinternet`	0.925	0.948	0.922	0.943
	`gesso`	0.906	0.937	0.903	0.932
F_1	`pglmm (1 RE)`	**0.123**	**0.120**	**0.136**	**0.134**
	`pglmm (2 REs)`	**0.126**	**0.124**	**0.138**	**0.140**
	`glmnet`	0.115	0.091	0.119	0.094
	`glinternet`	0.109	0.085	0.116	0.094
	`gesso`	0.123	0.096	0.131	0.103

Model size is defined as $\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0)$.
FPR is defined as $\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0 \ \cap \beta_j = 0) / \sum_{j=1}^{p} \mathbb{I}(\beta_j = 0)$.
TPR is defined as $\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0 \ \cap \beta_j \neq 0) / \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)$.
FDR is defined as $\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0 \ \cap \beta_j = 0) / \sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0)$.
F_1 is defined as $2 \times \left(\frac{1}{1-FDR} + \frac{1}{TPR}\right)^{-1}$.

Table 4.4: Results for the prediction accuracy of a binary outcome on test sets. 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.

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

Table 4.5: 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

Table 4.6: 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. 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.

			`pglmm`			`gesso`		`glmnet`
Chromosome	Position	OR_β	OR_γ	$OR_{\beta+\gamma}$	OR_β	OR_γ	$OR_{\beta+\gamma}$	OR_β
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	-

Table 4.7: Area under the roc curve (AUC), model size and computational time for the analysis of TMD.

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

4.5 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 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 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 hierachichal 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. [2018], and the use of sparse GSMs to adjust for the sample relatedness (Jiang et al. [2019]).

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 (Zhao et al. [2009]), 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) (Fan and Li [2001]) and Minimax Concave Penalty (MCP) (Zhang [2010]) 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 (Meinshausen [2007]).

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 (Janková et al. [2020]). Although there exist graphical and numerical methods for checking the adequacy of GLMMs (Pan and Lin [2005]), 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 (Shen et al. [2013]), or replacing it by a lasso regularization to perform selection of individual random effects (Hui et al. [2017], Bondell et al. [2010]).

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 S.B. 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 authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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.].

Supplemental material

Our Julia package PenalizedGLMM and codes for simulating data are available on github. The data from the real data application that supports the findings of this study are available from the corresponding author upon reasonable request.

References

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3), June 2013. ISSN 0090-5364. 10.1214/13-aos1096. URL http://dx.doi.org/10.1214/13-AOS1096.

Howard D. Bondell, Arun Krishna, and Sujit K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, December 2010. ISSN 1541-0420. 10.1111/j.1541-0420.2010.01391.x. URL http://dx.doi.org/10.1111/j.1541-0420.2010.01391.x.

Norman E. Breslow and David G. Clayton. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421):9–25, March 1993. ISSN 0162-1459, 1537-274X. 10.1080/01621459.1993.10594284.

C. H. Bueno, D. D. Pereira, M. P. Pattussi, P. K. Grossi, and M. L. Grossi. Gender differences in temporomandibular disorders in adult populational studies: A systematic review and meta-analysis. *Journal of Oral Rehabilitation*, 45(9):720–729, June 2018. 10.1111/joor.12661. URL https://doi.org/10.1111/joor.12661.

Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), February 2015. 10.1186/s13742-015-0047-8. URL https://doi.org/10.1186/s13742-015-0047-8.

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.

David R. Cox. Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1, April 1984. 10.2307/1403235. URL https://doi.org/10.2307/1403235.

Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, December 2011. 10 .1038/nmeth.1785. URL https://doi.org/10.1038/nmeth.1785.

Frank Dudbridge and Olivia Fletcher. Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics*, 95(3):301–307, September 2014. 10.1016/j.ajhg.2014.07.014. URL https://doi.org/10.1016/j.ajhg.2014.07.014.

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.

Kuangnan Fang, Jingmao Li, Qingzhao Zhang, Yaqing Xu, and Shuangge Ma. Pathological imaging-assisted cancer gene–environment interaction analysis. *Biometrics*, May 2023. 10.1111/biom.13873. URL https://doi.org/10.1111/biom.13873.

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.

Gabriel E. Hoffman. Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PLoS ONE*, 8(10):e75707, October 2013. 10.1371/journal.pone.0075707. URL https://doi.org/10.1371/journal.pone.0075707.

Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, June 2009. 10.1371/journal.pgen.1000529. URL https://doi.org/10.1371/journal.pgen.1000529.

Liuyi Hu, Wenbin Lu, Jin Zhou, and Hua Zhou. MM ALGORITHMS FOR VARIANCE COMPONENT ESTIMATION AND SELECTION IN LOGISTIC LINEAR MIXED MODEL. *Statistica Sinica*, 2019. 10.5705/ss.202017.0220. URL https://doi.org/10.5705/ss.202017.0220.

Francis K. C. Hui, Samuel Müller, and A. H. Welsh. Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association*, 112(519):1323–1333, June 2017. ISSN 1537-274X. 10.1080/01621459.2016.1215989. URL http://dx.doi.org/10.1080/01621459.2016.1215989.

Jana Janková, Rajen D. Shah, Peter Bühlmann, and Richard J. Samworth. Goodness-of-fit Testing in High Dimensional Generalized Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):773–795, 05 2020. ISSN 1369-7412. 10.1111/rssb.12371. URL https://doi.org/10.1111/rssb.12371.

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.

Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, July 2015. 10.1080/10618600.2014.938812. URL https://doi.org/10.1080/10618600.2014.938812.

William Maixner, Luda Diatchenko, Ronald Dubner, Roger B. Fillingim, Joel D. Greenspan, Charles Knott, Richard Ohrbach, Bruce Weir, and Gary D. Slade. Orofacial pain prospective evaluation and risk assessment study – the OPPERA study. *The Journal of Pain*, 12(11):T4–T11.e2, November 2011. 10.1016/j.jpain.2011.08.002. URL https://doi.org/10.1016/j.jpain.2011.08.002.

Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009. ISSN 1476-4687. 10.1038/nature08494. URL http://dx.doi.org/10.1038/nature08494.

Andries T. Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):e1608, February 2018. 10.1002/mpr.1608. URL https://doi.org/10.1002/mpr.1608.

Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, September 2007. 10.1016/j.csda.2006.12.019. URL https://doi.org/10.1016/j.csda.2006.12.019.

Bhramar Mukherjee, Jaeil Ahn, Stephen B. Gruber, Malay Ghosh, and Nilanjan Chatterjee. Case-Control Studies of Gene-Environment Interaction: Bayesian Design and Analysis. *Biometrics*, 66(3):934–948, November 2009. 10.1111/j.1541-0420.2009.01357.x. URL https://doi.org/10.1111/j.1541-0420.2009.01357.x.

Zhiying Pan and D. Y. Lin. Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61(4):1000–1009, December 2005. ISSN 1541-0420. 10.1111/j.1541-0420.2005.00365.x. URL http://dx.doi.org/10.1111/j.1541-0420.2005.00365.x.

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.

Junyang Qian, Yosuke Tanigawa, Wenfei Du, Matthew Aguirre, Chris Chang, Robert Tibshirani, Manuel A. Rivas, and Trevor Hastie. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genetics*, 16(10):e1009141, October 2020. 10.1371/journal.pgen.1009141. URL https://doi.org/10.1371/journal.pgen.1009141.

Xia Shen, Moudud Alam, Freddy Fikse, and Lars Rönnegård. A novel generalized ridge

regression method for quantitative genetics. *Genetics*, 193(4):1255–1268, April 2013. 10 .1534/genetics.112.146720. URL https://doi.org/10.1534/genetics.112.146720.

M. Shi, K. M. O'Brien, and C. R. Weinberg. Interactions between a polygenic risk score and non-genetic risk factors in young-onset breast cancer. *Scientific Reports*, 10(1), February 2020. 10.1038/s41598-020-60032-3. URL https://doi.org/10.1038/s41598-020-60032-3.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, April 2013. ISSN 1537-2715. 10.1080/10618600.2012.681250. URL http://dx.doi.org/10.1080/10618600.2012.681250.

Shad B. Smith, Marc Parisien, Eric Bair, Inna Belfer, Anne-Julie Chabot-Doré, Pavel Gris, Samar Khoury, Shannon Tansley, Yelizaveta Torosyan, Dmitri V. Zaykin, Olaf Bernhardt, Priscila de Oliveira Serrano, Richard H. Gracely, Deepti Jain, Marjo-Riitta Järvelin, Linda M. Kaste, Kathleen F. Kerr, Thomas Kocher, Raija Lähdesmäki, Nadia Laniado, Cathy C. Laurie, Cecelia A. Laurie, Minna Männikkö, Carolina B. Meloto, Andrea G. Nackley, Sarah C. Nelson, Paula Pesonen, Margarete C. Ribeiro-Dasilva, Celia M. Rizzatti-Barbosa, Anne E. Sanders, Christian Schwahn, Kirsi Sipilä, Tamar Sofer, Alexander Teumer, Jeffrey S. Mogil, Roger B. Fillingim, Joel D. Greenspan, Richard Ohrbach, Gary D. Slade, William Maixner, and Luda Diatchenko. Genome-wide association reveals contribution of MRAS to painful temporomandibular disorder in males. *Pain*, 160(3):579–591, November 2018. 10.1097/j.pain.0000000000001438. URL https://doi.org/10.1097/j.pain.0000000000001438.

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.

Jae Hoon Sul, Michael Bilow, Wen-Yun Yang, Emrah Kostem, Nick Furlotte, Dan He, and Eleazar Eskin. Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models. *PLOS Genetics*, 12(3): e1005849, March 2016. ISSN 1553-7404. 10.1371/journal.pgen.1005849.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, July 2017. ISSN 0002-9297. 10.1016/j.ajhg.2017.06.005. URL http://dx.doi.org/10.1016/j.ajhg.2017.06.005.

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.

Natalia Zemlianskaia, W. James Gauderman, and Juan Pablo Lewinger. A scalable hierarchical lasso for gene–environment interactions. *Journal of Computational and Graphical Statistics*, pages 1–13, March 2022. 10.1080/10618600.2022.2039161. URL https://doi.org/10.1080/10618600.2022.2039161.

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.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A), December 2009. ISSN 0090-5364. 10.1214/07-aos584. URL http://dx.doi.org/10.1214/07-AOS584.

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 2019. 10.1080/10618600.2018.1529601. URL https://doi.org/10.1080/10618600.2018.1529601.

Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A. Gagliano, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, August 2018. ISSN 1546-1718. 10.1038/s41588-018-0184-y. URL http://dx.doi.org/10.1038/s41588-018-0184-y.

Chapter 5

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

Preamble to Manuscript 3.

In the previous two chapters, a general framework was proposed for accounting for multiple sources of confounding in multivariable regularized models applied to genetic association studies. In this manuscript, we extended the proposed methodology to analysis of longitudinal outcomes. This work was first 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.

The original contributions of this chapter are i) proposing a methodology based on regularized penalized quasi-likelihood (PQL) estimation to perform selection of genetic predictors in sparse regularized longitudinal mixed models, ii) studying the performance of the AIREML algorithm in the estimation of variance components in multilevel mixed models for longitu-

dinal genetic studies of both continuous and binary outcomes in the presence of population structure and family relatedness, and iii) applying the proposed methodology to a real case study to determine the genetic contribution of emotional and behavioural difficulties in childhood and adolescents from two longitudinal cohorts from the province of Quebec.

The manuscript presented in this chapter will be submitted to a statistical journal soon after the submission of this thesis.

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, $(\boldsymbol{\theta}^\intercal, \boldsymbol{\beta}^\intercal)^\intercal$ a $(c+p) \times 1$ column vector of fixed covariate and additive genotype effects including the intercept. We assume that $\mathbf{b}_0 = (b_{01}, ..., b_{0m})^\intercal \sim \mathcal{N}(0, \sum_{k=1}^{K} \tau_k \mathbf{V}_k)$ is an $m \times 1$ column vector of random intercepts, $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)^\intercal$ 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 $\mathbf{b}_{1i} = (b_{11i}, b_{12i}, ..., b_{1ri})^\intercal \sim \mathcal{N}(0, \mathbf{D}(\psi))$ 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 $\text{Var}(y_{ij}|\mathbf{C}_{ij}, \mathbf{G}_i, \mathbf{Z}_{ij}, b_{0i}, \mathbf{b}_{1i}) = \phi w_{ij}^{-1} \nu(\mu_{ij})$, 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 $\mathbf{b}_1 = (\mathbf{b}_{11}^\intercal, ..., \mathbf{b}_{1r}^\intercal)$, such that the stacked random individual effects vector is $\mathbf{b} = (\mathbf{b}_0^\intercal, \mathbf{b}_{11}^\intercal, ..., \mathbf{b}_{1r}^\intercal)^\intercal \sim N(0, \text{diag}\left\{\sum_{k=1}^{K} \tau_k \mathbf{V}_k, \mathbf{D} \otimes \mathbf{I}_m\right\})$,

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 k^{th} 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 $(\boldsymbol{\theta}^\intercal, \boldsymbol{\beta}^\intercal, \phi, \boldsymbol{\tau}, \boldsymbol{\psi})$ 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

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

and $\boldsymbol{\Theta} = (\boldsymbol{\theta}^\intercal, \boldsymbol{\beta}^\intercal)^\intercal$, $\mathbf{W} = \phi^{-1}\boldsymbol{\Delta}^{-1} = \phi^{-1}\text{diag}\left\{\frac{a_{ij}}{\nu(\mu_{ij})[g'(\mu_{ij})^2]}\right\}$ 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 $\sum_{i,j} ql_{ij}(\boldsymbol{\Theta}, \mathbf{b}) - \frac{1}{2}\mathbf{b}^\intercal\left(\text{diag}\left\{\sum_{k=1}^{K}\tau_k\mathbf{V}_k, \mathbf{D}\otimes\mathbf{I}_m\right\}\right)^{-1}\mathbf{b}$.

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 $(\boldsymbol{\Theta}, \phi, \boldsymbol{\tau}, \boldsymbol{\psi})$:

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

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

$$y_{ij} = \text{logit}(\pi_{0k}) - \log(1.3) \times Sex_i + \log(1.05) \times Age_{ij} + \sum_{s=1}^{100} \beta_s \cdot G_{is} + b_{0i} + \mathbf{Z}_{ij}\mathbf{b}_{1i} + \epsilon_{ij}, \tag{5.4}$$

where G_{is} is the (standardized) number of alleles for the s^{th} 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

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

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.

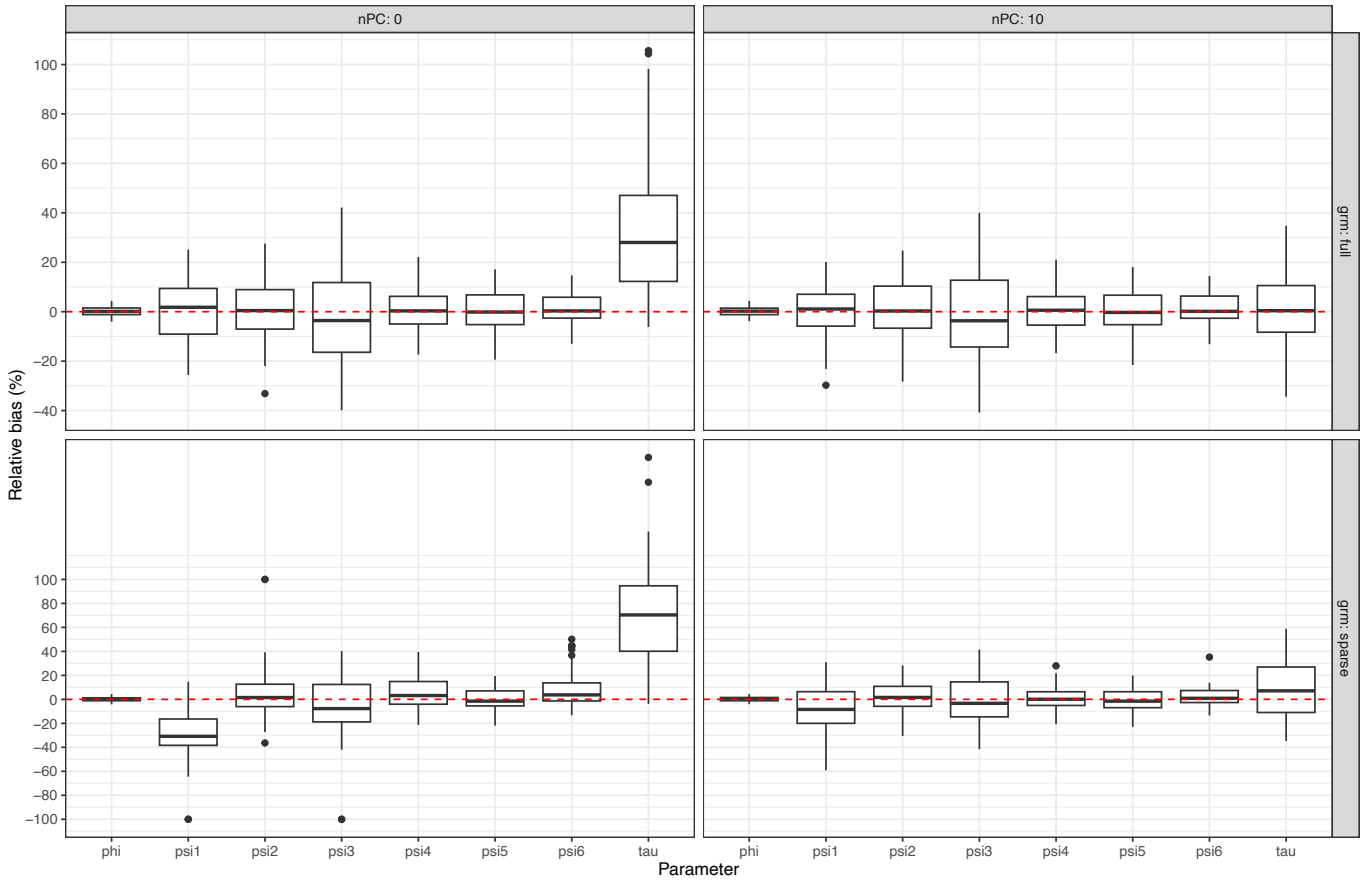


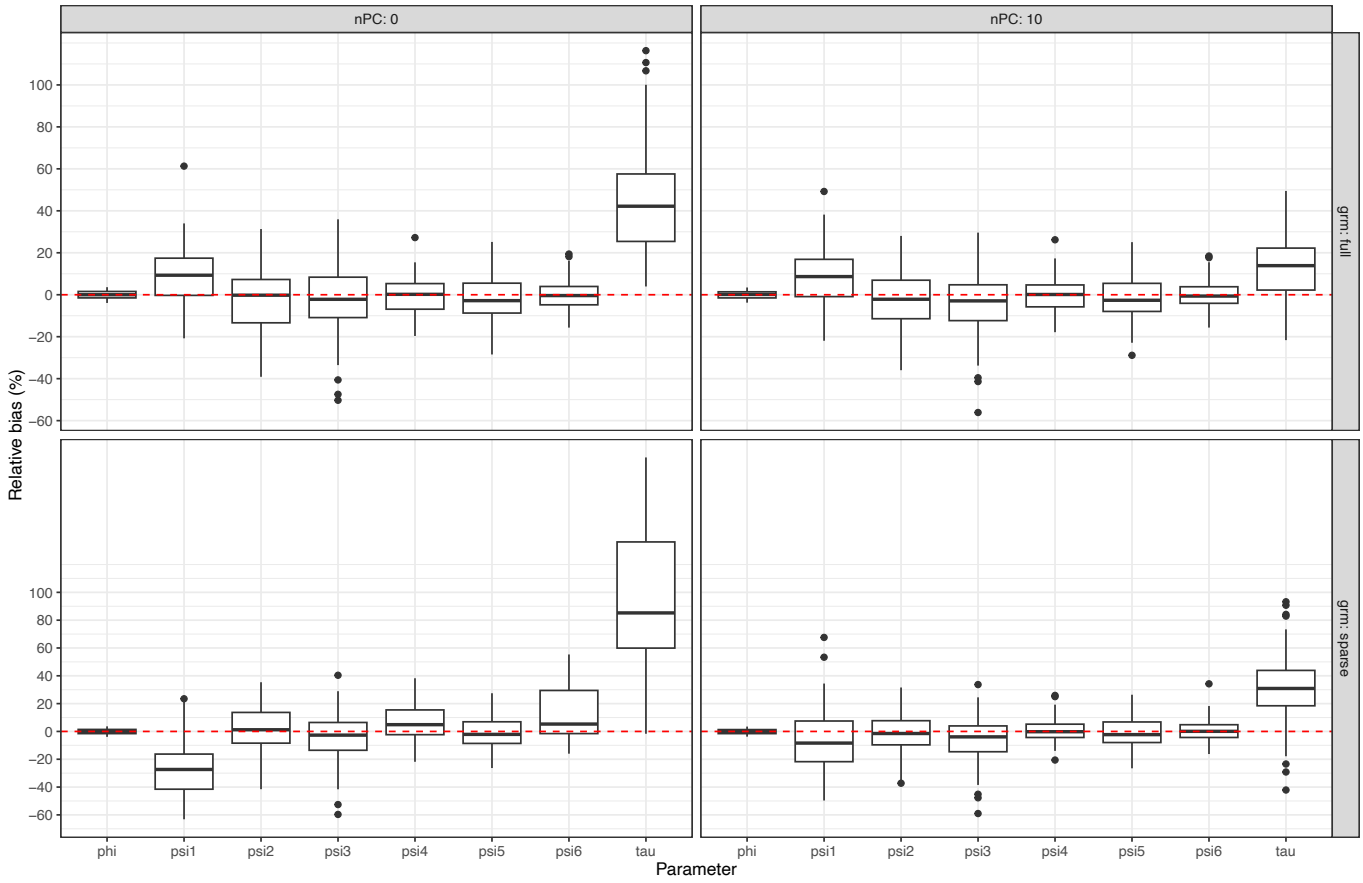
Table 5.2: Computation time in minutes to fit the AI-REML algorithm as a function of the modelling strategy for the simulation model with no causal predictor (0% heritability) and for the simulation model with 100 causal predictors explaining 2% of heritability for both continuous and binary phenotypes. We present the median value with IQR in brackets.

