The University of Auckland

Doctoral Thesis

A penalized linear mixed model with generalized method of moments estimators for complex phenotype prediction

---

Author:
Xiaqiong Wang

Supervisor:
Dr. Yalu Wen
Dr. Kathy Ruggiero

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Statistics, the University of Auckland, 2022.

Abstract

Linear mixed models have long been the method of choice for risk prediction analysis on high-dimensional data, where random effect terms are used to capture predictive effects from multiple markers. However, it remains computationally challenging to simultaneously model a large number of variables that can be noise or have predictive effects of complex forms. In this thesis, we first develop a penalized linear mixed model with generalized method of moments estimators for prediction analyses. The proposed method adopts the generalized method of moments estimators to improve computational efficiency and uses the L_1 penalty to select predictors. We show that generalized method of moments estimators have oracle properties, including variable selection consistency, estimation consistency, and asymptotic normality. We further develop a hybrid screening rule that constitutes of the sequential strong rule and the enhanced dual polytope projection rule to reduce data dimension and improve computational efficiency. The proposed hybrid screening rule projects solutions to the objective function of the proposed penalized linear mixed model into the dual space, and then uses the sequential strong rule and the enhanced dual polytope projection rule to detect inactive variables in the space. We show that the hybrid screening rule aligns well with the proposed downstream prediction model, and it can correctly and efficiently discard a large number of variables with no predictive effects in the corresponding penalized linear mixed model. Lastly, we incorporate multiple kernels into the proposed penalized linear mixed model to model high-dimensional multi-omics data, where the interactive roles of multi-omics data and their complex types of predictive effects are captured. Through extensive simulation studies, we have demonstrated that the proposed methods are computationally efficient and can be applied to genome-wide data. They can capture predictive effects of complex forms and outperform competing linear mixed models. In the prediction analyses of PET-imaging outcomes using high-dimensional omics data, we find that the proposed method has better prediction performance than commonly used methods, and our analyses show that genetic variants on *APOE*, *APOC1*, *TOMM40* and *FADS3* genes are highly predictive.

Acknowledgements

Foremost, I would like to express my deep and sincere gratitude to my supervisor Dr. Yalu Wen, for all of her support and wisdom throughout this project, from inception to completion, as well as for her incredible and unwavering patience in working with me over the past three and a half years. Her dynamism, vision, sincerity, and motivation have deeply inspired me. She has taught me the methodology to carry out the research and to present the research works as clearly as possible. It is a great privilege and honor to work and study under her guidance. I am extremely grateful for what she offered me. I would also like to thank her for her friendship, empathy, and caring. I could not have imagined having a better advisor and mentor for my PhD study.

Additionally, I thank and acknowledge my co-supervisor Dr. Kathy Ruggiero, my committee members, and other research staff for their help, expertise, and encouragement. Their insightful comments, feedback, and advice were influential and essential throughout the thesis writing process.

I would like to thank my colleagues for the simulation discussions, for the research days we were working together, and for all the fun we have had over the past three years. I would also like to thank my friends for their encouragement and companionship in my daily life.

I would also like to give special thanks to Precision Driven Health for the financial support they have provided. This scholarship offered me educational opportunities and allowed me to focus on my studies. I am truly grateful for the scholarship support.

My sincere thanks also goes to my boyfriend for his love, understanding, and continuous support in completing this research work. Also, I express my thanks to my younger brother for his support. Last but not least, I would like to thank my father and my mother, for giving birth to me and supporting me spiritually throughout my life. Without the constant support, encouragement and understanding of my family, it would not have been possible for me to achieve my educational goals.

Contents

List of Figures

List of Tables

List of Abbreviations

AD	Alzheimer's disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AV45	florbetapir (a radiotracer used in PET imagery)
BIC	Bayesian information criteria
BLMM	Bayesian linear mixed model
DPP	Dual Polytope Projection
EDPP	enhanced Dual Polytope Projection
FDG	fluorodeoxyglucose (a radiotracer used in PET imagery)
gBLUP	genomic best linear unbiased prediction (method)
GIC	generalized information criterion
GMM	generalized method of moments
GWAS	genome-wide association studies
HpLMMGMM	a hybrid screening rule designed for the penalized linear mixed model with generalized method of moments estimators
JIVE	joint and individual variation explained
KKT	Karush-Kuhn-Tucker (conditions)
LMMs	linear mixed models
LCPUFA	long-chain polyunsaturated fatty acids
LOAD	late-onset Alzheimer's disease
MCI	mild cognitive impairment
MCMC	Markov chain Monte Carlo (algorithm)
MINQUE	minimum norm quadratic unbiased estimation
MIVQUE	minimum variance invariant quadratic unbiased estimation
MKLMM	multi-kernel linear mixed model
MLE	maximum likelihood estimators
MpLMMGMM	a penalized linear mixed model with generalized method of moments estimators for multi-omics data
MSE	mean square error

non-CLIA	non-Clinical Laboratory Improvements Amendments
PET	positron emission tomography
pLMMGMM	a penalized linear mixed model with generalized method of moments estimators
RBF	radial basis function
REML	restricted maximum likelihood estimators
SAFE	SAfe Feature Elimination (rule)
SIS	sure independence screening
SNP	single nucleotide polymorphism
SSR	sequential strong rule

Chapter 1

Introduction

Precision medicine, an emerging model of health care, aims at providing effective treatments tailored according to individual differences, including biomarkers at various molecular levels and other environmental factors (Ashley, 2015; Collins and Varmus, 2015). Accurate disease risk prediction that recognizes individual differences is an essential step in the modern quest for precision medicine. Complex human diseases, such as type 2 diabetes, cancer and coronary heart disease, are regulated by multiple biological pathways at different molecular levels (Kirchner et al., 2013). For example, both genetic regulation and epigenetic regulatory mechanisms (e.g, DNA methylation) can play a key role in cancer cell formation and growth (Das and Singal, 2004). With the recent advances in high-throughput biotechnologies and the initiation of large-scale precision medicine programmes (e.g., the All of Us Research Program), multi-omics data (e.g., genome, transcriptome, methylome, epigenome, proteome and metabolome) that can reflect different aspects of diseases become increasingly available. They are anticipated to advance our understanding of complex human diseases as well as promote precise prediction and treatment strategies (Boekel et al., 2015). The use of findings from ongoing multi-omics studies and other existing knowledge to accurately predict disease risk is expected to revolutionize the current trial-and-error practice of medicine. However, the huge amount of noise (Byrnes et al., 2013), the high computational cost (Weissbrod et al., 2016; Wen and Lu, 2020) and the complex relationships among multi-omics data (Hasin et al., 2017) pose significant analytical challenges. An integrative risk prediction framework that can efficiently detect predictors from ultra-high dimensional multi-omics data and account for their complex relationships is urgently needed (Morris and Baladandayuthapani, 2017; Ritchie et al., 2015; Zeng and Lumley, 2018).

Linear mixed models (LMMs) and their extensions have long been used for prediction analysis on high-dimensional data (Speed and Balding, 2014; Weissbrod et al.,

2016; Wen and Lu, 2020; Yang et al., 2010). Instead of estimating effect sizes from each variable, LMMs model cumulative predictive effects from a group of variables through random effect terms whose variance-covariance structures encode the assumed relationships between predictors and outcomes (Speed and Balding, 2014; Weissbrod et al., 2016). Therefore, LMMs have substantially reduced the number of model parameters, making them applicable in the analysis of high-dimensional data. Within the LMM framework, the genomic best linear unbiased prediction (gBLUP) method, a seminal work proposed by Harris et al., 2008, has been used extensively in genomic risk prediction studies. For example, Yang et al., 2010 used the gBLUP method to predict human heights, where a single random effect term was used to model cumulative predictive effects from all genetic variants. Although efficient to implement, gBLUP assumes that effect sizes of all genetic variants follow the same normal distribution, which is too simple to be realistic. For example, single nucleotide polymorphisms (SNPs) located at different genetic regions (e.g. coding and intron SNPs) can have various type of effect sizes (Speed and Balding, 2014). Recently, Speed and Balding, 2014 proposed the MultiBLUP, where gBLUP is extended to have multiple random effect terms with each capturing the predictive effects from different genomic regions that can be defined by various criteria (e.g., gene or pathway annotations). To account for complex types of predictive effects, Weissbrod et al., 2016 and Wen and Lu, 2020 further generalized MultiBLUP, where predictors are embedded into the reproducing kernel Hilbert space and kernel functions are used to capture non-linear effects. Bayesian LMMs have become more popular in recent years and they can accommodate various disease models by assigning different prior distributions (Habier et al., 2011; Zhao et al., 2006; Zhou et al., 2013). For example, BayesA assumes the variance of each genetic effect is different, and thus uses a scaled univariate student's t distribution (i.e., $\beta_i \sim t(0, t, \sigma_a^2)$) as its prior (Habier et al., 2011). BayesC assumes that a portion (π) of genetic variants have no predictive effects and the remaining $(1-\pi)$ variants have effects on the trait. Therefore, each genetic variant is assumed to follow a mixture of distributions that has a point mass at zero with probability π and a normal distribution with probability $1-\pi$ (i.e., $\beta_i \sim \pi\delta_0 + (1-\pi)N(0, \sigma_a^2)$) (Habier et al., 2011). Indeed, Bayesian LMMs have been widely used for prediction analysis (Dunson, 2001; Hai and Wen, 2020; Zeng and Zhou, 2017; Zhou et al., 2013). For example, Zeng and Zhou, 2017 proposed the latent Dirichlet process regression model to predict eight complex traits in a human cohort, where the non-parametric Dirichlet process is used to efficiently model the effect size distributions, which can be of any form. Zhou et al., 2013 proposed a Bayesian sparse

linear mixed model (BSLMM), which consists of a linear mixed model and a sparse regression model, to predict five phenotypes using two human Genome-Wide Association Studies (GWAS) data sets. The BSLMM sets sparsity-inducing priors for the distribution of genetic effects and uses a Markov chain Monte Carlo (MCMC) algorithm for posterior inference. Rather than using a MCMC algorithm, which is computationally demanding, Hai and Wen, 2020 developed a Bayesian LMM (BLMM) for risk prediction, where a variational Bayesian algorithm that provides an analytical approximation to the posterior distribution is used. Bayesian LMMs offer the flexibility to model a range of phenotypes, but their performance can still be sensitive to the underlying disease model (Zeng and Zhou, 2017; Zhou et al., 2013) and the adopted priors (Gianola, 2013). In addition, Bayesian LMMs tend to be computationally expensive. While LMM-based methods have great potential in modeling high-dimensional multi-omics data, there are several key challenges, including: 1) how to reduce the impact of noise in high-dimensional data; 2) how to efficiently estimate parameters for LMMs with a large number of random effects; 3) how to reduce computational resources needed to handle big multi-omics data; and 4) how to efficiently model complex predictive effects while accounting for the intrinsic dependencies and interactive roles of multi-omics data.

The fundamental assumption used in LMM-based prediction models is that individuals with similar molecular profiles would have similar phenotypes. LMM-based prediction models model cumulative predictive effects from a group of variables through random effect terms whose variance-covariance structures determined by similarity matrices reflect the assumed relationships. Therefore, the performance of a LMM-based risk prediction model highly depends on whether the estimated molecular similarity matrices can capture the predictive patterns of disease-associated markers. For high-dimensional data, the majority of the variables are noise, and thus constructing molecular similarity matrices using all measured variables can dilute the signals, leading to unstable and less accurate estimates of the molecular similarities of disease-relevant markers. To reduce the impact of noise, dimension reduction and variable selection have been considered as an indispensable step for the analysis of high-dimensional data. However, for LMM-based models, most theoretical studies only focus on the selection of fixed effects. For example, Schelldorfer et al., 2011 proposed an L_1 penalized maximum likelihood estimator for LMMs to select the relevant fixed effects in a high-dimensional setting, where the estimation consistency and oracle optimality are

presented. Rohart et al., 2014 consistently selected fixed effects in LMMs with L_1 penalization by using a multi-cycle expectation conditional maximization algorithm, where random effects are considered as missing values in the LMMs. Despite the advances achieved in selecting fixed effect terms, selection of random effects is still challenging. In early studies, traditional methods (e.g, forward/stepwise selection) were used for their selections. However, they lack theoretical guarantee and statistical stability (Pan and Huang, 2014). Information criteria, such as Bayesian information criteria (BIC) and generalized information criterion (GIC), have also been adopted for variable selections. However, they can only maintain consistency for fixed effects and perform substantially worse for the selection of random effects (Pu and Niu, 2006). Recently, Wen and Lu, 2020 and Li et al., 2020 imposed L_1 penalties on the random effect terms for variable selection, where they showed that their methods can achieve estimation and selection consistency and their estimated effects are normally distributed. While imposing L_1 penalties on the random effect terms can enable the variable selection, it becomes computationally challenging for parameter estimations. Indeed, both Wen and Lu, 2020 and Li et al., 2020 could only model a few random effect terms, limiting the application of their methods for genome-wide studies.

The second challenge for existing LMMs with multiple random effects is their parameter estimation. Existing LMMs usually obtain the maximum likelihood estimators (MLE) or the restricted maximum likelihood estimators (REML), both of which are commonly estimated via the Newton-Raspon or expectation-maximization algorithms (Speed and Balding, 2014; VanRaden, 2008; Yang et al., 2010). While MLE and REML are statistically efficient, they can be computationally demanding, especially for LMMs with a large number of random effects. As shown in Wang and Wen, 2021, traditional LMMs can usually handle no more than 10 random effects, which greatly limit their ability to capture complex predictive effects from different groups of variables. The penalized LMMs also suffer from high computational cost. Although the models adopted a one-step approximation procedure to reduce the complexity of their objective function (Fan and Li, 2001; Zou and Li, 2008), their parameter estimates depend on the initial values that are estimated via MLE/REML, and thus is computationally expensive. Generalized method of moments (GMM) is a long-existing alternative to MLE/REML. For example, ANOVA proposed by Fisher, 1992 estimates variance components using the method of moments, where the observed mean squares are equated to their expectations to solve for the variance components. If the data are balanced, the ANOVA estimators are unbiased and have minimum variance among

all the unbiased estimators that are quadratic. However, ANOVA estimators can only keep unbiased property for balanced data and it requires the normality assumption (Swallow and Monahan, 1984). Rao, 1972 proposed the minimum norm quadratic unbiased estimation (MINQUE) method to estimate the variance and covariance components in models, where unbiased quadratic estimators are obtained by minimising a Euclidean norm that measures the size of the covariance matrix of the estimators. MINQUE does not require the normality assumptions built into MLE/REML, and the equations do not need to be solved iteratively (Swallow and Monahan, 1984). Therefore, MINQUE offers more flexibility with reduced computational cost. In addition, Rao, 1971b proposed the minimum variance invariant quadratic unbiased estimation (MIVQUE) method to estimate variance components, finding the minimum variance estimator among all quadratic unbiased estimators of the variance components. Both MINQUE and MIVQUE can be used for balanced and unbalanced data, and MINQUE is equivalent to MIVQUE under normality assumptions (Khuri and Sahai, 1985). While the GMM estimators are less statistically efficient than MLE/REML, they can substantially benefit from their high computational efficiency, as the objective function of GMM can be changed into a quadratic form that is much easier to optimize. Indeed, GMM has been used in genetic studies (Zhou, 2017; Zhu, 1995; Zhu and Weir, 1996). For example, Mathew et al., 2018 used ANOVA to estimate variance components and heritability of biomass allocation and related traits in 99 genotypes of wheat and one triticale. Reimherr and Nicolae, 2016 examined the long-term effects of daily asthma medications on children using a LMM, where variance components were estimated using the MINQUE method. Gianola et al., 2018 predicted the milk yield of Italian Brown Swiss cattle by a LMM, where variance components are estimated via the MINQUE method. El-Moghazy et al., 2015 investigated some factors affecting body weight traits of Zaraibi goat and used the MIVQUE method for mixed models to estimate variance components of random effects. While GMMs have demonstrated their computational efficiencies in estimating parameters in LMMs, they have rarely been used in penalized LMMs, and thus their theoretical properties are not well studied.

One of the features of multi-omics data is their ultra-high dimensionality. LMM-based models attempt to reduce the impact of high-dimension by grouping biomarkers into groups and estimating the cumulative predictive effects from all markers within the group. However, for ultra-high dimensional multi-omics data with intrinsic dependencies across omics, it is natural to have a large number of groups, leading to a LMM with a huge amount of random effects. Neither traditional methods (e.g., information

criteria) nor the most recently developed penalization-based methods are capable of dealing with such data directly. For example, as noted by Wang et al., 2015, when using penalized LMMs, ultra-high dimensional data may not be able to load into the memory. While empirical screening criteria like those adopted in MultiBLUP and MKLMM can be applied to first reduce the number of random effects, there is no guarantee that the discarded variables have no predictive effects. More recently, screening rules have been proposed to reduce the data dimension so that the number of variables that are fed into the prediction models is at a manageable size (Fan and Lv, 2008; Ghaoui et al., 2010; Tibshirani et al., 2012; Wang et al., 2015; Xiang et al., 2016). The sure independence screening rule (SIS) proposed by Fan and Lv, 2008 is one of the seminal works in this area. Fan and Lv, 2008 first proposed the correlation ranking procedure within the context of linear model with Gaussian variables (Fan and Lv, 2008), and then extended it to the generalized linear models, where marginal maximum likelihood is considered (Fan et al., 2009). SIS has also been extended to other models, such as nonparametric additive models (Fan et al., 2011) and the Cox proportional hazard model (Zhao and Li, 2012). SIS has a sure screening property; i.e., with probability tending to 1, SIS can retain all important variables in the model. While attractive and easy to implement, SIS and its extensions cannot consider the joint effects from all predictors as they only focus on the marginal effects of each predictor. In addition, they are not designed for penalized models, leading to a model with redundant variables. To accommodate these issues, screening rules have been developed for penalized models (Ghaoui et al., 2010; Ndiaye et al., 2015; Tibshirani et al., 2012; Wang et al., 2015). According to the guarantees of correctness for discarded variables, these screening rules can be broadly grouped into heuristic screening rules and safe screening rules (Wang et al., 2015). The strong rule proposed by Tibshirani et al., 2012 is one of the well-known heuristic screening rules. Given a penalty value of λ, it discards $\boldsymbol{X}_i$ when $\mid \boldsymbol{X}_i^T\boldsymbol{y} \mid < 2\lambda - \lambda_{max}$, where $\lambda_{max} = \max_i \mid \boldsymbol{X}_i^T\boldsymbol{y} \mid$ and $\boldsymbol{y}$ is the outcome vector (Tibshirani et al., 2012). Therefore, for each penalty, the strong rule can discard a large number of variables and reduce the complexity of penalized models used in downstream analysis. While useful in practice, the strong rule can also mistakenly remove relevant variables, which may reduce the accuracy of a prediction model (Wang et al., 2015). In contrast, as the name indicates, safe screening rules can guarantee that the discarded variables are noise (Wang et al., 2015). The SAfe Feature Elimination (SAFE) rule proposed by Ghaoui et al., 2010 is a pioneer work of safe rules, where the common Lasso problem (i.e., argmin $\frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2+\lambda||\boldsymbol{\beta}||_1$, where $\boldsymbol{X}$ is the variable matrix and $\boldsymbol{\beta}$ is the unknown

coefficients vector) was transformed to a dual problem (Kim et al., 2007):

$$\text{argmax } \frac{1}{2}||\boldsymbol{y}||_2^2 - \frac{\lambda^2}{2}||\boldsymbol{\theta} - \frac{\boldsymbol{y}}{\lambda}||_2^2, \quad s.t. \ |\boldsymbol{X}_i^T\boldsymbol{\theta}| \leq 1$$

where $\boldsymbol{\theta}$ is the dual variable. We use $\boldsymbol{\theta}^*$ to denote the optimal solution of the dual problem, and thus $\boldsymbol{\theta}^* = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^*$ beings the optimal solution of the common Lasso problem. When the condition (i.e., $|\boldsymbol{X}_i^T\boldsymbol{\theta}| < 1$) is satisfied, the ith variable will be removed. As in practice the optimal solution $\boldsymbol{\theta}^*$ is unknown in advance, the screening rule usually aims at finding a set $\boldsymbol{\Theta}$, such that $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\boldsymbol{X}_i^T\boldsymbol{\theta}| < 1$. Obviously, the smaller the $\boldsymbol{\Theta}$ is, the more efficient the screening rule is (Wang et al., 2015). The set $\boldsymbol{\Theta}$ derived from the SAFE rule is a sphere that is not very effective, and thus SAFE can only discard a limited number of inactive variables. Recent efforts have been made to improve the set $\boldsymbol{\Theta}$. For example, the Dual Polytope Projection (DPP) rule proposed by Wang et al., 2015 is designed based on the uniqueness and non-expansiveness of the optimal dual solutions, and thus the problem in the dual space becomes convex. Therefore, for a given value of λ, DPP removes $\boldsymbol{X}_i$ when the following condition is satisfied, $|\ \boldsymbol{X}_i^T\frac{\boldsymbol{y}}{\lambda_{max}}\ | < 1 - (\frac{1}{\lambda} - \frac{1}{\lambda_{max}})\|\boldsymbol{y}\|_2\|\boldsymbol{X}_i\|_2$. While the safe screening rules can guarantee that inactive sets only include non-relevant variables, they usually discard a moderate number of variables and the remaining can still be large for the final optimization problem, limiting their applications for the analysis of ultra-high dimensional data.

For penalized models, usually a full regularization path is considered and the optimal penalty parameter is selected based on some pre-defined criteria (e.g., BIC and cross validation errors) (Li et al., 2020; Wang and Wen, 2021; Wen and Lu, 2020). The standard versions for both safe and strong rules are designed for a given value of λ, and thus it can be computationally expensive to apply these rules to a series values of penalty parameters. As for penalized models, the size of active sets increases as the penalty parameter decreases. To utilize this property, sequential screening rules have been proposed, such as the sequential strong rules (SSR) (Tibshirani et al., 2012) and the enhanced DPP (EDPP) rule (Wang et al., 2015), where inactive sets are efficiently updated via a grid of decreasing tuning parameter values $\lambda_1, \lambda_2, ..., \lambda_K$. Specifically, SSR obtains the screening results at λ_{k+1} based on the solution of $\hat{\boldsymbol{\beta}}(\lambda_k)$ at λ_k, and the ith variable at λ_{k+1} is discarded if $|\boldsymbol{X}_i^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_k))| < 2\lambda_{k+1} - \lambda_k$. The EDPP rule

utilizes a similar idea but the screening condition is:

$$\left| \boldsymbol{X}_i^T \left(\frac{\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_k)}{\lambda_k} + \frac{1}{2}\boldsymbol{v}_2^{\perp}(\lambda_{k+1}, \lambda_k) \right) \right| < 1 - \frac{1}{2} \|\boldsymbol{X}_i\|_2 \| \boldsymbol{v}_2^{\perp}(\lambda_{k+1}, \lambda_k)\|_2$$

where $\boldsymbol{v}_2^{\perp}(\lambda_{k+1}, \lambda_k) = \boldsymbol{v}_2(\lambda_{k+1}, \lambda_k) - \frac{<\boldsymbol{v}_1(\lambda_k), \boldsymbol{v}_2(\lambda_{k+1}, \lambda_k)>}{\|\boldsymbol{v}_1(\lambda_k)\|_2^2}\boldsymbol{v}_1(\lambda_k)$,

$$\boldsymbol{v}_1(\lambda_k) = \begin{cases} \dfrac{\boldsymbol{y}}{\lambda_k} - \dfrac{\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_k)}{\lambda_k}, & 0 < \lambda_k < \lambda_{max}; \\ \text{sign}(\boldsymbol{X}_*^T\boldsymbol{y})\boldsymbol{X}_*, & \lambda_k = \lambda_{max}, \text{ and } \boldsymbol{X}_* = \text{argmax}_{\boldsymbol{X}_i} \mid \boldsymbol{X}_i^T\boldsymbol{y} \mid. \end{cases}$$

and

$$\boldsymbol{v}_2(\lambda_{k+1}, \lambda_k) = \frac{\boldsymbol{y}}{\lambda_{k+1}} - \frac{\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_k)}{\lambda_k}.$$

Although sequential strong and safe rules have improved the computational efficiencies, they still come with their own drawbacks. Specifically, strong-based rules can discard a large number of variables but there is no guarantee that the discarded variables are all inactive, while safe-based rules only screen out inactive variables but the remaining variables can still be large. Therefore, a hybrid sequential screening rule that can take advantages of both sequential strong and safe rules is needed for the analysis of ultra-high dimensional multi-omics data (Zeng et al., 2021).

Unlike single-layer omics data, multi-omics can provide a more comprehensive view of various diseases if integrated appropriately. For example, both genetic and epigenetic variations were found to alter the expression of oncogenes or tumor-suppressor genes in cervical cancer (Xu et al., 2019b). While multi-omics data offer great opportunities to systematically investigate a deep catalogue of biomarkers at different molecular levels, their intrinsic dependencies and their complex relationships with disease outcomes have brought tremendous analytical challenges. Many integrative methods have been proposed to tackle this issue (Bersanelli et al., 2016; Huang et al., 2017; Morris and Baladandayuthapani, 2017; Zeng and Lumley, 2018). For example, the non-negative matrix factorization method (Zhang et al., 2012) projects multi-omics data onto a common basis space to capture the coherent patterns among multi-omics data. Joint and Individual Variation Explained (JIVE) decomposes the original multi-omics data into three parts: joint variation, specific variation and noise. The extracted joint variation

from mRNA expression and miRNA expression data was used for cancer patient classification (Lock et al., 2013). Exploratory multivariate analysis tools have also been used for integrative analysis. For example, canonical correlation analysis constructs a set of linear combinations of all variables within each omics data and searches the optimal linear combination by maximizing the correlations between each canonical variate pair. Therefore, the most expressive elements of canonical vectors can be used to reflect the relationships among omics data (Meng et al., 2016). Similarly, partial least squares considers the covariance rather than correlation (Chen and Zhang, 2016). Bayesian methods that are more flexible in modeling various data types have gained popularities for integration in recent years (Imoto et al., 2004; Zhao et al., 2012). For example, the Patient Specific Data Fusion method applies a Bayesian non-parametric model (i.e., a two-level hierarchy Dirichlet Process model) to integrate copy number variation and expression data for discovering prognostic cancer subtypes (Yuan et al., 2011). Shen et al., 2009 proposed the iCluster method to integrate copy number and gene expression data for identifying subtypes of breast cancer and lung cancer. Network-based methods, another type of integration approach, can offer better interpretation of the model that can facilitate the understanding of the underlying mechanisms of complex diseases (Huang et al., 2017). For example, the similarity network fusion method proposed by Wang et al., 2014a builds patient similarity networks by integrating DNA methylation, mRNA and miRNA expression data to detect three glioblastoma multiforme subtypes. The smoothed t-statistic support vector machine (stSVM) method proposed by Cun and Fröhlich, 2013 smooths the gene-wise statistics (i.e., t-statistics) from miRNA data and gene expression data over a molecular network, which is integrated by the protein-protein interaction network and the miRNA-target gene network. A random walk kernel is used for smoothing and a permutation test is used to select significant genes that will be used to train a support vector machine classifier (Cun and Fröhlich, 2013). Recently, Bayesian network has been used for integration. Conexic proposed by Akavia et al., 2010 first integrates gene expression data and DNA copy number variations into modules in the form of regression trees, and then evaluates each module based on a Bayesian scoring function, where driver mutations in cancer can be detected by searching for the highest scoring module. The PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) (Vaske et al., 2010), another Bayesian network-based method, infers the activities of each biological pathway using a probability score calculated from factor graphs, where copy number variations, gene expression, DNA methylation and epigenetic data are integrated. In

recent years, deep learning methods, which have become the state-of-the-art method in many fields (e.g., the analysis of electronic health records and medical images), have also made their ways into the integrative analysis of multi-omics data (Chaudhary et al., 2018; Grapov et al., 2018; Kang et al., 2022). For example, Seal et al., 2020 used a deep learning model to integrate genomic, epigenomic and transcriptomic data and then extracted significant features from the integrative multi-omics data for disease classification. Xu et al., 2019a integrated gene expression, miRNA expression and DNA methylation data using a deep learning model for the detection of breast cancer, glioblastoma multiforme and ovarian cancer. While existing integrative methods have been widely used for multi-omics data analysis, they are mainly designed for discovering coherent patterns that can be used for understanding molecular mechanisms and/or disease classification. Therefore, they are not directly applicable for prediction studies, especially for continuous outcomes.

The rest of the thesis is arranged as follows. In chapter 2, we develop a penalized LMM with GMM estimators for the prediction analysis of genomic data, where we focused on improving the computational efficiency of penalized LMMs with a large number of random effects. In chapter 3, we develop a hybrid screening rule for penalized LMMs with GMM estimators, where the hybrid screening rule is designed to reduce the number of parameters entered into the penalized LMMs. In chapter 4, we extend the penalized LMM with GMM estimators to multi-omics data analysis, where we mainly focused on efficiently modeling complex predictive effects from multi-omics data within our proposed LMM framework. Finally, we present the summary and future work in the last chapter.

Chapter 2

A penalized linear mixed model with generalized method of moments estimators for complex phenotype prediction using genomic data

2.1 Introduction

Accurate disease risk prediction plays an important role in precision medicine, an emerging model of healthcare that tailors treatment strategies based on individuals' profiles (Ashley, 2015). Over the past decades, genome-wide association and whole-genome sequencing studies have detected many disease-associated genetic variants. While it is hoped that these genetic findings can facilitate the ongoing risk prediction studies (Speed and Balding, 2014; Weissbrod et al., 2016; Wen and Lu, 2020), existing genetic risk prediction models can only explain a small proportion of the heritability. The complex relationships between predictors and phenotypes (Speed and Balding, 2014; Weissbrod et al., 2016), the huge amount of noise in high-dimensional genetic data (Byrnes et al., 2013), and the high computational cost (Weissbrod et al., 2016; Wen and Lu, 2020) greatly limit the prediction accuracy of existing models.

Linear mixed models (LMMs) and their extensions are the most widely used methods for risk prediction studies (Speed and Balding, 2014; Weissbrod et al., 2016; Wen and Lu, 2020). The genomic best linear unbiased prediction (gBLUP) method, which was first introduced by Harris et al., 2008 to predict milk production and then extended for the prediction of human traits (Yang et al., 2010), is one of the earliest methods within the LMM framework. gBLUP assumes that each genetic variant acts in an

additive manner and their effect sizes follow the same normal distribution. gBLUP is equivalent to a LMM with one random effect term, in which the variance-covariance structure encodes the assumed linear additive relationships. gBLUP only needs to estimate one parameter associated with the random effect term, making it computationally efficient. While easy to implement, the modeling assumptions in gBLUP are too simple, and thus efforts have been made to relax them. For example, Speed and Balding, 2014 have shown that genetic variants from different regions (e.g., eQTLs, intron SNPs and coding) can have different effect sizes, and thus extended the gBLUP to MultiBLUP, where the genome is split into multiple regions (e.g., based on gene or pathway annotations) with each being modeled by a random effect term with its own variance parameter. Converging evidence suggests that epistasis widely exists (Buil et al., 2015; Moore and Williams, 2009). Thus, Weissbrod et al., 2016 generalized the MultiBLUP further by embedding cumulative predictive effects from each region into the reproducing kernel Hilbert space, where a pre-specified kernel function is used to construct variance-covariance matrices for each region, making it capable of capturing non-linear predictive effects. Recently, Wen and Lu, 2020 incorporated the multi-kernel learnings into the LMMs, where complex predictive effects can be efficiently captured. While the advances in LMM-based methods offer greater flexibility in modeling complex diseases, their levels of successes are largely limited, mainly due to the high computational cost when a large number of random effects are used to accommodate different types of predictive effects.

While high-dimensional genomic data allow for thorough investigations of disease etiologies, they also contain a huge amount of noise. Within the LMM framework, using all genetic regions, including those that only harbour noise variants, to build risk prediction models not only substantially increases the computational complexity but also reduces the robustness and accuracy of the models. Byrnes et al., 2013 have already demonstrated that in the absence of good biological annotations, variable selection is an efficient way to improve prediction accuracy, especially for high-dimensional data. Conventional variable selection methods (e.g., BIC, GIC, and forward selection) as well as those empirical criteria employed in LMMs can all be used to select predictive regions (Speed and Balding, 2014; Weissbrod et al., 2016). However, there is no theoretical guarantee for their optimal performance. L_1 penalization is a common technique used for simultaneously selecting predictive variables and estimating their effect sizes (Ghosh and Chinnaiyan, 2005; Ma et al., 2007; Sun and Wang, 2012; Wu et al., 2009). Recently, L_1 penalization has been introduced into LMM-based risk prediction models, where

penalties are imposed on the random effect terms to allow for consistent and efficient selection of predictive regions (i.e., random effects) (Li et al., 2020; Wen and Lu, 2020). While these advances can reduce the impact of noise and improve prediction accuracy, their parameter estimations can be computationally demanding. Existing algorithms usually employ the one-step approximation procedure (Fan and Li, 2001; Zou and Li, 2008) and their performance depends on the initial values, which are usually set to be either the maximum likelihood estimators (MLE) or the restricted maximum likelihood estimators (REML). However, both MLE and REML can themselves be hard to obtain for LMMs when a large number of random effects is being considered. As a consequence, existing penalized LMMs can only handle a few regions with limited types of predictive effects, leading to sub-optimal prediction performance.

The parameters in LMM-based models are usually estimated with either MLE or REML (Speed and Balding, 2014; VanRaden, 2008; Yang et al., 2010), where Newton-Raphson or expectation-maximization algorithms are commonly used to optimize the objective functions. Although both MLE and REML are statistically efficient, their estimation procedure involves repeatedly inversing the variance-covariance matrix, making it computationally prohibitive to consider a large number of random effects. Recently, simulated annealing algorithms have also been introduced to optimize the objective function of LMMs (Weissbrod et al., 2016). However, the performance of simulated annealing algorithms is determined by empirical criteria, and thus the algorithm will be very likely to find a local optimal or take a very long time to find a global optimal. The generalized method of moments (GMM) is a long-existing alternative to REML/MLE for LMMs (Rao, 1970; Rao, 1971a; Rao, 1972), where statistical efficiency is traded with computational efficiency. GMM is a promising alternative for penalized LMMs with multiple random effects because it can change the objective function into a quadratic form, which is much easier to optimize. Indeed, GMM estimators have been used in LMMs for the estimation of variance components. For example, Zhu and Weir, 1996 used the minimum norm quadratic unbiased estimation (MINQUE) method to estimate variance components for maternal and paternal effects in a bio-model for diallel crosses. Zhou, 2017 unified the method of moments, the MINQUE criterion, the Haseman-Elston regression and the LD score regression to estimate the heritability of height using the Australian GWAS data. Pazokitoroudi et al., 2019 presented a randomized multi-component version of the classical Haseman-Elston regression, which uses method-of-moments to estimate heritability of 22 complex traits using the UK Biobank data. While GMMs have demonstrated their computational efficiencies in

variance component estimation, they have not been used for parameter estimations in penalized LMMs with multiple random effects, and their theoretical properties are rarely studied.

To address these issues, we developed a penalized LMM with GMM estimators (referred to as pLMMGMM) to simultaneously select predictors and estimate their effect sizes in the prediction analysis. Similar to existing LMMs, the proposed method splits the genome into multiple regions and models the cumulative predictive effects for each region via random effect terms. Fundamentally different from existing LMMs that rely on MLE or REML, our method estimates its parameters using GMM, making it much more computationally efficient. Therefore, our method can: 1) use a data-driven approach to choose appropriate kernel functions to reflect different types of relationships between predictors and outcomes, and 2) simultaneously and efficiently model a large number of regions (i.e., random effects) and detect those that are predictive. In the following sections, we will first present the pLMMGMM model and its theoretical properties, in section 2. In section 3, we conduct the extensive simulation studies to evaluate the model's empirical performance, and further compare its prediction accuracy with commonly used methods. In section 4, we illustrate the practical utility of our method by analyzing a data set obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) study (Saykin et al., 2010).

2.2 Methods

LMMs have long been used for risk prediction analysis on high-dimensional genomic data (De Los Campos et al., 2013; Speed and Balding, 2014; Weissbrod et al., 2016). For completeness, we first present the LMMs used for prediction research, and then introduce our penalized LMM where the parameters are estimated using GMM. Finally, we show the theoretical properties of our proposed GMM-based estimators.

2.2.1 A linear mixed model for risk prediction using genomic data

The fundamental assumption in LMMs is that genetically similar individuals can have similar phenotypes. As genetic variants located at different locations (e.g., eQTLs, intron SNPs and coding) can have different effect sizes (Speed and Balding, 2014), we first split the genome into R regions based on some criteria (e.g., gene annotation and

pathway), and model the outcomes $\boldsymbol{Y}$ as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{R} \boldsymbol{g}_i + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim N(0, \sigma_0^2 \boldsymbol{I}_n) \quad \boldsymbol{g}_i \sim N(0, \sigma_i^2 \boldsymbol{K}_i) \tag{2.1}$$

where $\boldsymbol{X}$ is an $n \times P$ matrix of the demographic variables (e.g., age and gender) and $\boldsymbol{\beta}$ is their effect sizes; $\boldsymbol{g}_i$ is the cumulative predictive effect from the ith region; and $\boldsymbol{K}_i$ is a $n \times n$ kernel matrix measuring the genetic similarities of region i.

Similar to existing LMMs, the variance-covariance matrix for each cumulative effect implicitly determines the assumed relationship between predictors and the outcome. For example, if a linear kernel is used for each region (i.e., $\boldsymbol{K}_i = \boldsymbol{G}_i \boldsymbol{G}_i^T$ with $\boldsymbol{G}_i$ is an $n \times p_i$ genotype matrix for region i), the proposed model in equation 2.1 is equivalent to:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{R} \boldsymbol{G}_i \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma_0^2 \boldsymbol{I}_n)$$

where $\boldsymbol{\gamma}_i \sim N(0, \sigma_i^2 \boldsymbol{I}_{p_i})$. Therefore, a linear kernel implicitly assumes that there is a linear additive relationships between predictors and the outcome. To accommodate more complex relationships, we extended the single kernel into multiple kernels, where multiple kernel matrices are combined in a data-driven manner. For example, when both linear and pair-wise interaction effects are considered, we set the variance-covariance matrix to be $\sigma_i^2 \boldsymbol{K}_i = \sigma_{i1}^2 \boldsymbol{K}_{i1} + \sigma_{i2}^2 \boldsymbol{K}_{i2}$, where $\boldsymbol{K}_{i1}$ is a linear kernel designed to capture additive effects and $\boldsymbol{K}_{i2}$ is the polynomial kernel with 2 degrees of freedom designed to model the pairwise interaction effects. More kernels (e.g., the RBF kernel) can be added to the candidate kernel set to efficiently model various types of predictive effects. By using random effects to capture cumulative predictive effects and kernelizing the covariance matrices, our proposed method not only reduces the model parameters but also offers a very flexible framework for modeling traits with various underlying genetic architecture (Li et al., 2020; Wen and Lu, 2020).