		Binary phenotypes		Continuous phenotypes	
Number of PCs	GRM	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.

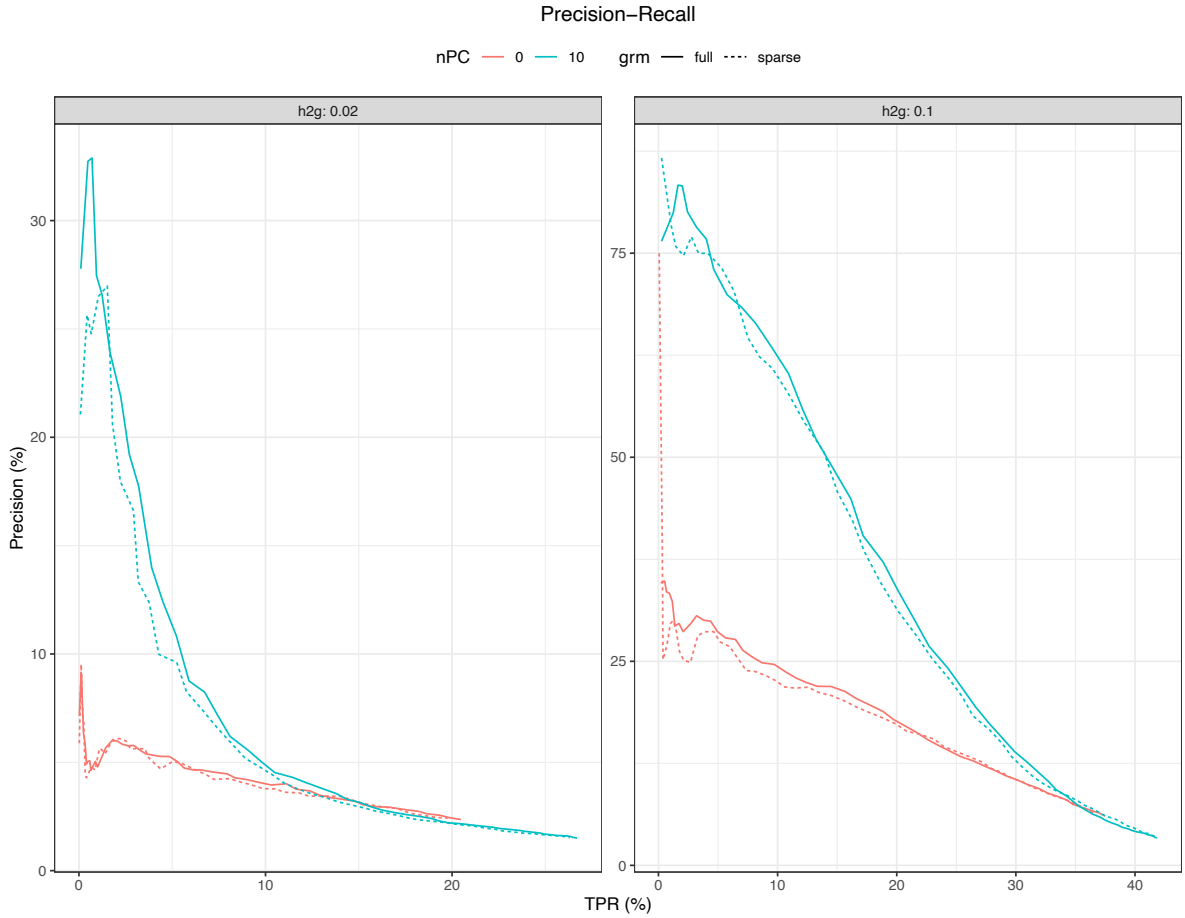


Selection of genetic predictors

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.



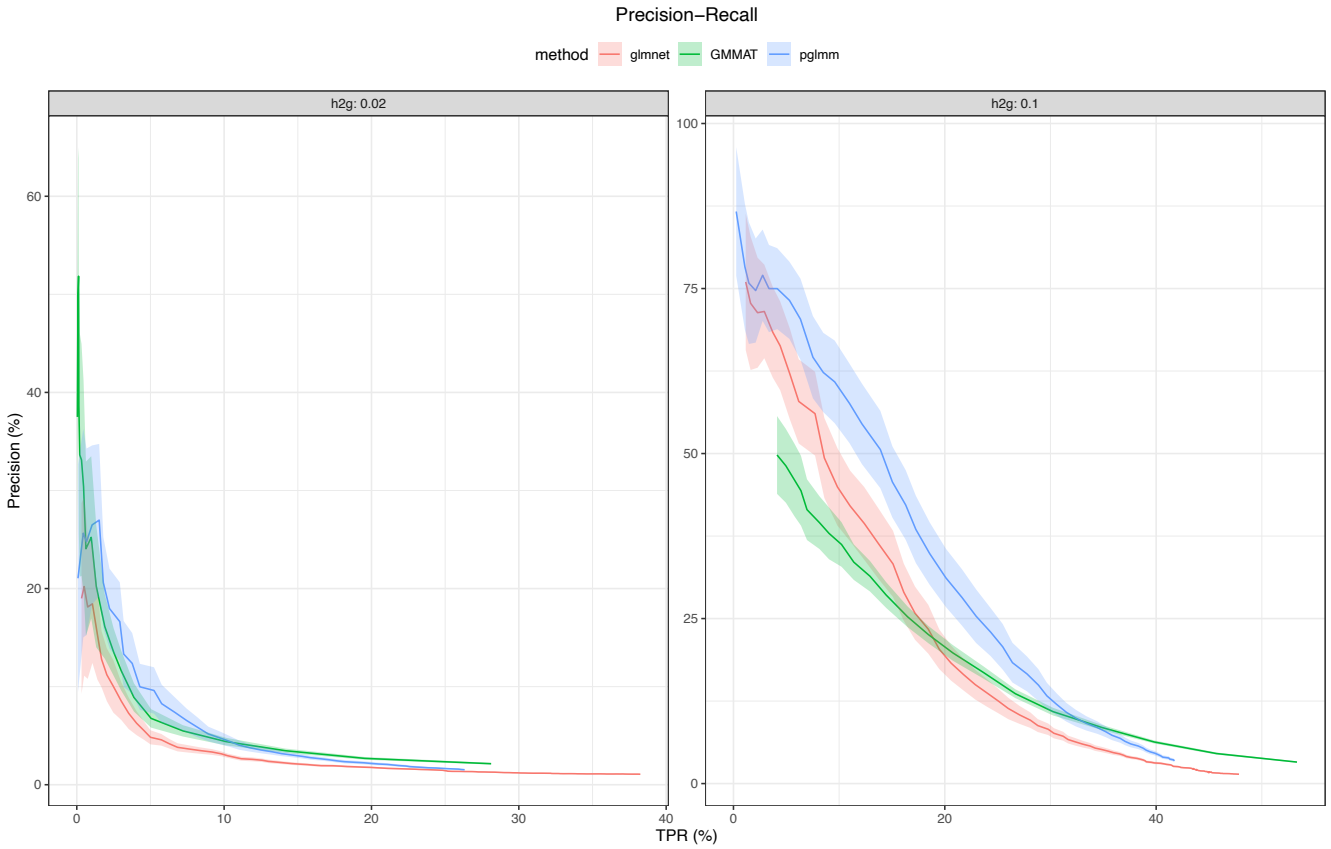
ing values of λ. For these two methods, we used a grid of 100 values of λ on the log10 scale with $\lambda_{min} = 0.01\lambda_{max}$, where λ_{max} is chosen such that no predictors are selected in the model. We also compared the performance of our method with that of GMMAT, using the freely available `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.

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

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.

		Lasso mixed model			Adaptive lasso mixed model		
Externalizing score	Analysis	Number of selected genetic predictors	MSPE	R^2_{MSPE}	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

hierarchichal 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. OpenMendel: 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.

Chapter 6

Conclusion

6.1 Summary

In this thesis, I have proposed computationally efficient mixed-effects models for variable selection in the context of genetic association studies to address the spurious selection of variables due to confounding by population stratification and close relatedness, and the computational complexity associated with fitting high-dimensional mixed models.

In the first manuscript presented in Chapter 3, a lasso regularized GLMM for selecting important predictors and estimating their effects in high-dimensional GWAS data was proposed and implemented. Building on the work of Bhatnagar et al. [2020a] for lasso regularized LMMs, I relaxed the normality assumption of the phenotype distribution, and obtained an analytical form for the loss function by using PQL estimation. In PQL estimation, because the random effects vector is treated as a fixed effect parameter, it is estimated in a similar manner to other fixed effects as in a GLM. Thus, minimizing the lasso regularized PQL loss function with respect to the fixed and random effects respectively is equivalent to iteratively solving two penalized WLS problems. By using the spectral decomposition of the variance-covariance matrix to rotate the data, the estimate for the random effects vector can

be obtained as the solution to a generalized ridge WLS problem where the fixed effect parameters of the model are held constant. By profiling out the random effects vector estimate from the objective function and replacing it by its PQL estimate, the solution for the genetic predictors fixed effects are obtained by solving a simple lasso regularized WLS regression. Thus, one could use an existing implementation of lasso regularization such as `glmnet` to obtain estimates for the genetic predictors fixed effects. Nonetheless, I decided to implement the coordinate descent algorithm using the `julia` programming language (Bezanson et al. [2017]) to make it reusable and adaptable for the work presented in Chapters 4 and 5.

A key distinction with the approach of Bhatnagar et al. [2020a] is that I proposed estimating the variance components of the model only once under the null hypothesis of no genetic effects. This is the typical approach in mixed-model association tests and is known as the P3D (population parameters previously determined) method (Zhang et al. [2010]). While I did not assess the implications of this approach on the estimation and selection accuracy of my proposed model, it is important to note that the impact of reestimating the variance components on the computational requirements would be high. Indeed, by proposing to use a lower bound on the variance-covariance matrix of the working data (Böhning and Lindsay [1988]), I showed that a single spectral decomposition of this matrix is needed, similarly to the factored spectrally transformed linear mixed model (FaST-LMM) algorithm proposed by Lippert et al. [2011]. For GLMMs, this is only possible if the variance components of the model are considered to be known, or estimated once under the null model. Moreover, I showed in Section 3.2.2 of Chapter 3 that including a random effect with variance-covariance proportional to a known GRM is equivalent to a ridge regularized regression where each PC's fixed effect is shrinked by a factor $\hat{\tau_1}^{-1}\Lambda_i^{-1}$ with $\hat{\tau_1}^{-1}$ the estimated variance component for the polygenic random effect and Λ_i the i^{th} eigenvalue of the GRM. Thus, an alternative approach to estimating the polygenic random effect variance under the null model would be to consider τ as an additional tuning parameter of the model, similarly to the lasso tuning parameter λ. An analogous approach has been proposed in the literature for LMMs

by Runcie and Crawford [2019].

By simulating random genotypes from the BN-PSD admixture model (Ochoa and Storey [2021]) for 10 or 20 subpopulations with one dimensional geography or independent subpopulations, we showed that including a random effect with variance-covariance structure proportional to the GRM has better selection, estimation and prediction accuracy than a logistic lasso with PC adjustment when the number of subpopulations was greater than the number of PCs included. I also showed that the lasso penalized LMM proposed by Bhatnagar et al. [2020a] was unable to estimate predictor effects with accuracy for binary responses, which greatly decreased its predictive performance compared to our proposed lasso penalized GLMM. In the second simulation scenario, I sampled real genotype data from a subset of related individuals from the UK Biobank data, and showed that our proposed model effectively led to sparser models with higher precision and prediction accuracy than the lasso LMM from Bhatnagar et al. [2020a] and a standard logistic lasso model with or without PC adjustment. It was also demonstrated that using AIC as a model selection strategy led to similar prediction performance than cross-validation, with even sparser models. Finally, I demonstrated through the analysis of two polygenic binary traits in a subset of 6731 related individuals from the UK Biobank data that our method achieved higher predictive performance, while also selecting consistently fewer predictors than a logistic lasso with PC adjustment. Using sequential strong rules for solving the lasso problem such that most of the 320,000 analyzed predictors were discarded from the optimization problem at each iteration [Tibshirani et al., 2012], fitting the full lasso path for the real case study took a median time of 53 minutes. Thus, for relatively small sample sizes, our proposed sparse regularized logistic mixed model remains computationally efficient even when the number of genetic variants is very large.

In the second manuscript presented in Chapter 4, a unified approach based on regularized PQL estimation was derived for selecting important main genetic and GEI effects in high-

dimensional GWAS data, while imposing hierarchy between main and interaction effects, and accounting for shared environmental exposure that may induce additional relatedness between individuals. To perform hierarchical selection of main genetic and GEI effects, that is an interaction term can only be selected in the model if the corresponding main effects are different from zero, a sparse group lasso penalty was added to the PQL loss function. Indeed, by regrouping together the main genetic and GEI effects and applying a group lasso penalty, it ensured that hierarchical selection was maintained. Adding an additional lasso penalty for the GEI effects allowed for main genetic effects to be selected in the model without any interaction effect. We derived a proximal Newton-type algorithm with BCD to find coordinate-wise updates of the fixed and random effects vector. As opposed to the work presented in Chapter 3, we did not profile out the random effects vector estimate from the objective function, because this approach requires rotating the phenotype and genotype using the eigenvectors of the random effects covariance matrix which is computationally demanding and requires storing in memory a very large amount of data. In fact, by showing that the random effects vector estimate is obtained as the solution to a generalized ridge WLS problem, we proposed estimating it using coordinate descent rather than Newton's method, which does not require using a lower bound on the variance-covariance matrix of the working data. Thus, our proposed methodology is applicable to other exponential family distributions such as the Poisson distribution for which no lower bound for the variance-covariance matrix exists.

Simulations results showed that including and additional GRM to account for the shared environmental exposure, as first proposed by Sul et al. [2016] when testing for association in single-SNP mixed models, reduced the false positive rate (FPR) for selection of both GEI and main effects. Using the F_1 score as a balanced measure of the false discovery rate (FDR) and true positive rate (TPR), we showed that in the hierarchical simulation scenarios, our proposed method outperformed comparative methods for retrieving important GEI effects. Moreover, using real data from the OPPERA study to explore the performance of our method

in selecting important predictors of TMD, we were able to retrieve a previously reported significant loci in a combined or sex-segregated GWAS.

In the third manuscript presented in Chapter 5, the methodology derived in the previous two manuscripts was extended 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. The performance of the AIREML algorithm when simulating population structure and subjects relatedness for continuous and binary traits was studied, and it was shown that using the PC-AiR method (Conomos et al. [2015]) to calculate PCs that account for genetic correlations due to distant common ancestry and the PC-Relate method (Conomos et al. [2016]) 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 for continuous traits. For binary traits, the polygenic random effect heritability was consistently overestimated, even when simulating data under no genetic association, that is when the model was correctly specified. It is reported in the literature that PQL estimation results in biased estimators of both regression coefficients and variance components in GLMMs, especially for binary outcomes (Jang and Lim [2009]), and first-order and second-order correction procedures were proposed to correct for this bias (Breslow and Lin [1995], Lin and Breslow [1996]). Although accurate estimation of variance components in mixed models is often not of interest or extensively discussed in genetic association studies, it was demonstrated in the simulations of Chapter 5 that increased relatives biases were associated with an increase of the FDRs in the penalized model. Moreover, overestimation of the polygenic heritability results in overfitting of the contribution of the GRM in explaining the observed phenotypic variance, meaning 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 observations that were not used in the training model, such as

cross-validation test sets or replication data sets.

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 the GMMAT model proposed by Chen et al. [2016] and that of a lasso penalized model without any random effect. In a real-world case study aimed at identifying key genetic predictors of aggression, hyperactivity, and opposition externalizing behaviours in children and adolescents from the Quebec Longitudinal Study of Child Development (QLSCD) and Quebec Newborn Twin Study (QNTS) cohorts, we applied both the GMMAT univariate model and our proposed penalized mixed model. The two methods identified different sets of potentially important SNPs. To address missing data due to non-response or loss to follow-up, we used a single imputation method and compared findings with a complete case analysis. For opposition externalizing behaviour, two SNPs (rs116541675, rs145226968) were genome-wide significant (p-value $< 5 \times 10^{-8}$) in both the single imputation and complete case analyses. For hyperactivity and aggression behaviors, we identified nine SNPs that were significant only in the complete case analysis. Additionally, we demonstrated the utility of our proposed method in predicting externalizing behavior scores in children from the combined QLSCD and QNTS cohorts. Using a mean squared prediction error based definition of the R^2 coefficient of determination as proposed by Staerk et al. [2024], we showed that the predictions were well-calibrated, and that including genetic effects improved prediction accuracy compared to a model without genetic contributions. Furthermore, we fitted an adaptive lasso penalized mixed model, with weights inversely proportional to effect size estimates from elastic-net regularized regression. This model resulted in sparser solutions than the lasso mixed model, while maintaining comparable accuracy in the predicted scores.

6.2 Limitations and future directions

Important limitations of the methodology developed in this thesis are discussed in this section.

The regularized mixed models framework presented in Chapter 3 to Chapter 5 relies upon the lasso and sparse group lasso penalties as a regularization technique for simultaneous estimation and variable selection. However, because the lasso shrinkage produces biased estimates for larger coefficients, Fan and Li [2001] demonstrated that the oracle property does not hold for the lasso and proposed a SCAD penalty for variable selection. The SCAD penalty retains the penalization rate and bias of the lasso for small coefficients, but continuously relaxes the rate of penalization as the absolute value of the coefficient increases through the use of a second tuning parameter (usually denoted γ) that controls the concavity of the penalty. Another folded concave penalty, the MCP, was proposed by Zhang [2010] to address the bias and non-consistency of the lasso regularization estimates. However, because the penalty functions for SCAD and MCP are concave, there are numerical challenges in fitting these models. Indeed, concave objective functions are difficult to optimize and often produce unstable solutions. Breheny and Huang [2011] discussed the convergence properties of coordinate descent algorithms for fitting MCP and SCAD penalties within the linear regression framework, and proposed a fast, efficient and stable algorithm available in an open-source R package called `ncvreg`. An alternative method is to fit adaptive lasso mixed models as presented in the real case study of Chapter 5, where adaptive weights are used for penalizing coefficients differently, as proposed by Zou [2006]. Future work could entail allowing for concave penalties for selection of genetic predictors in high dimensional genetic association studies, as was done in (Chung et al. [2019], Huang et al. [2013]). Concave penalties can also be combined in a manner analogous to the sparse group lasso to perform hierarchical variable selection for linear and logistic regression models (Buch et al. [2024]).

Another limitation of the lasso regularization is that it is known that when correlations

between predictors are strong and effect sizes are small, the FDR for the models selected by the lasso procedure can be important (Su et al. [2017]). Indeed, when the correlations between predictors is high, as it is the case with SNPs in blocks of LD, the lasso will usually select only one predictor randomly from a group of correlated predictors. Moreover, it is empirically expected that the prediction performance of ridge regression (Hoerl and Kennard [1970]) is superior to that of the lasso regression (Tibshirani [1996]) when pair-wise correlations between predictors is important. By adding a constraint on the ℓ_2 norm of the predictors coefficients, ridge regularization ensures that correlated predictors coefficients estimates are shrunk towards a common value when there is multicollinearity. However, ridge regression does not induce sparsity in the predictors coefficients, hence it cannot perform automatic variable selection. In high-dimensional models, where the number of predictors is much greater than sample size, a convex combination of the lasso and ridge regularization, referred to as the elastic-net penalty (Zou and Hastie [2005]), can be used to induce sparsity in the predictors coefficients estimates while shrinking predictors' effect sizes that are highly correlated. An adaptive elastic-net penalty framework has also been proposed by Zou and Zhang [2009] to combine the strengths of the quadratic regularization and the adaptively weighted lasso shrinkage. The authors further established the oracle property of the adaptive elastic-net under weak regularity conditions. An alternative approach was proposed by Huang et al. [2013] in which the authors proposed a novel penalty, the smoothed minimax concave penalty (SMCP) that is a combination of the MCP and a penalty consisting of the squared differences of the absolute effects of adjacent markers. Their proposed penalization framework explicitly uses correlation between adjacent markers and penalizes the differences of genetic effects at adjacent SNPs with high correlations. Future work could consider extending our proposed penalized mixed models framework to other penalizations such as the SMCP or adaptive elastic-net penalties to address the challenges stemming from the presence of causal SNPs in LD.