2.2.2 A penalized linear mixed model with generalized method of moments estimators using genomic data

Converging evidence has shown that not all genetic variants and regions are predictive (Li et al., 2020; Speed and Balding, 2014; Weissbrod et al., 2016; Wen et al., 2016; Wen and Lu, 2020). Including noise can reduce the robustness and accuracy of the prediction model. For our proposed model, if region i is not predictive, then there

are no variations for its cumulative predictive effect $\boldsymbol{g}_i$ and thus $\sigma_i^2 = 0$. Therefore, selecting predictive regions is equivalent to determining which σ_i^2 are not zero.

L_1 penalty is a commonly used technique for simultaneously selecting predictors and estimating their effect sizes. For example, Wen and Lu, 2020 added an L_1 penalty to the log likelihood function of LMMs to simultaneously select predictive regions and estimate their effect sizes:

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \frac{1}{2} log|\boldsymbol{\Sigma}| + \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{P+R+1} \omega_i(|\phi_i|)$$

where $\boldsymbol{\Sigma} = \sigma_0^2 \boldsymbol{I}_n + \sum_{i=1}^{R} \sigma_i^2 \boldsymbol{K}_i$; $\boldsymbol{\sigma}^2 = (\sigma_0^2, \sigma_1^2, \cdots, \sigma_R^2)$; and $\boldsymbol{\phi} = (\boldsymbol{\sigma}^2, \boldsymbol{\beta})$. While capable of detecting predictive regions, this method can only consider a limited number of genetic regions in practice, mainly due to their high computational cost.

To simultaneously model predictive effects from a large number of genetic regions (i.e., random effects in LMMs), we propose to use the GMM to select predictive regions and estimate their effect sizes. Clearly, the variance of $\boldsymbol{Y}$ depends on covariates $\boldsymbol{X}$, and thus we propose to follow the same procedure developed by Pazokitoroudi et al., 2019 to choose an $\boldsymbol{A}$ matrix, such that the variance of $\boldsymbol{A}^T\boldsymbol{Y}$ is independent of the covariates $\boldsymbol{X}$. Let $\boldsymbol{V} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is a symmetric and idempotent of rank $n - P$ matrix. We consider the eigen decomposition of $\boldsymbol{V} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^T$, where $\boldsymbol{D}$ is a diagonal matrix with $n - P$ ones and P zeros on the diagonal. We set the $\boldsymbol{A}$ matrix as the first $n - P$ columns of $\boldsymbol{E}$. Therefore, $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{V}$, $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{A}^T\boldsymbol{X} = 0$, and

$$\text{var}(\boldsymbol{A}^T\boldsymbol{Y}) = \boldsymbol{A}^T \sum_{i=1}^{R} \sigma_i^2 \boldsymbol{K}_i \boldsymbol{A} + \sigma_0^2 \boldsymbol{I}_{n-P} \tag{2.2}$$

Traditionally, the parameters in equation 2.2 are estimated using REML estimators that can be computationally infeasible when the number of regions is large. To overcome the computational bottleneck, we propose to trade statistical efficiency with computational efficiency, and develop a penalized GMM estimator for its parameter estimation:

$$\hat{\boldsymbol{\sigma}}^2 = \underset{\boldsymbol{\sigma}^2 \geq 0}{\text{argmin}} \frac{1}{2} ||\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A} - \boldsymbol{A}^T \sum_{i=1}^{R} \sigma_i^2 \boldsymbol{K}_i \boldsymbol{A} - \sigma_0^2 \boldsymbol{I}_{n-P}||_F^2 + \lambda \sum_{i=1}^{R} \sigma_i^2, \quad \lambda > 0 \tag{2.3}$$

By using the proposed penalized GMM estimators, the objective function is in a quadratic form that is much easier to optimize than traditional REML or MLE

estimators. It is straightforward to see that the parameters can be easily obtained by solving equation 2.4 (the details are shown in appendix A.1.1 of the Supplementary Materials):

$$\hat{\boldsymbol{\sigma}}^2 = \underset{\boldsymbol{\sigma}^2 \geq 0}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^2||_F^2 + \lambda \sum_{i=1}^{R} \sigma_i^2, \quad \lambda > 0 \tag{2.4}$$

where $\boldsymbol{M} = vec(\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A})$; $\boldsymbol{T}_i = vec(\boldsymbol{A}^T\boldsymbol{K}_i\boldsymbol{A})$; $\boldsymbol{T}_i$ is the ith column of $\boldsymbol{T}$ matrix; and thus $\boldsymbol{T} = (\boldsymbol{T}_0, \boldsymbol{T}_1, ..., \boldsymbol{T}_R)$. $vec(.)$ is the vectorization of a matrix. Equation 2.4 can be solved by the coordinate descent algorithm implemented in `glmnet` R package (Friedman et al., 2010). The optimal tuning parameter λ is chosen by cross validation.

Let $\boldsymbol{Y}_a = (\boldsymbol{Y}_p, \boldsymbol{Y})$, where $\boldsymbol{Y}$ is the $n \times 1$ vector of outcomes in the training data and $\boldsymbol{Y}_p$ is $n_p \times 1$ vector of outcomes to be predicted. Given the parameter estimates for $\hat{\boldsymbol{\sigma}}^2$ and $\hat{\boldsymbol{\beta}}$, the variance of $\boldsymbol{Y}_a$ can be directly derived as $\hat{\boldsymbol{\Sigma}}_{Y_a} = \sum \boldsymbol{K}_i^a \hat{\sigma}_i^2 + \boldsymbol{I}_{n+n_p}\hat{\sigma}_0^2$, where $\boldsymbol{K}_i^a$ is the $(n_p+n) \times (n_p+n)$ genetic similarity matrix calculated from all samples. The variance of $\boldsymbol{Y}_a$ can be written as:

$$\hat{\boldsymbol{\Sigma}}_{Y_a} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{pp} & \hat{\boldsymbol{\Sigma}}_{po} \\ \hat{\boldsymbol{\Sigma}}_{op} & \hat{\boldsymbol{\Sigma}}_{oo} \end{bmatrix}$$

where $\hat{\boldsymbol{\Sigma}}_{pp}$ and $\hat{\boldsymbol{\Sigma}}_{oo}$ are the variance matrices for the testing and training samples, respectively; and $\hat{\boldsymbol{\Sigma}}_{po}$ is the covariance matrix between testing and training samples. Therefore, the predictive values for the testing samples can be calculated as:

$$\boldsymbol{Y}_p = \boldsymbol{X}_p\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Sigma}}_{po}\hat{\boldsymbol{\Sigma}}_{oo}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_o\hat{\boldsymbol{\beta}})$$

where $\boldsymbol{X}_p$ and $\boldsymbol{X}_o$ are the covariates of the testing samples and training samples, respectively.

2.2.3 Theoretical properties

We investigated the theoretical properties of our proposed method, including the selection consistency, estimation consistency and asymptotic normality. We will explore whether our model can choose the right predictive variables, and establish the asymptotic distribution of our estimators.

Notations and assumptions

Let $S_1 = \{i \in \{0, 1, 2, ..., R\} : \sigma_i^2 \neq 0\}$ denote the set of all predictive regions, and $S_0 = \{i \in \{0, 1, 2, ..., R\} : \sigma_i^2 = 0\}$ be the set of regions that just harbour noise variants. Let $S(\lambda) = \{i \in \{0, 1, 2, ..., R\} : \hat{\sigma}_i^2(\lambda) \neq 0\}$ be the estimated set of predictive regions for a given value of λ that is selected independent of predictors and outcomes. For simplicity and without loss of generality, we assume the first q regions are predictive and the remaining $R - q$ regions are noise. Let $\boldsymbol{T}(1) = (\boldsymbol{T}_0, \boldsymbol{T}_1, ..., \boldsymbol{T}_q)$ and $\boldsymbol{T}(2) = (\boldsymbol{T}_{q+1}, \boldsymbol{T}_{q+2}, ..., \boldsymbol{T}_R)$. Let $\boldsymbol{C} = \frac{1}{N}\boldsymbol{T}^T\boldsymbol{T}$, where N is the number of rows in $\boldsymbol{T}$ matrix. Therefore, $\boldsymbol{C}$ can be written as:

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix}$$

where $\boldsymbol{C}_{11} = \frac{1}{N}\boldsymbol{T}(1)^T\boldsymbol{T}(1)$; $\boldsymbol{C}_{12} = \frac{1}{N}\boldsymbol{T}(1)^T\boldsymbol{T}(2)$; $\boldsymbol{C}_{21} = \frac{1}{N}\boldsymbol{T}(2)^T\boldsymbol{T}(1)$; and $\boldsymbol{C}_{22} = \frac{1}{N}\boldsymbol{T}(2)^T\boldsymbol{T}(2)$.

Similar to Wu et al., 2014, we assumed the following conditions for our method:

Assumption 1. *There exist constants* $0 \leq c_1 < c_2 \leq 1$, $0 \leq c_3 < c_2 - c_1$, *and* $M_1, M_2, M_3 > 0$, *such that the following conditions hold:*

$$\begin{cases} \frac{\boldsymbol{T}_i{}^T\boldsymbol{T}_i}{N} \leq M_1, \quad \forall i \\ \alpha^T C_{11}\alpha \geq M_2, \quad \forall ||\alpha||_2^2 = 1 \\ q + 1 = O(N^{c_1}) \\ R = O(e^{N^{c_3}}) \\ N^{\frac{1-c_2}{2}} \min_{i=0,1,2,...,q} \mid \sigma_i^2 \mid \geq M_3 \\ \frac{1}{N} \max_{1 \leq i \leq N} \boldsymbol{T}_i{}^T\boldsymbol{T}_i \to 0, \quad as \quad N \to \infty \end{cases}$$

Assumption 2. *Nonnegative irrepresentable condition: there exists a positive constant vector* $\boldsymbol{\rho}$, *such that:*

$$\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\mathbf{1} \leq \mathbf{1} - \boldsymbol{\rho}$$

Assumption 3. *Restricted eigenvalue condition: there exists constant* k_m, *such that:*

$$\begin{cases} \frac{||\boldsymbol{T}\boldsymbol{\sigma}^2||_2^2}{N} \geq k_m ||\boldsymbol{\sigma}^2||_2^2 \\ \sum_{i \in S_0} |\sigma_i^2| \leq 3 \sum_{i \in S_1} |\sigma_i^2| \end{cases}$$

Assumption 4. *Column normalization condition:*

$$\frac{||\boldsymbol{T}_i||_2}{\sqrt{N}} \leq 1, \quad i = 0, 1, 2, ..., R$$

Theoretical properties

Theorem 2.1 (Variable selection consistency)**.** *Under assumptions 1 and 2, the pLMMGMM method can have variable selection consistency. In particular, when* $\lambda \propto N^{\frac{1+c^4}{2}}$ *where* $c_3 < c_4 < c_2 - c_1$*, the following condition holds:*

$$P(S(\lambda) = S_1) \geq 1 - o(e^{-N^{c_3}}) \to 1, \quad as \quad N \to \infty$$

Theorem 2.1 demonstrates that the proposed pLMMGMM can consistently select the predictive regions even when the number of regions (i.e., R) increases faster than the sample size (N) at an exponential speed. The proof for theorem 2.1 can be seen in appendix A.1.2 of the Supplementary Materials.

Theorem 2.2 (Variable estimation consistency)**.** *Under assumptions 3 and 4, the pLMMGMM method can have variable estimation consistency. In particular, when the regularization parameter* $\lambda = 4\sigma_\omega^2\sqrt{\frac{\log R}{N}}$*, there exists constants* $v_1, v_2 > 0$ *such that, with probability at least* $1 - v_1 \exp(-v_2 N\lambda^2)$*:*

$$||\hat{\sigma}^2 - \sigma^2||_2^2 \to 0, \quad as \quad N \to \infty$$

Theorem 2.2 indicates that the penalized GMM estimators have variable estimation consistency. The proof for theorem 2.2 can be seen in appendix A.1.3 of the Supplementary Materials.

Theorem 2.3 (Asymptotic normality)**.** *Under the same setting in theorem 2.1, the nonzero estimator* $\hat{\boldsymbol{\sigma}}^2(1)$ *is asymptotically normal:*

$$\sqrt{N}(\hat{\boldsymbol{\sigma}}^2(1) - \boldsymbol{\sigma}^2(1)) \sim N(0, \sigma_\omega^2 \boldsymbol{C}_{11}^{-1})$$

Theorem 2.3 indicates that the penalized GMM estimators for those nonzero parameters are asymptotically normally distributed. The proof for theorem 2.3 can be seen in appendix A.1.4 of the Supplementary Materials.

2.3 Simulation studies

We investigated the performance of the proposed pLMMGMM through extensive simulation studies, where the impacts of noise and the underlying disease models were evaluated. We considered sample sizes of 500 and 1000. For each setting, we randomly chose 70% of the samples to train the model, and used the remaining samples to assess the model's prediction accuracy, which was measured by both Pearson correlations and mean square errors (MSE). We further compared our method with two widely used methods, including gBLUP with its default settings (Yang et al., 2010) and MKLMM (Weissbrod et al., 2016). Note that by default, MKLMM only includes a pre-determined number of regions which are selected based on the rank of Likelihood Ratio (LR) for each region. Therefore, for a fair comparison, we compared our method with MKLMM under two settings, where the screening step is included (i.e., the default of MKLMM where the top 5% regions based on the likelihood ratio test are chosen) or omitted (i.e., all regions are considered jointly). We denoted these two settings as MKLMM and MKLMMpre, respectively. We did not compare our method with MultiBLUP, mainly because MultiBLUP is equivalent to MKLMM with a linear kernel. To evaluate whether the proposed method can select predictive regions, we calculated the sensitivity and specificity for our method. For all simulations, to mimic the real human genome, we directly obtained genotypes from the 1000 Genome Project (The 1000 Genomes Project Consortium, 2015), and constructed each region with 30 randomly selected SNPs that are within 75Kb.

2.3.1 Scenario I: the impact of the number of noise regions

In this set of simulations, we gradually increased the number of noise regions to evaluate their impact. In particular, we randomly set two regions as causal, and simulated the outcomes under an additive model:

$$Y_i = \sum_j G_{ij}^1 \beta_{ij}^1 + \sum_j G_{ij}^2 \beta_{ij}^2 + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_0^2)$; $G_{ij}^k, k \in \{1, 2\}$ represents the jth SNPs on the kth causal region for individual i; and $\beta_{ij}^k \sim N(0, \sigma_k^2)$ is their corresponding effect. It is straightforward to show that:

$$\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{K}_1\sigma_1^2 + \boldsymbol{K}_2\sigma_2^2 + \boldsymbol{I}_n\sigma_0^2) \tag{2.5}$$

where $\boldsymbol{K}_k = \boldsymbol{G}^k \boldsymbol{G}^{k^T}$ and $\boldsymbol{G}^k$ is the genotype matrix for region k. Therefore, as shown in equation 2.5, we simulated the outcomes based on a multivariate normal distribution.

We gradually increased the number of noise regions from 0 to 98 (i.e., the total number of regions ranges from 2 to 100). For each setting, we conducted 1000 Monte Carlo simulations, and calculated the prediction accuracy based on testing samples. We reported the Pearson correlation, MSE and the computational cost for each method. We further calculated the probabilities of correctly detecting causal and noise regions for our method.

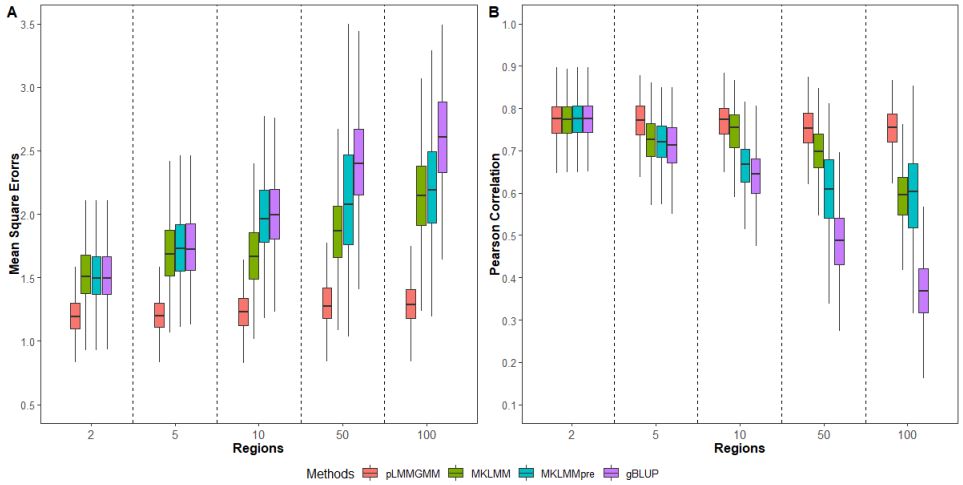

FIGURE 2.1: The impact of the number of noise regions on Pearson correlations and MSEs ($n = 500$)

The Pearson correlations and MSEs for sample sizes of 500 and 1000 are shown in Figure 2.1 and Supplementary Figure A.1, respectively. Among all the scenarios considered, pLMMGMM performs the best. When there are no noise regions, our proposed method has similar levels of Pearson correlations as those of gBLUP and MKLMM, but its MSEs tend to be smaller. As the number of noise regions increases, the prediction accuracy of the proposed method remains roughly stable whereas the performance of the other methods decreases to some extent, with gBLUP being affected the most and MKLMM the least. gBLUP assumes that all genetic variants act in an additive manner and their effect sizes follow the same normal distribution. Therefore, as the number of noise regions increases, the assumption of gBLUP is severely violated, and thus its performance dropped the most. In contrast, MKLMM allows genetic variants from different regions to have different effects, and thus its variance component estimates in LMM can differ substantially for variants located on different regions. Therefore, MKLMM has some capacities to deal with noise regions and so tends to perform better

than gBLUP. Comparing MKLMM under the two settings, the default setting has better performance than the setting where the screening step is omitted. The employed pre-screening procedure can limit the number of regions, and thus reduces the number of random effects in LMM, which improves the robustness and computational efficiency of MKLMM. However, the screening step implemented in MKLMM considers one region at a time and relies on some empirical criteria for region pre-selection, both of which may lead to a sub-optimal prediction model.

TABLE 2.1: The chances of selecting two predictive regions as the number of noise regions increases ($n = 500$)

Regions	Sensitivity	Specificity
5	1.000	0.905
10	0.999	0.884
50	1.000	0.909
100	0.999	0.932

Comparing pLMMGMM with both gBLUP and MKLMM, our method can jointly consider a large number of regions and select those that are predictive. Therefore, its performance is relatively robust against noise. Indeed, excluding the noise regions not only improves prediction accuracy but also improves the robustness of the model. With regards to variable selection, our pLMMGMM method not only correctly detects causal regions but also identifies those regions that are noise. The results for variable selection under sample sizes of 500 and 1000 are summarized in Table 2.1 and Supplementary Table A.1, respectively.

One of the benefits of using GMM is improved computational efficiencies, and thus we compared the running time of our method and MKLMMpre, where the same LMM model was fitted using the REML estimator. While the Newton-Raphson algorithm is the most widely used method for optimizing the objective function of REML estimators, it can barely converge when the number of random effects is large. Indeed, the convergence rates for 2, 5, 10, 50 and 100 regions (i.e., random effects) are 100%, 15.5%, 7.9%, 0% and 0%, respectively. REML estimated using the simulated annealing algorithm can converge. However, there is no guarantee that the simulated anneal algorithm will achieve the global optimum. As shown in Figure 2.1, MKLMMpre has lower prediction accuracy than our proposed method. In addition, the computational time of REML grows much faster than the GMM-based estimators as the number of random effects increases, regardless of the sample sizes considered (Figure 2.2 and Supplementary Figure A.2).

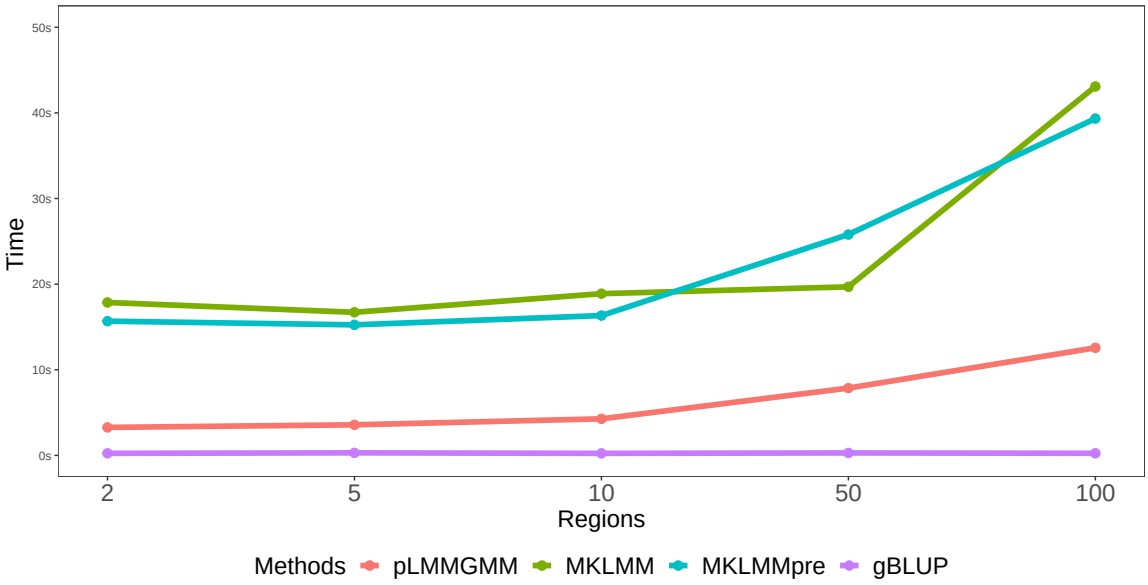

FIGURE 2.2: The impact of the number of noise regions on computational time ($n = 500$)

2.3.2 Scenario II: the impact of disease models

Complex traits and diseases are affected by a large number of genes through complicated biological pathways that are usually unknown in advance (Chatterjee et al., 2013). It has long been recognized that a risk prediction model with flexible modeling assumptions is more robust and accurate across a range of phenotypes with different genetic architectures. In this set of simulations, we evaluated the performance of our proposed method given different disease models. As in Scenario I, we considered two causal genes and simulated the outcomes using equation 2.5, where the kernel matrices $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ were used to reflect different disease models. Specifically, we considered five disease models: 1)$L + L$: genetic variants on both regions have linear additive effects (i.e., $k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$); 2) $R + R$: predictors from both regions have non-linear predictive effects (i.e., $k_{rbf}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\frac{1}{2}||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2^2\right]$); 3) $P + P$: both regions harbour variants with pair-wise interaction effects (i.e., $k_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = (< \boldsymbol{x}_1, \boldsymbol{x}_2 >)^2$); 4) $L + R$: genetic variants on the first and second regions have linear additive effects and non-linear effects, respectively; and 5) $L+P$: predictors on the first and second regions have linear additive and pair-wise interaction effects, respectively. The details for each model setting and the corresponding kernels are summarized in Table 2.2. In addition to these two causal regions, we also simulated 48 noise regions (i.e, the total number of regions were 50) for this set of simulations.

TABLE 2.2: Disease models description

Disease Models	Description	$\boldsymbol{K}_1$	$\boldsymbol{K}_2$
$S_1 : L + L$	Linear additive effects	$k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$	$k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$
$S_2 : R + R$	Non-linear effects.	$k_{rbf}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\frac{1}{2}\|\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\|_2^2\right]$	$k_{rbf}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\frac{1}{2}\|\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\|_2^2\right]$
$S_3 : P + P$	Pair-wise interaction effects	$k_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = (< \boldsymbol{x}_1, \boldsymbol{x}_2 >)^2$	$k_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = (< \boldsymbol{x}_1, \boldsymbol{x}_2 >)^2$
$S_4 : L + R$	Linear and non-linear effects	$k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$	$k_{rbf}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\frac{1}{2}\|\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\|_2^2\right]$
$S_5 : L + P$	Linear and pair-wise interaction	$k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$	$k_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = (< \boldsymbol{x}_1, \boldsymbol{x}_2 >)^2$

For MKLMM, unlike the first simulation where only linear kernel was used, we used the default setting where the most appropriate kernels are selected in a data-driven manner. Note that since MKLMMpre performs worse than MKLMM that adopts an empirical screening rule (Figure 2.1 and Supplementary Figure A.1), we only compared our method to MKLMM with screening implemented. Similar to Scenario I, we conducted 1000 Monte Carlo simulations for each setting, and summarized the prediction accuracy for all the methods considered. For our proposed method, we also calculated the probabilities of detecting causal and non-causal regions.

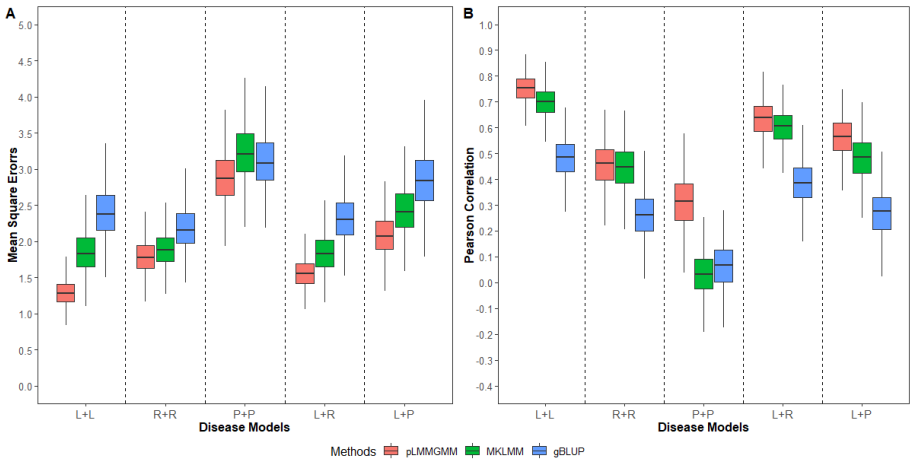

FIGURE 2.3: The impact of disease models. $L + L$: genetic variants on both regions have linear additive effects. $R + R$: predictors from both regions have non-linear predictive effects. $P + P$: both regions harbor variants with pair-wise interaction effects. $L + R$: genetic variants on the first and second regions have linear additive and non-linear effects, respectively. $L+P$: predictors on the first and second regions have linear additive and pair-wise interaction effects, respectively ($n = 500$)

As shown in Figure 2.3 and Supplementary Figure A.3, the proposed pLMMGMM has the lowest MSEs and highest Pearson correlations of all the methods considered, which indicates that our method is quite robust against different disease models. This is mainly because our method not only selects predictive genetic regions but also accounts

for non-linear effects through selecting appropriate kernel functions from the candidate kernel set.

While MKLMM is designed to capture non-linear effects, in practice it can barely capture them. Indeed, for most of the non-linear models that we considered, the chance of selecting only linear kernels by the adaptive MKLMM is extremely high ($> 99\%$). For example, for the disease model that has pair-wise interaction effects (i.e., $P+P$) on both causal regions, the chance of selecting the ideal polynomial kernel for MKLMM is close to 0%, whereas our proposed method had an average of 76% chance of choosing the appropriate polynomial kernel. As a consequence, the adaptive MKLMM performs in a way that is very similar to a MKLMM that only considers the linear kernel, and can perform worse than the MKLMM with the most appropriate kernel employed. See Supplementary Figures A.4 and A.5, where the most appropriate kernels are used for MKLMM (i.e., the kernels are pre-determined based on the underlying disease etiology).

TABLE 2.3: The chances of selecting two predictive regions under different disease models ($n = 500$)

Disease Models	Sensitivity	Specificity
$S_1 : L + L$	0.999	0.919
$S_2 : R + R$	0.969	0.951
$S_3 : P + P$	0.794	0.980
$S_4 : L + R$	0.959	0.931
$S_5 : L + P$	0.917	0.949

One of the key features of MKLMM is that it screens the genome and selects regions that are predictive to build the risk prediction model. However, its ability to detect causal regions also depends on the underlying disease model. For example, for the $P+P$ disease model, the chance of adaptive MKLMM selecting any of those two causal regions is very low (i.e., only noise regions are used for prediction), leading to a prediction model that performs even worse than gBLUP. For other disease models, although adaptive MKLMM is unlikely to efficiently capture those non-linear effects (i.e., it is not able to choose the most appropriate kernel), the adaptive MKLMM has a relatively high chance of detecting the causal regions and thus reducing the impact of noise. As a result, the adaptive MKLMM can outperform gBLUP under these settings, as gBLUP utilizes all regions and ignores the impact of noise. To evaluate whether our method can detect causal regions under different underlying disease models, we calculated the sensitivity and specificity of our method, and the results are summarized in Table 2.3

and Supplementary Table A.2. On average, our method achieves a sensitivity of 93% and a specificity of 95% among all the models considered with samples of 500, and a sensitivity of 99% and a specificity of 93% among all the models considered with samples of 1000.

2.4 Real data application

We used our proposed pLMMGMM method to predict positron emission tomography (PET) imaging outcomes, including FDG and AV45, using the whole-genome sequencing data obtained from the ADNI. ADNI is a longitudinal study designed for the prevention and treatment of Alzheimer's disease (AD) (Mueller et al., 2005). It measures clinical, imaging, genetic and biochemical biomarkers from each participant to investigate the pathology of AD. DNA samples from 818 participants aged between 55 and 90 were collected and sequenced on the Illumina HiSeq2000 at a non-Clinical Laboratory Improvements Amendments (non-CLIA) laboratory (Saykin et al., 2015). Genetic variants with missing rate large than 1% were first excluded, and then the remaining variants were annotated based on GRch37 assembly. We selected 95 AD-related genes based on existing literature (details are listed in Supplementary Table A.3, and a total of 117,668 variants were included in the final analyses.

For our analyses, we are interested in predicting PET-imaging outcomes, including FDG and AV45, using the whole-genome sequencing data. We removed individuals who are either correlated or have missing outcomes, and the distributions of FDG and AV45 for the remaining samples ($n = 639$ for FDG and $n = 501$ for AV45) are shown in Supplementary Figure A.6. To evaluate the prediction accuracy, we randomly split the samples into testing ($n = 100$) and training sets, where the training samples were used to train the prediction model and the remaining samples were used to calculate the Pearson correlations and MSEs. To reduce the risk of chance finding, we repeated this process 100 times.

The prediction accuracies for FDG and AV45 are shown in Figure 2.4. For both FDG and AV45, pLMMGMM has lower MSEs and higher Pearson correlations than both gBLUP and MKLMM. This indicates that simultaneously considering multiple genes and excluding those that are not predictive can improve the robustness and accuracy of the risk prediction model.

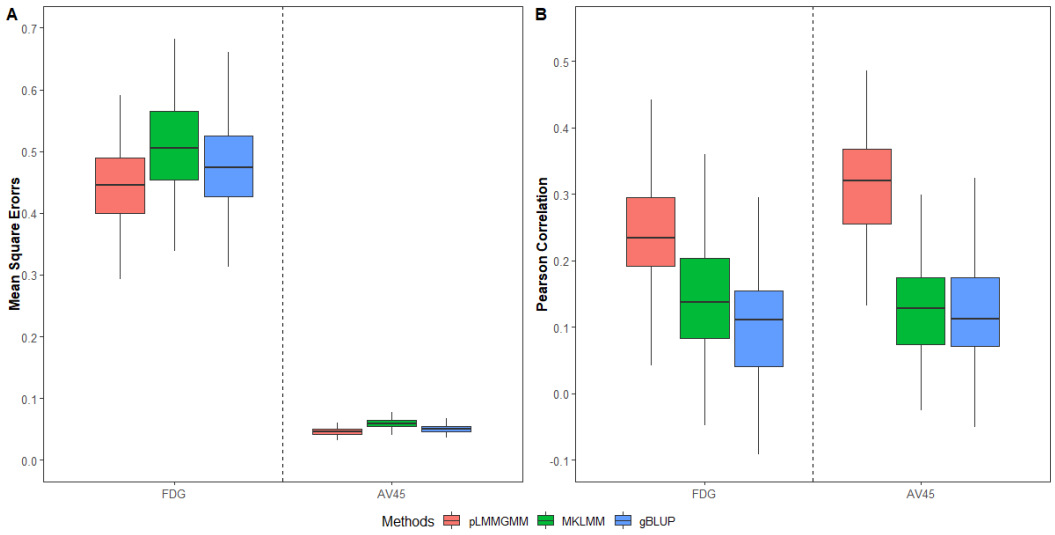

FIGURE 2.4: Accuracy comparisons for FDG and AV45

TABLE 2.4: The top three genes highly selected for FDG and AV45

Genes	Chromosome	Start Position	End Position	Probability(FDG)	Probability(AV45)
APOC1	19	45417920	45422606	1	1
APOE	19	45409038	45412650	1	0.97
TOMM40	19	45394476	45406946	0.28	1

Table 2.4 lists the three genes that were most highly selected by pLMMGMM for either AV45 or FDG, and the selection details of all the genes are shown in Supplementary Table A.3. For FDG, both *APOC1* and *APOE* were selected 100% of the time, and the remaining genes were selected, on average, less than 1% of the time. For AV45, *APOC1*, *APOE* and *TOMM40* were selected more than 97% of the time and the remaining genes were selected less than 5% of the time. The highly selected genes, *APOC1*, *APOE* and *TOMM40*, are well-known AD-related genes (Ossenkoppele et al., 2013; Roses, 2010). All three genes are located on chromosome 19 and are widely known as genetic risk factors of AD. For example, Duijn et al., 1994 found that *APOE* $\epsilon 4$ was highly associated with a group of 175 early-onset AD patients. Zhou et al., 2014b found the association of $rs11568822$ on *APOC1* gene with the increased AD risk in Caucasians, Asians and Caribbean Hispanics. Huang et al., 2016 found that $rs2075650$ on *TOMM40* is associated with AD patients for Caucasian and Asian subjects.

2.5 Discussion

In this work, we presented a novel and computationally efficient penalized LMM with GMM estimators for prediction modeling using high-dimensional genomic data. The proposed pLMMGMM first splits the genome into multiple regions and then adopts multiple kernels for each region to capture complex predictive effects. pLMMGMM simultaneously models the joint predictive effects from all variants within each region, and efficiently select those regions that are predictive via a GMM-based estimator. Through theoretical proof, we have shown that our proposed method can achieve consistency of variable selection and variable estimation. We have also shown that our proposed GMM estimators are asymptotically normal. Through extensive simulation studies and the analysis of the ADNI data set, we have demonstrated that our method: 1) is more accurate and robust against various underlying disease models; 2) can accurately detect predictive regions; and 3) is much more computationally efficient, especially when the number of regions is large (i.e., the number of random effects is large).

Genomic data are high-dimensional and contain a large amount of noise. Including noise variants in the analyses can reduce the robustness and accuracy of a risk prediction model (Byrnes et al., 2013). Within the LMM framework, gBLUP cannot perform variable selection (Yang et al., 2010), and has low prediction accuracy when a large amount of noise present. Other LMM-based methods (e.g., MultiBLUP and MKLMM) select regions based on empirical criteria (Speed and Balding, 2014; Weissbrod et al., 2016), which cannot guarantee the optimal prediction performance. Existing penalized LMMs can detect predictive regions and estimate their effects, but they can only handle a very limited number of regions (Li et al., 2020; Wen and Lu, 2020). On contrary, our proposed model can handle a large number of regions and efficiently remove those that are not predictive. We have proved that the probability to correctly identify all predictive regions approaches 1 when the sample size is large. In addition, the effect estimates for these predictive regions are unbiased and asymptotically normal. As shown in the first simulation (Figure 2.1 and Supplementary Figure A.1), the prediction accuracy for pLMMGMM remains stable as the amount of noise increases, whereas it can be greatly affected for other methods (i.e., gBLUP and MKLMM). Furthermore, the proposed pLMMGMM has achieved relatively high sensitivity and specificity, regardless of the number of noise regions (Table 2.1 and Supplementary Table A.1).

The underlying genetic etiology for many human diseases is complicated and is

usually unknown in advance. While it is widely accepted that models with flexible modeling assumptions can achieve more robust and accurate prediction performance across a range of phenotypes (Speed and Balding, 2014; VanRaden, 2008; Yang et al., 2010), existing LMMs mainly focus on linear relationships and thus their performance can be sub-optimal when non-linear predictive effects are present. The recent development in LMMs aims at capturing these non-linear predictive effects by embedding them into the reproducing kernel Hilbert space, where appropriate kernel functions that reflect the underlying disease etiology are used (Weissbrod et al., 2016). However, how to pre-choose appropriate kernel functions can be challenging in practice, as they can be quite disease/trait dependent. In this work, we adopted the idea used in multi-kernel learning algorithms and put multiple kernels into a candidate set, from which appropriate kernel(s) are chosen through a data-driven approach. Through simulations, we have shown that our proposed pLMMGMM has robust and accurate prediction performance across a range of disease models (Figure 2.3 and Supplementary Figure A.3). In addition, our model has relatively high sensitivity and specificity in correctly detecting prediction regions that harbour genetic variants with various types of predictive effects (Table 2.3 and Supplementary Table A.2).

Computational efficiency is one of the major bottlenecks for LMMs with multiple random effects (Li et al., 2020; Weissbrod et al., 2016; Wen and Lu, 2020). Traditional methods usually obtain the MLE/REML estimators, both of which can be computationally demanding, especially when the number of multiple random effects is large. In our work, we traded computational efficiency with statistical efficiency, and proposed to obtain parameter estimates via GMM. By using GMM, our objective function is much easier to optimize. Indeed, REML with a traditional Newton-Raphson algorithm can barely converge when the number of random effects is above 10. Even for REML with a simulated anneal algorithm where the global optimum is not guaranteed (Goffe et al., 1994), the computational time increases at a much faster rate compared with our proposed method (Figure 2.2 and Supplementary Figure A.2). The computational efficiency allows our proposed method to jointly model a large number of genetic regions and consider various forms of predictive effects, both of which can be important for improving risk prediction models.

In the prediction analyses of FDG and AV45, we found that the proposed method outperformed both MKLMM and gBLUP (Figure 2.4), indicating that the designed variable selection and multi-kernel learning in pLMMGMM can improve the prediction accuracy. In addition, the predictive genes selected by the pLMMGMM method are

also quite consistent (Supplementary Table A.3). The three well-known AD-related genes, *APOC1*, *APOE* and *TOMM40*, are highly selected ($> 97\%$) for both FDG and AV45. The *APOE* gene encodes apolipoprotein E which is involved in cholesterol transport (Zannis et al., 1993), and high levels of cholesterol play a significant role in the pathogenesis of AD (Puglielli et al., 2003). Indeed, the *APOE* gene has been identified as a major genetic risk factor for AD in existing literature (Poirier et al., 1993; Strittmatter et al., 1993). For example, Tang et al., 1998 found that the presence of *APOE* $\epsilon 4$ is a determinant risk factor of AD in Caucasians; and Graff-Radford et al., 2002 reported that one or two copies of *APOE* $\epsilon 4$ affect the risk of AD for African Americans. The *APOC1* gene encodes apolipoprotein C1, a member of the apolipoprotein family, and it affects the cholesterol metabolism that is involved in AD pathology (Poirier et al., 1993). In addition, it has also been found that the $rs4420638$ polymorphism on *APOC1* increases the accumulation of homocysteine, and thus affects AD risk (Prendecki et al., 2018). *TOMM40* regulates mitochondrial function and it is also a candidate gene for AD (Bagnoli et al., 2013; Huang et al., 2016; Ma et al., 2013). For example, Roses, 2010 found that $rs10524523$ on *TOMM40* is highly associated with late-onset AD. In addition, Prendecki et al., 2018 proposed that $rs10524523$ on *TOMM40* can affect oxidative damage and thus influence the onset and progression of AD. While our models improved prediction accuracy for both AV45 and FDG, additional replication studies are needed to further investigate the performance of our models.

In summary, we have proposed GMM-based penalized LMMs for risk prediction analyses on high-dimensional genomic data, where the variable selection consistency, variable estimation consistency and asymptotic normality of non-zero parameters have been established. Our proposed pLMMGMM method is highly computationally efficient. It can simultaneously consider a large number of genetic regions and accurately detect those that are predictive. In addition, our proposed method can accommodate various disease models, as it can select appropriate kernel functions that best reflect the underlying disease model via a data-driven approach. Similar to other existing literature (Speed and Balding, 2014; Weissbrod et al., 2016), the proposed pLMMGMM method only focus on continuous outcomes. It would be of interest to develop a generalized framework that can analyse binary outcomes or Poisson outcomes. Although our method can reduce the computational cost, it can still be computationally expensive for ultrahigh-dimensional data (e.g., multi-omics data). These will be the future directions of our research. The R-package is available at

https://github.com/XiaQiong/GMMLasso.

Chapter 3

A hybrid screening rule designed for the penalized linear mixed model with generalized method of moments estimators

3.1 Introduction

Precision medicine that takes individuals' differences into account has been recently initiated to provide tailored and effective health care (Ho et al., 2019). Accurate risk prediction models can not only detect individuals that are at high risk (Abraham and Inouye, 2015), but also provide precise diagnosis, interventions and treatment, playing a pivotal role on precision medicine (Ashley, 2015). Over the past decades, information from genomic data has been incorporated into the traditional risk prediction models, which has improved the prediction performance (Speed and Balding, 2014; Weissbrod et al., 2016; Wen and Lu, 2020). However, the high dimensionality of genomic data and the complex relationships between predictors and outcomes have imposed significant challenges for risk prediction models.

Linear mixed models (LMMs) have long been the method of choice for risk prediction analysis of genomic data (Li et al., 2020; Speed and Balding, 2014; Weissbrod et al., 2016; Yang et al., 2010). The well-known genomic best linear unbiased prediction (gBLUP) method, an equivalence to a LMM with one random effect term, was first proposed for the prediction of milk production (Harris et al., 2008) and then human traits (Yang et al., 2010). The assumptions in gBLUP are that effect sizes of all genetic variants are from the same normal distribution, and all genetic variants affect traits in

a linear manner. However, these assumptions can be too simply to be realistic. Speed and Balding, 2014 proposed the MultiBLUP, an extension of gBLUP, where genetic variants are allowed to have different effect size distributions. Different from gBLUP, MultiBLUP first groups genetic variants into regions based on some criteria (e.g., gene or pathway annotations) and then each region is screened using the likelihood ratio test to detect those that are significantly predictive. Finally, MultiBLUP builds risk prediction models using a LMM with multiple random effects, with each corresponding to a predictive region. While MultiBLUP has relaxed the assumptions of gBLUP, it only aims at modeling linear additive effects from chosen regions, and thus the adopted empirical screening criteria can be crucial for the final prediction performance of MultiBLUP. To reduce the impact of screening, Wen and Lu, 2020 and Li et al., 2020 have recently introduced an L_1 penalty into the random effects for LMM with multiple random effects, where predictive regions and their effects can be detected and modeled simultaneously. In addition, through theoretical investigations, they have shown that their methods can achieve selection and estimation consistency.

While the advances in LMM-based methods can facilitate risk prediction studies, the parameter estimations in LMMs can be computationally demanding, especially for a LMM with multiple random effects (Speed and Balding, 2014; Weissbrod et al., 2016; Wen and Lu, 2020). Most existing LMMs either use maximum likelihood estimators (MLE) or restricted maximum likelihood estimators (REML), both of which are usually obtained by Newton-Raphson or expectation-maximization algorithms that need repeat calculation of matrix inversion. Recently, Weissbrod et al., 2016 used a simulated annealing algorithm to optimize the objective function of LMMs, and thus their method can consider a large number of regions. However, the performance of simulated annealing algorithms is determined by empirical criteria. Therefore, the algorithm may simply find a local optimal or take an extremely long time to obtain the global optimal. In contrast, the generalized method of moments (GMM) has been a long-existing alternative for LMMs (Rao, 1970; Rao, 1971a; Rao, 1972). For example, Rao, 1972 proposed the minimum norm quadratic unbiased estimation (MINQUE) method to estimate the variance and covariance components in LMMs. In addition, Rao, 1971b proposed the minimum variance invariant quadratic unbiased estimation (MIVQUE) method for the estimations of variance components, finding the minimum variance estimator among all quadratic unbiased estimators. Although GMM is less statistically efficient than MLE/REML, it is computationally efficient. GMM transforms the objective function of LMMs into a quadratic form that is much easier to optimize. GMM

has been used in genetic studies (Zhou, 2017; Zhu, 1995; Zhu and Weir, 1996). For example, Zhu and Weir, 1996 predicted maternal and paternal effects of five plants in a bio-model for diallel crosses, where the MINQUE method is used for estimating variance and covariance components. Reimherr and Nicolae, 2016 examined the long-term effects of daily asthma medications on children using a LMM, where variance components are estimated using the MINQUE method. Henderson, 1985 predicted two genetic merits of five animals using a LMM, where MIVQUE is used for variance estimation of additive and non-additive genetic effects. Recently, Wang and Wen, 2021 also developed a GMM-based method to estimate parameters in penalized LMMs (denoted as pLMMGMM). They showed that their method can achieve estimation and variable selection consistency, and their GMM estimators are asymptotically normally distributed. Unlike traditional LMMs, which can handle a very limited number of random effects (≤ 10), the number of random effects that pLMMGMM can handle is much larger (≈ 100).

While treating all genetic variants as if they were predictive can dilute the signals from relevant markers, including only a very limited number of genetic variants cannot fully describe the genetic diversity and may only explain a very small fraction of heritability (De Los Campos et al., 2010). For example, Holzapfel et al., 2010 considered several highly significantly single nucleotide polymorphisms (SNPs) on *TMEM18* and *SH2B1* genes to predict body mass index, and found that these SNPs only account for about 0.006% of the variability. Genome-wide data, which allow for the consideration of the entire human genome, have great potential in improving prediction models as they can detect additional predictive variants and model their effects (De Los Campos et al., 2010). For example, Seshadri et al., 2010 identified two new loci (i.e., $rs744373$ near *BIN1* and $rs597668$ near *EXOC3L2/BLOC1S3/MARK4*) that are associated with late-onset Alzheimer's disease (AD) through genome-wide association studies and found that these two loci further explained variation in susceptibility to AD. While genome-wide data allow for systematic evaluations of the predictive effects from all genetic variants, their use can be a double-edged sword. On one hand, they offer a more comprehensive view of the human genome, if analyzed appropriately. On the other hand, the huge amount of measured variants from genome-wide data not only increases the impact of noise but also makes traditional methods not applicable primarily due to its heavy computation. This can lead to a prediction model that requires a huge amount of computational resources yet is much less robust and accurate. Existing methods that use genome-wide data often employ very simple assumptions to reduce

the model complexity and computational burden (Yang et al., 2010; Yang and Zhou, 2020). For example, gBLUP assumes that the effect sizes from all genetic variants come from the same normal distribution, and thus can efficiently model genome-wide data through a single parameter (i.e., a random effect term in LMM) (Harris et al., 2008; Yang et al., 2010). Similarly, the Deterministic Bayesian Sparse Linear Mixed Model (DBSLMM) first detects the markers with large effects through marginal associations obtained via simple analysis (e.g., regression), and then treats them as fixed effects with the remaining variants being modeled similarly to gBLUP. By assuming that all large-effect predictors have marginal additive effects, DBSLMM significantly simplifies the model, making it scalable to the prediction analysis of large-scale data (Yang and Zhou, 2020). While simple assumptions allow for the consideration of genome-wide data, they may lead to sub-optimal prediction performance. For example, both gBLUP and DBSLMM assume linear relationships, and thus neither model is capable of modeling complex effects (Yang et al., 2010; Yang and Zhou, 2020) and their performance can drop substantially when data severely violate the adopted model assumptions. Using an empirical screening rule that is designed to align well with the downstream task is another common approach for handling genome-wide data (Speed and Balding, 2014; Weissbrod et al., 2016). For example, both MKLMM and MultiBLUP first fit a LMM with single random effect for each region, and only regions that are significant from those single-region LMMs are included as random effects in the final LMM (Speed and Balding, 2014; Weissbrod et al., 2016). By employing this empirical screening rule, MultiBLUP and MKLMM substantially reduce the number of model parameters, making both models applicable to use with genome-wide data. While an empirical screening rule is useful in practice, the rule may discard predictive regions, leading to a low prediction accuracy. Penalized LMMs that can simultaneously detect predictive markers and model their effects have shown their advantages in prediction studies (Li et al., 2020; Wen and Lu, 2020). However, they have not been able to scale to genome-wide data so far, even for the recently developed pLMMGMM (Wang and Wen, 2021). Existing penalized LMMs usually go with a candidate gene approach, where genetic variants from genome-wide data are first filtered according to existing literature (Hai and Wen, 2020; Wang and Wen, 2021; Wen and Lu, 2020). Although these candidate gene approaches tend to be consistent, they only consider existing knowledge and can overlook important predictors that have not been reported yet. Therefore, a screening rule designed for penalized LMMs can be of great importance for the prediction analysis of genome-wide data.

Screening rules, especially those designed for penalized models, can reduce the number of variables entered into the optimization process, and thus can substantially reduce the model complexity, making penalized models applicable to genome-wide data (Fan and Lv, 2008; Ghaoui et al., 2010; Tibshirani et al., 2012; Wang et al., 2015; Xiang et al., 2016). Fan and Lv, 2008 proposed the sure independence screening (SIS) method, one of the pioneering works in the field, for linear models, where marginal correlations are used to rank the importance of each variable. Fan and Lv, 2008 established the sure screening property of their procedure under a Gaussian linear model framework and the probability that all important variables survive approaches to 1. Fan et al., 2009 further proposed a SIS procedure for generalized linear models based on marginal likelihood estimates, and SIS has also been extended for other models, such as nonparametric additive models (Fan et al., 2011) and a cox proportional hazard model (Zhao and Li, 2012). Although easy to implement, SIS and its extensions only consider marginal effects and cannot take their joint effects into account. Moreover, SIS is not designed for penalized models, and thus those redundant variables selected by SIS are kept in the final prediction models.

Unlike SIS and its extensions, the strong and safe rules are designed for penalized models (Tibshirani et al., 2012; Wang et al., 2015). A strong rule that is designed for a penalized model with L_1 penalty (i.e., Lasso type of problem) can discard the ith variable (denoted by $\boldsymbol{X}_i$) for a given penalty value of λ if $\mid \boldsymbol{X}_i^T\boldsymbol{y} \mid < 2\lambda - \lambda_{max}$, where $\lambda_{max} = \max_i \mid \boldsymbol{X}_i^T\boldsymbol{y} \mid$ and $\boldsymbol{y}$ is the vector of outcomes (Tibshirani et al., 2012). The strong rule is simple and can screen out a large amount of inactive variables. However, it can also mistakenly discard variables that are indeed predictive, leading to a sub-optimal prediction model (Wang et al., 2015). Unlike strong rules, safe screening rules that are also designed for penalized models can guarantee that the discarded variables have no effects (Wang et al., 2015) for each penalty λ. These safe screening rules are usually based on exploring geometric properties of the dual Lasso problem. The objective function of a penalized model with L_1 penalty is usually of the form: argmin $\frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1$, where $\boldsymbol{X}$ is the variable matrix and $\boldsymbol{\beta}$ is the unknown coefficients vector. The dual form (Kim et al., 2007) can be written as:

$$\text{argmax } \frac{1}{2}||\boldsymbol{y}||_2^2 - \frac{\lambda^2}{2}||\boldsymbol{\theta} - \frac{\boldsymbol{y}}{\lambda}||_2^2, \quad s.t. \ |\boldsymbol{X}_i^T\boldsymbol{\theta}| \leq 1$$

where $\boldsymbol{\theta}$ is the dual variable and the coefficient of the ith variable is zero when the optimal condition (i.e., $|\boldsymbol{X}_i^T\boldsymbol{\theta}| < 1$) is satisfied. The optimal solution $\boldsymbol{\theta}^*$ is generally

unknown, and a set $\boldsymbol{\Theta}$ that contains optimal solution $\boldsymbol{\theta}^*$ is often used for screening. Indeed, the set $\boldsymbol{\Theta}$ determines the effectiveness of the screening rule. The smaller the set $\boldsymbol{\Theta}$ is, the more effective the screening rule is (Wang et al., 2015), which motivates different safe screening rules being proposed to find a smaller $\boldsymbol{\Theta}$. The pioneering work of safe screening rules is the SAfe Feature Elimination (SAFE) rule, and it bounds the optimal solution $\boldsymbol{\theta}^*$ of the dual problem within a sphere (Ghaoui et al., 2010). The Dome rule further shrinks the set $\boldsymbol{\Theta}$ via bounding the optimal solution $\boldsymbol{\theta}^*$ into a spherical dome region (Xiang and Ramadge, 2012). Both SAFE and Dome rules are based on the estimation of the dual optimal solution that is unknown in advance, therefore it can be quite challenging to accurately estimate them (Wang et al., 2015). Recent efforts have been made to address this issue. For example, the Dual Polytope Projection (DPP) rule is designed based on the uniqueness and non-expansiveness of the optimal dual solutions (Wang et al., 2015), and thus the problem in the dual space becomes convex. DPP can discard more inactive variables than SAFE rule as it bounds the optimal solution within a smaller region. At a given value of λ, DPP screens out the ith variable when $| \boldsymbol{X}_i^T \frac{\boldsymbol{y}}{\lambda_{max}} | < 1 - (\frac{1}{\lambda} - \frac{1}{\lambda_{max}}) \|\boldsymbol{y}\|_2 \|\boldsymbol{X}_i\|_2$ is met. While safe rules can guarantee that removed variables are inactive in the corresponding penalized models, they can only discard a moderate number of variables and the remaining number of variables can still be large, making safe rules not appealing for ultra-high dimensional data analysis.

For penalized models, the optimal tuning parameter λ is usually selected from a sequence of candidates, and thus the inactive sets have to be determined accordingly. While basic screening rules (e.g., strong rule and DPP rule) can be applied to discard variables for each penalty, it is computationally expensive. It has been recognized that the size of inactive set increases as the penalty parameter λ decreases. Therefore, the basic screening rules have been extended into a sequential version, where inactive sets are efficiently updated via a grid of decreasing regularization parameters $\lambda_1 > \lambda_2 > ... > \lambda_K$. For example, the sequential strong rule (SSR) obtains the screening results at λ_{k+1} based on the solution of $\hat{\boldsymbol{\beta}}(\lambda_k)$ at λ_k, and the ith variable at λ_{k+1} is discarded when $|\boldsymbol{X}_i^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_k))| < 2\lambda_{k+1} - \lambda_k$. Similarly, the enhanced DPP (EDPP) rule determines the inactive set at penalty λ_{k+1} based on a known optimal solution $\hat{\boldsymbol{\beta}}(\lambda_k)$ at λ_k. While these sequential screening rules reduce the computational complexity, they still suffer the same problems as their basic versions, where the strong rule cannot guarantee that discarded variables have no effects and the safe rules can only discard a limited number of variables. Therefore, neither sequential strong or safe rules can be

directly incorporated into the penalized LMMs for prediction analysis of genome-wide data.

Many complex diseases, such as type 2 diabetes and cancers, are affected by multiple genetic factors through complex biological pathways (Kirchner et al., 2013). One of the fundamental problems in risk prediction studies is how to detect and model the underlying genetic architecture of complex traits (Ho and Hsu, 2015), especially those non-linear effects. Existing studies mainly focus on detecting predictors with linear additive effects (Chatterjee et al., 2016; Hill et al., 2008). For example, Reyes-Gibby et al., 2009 evaluated 59 SNPs in 37 inflammation genes in newly diagnosed non-Hispanic Caucasian lung cancer patients, showing a linear effect on their pain severity. Collado-Hidalgo et al., 2008 found that *IL1B*-511 (C/T) polymorphism on the *IL1B* gene may have a linear relationship with persistent fatigue in the aftermath of breast cancer. However, as indicated by Ho and Hsu, 2015, non-linear effects (e.g., epistasis) widely exist. For example, Badano et al., 2006 identified a novel locus, *MGC1203*, which interacts with the *BBS4* gene and has an epistatic effect on the developmental phenotype of Bardet–Biedl syndrome. Using combinatorial RNA in human breast epithelial cells, Wang et al., 2014b found hundreds of genetic interactions (i.e., epistasis) among 67 genes that are frequently altered in breast cancer as well as other cancer types. Therefore, recent efforts have been made to develop analytical methods that can capture non-linear effects efficiently for prediction studies (Li et al., 2020; Weissbrod et al., 2016; Wen and Lu, 2020). Within the LMM framework, Weissbrod et al., 2016 proposed MKLMM, an extension of MultiBLUP, to model both linear and non-linear effects for prediction analyses. They first embed genetic variants into the reproducing kernel Hilbert space and then use kernel functions (e.g., linear, radial basis function and polynomial kernels) to accommodate various types of predictive effects. Wen and Lu, 2020 and Li et al., 2020 further extended MKLMM so that predictive markers as well as their types of effects can be determined via a data-driven manner. Although these methods enable the detection of non-linear effects, they suffer from the same computational challenges faced by LMMs. For example, to consider non-linear effects, multi-kernel LMMs model each region using multiple kernel functions with each corresponding to a random effect term in LMMs (Li et al., 2020; Weissbrod et al., 2016; Wen and Lu, 2020). As a consequence, multi-kernel LMMs can have a much larger number of random effects than a LMM that only considers linear relationships. Therefore, multi-kernel LMMs require the selection of both predictive regions and their types of effects (i.e., kernels). This can lead to much heavier computation, especially for the

analysis of genome-wide data.

To address these issues, we have incorporated a hybrid screening rule into a penalized LMM with GMM estimators (denoted as HpLMMGMM) to simultaneously select predictors from genome-wide data and estimate their predictive effects. The proposed HpLMMGMM first applies the designed hybrid screening rule to discard a large amount of noise variables, and then fits penalized LMMs with GMM estimators based on the remaining variables. The HpLMMGMM method can be applied to genome-wide data and efficiently captures both linear and non-linear predictive effects. In section 2, we present our hybrid screening rule designed for penalized LMM with GMM estimators. The results from the simulation studies and the analysis of whole-genome sequencing data obtained from the ADNI study are summarized in sections 3 and 4, respectively. We summarize our findings in section 5.

3.2 Methods

3.2.1 A penalized linear mixed model with generalized method of moments estimators

For completeness, we first briefly describe the penalized LMM with GMM estimators (pLMMGMM) developed by Wang and Wen, 2021. Similar to existing LMM-based methods (Li et al., 2020; Speed and Balding, 2014; Weissbrod et al., 2016; Wen and Lu, 2020), pLMMGMM first splits the genome into multiple regions and estimates the cumulative predictive effects of all genetic variants within the region. It models the outcomes as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{R} \boldsymbol{g}_i + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma_0^2 \boldsymbol{I}_n)$$

where $\boldsymbol{X}$ is an $n \times P$ matrix of the demographic variables (e.g., age and gender) and $\boldsymbol{\beta}$ is their effect sizes; R is the total number of regions considered; $\boldsymbol{g}_i$ models the cumulative predictive effects from all genetic variants within the ith region, where $\boldsymbol{g}_i \sim N(0, \sum_m^M \boldsymbol{K}_{im}\sigma_{im}^2)$ and M is the total number of kernels considered. $\boldsymbol{K}_{im}$ is a kernel matrix calculated based on the mth kernel for region i and measures the genetic similarities based on the mth kernel that reflects the assumed relationships between predictors and outcomes. For example, if both linear relationship and pair-wise interaction effects are considered, then a linear kernel (i.e., $\boldsymbol{K}_{i1} = \boldsymbol{G}_i\boldsymbol{G}_i^T$) and polynomial kernel

with 2 degrees of freedom (i.e., $\boldsymbol{K}_{i2} = (\boldsymbol{G}_i\boldsymbol{G}_i^T)^2$) are used. Therefore, the cumulative predictive effects from the ith regions are assumed to be $\boldsymbol{g}_i \sim N(0, \boldsymbol{K}_{i1}\sigma_{i1}^2 + \boldsymbol{K}_{i2}\sigma_{i2}^2)$.

To select predictive regions and appropriate kernel functions that reflect the underlying relationships, pLMMGMM imposes an L_1 penalty on the associated parameters $\boldsymbol{\sigma}^2 = (\sigma_{11}^2, \cdots, \sigma_{1M}^2, \cdots, \sigma_{R1}^2, \cdots, \sigma_{RM}^2)$. Rather than using the traditional MLE/REML estimators that are computationally expensive, pLMMGMM obtains its parameters based on the GMM estimator with the corresponding objective function defined as:

$$\hat{\boldsymbol{\sigma}}^2 = \underset{\boldsymbol{\sigma}^2 \geq 0}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A} - \boldsymbol{A}^T \sum_{i=1}^{R} \sigma_i^2 \boldsymbol{K}_i\boldsymbol{A} - \sigma_0^2 \boldsymbol{I}_{n-P} \right\|_2^2 + \lambda \left\| \boldsymbol{\sigma}^2 \right\|_1 \tag{3.1}$$

where $\lambda > 0$; $\|\cdot\|_i, i \in \{1, 2\}$ denotes the $\ell - i$ norm; and the $\boldsymbol{A}$ matrix is chosen such that $\boldsymbol{A}^T\boldsymbol{Y}$ is independent of the covariates $\boldsymbol{X}$. To further simplify the notations, let $\boldsymbol{M} = vec(\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A})$, $\boldsymbol{T}_i = vec(\boldsymbol{A}^T\boldsymbol{K}_i\boldsymbol{A})$ is the ith column of $\boldsymbol{T}$ matrix, and $vec(.)$ is the vectorization of a matrix. The objective function 3.1 can be rewritten as:

$$\hat{\boldsymbol{\sigma}}^2 = \underset{\boldsymbol{\sigma}^2 \geq 0}{\operatorname{argmin}} \frac{1}{2} \left\| \boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^2 \right\|_2^2 + \lambda \sum_{i=1}^{R} \sigma_i^2, \quad \lambda > 0 \tag{3.2}$$

3.2.2 A hybrid screening rule designed for the penalized linear mixed model with generalized method of moments estimators

It is straightforward to see the optimization problem in equation 3.2 is a standard non-negative Lasso problem. Therefore, motivated by the idea in Zeng et al., 2021, we propose to design a hybrid screening rule, where SSR and EDPP rules are combined to correctly and effectively discard a large amount of noise. The SSR rule can mistakenly screen out predictive variables, and thus in practice Karush-Kuhn-Tucker (KKT) conditions have to be checked for each variable, which can substantially increase the computational burden. Therefore, our idea is to first apply the EDPP rule to discard a moderate number of variables that are guaranteed to be inactive, and then adopt the SSR rule on the remaining variables so that the KKT checking is only employed for a fraction of variables.

Suppose that we are given a sequence of decreasing penalties $\lambda_1 > \lambda_2 > ... > \lambda_K$ and the parameter estimates at λ_k (denoted as $\hat{\boldsymbol{\sigma}}^2(\lambda_k)$), we will first apply the EDPP rule to discard a moderate number of variables at λ_{k+1}. Specifically, the ith variable is

treated as inactive at λ_{k+1} if the following condition holds:

$$\left|\boldsymbol{T}_i^T\left(\frac{\boldsymbol{M}-\boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_k)}{\lambda_k}+\frac{1}{2}\boldsymbol{v}_2^{\perp}(\lambda_{k+1},\lambda_k)\right)\right|<1-\frac{1}{2}\left\|\boldsymbol{T}_i\right\|_2\left\|\boldsymbol{v}_2^{\perp}(\lambda_{k+1},\lambda_k)\right\|_2 \tag{3.3}$$

where $\boldsymbol{v}_2^{\perp}(\lambda_{k+1},\lambda_k)=\boldsymbol{v}_2(\lambda_{k+1},\lambda_k)-\frac{<\boldsymbol{v}_1(\lambda_k),\boldsymbol{v}_2(\lambda_{k+1},\lambda_k)>}{\|\boldsymbol{v}_1(\lambda_k)\|_2^2}\boldsymbol{v}_1(\lambda_k)$; $<\cdot>$ is the inner product; and $\boldsymbol{v}_2(\lambda_{k+1},\lambda_k)=\frac{\boldsymbol{M}}{\lambda_{k+1}}-\frac{\boldsymbol{M}-\boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_k)}{\lambda_k}$. $\boldsymbol{v}_1(\lambda_k)=\frac{\boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_k)}{\lambda_k}$ when $0<\lambda_k<\lambda_{max}$, and it equals to $\text{sign}(\boldsymbol{T}_*^T\boldsymbol{M})\boldsymbol{T}_*$ when $\lambda_k=\lambda_{max}$ and $\boldsymbol{T}_*=\text{argmax}_{\boldsymbol{T}_i}\mid\boldsymbol{T}_i^T\boldsymbol{M}\mid$. The details of the derivation of our employed EDPP rule are shown in Supplementary Materials B.1.2.

$S_{\lambda_{k+1}}^{EDPP}:=\{\boldsymbol{T}_i:\left|\boldsymbol{T}_i^T\left(\frac{\boldsymbol{M}-\boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_k)}{\lambda_k}+\frac{1}{2}\boldsymbol{v}_2^{\perp}(\lambda_{k+1},\lambda_k)\right)\right|\geq 1-\frac{1}{2}\left\|\boldsymbol{T}_i\right\|_2\left\|\boldsymbol{v}_2^{\perp}(\lambda_{k+1},\lambda_k)\right\|_2\}$ represents the set of remaining variables after applying the EDPP rule. As the size of set $S_{\lambda_{k+1}}^{EDPP}$ can be large for high-dimensional data, we propose to apply the SSR rule on $\boldsymbol{T}_i\in S_{\lambda_{k+1}}^{EDPP}$ to further discard inactive variables. Specifically, $\boldsymbol{T}_i$ is discarded by our adopted SSR rule when the following conditions are met:

$$\left|\boldsymbol{T}_i^T\frac{\boldsymbol{r}}{n}\right|<2\lambda_{k+1}-\lambda_k \tag{3.4}$$

where n is the sample size; $\boldsymbol{r}=\boldsymbol{M}-\boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_k)$ and $\boldsymbol{T}_i\in S_{\lambda_{k+1}}^{EDPP}$. The detailed derivations can be found in Supplementary Materials B.1.1.

Let $S_{\lambda_{k+1}}^{SSR}$ denote the remaining variables after applying the SSR rule, and thus $S_{\lambda_{k+1}}^{SSR}:=\{\boldsymbol{T}_i:\left|\boldsymbol{T}_i^T\frac{\boldsymbol{r}}{n}\right|\geq 2\lambda_{k+1}-\lambda_k$ and $\boldsymbol{T}_i\in S_{\lambda_{k+1}}^{EDPP}\}$ by definition. Since the SSR rule can mistakenly screen out active predictors, KKT conditions must be checked for all discarded variables; i.e., $\boldsymbol{T}_i\in S_{\lambda_{k+1}}^{EDPP}\cap\bar{S}_{\lambda_{k+1}}^{SSR}$. Therefore, we first fit model 3.2 with inputs being $\{\boldsymbol{T}_i:\boldsymbol{T}_i\in S_{\lambda_{k+1}}^{SSR}\}$ to estimate $\hat{\boldsymbol{\sigma}}^2(\lambda_{k+1})$, and then check whether each of the discarded variable satisfies the KKT condition:

$$\left|\frac{\boldsymbol{T}_i^T\boldsymbol{r}_{\lambda_{k+1}}}{n}\right|<\lambda_{k+1}$$

where $\boldsymbol{r}_{\lambda_{k+1}}=\boldsymbol{M}-\boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_{k+1})$ and $\boldsymbol{T}_i\in S_{\lambda_{k+1}}^{EDPP}\cap\bar{S}_{\lambda_{k+1}}^{SSR}$. Let $S_{\lambda_{k+1}}^{SSR,KKT}$ denote the set of all mistakenly discarded variables by the SSR rule, and thus $S_{\lambda_{k+1}}^{SSR,KKT}:=\{\boldsymbol{T}_i:\left|\frac{\boldsymbol{T}_i^T\boldsymbol{r}_{\lambda_{k+1}}}{n}\right|\geq\lambda_{k+1}$ and $\boldsymbol{T}_i\in S_{\lambda_{k+1}}^{EDPP}\cap\bar{S}_{\lambda_{k+1}}^{SSR}\}$. If there are active variables being screened out (i.e., $S_{\lambda_{k+1}}^{SSR,KKT}\neq\emptyset$), we first update the input set as $S_{\lambda_{k+1}}^{SSR}=S_{\lambda_{k+1}}^{SSR}\bigcup S_{\lambda_{k+1}}^{SSR,KKT}$. We then refit model 3.2 and re-check the KKT conditions with the updated input set $S_{\lambda_{k+1}}^{SSR}$. We repeat this process until $S_{\lambda_{k+1}}^{SSR,KKT}=\emptyset$. The final variables included

in model 3.2 at λ_{k+1} are $S^{SSR}_{\lambda_{k+1}}$ when $S^{SSR,KKT}_{\lambda_{k+1}} = \emptyset$. The detailed procedure of our proposed hybrid screening rule is shown in algorithm 1.

Algorithm 1 The hybrid screening rule

Input: $\boldsymbol{T}$, $\boldsymbol{M}$, $\lambda_{max} = \lambda_0 > \lambda_1 > ... > \lambda_K$, $\lambda_{max} = \max_i \mid \boldsymbol{T}_i^T \boldsymbol{M} \mid$, n is the sample size

Initialization: $\hat{\boldsymbol{\sigma}}^2(\lambda_0) = \mathbf{0}$

for $k \in \{0, 1, ..., K\}$ **do**

EDPP screening: $S^{EDPP}_{\lambda_{k+1}} := \{\boldsymbol{T}_i : \left| \boldsymbol{T}_i^T \left(\frac{\boldsymbol{M} - \boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_k)}{\lambda_k} + \frac{1}{2}\boldsymbol{v}_2^{\perp}(\lambda_{k+1}, \lambda_k) \right) \right| \geq 1 - \frac{1}{2} \left\| \boldsymbol{T}_i \right\|_2 \left\| \boldsymbol{v}_2^{\perp}(\lambda_{k+1}, \lambda_k) \right\|_2 \}$

SSR screening: $S^{SSR}_{\lambda_{k+1}} := \{\boldsymbol{T}_i : \left| \boldsymbol{T}_i^T \frac{\boldsymbol{M} - \boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_k)}{n} \right| \geq 2\lambda_{k+1} - \lambda_k \text{ and } \boldsymbol{T}_i \in S^{EDPP}_{\lambda_{k+1}} \}$

while $S^{SSR,KKT}_{\lambda_{k+1}} \neq \emptyset$ **do**

Estimate $\hat{\boldsymbol{\sigma}}^2(\lambda_{k+1})$ via equation 3.2 with inputs being $\{\boldsymbol{T}_i : \boldsymbol{T}_i \in S^{SSR}_{\lambda_{k+1}}\}$

Update residual: $\boldsymbol{r}_{\lambda_{k+1}} = \boldsymbol{M} - \boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda_{k+1})$

Check KKT: $S^{SSR,KKT}_{\lambda_{k+1}} := \{\boldsymbol{T}_i : \left| \frac{\boldsymbol{T}_i^T \boldsymbol{r}_{\lambda_{k+1}}}{n} \right| \geq \lambda_{k+1} \text{ and } \boldsymbol{T}_i \in S^{EDPP}_{\lambda_{k+1}} \cap \bar{S}^{SSR}_{\lambda_{k+1}} \}$

if $S^{SSR,KKT}_{\lambda_{k+1}} \neq \emptyset$ **then**

$S^{SSR}_{\lambda_{k+1}} = S^{SSR}_{\lambda_{k+1}} \bigcup S^{SSR,KKT}_{\lambda_{k+1}}$

end if

end while

end for

3.2.3 Prediction

To choose an appropriate penalty parameter that performs the best for prediction, we consider a sequence of penalty values in a decreasing order, $\lambda_1 > \lambda_2 > ... > \lambda_K$. For a given value of λ_k, we first apply the proposed hybrid screening rule to prune a large number of inactive variables. We then fit model 3.2 with all inputs that are in the set $S^{SSR}_{\lambda_k}$. Bayesian information criteria (BIC) are calculated for model 3.2 at λ_k. The optimal penalty parameter λ^* is selected based on BIC and the variance components estimates at the optimal penalty parameter λ^* are used for prediction. We denote the parameter estimates at the optimal value of λ^* as $\hat{\boldsymbol{\sigma}}^2(\lambda^*)$.

Let $\boldsymbol{Y}_a = (\boldsymbol{Y}_p, \boldsymbol{Y})$, where $\boldsymbol{Y}$ is the $n \times 1$ vector of outcomes in the training data and $\boldsymbol{Y}_p$ is $n_p \times 1$ vector of outcomes in the testing data. After obtaining the variance components estimates $\hat{\boldsymbol{\sigma}}^2(\lambda^*)$, the variance of $\boldsymbol{Y}_a$ can be directly derived as $\hat{\boldsymbol{\Sigma}}_{Y_a} =$

$\sum_i^R \sum_m^M \boldsymbol{K}_{im}\hat{\sigma}^2_{im}(\lambda^*) + \boldsymbol{I}_{n_p+n}\hat{\sigma}^2_0(\lambda^*)$, where $\boldsymbol{K}_{im}$ is the $(n_p + n) \times (n_p + n)$ genetic similarity matrix measured by the mth kernel for region i calculated from all samples. The variance of $\boldsymbol{Y}_a$ can be written as:

$$\hat{\boldsymbol{\Sigma}}_{Y_a} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{pp} & \hat{\boldsymbol{\Sigma}}_{po} \\ \hat{\boldsymbol{\Sigma}}_{op} & \hat{\boldsymbol{\Sigma}}_{oo} \end{bmatrix}$$

where $\hat{\boldsymbol{\Sigma}}_{pp}$ and $\hat{\boldsymbol{\Sigma}}_{oo}$ are the variance matrices for the testing and training samples, respectively. $\hat{\boldsymbol{\Sigma}}_{po}$ is the covariance between the testing and training samples. Therefore, the predictive values for the testing samples can be calculated as:

$$\boldsymbol{Y}_p = \boldsymbol{X}_p\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Sigma}}_{po}\hat{\boldsymbol{\Sigma}}_{oo}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_o\hat{\boldsymbol{\beta}})$$

where the $\boldsymbol{X}_p$ and $\boldsymbol{X}_o$ are the covariates of the testing and training samples, respectively.

3.3 Simulation studies

We conducted extensive simulation studies to evaluate the impact of data dimension and the underlying disease models on the performance of HpLMMGMM. We further compared our method with two widely used methods that can be applied to genome-wide data, including gBLUP (Yang et al., 2010) under its default setting and MKLMM (Weissbrod et al., 2016). Note that we did not compare our method with MultiBLUP, as MultiBLUP is equivalent to MKLMM with a linear kernel. MKLMM requires the user to specify the number of chosen regions. Therefore, we considered three settings. We first set the number of chosen regions equal to the top 5% of regions based on the likelihood ratio test that is employed to screen the regions. This is mainly because MKLMM first divides all the genetic variants into regions and then discards those whose likelihood is among the bottom 95%. We then considered the setting where the number of selected regions is 9 (denoted as MKLMM9). This is mainly because Weissbrod et al., 2016 pointed out that no improvement in prediction is observed for more than 9 regions. Finally, we considered the setting where the pre-specified number of regions is set to be 2 (denoted as MKLMM2). This is mainly because for all our simulations, the number of casual regions is equal to 2. To mimic the real human genome, we directly obtained genomic data from the 1000 Genome Project (The 1000 Genomes Project Consortium, 2015), and built genetic regions based on 10 randomly selected SNPs that are within 75Kb. For all settings, we conducted 100 Monte Carlo

simulations and considered sample sizes of 500 and 1000, where 65% of the samples were randomly selected for training and the remaining samples were used to gauge the prediction performance, measured by both Pearson correlations and mean square errors (MSEs). We further evaluated the efficiency of variable screening rule employed by both our method and MKLMM.