A limitation of the proposed methodology pertains to the computational burden of esti-

mating variance components for high-dimensional mixed models, which relies on performing a spectral or Cholesky decomposition of the covariance matrix. We partially circumvented this challenge by estimating the random effects variance parameters only once under the null hypothesis of no genetic association, as typically done in genetic univariate association tests, such that the number of decompositions required remained low. However, for very large sample sizes, computing a Cholesky or spectral decomposition of the covariance matrix can be very burdensome, with complexity of $O(n^3)$, where n is the sample size. For low dimensional confounding effects such as population structure, adjusting for the top eigenvectors (PCs) of an ancestry representative GRM is usually enough. On the other hand, for higher dimensional confounding effects such as family or cryptic relatedness, including a random effect with variance-covariance proportional to the GRM is warranted. A possible approach would be to use the implicitly restarted Arnoldi method (Abraham et al. [2017]) to obtain only the top k eigenvectors of the covariance matrix, instead of computing its full spectral decomposition. However, estimating the dimensionality of real datasets, and thus the number of eigenvectors to include is still a challenging open question, notably because estimated eigenvalues have biased distributions (Yao and Ochoa [2022]). Another alternative that we followed in Chapter 5 is to use a sparse GRM to adjust for the sample relatedness combined with efficient sparse matrix-based algorithms for parameter estimation (Jiang et al. [2019]). However, because a sparse GRM cannot incorporate polygenic effects, as contributions from small effects are effectively replaced by zero, it may be less powerful than using a dense GRM (Bi et al. [2021]). Finally, another solution to explore 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. [2018].

In GWAS, confidence intervals for the effect size estimates of predictors and their related standard errors and p-values are of primary interest. A limitation of penalized multivariable regression models is that they do not allow to perform inference on the individual statistical significance of the selected predictors. Indeed, the distribution of the lasso estimator is

known only asymptotically in the case where the sample size n is much smaller than the number of predictors p (Knight and Fu [2000]), which is not useful in the context of genetic association studies. Inference based on the lasso estimator is still an open question, but we can cite here the work of Lee et al. [2016] who derived optimal and exact confidence intervals for post-selection inference using a truncated Gaussian distribution. However, their proposed post-selection intervals for regression coefficients have $1 - \alpha$ coverage conditional on the selected model, where α is the size of the test. In the case where the oracle property of the model selection procedure is not guaranteed, such as for the lasso when the number of predictors is much greater than the sample size, these conditional confidence intervals might not be suitable as they ignore the uncertainty associated with the model selection procedure. An alternative approach is the use of data-splitting methods, where an independent sample from the one that was used to fit the regularized model is used for performing post-selection inference (Cox [1975]). This is a popular approach in the Mendelian randomization literature (Grant and Burgess [2020], Zhao et al. [2019]) and it has been shown that the correct type I error rate will be retained given that multiple independent samples measuring the traits of interest are available.

Another limitation of our proposed methodology is that for continuous outcomes, we assumed that important genetic predictors only affect the conditional mean response of a phenotype. However, the influence of a genetic variant can extend beyond the mean, affecting both the lower and upper tails of the phenotype distribution. A robust alternative would be the use of quantile regression (QR) which is a generalization of the least absolute deviations regression to other percentiles than the median. More formally, QR allows to model the conditional quantile of a response as a function of the covariates. Compared to least squares regression, QR offers some uniques advantages such as (1) identifying variants with heterogeneous effects across quantiles of the phenotype distribution; (2) accommodating a wide range of outcome distributions including non-normal distributions; and (3) providing robustness against outliers and flexibility to model nonlinear relationships. Recently, QR has been used

to discover variants with larger effects on high-risk subgroups of individuals but with lower or no contribution overall for 39 quantitative traits in the UK Biobank (Wang et al. [2024]). Several authors have recently addressed the limitations related to the computational burden of fitting QR due to the non smoothness of the loss function, and extended QR models to high-dimensional sparse regularized problems (Pietrosanu et al. [2021], Gu et al. [2018], Yu and Lin [2017]). They proposed combining fast alternating direction method of multipliers (ADMM) algorithms with coordinate descent steps to solve large-scale penalized QR problems. In addition, I find relevant the work of Mendez-Civieta et al. [2020] who proposed an adaptive sparse group lasso methodology to the QR framework, and the work of Koenker [2004] who derived a penalized QR framework for longitudinal data. Future work includes extending our proposed methodology to the QR framework, relying on the work from the aforementioned authors.

6.3 Concluding remarks

Genetic association studies, with sample sizes and number of predictors ever increasing, offer interesting challenges for variable selection in high dimensional problems. This thesis focused on the development of computationally efficient multivariable regularized methods to account for different sources of confounding in genetic association studies. I have proposed models based on regularized PQL estimation to fit GLMMs to high dimensional GWAS data and developed efficient algorithms implemented in the `julia` programming language to make the proposed models applicable in real-life case studies where the number of candidate genetic predictors is very high. The proposed methods may be used to perform variable selection or to construct prediction scores when multiple levels of correlation between observations used to fit the model is observed. The methods may also be used in other areas of research related to genetic association studies, such as Mendelian randomization studies, given that post-selection confidence intervals with the desired level of coverage can be constructed.

Appendices

APPENDIX A

Appendix to Manuscript 1

A.1 Estimation of Variance Component Parameters

In what follows, we provide a near verbatim of Appendix A from Chen et al. [2016] that details the AI-REML algorithm to estimate the variance components and fixed effects for the non-genetic covariates under the assumption of no genetic association. To be consistent with the remainder of the manuscript, we slightly changed the notation from the original derivation where appropriate.

If ϕ and $\boldsymbol{\tau}$ are known, we jointly choose $\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau})$, $\hat{\boldsymbol{\gamma}}(\phi, \boldsymbol{\tau})$ and $\hat{\boldsymbol{b}}(\phi, \boldsymbol{\tau})$ to minimize (5.2), then $\hat{\boldsymbol{b}}(\phi, \boldsymbol{\tau}) = \tilde{\boldsymbol{b}}(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \hat{\boldsymbol{\gamma}}(\phi, \boldsymbol{\tau}))$ because $\tilde{\boldsymbol{b}}$ maximizes $f(\boldsymbol{b})$ for given $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$. Assuming that the weights in $\boldsymbol{W}$ vary slowly with the conditional mean, the derivatives of (5.2) at $\boldsymbol{\gamma} = 0$ with respect to $(\boldsymbol{\alpha}, \boldsymbol{b})$ are given by

$$\frac{\partial ql(\boldsymbol{\alpha}, \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \boldsymbol{\alpha}} = -\sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{\phi \nu(\mu_i)} \frac{1}{g'(\mu_i)} \boldsymbol{X}_i^{\intercal} = -\boldsymbol{X}^{\intercal} \boldsymbol{W} \Delta (\boldsymbol{y} - \boldsymbol{\mu}),$$

$$\frac{\partial ql(\boldsymbol{\alpha}, \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \boldsymbol{b}} = -\sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{\phi \nu(\mu_i)} \frac{1}{g'(\mu_i)} \boldsymbol{Z}_i^{\intercal} + \left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s \right)^{-1} \boldsymbol{b} = \left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s \right)^{-1} \boldsymbol{b} - \boldsymbol{W} \Delta (\boldsymbol{y} - \boldsymbol{\mu}),$$

where $\Delta = \text{diag}(g'(\mu_i))$ and $\boldsymbol{Z}_i$ is a $n \times 1$ vector of indicators such that $b_i = \boldsymbol{Z}_i \boldsymbol{b}$. Defining

the working vector $\tilde{\boldsymbol{Y}}$ with elements $\tilde{Y}_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$, the solution of

$$\begin{cases} \boldsymbol{X}^\intercal \boldsymbol{W} \Delta(\boldsymbol{y} - \boldsymbol{\mu}) = 0 \\ \boldsymbol{W}\Delta(\boldsymbol{y} - \boldsymbol{\mu}) = \left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s\right)^{-1} \boldsymbol{b} \end{cases}$$

can be written as the solution to the system

$$\begin{bmatrix} \boldsymbol{X}^\intercal \boldsymbol{W} \boldsymbol{X} & \boldsymbol{X}^\intercal \boldsymbol{W} \\ \boldsymbol{W}\boldsymbol{X} & \left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s\right)^{-1} + \boldsymbol{W} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\intercal \boldsymbol{W} \tilde{\boldsymbol{Y}} \\ \boldsymbol{W}\tilde{\boldsymbol{Y}} \end{bmatrix}.$$

Let $\boldsymbol{\Sigma} = \boldsymbol{W}^{-1} + \sum_{s=1}^{S} \tau_s \boldsymbol{V}_s$, $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^\intercal \boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\intercal\boldsymbol{\Sigma}^{-1}$, then

$$\begin{cases} \hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^\intercal \boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\intercal\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{Y}} \\ \hat{\boldsymbol{b}} = \left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s\right)\boldsymbol{\Sigma}^{-1}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right) \end{cases}.$$

Of note, we have that

$$\begin{aligned} \tilde{\boldsymbol{Y}} - \hat{\boldsymbol{\eta}} &= \tilde{\boldsymbol{Y}} - \boldsymbol{X}\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{b}} \\ &= \left\{\boldsymbol{I} - \left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s\right)\boldsymbol{\Sigma}^{-1}\right\}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right) \\ &= \boldsymbol{W}^{-1}\boldsymbol{\Sigma}^{-1}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right) \\ &= \boldsymbol{W}^{-1}\boldsymbol{P}\tilde{\boldsymbol{Y}}. \end{aligned}$$

The log integrated quasi-likelihood function in (5.2) evaluated at $(\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})$ becomes

$$\begin{aligned} ql(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau}) &= -\frac{1}{2}\log\left|\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s \boldsymbol{W} + \boldsymbol{I}\right| - \frac{1}{2}\sum_{i=1}^{n} \frac{a_i(y_i - \hat{\mu}_i)^2}{\phi\nu(\hat{\mu}_i)} - \frac{1}{2}\hat{\boldsymbol{b}}^{\intercal}\left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s\right)\hat{\boldsymbol{b}} \\ &= -\frac{1}{2}\log|\boldsymbol{\Sigma}\boldsymbol{W}| - \frac{1}{2}(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{\eta}})^{\intercal}\boldsymbol{W}(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{\eta}}) \\ &\quad - \frac{1}{2}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right)^{\intercal}\boldsymbol{\Sigma}^{-1}\left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s\right)\boldsymbol{\Sigma}^{-1}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right) \\ &= -\frac{1}{2}\log|\boldsymbol{W}| - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\tilde{\boldsymbol{Y}}^{\intercal}\boldsymbol{P}\boldsymbol{W}^{-1}\boldsymbol{P}\tilde{\boldsymbol{Y}} \\ &\quad - \frac{1}{2}\tilde{\boldsymbol{Y}}^{\intercal}\boldsymbol{P}\left(\sum_{s=1}^{S} \tau_s \boldsymbol{V}_s\right)\boldsymbol{P}\tilde{\boldsymbol{Y}} \\ &= c - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\tilde{\boldsymbol{Y}}^{\intercal}\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}\tilde{\boldsymbol{Y}} \\ &= c - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\tilde{\boldsymbol{Y}}^{\intercal}\boldsymbol{P}\tilde{\boldsymbol{Y}}. \end{aligned}$$

Similarly, the restricted maximum likelihood (REML) version is

$$ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau}) = c_R - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log\left|\boldsymbol{X}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right| - \frac{1}{2}\tilde{\boldsymbol{Y}}^{\intercal}\boldsymbol{P}\tilde{\boldsymbol{Y}}.$$

We need to maximize $ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})$ with respect to $\phi, \boldsymbol{\tau}$. Let $\boldsymbol{V}_0 = \text{diag}\{a_i^{-1}\nu(\mu_i)[g'(\mu_i)^2]\} = \phi^{-1}\boldsymbol{W}^{-1}$, then $\boldsymbol{\Sigma} = \phi\boldsymbol{V}_0 + \sum_{s=1}^{S} \tau_s \boldsymbol{V}_s$, and the first derivatives of $ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})$ with respect to ϕ and τ_s are

$$\frac{\partial ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \phi} = \frac{1}{2}\left\{\tilde{\boldsymbol{Y}}^{\intercal}\boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\tilde{\boldsymbol{Y}} - tr(\boldsymbol{P}\boldsymbol{V}_0)\right\} \tag{A.1}$$

$$\frac{\partial ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \tau_s} = \frac{1}{2}\left\{\tilde{\boldsymbol{Y}}^{\intercal}\boldsymbol{P}\boldsymbol{V}_s\boldsymbol{P}\tilde{\boldsymbol{Y}} - tr(\boldsymbol{P}\boldsymbol{V}_s)\right\}, \tag{A.2}$$

since one can show that

$$\frac{\partial \boldsymbol{P}}{\partial \phi} = -\boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}, \qquad \frac{\partial \boldsymbol{P}}{\partial \tau_s} = -\boldsymbol{P}\boldsymbol{V}_s\boldsymbol{P}.$$

$\hat{\phi}$ and $\hat{\boldsymbol{\tau}}$ are estimated by finding the solutions of (A.1) and (A.2) equal to zero. Let $\boldsymbol{\theta} =$

$(\phi, \boldsymbol{\tau})$, and recall that in the REML iterative process, $\hat{\boldsymbol{\theta}}$ at the (i+1)th iteraton is updated by $\hat{\boldsymbol{\theta}}^{(i+1)} = \hat{\boldsymbol{\theta}}^{(i)} + J(\hat{\boldsymbol{\theta}}^{(i)})^{-1} S(\hat{\boldsymbol{\theta}}^{(i)})$, where $S(\boldsymbol{\theta}) = \frac{\partial ql_R(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and $J(\boldsymbol{\theta}) = -\frac{\partial^2 ql_R(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}$. The elements of the observed information matrix $J(\boldsymbol{\theta})$ are

$$
\begin{aligned}
-\frac{\partial^2 ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \phi^2} &= \tilde{\boldsymbol{Y}}^\intercal \boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\tilde{\boldsymbol{Y}} - \frac{1}{2} tr(\boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\boldsymbol{V}_0) \\
-\frac{\partial^2 ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \phi \partial \tau_s} &= \tilde{\boldsymbol{Y}}^\intercal \boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\boldsymbol{V}_s\boldsymbol{P}\tilde{\boldsymbol{Y}} - \frac{1}{2} tr(\boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\boldsymbol{V}_s) \\
-\frac{\partial^2 ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \tau_l \partial \tau_s} &= \tilde{\boldsymbol{Y}}^\intercal \boldsymbol{P}\boldsymbol{V}_l\boldsymbol{P}\boldsymbol{V}_s\boldsymbol{P}\tilde{\boldsymbol{Y}} - \frac{1}{2} tr(\boldsymbol{P}\boldsymbol{V}_l\boldsymbol{P}\boldsymbol{V}_s).
\end{aligned}
$$

The elements of the expected information matrix are

$$
\begin{aligned}
E\left(-\frac{\partial^2 ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \phi^2}\right) &= \frac{1}{2} tr(\boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\boldsymbol{V}_0) \\
E\left(-\frac{\partial^2 ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \phi \partial \tau_s}\right) &= \frac{1}{2} tr(\boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\boldsymbol{V}_s) \\
E\left(-\frac{\partial^2 ql_R(\hat{\boldsymbol{\alpha}}(\phi, \boldsymbol{\tau}), \boldsymbol{\gamma} = 0, \phi, \boldsymbol{\tau})}{\partial \tau_l \partial \tau_s}\right) &= \frac{1}{2} tr(\boldsymbol{P}\boldsymbol{V}_l\boldsymbol{P}\boldsymbol{V}_s).
\end{aligned}
$$

The average information matrix $\boldsymbol{AI}$ is defined as the average of the observed information $J(\boldsymbol{\theta})$ and the expected information

$$
\begin{aligned}
\boldsymbol{AI}_{\phi\phi} &= \frac{1}{2} \tilde{\boldsymbol{Y}}^\intercal \boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\tilde{\boldsymbol{Y}}, \\
\boldsymbol{AI}_{\phi\tau_s} &= \frac{1}{2} \tilde{\boldsymbol{Y}}^\intercal \boldsymbol{P}\boldsymbol{V}_0\boldsymbol{P}\boldsymbol{V}_s\boldsymbol{P}\tilde{\boldsymbol{Y}}, \\
\boldsymbol{AI}_{\tau_s\tau_l} &= \frac{1}{2} \tilde{\boldsymbol{Y}}^\intercal \boldsymbol{P}\boldsymbol{V}_s\boldsymbol{P}\boldsymbol{V}_l\boldsymbol{P}\tilde{\boldsymbol{Y}}.
\end{aligned}
$$

Let $\boldsymbol{\theta}$ be the variance component and dispersion parameters to estimate, that is when $\phi \neq 1$, $\boldsymbol{\theta} = (\phi, \boldsymbol{\tau})$, and $\boldsymbol{AI}$ is a $(S+1) \times (S+1)$ matrix. For binary data, $\phi = 1$, $\boldsymbol{\theta} = \boldsymbol{\tau}$, and $\boldsymbol{AI}$ is a $S \times S$ matrix containing only $\boldsymbol{AI}_{\tau_s\tau_l}$. We use the following algorithm to estimate $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{b}$:

Algorithm 1 AI-REML algorithm

1. *Initialization*
   Fit a generalized linear model with $\boldsymbol{\tau} = 0$ and get $\hat{\boldsymbol{\alpha}}^{(0)}$ and working vector $\tilde{\boldsymbol{Y}}^{(0)}$;
   Use $\boldsymbol{\theta}^{(0)} = Var(\tilde{\boldsymbol{Y}}^{(0)})/S$ (if $\phi = 1$) or $\boldsymbol{\theta}^{(0)} = Var(\tilde{\boldsymbol{Y}}^{(0)})/(S+1)$ (if $\phi \neq 1$) as the initial value of $\boldsymbol{\theta}$;
   For each $s = 0, 1, ..., S,$ update $\boldsymbol{\theta}$ using $\theta_s^{(1)} = \theta_s^{(0)} + 2n^{-1}\{\theta_s^{(0)}\}^2(\partial ql_R(\boldsymbol{\theta}^{(0)})/\partial\theta_s)$;

2. *Iteration*
   for $t = 1, 2, ...,$ *until convergence* **do**
   Update $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \{\boldsymbol{AI}^{(t)}\}^{-1}(\partial ql_R(\boldsymbol{\theta}^{(t)})/\partial\boldsymbol{\theta})$;
   Calculate $\hat{\boldsymbol{\alpha}}^{(t+1)}$ and $\hat{\boldsymbol{b}}^{(t+1)}$ using $\tilde{\boldsymbol{Y}}^{(t)}$ and $\boldsymbol{\theta}^{(t+1)}$;
   Update $\tilde{\boldsymbol{Y}}^{(t+1)}$ using $\hat{\boldsymbol{\alpha}}^{(t+1)}$ and $\hat{\boldsymbol{b}}^{(t+1)}$;

Convergence is defined using $2\max\{|\hat{\boldsymbol{\alpha}}^{(t)} - \hat{\boldsymbol{\alpha}}^{(t-1)}|/(|\hat{\boldsymbol{\alpha}}^{(t)}| + |\hat{\boldsymbol{\alpha}}^{(t-1)}|), |\hat{\boldsymbol{\theta}}^{(t)} - \hat{\boldsymbol{\theta}}^{(t-1)}|/(|\hat{\boldsymbol{\theta}}^{(t)}| + |\hat{\boldsymbol{\theta}}^{(t-1)}|)\} \leq$ tolerance

A.2 Cyclic coordinate Descent for PQL Regularized Parameters

Assuming that the variance components and dispersion parameters are known, we fit the full GLMM (5.1) with lasso regularization on $\boldsymbol{\beta} = (\boldsymbol{\alpha}^\intercal, \boldsymbol{\gamma}^\intercal)^\intercal$ to obtain PQL regularized estimates for $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{b}}$. At each iteration, we cycle through the coordinates and minimize the objective function (5.3) with respect to one coordinate only. Suppose we have estimates $\tilde{\boldsymbol{\beta}}$ and we wish to partially optimize (5.3) with respect to $\tilde{\boldsymbol{b}}$. The gradient and Hessian of ℓ_{PQL} with respect to $\tilde{\boldsymbol{b}}$ at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ are given by

$$\begin{aligned}\nabla_{\tilde{\boldsymbol{b}}}\ell_{PQL}(\tilde{\boldsymbol{\beta}}, \hat{\phi}, \hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}}) &= \sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{\hat{\phi}\nu(\mu_i)} \frac{1}{g'(\mu_i)} \boldsymbol{Z}_i^\intercal - \left(\sum_{s=1}^{S} \hat{\tau}_s \boldsymbol{V}_s\right)^{-1} \tilde{\boldsymbol{b}} \\ &= \boldsymbol{W}\Delta(\boldsymbol{y} - \boldsymbol{\mu}) - \left(\sum_{s=1}^{S} \hat{\tau}_s \boldsymbol{V}_s\right)^{-1} \tilde{\boldsymbol{b}},\end{aligned}$$

and

$$\nabla^2_{\tilde{\boldsymbol{b}}}\ell_{PQL}(\tilde{\boldsymbol{\beta}},\hat{\phi},\hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}}) = -\boldsymbol{W} - \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)^{-1},$$

where $\Delta = \text{diag}(g'(\mu_i))$ and $\boldsymbol{Z}_i$ is a $n \times 1$ vector of indicators such that $b_i = \boldsymbol{Z}_i\boldsymbol{b}$. We form a quadratic approximation of $\ell_{PQL}(\tilde{\boldsymbol{\beta}},\hat{\phi},\hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}})$ around current iterate $\tilde{\boldsymbol{b}}$, which yields

$$f(\boldsymbol{b}) := \ell_{PQL}(\tilde{\boldsymbol{\beta}},\hat{\phi},\hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}}) + (\boldsymbol{b}-\tilde{\boldsymbol{b}})^\intercal\nabla\ell_{PQL}(\tilde{\boldsymbol{\beta}},\hat{\phi},\hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}}) + \frac{1}{2}(\boldsymbol{b}-\tilde{\boldsymbol{b}})^\intercal\nabla^2\ell_{PQL}(\tilde{\boldsymbol{\beta}},\hat{\phi},\hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}})(\boldsymbol{b}-\tilde{\boldsymbol{b}}).$$