3.3.1 Scenario I: the impact of data dimension

The genome-wide data are high-dimensional, and there are approximately 20,000 genes for human genomes. To evaluate the impact of data dimension, we gradually increased the number of regions from 5000 to 20000 with the number of SNPs increasing from 50,000 to 200,000. We randomly selected two regions and set them as causal. We simulated the outcomes under an additive model:

$$Y_i = \sum_j G_{ij}^1 \beta_{ij}^1 + \sum_j G_{ij}^2 \beta_{ij}^2 + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_0^2)$; $G_{ij}^k, k \in \{1, 2\}$ is the jth genotype on the kth causal region for individual i; and $\beta_{ij}^k \sim N(0, \sigma_k^2)$ is the effect size of genetic variants. It is straightforward to show that:

$$\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{K}_1\sigma_1^2 + \boldsymbol{K}_2\sigma_2^2 + \boldsymbol{I}_n\sigma_0^2) \tag{3.5}$$

where $\boldsymbol{K}_k = \boldsymbol{G}^k\boldsymbol{G}^{k^T}$ and $\boldsymbol{G}^k$ is the genotype matrix for region k. Therefore, we simulated the outcomes based on a multivariate normal distribution following equation 3.5. The details of the simulation setting are summarized in Supplementary Table B.1. We reported the Pearson correlations and MSEs based on the testing samples. We further presented the average total number of selected regions and the number of causal regions among the selected regions for both the hybrid screening rule used by HpLMMGMM and the empirical screening rule employed by MKLMM.

Pearson correlations and MSEs for sample sizes of 500 and 1000 are shown in Figure 3.1 and Supplementary Figure B.1, respectively. Of all the scenarios considered, HpLMMGMM performed the best, followed by MKLMM and gBLUP. MKLMM2 and MKLMM9 always performed the worst. This is mainly because the causal regions are not ranked the top based on the likelihood ratio test statistics, and thus both MKLMM2 and MKLMM9 essentially built the prediction models based on only noise. This leads to almost no prediction power regardless of the number of regions considered. For MKLMM and gBLUP, their prediction performance decreased as the number of

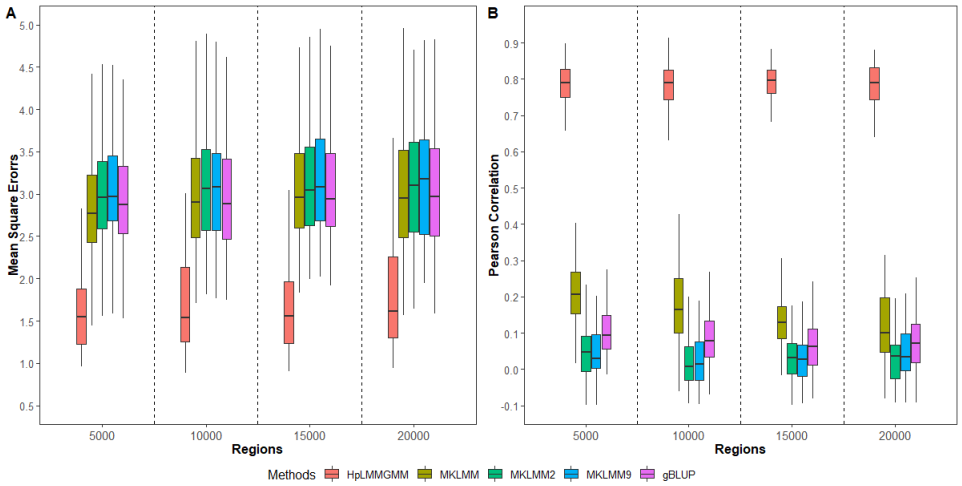

FIGURE 3.1: The impact of data dimension on Pearson correlations and MSEs ($n = 500$)

simulated regions increased. gBLUP models the cumulative predictive effects from all measured genetic variants and ignores the impact of noise. As a consequence, its performance can worsen as the amount of noise accumulates. For the MKLMM method, the screening rule it employs can help to reduce the impact of noise, and thus MKLMM tends to have better prediction accuracy than gBLUP. However, as the number of regions increases, the remaining regions included in the prediction model can still include lots of noise, leading to reduced prediction performance. For the proposed HpLMMGMM, the hybrid screening rule can screen out lots of inactive regions and the penalization used in LMM can further fine-tune the selected regions to achieve an optimal prediction performance. As shown in Figure 3.1 and Supplementary Figure B.1, HpLMMGMM maintains a robust prediction performance as the number of regions approaches to genome-wide level.

TABLE 3.1: The number of selected total and causal regions as the input data dimension increases ($n = 500$)

Regions	Number of Total Regions Selected (Number of Causal Regions Selected)			
	MKLMM	MKLMM2	MKLMM9	Hybrid Screening Rule
5000	292.84 (1.99)	2 (0)	9 (0)	13.57 (1.99)
10000	589.64 (1.96)	2 (0)	9 (0)	17.06 (1.99)
15000	884.73 (1.99)	2 (0)	9 (0)	16.71 (1.98)
20000	1178.02 (1.96)	2 (0)	9 (0)	17.46 (1.98)

Table 3.1 and Supplementary Table B.2 present the efficiencies of the screening rules employed by both MKLMM and our proposed method. Apparently, neither of the two

causal regions have been ranked among the top 9, and thus MKLMM with 2 or 9 regions failed to capture their predictive effects. This means that MKLMM2 and MKLMM9 had no predictive power (Figure 3.1 and Supplementary Figure B.1). MKLMM screens out the bottom 95% of regions based on the rank of the Likelihood Ratio (LR) test for each region. Although MKLMM could detect the predictive markers, the number of noise regions also increased with the data dimension. Therefore, MKLMM would fit a LMM with lots of random effects that are not associated with the outcomes. This not only increases computational complexity, but also reduces the accuracy and robustness of the prediction model. The proposed hybrid screening rule, however, could tease out a huge amount of noise while still keeping the causal regions. Indeed, when the number of regions increase from 5,000 to 20,000, the number of kept regions for the hybrid screening rule remained relatively stable (i.e., within 20), whereas the number of kept regions increased from 293 to 1,178 for MKLMM. Therefore, comparing with MKLMM, the data dimension and the amount of noise had little impact on the performance of HpLMMGMM. This makes the proposed method robust enough to be applied to high-dimensional data.

3.3.2 Scenario II: the impact of disease models

Complex human diseases are affected by a wide array of genes through a complicated biological system that is usually unknown in advance (Chatterjee et al., 2013). Therefore, we evaluated the performance of our proposed method under various underlying disease models. Similar to Scenario I, we considered two causal regions and generated the outcomes using equation 3.5. We used different kernel matrices to reflect the assumed relationships between predictors and outcomes, and simulated five disease models with genetic variants. Specifically, we considered simulations when: 1) both causal regions have linear additive effects and linear kernels are used to simulate outcomes (denoted as model $L + L$); 2) both causal regions have non-linear effects and radial basis function (RBF) kernels are used (denoted as model $R + R$); 3) both causal regions only have pair-wise interaction effects and polynomial kernels of 2 degrees are employed (denoted as model $P + P$); 4) one causal region has linear additive effects and the other causal region has non-linear effects (denoted as model $L + R$); and 5) one causal region has linear effects and the other causal region have pair-wise interaction effects (denoted as $L + P$). The details of the simulation settings are summarized in Supplementary Table B.3. We simulated 5,000 regions for each disease model setting.

Similar to Scenario I, we reported the Pearson correlations and MSEs based on testing samples, and calculated the efficiencies of variable screening. However, unlike the first set of simulations which only considered linear kernel for the MKLMM method, we used the adaptive setting in this set of simulations, where a linear kernel, a RBF kernel, a polynomial kernel of 2 degrees and a neural network kernel are included in the candidate kernel set and the most appropriate kernels are selected in a data-driven manner.

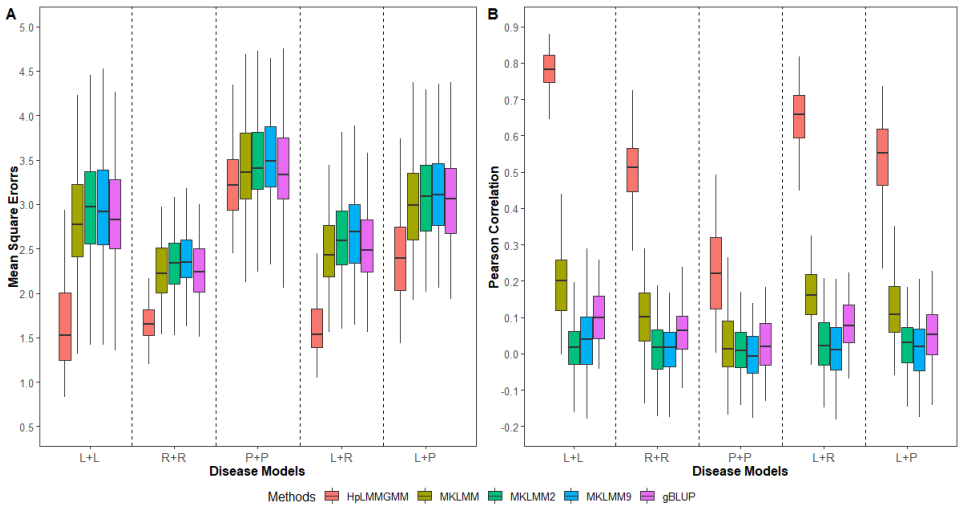

FIGURE 3.2: The impact of disease models. $L + L$: genetic variants on both regions have linear additive effects. $R + R$: predictors from both regions have non-linear predictive effects. $P + P$: both regions harbor variants with pair-wise interaction effects. $L + R$: genetic variants on the first and second regions have linear additive and non-linear effects, respectively. $L+P$: predictors on the first and second regions have linear additive and pair-wise interaction effects, respectively ($n = 500$)

As shown in Figure 3.2 and Supplementary Figure B.2, the proposed HpLMMGMM performed the best among all simulations considered. Our method had the lowest MSEs and highest Pearson correlations regardless of disease models, indicating that our method can maintain robust performance across a range of phenotypes. Comparing the other methods, adaptive MKLMM and gBLUP performed better than adaptive MKLMM with 2 or 9 chosen regions. While the screening rule employed by adaptive MKLMM is designed to reduce the impact of noise as well as improve the computational efficiencies of LMMs, it can mistakenly discard predictive regions, leading to low prediction performance. As shown in Table 3.2 and Supplementary Table B.4, when the number of chosen regions is 2 or 9, the screening process employed by adaptive

MKLMM can barely keep any causal regions. This explains why adaptive MKLMM2 and adaptive MKLMM9 have even worse prediction performance than gBLUP that keeps all genetic variants. Comparing gBLUP with adaptive MKLMM that keeps the 5% of regions with the highest likelihood ratio test statistics, adaptive MKLMM has better performance in all of the disease models except the $P + P$ model. This is primarily because gBLUP keeps all measured genetic variants in the prediction modeling and thus is not able to reduce the impact of noise, leading to a less accurate prediction model. For the $P + P$ model, adaptive MKLMM is quite likely to screen out some of the causal regions (Table 3.2 and Supplementary Table B.4), resulting in reduced prediction accuracy. Comparing adaptive MKLMM with 5% regions chosen with our designed hybrid rule, our method has much higher probability of correctly teasing out the noise while maintaining the causal regions in the selected set, regardless of the underlying disease models. Indeed, the number of kept regions after screening is only 12 for our method, whereas it can go up to 295 regions for adaptive MKLMM with 5% regions chosen (Table 3.2), on average. Including lots of noise regions in the final prediction analysis not only increases the computational burden for LMMs, but also reduces the robustness and accuracy of the prediction.

TABLE 3.2: The number of selected regions and the number of causal regions within the selected regions under different disease models ($n = 500$)

Disease Models	Number of Total Regions Selected (Number of Causal Regions Selected)			
	MKLMM	MKLMM2	MKLMM9	Hybrid Screening Rule
$S_1 : L + L$	294.12 (2.00)	2 (0)	9 (0)	12.11 (1.96)
$S_2 : R + R$	295.02 (1.97)	2 (0)	9 (0)	13.88 (1.88)
$S_3 : P + P$	295.54 (1.25)	2 (0)	9 (0)	8.95 (1.92)
$S_4 : L + R$	294.21 (1.97)	2 (0)	9 (0)	13.77 (1.83)
$S_5 : L + P$	294.11 (1.62)	2 (0)	9 (0)	10.45 (1.96)

Our proposed hybrid screening rule can capture non-linear effects and shed light on the underlying relationships through selecting appropriate kernel functions from the candidate set. The probability of selecting the most appropriate kernels for both adaptive MKLMM and our method is shown in Table 3.3 and Supplementary Table B.5 . Note that we did not report the screening results of kernel selection for adaptive MKLMM2 and adaptive MKLMM9, since neither method could correctly detect any predictive regions. Although adaptive MKLMM is designed to capture non-additive effects through a data-driven manner, it can barely select any non-linear kernels from

the candidate set in practice. For example, for the disease model $P + P$ where both causal regions have pair-wise interaction effects, the probability of selecting the polynomial kernel for adaptive MKLMM is close to 0%, whereas it is above 90% for the proposed hybrid screening rule. Similarly, under the disease model $R+R$, the screening rule employed by adaptive MKLMM mainly chooses linear kernel for prediction ($\approx 0\%$ for using the RBF kernel), whereas our method has over 85% of chance to use the most appropriate kernels. Therefore, the proposed method has better capability of capturing non-linear effects, leading to a prediction model with much higher accuracy.

TABLE 3.3: The chance of selecting the most appropriate kernels under different disease models ($n = 500$)

Disease	MKLMM		Hybrid Screening Rule	
Models	1st Causal Region	2nd Causal Region	1st Causal Region	2nd Causal Region
$S_1 : L + L$	1.00	1.00	0.98	0.98
$S_2 : R + R$	0	0	0.87	0.85
$S_3 : P + P$	0	0	0.94	0.98
$S_4 : L + R$	0.99	0	1.00	0.73
$S_5 : L + P$	0.97	0	1.00	0.96

*Note: results from MKLMM2 and MKLMM9 are not reported. Neither of the causal regions can be kept, and thus the chances of selecting the most appropriate kernels are 0.

3.4 Real data application

AD is a common neurodegenerative condition and accounts for 60% to 70% of dementia cases; it is becoming a growing health problem worldwide due to population aging (Cuingnet et al., 2011). An early and accurate diagnosis of AD has been widely regarded as a critical step for AD treatment. The ADNI study, which was launched in 2003 by multiple public and private organizations, has provided great opportunities for systematic investigations of AD (Cuingnet et al., 2011). ADNI has collected clinical assessment, biochemical biomarkers, and results from magnetic resonance imaging and positron emission tomography (PET) for early and accurate diagnosis of AD (Mueller et al., 2005).

In this study, we are interested in using whole-genome sequencing data to predict baseline PET-imaging outcomes, including AV45 and FDG. We excluded individuals who do not have genetic data and/or are missing baseline phenotype measurements. A total of 639 and 501 samples were kept for the analyses of FDG and AV45, respectively. The distributions of these phenotypes are shown in Supplementary Figure

B.3. Whole-genome sequencing of the genomic data has been performed on the Illumina HiSeq2000 at a non-Clinical Laboratory Improvements Amendments (non-CLIA) laboratory (Saykin et al., 2015). We removed genetic variants with more than 1% of missing rate and annotated them based on GRch37 assembly. A total of 20,350 genes with 8,041,596 genetic variants and 20,617 genes with 7,702,415 variants were analyzed for FDG and AV45, respectively. To avoid over-fitting, we randomly set 100 individuals as testing samples and used the remaining samples to train the predictive models. We evaluated the prediction accuracy based on the testing samples. To reduce the risk of chance finding, this process was repeated 100 times.

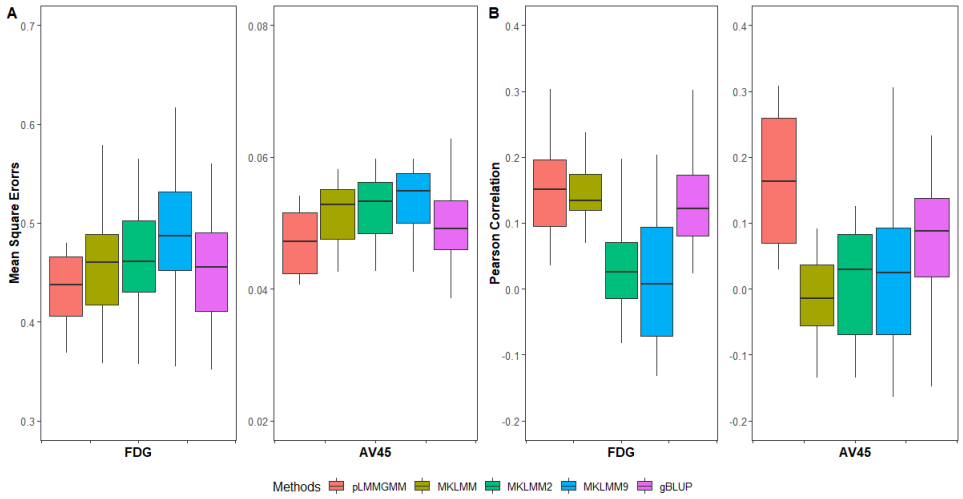

FIGURE 3.3: Accuracy comparisons for FDG and AV45

The prediction accuracies for FDG and AV45 are shown in Figure 3.3. For FDG, MKLMM2 and MKLMM9 were the worst performers of all the methods. MKLMM had a similar performance to that of the gBLUP method. For AV45, MKLMM was the worst performer regardless of the pre-specified number of region. For both FDG and AV45, HpLMMGMM performed the best, having the lowest MSEs and highest Pearson correlations of all the methods. These results clearly indicate that correctly detecting predictive regions and simultaneously modeling their linear and non-linear effects can improve the robustness and accuracy of a prediction model.

The MKLMM that included the top 5% of regions selected about 1700 genes from an initial data set of more than 20,000 genes in each replicate, and about 400 genes were consistently selected among 100 replicates. However, the prediction accuracy of MKLMM is generally lower than the proposed method (HpLMMGMM), and this suggests that a substantial number of the genes selected by MKLMM could still be

noise, leading to reduced the method's accuracy. Furthermore, most the genes selected by MKLMM have not previously been reported to be associated with AD. Indeed, some well-known AD-related genes had a low probability of being selected by MKLMM. For example, the probability of detecting *APOC1* *APOE* and *TOMM40* genes were close to 0%. Similarly, there is little evidence in the current literature that the genes kept by MKLMM2 and MKLMM9 are actually associated with AD. Indeed, these selected genes have almost no prediction power, as shown in Figure 3.3. Unlike MKLMM, the proposed hybrid screening rule aligns well with the downstream prediction tasks and tends to keep only a small proportion of genes with most of them being predictive. In particular, we found there was a high probability that well-known AD-related genes are kept by the proposed screening rule. For example, *APOC1*, *APOE* and *TOMM40* genes can be selected by more than 96% of the time among 100 replicates.

The final prediction model HpLMMGMM used was the penalized LMM with GMM estimators, and thus the model had the capability to further fine-tune the gene set determined by the hybrid screening rule and so improve its predictive powers. Supplementary Tables B.6 and B.7 list the genes that were selected by the final prediction model at least once for FDG and AV45, respectively. For FDG, a total of 120 genes were selected by the final prediction model at least once among 100 replicates. *APOE*, *APOC1* and *FADS3* genes were selected more than 95% of the time, whereas the remaining 117 genes were averagely selected less than 7% of the time. For AV45, a total of 124 genes were selected by the final prediction model at least once among 100 replicates. *APOE*, *APOC1* and *TOMM40* were selected more than 94% of the time, whereas the others were selected less than 5% of the time, on average. From our genome-wide analysis with more than 20,000 genes considered, all the most commonly selected genes for both FDG and AV45 are well-known AD-related genes. For example, both *APOC1* and *APOE* were highly selected for both FDG and AV45 by our HpLMMGMM method (i.e., $> 94\%$). Converging evidence has show that these two genes are major genetic risk factors for AD (Ossenkoppele et al., 2013; Roses, 2010). *APOE* $\epsilon 4$ is a major risk factor for late-onset AD (LOAD) and an increasing number of *APOE* $\epsilon 4$ alleles increases the LOAD risk (Corder et al., 1993). *APOC1* participates in the biological processes of cholesterol metabolism, whose deterioration has an important impact on the development of AD (Zhou et al., 2014b). The $rs11568822$ polymorphism on *APOC1* also increases the risk of AD in Caucasians, Asians and Caribbean Hispanics (Zhou et al., 2014b). In addition to these two genes which are predictive for both FDG and AV45, *FADS3* and *TOMM40*, which are respectively predictive

for FDG and AV45, have also been reported to be associated with AD. The *FADS3* gene has been found to be related to AD via long-chain polyunsaturated fatty acids (LCPUFA), which are involved in the pathophysiology of neurodegenerative diseases, including AD (Schuchardt et al., 2016). *TOMM40* encodes a mitochondrial protein whose dysfunction is closely related to aging diseases (e.g, LOAD) (Bagnoli et al., 2013).

3.5 Discussion

In this work, we have designed a hybrid screening rule and further incorporated it into a penalized LMM with GMM estimators for prediction analysis on genome-wide data. The proposed HpLMMGMM first splits the genome into multiple regions based on various criteria (e.g., gene and pathway annotations), where the complex predictive effects (i.e., linear and non-linear effects) for each region are captured by multiple kernel functions. We then developed a hybrid screening rule to reduce the data dimension, where the number of inputs for the downstream prediction model (i.e., penalized LMM with GMM estimators) has been substantially reduced, making our method applicable to the analysis of genome-wide data. Through extensive simulation studies and the analysis of the ADNI data set, we have demonstrated that our method can: 1) be applied to genome-wide data; 2) efficiently capture both linear and non-linear predictive effects; and 3) have robust and accurate prediction performance across a range of phenotypes and dimensions of input data.

Whole-genome data are high-dimensional and contain a large number of non-relevant variables. Therefore, variable selection has been regarded as an indispensable step in the analysis of whole-genome data. Variable selection can not only reduce the impact of noise from high-dimensional data, but also improves the computational efficiencies that are of great importance for large data analyses (Byrnes et al., 2013). Existing LMM-based models that can be applied to genome-wide data either ignore the impact of noise and treat all genetic variants in a similar manner (e.g., gBLUP (Yang et al., 2010)) or employ empirical screening rules for variable selection (e.g., MultiBLUP and MKLMM (Speed and Balding, 2014; Weissbrod et al., 2016)). Therefore, their performance depends on the underlying disease models and the amount of noise in the data, leading to less robust prediction performance. For example, MKLMM with the top 5% of regions selected achieved a similar level of prediction performance as gBLUP for FDG (i.e., the median of the Pearson correlations were all around 0.12),

but it had almost no predictive power for AV45, where the median of the Pearson correlations were -0.02 and 0.09 for MKLMM and gBLUP, respectively. While some existing penalized LMMs intend to improve the robustness of prediction via enabling the variable selection within the prediction modeling, they can only handle a limited number of regions and it is computationally prohibitive to apply them to genome-wide data (Li et al., 2020; Wen and Lu, 2020). This is partially due to the fact that their parameter estimations rely on MLE/REML, which can be extremely computational demanding even for a moderate number of random effects. Unlike existing methods, the proposed HpLMMGMM incorporated a hybrid screening rule into the prediction modeling framework, where the data dimension is reduced substantially before building the prediction model and the parameter estimation in the final model relies on a much more computationally efficient GMM estimator. The designed hybrid screening rule aligns well with the downstream prediction task, and the discarded variables are guaranteed to have no predictive power in the corresponding penalized LMM with parameters estimated by GMM. This property makes HpLMMGMM much more appealing in modeling high-dimensional data, as it can keep the number of variables at a manageable size while maintaining the same level of prediction performance. As shown in the Scenario I, as the number of regions increase from 5,000 to 20,000 (i.e., genome-wide level), the number of regions kept by the designed hybrid screening rule remained small (i.e., about 16), whereas the number of regions increased from 293 to 1,178 for the empirical screening rule adopted by MKLMM (Table 3.1). Correspondingly, the prediction accuracy for HpLMMGMM remained relatively stable, whereas it decreased substantially for MKLMM as the number of regions increases. Furthermore, our proposed hybrid screening rule has very high chance of screening out the noise regions. Indeed, HpLMMGMM detected the true causal regions more than 99.25% of the time and incorrectly identified noise regions as causal only 0.15% of the time. Therefore, our proposed HpLMMGMM can effectively screen out noise using its designed hybrid screening rule, and the remaining regions can be jointly modeled by the prediction model (i.e., penalized LMM with GMM estimator) employed in HpLMMGMM. This makes HpLMMGMM capable of modeling genome-wide data while maintaining high and robust prediction accuracy (Figure 3.1 and Supplementary Figure B.1).

The underlying etiology for human diseases can be quite complex and it is unknown in advance. Therefore, traditional LMMs (e.g., gBLUP and MultiBLUP) that only focus on capturing linear additive effects can have reduced prediction accuracy when other types of predictive effects are present (Speed and Balding, 2014; VanRaden,

2008; Yang et al., 2010). Multi-kernel LMM is designed to capture complex effects through utilizing various kernel functions. However, in practice, Multi-kernel LMM tends to select only linear kernels even when non-linear effects truly exist (Table 3.3 and Supplementary Table B.5). In the contrast, although our method also applies multiple kernels to capture various types of predictive effects, it has a much higher chance of choosing the most appropriate kernels, which can shed light on the underlying disease model. For example, for disease model $L + P$ where one of the causal regions only has pair-wise interaction effects, the hybrid screening rule has 96% probability of using polynomial kernels that clearly suggest the presence of non-linear effects for the prediction, whereas adaptive MKLMM always chooses to use the linear kernel. Similarly, for model $L+R$, our hybrid screening has 73% chance of using the RBF kernel for prediction, whereas it is close to 0% for adaptive MKLMM. In practice, having the ability to choose the most appropriate kernels for prediction can lead to robust and accurate performance across a range of phenotypes with different underlying genetic architecture.

In the prediction analyses of FDG and AV45, HpLMMGMM achieved a better prediction performance than those of commonly used methods (Figure 3.3). The designed hybrid screening rule can substantially reduce the impact of noise and the multiple kernels used in HpLMMGMM can capture both linear and non-linear predictive effects, both of which have facilitated the prediction modeling. The empirical screening rule used by MKLMM selected a large number of genes (i.e., about 1700 genes, on average) in each replicate and about 400 genes were constantly kept among 100 replicates. As shown in Figure 3.3, these selected genes may include a large number of noise, which greatly reduces the accuracy of the prediction. Indeed, well-known AD-related genes (e.g.,*APOE* and *APOC1*) were barely selected. In contrast, the hybrid screening rule selected, on average, 34 genes and 36 genes in each replicate for FDG and AV45, respectively. Moreover, from 100 replicates, only 5 genes and 3 genes were detected more than 90% of the time for FDG and AV45, and most of these selected genes have been reported to be AD-related genes in existing studies (i.e.,*APOE*, *APOC1* and *TOMM40*), which substantially improves the prediction accuracy of our method (Figure 3.3). The input for final prediction models is determined by screening rules, and thus its effectiveness can substantially affect the prediction performance and the interpretation of corresponding models.

The genes kept by the screening rule are not guaranteed to be predictive, and thus the HpLMMGMM prediction framework utilizes a penalized LMM that can further

fine-tune the set of predictive genes for improved prediction modeling. All the genes that were consistently selected by HpLMMGMM have been previously reported to be associated with AD. For example, the *APOE* gene has been identified as the major susceptibility gene for AD (Poirier et al., 1993; Strittmatter et al., 1993). It encodes apolipoprotein E which is related to cholesterol metabolism, and mounting evidence has shown that the accumulation of cholesterol in the brain (Puglielli et al., 2003) is associated to AD. Furthermore, other studies have found that the *APOE* $\epsilon 4$ allele is overrepresented among AD patients (Poirier et al., 1993; Strittmatter et al., 1993). For example, Zhou et al., 2014a reported that *APOE* $\epsilon 4$ allele increases the risk of cognitive decline in Chinese LOAD patients. Saunders et al., 1993 also found a strong connection between the *APOE* $\epsilon 4$ allele and the risk of AD. The *APOC1* gene encodes apolipoprotein C1, which is also a member of apolipoprotein family, and thus it affects AD in a similar way as *APOE*. Indeed, many studies suggest *APOC1* along with *APOE* affect the risk of AD (Ki et al., 2002; Lucatelli et al., 2011; Shi et al., 2004). For example, Bertram et al., 2007 proposed that $rs11568822$ on the *APOC1* gene is associated with AD due to the linkage disequilibrium between the *APOC1* with *APOE*. The *APOC1 Hpal+* variant has been found to be associated with AD in Caribbean Hispanic individuals, mainly due to strong linkage disequilibrium between the *APOC1* and the *APOE* $\epsilon 4$ allele (Tycko et al., 2004). In addition, *APOC1* can be independently related to AD. For example, the *APOC1 H2* allele was found to be significantly associated with LOAD in the Korean population (Ki et al., 2002). The *FADS3* gene has been reported to be associated with aging diseases via affecting LCPUFA metabolism, which plays a key role in neuronal membrane integrity and function within the brain (Schuchardt et al., 2016). Specifically, $rs174455$ on *FADS3* was found to be associated with LCPUFA metabolism and further affects the cognitive ability of patients with mild cognitive impairment (MCI) (Schuchardt et al., 2016). *TOMM40* is one of the candidate genes related to the pathogenesis of AD (Ma et al., 2013). Recent, several SNPs on the *TOMM40* gene have been found to be significantly associated with AD in genome-wide association studies (Kim et al., 2011; Potkin et al., 2009; Shen et al., 2010). For example, Bagnoli et al., 2013 found the association of $rs157581$ on *TOMM40* with AD in Italian population and Huang et al., 2016 found the association between $rs2075650$ on *TOMM40* and AD risk for Caucasian and Asian patients.

In summary, we have incorporated a hybrid screening rule into a penalized LMM with GMM estimators for risk prediction analyses on high-dimensional genetic data.

Our proposed HpLMMGMM method can be applied to genome-wide data and efficiently captures both linear and non-linear predictive effects, leading to a better prediction model across a range of phenotypes. HpLMMGMM can effectively select variables from high-dimensional genome-wide data but ignore the large sample size (e.g.,10,000) problem. This will be the future direction of our research.

Chapter 4

A penalized linear mixed model with generalized method of moments estimators for the prediction analysis of multi-omics data

4.1 Introduction

Accurately predicting disease risk, which can facilitate the delivery of tailored treatments, plays a key role towards precision medicine (Ashley, 2015). Recent emerging high-dimensional multi-layer omics data (e.g., genome, transcriptome, methylome and proteome data) has provided unprecedented opportunities to comprehensively investigate the role of a deep catalogue of predictors in disease risk prediction (Boekel et al., 2015). However, the complex relationships among multi-layer omics data and their high-dimensionality have brought tremendous analytical and computational challenges (Morris and Baladandayuthapani, 2017; Ritchie et al., 2015; Zeng and Lumley, 2018).

Existing integrative methods are mainly designed for discovering coherent patterns among multi-omics data (Bersanelli et al., 2016; Huang et al., 2017; Morris and Baladandayuthapani, 2017; Zeng and Lumley, 2018). For example, the non-negative matrix factorization method (Zhang et al., 2012) projects multi-omics data onto a common basis space so that their consistent information can be captured. Canonical correlation analysis, an exploratory multivariate analysis tool, finds linear combinations of all variables within each omics data that maximize the correlations between each canonical variate pair. Therefore, the most expressive elements of canonical vectors reflect the relationships among omics data. Partial least squares utilizes a similar idea, but

considers covariance rather than correlation (Chen and Zhang, 2016). To further consider prior biological knowledge, Bayesian models have been introduced for the integrative analysis of omics data (Huang et al., 2017; Wang et al., 2012). For example, Integrative Bayesian Analysis of Genomics (iBAG) developed by Wang et al., 2012, integrates gene expression and methylation data in the Bayesian framework to explore their associations with clinical outcomes. Wang et al., 2019 proposed the integrative risk gene selector (iRIGS), a Bayesian framework that integrates multi-omics data and gene networks, to select risk genes from genome wide association studies. Recently, network-based methods, which can reflect complex inter-relationships in a network and facilitate model interpretation, have been used in the integrative analysis (Huang et al., 2017; Zhou et al., 2020). For example, similarity network fusion method proposed by Wang et al., 2014a constructs a sample-by-sample similarity matrix from each data type, and then uses a graph diffusion algorithm to fuse these similarity matrices into a comprehensive network that is further used for patient detection. Lemon-Tree, an integrative multi-omics network analysis, first finds co-expressed gene clusters, and then reconstructs regulatory programs that include a set of regulator genes as network modules by fuzzy decision trees. Finally, a probabilistic score is calculated for each regulatory program, and the ones with high probabilistic scores are selected as potential disease drivers (Bonnet et al., 2015). Although the existing integrative analysis has facilitated the detection of coherent patterns embedded in multi-omics data, they usually focus on a particular gene/pathway and thus cannot be directly applied to the analysis of high-dimensional multi-omics data.

Complex human diseases/traits manifest themselves at various molecular levels and they are usually regulated by a number of pathways (Subramanian et al., 2020). Therefore, jointly modeling a large number of predictors at various molecular levels while accounting for their complex inter-relationships is a critical step for an accurate prediction model (Morris and Baladandayuthapani, 2017). While high-dimensional multi-layer omics data have provided the essential information, their ultra-high dimensionality has made it computationally challenging to jointly analyze them. Existing integrative methods usually only focus on specific genes or pathways, and they are mainly designed for detecting disease-associated variables. For example, Meng et al., 2014 integrated transcriptomic and proteomic data in the NCI-60 cancer cell line panel and found that the leukemia extravasation signaling pathway is highly related to metastasis in leukemia cell lines. Vaske et al., 2010 showed that the estrogen- and ErbB2-related pathways are associated with breast cancer through integrating copy number variations, gene

expression and DNA methylation data. While existing integrative methods have shed light on the underlying disease etiology, they can only model a limited number of variables (e.g., one specific pathway) and thus cannot directly be applied for prediction analyses. This is mainly because an accurate risk prediction model requires the joint consideration of a large number of predictors from multiple candidate pathways, and utilizing information from only one disease-associated pathway is unlikely to produce an accurate prediction model. For example, immune response, lipid metabolism and cell differentiation pathways are all associated Alzheimer's disease (AD). Using information from the immune response pathway alone is not enough to accurately predict AD risk. Therefore, an integrative method that can simultaneously model a large number of variables from different layers of omics data is urgently needed for prediction research.

Linear mixed models (LMMs) have great potential in modeling high-dimensional multi-omics data. Indeed, LMMs have already long been used for prediction analysis on high-dimensional genomic data (Speed and Balding, 2014; VanRaden, 2008; Weissbrod et al., 2016; Yang et al., 2010). For example, the genomic best linear unbiased prediction (gBLUP) method uses a single random effect term to model cumulative predictive effects from all measured genetic variants (Yang et al., 2010). Both MultiBLUP and multi-kernel LMM adopt multiple random effect terms to estimate the joint predictive effects from multiple genetic regions with each harboring many variants (Speed and Balding, 2014; Weissbrod et al., 2016). Recently, to account for non-linear predictive effects, Wen and Lu, 2020 introduced a penalized multi-kernel LMM, where kernel functions are used to model complex jointly predictive effects from multiple genetic variants and penalization is used to select predictive regions. The basic rationale for these LMM-based models is that genetically similarly individuals can have similar phenotypes. Therefore, instead of estimating effect sizes for each genetic variant, LMMs aim at capturing cumulative predictive effects from a large number of predictors through their estimated genetic similarity, which can substantially reduce the number of model parameters, making it applicable for the analysis of genome-wide data. A similar idea can be applied for the prediction analysis of multi-omics data, where genetic similarities are replaced by omic-similarities that can be measured by various kernel functions.

While LMM-based models are promising for the analysis of high-dimensional multi-layer omics data, they can have limited predictive power if a large amount of noise is

present. Recent work has shown that excluding noise when estimating genetic similarities can not only facilitate model interpretation, but also improve the robustness and accuracy of a prediction model. Adding an L_1 penalty to the objective function is a commonly adopted approach to reduce the impact of noise. For example, Wen and Lu, 2020 proposed a penalized multi-kernel LMM to predict phenotypes based on high-dimensional genomic data, and Li et al., 2020 extended this method for the prediction analysis on multi-omics data. While these methods have improved the accuracy of prediction models, their parameter estimation can be extremely computationally demanding. This is mainly because for penalized LMMs, obtaining the maximum likelihood estimator (MLE) or the restricted maximum likelihood estimator (REML) (Weissbrod et al., 2016; Wen and Lu, 2020), which are usually estimated by Newton-Raphson or expectation-maximization algorithms, is computationally expensive. Generalized method of moments (GMM) is a promising alternative for the estimation of variance components for penalized LMMs, as it can change the objective function into a quadratic form that is much easier to optimize (Rao, 1970; Rao, 1971a; Rao, 1972). For example, Zhu and Weir, 1996 used the minimum norm quadratic unbiased estimation method to estimate variance components for maternal and paternal effects in a bio-model for diallel crosses. We recently developed a GMM-based LMM for the prediction analysis of genomic data, where we showed that the GMM-based estimators can accurately detect prediction genetic regions and improved the prediction accuracy of LMM-based prediction models (Wang and Wen, 2021).

In this paper, we propose a penalized LMM with GMM estimators (MpLMMGMM) for the prediction analysis of multi-omics data. The proposed MpLMMGMM model can: 1) account for complex inter/intra-relationships among multi-omics data; 2) detect predictive biomarkers; and 3) substantially reduce the computational cost of penalized LMMs. In the following sections, we first present the MpLMMGMM method and then compare its prediction accuracy with commonly used methods (i.e., OmicKrig) through simulation studies. Finally, we use the proposed method to analyze the multi-omics data obtained from the ADNI (Saykin et al., 2010).