This leads to the Newton's updates

$$\begin{aligned}
\hat{\boldsymbol{b}} &= \tilde{\boldsymbol{b}} + \left[-\nabla^2_{\tilde{\boldsymbol{b}}}\ell_{PQL}(\tilde{\boldsymbol{\beta}},\hat{\phi},\hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}})\right]^{-1}\nabla_{\tilde{\boldsymbol{b}}}\ell_{PQL}(\tilde{\boldsymbol{\beta}},\hat{\phi},\hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}}) \\
&= \tilde{\boldsymbol{b}} + \left(\boldsymbol{W} + \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)^{-1}\right)^{-1}\left(\boldsymbol{W}\Delta(\boldsymbol{y}-\boldsymbol{\mu}) - \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)^{-1}\tilde{\boldsymbol{b}}\right) \\
&= \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)\left(\boldsymbol{W}^{-1} + \sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)^{-1}\left(\Delta(\boldsymbol{y}-\boldsymbol{\mu}) + \tilde{\boldsymbol{b}}\right). \qquad \text{(A.3)}
\end{aligned}$$

Defining the working vector $\tilde{\boldsymbol{Y}}$ with elements $\tilde{Y}_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$, the solution of (A.3) is equal to

$$\hat{\boldsymbol{b}} = \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)\boldsymbol{\Sigma}^{-1}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\right), \qquad \text{(A.4)}$$

where $\tilde{\boldsymbol{X}} = [\boldsymbol{X} \quad \boldsymbol{G}]$ and $\boldsymbol{\Sigma} = \boldsymbol{W}^{-1} + \sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s$. Because the weights $\boldsymbol{W}$ are being updated repeatedly, the solution (A.4) requires inverting a different variance-covariance matrix $\boldsymbol{\Sigma}$ at each iteration, with complexity $O(n^3)$. In modern large-scale data sets, the sample size n can be very large, thus we want to avoid costly matrix inversions. For binary traits, we have that

$$\nabla^2_{\tilde{\boldsymbol{b}}}\ell_{PQL}(\tilde{\boldsymbol{\beta}},\hat{\phi},\hat{\boldsymbol{\tau}}|\tilde{\boldsymbol{b}}) \succeq -0.25\boldsymbol{I}_n - \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)^{-1}.$$

Therefore, we can replace the hessian by its lower-bound in the quadratic approximation $f(\boldsymbol{b})$ (Böhning and Lindsay [1988]). This leads to a minorization-maximization (MM) algorithm (Hunter and Lange [2004]) with updates

$$\begin{aligned}\hat{\boldsymbol{b}} &= \tilde{\boldsymbol{b}} + \left(0.25\boldsymbol{I}_n + \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)^{-1}\right)^{-1}\left(\boldsymbol{W}\Delta(\boldsymbol{y}-\boldsymbol{\mu}) - \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)^{-1}\tilde{\boldsymbol{b}}\right) \\ &= \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)\left(4\boldsymbol{I}_n + \sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)^{-1}\left(4(\boldsymbol{y}-\boldsymbol{\mu}) + \tilde{\boldsymbol{b}}\right). \end{aligned} \tag{A.5}$$

Redefining the working vector $\tilde{\boldsymbol{Y}}$ with elements $\tilde{Y}_i = \eta_i + 4(y_i - \mu_i)$, the solution of (A.5) is equal to

$$\hat{\boldsymbol{b}} = \left(\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s\right)\tilde{\boldsymbol{\Sigma}}^{-1}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\right), \tag{A.6}$$

where $\tilde{\boldsymbol{\Sigma}} = 4\boldsymbol{I}_n + \sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s$.

Let $\sum_{s=1}^{S}\hat{\tau}_s\boldsymbol{V}_s = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\intercal$ be the associated eigen-spectral decomposition of the variance-covariance matrix of $\boldsymbol{b}$, where $\boldsymbol{U}_{n\times n}$ is an orthonormal matrix of eigenvectors and $\boldsymbol{D}_{n\times n}$ is a diagonal matrix of eigenvalues, such that (A.6) can be rewritten as

$$\hat{\boldsymbol{b}} = \boldsymbol{U}\left(4\boldsymbol{D}^{-1} + \boldsymbol{I}_n\right)^{-1}\boldsymbol{U}^\intercal\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\right). \tag{A.7}$$

By rotating the random effect $\boldsymbol{\delta} = \boldsymbol{U}^\intercal\boldsymbol{b}$, we have that (A.7) is equivalent to solving the following generalized ridge regression problem

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}}\ \frac{1}{4}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{U}\boldsymbol{\delta}\right)^\intercal\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \boldsymbol{U}\boldsymbol{\delta}\right) + \boldsymbol{\delta}^\intercal\boldsymbol{D}^{-1}\boldsymbol{\delta}.$$

Consider now a coordinate descent step for $\boldsymbol{\beta}$. That is, suppose we have updates $\tilde{\boldsymbol{b}}$ and $\tilde{\boldsymbol{\beta}}_l$ for $l \neq j$, and we wish to partially optimize with respect to β_j. We would like to compute

the gradient at $\beta_j = \tilde{\beta}_j$, which only exists if $\tilde{\beta}_j \neq 0$. If $\tilde{\beta}_j > 0$, then

$$\frac{\partial Q_\lambda(\boldsymbol{\beta}, \boldsymbol{b})}{\partial \beta_j}|_{(\boldsymbol{b},\boldsymbol{\beta})=(\tilde{\boldsymbol{b}},\tilde{\boldsymbol{\beta}})} = -\sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{\hat{\phi}\nu(\mu_i)} \frac{1}{g'(\mu_i)} \tilde{X}_{ij} + \lambda v_j = -\tilde{\boldsymbol{X}}_j^\intercal \boldsymbol{W} \Delta(\boldsymbol{y} - \boldsymbol{\mu}) + \lambda v_j, \quad \text{(A.8)}$$

where $\tilde{\boldsymbol{X}}_j$ is a $n \times 1$ column vector for predictor j. Recall that for binary traits, we defined the working vector $\tilde{\boldsymbol{Y}}$ such that $\boldsymbol{y} - \boldsymbol{\mu} = \frac{1}{4}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta} - \boldsymbol{b})$. Moreover, for binary traits with logistic link function, we have $\phi = 1$ and $\boldsymbol{W} = \boldsymbol{\Delta}^{-1}$. Thus, plugging $\tilde{\boldsymbol{b}} = \hat{\boldsymbol{b}}$ from (A.7) and solving (A.8) leads to

$$\begin{aligned}
&-\frac{1}{4}\tilde{\boldsymbol{X}}_j^\intercal \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{b}}\right) + \lambda v_j = 0 \\
\iff &-\frac{1}{4}\tilde{\boldsymbol{X}}_j^\intercal \boldsymbol{U} \left(\boldsymbol{I}_n - \left(4\boldsymbol{D}^{-1} + \boldsymbol{I}_n\right)^{-1}\right) \boldsymbol{U}^\intercal \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\right) + \lambda v_j = 0 \\
\iff &-\tilde{\boldsymbol{X}}_j^\intercal \boldsymbol{U} \left(4\boldsymbol{I}_n + \boldsymbol{D}\right)^{-1} \boldsymbol{U}^\intercal \left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}\right) + \lambda v_j = 0.
\end{aligned}$$

Finally, isolating β_j yields

$$\begin{aligned}
\hat{\beta}_j &= \frac{\tilde{\boldsymbol{X}}_j^\intercal \boldsymbol{U} \left(4\boldsymbol{I}_n + \boldsymbol{D}\right)^{-1} \boldsymbol{U}^\intercal \left(\tilde{\boldsymbol{Y}} - \sum_{l \neq j} \tilde{\boldsymbol{X}}_l \tilde{\beta}_l\right) - \lambda v_j}{\tilde{\boldsymbol{X}}_j^\intercal \boldsymbol{U} \left(4\boldsymbol{I}_n + \boldsymbol{D}\right)^{-1} \boldsymbol{U}^\intercal \tilde{\boldsymbol{X}}_j} \\
&= \frac{\sum_{i=1}^{n} \frac{1}{4+\Lambda_i} \tilde{X}^*_{ij} \left(\tilde{Y}^*_i - \sum_{l \neq j} \tilde{X}^*_{il} \tilde{\boldsymbol{\beta}}_l\right) - \lambda v_j}{\sum_{i=1}^{n} \frac{1}{4+\Lambda_i} \tilde{X}^{*2}_{ij}},
\end{aligned} \quad \text{(A.9)}$$

where Λ_i are the eigenvalues of $\sum_{s=1}^{S} \hat{\tau}_s \boldsymbol{V}_s$, $\tilde{\boldsymbol{Y}}^* = \boldsymbol{U}^\intercal \tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{X}}^* = \boldsymbol{U}^\intercal \tilde{\boldsymbol{X}}$. By proceeding in a similar way for $\tilde{\beta}_j < 0$, one can show (Friedman et al. [2007]) that the coordinate-wise update for β_j has the form

$$\hat{\beta}_j = \frac{S\left(\sum_{i=1}^{n} \frac{1}{4+\Lambda_i} \tilde{X}^*_{ij} \left(\tilde{Y}^*_i - \sum_{l \neq j} \tilde{X}^*_{il} \tilde{\boldsymbol{\beta}}_l\right), \lambda v_j\right)}{\sum_{i=1}^{n} \frac{1}{4+\Lambda_i} \tilde{X}^{*2}_{ij}}, \quad \text{(A.10)}$$

where $S(z, \gamma)$ is the soft-thresholding operator:

$$\text{sign}(z)(|z| - \gamma)_+ = \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z| \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z| \\ 0 & \text{if } \gamma \geq |z|. \end{cases}$$

Finally, the updates for $\boldsymbol{\eta}$ are given by

$$\begin{aligned} \hat{\boldsymbol{\eta}} &= \tilde{\boldsymbol{Y}} - \left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{b}}\right) \\ &= \tilde{\boldsymbol{Y}} - \boldsymbol{U}\left\{\boldsymbol{I}_n - (4\boldsymbol{D}^{-1} + \boldsymbol{I}_n)^{-1}\right\}\left(\tilde{\boldsymbol{Y}}^* - \tilde{\boldsymbol{X}}^*\hat{\boldsymbol{\beta}}\right) \\ &= \tilde{\boldsymbol{Y}} - \boldsymbol{U}\left(\frac{1}{4}\boldsymbol{D} + \boldsymbol{I}_n\right)^{-1}\left(\tilde{\boldsymbol{Y}}^* - \tilde{\boldsymbol{X}}^*\hat{\boldsymbol{\beta}}\right). \end{aligned} \tag{A.11}$$

We performed additional simulations in Appendix A.4 and show that coefficients estimates for $\boldsymbol{\beta}$ obtained by replacing the hessian by a lower bound are similar to those obtained by repeatedly inverting the full hessian matrix at each iteration.

The cyclic coordinate descent algorithm to obtain regularized PQL estimates for $\boldsymbol{\beta} = (\boldsymbol{\alpha}^\intercal, \boldsymbol{\gamma}^\intercal)^\intercal$ and $\boldsymbol{b}$ is as follows:

Algorithm 2 Cyclic coordinate descent for regularized PQL estimation

1. *Initialization*
   Set $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\boldsymbol{\alpha}}^\intercal, \mathbf{0}^\intercal)$ and $\hat{\boldsymbol{b}}^{(0)} = \hat{\boldsymbol{b}}$, where $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}}$ are the estimates from the AI-REML algorithm;
   Calculate $\hat{\boldsymbol{\eta}}^{(0)} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{(0)} + \hat{\boldsymbol{b}}^0$, $\tilde{\boldsymbol{Y}}^{*(0)} = \boldsymbol{U}^\intercal\tilde{\boldsymbol{Y}}^{(0)}$ and $\tilde{\boldsymbol{X}}^* = \boldsymbol{U}^\intercal\tilde{\boldsymbol{X}}$;
2. *Iteration*
   for $\lambda = \lambda_{max}$ *to* λ_{min} **do**
      for $t = 1, 2, ...,$ *until outer-loop convergence* **do**
         for $j = 1, ..., m + p$ **do**
            Calculate

$$\hat{\boldsymbol{\beta}}_j^{(t)} = \frac{S\left(\sum_{i=1}^n \frac{1}{4+\Lambda_i}\tilde{X}_{ij}^*\left(\tilde{Y}_i^{*(t-1)} - \sum_{l\neq j}\tilde{X}_{il}^*\hat{\boldsymbol{\beta}}_l^{(t-1)}\right), \lambda v_j\right)}{\sum_{i=1}^n \frac{1}{4+\Lambda_i}\tilde{X}_{ij}^{*^2}},$$

            until inner-loop convergence;
         Calculate $\hat{\boldsymbol{\eta}}^{(t)} = \tilde{\boldsymbol{Y}}^{(t-1)} - \boldsymbol{U}\left(\frac{1}{4}\boldsymbol{D} + \boldsymbol{I}_n\right)^{-1}\left(\tilde{\boldsymbol{Y}}^{*(t-1)} - \tilde{\boldsymbol{X}}^*\hat{\boldsymbol{\beta}}^{(t)}\right)$;
      Update $\tilde{\boldsymbol{Y}}^{(t)}$ and $\tilde{\boldsymbol{Y}}^{*(t)}$ using $\hat{\boldsymbol{\eta}}^{(t)}$;
   Set $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}^{(t)}$, $\tilde{\boldsymbol{Y}}^{(0)} = \tilde{\boldsymbol{Y}}^{(t)}$ and $\tilde{\boldsymbol{Y}}^{*(0)} = \tilde{\boldsymbol{Y}}^{*(t)}$ as warm starts for next λ;

For inner-loop convergence, we use the same criteria as Friedman et al. [2007], that is after a complete cycle of coordinate descent we look at

$$\max_j \Delta_j = \max_j \sum_{i=1}^n \frac{1}{4+\Lambda_i}\tilde{X}_{ij}^{*2}(\hat{\beta}_j^{(t-1)} - \hat{\beta}_j^{(t)})^2,$$

which measures the maximum weighted sum of squares of changes in fitted values for all coefficients. If $\max_j \Delta_j$ is smaller than tolerance, we stop the coordinate descent loop. For outer-loop convergence, we calculate the fractional change in the loss function $-l_{PQL}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \hat{\phi}, \hat{\boldsymbol{\tau}})$ and declare convergence if its value is smaller than tolerance.

A.3 Model selection

Approaches to selecting the optimal tuning parameter in regularized models are of primary interest since in real data analysis, the underlying true model is unknown. A popular strategy is to select the value that minimizes out-of-sample prediction error, e.g., cross-validation (CV), which is asymptotically equivalent to the Akaike information criterion (AIC) (Akaike [1998], Yang [2005]). While being conceptually attractive, CV becomes computationally expensive for very high-dimensional data. Moreover, in studies where the proportion of related subjects is important, either by known or cryptic relatedness, the CV prediction error is no longer an unbiased estimator of the generalization error (Rabinowicz and Rosset [2020]). Through simulation studies and real data analysis, Wang et al. [2020] found that LD and minor allele frequencies (MAF) differences between ancestries could explain between 70 and 80% of the loss of relative accuracy of European-based prediction models in African ancestry for traits like body mass index and type 2 diabetes. Thus, there is no clear approach to how multiple admixed and/or similar populations should be split when using CV to minimize out-of-sample prediction error.

Alternatively, we can select the optimal value of the tuning parameter by optimizing the generalized information criterion (GIC) with an appropriate model complexity penalty a_n, defined as

$$\text{GIC}_\lambda = -2\ell_{PQL} + a_n \cdot \hat{df}_\lambda, \tag{A.12}$$

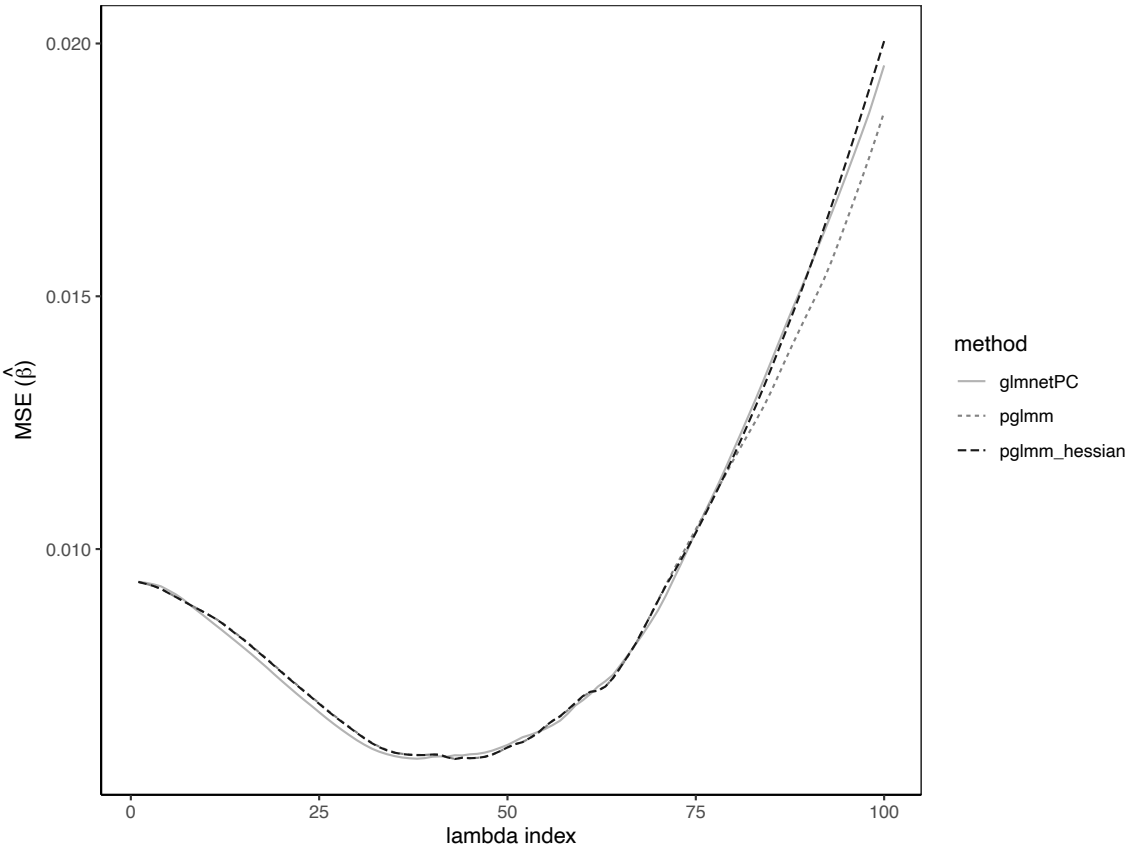
where ℓ_{PQL} is defined in (5.3), and $\hat{df}_\lambda = |\{1 \leq k \leq p : \hat{\beta}_k \neq 0\}| + \dim(\hat{\boldsymbol{\tau}})$ is the number of nonzero fixed-effects coefficients (Zou et al. [2007]) plus the number of variance components. The choice of a_n becomes crucial for effectively identifying the true model in high-dimensional data. Fan and Tang [2012] have proposed using a high-dimensional Bayesian information criterion (HDBIC) with $a_n = \log(\log n)\log p$. However, in our simulations and analysis of real data, our findings were that using $a_n = 2$ (AIC) was an appropriate model complexity

penalty, and that using $a_n = \log(n)$ (BIC, Schwarz [1978]) resulted in sparse models with almost no predictors. Hence, we did not investigate in this work the use of more severe model complexity penalties such as the HDBIC, but in our software implementation we allow to choose between AIC, BIC and HDBIC for selecting the best model.

A.4 Comparison of coefficient estimates using a lower-bound algorithm

We performed additional simulations to evaluate the impact on the coefficient estimates of taking a lower-bound of the variance-covariance matrix. More specifically, we simulated random genotypes from the BN-PSD admixture model for 10 intermediate populations of the 1D linear admixture model, with a total of $n = 2500$ samples, $p = 5000$ candidate SNPs where we randomly selected $c = 1\%$ to be causal. We fitted the full lasso path for 100 values of the regularization parameter λ. We report in Figure 1 the median mean squared-error (MSE) for 20 replications, defined as $\text{MSE}(\hat{\boldsymbol{\beta}}) = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p$, when we repeatedly inverted the full variance-covariance matrix (`pglmm_hessian`), and when we replaced the hessian by a lower-bound (`pglmm`). We also included in the comparison the coefficient estimates for a logistic lasso with 10 PCs (`glmnetPC`). We can see that all 3 methods lead to similar estimates on the full lasso path as measured by the median $\text{MSE}(\hat{\boldsymbol{\beta}})$. For models with large number of active predictors, the estimates based on the lower-bound method are closer to the true estimates compared to the other methods. We note that for binary responses, the existing lower-bound on the variance-covariance matrix not only provides an increasing computational advantage over Newton–Raphson, but also guarantees linear convergence to the optimal solution (Böhning and Lindsay [1988]). Thus, using a lower-bound on the hessian of the loss function for logistic regression is commonly done in majorization-minimization (MM) algorithms (Hunter and Lange [2004], Hu et al. [2019]).

Figure A.1: Median MSE($\hat{\boldsymbol{\beta}}$) for 20 replications of the simulated genotype with 1d linear admixture and $K = 10$ subpopulations. Compared methods are `pglmm` with a lower-bound algorithm, `pglmm_hessian` where we repeatedly invert the full variance-covariance matrix and logistic lasso with 10 PCs (`glmnetPC`).



A.5 Confounding from population structure

Figure A.2: Correlation heatmap between the first 20 PCs and $K = 20$ indicator functions identifying the independent subpopulations from the simulated genotype. We used the absolute value or the Pearson's correlation coefficient for the color scaling, and displayed the value whenever $|r^2| > 0.2$.

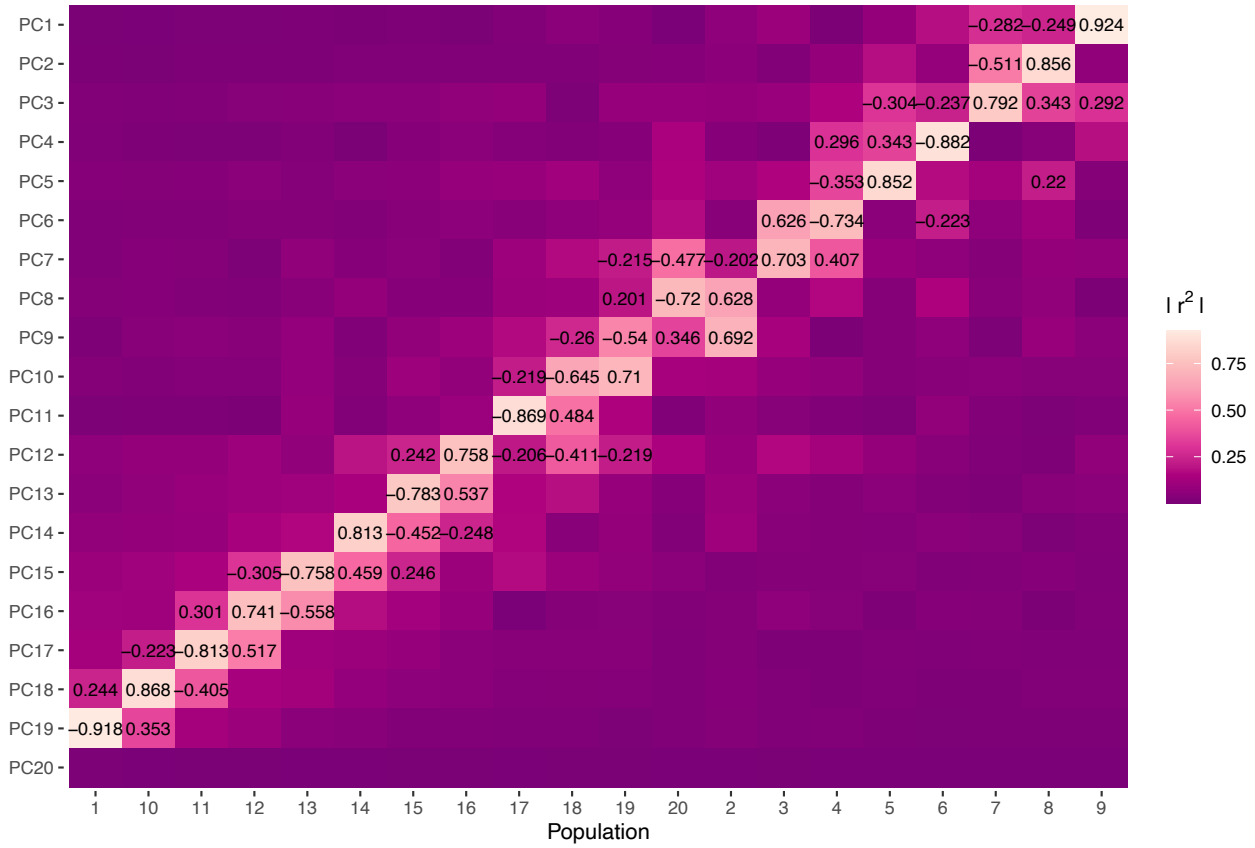
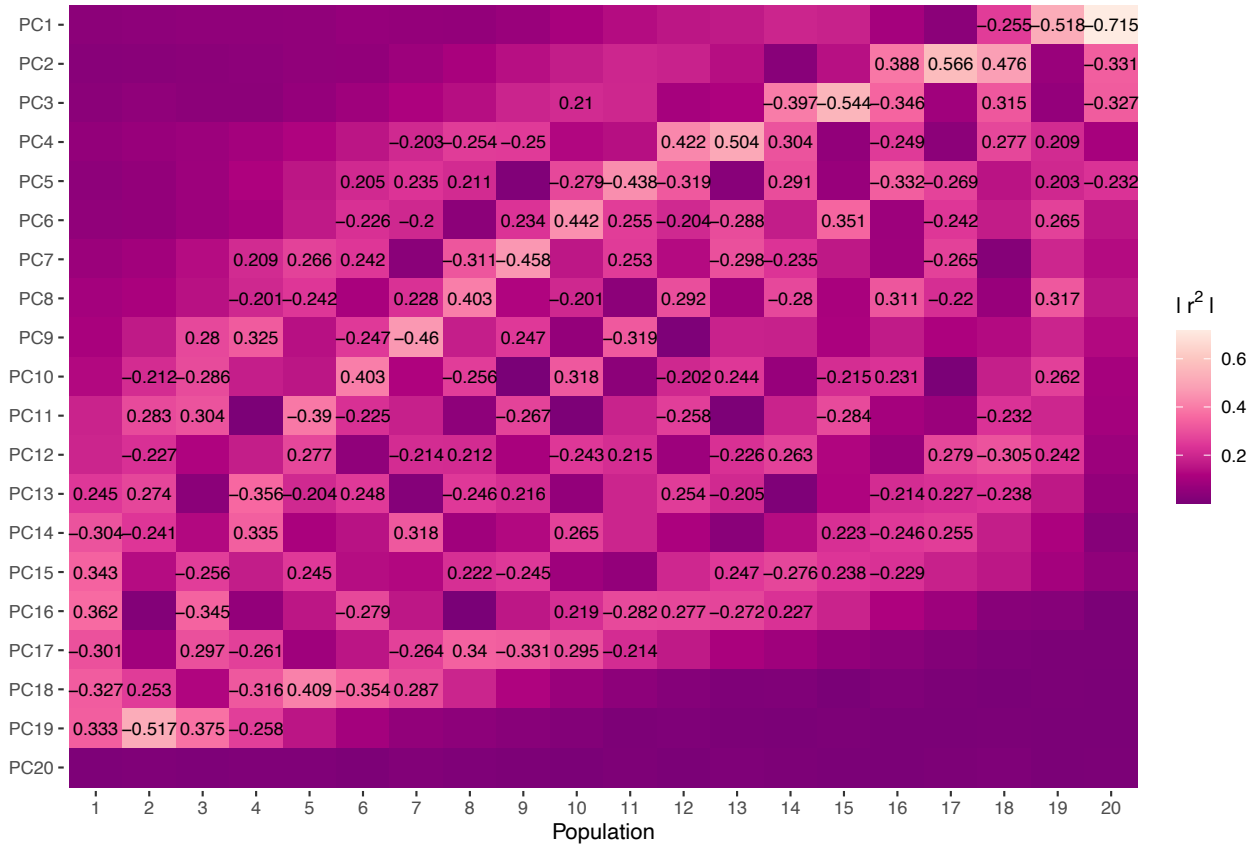


Figure A.3: Correlation heatmap between the first 20 PCs and $K = 20$ indicator functions identifying subpopulations from the simulated genotype with 1d linear admixture. We used the absolute value or the Pearson's correlation coefficient for the color scaling, and displayed the value whenever $|r^2| > 0.2$.



APPENDIX B

Appendix to Manuscript 2

B.1 Updates for $\tilde{\boldsymbol{\delta}}$

The gradient and Hessian of $f(\boldsymbol{\Theta};\boldsymbol{\delta})$ are given by

$$\nabla_{\boldsymbol{\delta}} f(\boldsymbol{\Theta};\boldsymbol{\delta}) = -\hat{\phi}^{-1}\boldsymbol{U}^{\intercal}(\boldsymbol{y}-\boldsymbol{\mu}) + \boldsymbol{\Lambda}^{-1}\boldsymbol{\delta},$$
$$\nabla^2_{\boldsymbol{\delta}} f(\boldsymbol{\Theta};\boldsymbol{\delta}) = \hat{\phi}^{-1}\boldsymbol{U}^{\intercal}\boldsymbol{\Delta}^{-1}\boldsymbol{U} + \boldsymbol{\Lambda}^{-1}.$$

This leads to the Newton updates

$$\begin{aligned}
\tilde{\boldsymbol{\delta}}^{(t+1)} &= \tilde{\boldsymbol{\delta}}^{(t)} - [\nabla^2_{\boldsymbol{\delta}} f(\boldsymbol{\Theta}|\tilde{\boldsymbol{\delta}}^{(t)})]^{-1}\nabla_{\boldsymbol{\delta}} f(\boldsymbol{\Theta}|\tilde{\boldsymbol{\delta}}^{(t)}) \\
&= \tilde{\boldsymbol{\delta}}^{(t)} + \left[\hat{\phi}^{-1}\boldsymbol{U}^{\intercal}\boldsymbol{\Delta}^{-(t)}\boldsymbol{U} + \boldsymbol{\Lambda}^{-1}\right]^{-1}\left(\hat{\phi}^{-1}\boldsymbol{U}^{\intercal}(\boldsymbol{y}-\boldsymbol{\mu}^{(t)}) - \boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{\delta}}^{(t)}\right) \\
&= \left[\boldsymbol{U}^{\intercal}\boldsymbol{\Delta}^{-(t)}\boldsymbol{U} + \hat{\phi}\boldsymbol{\Lambda}^{-1}\right]^{-1}\boldsymbol{U}^{\intercal}\boldsymbol{\Delta}^{-(t)}\left(\boldsymbol{\Delta}^{(t)}(\boldsymbol{y}-\boldsymbol{\mu}^{(t)}) + \boldsymbol{U}\tilde{\boldsymbol{\delta}}^{(t)}\right), \qquad \text{(B.1)}
\end{aligned}$$

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

be rewritten as

$$\tilde{\boldsymbol{\delta}}^{(t+1)} = \left[\boldsymbol{U}^\intercal \boldsymbol{\Delta}^{-(t)} \boldsymbol{U} + \hat{\phi}\boldsymbol{\Lambda}^{-1}\right]^{-1} \boldsymbol{U}^\intercal \boldsymbol{\Delta}^{-(t)} \left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\Theta}^{(t)}\right),$$

which can be equivalently obtained as the solutions to the following generalized ridge weighted least-squares (WLS) problem

$$\tilde{\boldsymbol{\delta}}^{(t+1)} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}}\ \hat{\phi}^{-1} \left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\Theta}^{(t)} - \boldsymbol{U}\boldsymbol{\delta}\right)^\intercal \boldsymbol{\Delta}^{-(t)} \left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\Theta}^{(t)} - \boldsymbol{U}\boldsymbol{\delta}\right) + \boldsymbol{\delta}^\intercal \boldsymbol{\Lambda}^{-1} \boldsymbol{\delta}. \tag{B.2}$$

Equation (B.2) is analogous to the principal component ridge regression (PCRR) model (Ødegård et al. [2018]), and demonstrates that PCA and MMs indeed share the same underlying model. At last, to solve (B.2) without repeatedly inverting the $n \times n$ matrix $\boldsymbol{\Sigma}^{(t)} := \boldsymbol{U}^\intercal \boldsymbol{\Delta}^{-(t)} \boldsymbol{U} + \hat{\phi}\boldsymbol{\Lambda}^{-1}$, we propose using a coordinate descent algorithm (Kooij [2007]), for which each coordinate's updates are given, for $j = 1, ..., n$, by

$$\tilde{\delta}_j \leftarrow \frac{\sum_{i=1}^n w_i U_{ij} \left(\tilde{Y}_i - \boldsymbol{X}_i \boldsymbol{\Theta}^{(t)} - \sum_{l \neq j} U_{il} \tilde{\delta}_l\right)}{\sum_{i=1}^n w_i U_{ij}^2 + \hat{\phi}\Lambda_j^{-1}}, \tag{B.3}$$

where $w_i = \boldsymbol{\Delta}_{ii}^{-(t)}$.

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

Since the objective function in (4.5) 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. [2010b] for estimation of generalized linear models with convex penalties. Let again $\boldsymbol{X}\boldsymbol{\Theta} = \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{D}\alpha + \boldsymbol{G}\boldsymbol{\beta} + (\boldsymbol{D} \odot \boldsymbol{G})\boldsymbol{\gamma}$ and $\boldsymbol{\Theta}^{(t)}$ be the current iterate,

the iterative step reduces to

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

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

$$\begin{aligned}\boldsymbol{\Theta}^{(t+1)} &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left\{ \frac{1}{2s_t} \left\| \boldsymbol{\Theta} - \left[\boldsymbol{X}^\intercal \boldsymbol{\Delta}^{-(t)} \boldsymbol{X} \right]^{-1} \boldsymbol{X}^\intercal \boldsymbol{\Delta}^{-(t)} \left(\tilde{\boldsymbol{Y}} - \boldsymbol{U}\tilde{\boldsymbol{\delta}}^{(t+1)} \right) \right\|_2^2 + g(\boldsymbol{\Theta}) \right\} \\ &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left\{ \frac{1}{2s_t} \sum_{i=1}^{n} w_i \left(\tilde{Y}_i - \boldsymbol{X}_i \boldsymbol{\Theta} - \boldsymbol{U}_i \tilde{\boldsymbol{\delta}}^{(t+1)} \right)^2 + (1-\rho)\lambda \sum_j \| (\beta_j, \gamma_j) \|_2 + \rho\lambda \sum_j |\gamma_j| \right\},\end{aligned} \tag{B.4}$$

where $w_i = \boldsymbol{\Delta}_{ii}^{-(t)}$. We use block coordinate descent and minimize (B.4) with respect to each component of $\boldsymbol{\Theta} = (\boldsymbol{\theta}^\intercal, \alpha^\intercal, \boldsymbol{\beta}^\intercal, \boldsymbol{\gamma}^\intercal)^\intercal$. In practice, we set $s_t = 1$ and do not perform step-size optimization. We present in Appendix B.5 the detailed derivations and our block coordinate descent algorithm to obtain PQL regularized estimates for $\boldsymbol{\Theta}$.

B.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. [2012] 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 in Liang et al. [2024]. 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

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

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

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

B.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 linear mixed model

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

where $\hat{\boldsymbol{\Theta}}=\{\hat{\Theta}_k\neq 0|1\leq k\leq 2p+m+1\}$ is the set of non-null predictors, and $\boldsymbol{\epsilon}=g'(\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})\sim\mathcal{N}(0,\boldsymbol{W}^{-1})$, with $\boldsymbol{W}=\phi^{-1}\text{diag}\left\{\frac{a_i}{\nu(\mu_i)[g'(\mu_i)^2]}\right\}$ the diagonal matrix containing weights for each observation. Let $\tilde{\boldsymbol{Y}}_s$ be the latent working vector in a testing set of n_s individuals with predictor set $\boldsymbol{X}_s$. Similar to Bhatnagar et al. [2020b], we assume that the marginal joint distribution of $\tilde{\boldsymbol{Y}}_s$ and $\tilde{\boldsymbol{Y}}$ is multivariate Normal :

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

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

from standard normal theory that

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

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

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

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

B.5 Proximal Newton method

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

$$\begin{aligned}\boldsymbol{\Theta}^{(t+1)} &= \underset{\boldsymbol{\Theta}}{\text{argmin}}\left\{\frac{1}{2s_t}\left\|\boldsymbol{\Theta}-\left[\boldsymbol{X}^\intercal\boldsymbol{\Delta}^{-(t)}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^\intercal\boldsymbol{\Delta}^{-(t)}\left(\tilde{\boldsymbol{Y}}^{(t)}-\boldsymbol{U}\tilde{\boldsymbol{\delta}}^{(t+1)}\right)\right\|_2^2+g(\boldsymbol{\Theta})\right\}\\ &= \underset{\boldsymbol{\Theta}}{\text{argmin}}\left\{\frac{1}{2s_t}\sum_{i=1}^{n}w_i\left(\tilde{Y}_i-\boldsymbol{X}_i\boldsymbol{\Theta}-\boldsymbol{U}_i\tilde{\boldsymbol{\delta}}^{(t+1)}\right)^2+(1-\rho)\lambda\sum_j\|(\beta_j,\gamma_j)\|_2+\rho\lambda\sum_j|\gamma_j|\right\},\end{aligned} \tag{B.5}$$

with $w_i=\boldsymbol{\Delta}_{ii}^{-(t)}$. We use block coordinate descent and minimize (B.5) with respect to each component of $\boldsymbol{\Theta}=(\boldsymbol{\theta}^\intercal,\alpha^\intercal,\boldsymbol{\beta}^\intercal,\boldsymbol{\gamma}^\intercal)^\intercal$. Suppose we have estimates $\tilde{\theta}_l$ for $l\neq j$, $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\delta}}$, it

is straightforward to show that the updates for θ_j and α are given by

$$\tilde{\theta_j} \leftarrow \frac{\sum_{i=1}^n w_i Z_{ij} \left(\tilde{Y}_i - \sum_{l \neq j} Z_{il} \tilde{\theta}_l - D_i \tilde{\alpha} - \boldsymbol{G}_i \tilde{\boldsymbol{\beta}} - (D_i \odot \boldsymbol{G}_i) \tilde{\boldsymbol{\gamma}} - \boldsymbol{U}_i \tilde{\boldsymbol{\delta}} \right)}{\sum_{i=1}^n w_i Z_{ij}^2},$$
$$\tilde{\alpha} \leftarrow \frac{\sum_{i=1}^n w_i D_i \left(\tilde{Y}_i - \boldsymbol{Z}_i \tilde{\boldsymbol{\theta}} - \boldsymbol{G}_i \tilde{\boldsymbol{\beta}} - (D_i \odot \boldsymbol{G}_i) \tilde{\boldsymbol{\gamma}} - \boldsymbol{U}_i \tilde{\boldsymbol{\delta}} \right)}{\sum_{i=1}^n w_i D_i^2}.$$

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

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

where we define the subgradients

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

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

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

Since $\|\boldsymbol{v}\|_2 \leq 1$, 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

$$\left(\sum_{i=1}^{n} w_i G_{ij} r_{i;-j}\right)^2 + \left(S_{\rho\lambda s_t}\left(\sum_{i=1}^{n} w_i (D_i \odot G_{ij}) r_{i;-j}\right)\right)^2 \leq ((1-\rho)\lambda s_t)^2, \tag{B.6}$$

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

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

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

$$\begin{bmatrix} \sum_{i=1}^{n} w_i G_{ij}(r_{i;-j} - D_i G_{ij} \tilde{\gamma}_j) \\ \sum_{i=1}^{n} w_i (D_i \odot G_{ij})(r_{i;-j} - G_{ij}\tilde{\beta}_j) - \rho\lambda s_t u \end{bmatrix} = \left(\begin{bmatrix} \sum_{i=1}^{n} w_i G_{ij}^2 & 0 \\ 0 & \sum_{i=1}^{n} w_i (D_i \odot G_{ij})^2 \end{bmatrix} + \frac{(1-\rho)\lambda s_t}{\sqrt{\tilde{\beta}_j^2 + \tilde{\gamma}_j^2}} \boldsymbol{I}_2 \right) \begin{bmatrix} \tilde{\beta}_j \\ \tilde{\gamma}_j \end{bmatrix}. \tag{B.7}$$

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

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

There is no closed-form solution for (B.7) if both $\tilde{\gamma}_j$ and $\tilde{\beta}_j$ are non-null. In this case, we can replace (B.5) by a surrogate objective function using a majorization-minorization algorithm (Wu and Lange [2008]). 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

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

from where we derive the majorization-minimization iterative step

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

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

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

where we defined $\tilde{s}_t = s_t / \|\beta_j^{(t)}, \gamma_j^{(t)}\|_2$. Algorithm 3 below summarizes our block coordinate descent (BCD) procedure to obtain regularized estimates for the fixed effects vector $\boldsymbol{\Theta} = (\boldsymbol{\theta}^\intercal, \alpha^\intercal, \boldsymbol{\beta}^\intercal, \boldsymbol{\gamma}^\intercal)^\intercal$.

Algorithm 3 BCD algorithm to minimize the PQL loss function of the GEI model (4.5) with mixed lasso and group lasso penalties for GLMMs.