4.2 Methods

4.2.1 A linear mixed model for prediction analysis using multi-omics data

Suppose we have a sample of n individuals. Let $\boldsymbol{Y}$ be the $n \times 1$ outcome vector and $\boldsymbol{X}_d$ be a $n \times P_d$ matrix of demographic variables (e.g., age and gender). We split the genome into R sets that can be defined by various criteria (e.g., gene and pathway annotations), and use $\boldsymbol{O}_i$ to denote the joint predictive effects from all predictors in the ith set. We model the outcomes as:

$$\boldsymbol{Y} = \boldsymbol{X}_d\boldsymbol{\beta}_d + \sum_{i=1}^{R} \boldsymbol{O}_i + \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{\epsilon} \sim N(0, \sigma_0^2 \boldsymbol{I}_n) \tag{4.1}$$

For notation simplicity and without loss of generality, we use the gene annotation to define the set and only considered gene expression, genomic and methylation data. Correspondingly, equation 4.1 can be written as:

$$\boldsymbol{Y} = \boldsymbol{X}_d\boldsymbol{\beta}_d + \sum_{i=1}^{R} \boldsymbol{e}_i + \sum_{i=1}^{R} \boldsymbol{g}_i + \sum_{i=1}^{R} \boldsymbol{m}_i + \sum_{i=1}^{R} \boldsymbol{O}_i^{inter} + \boldsymbol{\epsilon} \tag{4.2}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma_0^2 \boldsymbol{I}_n)$. $\boldsymbol{e}_i$, $\boldsymbol{g}_i$, $\boldsymbol{m}_i$, and $\boldsymbol{O}_i^{inter}$ are predictive effects of gene expression data, genomic data, methylation data and their interactions in set i. Similar to LMM-based models designed for the analysis of genomic data (Weissbrod et al., 2016), we assume individuals with similar molecular profiles have similar phenotypes, and model the joint predictive effects from a large number of predictors within each omics layer using random effect terms, where $\boldsymbol{g}_i \sim N(0, \boldsymbol{K}_{g,i}\sigma_{g,i}^2)$, $\boldsymbol{m}_i \sim N(0, \boldsymbol{K}_{m,i}\sigma_{m,i}^2)$ and $\boldsymbol{O}_i^{inter} \sim N(0, \boldsymbol{K}_{inter,i}\sigma_{inter,i}^2)$. Here $\boldsymbol{K}_{g,i}$, $\boldsymbol{K}_{m,i}$ and $\boldsymbol{K}_{inter,i}$ respectively measure the similarities among genomic data, methylation data and their interactions for the set i. While the predictive effects from gene expression data can also be modeled in a similar fashion, we propose to use the fixed effect defined as $\boldsymbol{e}_i = \boldsymbol{E}_i \times \gamma_i$ instead, where $\boldsymbol{E}_i$ represents the gene expression level for the set i and γ_i is the corresponding effect. This is mainly because when the number of predictors within the set is very limited, using a fixed effect term is more efficient than the corresponding random effect model. Therefore,

equation 4.2 can be written as:

$$\boldsymbol{Y} = \boldsymbol{X}_d\boldsymbol{\beta}_d + \sum_{i=1}^{R}\boldsymbol{E}_i\gamma_i + \sum_{i=1}^{R}\boldsymbol{g}_i + \sum_{i=1}^{R}\boldsymbol{m}_i + \sum_{i=1}^{R}\boldsymbol{O}_i^{inter} + \boldsymbol{\epsilon} \tag{4.3}$$

where $\boldsymbol{g}_i \sim N(0, \boldsymbol{K}_{g,i}\sigma^2_{g,i})$, $\boldsymbol{m}_i \sim N(0, \boldsymbol{K}_{m,i}\sigma^2_{m,i})$, and $\boldsymbol{O}_i^{inter} \sim N(0, \boldsymbol{K}_{inter,i}\sigma^2_{inter,i})$.

The proposed modeling framework is very flexible and can accommodate various disease model assumptions. For example, if only linear effects from all omics layers are considered, then both genomic and methylation similarities can be measured using linear kernels, $\boldsymbol{K}_{g,i} = \boldsymbol{G}_i\boldsymbol{G}_i^T/p_{g,i}$ and $\boldsymbol{K}_{m,i} = \boldsymbol{M}_i\boldsymbol{M}_i^T/p_{m,i}$, $\forall i \in \{1, \cdots, R\}$, where $\boldsymbol{G}_i$ and $\boldsymbol{M}_i$ are $n \times p_{g,i}$ genotype and $n \times p_{m,i}$ methylation matrices for set i, respectively. By using linear kernels, our proposed model is equivalent to:

$$\boldsymbol{Y} = \boldsymbol{X}_d\boldsymbol{\beta}_d + \sum_{i=1}^{R}\boldsymbol{E}_i\gamma_i + \sum_{i=1}^{R}\sum_{j=1}^{p_{g,i}}\boldsymbol{G}_{ij}\gamma^g_{ij} + \sum_{i=1}^{R}\sum_{j=1}^{p_{m,i}}\boldsymbol{M}_{ij}\gamma^m_{ij} + \boldsymbol{\epsilon}$$

where $\gamma^g_{ij} \sim N(0, \sigma^2_{g,i}/p_{g,i})$, $\gamma^m_{ij} \sim N(0, \sigma^2_{m,i}/p_{m,i})$, $\boldsymbol{\epsilon} \sim N(0, \sigma^2_0\boldsymbol{I}_n)$, $\boldsymbol{G}_{ij}$ ($\boldsymbol{M}_{ij}$) is the jth column of $\boldsymbol{G}_i$ ($\boldsymbol{M}_i$), and γ^g_{ij} (γ^m_{ij}) is their corresponding effect. Similarly, if only pairwise interaction between genomic and methylation is considered, then we can set $\boldsymbol{O}_i^{inter} = \boldsymbol{K}_{g,i} \circ \boldsymbol{K}_{m,i}$, where $\circ$ is the Hadamard product.

4.2.2 A penalized linear mixed model with generalized method of moments estimators using multi-omics data

Recent work has indicated that not all measured variables from multi-omics data are predictive (Li et al., 2020; Wen et al., 2016; Wen and Lu, 2020), and thus variable selection can be of great importance for the robustness and accuracy of a prediction model (Byrnes et al., 2013). Adding an L_1 penalty into the objective function is a commonly adopted approach for simultaneous variable selection and parameter estimation (Li et al., 2020; Wen and Lu, 2020; Wu et al., 2009). For high-dimensional multi-omics data, it is essential to perform variable selection at each omics layer. Therefore, we proposed to add an L_1 penalty on both the fixed effect (e.g., for the selection of gene expression data) and random effect terms (e.g., for the selection of genomic and methylation data). While REML is widely used to estimate parameters for LMMs (Speed and Balding, 2014; VanRaden, 2008; Yang et al., 2010), it is computationally expensive, especially for LMMs with a large number of random effects. Indeed, it is

computationally prohibitive to consider a large number of random effects for REML and MLE. Therefore, following a similar idea in Wang and Wen, 2021, we proposed to use the GMM to estimate model parameters, and thus the objective function for model 4.3 can be written as:

$$(\hat{\boldsymbol{\beta}}_d, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\sigma}}^2) = \underset{\boldsymbol{\beta}_d, \boldsymbol{\gamma}, \boldsymbol{\sigma}^2}{\operatorname{argmin}} \frac{1}{2} ||\boldsymbol{Z}\boldsymbol{Z}^T - \sum_{i=1}^{R} \sum_{j \in (g,m)} \boldsymbol{K}_{j,i} \sigma_{j,i}^2 - \sigma_0^2 \boldsymbol{I}_n||_F^2 + \lambda_1 \sum_{i=1}^{R} \sum_{j \in (\mathrm{g},\, \mathrm{m})} \sigma_{j,i}^2 + \lambda_2 \sum_{i=1}^{R} |\gamma_i| \tag{4.4}$$

where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_R)$; $\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{X}_d \boldsymbol{\beta}_d - \sum_{i=1}^{R} \boldsymbol{E}_i \gamma_i$; $\lambda_i > 0, i \in \{1, 2\}$ is the penalty; and $\boldsymbol{\sigma}^2 = (\sigma_0^2, \sigma_{g,1}^2, \cdots, \sigma_{g,R}^2, \sigma_{m,1}^2, \cdots, \sigma_{m,R}^2)$.

We used an iterative procedure to estimate parameters in the random (i.e., $\boldsymbol{\sigma}^2$) and fixed effects (i.e., $\boldsymbol{\beta}_d$ and $\boldsymbol{\gamma}$). During iteration step $t+1$, we first updated the random effect term as:

$$\hat{\boldsymbol{\sigma}}^{2,t+1} = \underset{\boldsymbol{\sigma}^2 \geq 0}{\operatorname{argmin}} \frac{1}{2} ||\boldsymbol{Z}_t \boldsymbol{Z}_t^T - \sum_{i=1}^{R} \sum_{j \in (g,m)} \boldsymbol{K}_{j,i} \sigma_{j,i}^2 - \sigma_0^2 \boldsymbol{I}_n||_F^2 + \lambda_1 \sum_{i=1}^{R} \sum_{j \in (\mathrm{g},\, \mathrm{m})} \sigma_{j,i}^2, \quad \lambda_1 > 0 \tag{4.5}$$

where $\boldsymbol{Z}_t = \boldsymbol{Y} - \boldsymbol{X}_d \boldsymbol{\beta}_d^t - \sum_{i=1}^{R} \boldsymbol{E}_i \gamma_i^t$. Given the parameter estimates for the random effect term during step $t+1$, we updated the parameters associated with fixed effects as:

$$(\hat{\boldsymbol{\beta}}_d^{t+1}, \hat{\boldsymbol{\gamma}}^{t+1}) = \underset{\boldsymbol{\beta}_d, \boldsymbol{\gamma}}{\operatorname{argmax}} -\frac{1}{2} \log |\boldsymbol{\Sigma}_{t+1}| - \frac{1}{2} \boldsymbol{Z}^T \boldsymbol{\Sigma}_{t+1}^{-1} \boldsymbol{Z} - \lambda_2 \sum_{i=1}^{R} |\gamma_i|, \quad \lambda_2 > 0 \tag{4.6}$$

where $\boldsymbol{\Sigma}_{t+1} = \sum_{i=1}^{R} \sum_{j \in (g,m)} \boldsymbol{K}_{j,i} \sigma_{j,i}^{2,t+1} + \sigma_0^{2,t+1} \boldsymbol{I}_n$. The details of the proposed estimation procedure is shown in algorithm 2.

Compared to penalized LMMs that rely on REML estimators, our proposed objective function during each of the iteration step is much easier to optimize. Therefore, our proposed algorithm is computationally efficient. As opposed to existing LMMs that can only consider a limited number of random effects (i.e., usually ≤ 10 (Wang

and Wen, 2021)), our proposed method can jointly consider a large number of regions (i.e., random effects) and efficiently detect those that are predictive.

Algorithm 2 Procedure for the parameter estimation

Initialization: at step $t = 0$:

Set $\boldsymbol{\sigma}^{2,0} = 0$

Estimate $(\hat{\boldsymbol{\beta}}_d^0, \hat{\boldsymbol{\gamma}}^0) = \text{argmin}\, \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}_d\boldsymbol{\beta}_d - \sum_{i=1}^R \boldsymbol{E}_i\gamma_i||_F^2 + \lambda_2 \sum_{i=1}^R |\gamma_i|$

while the changes of parameters (i.e., $\boldsymbol{\beta}_d$, $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}^2$) estimation are not simultaneously negligible **do**

$t = t + 1$

Update $\hat{\boldsymbol{\sigma}}^{2,t}$ via equation 4.5

Update $(\hat{\boldsymbol{\beta}}_d^t, \hat{\boldsymbol{\gamma}}^t)$ via equation 4.6

end while

Let $\boldsymbol{Y}_a = (\boldsymbol{Y}_p, \boldsymbol{Y})$, where $\boldsymbol{Y}_p$ is $n_p \times 1$ vector of outcomes to be predicted. Given the parameter estimates for $\hat{\boldsymbol{\sigma}}^2$, $\hat{\boldsymbol{\beta}}_d$ and $\hat{\boldsymbol{\gamma}}$, the variance of $\boldsymbol{Y}_a$ can be directly derived as $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_a} = \sum_{i=1}^R \sum_{j\in(g,m)} \boldsymbol{K}_{j,i}\hat{\sigma}_{j,i}^2 + \hat{\sigma}_0^2\boldsymbol{I}_n$. The variance of $\boldsymbol{Y}_a$ can be further written as:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_a} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{pp} & \hat{\boldsymbol{\Sigma}}_{po} \\ \hat{\boldsymbol{\Sigma}}_{op} & \hat{\boldsymbol{\Sigma}}_{oo} \end{bmatrix}$$

where $\hat{\boldsymbol{\Sigma}}_{pp}$ and $\hat{\boldsymbol{\Sigma}}_{oo}$ respectively denote the variance of testing and training samples, and $\hat{\boldsymbol{\Sigma}}_{po}$ is their covariance. Using the conditional distribution formula of the multivariate normal distribution, the predictive values for the testing samples can be calculated as:

$$\boldsymbol{Y}_p = \boldsymbol{X}_{d,p}\hat{\boldsymbol{\beta}}_d + \sum_{i=1}^R \boldsymbol{E}_{i,p}\hat{\gamma}_i + \hat{\boldsymbol{\Sigma}}_{po}\hat{\boldsymbol{\Sigma}}_{oo}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_d\hat{\boldsymbol{\beta}}_d - \sum_{i=1}^R \boldsymbol{E}_i\hat{\gamma}_i)$$

where $\boldsymbol{X}_{d,p}$ ($\boldsymbol{X}_d$) and $\boldsymbol{E}_{i,p}, i \in \{1, \cdots, R\}$ ($\boldsymbol{E}_i$) denote the demographic variables and gene expression levels in the testing (training) samples, respectively.

4.3 Simulation studies

We conducted extensive simulation studies to evaluate the performance of MpLMMGMM, and further compared it with OmicKrig, a commonly used method for prediction analysis of multi-omics data (Wheeler et al., 2014), under its default setting.

For all the simulation studies, we considered three types of omics data, including gene expression, DNA methylation and genotypes. For our proposed method, we grouped genetic variants and methylation levels according to the gene annotation, and modeled their effects using the random effect terms according to equation 4.3. For gene expression data, since they are summarized at the gene level (i.e., one expression level per gene), we modeled them using the fixed effects. For all the simulation scenarios, we used the 1000 Genome Project (The 1000 Genomes Project Consortium, 2015) to generate genomic data and randomly selected 30 SNPs that are within 75Kb in each region. In addition, 30 methylation levels were also included in each region. Both gene expression and methylation levels were simulated using the uniform distribution. We set the first three regions as causal and the remaining as noise. We considered sample sizes of 500 and 1000, where 70% of the samples are used for model training and the rest for model evaluations. The prediction accuracy was gauged according to both Pearson correlations and mean square errors (MSEs). For our proposed method, we also calculated the probability of correctly selecting predictive regions from each layer of omics data.

4.3.1 Scenario I: the impact of the number of noise regions

Converging evidence has suggested that a large number of variables collected from multi-omics data is noise. To evaluate their impact, we set three regions to be causal and gradually increased the number of noise regions from 7 to 97. We considered a disease model where three levels of omics data contributed to disease risk independently:

$$\boldsymbol{Y} = \sum_{i=1}^{3} \boldsymbol{E}_i \gamma_i + \sum_{i=1}^{3} \sum_{j=1}^{30} \boldsymbol{G}_{ij} \gamma_{ij}^{g} + \sum_{i=1}^{3} \sum_{j=1}^{30} \boldsymbol{M}_{ij} \gamma_{ij}^{m} + \boldsymbol{\epsilon} \tag{4.7}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma_0^2 \boldsymbol{I}_n)$. For region i, $\boldsymbol{E}_i$ is its gene expression data; $\boldsymbol{G}_{ij}$, $j \in \{1, \cdots, 30\}$ is its genotypes; and $\boldsymbol{M}_{ij}$ is the methylation levels. For region i, $\gamma_i, i \in \{1, 2, 3\}$ is the effect sizes of gene expression data; $\gamma_{ij}^{g} \sim N(0, \sigma_{g,i}^2 / p_{g,i}), \forall j$ is the effect sizes of genetic variants; and $\gamma_{ij}^{m} \sim N(0, \sigma_{m,i}^2 / p_{m,i}), \forall j$ is the effect sizes of methylation levels. The details of the simulation settings are shown in Supplementary Table C.1. It is straightforward to show that equation 4.7 is equivalent to:

$$\boldsymbol{Y} \sim N(\sum_{i=1}^{3} \boldsymbol{E}_i \gamma_i, \sum_{i=1}^{3} \sum_{j \in (g,m)} \boldsymbol{K}_{j,i} \sigma_{j,i}^2 + \boldsymbol{I}_n \sigma_0^2)$$

where $\boldsymbol{K}_{j,i}, j \in (g, m)$ is a kernel matrix calculated based on the linear kernel. Therefore, we simulated outcomes based on the multivariate normal distribution. For each model setting (i.e., different number of noise regions), we ran 1000 Monte Carlo replicates, and reported the Pearson correlations and MSEs calculated from the testing samples. We further calculated the average probability of correctly detecting causal predictors.

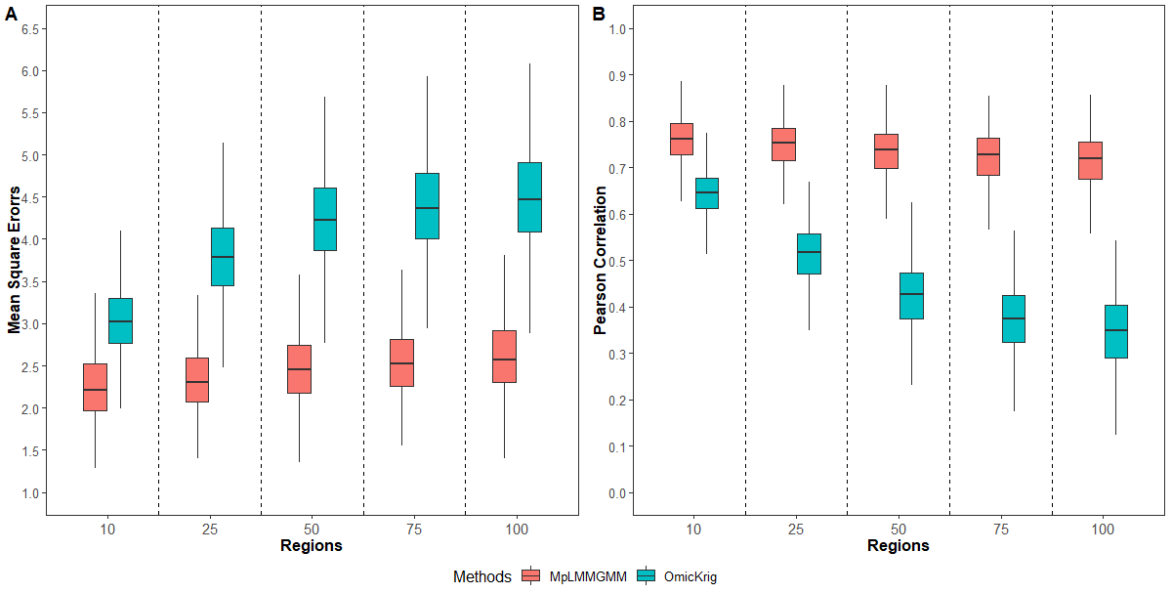

FIGURE 4.1: The impact of the number of noise regions on Pearson correlations and MSEs ($n = 500$)

TABLE 4.1: The chances of selecting causal regions as the number of noise regions increases ($n = 500$)

Regions	Gene Expression Data		Genomic Data		Methylation Data	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
10	0.999	0.919	0.928	0.901	0.924	0.969
25	1.000	0.971	0.924	0.917	0.923	0.968
50	0.998	0.984	0.895	0.929	0.911	0.975
75	0.996	0.987	0.906	0.940	0.899	0.977
100	0.995	0.990	0.887	0.948	0.894	0.979

Pearson correlations and MSEs for sample sizes of 500 and 1000 are shown in Figure 4.1 and Supplementary Figure C.1, respectively. Among all the scenarios considered,

MpLMMGMM performs better than the OmicKrig method. Of particular note, as the number of noise regions increases, the prediction accuracy of OmicKrig drops substantially, whereas it remains relatively stable for our proposed method. For example, the mean of the Pearson correlations dropped from 0.642 to 0.345 for OmicKrig, whereas it only changed from 0.757 to 0.712 for our method. Similarly, the MSEs increased from 3.043 to 4.502 for OmicKrig, while they barely changed for our method. In terms of the variable selection, our proposed method can choose the causal regions at a high probability while maintaining a low false positive rate, regardless of which layers of omics data we are exploring (Table 4.1 for $n = 500$ and Supplementary Table C.2 for $n = 1000$). This clearly indicates that our proposed method can significantly reduce the impact of noise, and thus can maintain robust performance as the amount of non-relevant variables increases. We consider the robustness against noise important, especially for the analysis of high-dimensional multi-layer omics data, as only a small proportion of measured variables are causal and they are usually unknown in advance.

4.3.2 Scenario II: the impact of disease models

Complex human diseases manifest themselves at various molecular levels (Chatterjee et al., 2013), and thus we evaluated the impact of disease models in this set of simulations. We set three regions to be causal and generated the outcomes as:

$$\boldsymbol{Y} \sim N(\sum_{i=1}^{3} \boldsymbol{E}_i \gamma_i, \sum_{i=1}^{3} \sum_{j \in (g,m,gm)} \boldsymbol{K}_{j,i} \sigma^2_{j,i} + \boldsymbol{I}_n \sigma^2_0)$$

We considered 7 disease models (Table 4.2), ranging from the simplest model where only one layer of omics data is associated with the outcomes to complex models where multiple layers of omics data jointly contribute to disease risk. The corresponding effect sizes under each disease model are summarized in Supplementary Table C.3. For each of the disease models, we considered 50 regions and generated 1000 Monte Carlo replicates for each model setting. As we did in the first round of simulations, we first used Pearson correlations and MSEs to gauge the prediction accuracy, and then calculated the probability of correctly detecting predictive markers. For comparison purposes, in addition to OmicKrig which models all layers of omics data, we also analyzed each simulated data using our proposed method, where only one layer of omics data is considered. Specifically, when only gene expression data are considered, our proposed method is equivalent to Lasso and we denoted this model as *Transcriptome*.

When only genomic or methylation data are considered, MpLMMGMM is equivalent to the pLMMGMM model proposed in Wang and Wen, 2021, and we denoted the genomic data only and methylation data only model as *Genome* and *Methylome*, respectively.

TABLE 4.2: The chances of selecting causal regions under different disease models ($n = 500$)

Disease	Gene Expression Data		Genomic Data		Methylation Data	
Models	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
$S_1 : E^{a}$	0.996	0.980	–	0.946	–	0.937
$S_2 : G^{b}$	–	0.994	0.983	0.930	–	0.986
$S_3 : M^{c}$	–	0.996	–	0.987	0.991	0.987
$S_4 : GM^{d}$	–	0.996	0.741	0.955	0.603	0.985
$S_5 : G + M^{e}$	–	0.995	0.893	0.946	0.896	0.986
$S_6 : E + G^{f}$	0.981	0.981	0.987	0.903	–	0.962
$S_7 : E + M^{g}$	0.984	0.981	–	0.965	0.993	0.962

[a] Only gene expression data is causal.
[b] Only genomic data is causal.
[c] Only methylation data is causal.
[d] Only the interaction between genomic and methylation data is causal.
[e] Both genomic and methylation data are causal.
[f] Both gene expression data and genomic data are causal.
[g] Both gene expression data and methylation data are causal.

Figure 4.2 and Supplementary Figure C.2 summarize the prediction accuracy for all the methods when using the sample sizes of 500 and 1000, respectively. Our proposed method outperforms OmicKrig under all the disease models considered. It has higher Pearson correlation coefficients and lower MSEs than OmicKrig. Although OmicKrig can simultaneously consider all layers of omics data, it treats all measured variables in a similar fashion and thus its performance can be greatly impacted when not all layers of omics data are predictive. On contrary, our proposed method has the capacity in selecting predictive variables at each omics layer, and thus maintains better prediction performance when a large number of noise is present or not all layers of omics data are predictive. As shown in Table 4.2 and Supplementary Table C.4 , our proposed MpLMMGMM method has high sensitivity and specificity for each layer of omics data. For example, when only methylation data are associated with the outcomes (i.e., disease model M), our model displayed 99.1% chance of correctly identifying the causal factors from the methylation data. With regards to the false positive, the model only has 0.4%, 1.3% and 1.3% chance of mislabeling noise variables as causal for gene expression, genomic and methylation data, respectively. Using our proposed method, we can

identify specific causal variants at each omics layer, providing a more comprehensive view of the disease etiology. The precise identification of causal factors from the corresponding omics layers can facilitate health practitioners to deliver tailed interventions. Furthermore, unlike OmicKrig which assumes each omics contributes independently to the traits, our proposed method can take the contributions from interactions into consideration (i.e., $\boldsymbol{K}_{gm}\sigma^2_{gm,i}$). As shown in Table 4.2, even for the models without marginal effects (i.e., disease model GM), the average chance of our method correctly detecting causal and noise regions are 67.2% and 97.9%, respectively. When building risk prediction models, our proposed method uses a data-driven approach to accurately select predictors from different omics layers, and thus reduces substantially the impact of noise. In addition, our proposed method can not only jointly model predictors at each omics layer, but also take the interaction effects between different omics layers into consideration. It can achieve robust and accurate prediction performance across a range of disease models (Figure 4.2 and Supplementary Figure C.2).

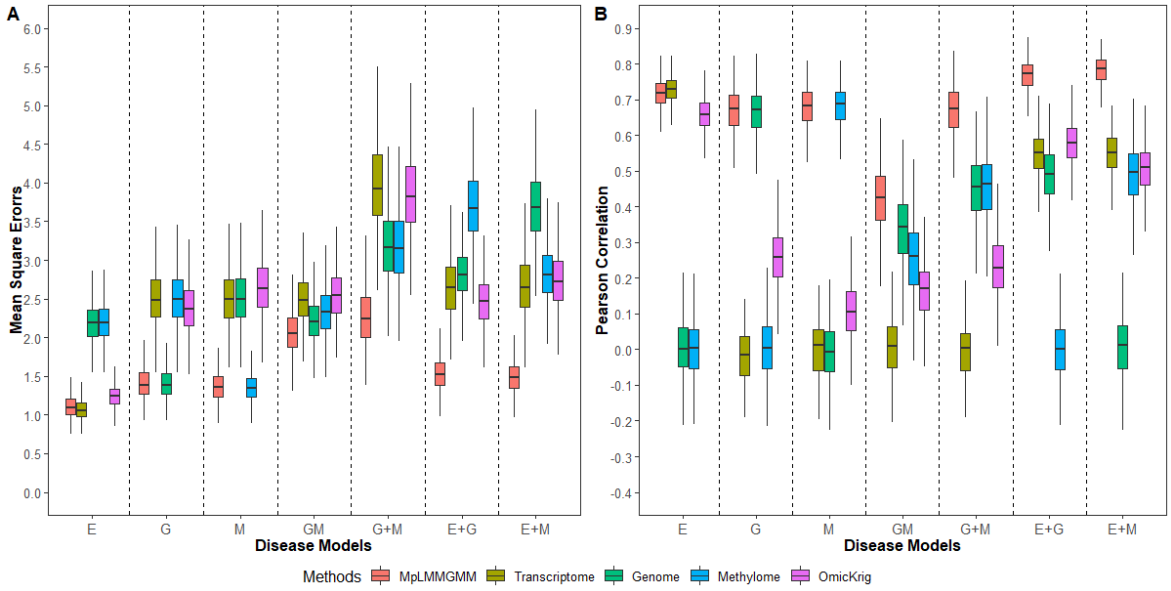

FIGURE 4.2: The impact of disease models ($n = 500$)

Comparing to the single-layer-based methods, when only one layer of omics data is associated with the outcomes (i.e., disease models E, G and M), our proposed method has similar performance to the models where only relevant omics data that contribute to disease risk are used. For example, when outcomes are only influenced by gene expression data (i.e., disease model E), our proposed method performs similarly to the

single-layer-based analysis where only gene expression data are used (i.e., *Transcriptome*), and it significantly outperforms the other single-layer-based methods where either genomic or methylation data are modeled. Similarly, when only genomic data are relevant to disease outcomes, our model has similar level of performance to *Genome* that only used genomic data, and it has much better performance than *Transcriptome* and *Methylome* where non-relevant layers of omics data are modeled. When multiple layers of omics data jointly affect the outcomes, as expected, our proposed method significantly outperformed the single-layer based methods. For example, for disease model $G + M$, where both genomic and methylation data are associated with the outcomes, our method performed better than the ones where only genomic or methylation data are used. This clearly indicates the advantages of jointly modeling multi-layer omics data, where predictors at various molecular levels can affect the outcomes. As shown in Figure 4.2 and Supplementary Figure C.2, our method has better and robust prediction performance, regardless of whether only one layer of omics data contributes to disease risk or multiple layers are relevant.

4.4 Real data application

We are interested in predicting PET-imaging outcomes, including FDG and AV45, using the whole-genome sequencing and gene expression data obtained from the ADNI. ADNI is a longitudinal study that collects biomarkers from control, mild cognitive impairment and AD patients to investigate prevention and treatment strategies for AD (Mueller et al., 2005). After removing correlated individuals, 808 subjects aged between 55 and 90 remained.

The whole-genome sequencing data were collected and sequenced on the Illumina HiSeq2000 at a non-Clinical Laboratory Improvements Amendments (non-CLIA) laboratory (Saykin et al., 2015). DNA samples come from study subjects in ADNI 2, which includes both newly recruited subjects and ADNI 1/GO continuing subjects. Gene expression data were collected from subjects in ADNI 2 at baseline for newly recruited subjects and 1st ADNI 2 visit for ADNO 1/GO continuing subjects, and then yearly. We annotated genetic variants based on GRch37 assembly, and selected 89 genes that have been reported to be associated with AD based on existing literature. We further filtered out genetic variants with missing rate larger than 1%, and a total of 59,666 variants remained in our final analyses. We focused on the baseline data, and only kept individuals with both genomic and gene expression data at the baseline. Therefore,

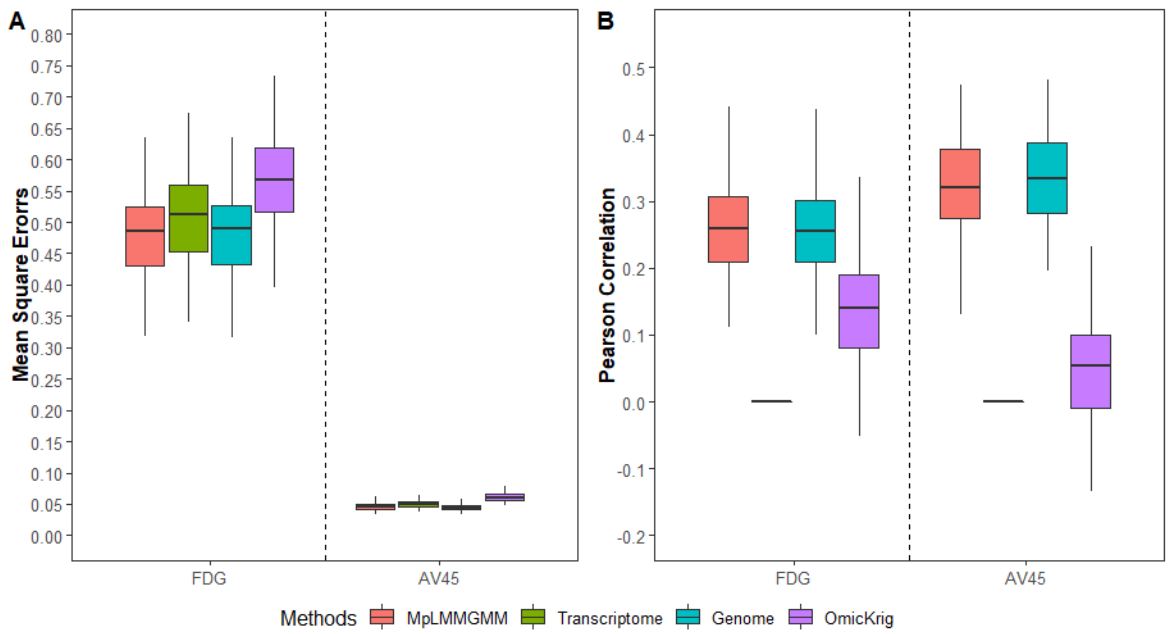

FIGURE 4.3: Accuracy comparisons for FDG and AV45

a total of 443 and 441 samples were analyzed for FDG and AV45, respectively. The distributions of FDG and AV45 for these samples are shown in Supplementary Figure C.3. We further randomly split the samples into training and validation sets ($n = 100$), where models are built based on the training samples and prediction accuracy is evaluated based on the validation set. We replicated this process 100 times to reduce the risk of chance findings.

The prediction accuracy for both FDG and AV45, including Pearson correlations and MSEs, is shown in Figure 4.3. Our proposed method has achieved better prediction performance than OmicKrig; i.e., it has higher Pearson correlations and lower MSEs than OmicKrig for both FDG and AV45. This result clearly indicates that filtering out the impact of noise can improve prediction accuracy. Comparing our proposed models built with multi-omics data with the ones built with single-layer omics data, our method has a similar level of prediction accuracy as the one built with genomic data only, but it has much better performance than the one where only gene expression data are modeled. This result indicates that genomic factors are the driving forces for the prediction of both FDG and AV45. Indeed, for both FDG and AV45, gene expression data have been rarely selected by our method (Supplementary Table C.5). Similarly, for the single-layer-based method where only gene expression data were modeled, only two genes are selected 1% of the time for FDG and eight genes are selected less than

7% of the time for AV45.

The selection details for our proposed method are shown in Supplementary Table C.5. For transcriptomic data, more than 88% of the genes were never selected from the 100 random replicates. For those genes that were selected at least once, the chance of their being selected was extremely low (i.e., 2% on average). For genomic data, three genes (i.e., *APOC1*, *APOE* and *TOMM40*) were selected more than 90% of the time, whereas the other genes averaged a less than 2% chance of being selected. All of the highly selected genes are well-known AD risk factors (Ossenkoppele et al., 2013; Roses, 2010). For example, *APOE* $\epsilon 4$ highly affects the risk of AD (Tang et al., 1998). The $rs4420638$ polymorphism on *APOC1* can increase the accumulation of homocysteine, and thus influences the risk of AD (Prendecki et al., 2018). The $rs10524523$ on *TOMM40* has also been reported to be associated with late-onset AD (Roses, 2010).

4.5 Discussion

In this work, we proposed a penalized LMM with the GMM estimator for prediction analysis on multi-omics data. The proposed MpLMMGMM groups multi-omics data into multiple regions that can be defined based on various criteria (e.g., gene and pathway annotations). It employs multiple random effect terms to model cumulative predictive effects from predictors at various molecular levels, and captures both linear and non-linear predictive effects through adopting multiple kernel functions. The proposed method uses a penalty term to enable the selection of predictive regions and omics layers, where the GMM estimator is used to expedite the model's computation. Through extensive simulation studies and analysis of the ADNI data set, we have demonstrated that our method: 1) is robust against noise; 2) has better prediction performance across a range of disease models; 3) can accurately detect predictors, including their interactions, from each layer of omics data; and 4) is computationally efficient.

Multi-omics data can be ultra-high dimensional, as single-layer omics data itself can already have millions of potential predictors. For example, the whole-genome sequencing for genomic and methylation data can each have millions of measured predictors. Treating variables obtained from all layers of omics data as predictive can not only increase the computational burden but also reduce the prediction accuracy (Byrnes et al., 2013). Therefore, variable selection is an essential step in the prediction analyses

of multi-omics data. Existing LMM-based methods either ignore the impact of noise (e.g., gBLUP) or reply on empirical criteria to perform variable screening (e.g., Multi-BLUP and MKLMM) (Speed and Balding, 2014; Weissbrod et al., 2016; Yang et al., 2010), both of which can result in poor and unstable performance. On the contrary, our proposed method can efficiently detect predictive variables at each omics layer, and simultaneously model their joint predictive effects. As the amount of noise increases, MpLMMGMM maintains stable and accurate prediction performance, whereas OmicKrig can be greatly affected (Figure 4.1 and Supplementary Figure C.1). Furthermore, as shown in Table 4.1 and Supplementary Table C.2, the sensitivity and specificity for the proposed MpLMMGMM method are relatively high, and they remain stable regardless of the amount of noise. This clearly indicates that the proposed method has achieved robust performance against noise, which is of great importance for an accurate risk prediction model.

Due to the advances in high-throughput biotechnologies, multi-omics data are becoming increasingly accessible. For example, the Cancer Genome Atlas project provides multiple molecular assays, including mRNA, DNA methylation and proteomics data, by profiling thousands of tumor samples (Cancer Genome Atlas Research Network et al., 2013). Although existing integrative methods have greatly facilitated our understanding of complex biological systems (Morris and Baladandayuthapani, 2017; Ritchie et al., 2015; Zeng and Lumley, 2018), they mainly focus on specific genes/pathways and thus have limited applicability to prediction research. This is mainly because complex human traits/diseases are usually affected by multiple genes/pathways at various molecular levels. Focusing on only a few factors can overlook the contributions from other predictors, leading to a model with low prediction accuracy. Therefore, jointly considering all potential predictors as well as their intra/inter-relationships is an essential step towards an accurate prediction model. To simultaneously model predictors at various omics layers, we extended the LMM framework, a widely used model for the analysis of genomic data, by introducing kernel functions to account for various types of predictive effects (e.g., pairwise interaction) and adopting penalization to detect predictors from all omics layers. As shown in the second simulation studies (Figure 4.2 and Supplementary Figure C.2), the proposed method outperforms the existing methods, especially when multiple layers of omics data jointly contribute to disease risk. In addition, the proposed method has much better interpretation than OmicKrig does. As shown in both Table 4.2 and Supplementary Table C.4, our model can correctly detect predictors and their interactions from the relevant omics layers, and thus greatly

facilitates the understanding of disease mechanisms. For example, when only one layer of omics data is predictive (e.g., disease models E, G and M), the proposed method can correctly detect causal regions from the corresponding omics layer and achieve a similar level of prediction accuracy as a model where only the disease-associated omics layer is used. Even for the models without marginal effects (i.e., disease model GM), our method can still detect causal regions and achieve better prediction performance than existing methods can.