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

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

Estimate τ_g, τ_d and ϕ under the null model (i.e. $\boldsymbol{\beta} = \boldsymbol{\gamma} = \boldsymbol{0}$) using the AI-REML algorithm;

Given $\hat{\tau}_g, \hat{\tau}_d$ and $\hat{\phi}$, perform spectral decomposition of the random effects covariance matrix $\hat{\tau}_g \boldsymbol{K} + \hat{\tau}_d \boldsymbol{K}^D = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\intercal$;

Initialize $\boldsymbol{\Theta}^{(0)} = (\boldsymbol{\theta}^{(0)\intercal}, \alpha^{(0)\intercal}, \boldsymbol{\beta}^{(0)\intercal}, \boldsymbol{\gamma}^{(0)\intercal})^\intercal$ and $\tilde{\boldsymbol{\delta}}^{(0)}$;

for $\lambda = \lambda_1, \lambda_2, ...$ **do**

  for *t=0,1,... until convergence* **do**

    Select a suitable step size s_t;

    Update $\boldsymbol{\mu}^{(t)} \leftarrow g^{-1}(\boldsymbol{X}\boldsymbol{\Theta}^{(t)} + \boldsymbol{U}\tilde{\boldsymbol{\delta}}^{(t)})$, $\boldsymbol{\Delta}^{(t)} \leftarrow \text{diag}(g'(\boldsymbol{\mu}^{(t)}))$ and $w_i \leftarrow \boldsymbol{\Delta}_{ii}^{-(t)}$ for $i = 1, ..., n$;

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

    /* **Inner loop to estimate $\tilde{\boldsymbol{\delta}}$**

    for *j=1,...,n until convergence* **do**

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

    Update $\boldsymbol{\mu}^{(t)} \leftarrow g^{-1}(\boldsymbol{X}\boldsymbol{\Theta}^{(t)} + \boldsymbol{U}\tilde{\boldsymbol{\delta}}^{(t+1)})$;

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

    /* **Inner loop to estimate $\boldsymbol{\Theta}^{(t+1)}$**

    for *k=1,...,m until convergence* **do**

$$\tilde{\theta_k} \leftarrow \frac{\sum_{i=1}^n w_i Z_{ij}\left(\tilde{Y}_i - \sum_{l\neq k} Z_{il}\tilde{\theta}_l - D_i\tilde{\alpha} - \boldsymbol{G}_i\tilde{\boldsymbol{\beta}} - (D_i \odot \boldsymbol{G}_i)\tilde{\boldsymbol{\gamma}} - \boldsymbol{U}_i\tilde{\boldsymbol{\delta}}\right)}{\sum_{i=1}^n w_i Z_{ik}^2},$$

$$\tilde{\alpha} \leftarrow \frac{\sum_{i=1}^n w_i D_i\left(\tilde{Y}_i - \boldsymbol{Z}_i\tilde{\boldsymbol{\theta}} - \boldsymbol{G}_i\tilde{\boldsymbol{\beta}} - (D_i \odot \boldsymbol{G}_i)\tilde{\boldsymbol{\gamma}} - \boldsymbol{U}_i\tilde{\boldsymbol{\delta}}\right)}{\sum_{i=1}^n w_i D_i^2};$$

  for *j=1,...,p until convergence* **do**

    Compute $r_{i;-j} = \tilde{Y}_i - \boldsymbol{Z}_i\tilde{\boldsymbol{\theta}} - D_i\tilde{\alpha} - \sum_{l\neq j} G_{il}\tilde{\beta}_l - \sum_{l\neq j}(D_i \odot G_{il})\tilde{\gamma}_l - \boldsymbol{U}_i\tilde{\boldsymbol{\delta}}$;

    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

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

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

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

APPENDIX C

Appendix to Manuscript 3

C.1 Supplementary Tables

Table C.1: Mean and standard deviation of the relative bias (%) of variance parameters estimated under the null model of no genetic association when simulating continuous phenotypes with no causal predictor.

		Number of PCs		
GRM	Variable	0	10	20
full	ϕ	0.13 (1.89)	0.14 (1.89)	0.15 (1.89)
	ψ_1	-0.49 (12.5)	0.01 (10.8)	-0.10 (10.6)
	ψ_2	0.89 (12.5)	1.05 (12.2)	1.10 (12.2)
	ψ_3	-1.49 (20.2)	-1.43 (20.3)	-1.86 (20.3)
	ψ_4	0.55 (8.28)	0.59 (7.94)	0.55 (7.97)
	ψ_5	-0.20 (9.66)	-0.41 (9.94)	-0.43 (10.1)
	ψ_6	0.97 (6.22)	1.04 (6.26)	0.99 (6.30)
	τ	33.0 (27.2)	1.64 (13.7)	1.19 (13.5)
sparse	ϕ	0.12 (1.89)	0.14 (1.91)	0.15 (1.90)
	ψ_1	-30.6 (23.6)	-5.86 (20.0)	-4.86 (21.2)
	ψ_2	6.29 (24.2)	1.16 (12.3)	1.27 (12.3)
	ψ_3	-5.53 (28.2)	-1.28 (20.2)	-1.68 (20.2)
	ψ_4	5.76 (14.2)	1.01 (9.11)	0.53 (8.31)
	ψ_5	-0.55 (9.89)	-0.68 (10.2)	-0.68 (10.3)
	ψ_6	10.1 (17.3)	1.84 (8.03)	1.12 (6.47)
	τ	72.7 (44.6)	8.68 (25.2)	6.35 (26.0)

Table C.2: Mean and standard deviation of the 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.

		Number of PCs		
GRM	Variable	0	10	20
full	ϕ	-0.08 (2.04)	-0.09 (1.99)	-0.09 (1.98)
	ψ_1	9.42 (14.4)	8.99 (13.9)	9.25 (14.0)
	ψ_2	-1.11 (16.0)	-1.85 (14.9)	-1.68 (15.1)
	ψ_3	-2.58 (18.6)	-4.16 (17.9)	-4.41 (17.9)
	ψ_4	-0.24 (9.24)	0.14 (8.89)	0.07 (8.96)
	ψ_5	-2.24 (11.1)	-1.79 (10.9)	-1.79 (11.0)
	ψ_6	0.15 (8.17)	0.17 (7.89)	0.14 (7.85)
	τ	46.3 (27.0)	13.5 (15.8)	12.6 (15.9)
sparse	ϕ	-0.06 (2.01)	-0.08 (1.97)	-0.08 (1.96)
	ψ_1	-26.7 (21.3)	-5.35 (25.0)	-4.45 (24.8)
	ψ_2	1.11 (17.8)	-1.59 (15.2)	-1.47 (15.5)
	ψ_3	-4.17 (20.6)	-5.26 (19.3)	-5.53 (19.3)
	ψ_4	6.69 (13.8)	0.54 (9.71)	0.52 (9.79)
	ψ_5	-1.89 (11.4)	-1.57 (11.1)	-1.51 (11.1)
	ψ_6	12.1 (19.4)	1.02 (9.29)	1.01 (9.26)
	τ	95.4 (47.2)	30.5 (29.6)	28.90 (29.0)

Table C.3: Mean and standard deviation of the relative bias (%) of variance parameters estimated under the null model of no genetic association when simulating binary phenotypes with no causal predictor.

		Number of PCs		
GRM	Variable	0	10	20
full	ψ_1	-7.29 (44.8)	16.2 (62.4)	16.2 (63.5)
	ψ_2	44.1 (32.1)	51.6 (38.5)	51.4 (38.7)
	ψ_3	-18.4 (40.7)	-6.85 (41.1)	-8.69 (40.5)
	ψ_4	-9.87 (32.0)	-19.4 (60.7)	-19.8 (60.6)
	ψ_5	-26.5 (30.1)	-35.4 (44.8)	-35.6 (44.9)
	ψ_6	-42.0 (20.8)	-36.3 (52.7)	-36.4 (51.8)
	τ	73.7 (47.2)	41.9 (37.3)	42.3 (37.6)
sparse	ψ_1	-54.6 (50.7)	4.18 (57.1)	7.17 (56.4)
	ψ_2	72.3 (36.2)	51.8 (37.5)	48.4 (36.6)
	ψ_3	-57.9 (44.7)	-19.2 (35.3)	-14.3 (37.6)
	ψ_4	-12.2 (30.6)	-18.1 (57.3)	-16.3 (56.1)
	ψ_5	-23.6 (28.5)	-36.2 (42.1)	-33.8 (41.5)
	ψ_6	-39.8 (26.5)	-36.9 (54.5)	-35.3 (54.7)
	τ	136 (81.7)	49.9 (57.2)	46.6 (60.3)

Table C.4: Mean and standard deviation of the relative bias (%) of variance parameters estimated under the null model of no genetic association when simulating binary phenotypes with 100 causal predictors explaining 2% of heritability.

		Number of PCs		
GRM	Variable	0	10	20
full	ψ_1	-5.20 (39.1)	9.49 (49.3)	9.55 (48.3)
	ψ_2	39.4 (25.3)	44.8 (35.0)	46.6 (35.2)
	ψ_3	-19.4 (43.4)	-17.2 (45.3)	-15.9 (45.9)
	ψ_4	-6.20 (19.8)	-19.0 (41.4)	-21.0 (43.2)
	ψ_5	-28.4 (25.2)	-31.2 (37.1)	-34.7 (39.2)
	ψ_6	-42.3 (18.8)	-43.4 (21.6)	-45.9 (21.5)
	τ	96.2 (39.0)	64.6 (37.7)	63.5 (37.5)
sparse	ψ_1	-54.4 (65.7)	-13.7 (71.7)	-18.3 (66.0)
	ψ_2	74.1 (32.1)	61.0 (37.2)	65.0 (36.3)
	ψ_3	-62.5 (47.6)	-36.7 (52.0)	-38.8 (52.5)
	ψ_4	-7.81 (39.8)	-19.3 (50.8)	-24.7 (41.2)
	ψ_5	-27.6 (26.9)	-33.3 (37.4)	-36.8 (39.5)
	ψ_6	-34.5 (53.7)	-40.4 (54.5)	-49.2 (22.1)
	τ	169 (71.4)	95.4 (69.2)	93.1 (67.8)

Table C.5: Point estimates of the residual variance ϕ, polygenic variance component τ and within-individual random effects variances and covariance parameters ψ_1, ψ_2 and ψ_3 estimated under the null model of no genetic association using the AIREML algorithm.

	Hyperactivity		Aggression		Opposition	
	CCA	SI	CCA	SI	CCA	SI
$\hat{\phi}$	0.119	0.078	0.070	0.049	0.098	0.063
$\hat{\tau}$	0.183	0.195	0.080	0.086	0.189	0.179
$\hat{\psi}_1$	0.308	0.505	0.350	0.365	0.237	0.413
$\hat{\psi}_2$	0.006	0.013	0.010	0.011	0.010	0.015
$\hat{\psi}_3$	-0.042	-0.078	-0.057	-0.061	-0.047	-0.075

CCA = Complete case analysis. SI = Single imputation.

Table C.6: Common SNPs selected by the penalized mixed model for all three externalizing scores.

							$\hat{\beta}$		
Analysis	SNP	CHR	POS	A1/A2	MAF	Gene : Consequence	Hyp	Aggr	Opp
CCA	rs4653589	1	224688598	G / A	0.172	CNIH3 : Intronic	0.013	0.013	0.015
	rs12123482	1	233105677	A / G	0.019	NTPCR : Missense	-0.006	-0.052	-0.034
	rs10491702	9	2227837	A / C	0.049	LOC107987043 : Intronic	0.019	0.001	0.005
SI	rs1181883	1	3677933	C / T	0.401	CCDC27 : Missense	0.005	$9x10^{-7}$	0.002
	rs6702929	1	25406674	A / C	0.437	None	0.005	0.007	0.007
	rs4653589	1	224688598	G / A	0.171	CNIH3 : Intronic	0.002	0.016	0.020
	rs12994424	2	26183142	T / C	0.101	KIF3C : Intronic	0.002	0.010	0.006
	rs10179260	2	226521752	G / A	0.085	NYAP2 : Intronic	-0.007	-0.003	-0.017
	rs2292997	3	183724072	A / G	0.099	ABCC5 : Intronic	-0.004	-0.006	-0.022
	rs1250109	4	1227951	C / T	0.161	CTBP1 : Intronic	-0.007	-0.027	-0.002
	rs4865047	4	56821806	T / C	0.096	CEP135 : Intronic	-0.003	-0.012	-0.013
	rs1490794	5	67168343	A / G	0.230	None	0.008	0.006	0.004
	rs4292570	6	82763865	C / T	0.200	LINC02542 : Intronic	-0.103	-0.008	-0.032
	rs9372528	6	119704829	T / C	0.069	None	0.015	0.0009	0.016
	rs547124	11	120052554	A / G	0.141	LOC124902773 : Intronic	-0.001	-0.022	-0.011
	rs12273131	11	123742913	T / C	0.232	None	-0.005	$-3x10^{-6}$	-0.013
	rs6571394	14	21236181	T / C	0.458	LOC107984671 : Intronic	-0.025	-0.023	-0.001
	rs4932232	15	90114309	G / A	0.405	None	-0.008	-0.043	-0.028
	rs2951665	17	32075015	T / C	0.485	ASIC2 : Intronic	0.003	0.002	0.008

CCA = Complete case analysis. SI = Single imputation. MAF = Minor allele frequency.
Hyp = Hyperactivity. Aggr = Aggression. Opp = Opposition.
Note 1: MAF was reported for the effect allele A1.

Table C.7: Common SNPs selected by the adaptive penalized mixed model for all three externalizing scores.

							$\hat{\beta}$		
Analysis	SNP	CHR	POS	A1/A2	MAF	Gene : Consequence	Hyp	Aggr	Opp
CCA	**rs4653589**	1	224688598	G / A	0.172	CNIH3 : Intronic	0.016	0.009	0.012
	rs12123482	1	233105677	A / G	0.019	NTPCR : Missense	-0.023	-0.092	-0.010
SI	**rs6702929**	1	25406674	A / C	0.437	None	0.004	0.005	0.004
	rs921197	1	30319889	G / A	0.473	None	-0.0006	-0.002	-0.006
	rs4653589	1	224688598	G / A	0.171	CNIH3 : Intronic	0.008	0.010	0.013
	rs12123482	1	233105677	A / G	0.020	NTPCR : Missense	-0.019	-0.034	-0.037
	rs13027447	2	139943532	C / T	0.181	None	0.018	0.004	0.013
	rs10179260	2	226521752	G / A	0.085	NYAP2 : Intronic	-0.001	-0.0008	-0.016
	rs9819889	3	134584764	A / G	0.322	EPHB1 : Intronic	0.0001	0.018	0.012
	rs11922733	3	183260759	T / C	0.057	KLHL6 : Intronic	0.003	0.0003	0.008
	rs1250109	4	1227951	C / T	0.161	CTBP1 : Intronic	-0.011	-0.028	-0.002
	rs4835163	4	150348995	T / C	0.143	IQCM : Intronic	0.013	0.007	0.009
	rs7716386	5	9964861	A / G	0.183	LOC107986405 : Intronic	-0.0001	-0.036	-0.028
	rs4292570	6	82763865	C / T	0.200	LINC02542 : Intronic	-0.100	-0.027	-0.035
	rs2986977	10	7950558	A / G	0.255	TAF3 : Intronic	-0.013	-0.004	-0.015
	rs547124	11	120052554	A / G	0.141	LOC124902773 : Intronic	-0.004	-0.023	-0.009
	rs7138693	12	75257224	G / T	0.278	LOC105369842 : Non coding	-0.015	-0.023	-0.008
	rs6571394	14	21236181	T / C	0.458	LOC107984671 : Intronic	-0.053	-0.050	-0.010
	rs4932232	15	90114309	G / A	0.405	None	-0.007	-0.045	-0.020

CCA = Complete case analysis. SI = Single imputation. MAF = Minor allele frequency.
Hyp = Hyperactivity. Aggr = Aggression. Opp = Opposition.
SNPs in bold are SNPs that were also selected by the penalized mixed model (Table C.6).
Note 1: MAF was reported for the effect allele A1.

C.2 Supplementary Figures

Figure C.1: Relative bias of variance parameters estimated under the null model of no genetic association when simulating binary 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.

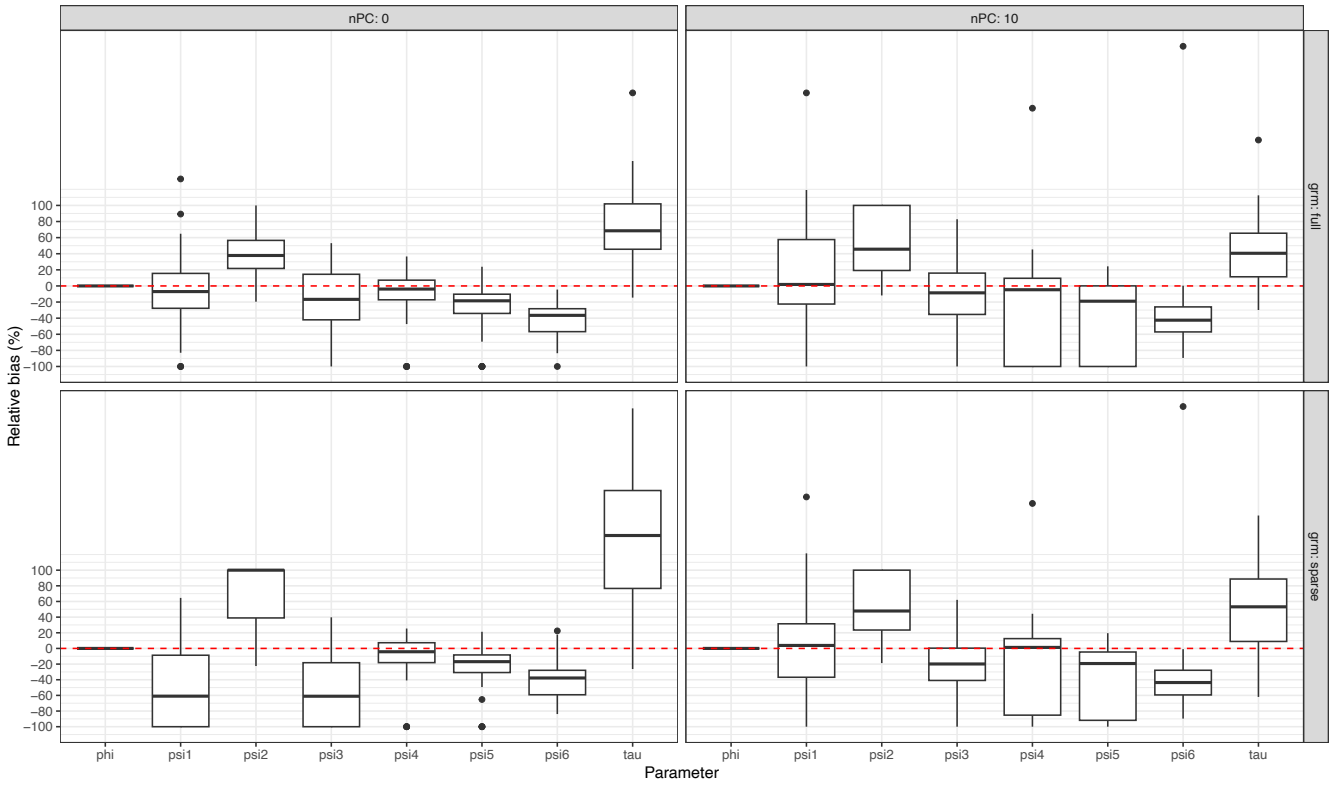


Figure C.2: Relative bias of variance parameters estimated under the null model of no genetic association when simulating binary phenotypes with 100 causal predictors explaining 2% of heritability. Top-left panel: Model with a full GRM and no PC to control for genetic ancestry. Top-right panel: Model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

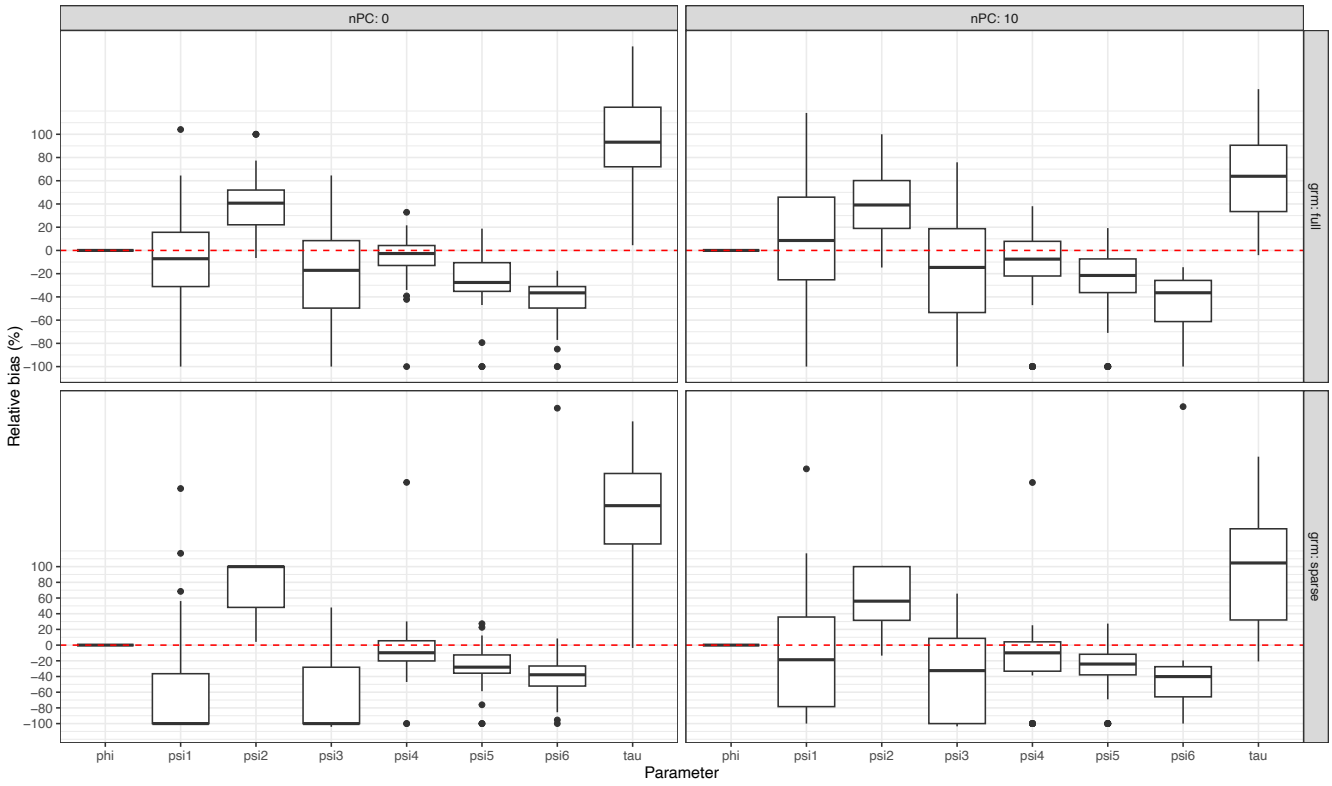


Figure C.3: Precision-recall curve for selection of genetic predictors for our proposed method as a function of the modelling strategy. The left and right panels illustrate the average performance of the method over 50 replications for the simulation model with binary phenotypes and 100 causal predictors explaining 2% and 10% of heritability respectively.