Computational efficiency is one of the major challenges for penalized LMMs with a large number of random effects (Li et al., 2020; Weissbrod et al., 2016; Wen and Lu, 2020). While MLE and REML are widely used in the parameter estimations for LMMs, it is computationally demanding, especially when the number of random effects is large. This is mainly because the objective function of penalized REML/MLE is non-convex, and it has to repeatedly calculate the inverse of the $n \times n$ matrix. To expedite its computation, we adopted the GMM estimators and the objective functions to obtain GMM estimators are in a quadratic form, which is much easier to optimize. The computational efficiency of GMM allows us to jointly model a large number of regions and account for various non-linear effects. As shown in the simulations in Scenario I, MpLMMGMM can simultaneously model 100 random effect terms (e.g., the number of regions $\geq$ 50), whereas other existing LMMs can only consider a limited number of random effects (i.e., usually $\leq$ 10 (Wang and Wen, 2021)). The computational time as the number of random effects increases for our proposed method is shown in Supplementary Figures C.4 and C.5.

In the prediction analysis of PET-imaging outcomes based on genomic and gene expression data, we found that baseline FDG and AV45 are mainly predicted by the genomic data. Our method consistently found that genotypes on *APOC1*, *APOE* and *TOMM40* are highly predictive. *APOE* has been identified as a major genetic risk factor for AD. The apolipoprotein E is encoded by *APOE* gene on chromosome 19, and is involved in cholesterol transport (Zannis et al., 1993), which affects the pathogenesis of AD (Puglielli et al., 2003). The *APOE* $\epsilon 4$ is also found to be a determinant risk factor for AD (Graff-Radford et al., 2002; Duijn et al., 1994). The *APOC1* gene located on chromosome 19 encodes apolipoprotein *C1*, which takes part in the metabolism of brain cholesterol. Researchers have found that the deterioration of brain cholesterol is associated with AD (Poirier et al., 1993). In addition, the $rs11568822$ polymorphism on *APOC1* increases the risk of AD in Caucasians, Asians and Caribbean Hispanics (Zhou et al., 2014b). *TOMM40* encodes a translocase (i.e., $Tom40$) which causes the

accumulation of 29 amyloid precursor proteins during mitochondrial biogenesis, and thus affects the mitochondrial dysfunction in late-onset AD (Roses, 2010). In addition, the $rs2075650$ and $rs10524523$ polymorphisms on *TOMM40* were also found to be associated with AD (Huang et al., 2016; Prendecki et al., 2018).

In summary, we have developed a penalized LMMs with GMM estimators for risk prediction analysis on multi-omics data. Our method is robust against noise and can capture predictive markers, including their interactions, from relevant omics layers. It has better prediction performance than the commonly used methods. While our proposed method has achieved better prediction performance, there are several limitations. Similar to existing literature (Speed and Balding, 2014; Weissbrod et al., 2016), MpLMMGMM only focuses on continuous outcomes. It would be of interest to develop a generalized LMM framework for outcomes that come from the exponential family (e.g., binary and Poisson). In addition, a penalized LMM with adaptive Lasso where different penalties are imposed on each coefficient to meet the smaller bias and better sparsity, can be proposed to adopt a data-driven approach to select predictive regions. The R-package implementing the proposed method is available at the GitHub (https://github.com/XiaQiong/MpLMMGMM).

Chapter 5

Summary and future Work

5.1 Summary

This dissertation considers the penalized linear mixed model (LMM) with generalized method of moments (GMM) estimators for the prediction analysis using high-dimensional data. LMMs and their extensions have long been used for prediction analysis on high-dimensional genetic data. They usually focus on modeling simple linear relationships, and their parameter estimations rely on MLE or REML. However, the computation of penalized LMMs can be demanding, especially when the number of random effects is large and/or complex types of relationships are modeled. Therefore, we have proposed to use GMM to estimate the parameters in penalized LMM, where multi-kernel learning and variable screening are further incorporated for improved prediction performance. The new modeling framework performed significantly better than existing LMM-based methods in both simulated and real data examples.

In chapter 1, previous work on prediction analysis based on high-dimensional multi-layer omics data was reviewed. Particularly, we focused on the widely used LMMs and their challenges in modeling high-dimensional data, including variable selection, parameter estimation and non-linear effect modeling. We first briefly described the commonly used variable selection method, and then presented a detailed discussion about regularization and screening rules, including SIS, strong and safe rules. Regarding parameter estimations, the traditional maximum likelihood and restricted maximum likelihood estimators used in LMMs and penalized LMMs were discussed first, and then the GMM estimators for variance component estimation, including ANOVA, MINQUE, and MIVQUE. Finally, we discussed existing methods for integrating multi-layer omics data, where the complex inter/intra-relationships among multi-omics data can be efficiently captured.

In chapter 2, a new penalized LMM that estimates its parameters using the GMM was presented. The major challenge in estimating parameters for penalized LMMs with multiple random effects is computation. The objective function to obtain MLE or REML estimators is not easy to optimize and its traditional optimization algorithms, such as Newton Raphson and expectation-maximization, cannot handle a large number of random effects. To overcome this, we utilize the GMM estimators, where the objective function is transformed into a quadratic form that is much easier to optimize. We showed that the GMM estimators for penalized LMM have the desired oracle property. Empirical studies using simulated and real-world data showed that the new penalized LMM with GMM estimators performs remarkably better than existing methods. It can not only simultaneously model a large number of random effects, but also efficiently detect those regions that are relevant. For the prediction analysis of PET-imaging outcomes based on a candidate gene approach, the new method explained more heritability than existing methods, and it detected that genetic variants on *APOC1*, *APOE* and *TOMM40* genes are highly important for the predictions.

In chapter 3, a hybrid screen rule, which is built based on the sequential strong rule and enhanced DPP rule and incorporated into the penalized LMM with GMM estimators, is described. Penalized LMMs use multiple random effects to capture joint predictive effects from multiple markers. For the analysis of genome-wide data, penalized LMMs can have tens of thousands of random effects, which makes it computational infeasible. To address this challenge, we developed a hybrid screening rule and further incorporated it into the penalized LMMs with GMM estimators. We showed that the designed hybrid screening rule aligns well with the penalized LMM with GMM estimators. In particular, for each given value of penalty, variables that have been screened out by the hybrid screening rule are guaranteed to have no predictive effects in the corresponding penalized LMM with parameters estimated by GMM. Empirical studies showed that the hybrid screening rule can reduce the data dimension to a manageable size for the downstream prediction task. Our model can not only consistently detect predictive variables while keeping the number of selected non-relevant markers small, but also screen out inappropriate kernels to facilitate the final prediction. For the analysis of PET-imaging outcomes based on genome-wide data, the proposed screening rule screened out more than 95% of the 20,000 possible genes and the final predictions had much better and more robust performance than the performance of competing methods. The highly predictive genes detected from this genome-wide analysis were *APOC1*, *APOE*, *TOMM40* and *FADS3*, all of which are known to be

related to Alzheimer's disease (AD).

In chapter 4, a penalized LMM with the GMM estimator was developed for prediction analysis on high-dimensional multi-layer omics data. Multi-omics data are high-dimensional and they have complex inter/intra relationships, which impose significant analytical and computational challenges for prediction analysis. To address these challenges, we extended the penalized LMM with GMM estimators for the prediction analysis of multi-omics data. We used multiple kernel functions to capture complex predictive effects at various molecular levels and used a penalty term to select predictors from relevant omics layers, where GMM is used for parameters estimation. We showed that the computational efficiency of GMM allows our method to jointly model a large number of predictors from multiple omics layers. Extensive simulation studies showed that our method has better predictive performance than other existing methods. Our method can not only account for complex relationships among multi-omics data but also has the capabilities of selecting predictive regions, including their interactions, from the relevant omics layer. In the analysis of the ADNI data set, we used genomic data and gene expression data to predict PET-imaging outcomes, and we found that our proposed method has better prediction performance than other competing method. In addition, our method consistently found that genotypes on *APOC1*, *APOE* and *TOMM40* are highly predictive.

5.2 Future work

The aim of our research was to effectively select variables from high-dimensional data and ignore the large sample size (e.g.,10,000) problem. However, in the era of big data, the sample size also increases at an exponential speed. For example, The UK Biobank can provide the whole-genome sequencing data from 314,278 participants for an AD genetic study, which provides more opportunities to elucidate the biological mechanisms underlying AD. Our method cannot be used when the sample size is very large, mainly because of the complexity of objective functions. After the vectorization of elements in our objective function, the sample size becomes the square of the previous sample size. Thus, as a next step in the development of our model, we would like to derive a new optimization process of objective functions instead of its vectorization or apply a more efficient algorithm to solve the objective functions.

Another problem is that our method only focuses on continuous outcomes. However, some disease outcomes used in prediction analysis are categorical. For example, in the

studies of Gestational trophoblastic diseases, the presence of gestational trophoblastic neoplasia can be regarded as an outcome, so the result is binary (Dandis et al., 2020). Therefore, it would be of interest to develop a generalized LMM framework for outcomes that come from the exponential family (e.g., binary and Poisson), where various link functions can be applied to the outcomes. For example, we could use a logistic link function for binary outcomes and use the probability mass function for the Poisson outcomes.

In addition, our method applies a $L1$ penalty (i.e., Lasso) to select predictive markers, where the Lasso imposes the same penalty on each coefficient. However, for small bias and better sparsity, small penalties should be applied to large coefficients and large penalties should be applied to small coefficients. Thus, in the future, we hope to develop a penalized LMM with adaptive Lasso and further use GMM for parameter estimation, which will adopt a data-driven approach to select predictive regions as well as predictive layers of omics data.

To summarize, our future research will explore the large sample size, categorical outcomes and the selection of penalty techniques.

Appendix A

A penalized linear mixed model with generalized method of moments estimators for complex phenotype prediction using genomic data

A.1 Proofs of theorems

A.1.1 The derivation of objective function

Recall the objective function defined in equation 2.3 of the main text is:

$$\hat{\boldsymbol{\sigma}}^2 = \underset{\boldsymbol{\sigma}^2 \geq 0}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A} - \boldsymbol{A}^T\sum_{i=1}^{R}\sigma_i^2\boldsymbol{K}_i\boldsymbol{A} - \sigma_0^2\boldsymbol{I}_{n-P}||_F^2 + \lambda\sum_{i=1}^{R}\sigma_i^2, \quad \lambda > 0$$

Let $L = \frac{1}{2}||\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A} - \boldsymbol{A}^T\sum_{i=1}^{R}\sigma_i^2\boldsymbol{K}_i\boldsymbol{A} - \sigma_0^2\boldsymbol{I}_{n-P}||_F^2 = \frac{1}{2}||\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A} - \boldsymbol{A}^T\sum_{i=0}^{R}\sigma_i^2\boldsymbol{K}_i\boldsymbol{A}||_F^2$ be the loss function, and $\boldsymbol{K}_0 = \boldsymbol{I}_{n-P}$. Then the loss function L can be rewritten as:

$$\begin{aligned}
L =& \frac{1}{2}[tr(\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A}) + \sum_i\sum_j\sigma_i^2\sigma_j^2 tr(\boldsymbol{A}^T\boldsymbol{K}_i\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{K}_j\boldsymbol{A}) \\
& - 2\sum_i\sigma_i^2 tr(\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{K}_i\boldsymbol{A})] \\
=& \frac{1}{2}[\boldsymbol{M}^T\boldsymbol{M} + \sum_i\sum_j\sigma_i^2\sigma_j^2\boldsymbol{T}_i{}^T\boldsymbol{T}_j - 2\sum_i\sigma_i^2\boldsymbol{M}^T\boldsymbol{T}_i] \\
=& \frac{1}{2}||\boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^2||_F^2
\end{aligned}$$

As noted in the main text, $\boldsymbol{M} = vec(\boldsymbol{A}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{A})$, $\boldsymbol{T}_i = vec(\boldsymbol{A}^T\boldsymbol{K}_i\boldsymbol{A})$. Therefore, the estimation of equation 2.3 in the main text is equal to solving the equation $\boldsymbol{M} = \boldsymbol{T}\boldsymbol{\sigma}^2 + \boldsymbol{\omega}$ and $\boldsymbol{\omega} \sim N(0, \sigma_\omega^2 \boldsymbol{I})$. Our penalized objective function can be rewritten as:

$$\hat{\boldsymbol{\sigma}}^2 = \underset{\boldsymbol{\sigma}^2 \geq 0}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^2||_F^2 + \lambda \sum_{i=1}^{R} \sigma_i^2, \quad \lambda > 0$$

A.1.2 Proof of theorem 2.1

Using the KKT (Karush-Kuhn-Tucker) conditions, $\boldsymbol{\sigma}^2 = (\sigma_0^2, \sigma_1^2, \cdots, \sigma_R^2)$ is the non-negative Lasso estimator for given λ if and only if:

$$-2\boldsymbol{T}^T(\boldsymbol{M} - \boldsymbol{T}\hat{\boldsymbol{\sigma}}^2) + \lambda\mathbf{1} - \boldsymbol{r} = 0 \text{ , subject to } \begin{cases} \hat{\boldsymbol{\sigma}}^2 \geq 0 \\ \boldsymbol{r} \geq 0 \\ r_i \hat{\sigma_i}^2 = 0 \quad \forall i \end{cases} \tag{A.1}$$

Replace $\boldsymbol{M}$ by $\boldsymbol{T}\boldsymbol{\sigma}^2 + \boldsymbol{\omega}$ and recall $\boldsymbol{C} = \frac{1}{N}\boldsymbol{T}^T\boldsymbol{T}$. Let $\boldsymbol{W} = \boldsymbol{T}^T\boldsymbol{\omega}/\sqrt{N}$. Then the equation A.1 can be rewritten as:

$$2\sqrt{N}(\boldsymbol{C}(\sqrt{N}(\hat{\boldsymbol{\sigma}}^2 - \boldsymbol{\sigma}^2)) - \boldsymbol{W}) = \boldsymbol{r} - \lambda\mathbf{1} \text{ , subject to } \begin{cases} \hat{\boldsymbol{\sigma}}^2 \geq 0 \\ \boldsymbol{r} \geq 0 \\ r_i \hat{\sigma_i}^2 = 0 \quad \forall i \end{cases} \tag{A.2}$$

Let $\boldsymbol{\sigma}^2(1), \boldsymbol{\sigma}^2(2), \hat{\boldsymbol{\sigma}}^2(1), \hat{\boldsymbol{\sigma}}^2(2), \boldsymbol{W}(1), \boldsymbol{W}(2)$ and $\boldsymbol{r}(1), \boldsymbol{r}(2)$ are the $q+1$ nonzero elements and the $R-q$ zero elements of $\boldsymbol{\sigma}^2$, $\hat{\boldsymbol{\sigma}}^2$, $\boldsymbol{W}$ and $\boldsymbol{r}$. If there exist $\hat{\boldsymbol{\sigma}}^2$ meet equation A.2, and $\hat{\boldsymbol{\sigma}}^2(1) > 0$, $\hat{\boldsymbol{\sigma}}^2(2) = 0$. Then $S_1 = S(\lambda)$. That is:

$$2\sqrt{N}(\boldsymbol{C}_{11}(\sqrt{N}(\hat{\boldsymbol{\sigma}}^2(1) - \boldsymbol{\sigma}^2(1))) - \boldsymbol{W}(1)) + \lambda\mathbf{1} = 0 \tag{A.3}$$

$$2\sqrt{N}(\boldsymbol{C}_{21}(\sqrt{N}(\hat{\boldsymbol{\sigma}}^2(1) - \boldsymbol{\sigma}^2(1))) - \boldsymbol{W}(2)) + \lambda\mathbf{1} - r(2) = 0 \tag{A.4}$$

Thus, the existence of $\hat{\boldsymbol{\sigma}}^2$ of equation A.3 and A.4 can be implied by

$$\boldsymbol{C}_{11}^{-1}(\boldsymbol{W}(1) - \frac{\lambda}{2\sqrt{N}}\mathbf{1}) + \sqrt{N}\boldsymbol{\sigma}^2(1) > 0$$

$$\lambda\mathbf{1} + 2\sqrt{N}[\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}(\boldsymbol{W}(1) - \frac{\lambda}{2\sqrt{N}}\mathbf{1}) - \boldsymbol{W}(2)] \geq 0$$

Let $\boldsymbol{Z} = (Z_0, Z_1, Z_2, ...Z_q)^T = \boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1)$ and $\boldsymbol{\xi} = (\xi_{q+1}, \xi_{q+2}, ...\xi_R)^T = \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1) - \boldsymbol{W}(2)$, and then

$$\boldsymbol{Z} > -\sqrt{N}\boldsymbol{\sigma}^2 + \frac{\lambda}{2\sqrt{N}}\boldsymbol{C}_{11}^{-1}\mathbf{1}$$

$$\boldsymbol{\xi} > -\frac{\lambda}{2\sqrt{N}}(\mathbf{1} - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\mathbf{1})$$

Based on the Nonnegative irrepresentable condition:

$$\boldsymbol{\xi} > -\frac{\lambda}{2\sqrt{N}}\boldsymbol{\rho}$$

Then the probability is:

$$P(S(\lambda) \neq S_1) \leq \sum_{i=0}^{q} P(Z_i \leq -\sqrt{N}\boldsymbol{\sigma}_i^2 + \frac{\lambda}{2\sqrt{N}}[\boldsymbol{C}_{11}^{-1}\mathbf{1}]_i) + \sum_{i=q+1}^{R} P(\xi_i \leq -\frac{\lambda}{2\sqrt{N}}\rho_i)$$

Suppose that $\boldsymbol{Z} = \boldsymbol{H}_A^T\boldsymbol{\omega}$ and $\boldsymbol{H}_A^T = (h_0^a, h_1^a, h_2^a, ..., h_q^a)^T = \boldsymbol{C}_{11}^{-1}(N^{\frac{1}{2}}\boldsymbol{T}(1)^T)$. Then

$$\boldsymbol{H}_A^T\boldsymbol{H}_A = \boldsymbol{C}_{11}^{-1}(N^{\frac{1}{2}}\boldsymbol{T}(1)^T)(\boldsymbol{C}_{11}^{-1}(N^{\frac{1}{2}}\boldsymbol{T}(1)^T))^T = \boldsymbol{C}_{11}^{-1}$$

.

Let $\boldsymbol{e}_i$ be a vector with ith entry one and other zeros.

$$\begin{aligned} ||h_i^a||_2^2 &= \boldsymbol{e}_i^T\boldsymbol{H}_A^T\boldsymbol{H}_A\boldsymbol{e}_i \\ &= \boldsymbol{e}_i^T\boldsymbol{C}_{11}^{-1}\boldsymbol{e}_i \\ &\leq \frac{1}{M_2} \quad \forall i = 0, 1, 2, ..., q \end{aligned}$$

Suppose that $\boldsymbol{\xi} = \boldsymbol{H}_B^T\boldsymbol{\omega}$ and $\boldsymbol{H}_B^T = (h_{q+1}^b, h_{q+2}^b, ..., h_R^b)^T = \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}(N^{\frac{1}{2}}\boldsymbol{T}(1)^T) - (N^{\frac{1}{2}}\boldsymbol{T}(2)^T)$. Then

$$\boldsymbol{H}_B^T\boldsymbol{H}_B = \frac{1}{N}\boldsymbol{T}(2)^T(\boldsymbol{I} - \boldsymbol{T}(1)(\boldsymbol{T}(1)^T\boldsymbol{T}(1))^{-1}\boldsymbol{T}(1)^T)\boldsymbol{T}(2)$$

Let $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{T}(1)(\boldsymbol{T}(1)^T\boldsymbol{T}(1))^{-1}\boldsymbol{T}(1)^T$and $\boldsymbol{H}$ is a projection matrix and whose maximal eigen value $\lambda_{max}(\boldsymbol{H})$ is 1 or 0.

$$\begin{aligned} ||h_i^b||_2^2 &= \boldsymbol{e}_i^T\boldsymbol{H}_B^T\boldsymbol{H}_B\boldsymbol{e}_i \\ &\leq \frac{1}{N}\lambda_{max}(\boldsymbol{H})||\boldsymbol{T}(2)\boldsymbol{e}_i||_2^2 \\ &\leq \frac{1}{N}||\boldsymbol{T}_i||_2^2 \\ &\leq M_1 \quad \forall i = q+1, q+2, ..., R \end{aligned}$$

Because $\boldsymbol{\omega}_i$ are i.i.d. Gaussian variables, $\boldsymbol{Z}_i$ and $\boldsymbol{\xi}_i$ are also Gaussian variables and are bounded by secondary moment. Let $\Phi(t)$ denote the distribution function of standard Gaussian variable, then for any $t > 0$,

$$1 - \Phi(t) \leq t^{-1}e^{-\frac{1}{2}t^2}$$

for $\lambda \propto N^{\frac{1+c_4}{2}}$, by $\frac{\lambda}{2\sqrt{N}}[\boldsymbol{C}_{11}^{-1}\boldsymbol{1}]_i \leq \frac{\lambda}{2\sqrt{N}M_2}\sqrt{q+1}$, we have

$$\begin{aligned} \sum_{i=0}^{q} P(Z_i \leq -\sqrt{N}\boldsymbol{\sigma}_i^2 + \frac{\lambda}{2\sqrt{N}}[\boldsymbol{C}_{11}^{-1}\boldsymbol{1}]_i) &\leq \sum_{i=0}^{q} P(|Z_i| \geq \sqrt{N}\boldsymbol{\sigma}_i^2 - \frac{\lambda}{2\sqrt{N}}[\boldsymbol{C}_{11}^{-1}\boldsymbol{1}]_i) \\ &= (q+1)O(1 - \Phi((1+o(1))M_3M_2N^{\frac{c_2}{2}})) \\ &= o(e^{-N^{c_3}}) \end{aligned}$$

And also we can get

$$\begin{aligned} \sum_{i=q+1}^{R} P(\xi_i \leq -\frac{\lambda}{2\sqrt{N}}\rho_i) &\leq \sum_{i=q+1}^{R} P(|\xi_i| \geq \frac{\lambda}{2\sqrt{N}}\rho_i) \\ &= (R-q)O(1 - \Phi(\frac{1}{M_1}\frac{\lambda}{\sqrt{N}})\boldsymbol{\rho}) \\ &= o(e^{-N^{c_3}}) \end{aligned}$$

$$\begin{aligned} P(S(\lambda) \neq S_1) &= \sum_{i=0}^{q} P(Z_i \leq -\sqrt{N}\boldsymbol{\sigma}_i^2 + \frac{\lambda}{2\sqrt{N}}[\boldsymbol{C}_{11}^{-1}\boldsymbol{1}]_i) + \sum_{i=q+1}^{R} P(\xi_i \leq -\frac{\lambda}{2\sqrt{N}}\rho_i) \\ &\leq o(e^{-N^{c_3}}) + o(e^{-N^{c_3}}) \to 0, \quad as \quad N \to \infty \end{aligned}$$

$$P(S(\lambda) = S_1) \to 1, \quad as \quad N \to \infty$$

Variable selection consistency proof completed.

A.1.3 Proof of theorem 2.2

Under assumptions 3 and assumptions 4 in the main text, Negahban et al. (2012) establishes upper bounds for nonnegative Lasso estimation errors:

$$||\hat{\sigma}^2 - \sigma^2||_2^2 \leq \frac{64\sigma_\omega^2}{k_m^2} \frac{(q+1)\log R}{N}$$

$$||\hat{\sigma}^2 - \sigma^2||_1 \leq \frac{24\sigma_\omega^2}{k_m^2}(q+1)\sqrt{\frac{logR}{N}}$$

From these bounds, we know that our pLMMGMM method has estimation consistency, i.e., $||\hat{\sigma}^2 - \sigma^2||_2^2 \to 0$ if $\frac{(q+1)\log R}{N} \to 0$, as $N \to \infty$.

A.1.4 Proof of theorem 2.3

Based on the proof of theorem 2.1:

$$\sqrt{N}(\hat{\boldsymbol{\sigma}}^2(1) - \boldsymbol{\sigma}^2(1)) = \boldsymbol{C}_{11}^{-1}(\boldsymbol{W}(1) - \frac{\lambda}{\sqrt{N}}\mathbf{1})$$

Since $\omega_i \sim N(0, \sigma_\omega^2)$, then

$$\boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1) \sim N(0, \sigma_\omega^2 \boldsymbol{C}_{11}^{-1})$$

Then we have

$$\begin{aligned}
\hat{\Phi}(t) &= P(\sqrt{N}(\hat{\boldsymbol{\sigma}}^2(1) - \boldsymbol{\sigma}^2(1)) \leq t) \\
&= P(\boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1) - \frac{\lambda}{\sqrt{N}}\boldsymbol{C}_{11}^{-1}\mathbf{1} \leq t, S(\lambda) = S_1) \\
&+ P(\sqrt{N}(\hat{\boldsymbol{\sigma}}^2(1) - \boldsymbol{\sigma}^2(1)) \leq t, S(\lambda) \neq S_1) \\
&= P(\boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1) - \frac{\lambda}{\sqrt{N}}\boldsymbol{C}_{11}^{-1}\mathbf{1} \leq t) - P(\boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1) - \frac{\lambda}{\sqrt{N}}\boldsymbol{C}_{11}^{-1}\mathbf{1} \leq t, S(\lambda) \neq S_1) \\
&+ P(\sqrt{N}(\hat{\boldsymbol{\sigma}}^2(1) - \boldsymbol{\sigma}^2(1)) \leq t, S(\lambda) \neq S_1)
\end{aligned}$$

Based on the assumption 1 in the main text, we can prove that:

$$\frac{\lambda}{\sqrt{N}}\boldsymbol{C}_{11}^{-1}\mathbf{1} \to 0, \quad as \quad N \to \infty$$

then by Slutsky's theorem,

$$P(\boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1) - \frac{\lambda}{\sqrt{N}}\boldsymbol{C}_{11}^{-1}\mathbf{1} \leq t) = P(\boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1) \leq t) \sim N(0, \sigma_{\omega}^{2}\boldsymbol{C}_{11}^{-1}), \quad as \quad N \to \infty$$

In addition,

$$P(\boldsymbol{C}_{11}^{-1}\boldsymbol{W}(1) - \frac{\lambda}{\sqrt{N}}\boldsymbol{C}_{11}^{-1}\mathbf{1} \leq t, S(\lambda) \neq S_1) \quad \leq \quad P(S(\lambda) \neq S_1) \to 0, \quad as \quad N \to \infty$$

$$P(\sqrt{N}(\hat{\boldsymbol{\sigma}}^2(1) - \boldsymbol{\sigma}^2(1)) \leq t, S(\lambda) \neq S_1) \quad \leq \quad P(S(\lambda) \neq S_1) \to 0, \quad as \quad N \to \infty$$

Asymptotic normality proof completed.

A.2 Additional tables

TABLE A.1: The chances of selecting two predictive regions as the number of noise regions increases ($n = 1000$)

Regions	Sensitivity	Specificity
5	1.000	0.894
10	1.000	0.887
50	1.000	0.905
100	1.000	0.922

TABLE A.2: The chances of selecting two predictive regions under different disease models ($n = 1000$)

Disease Models	Sensitivity	Specificity
$S_1 : L + L$	1.000	0.911
$S_2 : R + R$	1.000	0.909
$S_3 : P + P$	1.000	0.980
$S_4 : L + R$	0.999	0.904
$S_5 : L + P$	0.999	0.947

TABLE A.3: The chances of selecting genes for FDG and AV45

Genes	Chromosome	Start Position	End Position	Probability(FDG)	Probability(AV45)
COL11A1	1	103342022	103574052	0	0
CR1	1	207669472	207815110	0	0
CR1L	1	207818457	207897036	0	0
FCER1G	1	161185086	161189038	0	0
FLVCR1	1	213031596	213072705	0	0
FLVCR1-AS1	1	213029945	213031480	0	0
GBP2	1	89571815	89591842	0	0
HSD11B1	1	209859524	209908295	0	0
NGF	1	115828536	115880857	0	0
PARP1	1	226548391	226595801	0	0
POU2F1	1	167190065	167396582	0	0
BIN1	2	127805598	127864903	0	0
LHCGR	2	48913912	48982880	0	0
LRP2	2	169983618	170219122	0	0
APOD	3	195295572	195311076	0	0
GSK3B	3	119540801	119813264	0	0
SST	3	187386693	187388201	0	0.01
ALB	4	74269971	74287129	0	0.02
COL25A1	4	109731876	110223799	0	0
ADRB2	5	148206155	148208197	0	0
ARSB	5	78073036	78282357	0	0
FGF1	5	141971742	142077635	0.24	0
FGF10	5	44305096	44388784	0	0
FGF10-AS1	5	44388833	44414091	0	0
FGF18	5	170846666	170884630	0	0
NDUFS4	5	52856464	52979171	0	0
PPP2R2B-IT1	5	146293769	146299069	0	0
AGER	6	32148744	32152099	0	0.01
HSPA1A	6	31783290	31785719	0	0
MICA	6	31367560	31383092	0	0.01
MICAL1	6	109765265	109787171	0	0
TBP	6	170863420	170881958	0	0
TBPL1	6	134273307	134308638	0	0
TREM2	6	41126243	41130924	0	0

TABLE A.3: The chances of selecting genes for FDG and AV45 *(continued)*

Genes	Chromosome	Start Position	End Position	Probability(FDG)	Probability(AV45)
CAV1	7	116164838	116201239	0	0
PON3	7	94989183	95025687	0	0.01
RELN	7	103112230	103629963	0	0
ADAM9	8	38854504	38962779	0	0.03
NAT1	8	18027970	18081198	0	0
NRG1	8	31497267	32622558	0	0
DAPK1	9	90112142	90323549	0	0
DFNB31	9	117164359	117267736	0	0
HSPA5	9	127997126	128003666	0	0.01
POMT1	9	134378288	134399193	0.05	0.04
RXRA	9	137218308	137332432	0	0
TLR4	9	120466452	120479769	0	0
CACNB2	10	18429605	18830688	0	0
MINPP1	10	89264222	89313218	0	0
TET1	10	70320116	70454239	0	0
TFAM	10	60144902	60158990	0	0
HBG2	11	5274420	5276011	0.07	0
ATF7	12	53901639	54020199	0.03	0
ATF7IP	12	14518565	14655869	0	0
OLR1	12	10310898	10324790	0	0
SLC11A2	12	51373565	51422058	0	0
KLF5	13	73629113	73651680	0	0
CINP	14	102814618	102829253	0	0
GNPNAT1	14	53241910	53258386	0	0
HNRNPC	14	21677295	21737638	0	0
MTHFD1	14	64854758	64926725	0.01	0
PNP	14	20937537	20946165	0	0
SEL1L	14	81937890	82000205	0	0
SERPINA1	14	94843083	94857029	0.21	0
SERPINA2	14	94829974	94833039	0	0
SERPINA3	14	95078713	95090390	0	0
SERPINA4	14	95027756	95036250	0	0
SERPINA5	14	95047705	95059457	0	0
SERPINA6	14	94770584	94789688	0	0
SERPINA9	14	94929057	94942670	0	0
SERPINA10	14	94749649	94759608	0	0
SERPINA11	14	94908800	94919122	0	0
SERPINA12	14	94953619	94984181	0	0
SERPINA13P	14	95107061	95113331	0	0
CHRNA3	15	78885394	78913637	0	0
MEF2A	15	100106132	100256629	0	0
MEFV	16	3292027	3306627	0	0
UBE2I	16	1359153	1377019	0	0
CCL3	17	34415602	34417506	0	0
CDK5R1	17	30814104	30818271	0	0
COX10	17	13972718	14111996	0	0
COX10-AS1	17	13932608	13972775	0	0
PNMT	17	37824233	37826728	0	0

TABLE A.3: The chances of selecting genes for FDG and AV45 *(continued)*

Genes	Chromosome	Start Position	End Position	Probability(FDG)	Probability(AV45)
APOC1	19	45417920	45422606	1	1
APOE	19	45409038	45412650	1	0.97
GNA11	19	3094407	3124000	0	0
TOMM40	19	45394476	45406946	0.28	1
DOPEY2	21	37536838	37666572	0	0
MCM3AP	21	47655038	47705308	0	0
MCM3AP-AS1	21	47649144	47671615	0	0
NCAM2	21	22370632	22912517	0	0
RUNX1-IT1	21	36410232	36411723	0.07	0
S100B	21	48018530	48025035	0.24	0
SAMSN1	21	15857548	15955723	0	0
SAMSN1-AS1	21	15954522	15970624	0.01	0
SEPT3	22	42372930	42394225	0	0

A.3 Additional figures

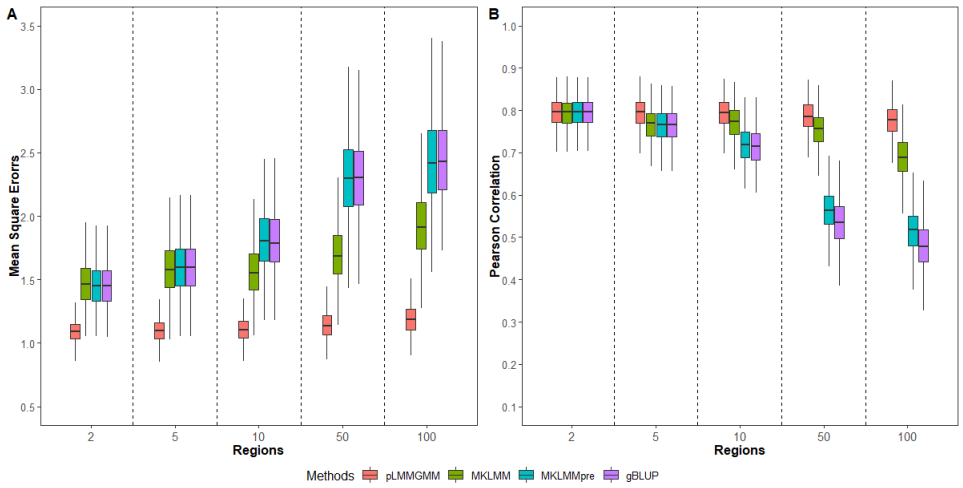

FIGURE A.1: The impact of the number of noise regions on Pearson correlations and MSEs ($n = 1000$)

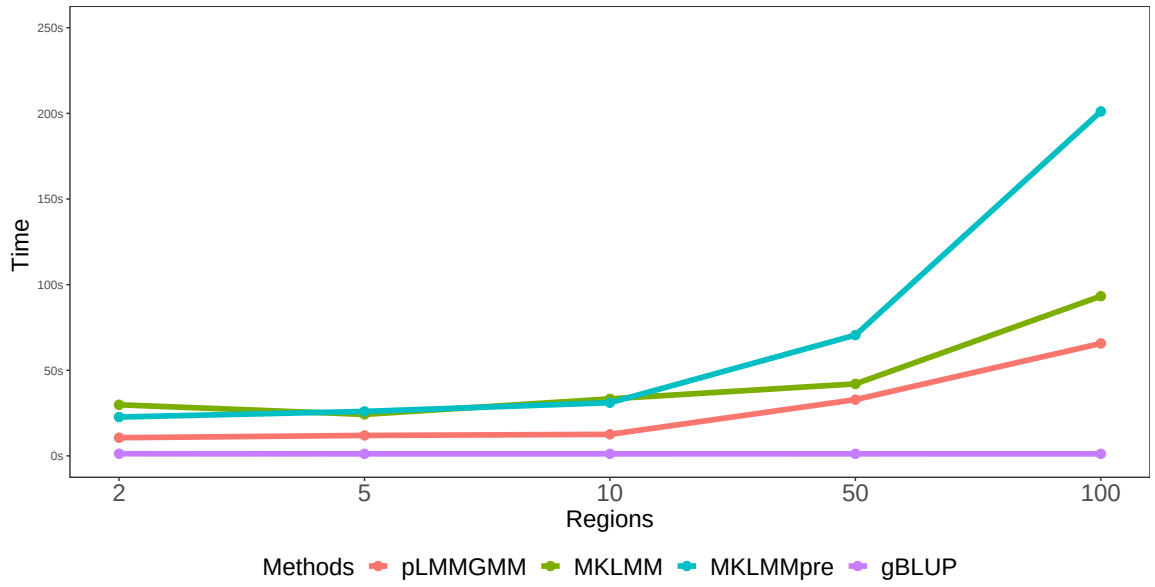

FIGURE A.2: The impact of the number of noise regions on computational time ($n = 1000$)

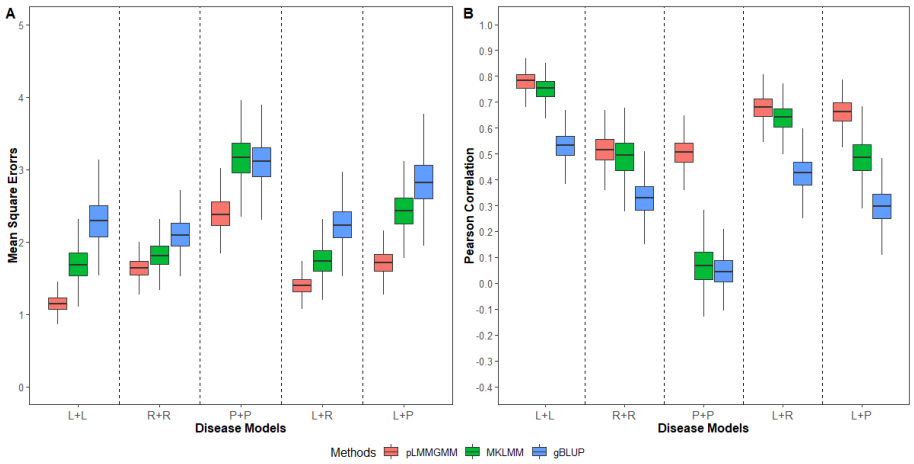

FIGURE A.3: The impact of disease models. $L + L$: genetic variants on both regions have linear additive effects. $R + R$: predictors from both regions have non-linear predictive effects. $P + P$: both regions harbor variants with pair-wise interaction effects. $L + R$: genetic variants on the first and second regions have linear additive and non-linear effects, respectively. $L+P$: predictors on the first and second regions have linear additive and pair-wise interaction effects, respectively ($n = 1000$)