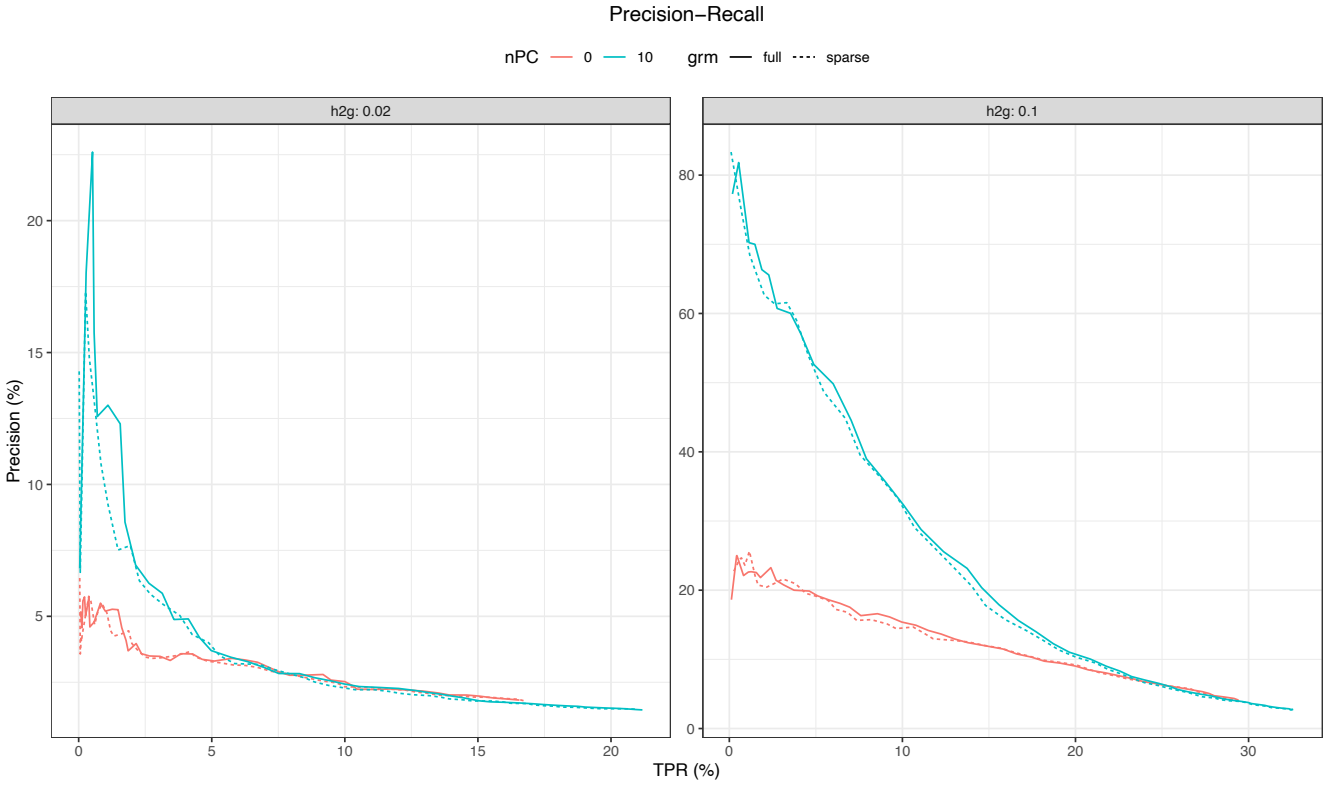


Figure C.4: Precision-recall curve for selection of genetic predictors for the three compared methods. The left and right panels illustrate the average performance with 95% confidence interval of the methods over 50 replications for the simulation model with binary phenotypes and 100 causal predictors explaining 2% and 10% of heritability respectively.

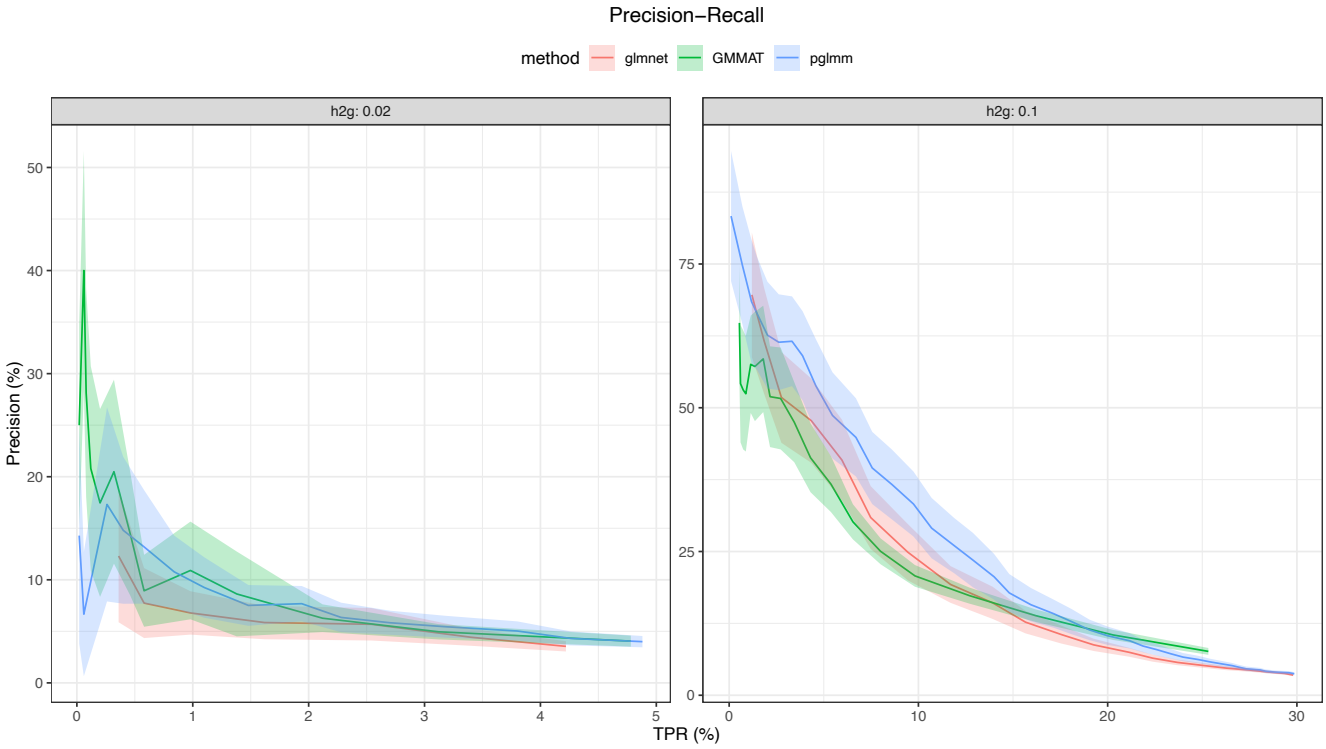


Figure C.5: Performance of the mixed lasso prediction model (R^2_{MSPE}) on the test set as a function of the number of selected genetic predictors. The left and right panels are respectively for the complete case analysis (CCA) and single imputation (SI) model. The dashed vertical lines represent the best model for each externalizing score.

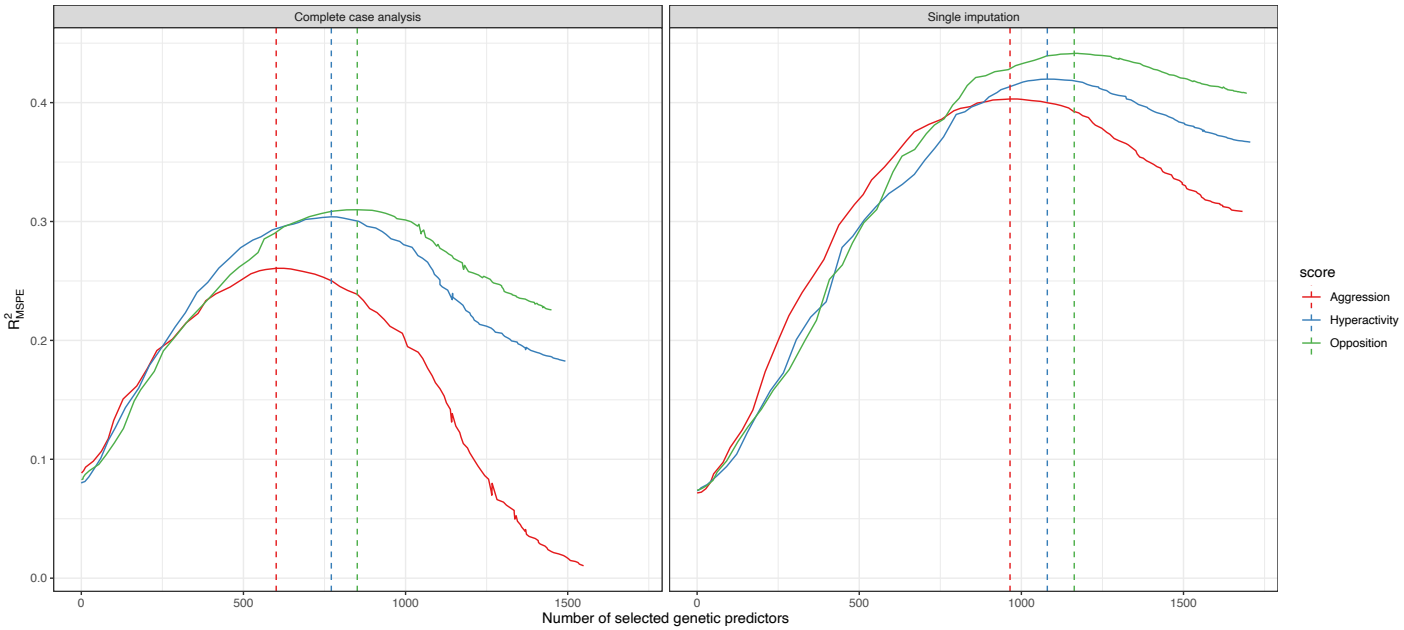
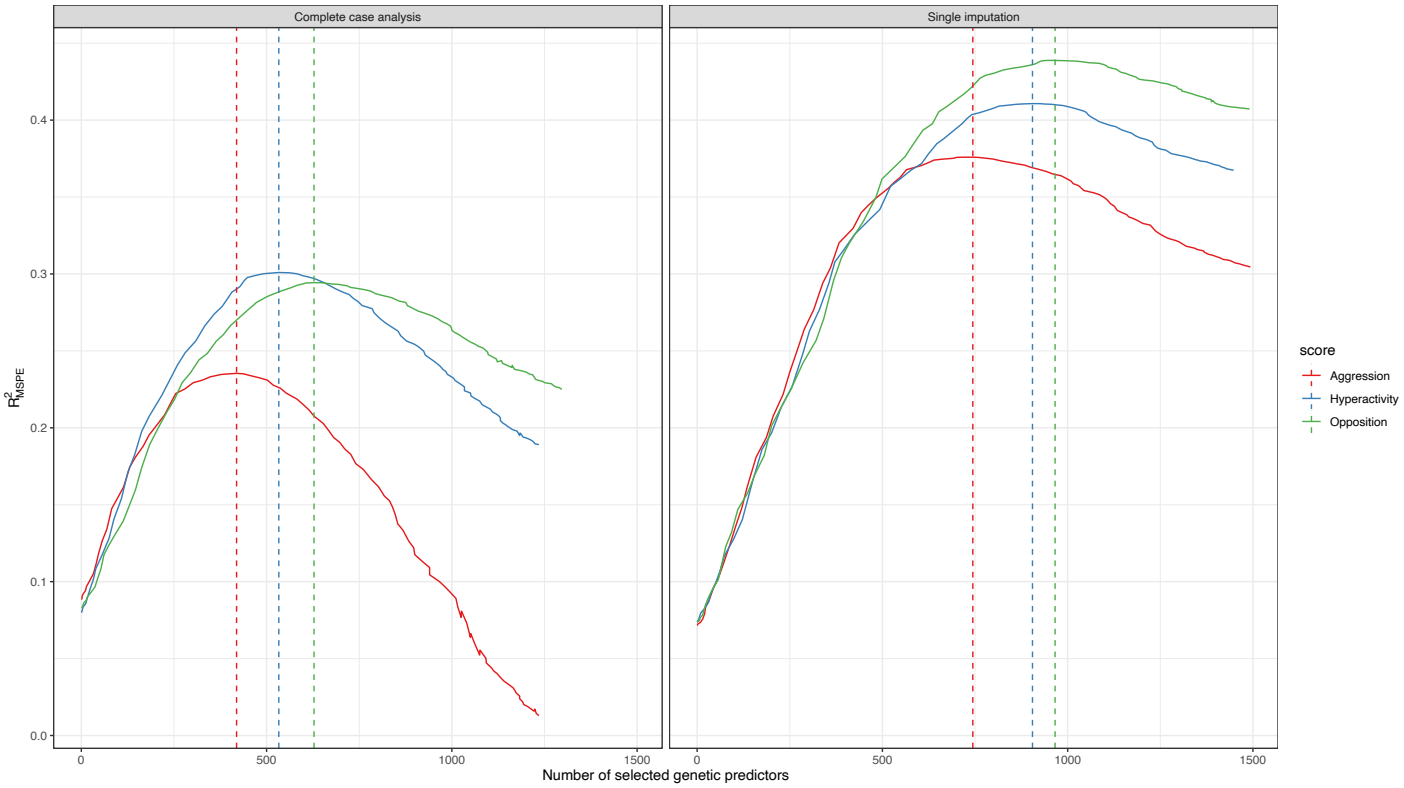


Figure C.6: Performance of the adaptive mixed lasso prediction model (R^2_{MSPE}) on the test set as a function of the number of selected genetic predictors. The left and right panels are respectively for the complete case analysis (CCA) and single imputation (SI) model. The dashed vertical lines represent the best model for each externalizing score. Adaptive weights were estimated by fitting an elastic-net model on the training data.



C.3 Genotype quality control

Genotyping was conducted using the Infinium PsychArray-24 v1.3 BeadChip. The quality control (QC) of genetic data was conducted in PLINK v1.90b5.3, PLINK v1.90b6.7 (Chang et al. [2015]), and R v3.4.3. Pre-imputation QC of genotype data consisted of the following steps:

1. Removal of SNPs with call rates $< 98\%$ or a minor allele frequency (MAF) $< 1\%$

2. Removal of individuals with genotyping rates $< 95\%$

3. Removal of sex mismatches

4. Removal of genetic duplicates

5. Removal of cryptic relatives with pi-hat≥ 12.5

6. Removal of genetic outliers with a distance from the mean of > 4 SD in the first eight multidimensional scaling (MDS) ancestry components

7. Removal of individuals with a deviation of the autosomal or X-chromosomal heterozygosity from the mean > 4 SD

8. Removal of non-autosomal variants

9. Removal of SNPs with call rates $< 98\%$ or a MAF $< 5\%$ or Hardy-Weinberg Equilibrium (HWE) test p-values $< 1 \times 10^{-3}$

10. Removal of A/T and G/C SNPs

11. Update of variant IDs and positions to the IDs and positions in the 1000 Genomes Phase 3 reference panel

12. Alignment of alleles to the reference panel

13. Removal of duplicated variants and variants not present in the reference panel

References

Gad Abraham, Yixuan Qiu, and Michael Inouye. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, May 2017. 10.1093/bioinformatics/btx299. URL https://doi.org/10.1093/bioinformatics/btx299.

Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998. ISBN 978-1-4612-1694-0. 10.1007/978-1-4612-1694-0_15. URL https://doi.org/10.1007/978-1-4612-1694-0_15.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. URL https://doi.org/10.1137/141000671.

Sahir R. Bhatnagar, Yi Yang, and Celia M. T. Greenwood. ggmix: Variable selection in linear mixed models for SNP data. R package version 0.0.1. 2020a. URL https://github.com/sahirbhatnagar/ggmix.

Sahir R. Bhatnagar, Yi Yang, Tianyuan Lu, Erwin Schurr, JC Loredo-Osti, Marie Forest, Karim Oualkacha, and Celia M. T. Greenwood. Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. *PLOS Genetics*, 16(5):e1008766, May 2020b. 10.1371/journal.pgen.1008766. URL https://doi.org/10.1371/journal.pgen.1008766.

Wenjian Bi, Wei Zhou, Rounak Dey, Bhramar Mukherjee, Joshua N. Sampson, and Seunggeun Lee. Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes. *The American Journal of Human Genetics*, 108 (5):825–839, May 2021. 10.1016/j.ajhg.2021.03.019. URL https://doi.org/10.1016/j.ajhg.2021.03.019.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3), June 2013. ISSN 0090-5364. 10.1214/13-aos1096. URL http://dx.doi.org/10.1214/13-AOS1096.

Dankmar Böhning and Bruce G. Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, December 1988. ISSN 0020-3157, 1572-9052. 10.1007/BF00049423.

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.

Howard D. Bondell, Arun Krishna, and Sujit K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, December 2010. ISSN 1541-0420. 10.1111/j.1541-0420.2010.01391.x. URL http://dx.doi.org/10.1111/j.1541-0420.2010.01391.x.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1), March 2011. ISSN 1932-6157. 10.1214/10-aoas388. URL http://dx.doi.org/10.1214/10-AOAS388.

Norman E. Breslow and David G. Clayton. Approximate Inference in Generalized Linear

Mixed Models. *Journal of the American Statistical Association*, 88(421):9–25, March 1993. ISSN 0162-1459, 1537-274X. 10.1080/01621459.1993.10594284.

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.

Gregor Buch, Andreas Schulz, Irene Schmidtmann, Konstantin Strauch, and Philipp S. Wild. Sparse group penalties for bi-level variable selection. *Biometrical Journal*, 66(4), May 2024. ISSN 1521-4036. 10.1002/bimj.202200334. URL http://dx.doi.org/10.1002/bimj.202200334.

C. H. Bueno, D. D. Pereira, M. P. Pattussi, P. K. Grossi, and M. L. Grossi. Gender differences in temporomandibular disorders in adult populational studies: A systematic review and meta-analysis. *Journal of Oral Rehabilitation*, 45(9):720–729, June 2018. 10.1111/joor.12661. URL https://doi.org/10.1111/joor.12661.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018. 10.1038/s41586-018-0579-z. URL https://doi.org/10.1038/s41586-018-0579-z.

Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), February 2015. 10.1186/s13742-015-0047-8. URL https://doi.org/10.1186/s13742-015-0047-8.

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.

Wonil Chung, Jun Chen, Constance Turman, Sara Lindstrom, Zhaozhong Zhu, Po-Ru Loh, Peter Kraft, and Liming Liang. Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. *Nature Communications*, 10(1), February 2019. ISSN 2041-1723. 10.1038/s41467-019-08535-0. URL http://dx.doi.org/10.1038/s41467-019-08535-0.

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 Develop-*

ment, 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.

David R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, August 1975. ISSN 1464-3510. 10.1093/biomet/62.2.441. URL http://dx.doi.org/10.1093/biomet/62.2.441.

David R. Cox. Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1, April 1984. 10.2307/1403235. URL https://doi.org/10.2307/1403235.

Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, December 2011. 10 .1038/nmeth.1785. URL https://doi.org/10.1038/nmeth.1785.

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.

Frank Dudbridge and Olivia Fletcher. Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics*, 95(3):301–307, September 2014. 10.1016/j.ajhg.2014.07.014. URL https://doi.org/10.1016/j.ajhg.2014.07.014.

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.

Yingying Fan and Cheng Yong Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552, December 2012. 10.1111/rssb.12001. URL https://doi.org/10.1111/rssb.12001.

Kuangnan Fang, Jingmao Li, Qingzhao Zhang, Yaqing Xu, and Shuangge Ma. Pathological imaging-assisted cancer gene–environment interaction analysis. *Biometrics*, May 2023. 10.1111/biom.13873. URL https://doi.org/10.1111/biom.13873.

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.

Rina Foygel and Mathias Drton. Exact block-wise optimization in group lasso and sparse group lasso for linear regression, 2010.

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, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2), December 2007. ISSN 1932-6157. 10.1214/07-AOAS131.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso, 2010a. URL https://arxiv.org/abs/1001.0736.

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

Abhik Ghosh and Magne Thoresen. Non-concave penalization in linear mixed-effect models and regularized selection of fixed effects. *AStA Advances in Statistical Analysis*, 102(2): 179–210, May 2017. ISSN 1863-818X. 10.1007/s10182-017-0298-z. URL http://dx.doi.org/10.1007/s10182-017-0298-z.

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.

Andrew J Grant and Stephen Burgess. An efficient and robust approach to mendelian randomization with measured pleiotropic effects in a high-dimensional setting. *Biostatistics*, 23(2):609–625, November 2020. ISSN 1468-4357. 10.1093/biostatistics/kxaa045. URL http://dx.doi.org/10.1093/biostatistics/kxaa045.

Andreas Groll and Gerhard Tutz. Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing*, 24(2):137–154, October 2012. ISSN 1573-1375. 10.1007/s11222-012-9359-z. URL http://dx.doi.org/10.1007/s11222-012-9359-z.

Yuwen Gu, Jun Fan, Lingchen Kong, Shiqian Ma, and Hui Zou. ADMM for High-Dimensional Sparse Penalized Quantile Regression. *Technometrics*, 60(3):319–331, May 2018. ISSN 1537-2723. 10.1080/00401706.2017.1345703. URL http://dx.doi.org/10.1080/00401706.2017.1345703.

David A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, June 1977. 10.1080/01621459.1977.10480998. URL https://doi.org/10.1080/01621459.1977.10480998.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 10.1080/00401706.1970.10488634. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

Gabriel E. Hoffman. Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PLoS ONE*, 8(10):e75707, October 2013. 10.1371/journal.pone.0075707. URL https://doi.org/10.1371/journal.pone.0075707.

Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, June 2009. 10.1371/journal.pgen.1000529. URL https://doi.org/10.1371/journal.pgen.1000529.

Liuyi Hu, Wenbin Lu, Jin Zhou, and Hua Zhou. MM ALGORITHMS FOR VARIANCE COMPONENT ESTIMATION AND SELECTION IN LOGISTIC LINEAR MIXED

MODEL. *Statistica Sinica*, 2019. 10.5705/ss.202017.0220. URL https://doi.org/10.5705/ss.202017.0220.

Jian Huang, Jin Liu, Shuangge Ma, and Kai Wang. Accounting for linkage disequilibrium in genome-wide association studies: a penalized regression method. *Statistics and Its Interface*, 6(1):99–115, 2013. ISSN 1938-7997. 10.4310/sii.2013.v6.n1.a10. URL http://dx.doi.org/10.4310/SII.2013.v6.n1.a10.

Francis K. C. Hui, Samuel Müller, and A. H. Welsh. Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association*, 112(519):1323–1333, June 2017. ISSN 1537-274X. 10.1080/01621459.2016.1215989. URL http://dx.doi.org/10.1080/01621459.2016.1215989.

David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, February 2004. 10.1198/0003130042836. URL https://doi.org/10.1198/0003130042836.

Joseph G. Ibrahim, Hongtu Zhu, Ramon I. Garcia, and Ruixin Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503, 2011. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/41242487.

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.

Jana Janková, Rajen D. Shah, Peter Bühlmann, and Richard J. Samworth. Goodness-of-fit Testing in High Dimensional Generalized Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):773–795, 05 2020. ISSN 1369-7412. 10 .1111/rssb.12371. URL https://doi.org/10.1111/rssb.12371.

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.

Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, March 2008. 10.1534/genetics.107.080101. URL https://doi.org/10.1534/genetics.107.080101.

Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, March 2010. 10.1038/ng.548. URL https://doi.org/10.1038/ng.548.

Yuwon Kim, Jinseog Kim, and Yongdai Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2):375–390, 2006. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24307549.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2674097.

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.

Roger Koenker. Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89, October 2004. ISSN 0047-259X. 10.1016/j.jmva.2004.05.006. URL http://dx.doi.org/10.1016/j.jmva.2004.05.006.

Anita Kooij. Prediction accuracy and stability of regression with optimal scaling transformations. *PhD thesis. Faculty of Social and Behavioural Sciences, Leiden University*, 01 2007.

Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9(1):1–20, 2000. ISSN 10618600. URL http://www.jstor.org/stable/1390605.

Jason D. Lee, Yuekai Sun, and Michael A. Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, January 2014. ISSN 1095-7189. 10.1137/130921428. URL http://dx.doi.org/10.1137/130921428.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3), June 2016. ISSN 0090-5364. 10.1214/15-aos1371. URL http://dx.doi.org/10.1214/15-AOS1371.

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.

Xiaoxuan Liang, Aaron Cohen, Anibal Sólon Heinsfeld, Franco Pestilli, and Daniel J. McDonald. sparsegl: An R package for estimating sparse group lasso. *Journal of Statistical Software*, 110(6), 2024. ISSN 1548-7660. 10.18637/jss.v110.i06. URL http://dx.doi.org/10.18637/jss.v110.i06.

Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, July 2015. 10.1080/10618600.2014.938812. URL https://doi.org/10.1080/10618600.2014.938812.

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.

Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, September 2011. 10.1038/nmeth.1681. URL https://doi.org/10.1038/nmeth.1681.

Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, Nick Patterson, and Alkes L Price. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290, February 2015. ISSN 1546-1718. 10.1038/ng.3190. URL http://dx.doi.org/10.1038/ng.3190.

Hillary M. Heiling, Naim U. Rashid, Quefeng Li, and Joseph G. Ibrahim. glmmPen: High Dimensional Penalized Generalized Linear Mixed Models. *The R Journal*, 15(4):106–128,

April 2024. ISSN 2073-4859. 10.32614/rj-2023-086. URL http://dx.doi.org/10.32614/rj-2023-086.

William Maixner, Luda Diatchenko, Ronald Dubner, Roger B. Fillingim, Joel D. Greenspan, Charles Knott, Richard Ohrbach, Bruce Weir, and Gary D. Slade. Orofacial pain prospective evaluation and risk assessment study – the OPPERA study. *The Journal of Pain*, 12 (11):T4–T11.e2, November 2011. 10.1016/j.jpain.2011.08.002. URL https://doi.org/10.1016/j.jpain.2011.08.002.

Nathalie Malo, Ondrej Libiger, and Nicholas J. Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *The American Journal of Human Genetics*, 82(2):375–385, February 2008. ISSN 0002-9297. 10.1016/j.ajhg.2007.10.012. URL http://dx.doi.org/10.1016/j.ajhg.2007.10.012.

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.

Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009. ISSN 1476-4687. 10.1038/nature08494. URL http://dx.doi.org/10.1038/nature08494.

Andries T. Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. A tutorial on conducting genome-wide as-

sociation studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):e1608, February 2018. 10.1002/mpr.1608. URL https://doi.org/10.1002/mpr.1608.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):53–71, January 2008. ISSN 1467-9868. 10.1111/j.1467-9868.2007.00627.x. URL http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x.

Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, September 2007. 10.1016/j.csda.2006.12.019. URL https://doi.org/10.1016/j.csda.2006.12.019.

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.

Bhramar Mukherjee, Jaeil Ahn, Stephen B. Gruber, Malay Ghosh, and Nilanjan Chatterjee. Case-Control Studies of Gene-Environment Interaction: Bayesian Design and Analysis. *Biometrics*, 66(3):934–948, November 2009. 10.1111/j.1541-0420.2009.01357.x. URL https://doi.org/10.1111/j.1541-0420.2009.01357.x.

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.

John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, April 2008. 10.1038/ng.139. URL https://doi.org/10.1038/ng.139.

Alejandro Ochoa and John D. Storey. Estimating FST and kinship for arbitrary population structures. *PLOS Genetics*, 17(1):e1009241, January 2021. 10.1371/journal.pgen.1009241. URL https://doi.org/10.1371/journal.pgen.1009241.

Luke J. O'Connor, Armin P. Schoech, Farhad Hormozdiari, Steven Gazal, Nick Patterson, and Alkes L. Price. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476, September 2019. 10.1016/j.ajhg.2019.07.003. URL https://doi.org/10.1016/j.ajhg.2019.07.003.

Jørgen Ødegård, Ulf Indahl, Ismo Strandén, and Theo H. E. Meuwissen. Large-scale genomic prediction using singular value decomposition of the genotype matrix. *Genetics Selection Evolution*, 50(1), February 2018. 10.1186/s12711-018-0373-2. URL https://doi.org/10.1186/s12711-018-0373-2.

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.

Ruth Ottman. Gene–environment interaction: Definitions and study design. *Preventive Medicine*, 25(6):764–770, November 1996. 10.1006/pmed.1996.0117. URL https://doi.org/10.1006/pmed.1996.0117.

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

Zhiying Pan and D. Y. Lin. Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61(4):1000–1009, December 2005. ISSN 1541-0420. 10.1111/j.1541-0420.2005.00365.x. URL http://dx.doi.org/10.1111/j.1541-0420.2005.00365.x.

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.

Matthew Pietrosanu, Jueyu Gao, Linglong Kong, Bei Jiang, and Di Niu. Advanced algorithms for penalized quantile and composite quantile regression. *Comput. Stat.*, 36(1):333–346, March 2021.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, July 2006. 10.1038/ng1847. URL https://doi.org/10.1038/ng1847.

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.

Florian Privé, Hugues Aschard, and Michael G. B. Blum. Efficient implementation of penalized regression for genetic risk prediction. *Genetics*, 212(1):65–74, February 2019. 10.1534/genetics.119.302019. URL https://doi.org/10.1534/genetics.119.302019.

Florian Privé, Bjarni J. Vilhjálmsson, and Hugues Aschard. Fitting penalized regressions on very large genetic data using snpnet and bigstatsr. October 2020. 10.1101/2020.10.30 .362079. URL https://doi.org/10.1101/2020.10.30.362079.

Junyang Qian, Yosuke Tanigawa, Wenfei Du, Matthew Aguirre, Chris Chang, Robert Tibshirani, Manuel A. Rivas, and Trevor Hastie. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genetics*, 16(10):e1009141, October 2020. 10.1371/journal.pgen.1009141. URL https://doi.org/10.1371/journal.pgen.1009141.

Assaf Rabinowicz and Saharon Rosset. Cross-validation for correlated data. *Journal of the American Statistical Association*, pages 1–14, September 2020. 10.1080/01621459.2020 .1801451. URL https://doi.org/10.1080/01621459.2020.1801451.

Barbara Rakitsch, Christoph Lippert, Oliver Stegle, and Karsten Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, November 2012. 10.1093/bioinformatics/bts669. URL https://doi.org/10.1093/bioinformatics/bts669.

Anna C. Reisetter and Patrick Breheny. Penalized linear mixed models for structured genetic data. *Genetic Epidemiology*, May 2021. 10.1002/gepi.22384. URL https://doi.org/10.1002/gepi.22384.

Daniel E. Runcie and Lorin Crawford. Fast and flexible linear mixed models for genome-wide genetics. *PLOS Genetics*, 15(2):e1007978, February 2019. ISSN 1553-7404. 10.1371/journal.pgen.1007978. URL http://dx.doi.org/10.1371/journal.pgen.1007978.

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.

Robert Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78 (4):719–727, 1991. ISSN 1464-3510. 10.1093/biomet/78.4.719. URL http://dx.doi.org/10.1093/biomet/78.4.719.

Jürg Schelldorfer, Lukas Meier, and Peter Bühlmann. GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477, April 2014. ISSN 1537-2715. 10.1080/10618600.2013.773239. URL http://dx.doi.org/10.1080/10618600.2013.773239.

Juürg Schelldorfer, Peter Bühlmann, and Sara van de Geer. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38 (2):197–214, May 2011. ISSN 1467-9469. 10.1111/j.1467-9469.2011.00740.x. URL http://dx.doi.org/10.1111/j.1467-9469.2011.00740.x.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), March 1978. 10.1214/aos/1176344136. URL https://doi.org/10.1214/aos/1176344136.

Xia Shen, Moudud Alam, Freddy Fikse, and Lars Rönnegård. A novel generalized ridge

regression method for quantitative genetics. *Genetics*, 193(4):1255–1268, April 2013. 10 .1534/genetics.112.146720. URL https://doi.org/10.1534/genetics.112.146720.

M. Shi, K. M. O'Brien, and C. R. Weinberg. Interactions between a polygenic risk score and non-genetic risk factors in young-onset breast cancer. *Scientific Reports*, 10(1), February 2020. 10.1038/s41598-020-60032-3. URL https://doi.org/10.1038/s41598-020-60032-3.

Noah Simon and Robert Tibshirani. Standardization and the group lasso penalty. *Statistica Sinica*, 22(3), 2012. ISSN 1017-0405. 10.5705/ss.2011.075. URL http://dx.doi.org/10.5705/ss.2011.075.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, April 2013. ISSN 1537-2715. 10.1080/10618600.2012.681250. URL http://dx.doi.org/10.1080/10618600.2012.681250.

Montgomery Slatkin. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, June 2008. ISSN 1471-0064. 10.1038/nrg2361. URL http://dx.doi.org/10.1038/nrg2361.

Shad B. Smith, Marc Parisien, Eric Bair, Inna Belfer, Anne-Julie Chabot-Doré, Pavel Gris, Samar Khoury, Shannon Tansley, Yelizaveta Torosyan, Dmitri V. Zaykin, Olaf Bernhardt, Priscila de Oliveira Serrano, Richard H. Gracely, Deepti Jain, Marjo-Riitta Järvelin, Linda M. Kaste, Kathleen F. Kerr, Thomas Kocher, Raija Lähdesmäki, Nadia Laniado, Cathy C. Laurie, Cecelia A. Laurie, Minna Männikkö, Carolina B. Meloto, Andrea G. Nackley, Sarah C. Nelson, Paula Pesonen, Margarete C. Ribeiro-Dasilva, Celia M. Rizzatti-Barbosa, Anne E. Sanders, Christian Schwahn, Kirsi Sipilä, Tamar Sofer, Alexander Teumer, Jeffrey S. Mogil, Roger B. Fillingim, Joel D. Greenspan, Richard Ohrbach, Gary D. Slade, William Maixner, and Luda Diatchenko. Genome-

wide association reveals contribution of MRAS to painful temporomandibular disorder in males. *Pain*, 160(3):579–591, November 2018. 10.1097/j.pain.0000000000001438. URL https://doi.org/10.1097/j.pain.0000000000001438.

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.

Julien St-Pierre, Karim Oualkacha, and Sahir Rai Bhatnagar. Hierarchical selection of genetic and gene by environment interaction effects in high-dimensional mixed models. *Statistical Methods in Medical Research*, December 2024. ISSN 1477-0334. 10.1177/09622802241293768. URL http://dx.doi.org/10.1177/09622802241293768.

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.

Weijie Su, Małgorzata Bogdan, and Emmanuel Candès. False discoveries occur early on the lasso path. *The Annals of Statistics*, 45(5), October 2017. ISSN 0090-5364. 10.1214/16-aos1521. URL http://dx.doi.org/10.1214/16-AOS1521.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1676. 10.1371/journal.pmed.1001779. URL http://dx.doi.org/10.1371/journal.pmed.1001779.

Jae Hoon Sul, Michael Bilow, Wen-Yun Yang, Emrah Kostem, Nick Furlotte, Dan He, and Eleazar Eskin. Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models. *PLOS Genetics*, 12(3): e1005849, March 2016. ISSN 1553-7404. 10.1371/journal.pgen.1005849.

Jae Hoon Sul, Lana S. Martin, and Eleazar Eskin. Population structure in genetic studies: Confounding factors and mixed models. *PLOS Genetics*, 14(12):e1007309, December 2018. 10.1371/journal.pgen.1007309. URL https://doi.org/10.1371/journal.pgen.1007309.

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.

Timothy A. Thornton and Justo Lorenzo Bermejo. Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genetic Epidemiology*, 38(S1), August 2014. ISSN 1098-2272. 10.1002/gepi.21819. URL http://dx.doi.org/10.1002/gepi.21819.

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.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems.

Journal of the Royal Statistical Society. Series B (Statistical Methodology), 74(2):245–266, 2012. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/41430939.

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, March 1986. ISSN 1537-274X. 10.1080/01621459.1986.10478240. URL http://dx.doi.org/10.1080/01621459.1986.10478240.

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.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, July 2017. ISSN 0002-9297. 10.1016/j.ajhg.2017.06.005. URL http://dx.doi.org/10.1016/j.ajhg.2017.06.005.

Patrik Waldmann, Maja Ferenčaković, Gábor Mészáros, Negar Khayatzadeh, Ino Curik, and Johann Sölkner. AUTALASSO: an automatic adaptive LASSO for genome-wide prediction. *BMC Bioinformatics*, 20(1), April 2019. 10.1186/s12859-019-2743-3. URL https://doi.org/10.1186/s12859-019-2743-3.

Chen Wang, Tianying Wang, Krzysztof Kiryluk, Ying Wei, Hugues Aschard, and Iuliana Ionita-Laza. Genome-wide discovery for biomarkers using quantile regression at biobank scale. *Nature Communications*, 15(1), July 2024. ISSN 2041-1723. 10.1038/s41467-024-50726-x. URL http://dx.doi.org/10.1038/s41467-024-50726-x.

Ying Wang, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher, and Loic Yengo. Theoretical

and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications*, 11(1), July 2020. 10.1038/s41467-020-17719-y. URL https://doi.org/10.1038/s41467-020-17719-y.

Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1), March 2008. 10.1214/07-aoas147. URL https://doi.org/10.1214/07-aoas147.

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, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, June 2010. ISSN 1546-1718. 10.1038/ng.608. URL http://dx.doi.org/10.1038/ng.608.

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.

Yuhong Yang. Can the Strengths of AIC and BIC Be Shared? A Conflict between Model Indentification and Regression Estimation. *Biometrika*, 92(4):937–950, 2005. ISSN 00063444. URL http://www.jstor.org/stable/20441246.

Yiqi Yao and Alejandro Ochoa. Limitations of principal components in quantitative genetic association models for human studies. March 2022. 10.1101/2022.03.25.485885. URL https://doi.org/10.1101/2022.03.25.485885.

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.

Liqun Yu and Nan Lin. ADMM for Penalized Quantile Regression in Big Data. *International Statistical Review*, 85(3):494–518, August 2017. ISSN 1751-5823. 10.1111/insr.12221. URL http://dx.doi.org/10.1111/insr.12221.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, December 2005. ISSN 1467-9868. 10.1111/j.1467-9868.2005.00532.x. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x.

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.

Natalia Zemlianskaia, W. James Gauderman, and Juan Pablo Lewinger. A scalable hierarchical lasso for gene–environment interactions. *Journal of Computational and Graphical Statistics*, pages 1–13, March 2022. 10.1080/10618600.2022.2039161. URL https://doi.org/10.1080/10618600.2022.2039161.

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.

Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A

Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, and Edward S Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360, March 2010. 10.1038/ng.546. URL https://doi.org/10.1038/ng.546.

Huaqing Zhao, Nandita Mitra, Peter A. Kanetsky, Katherine L. Nathanson, and Timothy R. Rebbeck. A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores (PCAPS). *Statistical Applications in Genetics and Molecular Biology*, 17(6), December 2018. 10.1515/sagmb-2017-0054. URL https://doi.org/10.1515/sagmb-2017-0054.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A), December 2009. ISSN 0090-5364. 10.1214/07-aos584. URL http://dx.doi.org/10.1214/07-AOS584.

Qingyuan Zhao, Yang Chen, Jingshu Wang, and Dylan S Small. Powerful three-sample genome-wide design and robust statistical inference in summary-data mendelian randomization. *International Journal of Epidemiology*, 48(5):1478–1492, July 2019. ISSN 1464-3685. 10.1093/ije/dyz142. URL http://dx.doi.org/10.1093/ije/dyz142.

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. OpenMendel: 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.

Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Rounak Dey, Maiken E. Gabrielsen, Brooke N. Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A. Gagliano, Aliya Gifford, Lisa A. Bastarache, Wei-Qi Wei, Joshua C. Denny, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer, and Seunggeun Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, August 2018. ISSN 1546-1718. 10.1038/s41588-018-0184-y. URL http://dx.doi.org/10.1038/s41588-018-0184-y.

Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, June 2012. 10.1038/ng.2310. URL https://doi.org/10.1038/ng.2310.

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.

Hui Zou and Trevor Hastie. Addendum: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67 (5):768–768, November 2005. ISSN 1467-9868. 10.1111/j.1467-9868.2005.00527.x. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00527.x.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of

parameters. *The Annals of Statistics*, 37(4), August 2009. ISSN 0090-5364. 10.1214/08-aos625. URL http://dx.doi.org/10.1214/08-AOS625.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5), October 2007. 10.1214/009053607000000127. URL https://doi.org/10.1214/009053607000000127.