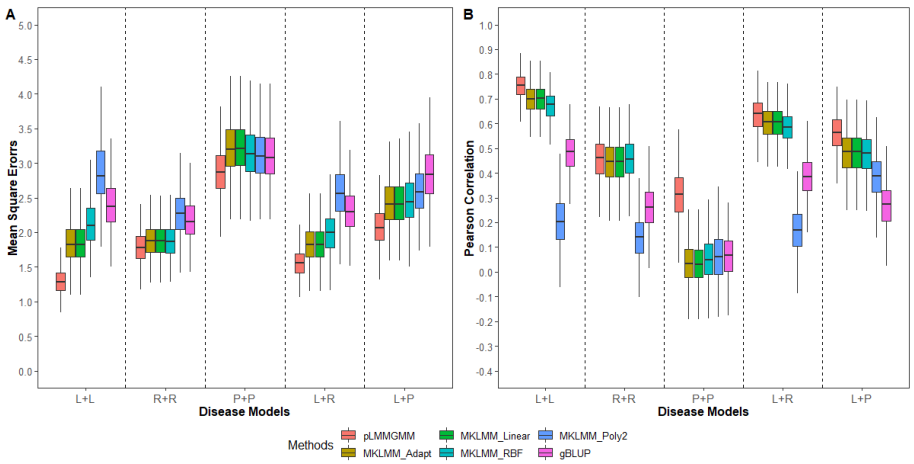

FIGURE A.4: The impact of disease models. $L + L$: genetic variants on both regions have linear additive effects. $R + R$: predictors from both regions have non-linear predictive effects. $P + P$: both regions harbor variants with pair-wise interaction effects. $L + R$: genetic variants on the first and second regions have linear additive and non-linear effects, respectively. $L+P$: predictors on the first and second regions have linear additive and pair-wise interaction effects, respectively ($n = 1000$)

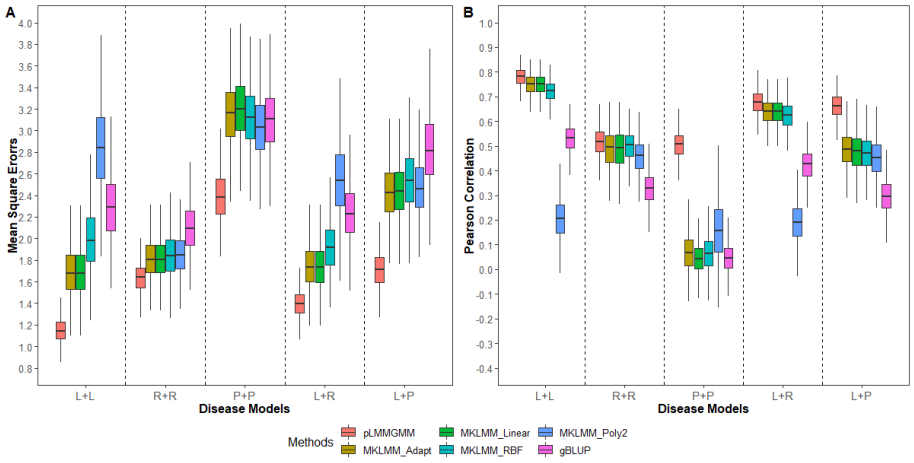

FIGURE A.5: The impact of disease models. $L + L$: genetic variants on both regions have linear additive effects. $R + R$: predictors from both regions have non-linear predictive effects. $P + P$: both regions harbor variants with pair-wise interaction effects. $L + R$: genetic variants on the first and second regions have linear additive and non-linear effects, respectively. $L+P$: predictors on the first and second regions have linear additive and pair-wise interaction effects, respectively ($n = 1000$)

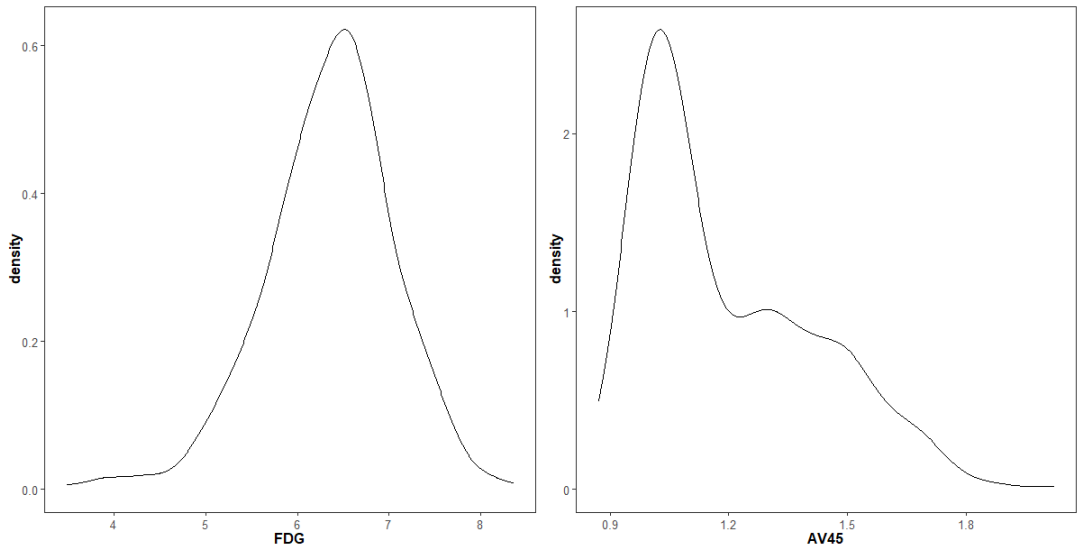

FIGURE A.6: The distribution for FDG and AV45

Appendix B

A hybrid screening rule designed for the penalized linear mixed model with generalized method of moments estimators

B.1 Screening rule

B.1.1 Sequential strong rule

For the derivation of SSR rule, we start with the KKT condition of the standard Lasso problem 3.1:

$$\boldsymbol{T}_i^T(\boldsymbol{M} - \boldsymbol{T}\hat{\boldsymbol{\sigma}}^2) = \lambda v_i,\ i = 1, 2, ...R \tag{B.1}$$

where v_i is the ith component of the subgradient of $\left\|\hat{\boldsymbol{\sigma}}^2\right\|_1$. Let $c_i(\lambda) = \boldsymbol{T}_i^T(\boldsymbol{M} - \boldsymbol{T}\hat{\boldsymbol{\sigma}}^2(\lambda))$ and assume that $c_i(\lambda)$ is non-expansive in λ, then:

$$\left|c_i(\lambda) - c_i(\tilde{\lambda})\right| \leq \left|\lambda - \tilde{\lambda}\right|,\ \textit{for any}\ \ \lambda\ \textit{and},\ \tilde{\lambda}\ \textit{and}\ i = 1, 2, ...R \tag{B.2}$$

Using the condition B.2, if we have $|c_i(\lambda_k)| < 2\lambda_{k+1} - \lambda_k$ then:

$$\begin{aligned} |c_i(\lambda_{k+1})| &\leq |c_i(\lambda_{k+1}) - c_i(\lambda_k)| + |c_i(\lambda_k)| \\ &< (\lambda_k - \lambda_{k+1}) + (2\lambda_{k+1} - \lambda_k) \\ &= \lambda_{k+1} \end{aligned}$$

which implies that $\hat{\sigma}_i^2(\lambda_{k+1}) = 0$ based on the KKT condition B.1.

B.1.2 Enhanced DPP rule

Firstly, we give the detailed derivation of the dual problem of Lasso. Based on the standard Lasso problem 3.2 in the main text, we introduce a new variable $\boldsymbol{Z} = \boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^{\boldsymbol{2}}$. Then the standard Lasso problem 3.2 was reformulated as:

$$\underset{\boldsymbol{\sigma}^2 \geq 0}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{Z}\|_2^2 + \lambda \left\|\boldsymbol{\sigma}^2\right\|_1$$

$$\textbf{s.t.} \quad \boldsymbol{Z} = \boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^{\boldsymbol{2}}$$

Thus the new likelihood function becomes:

$$L(\boldsymbol{\sigma}^2, \boldsymbol{Z}, \boldsymbol{\eta}) = \frac{1}{2} \|\boldsymbol{Z}\|_2^2 + \lambda \left\|\boldsymbol{\sigma}^2\right\|_1 + \boldsymbol{\eta}^T (\boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^{\boldsymbol{2}} - \boldsymbol{Z}) \tag{B.3}$$

and the objective function is:

$$\underset{\boldsymbol{\sigma}^2, \boldsymbol{Z}}{\operatorname{argmin}} L(\boldsymbol{\sigma}^2, \boldsymbol{Z}, \boldsymbol{\eta}) = \boldsymbol{\eta}^T \boldsymbol{M} + \underset{\boldsymbol{\sigma}^2}{\operatorname{argmin}} (-\boldsymbol{\eta}^T \boldsymbol{T}\boldsymbol{\sigma}^2 + \lambda \left\|\boldsymbol{\sigma}^2\right\|_1) + \underset{\boldsymbol{Z}}{\operatorname{argmin}} (\frac{1}{2} \|\boldsymbol{Z}\|_2^2 - \boldsymbol{\eta}^T \boldsymbol{Z}) \tag{B.4}$$

The objective function B.4 can be divided into two parts:

$$\underset{\boldsymbol{\sigma}^2}{\operatorname{argmin}} (-\boldsymbol{\eta}^T \boldsymbol{T}\boldsymbol{\sigma}^2 + \lambda \left\|\boldsymbol{\sigma}^2\right\|_1) \tag{B.5}$$

and

$$\underset{\boldsymbol{Z}}{\operatorname{argmin}} (\frac{1}{2} \|\boldsymbol{Z}\|_2^2 - \boldsymbol{\eta}^T \boldsymbol{Z}) \tag{B.6}$$

Firstly, we consider the optimization problem B.5. Let

$$f_1(\boldsymbol{\sigma}^2) = -\boldsymbol{\eta}^T \boldsymbol{T}\boldsymbol{\sigma}^2 + \lambda \left\|\boldsymbol{\sigma}^2\right\|_1 \tag{B.7}$$

and its subgradient:

$$\partial f_1(\boldsymbol{\sigma}^2) = -\boldsymbol{T}^T \boldsymbol{\eta} + \lambda \boldsymbol{v}$$

where $\|\boldsymbol{v}\|_\infty \leq 1$ and $\boldsymbol{v}$ is the subgradient of $\|\boldsymbol{\sigma}^2\|_1$. The necessary condition for f_1 to attain an optimum is to exist $\boldsymbol{\sigma}^{2*}$,$\boldsymbol{v}^*$ which satisfy:

$$\boldsymbol{v}^* = \frac{\boldsymbol{T}^T \boldsymbol{\eta}}{\lambda}, \quad \|\boldsymbol{v}\|_\infty \leq 1, \quad \boldsymbol{v}^{*T} \boldsymbol{\sigma}^{2*} = \left\|\boldsymbol{\sigma}^{2*}\right\|_1 \tag{B.8}$$

which is equivalent to

$$\left|\boldsymbol{T}_i^T\boldsymbol{\eta}\right| \leq \lambda, i = 1, 2, ..., R$$

Plugging the equations B.8 into equation B.7, we can get the optimal value of equation B.7 is 0.

Next, for the second optimal problem B.6, we let

$$f_2(\boldsymbol{Z}) = \frac{1}{2}\left\|\boldsymbol{Z}\right\|_2^2 - \boldsymbol{\eta}^T\boldsymbol{Z} \tag{B.9}$$

Clearly, the optimal value of equation B.9 is $-\frac{1}{2}\left\|\boldsymbol{\eta}\right\|_2^2$ when $\boldsymbol{Z} = \boldsymbol{\eta}$. Therefore, the objective function B.4 is rewritten as:

$$\begin{aligned} \underset{\boldsymbol{\eta}}{\text{argmin}} \quad & \boldsymbol{\eta}^T\boldsymbol{M} - \frac{1}{2}\left\|\boldsymbol{\eta}\right\|_2^2 \\ s.t. \quad & \left|\boldsymbol{T}_i^T\boldsymbol{\eta}\right| \leq \lambda, \quad i = 1, 2, ..., R \end{aligned}$$

which is equivalent to

$$\begin{aligned} \underset{\boldsymbol{\eta}}{\text{argmin}} \quad & \frac{1}{2}\left\|\boldsymbol{M}\right\|_2^2 - \frac{1}{2}\left\|\boldsymbol{\eta} - \boldsymbol{M}\right\|_2^2 \\ s.t. \quad & \left|\boldsymbol{T}_i^T\boldsymbol{\eta}\right| \leq \lambda, \quad i = 1, 2, ..., R \end{aligned} \tag{B.10}$$

let $\boldsymbol{\theta} = \frac{\boldsymbol{\eta}}{\lambda}$, function B.10 transforms to

$$\begin{aligned} \underset{\theta}{\text{argmin}} \quad & \frac{1}{2}\left\|\boldsymbol{M}\right\|_2^2 - \frac{\lambda^2}{2}\left\|\boldsymbol{\theta} - \frac{\boldsymbol{M}}{\lambda}\right\|_2^2 \\ s.t. \quad & \left|\boldsymbol{T}_i^T\boldsymbol{\theta}\right| \leq 1, \quad i = 1, 2, ..., R \end{aligned} \tag{B.11}$$

From the KKT condition, we have:

$$0 \in \partial_{\boldsymbol{\sigma}^2} L(\boldsymbol{\sigma}^{2*}, \boldsymbol{Z}^*, \boldsymbol{\theta}^*) = -\lambda\boldsymbol{T}^T\boldsymbol{\theta}^* + \lambda\boldsymbol{v}, \ \textit{where} \ \left\|\boldsymbol{v}\right\|_\infty \leq 1, \boldsymbol{v}^T\boldsymbol{\sigma}^{2*} = \left\|\boldsymbol{\sigma}^{2*}\right\|_1 \tag{B.12}$$

$$\nabla_{\boldsymbol{Z}} L(\boldsymbol{\sigma}^{2*}, \boldsymbol{Z}^*, \boldsymbol{\theta}^*) = \boldsymbol{Z}^* - \lambda\boldsymbol{\theta}^* = 0 \tag{B.13}$$

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\sigma}^{2*}, \boldsymbol{Z}^*, \boldsymbol{\theta}^*) = \lambda(\boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^{2*} - \boldsymbol{Z}^*) = 0 \tag{B.14}$$

From B.13 and B.14, we have:

$$\boldsymbol{M} = \boldsymbol{T}\boldsymbol{\sigma}^{2*} + \lambda\boldsymbol{\theta}^*$$

From B.12, we know there exists $\boldsymbol{v}^* \in \partial \left\|\boldsymbol{\sigma}^{2*}\right\|_1$ such that:

$$\boldsymbol{T}^T\boldsymbol{\theta}^* = \boldsymbol{v}^*, \left\|\boldsymbol{v}^*\right\|_\infty \leq 1, (\boldsymbol{v}^*)^T\boldsymbol{\sigma}^{2*} = \left\|\boldsymbol{\sigma}^{2*}\right\|_1$$

which is equivalent to

$$\left|\boldsymbol{T}_i^T\boldsymbol{\theta}^*\right| \leq 1, i = 1, 2, ..., R., (\boldsymbol{\theta}^*)^T\boldsymbol{T}\boldsymbol{\sigma}^{2*} = \left\|\boldsymbol{\sigma}^{2*}\right\|_1 \tag{B.15}$$

Thus, the equation B.15 can give a conclusion:

$$\boldsymbol{T}_i^T\boldsymbol{\theta}^*(\lambda) \in \begin{cases} 1 & if \quad \sigma_i^{2*} \neq 0 \\ (-1, 1) & if \quad \sigma_i^{2*} = 0 \end{cases} \tag{B.16}$$

Then we can conclude the following rule :

$$\left|\boldsymbol{T}_i^T\boldsymbol{\theta}^*(\lambda)\right| < 1 \quad \Rightarrow \quad \sigma_i^{2*} = 0 \quad \Rightarrow \quad \boldsymbol{T}_i \textbf{ is an noisy variable} \tag{B.17}$$

However, $\theta^*(\lambda)$ is generally unknown, we can not directly apply rule B.17 to identify the noise. Inspired by the SAFE rules (Ghaoui et al., 2010), we can first estimate a region $\boldsymbol{\Theta}$ which contains the $\boldsymbol{\theta}^*(\lambda)$. Then, the rule B.17 can be rewritten as follows:

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\boldsymbol{T}_i^T\boldsymbol{\theta}(\lambda)\right| < 1 \quad \Rightarrow \quad \sigma_i^{2*} = 0 \quad \Rightarrow \quad \boldsymbol{T}_i \textbf{ is an noisy variable} \tag{B.18}$$

Clearly, the final aim is to find a region $\boldsymbol{\Theta}$ which contains $\boldsymbol{\theta}^*(\lambda)$.

Based on the equation B.11, we can see that the dual optimal solution is the feasible point which is closest to $\frac{\boldsymbol{M}}{\lambda}$. Let the feasible set be $\boldsymbol{F}$ and $\boldsymbol{\theta}^*(\lambda)$ can be regarded as the projection of $\frac{\boldsymbol{M}}{\lambda}$ onto the polytope $\boldsymbol{F}$, i.e.,

$$\boldsymbol{\theta}^*(\lambda) = P_F\left(\frac{\boldsymbol{M}}{\lambda}\right) = \underset{\boldsymbol{\theta}\in\boldsymbol{F}}{\operatorname{argmin}} \left\|\boldsymbol{\theta} - \frac{\boldsymbol{M}}{\lambda}\right\|_2$$

where $P_F(\cdot)$ is the projection operator. It is easy to see that $\frac{\boldsymbol{M}}{\lambda}$ would be an interior point of $\boldsymbol{F}$ when λ is large enough, which implies that, for all $i = 1, 2, ...R$, we have $\left|\boldsymbol{T}_i^T\frac{\boldsymbol{M}}{\lambda}\right| < 1$. And $\boldsymbol{\theta}^*(\lambda)$ is also an interior point of $\boldsymbol{F}$ since $\boldsymbol{\theta}^*(\lambda) = \frac{\boldsymbol{M}}{\lambda}$. It is easy to see that $\left|\boldsymbol{T}_i^T\boldsymbol{\theta}^*(\lambda)\right| < 1$ for all $i = 1, 2, ...R$. Thus, we can conclude that $\boldsymbol{\sigma}^{2*}(\lambda) = 0$, under the assumption that λ is large enough.

For the Lasso problem, λ and λ_0 are two regularization parameters, we have:

$$\|\boldsymbol{\theta}^*(\lambda) - \boldsymbol{\theta}^*(\lambda_0)\|_2 \leq \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \cdot \|\boldsymbol{M}\|_2 \tag{B.19}$$

The equation B.19 implies that the dual optimal solution must be inside a ball centered at $\boldsymbol{\theta}^*(\lambda_0)$ with radius $\left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \cdot \|\boldsymbol{M}\|_2$. Then, Wang et al., 2015 proposed the theorem:

Theorem B.1. *For the Lasso problem, assume that the dual optimum at λ_0 (i.e., $\boldsymbol{\theta}^*(\lambda_0)$) is known. Let λ be a positive value different from λ_0. Then $\sigma_i^{2*}(\lambda) = 0$ if*

$$\left|\boldsymbol{T}_i^T \boldsymbol{\theta}^*(\lambda)\right| < 1 - \|\boldsymbol{T}_i\|_2 \|\boldsymbol{M}\|_2 \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right|$$

By setting $\lambda_0 = \lambda_{max}$ and $\boldsymbol{\theta}^*(\lambda_{max}) = \frac{\boldsymbol{M}}{\lambda_{max}}$, the basic DPP rule can be derived:

$$\left|\boldsymbol{T}_i^T \frac{\boldsymbol{M}}{\lambda_{max}}\right| < 1 - \left(\frac{1}{\lambda} - \frac{1}{\lambda_{max}}\right) \|\boldsymbol{T}_i\|_2 \|\boldsymbol{M}\|_2$$

Then Wang et al., 2015 further proposed the sequential DPP screening rule,

$$\left|\boldsymbol{T}_i^T \frac{\boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^{2*}(\lambda_k)}{\lambda_k}\right| < 1 - \|\boldsymbol{T}_i\|_2 \|\boldsymbol{M}\|_2 \left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right)$$

In order to improve the performance of the DPP screening rules, Wang et al., 2015 further proposed the EDPP screening rule based on projections of rays and the nonexpansiveness of the projection operators:

$$\left|\boldsymbol{T}_i^T \left(\frac{\boldsymbol{M} - \boldsymbol{T}\boldsymbol{\sigma}^{2*}(\lambda_k)}{\lambda_k} + \frac{1}{2} v_2^{\perp}(\lambda_{k+1}, \lambda_k)\right)\right| < 1 - \frac{1}{2} \|\boldsymbol{T}_i\|_2 \left\|v_2^{\perp}(\lambda_{k+1}, \lambda_k)\right\|_2 \tag{B.20}$$

Given $\boldsymbol{\sigma}^{2*}(\lambda_k)]$, the ith variable under λ_{k+1} will be discarded when the condition B.20 was met.

B.2 Additional tables

TABLE B.1: The effect sizes for the first simulation

Regions	σ_1^2	σ_2^2	σ_0^2
5000	1	1	1
10000	1	1	1
15000	1	1	1
20000	1	1	1

TABLE B.2: The number of selected total and causal regions as the input data dimension increases ($n = 1000$)

Regions	Number of Total Regions Selected (Number of Causal Regions Selected)			
	MKLMM	MKLMM2	MKLMM9	Hybrid Screening Rule
5000	289.92 (1.98)	2 (0)	9 (0)	6.06 (1.96)
10000	586.40 (1.96)	2 (0)	9 (0)	6.38 (2.00)
15000	878.57 (1.90)	2 (0)	9 (0)	6.23 (1.97)
20000	1102.00 (2.00)	2 (0)	9 (0)	7.47 (1.97)

TABLE B.3: Disease models description

Disease Models	Description	$\boldsymbol{K}_1$	$\boldsymbol{K}_2$	σ_1^2	σ_2^2	σ_0^2
$S_1 : L + L$	Linear additive effects	$k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$	$k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$	1	1	1
$S_2 : R + R$	Non-linear effects.	$k_{rbf}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\frac{1}{2}\|\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\|_2^2\right]$	$k_{rbf}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\frac{1}{2}\|\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\|_2^2\right]$	1	1	1
$S_3 : P + P$	Pair-wise interaction effects	$k_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = (< \boldsymbol{x}_1, \boldsymbol{x}_2 >)^2$	$k_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = (< \boldsymbol{x}_1, \boldsymbol{x}_2 >)^2$	1	1	1
$S_4 : L + R$	Linear and non-linear effects	$k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$	$k_{rbf}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left[-\frac{1}{2}\|\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|\|_2^2\right]$	1	1	1
$S_5 : L + P$	Linear and pair-wise interaction	$k_l(\boldsymbol{x}_1, \boldsymbol{x}_2) =< \boldsymbol{x}_1, \boldsymbol{x}_2 >$	$k_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = (< \boldsymbol{x}_1, \boldsymbol{x}_2 >)^2$	1	1	1

TABLE B.4: The number of selected regions and the number of causal regions within the selected regions under different disease models ($n = 1000$)

Disease	Number of Total Regions Selected (Number of Causal Regions Selected)			
Models	MKLMM	MKLMM2	MKLMM9	Hybrid Screening Rule
$S_1 : L + L$	290.77 (2.00)	2 (0)	9 (0)	8.10 (2.00)
$S_2 : R + R$	295.83 (1.97)	2 (0)	9 (0)	9.47 (1.83)
$S_3 : P + P$	296.60 (1.73)	2 (0)	9 (0)	5.94 (2.00)
$S_4 : L + R$	293.63 (1.93)	2 (0)	9 (0)	7.30 (1.87)
$S_5 : L + P$	292.83 (1.67)	2 (0)	9 (0)	5.20 (1.93)

TABLE B.5: The chance of selecting the most appropriate kernels under different disease models ($n = 1000$)

Disease	MKLMM		Hybrid Screening Rule	
Models	1st Causal Region	2nd Causal Region	1st Causal Region	2nd Causal Region
$S_1 : L + L$	1.00	1.00	1.00	1.00
$S_2 : R + R$	0	0	0.83	0.97
$S_3 : P + P$	0	0	1.00	0.94
$S_4 : L + R$	0.97	0	1.00	0.80
$S_5 : L + P$	1.00	0	1.00	0.93

*Note: results from MKLMM2 and MKLMM9 are not reported. Neither of the causal regions can be kept, and thus the chances of selecting the most appropriate kernels are 0.

TABLE B.6: The chances of selecting genes by HpLMMGMM for FDG

Genes	Chromosome	Start Position	End Position	Probability(FDG)
APOC1	19	45417920	45422606	0.98
APOE	19	45409038	45412650	0.96
FADS3	11	61640994	61659017	0.96
IQCF3	3	51860898	51864874	0.68
ZNF805	19	57752052	57774106	0.67
FAM193B	5	176946789	176981548	0.60
PSMD6	3	63996224	64009686	0.41
U2AF1	21	44513065	44527688	0.32
LOC728463	1	218517537	218519020	0.27
C3orf72	3	138666075	138672830	0.27
TRAPPC10	21	45432205	45526432	0.26
FSTL1	3	120113060	120169918	0.26
LOC101927364	16	51796429	51806557	0.24
ZNF320	19	53379424	53394599	0.24
TOMM7	7	22852250	22862471	0.23
OR10AD1	12	48596121	48597075	0.21
ZBTB1	14	64971291	65000408	0.20
LILRB1	19	55128383	55148981	0.17
PAPD4	5	78908242	78982471	0.15
UGT2B7	4	69962192	69978705	0.13
LDOC1L	22	44888449	44894005	0.13

TABLE B.6: The chances of selecting genes by HpLMMGMM for FDG *(continued)*

Genes	Chromosome	Start Position	End Position	Probability(FDG)
RFK	9	79000432	79009444	0.12
PTEN	10	89623194	89728532	0.11
GLUL	1	182350838	182361341	0.11
GPR108	19	6729924	6737633	0.11
H1FNT	12	48722762	48724062	0.11
FAM87B	1	752750	755214	0.10
NKX3-1	8	23536205	23540450	0.09
NDUFAB1	16	23592334	23607639	0.07
FILIP1L	3	99551987	99833357	0.07
CLDN24	4	184242916	184243579	0.05
DNAAF2	14	50091891	50101948	0.05
H19	11	2016405	2019065	0.05
ABCG8	2	44066102	44105605	0.05
LTB	6	31548335	31550202	0.04
C7orf55	7	139025195	139031065	0.04
TOR2A	9	130493802	130497628	0.04
TTC29	4	147628178	147867034	0.04
C16orf54	16	29753785	29757340	0.04
C20orf141	20	2795632	2796476	0.04
ZNF140	12	133657036	133684258	0.04
CCDC178	18	30517365	31020685	0.04
EPB41L4A-AS2	5	111755279	111756677	0.04
FKSG29	13	100003673	100004281	0.04
MFAP5	12	8798539	8815433	0.03
MYCBP	1	39328161	39339050	0.03
BAZ1A	14	35221936	35344853	0.03
PMCHL2	5	70671611	70681820	0.03
RPL32P3	3	129101676	129118282	0.03
SLC52A3	20	740723	749228	0.03
CFL2	14	35179587	35184029	0.03
F11R	1	160965000	160991133	0.03
HP09053	3	99535475	99542709	0.03
KCNH3	12	49932939	49952077	0.03

TABLE B.6: The chances of selecting genes by HpLMMGMM for FDG *(continued)*

Genes	Chromosome	Start Position	End Position	Probability(FDG)
KIAA1683	19	18367905	18385319	0.03
ARL6IP4	12	123464606	123467460	0.02
LOC101927332	15	98626207	98631982	0.02
PSMD6-AS2	3	63989697	63997917	0.02
RCN3	19	50030874	50046890	0.02
RUSC1-AS1	1	155290250	155293938	0.02
STAG3L4	7	66767624	66786513	0.02
SYCE1	10	135367403	135382876	0.02
TFAP2A-AS1	6	10412550	10416402	0.02
TOMM40	19	45394476	45406946	0.02
ZBTB25	14	64953554	64970554	0.02
CASP1	11	104896236	104905884	0.02
EFNA2	19	1286152	1301429	0.02
FAM168A	11	73111522	73309234	0.02
LINC00115	1	761585	762902	0.02
ARCN1	11	118443101	118473747	0.02
LOC646938	15	79044378	79045734	0.01
LOC102723641	17	72966783	72971823	0.01
LY75	2	160659867	160761267	0.01
ASPN	9	95218488	95244844	0.01
MARC2	1	220921675	220957596	0.01
MGAT4C	12	86373036	87232681	0.01
ACADM	1	76190031	76229363	0.01
MOBP	3	39509063	39570988	0.01
NDUFAF3	3	49057907	49060926	0.01
OR8G2	11	124095343	124096368	0.01
BPESC1	3	138823026	138844005	0.01
PROX1	1	214161277	214214847	0.01
PTRHD1	2	25013135	25016251	0.01
RPL37A	2	217363519	217366188	0.01
SEMA7A	15	74701629	74726299	0.01
C5orf20	5	134779903	134783038	0.01
C8G	9	139839697	139841426	0.01

TABLE B.6: The chances of selecting genes by HpLMMGMM for FDG *(continued)*

Genes	Chromosome	Start Position	End Position	Probability(FDG)
SPATS2	12	49760687	49921207	0.01
SPINT4	20	44350987	44354335	0.01
C9orf78	9	132589563	132597572	0.01
SUV420H1	11	67923506	67980784	0.01
TBC1D23	3	99979660	100044096	0.01
TEDDM1	1	182367251	182369751	0.01
TIMM21	18	71815745	71826204	0.01
TMEM14A	6	52535883	52551385	0.01
TMEM25	11	118401802	118417313	0.01
TNFSF8	9	117655622	117692875	0.01
TOR1A	9	132575220	132586441	0.01
TRPM4	19	49661015	49715098	0.01
TTC26	7	138818489	138876732	0.01
C16orf78	16	49407807	49433319	0.01
C22orf39	22	19428409	19435755	0.01
CHRNA4	20	61974661	61992748	0.01
CLEC4C	12	7882010	7902069	0.01
CMSS1	3	99536677	99897476	0.01
DNAJC25	9	114393631	114416631	0.01
DUSP16	12	12626215	12715448	0.01
ETNPPL	4	109663201	109684235	0.01
FAM170B-AS1	10	50329883	50359592	0.01
FJX1	11	35639734	35642421	0.01
FOXL2	3	138663065	138665982	0.01
GPC5-AS1	13	93353641	93373867	0.01
ANKRD20A1	9	67926760	67969840	0.01
GRB2	17	73314156	73401790	0.01
HID1	17	72946838	72968900	0.01
HIF1A	14	62162118	62214977	0.01
ICOSLG	21	45642877	45660887	0.01
ANXA2R	5	43039181	43040447	0.01
IL10	1	206940947	206945839	0.01
KIAA1024L	5	129083883	129100756	0.01

TABLE B.7: The chances of selecting genes by HpLMMGMM for AV45

Genes	Chromosome	Start Position	End Position	Probability(AV45)
TOMM40	19	45394476	45406946	1.00
APOC1	19	45417920	45422606	1.00
APOE	19	45409038	45412650	0.94
SPRR2G	1	153122057	153123427	0.49
LOC400863	21	35321229	35336262	0.46
CGB7	19	49557530	49558997	0.37
VOPP1	7	55538300	55640200	0.30
CPB2-AS1	13	46626982	46675482	0.30
PIN1P1	1	70385004	70386000	0.16
LINC00545	13	31456698	31457532	0.15
LINC01070	13	112851646	112855316	0.14
LOC101928269	21	37326976	37376965	0.12
PVRL2	19	45349392	45392485	0.12
OR5H14	3	97868229	97869162	0.11
VAMP1	12	6571403	6579843	0.11
ADAM28	8	24151579	24212726	0.11
NKX3-1	8	23536205	23540450	0.10
PRR23B	3	138737872	138739768	0.10
C16orf82	16	27078218	27080487	0.10
HGC6.3	6	168376603	168377619	0.09
LINC00316	21	46758504	46761905	0.09
FBXO33	14	39865576	39901704	0.07
GPATCH1	19	33571785	33621318	0.07
NT5C3A	7	33053724	33102409	0.06
PAFAH1B2	11	117014999	117048889	0.06
SEC61B	9	101984569	101992901	0.06
PRR23A	3	138722803	138725110	0.05
TPSAB1	16	1290677	1292555	0.05
TRAF3IP1	2	239229184	239309541	0.05
DHRS7	14	60611499	60632211	0.05
GPX2	14	65405869	65409623	0.05
ARL2BP	16	57279037	57287545	0.04
LOC100652768	11	117066328	117072630	0.04

TABLE B.7: The chances of selecting genes by HpLMMGMM for AV45 *(continued)*

Genes	Chromosome	Start Position	End Position	Probability(AV45)
NT5DC3	12	104166080	104234975	0.04
ZNF773	19	58011308	58019510	0.04
DDR1	6	30851860	30867933	0.04
APOA4	11	116691417	116694011	0.04
LOC100270746	6	26987144	26988085	0.02
LOC100996385	5	175476680	175489058	0.02
MYOC	1	171604556	171621773	0.02
PCNT	21	47744035	47865682	0.02
PDE6D	2	232597134	232646037	0.02
PSORS1C1	6	31082607	31107869	0.02
C2CD2L	11	118978059	118987834	0.02
SLC1A6	19	15060844	15121455	0.02
SLC30A5	5	68389775	68426899	0.02
SPATA13-AS1	13	24826886	24828577	0.02
TPTE2P5	13	41371120	41495886	0.02
U2SURP	3	142720371	142779567	0.02
C19orf48	19	51300949	51308110	0.02
CCDC37	3	126113781	126155398	0.02
CDSN	6	31082864	31088252	0.02
CFL1	11	65622284	65625804	0.02
CYFIP2	5	156693089	156822606	0.02
FMO5	1	146655883	146697230	0.02
FOLR3	11	71846770	71850934	0.02
FRRS1	1	100174258	100231349	0.02
H2AFX	11	118964584	118966177	0.02
HDGFL1	6	22569677	22570750	0.02
KLK4	19	51409607	51413994	0.02
LINC00471	2	232373136	232379050	0.02
LOC285074	2	87257797	87303536	0.01
LOC100128573	19	7537722	7538247	0.01
LOC101927332	15	98626207	98631982	0.01
LOC101928295	19	49871961	49891338	0.01
LOC101928744	7	41004276	41019537	0.01

TABLE B.7: The chances of selecting genes by HpLMMGMM for AV45 *(continued)*

Genes	Chromosome	Start Position	End Position	Probability(AV45)
LRIT1	10	85991275	86001217	0.01
LRRC25	19	18501953	18508415	0.01
LTB	6	31548335	31550202	0.01
MAD2L1	4	120980578	120988013	0.01
MAMDC2-AS1	9	72768049	72790804	0.01
MYCNOS	2	16076386	16081845	0.01
NEFH	22	29876180	29887277	0.01
NTF4	19	49564396	49567124	0.01
OR5E1P	11	7870597	7871118	0.01
BNIP2	15	59955061	59981642	0.01
PLAC9	10	81892257	81904784	0.01
POM121L12	7	53103348	53104618	0.01
PRR7	5	176873795	176883287	0.01
RAB3GAP2	1	220321609	220445843	0.01
RAB36	22	23487512	23506531	0.01
C1orf68	1	152691997	152692905	0.01
RPGRIP1L	16	53633817	53737771	0.01
C1S	12	7167979	7178335	0.01
SEC61A2	10	12171639	12211957	0.01
SETX	9	135136826	135230372	0.01
SIK3	11	116714117	116969131	0.01
SNAPC3	9	15422781	15461627	0.01
SNX2	5	122110690	122170234	0.01
SNX15	11	64794879	64808044	0.01
SNX24	5	122181159	122344902	0.01
TAOK2	16	29985187	30003582	0.01
TAPBPL	12	6561176	6571488	0.01
TMEM155	4	122680084	122686340	0.01
TMEM159	16	21169911	21191937	0.01
TMEM182	2	103378489	103434138	0.01
TSSK1B	5	112768250	112770728	0.01
UBA6-AS1	4	68566995	68588222	0.01
ZBTB40	1	22778343	22857650	0.01

TABLE B.7: The chances of selecting genes by HpLMMGMM for AV45 *(continued)*

Genes	Chromosome	Start Position	End Position	Probability(AV45)
CACNA1C-AS4	12	2329702	2332647	0.01
CCDC6	10	61548505	61666414	0.01
CELF5	19	3224700	3297073	0.01
CHRNA4	20	61974661	61992748	0.01
COA1	7	43670750	43769140	0.01
COL4A2-AS1	13	111154922	111160526	0.01
FAM192A	16	57186377	57219976	0.01
FBXO28	1	224301788	224349749	0.01
GCNT4	5	74323288	74326724	0.01
ABCC6P1	16	18582569	18609607	0.01
GS1-204I12.1	1	185527511	185597620	0.01
GTF2H4	6	30875976	30881880	0.01
HAVCR1	5	156456530	156485970	0.01
HINFP	11	118992232	119005765	0.01
HLA-DPA1	6	33032345	33048555	0.01
IL36A	2	113763448	113765621	0.01
INHBB	2	121103718	121109383	0.01
KBTBD8	3	67048726	67061632	0.01
KIAA1024L	5	129083883	129100756	0.01
APOC3	11	116700623	116703787	0.01
LAMP3	3	182840002	182880667	0.01
LAMTOR5	1	110943876	110950546	0.01
LAMTOR5-AS1	1	110950430	110958896	0.01
LINC00602	6	166401038	166403103	0.01
LINC01007	7	101206034	101212286	0.01

B.3 Additional figures

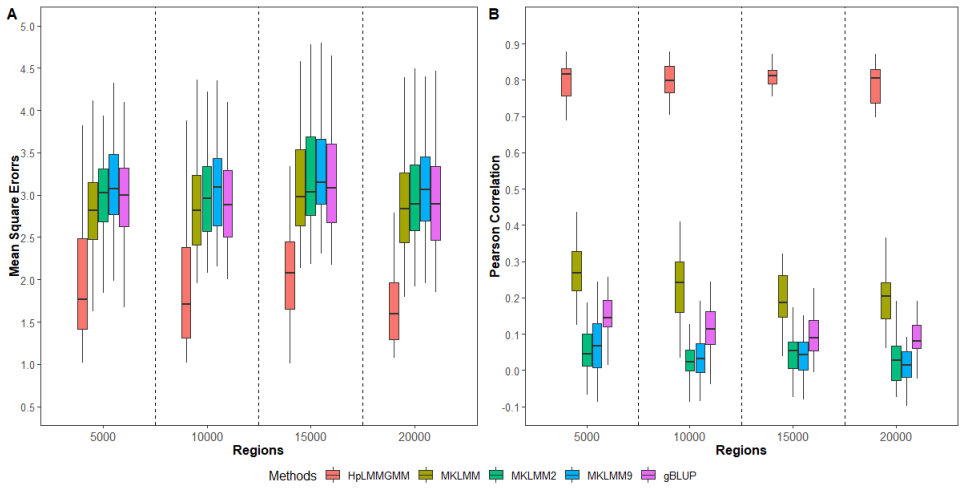

FIGURE B.1: The impact of data dimension on Pearson correlations and MSEs ($n = 1000$)

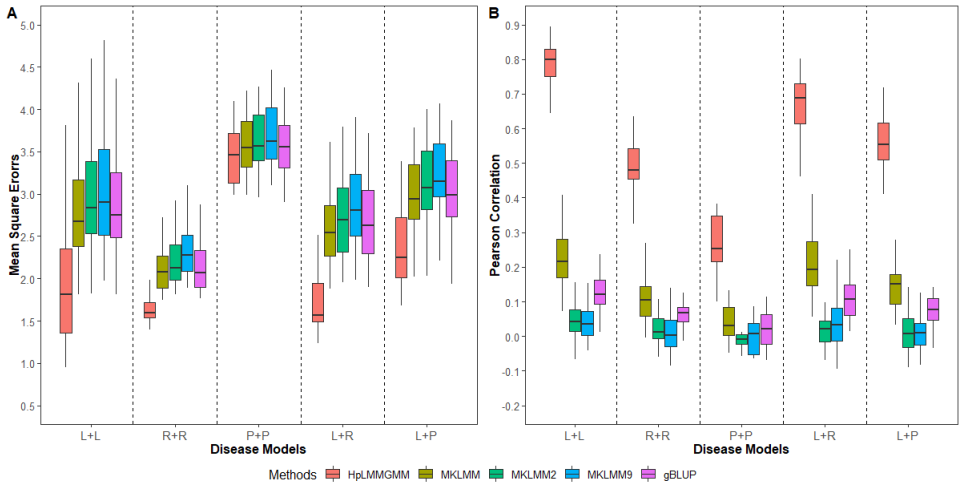

FIGURE B.2: The impact of disease models. $L + L$: genetic variants on both regions have linear additive effects. $R + R$: predictors from both regions have non-linear predictive effects. $P + P$: both regions harbor variants with pair-wise interaction effects. $L + R$: genetic variants on the first and second regions have linear additive and non-linear effects, respectively. $L+P$: predictors on the first and second regions have linear additive and pair-wise interaction effects, respectively ($n = 1000$)

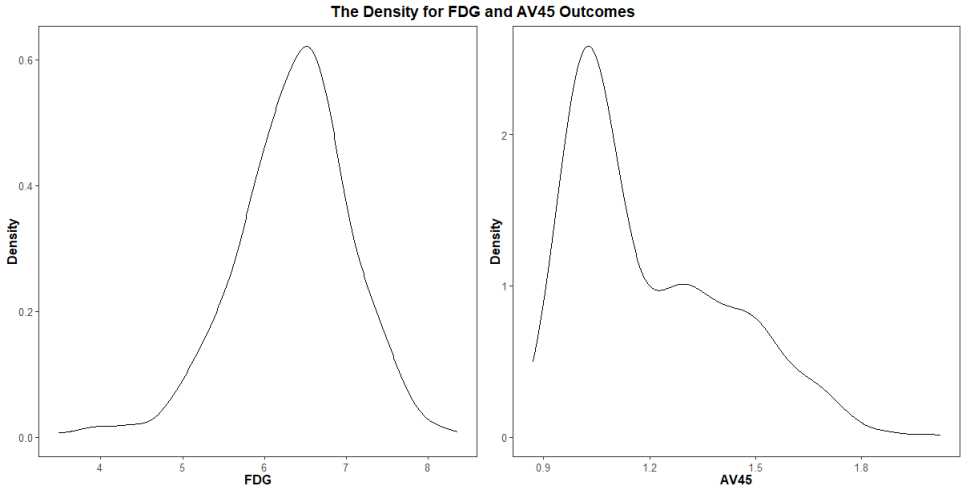

FIGURE B.3: The distribution for FDG and AV45

Appendix C

A penalized linear mixed model with generalized method of moments estimators for the prediction analysis of multi-omics data

C.1 Additional tables

TABLE C.1: The effect sizes for the first simulation

Parameters	1st Region	2nd Region	3rd Region
γ	0.8	1	1.2
σ_g^2	0.45	0.5	0.55
σ_m^2	0.45	0.5	0.55

TABLE C.2: The chances of selecting causal regions as the number of noise regions increases ($n = 1000$)

Regions	Gene Expression Data		Genomic Data		Methylation Data	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
10	1.000	0.926	0.997	0.878	0.998	0.960
25	1.000	0.968	0.998	0.893	0.998	0.964
50	1.000	0.984	0.996	0.912	0.998	0.969
75	1.000	0.988	0.993	0.920	0.999	0.971
100	1.000	0.989	0.996	0.927	0.999	0.974

TABLE C.3: The effect sizes for the second simulation

Parameters	E	G	M	GM	G+M	E+G	E+M
γ_1	0.8	0	0	0	0	0.8	0.8
γ_2	1.0	0	0	0	0	1.0	1.0
γ_3	1.2	0	0	0	0	1.2	1.2
$\sigma^2_{g,1}$	0	0.45	0	0	0.45	0.45	0
$\sigma^2_{g,2}$	0	0.5	0	0	0.5	0.5	0
$\sigma^2_{g,3}$	0	0.55	0	0	0.55	0.55	0
$\sigma^2_{m,1}$	0	0	0.45	0	0.45	0	0.45
$\sigma^2_{m,2}$	0	0	0.5	0	0.5	0	0.5
$\sigma^2_{m,3}$	0	0	0.55	0	0.55	0	0.55
$\sigma^2_{gm,1}$	0	0	0	0.45	0	0	0
$\sigma^2_{gm,2}$	0	0	0	0.5	0	0	0
$\sigma^2_{gm,3}$	0	0	0	0.55	0	0	0

TABLE C.4: The chances of selecting causal regions under different disease models ($n = 1000$)

Disease	Gene Expression Data		Genomic Data		Methylation Data	
Models	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
$S_1 : E^{\text{a}}$	1.000	0.982	–	0.946	–	0.934
$S_2 : G^{\text{b}}$	–	0.994	1.000	0.908	–	0.984
$S_3 : M^{\text{c}}$	–	0.995	–	0.986	1.000	0.987
$S_4 : GM^{\text{d}}$	–	0.996	0.962	0.942	0.949	0.985
$S_5 : G + M^{\text{e}}$	–	0.996	0.996	0.931	0.996	0.985
$S_6 : E + G^{\text{f}}$	1.000	0.980	0.999	0.871	–	0.956
$S_7 : E + M^{\text{g}}$	1.000	0.979	–	0.965	1.000	0.961

[a] Only gene expression data is causal.
[b] Only genomic data is causal.
[c] Only methylation data is causal.
[d] Only the interaction between genomic and methylation data is causal.
[e] Both genomic and methylation data are causal.
[f] Both gene expression data and genomic data are causal.
[g] Both gene expression data and methylation data are causal.

TABLE C.5: The chances of selecting genes by MpLMMGMM for FDG and AV45

Genes	Chromosome	Start Position	End Position	FDG[a]	FDG[b]	AV45[c]	AV45[d]
COL11A1	1	103342022	103574052	0	0	0	0
CR1	1	207669472	207815110	0	0	0	0.01

TABLE C.5: The chances of selecting genes by MpLMMGMM for FDG and AV45 *(continued)*

Genes	Chromosome	Start Position	End Position	FDG[a]	FDG[b]	AV45[c]	AV45[d]
CR1L	1	207818457	207897036	0	0	0	0
FCER1G	1	161185086	161189038	0	0	0	0
FLVCR1	1	213031596	213072705	0	0.01	0.05	0
FLVCR1-AS1	1	213029945	213031480	0	0	0	0
GBP2	1	89571815	89591842	0	0	0	0
HSD11B1	1	209859524	209908295	0	0	0	0
NGF	1	115828536	115880857	0	0	0	0
PARP1	1	226548391	226595801	0.01	0	0	0
POU2F1	1	167190065	167396582	0	0	0	0
BIN1	2	127805598	127864903	0	0	0	0.06
LHCGR	2	48913912	48982880	0	0	0	0
LRP2	2	169983618	170219122	0	0	0	0
APOD	3	195295572	195311076	0.01	0	0	0
GSK3B	3	119540801	119813264	0	0	0	0.02
SST	3	187386693	187388201	0	0	0	0
ALB	4	74269971	74287129	0.3	0	0.07	0
COL25A1	4	109731876	110223799	0	0	0	0
ADRB2	5	148206155	148208197	0.01	0	0	0
ARSB	5	78073036	78282357	0	0.01	0	0
FGF1	5	141971742	142077635	0.24	0.01	0	0
FGF10	5	44305096	44388784	0	0	0	0
FGF18	5	170846666	170884630	0	0	0	0
NDUFS4	5	52856464	52979171	0	0	0.05	0
AGER	6	32148744	32152099	0	0.01	0	0
HSPA1A	6	31783290	31785719	0	0	0	0
MICA	6	31367560	31383092	0.03	0	0.11	0
MICAL1	6	109765265	109787171	0.05	0	0	0
TBP	6	170863420	170881958	0.01	0	0	0
TBPL1	6	134273307	134308638	0	0	0	0
TREM2	6	41126243	41130924	0	0	0	0
CAV1	7	116164838	116201239	0	0	0	0
PON3	7	94989183	95025687	0	0	0.06	0.03
RELN	7	103112230	103629963	0	0	0	0
ADAM9	8	38854504	38962779	0	0	0.28	0
NAT1	8	18027970	18081198	0	0	0	0
NRG1	8	31497267	32622558	0	0.03	0	0
DAPK1	9	90112142	90323549	0	0	0	0

TABLE C.5: The chances of selecting genes by MpLMMGMM for FDG and AV45 *(continued)*

Genes	Chromosome	Start Position	End Position	FDG[a]	FDG[b]	AV45[c]	AV45[d]
DFNB31	9	117164359	117267736	0	0	0	0
HSPA5	9	127997126	128003666	0	0	0	0
POMT1	9	134378288	134399193	0.58	0	0.02	0
RXRA	9	137218308	137332432	0	0	0	0
TLR4	9	120466452	120479769	0	0	0	0
CACNB2	10	18429605	18830688	0.01	0	0	0
MINPP1	10	89264222	89313218	0	0	0	0
TET1	10	70320116	70454239	0	0	0	0
TFAM	10	60144902	60158990	0	0.02	0	0.01
HBG2	11	5274420	5276011	0.48	0	0	0
ATF7	12	53901639	54020199	0.02	0	0	0
ATF7IP	12	14518565	14655869	0	0	0.01	0
OLR1	12	10310898	10324790	0	0	0	0
SLC11A2	12	51373565	51422058	0	0	0	0
KLF5	13	73629113	73651680	0	0	0	0
CINP	14	102814618	102829253	0	0	0	0.03
GNPNAT1	14	53241910	53258386	0.01	0	0	0
HNRNPC	14	21677295	21737638	0	0	0.02	0
MTHFD1	14	64854758	64926725	0.06	0	0	0
PNP	14	20937537	20946165	0.01	0	0	0
SEL1L	14	81937890	82000205	0	0	0	0
SERPINA1	14	94843083	94857029	0.05	0	0	0
SERPINA3	14	95078713	95090390	0	0	0	0.01
SERPINA4	14	95027756	95036250	0	0.06	0	0.02
SERPINA5	14	95047705	95059457	0	0	0	0
SERPINA6	14	94770584	94789688	0.01	0	0.01	0.02
SERPINA9	14	94929057	94942670	0	0	0	0
SERPINA10	14	94749649	94759608	0	0	0	0
SERPINA11	14	94908800	94919122	0	0	0.06	0
SERPINA12	14	94953619	94984181	0	0	0	0
SERPINA13P	14	95107061	95113331	0.02	0	0	0
CHRNA3	15	78885394	78913637	0	0.06	0	0
MEF2A	15	100106132	100256629	0	0.03	0	0
MEFV	16	3292027	3306627	0	0	0	0
UBE2I	16	1359153	1377019	0	0	0.02	0
CCL3	17	34415602	34417506	0	0	0	0
CDK5R1	17	30814104	30818271	0	0	0	0

TABLE C.5: The chances of selecting genes by MpLMMGMM for FDG and AV45 *(continued)*

Genes	Chromosome	Start Position	End Position	FDG[a]	FDG[b]	AV45[c]	AV45[d]
COX10	17	13972718	14111996	0	0	0	0
PNMT	17	37824233	37826728	0	0.02	0	0
APOC1	19	45417920	45422606	1	0.01	1	0
APOE	19	45409038	45412650	0.9	0	0.95	0
GNA11	19	3094407	3124000	0.04	0.01	0	0
TOMM40	19	45394476	45406946	0.94	0.01	1	0.01
DOPEY2	21	37536838	37666572	0	0	0	0
MCM3AP	21	47655038	47705308	0	0	0	0
MCM3AP-AS1	21	47649144	47671615	0	0	0	0.01
NCAM2	21	22370632	22912517	0.01	0	0	0.01
S100B	21	48018530	48025035	0.02	0	0	0
SAMSN1	21	15857548	15955723	0	0	0	0.01
SEPT3	22	42372930	42394225	0	0	0	0

[a] The probability of genes being selected for FDG based on genetic data.

[b] The probability of genes being selected for FDG based on gene expression data.

[c] The probability of genes being selected for AV45 based on genetic data.

[d] The probability of genes being selected for AV45 based on gene expression data.

C.2 Additional figures

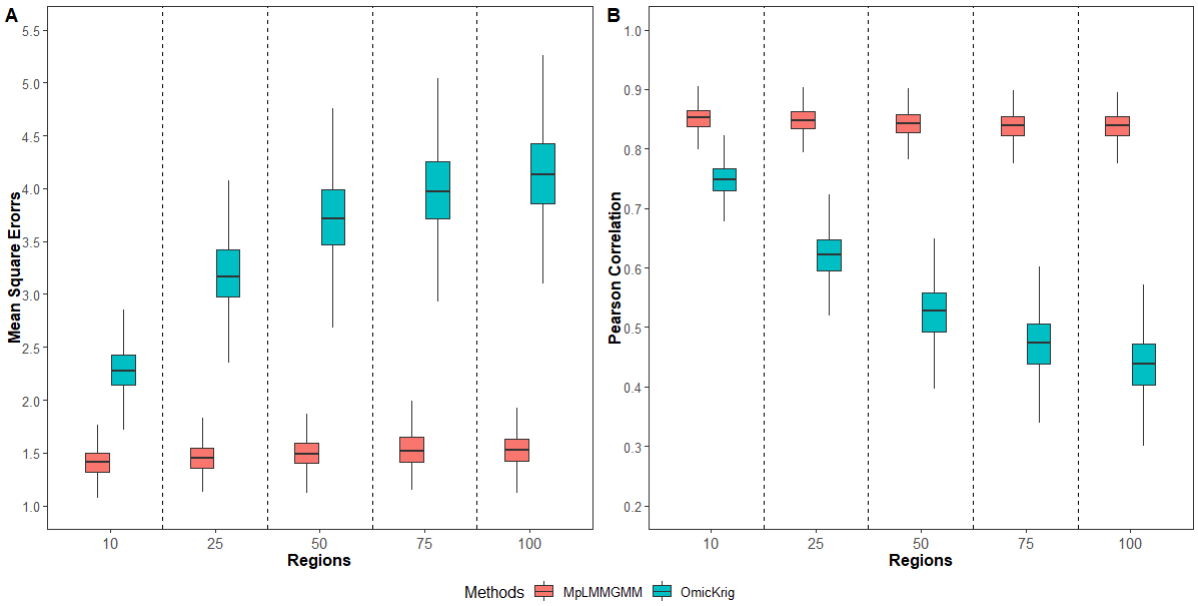

FIGURE C.1: The impact of the number of noise regions on Pearson correlations and MSEs ($n = 1000$)

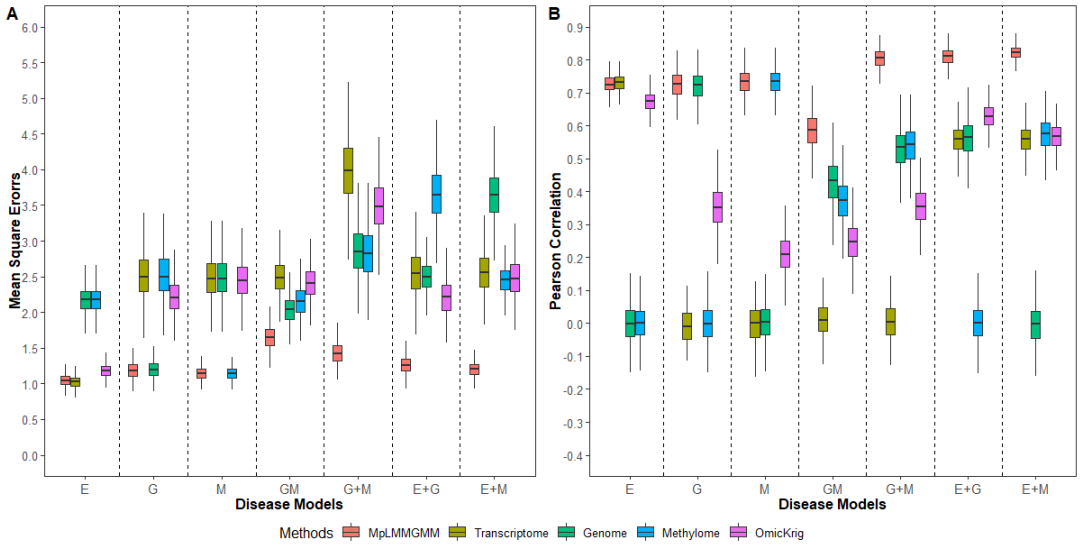

FIGURE C.2: The impact of disease models ($n = 1000$)

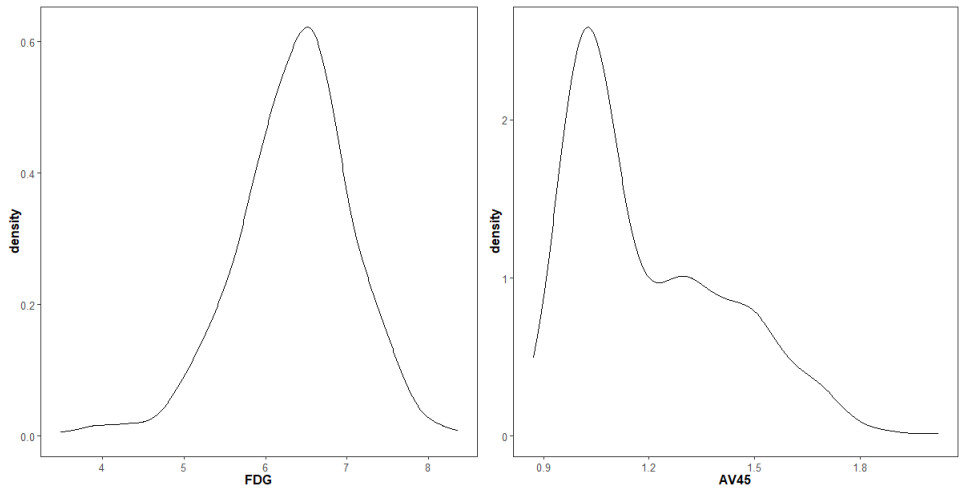

FIGURE C.3: The distribution of FDG and AV45

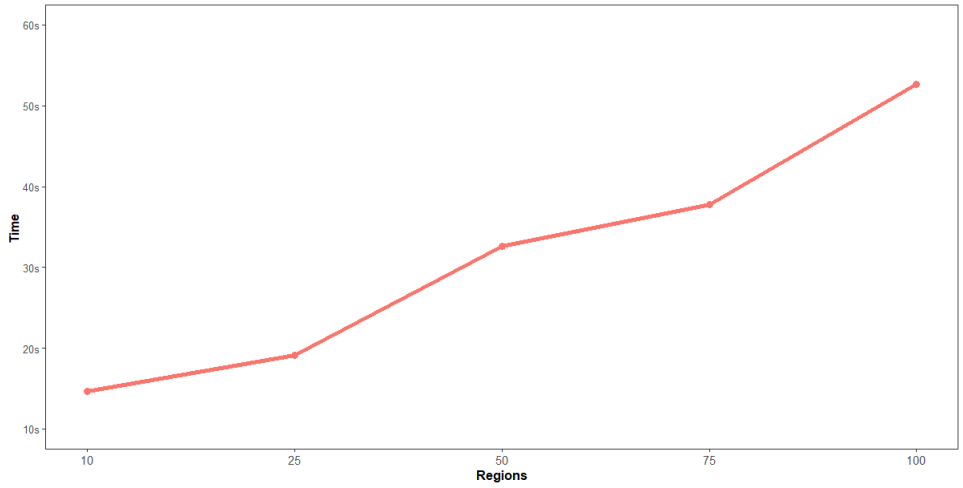

FIGURE C.4: The computational time as the number of random effects increases for MpLMMGMM ($n = 500$)

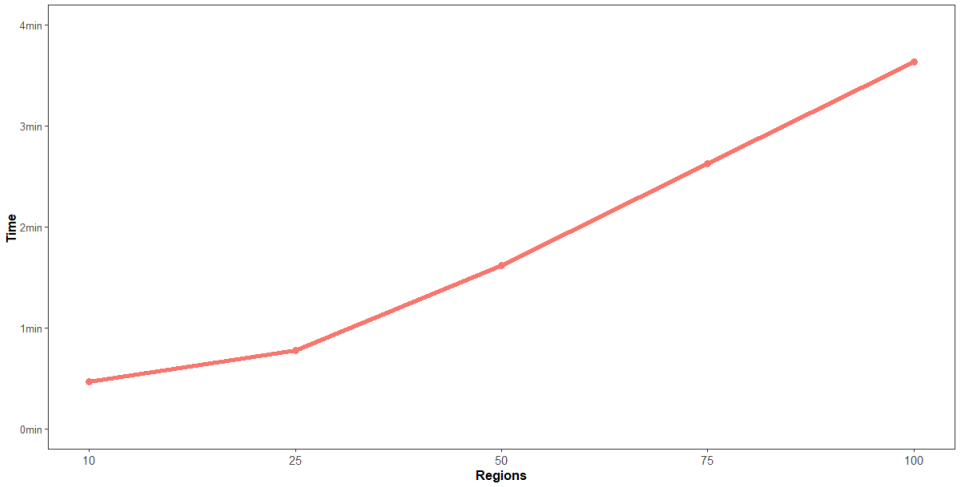

FIGURE C.5: The computational time as the number of random effects increases for MpLMMGMM ($n = 1000$)

Bibliography

Abraham, Gad and Michael Inouye (2015). "Genomic risk prediction of complex human disease and its clinical application". In: *Current Opinion in Genetics & Development* 33, pp. 10–16.

Akavia, Uri David et al. (2010). "An integrated approach to uncover drivers of cancer". In: *Cell* 143.6, pp. 1005–1017.

Ashley, Euan A (2015). "The precision medicine initiative: a new national effort". In: *The Journal of the American Medical Association* 313.21, pp. 2119–2120.

Badano, Jose L et al. (2006). "Dissection of epistasis in oligogenic Bardet–Biedl syndrome". In: *Nature* 439.7074, pp. 326–330.

Bagnoli, Silvia et al. (2013). "TOMM40 polymorphisms in Italian Alzheimer's disease and frontotemporal dementia patients". In: *Neurological Sciences* 34.6, pp. 995–998.

Bersanelli, Matteo et al. (2016). "Methods for the integration of multi-omics data: mathematical aspects". In: *BMC Bioinformatics* 17.2, pp. 167–177.

Bertram, Lars et al. (2007). "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database". In: *Nature Genetics* 39.1, pp. 17–23.

Boekel, Jorrit et al. (2015). "Multi-omic data analysis using Galaxy". In: *Nature Biotechnology* 33.2, pp. 137–139.

Bonnet, Eric et al. (2015). "Integrative multi-omics module network inference with Lemon-Tree". In: *PLoS Computational Biology* 11.2, e1003983.

Buil, Alfonso et al. (2015). "Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins". In: *Nature Genetics* 47.1, pp. 88–91.

Byrnes, Andrea E et al. (2013). "The value of statistical or bioinformatics annotation for rare variant association with quantitative trait". In: *Genetic Epidemiology* 37.7, pp. 666–674.

Cancer Genome Atlas Research Network John N Weinstein, Eric A Collisson et al. (2013). "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nature Genetics* 45, pp. 1113–1120.

Chatterjee, Nilanjan et al. (2013). "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies". In: *Nature Genetics* 45.4, pp. 400–405.

Chatterjee, Nilanjan et al. (2016). "Developing and evaluating polygenic risk prediction models for stratified disease prevention". In: *Nature Reviews Genetics* 17.7, pp. 392–406.

Chaudhary, Kumardeep et al. (2018). "Deep learning–based multi-omics integration robustly predicts survival in liver cancer". In: *Clinical Cancer Research* 24.6, pp. 1248–1259.

Chen, Jinyu and Shihua Zhang (2016). "Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data". In: *Bioinformatics* 32.11, pp. 1724–1732.

Collado-Hidalgo, Alicia et al. (2008). "Cytokine gene polymorphisms and fatigue in breast cancer survivors: Early findings". In: *Brain, Behavior, and Immunity* 22.8, pp. 1197–1200.

Collins, Francis S and Harold Varmus (2015). "A new initiative on precision medicine". In: *The New England Journal of Medicine* 372.9, pp. 793–795.

Corder, Elizabeth H et al. (1993). "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families". In: *Science* 261.5123, pp. 921–923.

Cuingnet, Rémi et al. (2011). "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database". In: *Neuroimage* 56.2, pp. 766–781.

Cun, Yupeng and Holger Fröhlich (2013). "Network and data integration for biomarker signature discovery via network smoothed t-statistics". In: *PLoS One* 8.9, e73074.

Dandis, Rana et al. (2020). "A tutorial on dynamic risk prediction of a binary outcome based on a longitudinal biomarker". In: *Biometrical Journal* 62.2, pp. 398–413.

Das, Partha M and Rakesh Singal (2004). "DNA methylation and cancer". In: *Journal of Clinical Oncology* 22.22, pp. 4632–4642.

De Los Campos, Gustavo et al. (2010). "Predicting genetic predisposition in humans: the promise of whole-genome markers". In: *Nature Reviews Genetics* 11.12, pp. 880–886.

De Los Campos, Gustavo et al. (2013). "Prediction of complex human traits using the genomic best linear unbiased predictor". In: *PLoS Genetics* 9.7, e1003608.

Duijn, Cornelia M van et al. (1994). "Apolipoprotein E4 allele in a population–based study of early–onset Alzheimer's disease". In: *Nature Genetics* 7.1, pp. 74–78.

Dunson, David B (2001). "Commentary: practical advantages of Bayesian analysis of epidemiologic data". In: *American Journal of Epidemiology* 153.12, pp. 1222–1226.

El-Moghazy, MM et al. (2015). "Genetic and non genetic factors affecting body weight traits in Zaraibi goat in Egypt". In: *Journal of Agricultural Research Kafr El-Shaikh University* 41.1, pp. 27–40.

Fan, Jianqing and Runze Li (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties". In: *Journal of the American Statistical Association* 96.456, pp. 1348–1360.

Fan, Jianqing and Jinchi Lv (2008). "Sure independence screening for ultrahigh dimensional feature space". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5, pp. 849–911.

Fan, Jianqing et al. (2009). "Ultrahigh dimensional feature selection: beyond the linear model". In: *The Journal of Machine Learning Research* 10, pp. 2013–2038.

Fan, Jianqing et al. (2011). "Nonparametric independence screening in sparse ultra-high-dimensional additive models". In: *Journal of the American Statistical Association* 106.494, pp. 544–557.

Fisher, Ronald Aylmer (1992). *Statistical Methods for Research Workers*. Springer.

Friedman, Jerome et al. (2010). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* 33.1, p. 1.

Ghaoui, Laurent El et al. (2010). "Safe feature elimination for the Lasso and sparse supervised learning problems". In: *arXiv preprint arXiv:1009.4219*.

Ghosh, Debashis and Arul M Chinnaiyan (2005). "Classification and selection of biomarkers in genomic data using Lasso". In: *BioMed Research International* 2005.2, pp. 147–154.

Gianola, Daniel (2013). "Priors in whole-genome regression: the Bayesian alphabet returns". In: *Genetics* 194.3, pp. 573–596.

Gianola, Daniel et al. (2018). "Prediction of complex traits: robust alternatives to best linear unbiased prediction". In: *Frontiers in Genetics* 9, p. 195.

Goffe, William L et al. (1994). "Global optimization of statistical functions with simulated annealing". In: *Journal of Econometrics* 60.1-2, pp. 65–99.

Graff-Radford, Neill R et al. (2002). "Association between apolipoprotein E genotype and Alzheimer disease in African American subjects". In: *Archives of Neurology* 59.4, pp. 594–600.

Grapov, Dmitry et al. (2018). "Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine". In: *OMICS: A Journal of Integrative Biology* 22.10, pp. 630–636.

Habier, David et al. (2011). "Extension of the Bayesian alphabet for genomic selection". In: *BMC Bioinformatics* 12.1, pp. 1–12.

Hai, Yang and Yalu Wen (2020). "A Bayesian linear mixed model for prediction of complex traits". In: *Bioinformatics* 36.22-23, pp. 5415–5423.

Harris, BL et al. (2008). "Genomic selection in New Zealand and the implications for national genetic evaluation". In: *Proc. Interbull Meeting, Niagara Falls, Canada.*

Hasin, Yehudit et al. (2017). "Multi-omics approaches to disease". In: *Genome Biology* 18.1, p. 83.

Henderson, CR (1985). "MIVQUE and REML estimation of additive and nonadditive genetic variances". In: *Journal of Animal Science* 61.1, pp. 113–121.

Hill, William G et al. (2008). "Data and theory point to mainly additive genetic variance for complex traits". In: *PLoS Genetics* 4.2, e1000008.

Ho, Chiu Man and Stephen DH Hsu (2015). "Determination of nonlinear genetic architecture using compressed sensing". In: *GigaScience* 4.1, s13742–015.

Ho, Daniel Sik Wai et al. (2019). "Machine learning SNP based prediction for precision medicine". In: *Frontiers in Genetics* 10, p. 267.

Holzapfel, Christina et al. (2010). "Genes and lifestyle factors in obesity: results from 12462 subjects from MONICA/KORA". In: *International Journal of Obesity* 34.10, pp. 1538–1545.

Huang, Hao et al. (2016). "The TOMM40 gene rs2075650 polymorphism contributes to Alzheimer's disease in Caucasian, and Asian populations". In: *Neuroscience Letters* 628, pp. 142–146.

Huang, Sijia et al. (2017). "More is better: recent progress in multi-omics data integration methods". In: *Frontiers in Genetics* 8.

Imoto, Seiya et al. (2004). "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks". In: *Journal of Bioinformatics and Computational Biology* 2.01, pp. 77–98.

Kang, Mingon et al. (2022). "A roadmap for multi-omics data integration using deep learning". In: *Briefings in Bioinformatics* 23.1.

Khuri, Andre I and Hardeo Sahai (1985). "Variance components analysis: a selective literature survey". In: *International Statistical Review/Revue Internationale de Statistique*, pp. 279–300.

Ki, Chang-Seok et al. (2002). "Genetic association of an apolipoprotein C1 (APOC1) gene polymorphism with late-onset Alzheimer's disease". In: *Neuroscience Letters* 319.2, pp. 75–78.

Kim, S et al. (2011). "Genome-wide association study of CSF biomarkers Aβ1-42, t-tau, and p-tau181p in the ADNI cohort". In: *Neurology* 76.1, pp. 69–79.

Kim, Seung-Jean et al. (2007). "An interior-point method for large-scale L_1-regularized least squares". In: *IEEE Journal of Selected Topics in Signal Processing* 1.4, pp. 606–617.

Kirchner, Henriette et al. (2013). "Epigenetic flexibility in metabolic regulation: disease cause and prevention?" In: *Trends in Cell Biology* 23.5, pp. 203–209.

Li, Jun et al. (2020). "Multi-kernel linear mixed model with adaptive Lasso for prediction analysis on high-dimensional multi-omics data". In: *Bioinformatics* 36.6, pp. 1785–1794.

Lock, Eric F et al. (2013). "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types". In: *The Annals of Applied Statistics* 7.1, p. 523.

Lucatelli, Juliana Fagion et al. (2011). "Genetic influences on Alzheimer's disease: evidence of interactions between the genes APOE, APOC1 and ACE in a sample population from the South of Brazil". In: *Neurochemical Research* 36.8, pp. 1533–1539.

Ma, Shuangge et al. (2007). "Supervised group Lasso with applications to microarray data analysis". In: *BMC Bioinformatics* 8.1, p. 60.

Ma, Xiao-Ying et al. (2013). "Association of TOMM40 polymorphisms with late-onset Alzheimer's disease in a Northern Han Chinese population". In: *Neuromolecular Medicine* 15.2, pp. 279–287.

Mathew, Isack et al. (2018). "Variance components and heritability of traits related to root: shoot biomass allocation and drought tolerance in wheat". In: *Euphytica* 214.12, pp. 1–12.

Meng, Chen et al. (2014). "A multivariate approach to the integration of multi-omics datasets". In: *BMC Bioinformatics* 15.1, p. 162.

Meng, Chen et al. (2016). "Dimension reduction techniques for the integrative analysis of multi-omics data". In: *Briefings in Bioinformatics* 17.4, pp. 628–641.

Moore, Jason H and Scott M Williams (2009). "Epistasis and its implications for personal genetics". In: *The American Journal of Human Genetics* 85.3, pp. 309–320.

Morris, Jeffrey S and Veerabhadran Baladandayuthapani (2017). "Statistical contributions to bioinformatics: design, modelling, structure learning and integration". In: *Statistical Modelling* 17.4-5, pp. 245–289.

Mueller, Susanne G et al. (2005). "The Alzheimer's Disease Neuroimaging Initiative". In: *Neuroimaging Clinics* 15.4, pp. 869–877.

Ndiaye, Eugene et al. (2015). "Gap safe screening rules for sparse multi-task and multi-class models". In: *Advances in Neural Information Processing Systems* 28, pp. 811–819.

Negahban, Sahand N et al. (2012). "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers". In: *Statistical Science* 27.4, pp. 538–557.

Ossenkoppele, Rik et al. (2013). "Differential effect of APOE genotype on amyloid load and glucose metabolism in AD dementia". In: *Neurology* 80.4, pp. 359–365.

Pan, Jianxin and Chao Huang (2014). "Random effects selection in generalized linear mixed models via shrinkage penalty function". In: *Statistics and Computing* 24.5, pp. 725–738.

Pazokitoroudi, Ali et al. (2019). "Scalable multi-component linear mixed models with application to SNP heritability estimation". In: *bioRxiv*, p. 522003.

Poirier, Judes et al. (1993). "Apolipoprotein E polymorphism and Alzheimer's disease". In: *The Lancet* 342.8873, pp. 697–699.

Potkin, Steven G et al. (2009). "Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease". In: *PLoS One* 4.8, e6501.

Prendecki, Michal et al. (2018). "Biothiols and oxidative stress markers and polymorphisms of TOMM40 and APOC1 genes in Alzheimer's disease patients". In: *Oncotarget* 9.81, p. 35207.

Pu, Wenji and Xu-Feng Niu (2006). "Selecting mixed-effects models based on a generalized information criterion". In: *Journal of Multivariate Analysis* 97.3, pp. 733–758.

Puglielli, Luigi et al. (2003). "Alzheimer's disease: the cholesterol connection". In: *Nature Neuroscience* 6.4, pp. 345–351.

Rao, C Radhakrishna (1970). "Estimation of heteroscedastic variances in linear models". In: *Journal of the American Statistical Association* 65.329, pp. 161–172.

— (1971a). "Estimation of variance and covariance components—MINQUE theory". In: *Journal of Multivariate Analysis* 1.3, pp. 257–275.

— (1971b). "Minimum variance quadratic unbiased estimation of variance components". In: *Journal of Multivariate Analysis* 1.4, pp. 445–456.

— (1972). "Estimation of variance and covariance components in linear models". In: *Journal of the American Statistical Association* 67.337, pp. 112–115.

Reimherr, Matthew and Dan Nicolae (2016). "Estimating variance components in functional linear models with applications to genetic heritability". In: *Journal of the American Statistical Association* 111.513, pp. 407–422.

Reyes-Gibby, Cielito C et al. (2009). "Role of inflammation gene polymorphisms on pain severity in lung cancer patients". In: *Cancer Epidemiology and Prevention Biomarkers* 18.10, pp. 2636–2642.

Ritchie, Marylyn D et al. (2015). "Methods of integrating data to uncover genotype–phenotype interactions". In: *Nature Reviews Genetics* 16.2, p. 85.

Rohart, Florian et al. (2014). "Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm". In: *Computational Statistics & Data Analysis* 80, pp. 209–222.

Roses, Allen D (2010). "An inherited variable poly-T repeat genotype in TOMM40 in Alzheimer disease". In: *Archives of Neurology* 67.5, pp. 536–541.

Saunders, Ann M et al. (1993). "Association of apolipoprotein E allele $\epsilon 4$ with late-onset familial and sporadic Alzheimer's disease". In: *Neurology* 43.8, pp. 1467–1467.

Saykin, Andrew J et al. (2010). "Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans". In: *Alzheimer's and Dementia* 6.3, pp. 265–273.

Saykin, Andrew J et al. (2015). "Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans". In: *Alzheimer's and Dementia* 11.7, pp. 792–814.

Schelldorfer, Jürg et al. (2011). "Estimation for high-dimensional linear mixed-effects models using L1-penalization". In: *Scandinavian Journal of Statistics* 38.2, pp. 197–214.

Schuchardt, Jan Philipp et al. (2016). "Genetic variants of the FADS gene cluster are associated with erythrocyte membrane LC PUFA levels in patients with mild cognitive impairment". In: *The Journal of Nutrition, Health & Aging* 20.6, pp. 611–620.

Seal, Dibyendu Bikash et al. (2020). "Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration". In: *Genomics* 112.4, pp. 2833–2841.

Seshadri, Sudha et al. (2010). "Genome-wide analysis of genetic loci associated with Alzheimer disease". In: *The Journal of the American Medical Association* 303.18, pp. 1832–1840.

Shen, Li et al. (2010). "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort". In: *Neuroimage* 53.3, pp. 1051–1063.

Shen, Ronglai et al. (2009). "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis". In: *Bioinformatics* 25.22, pp. 2906–2912.

Shi, J et al. (2004). "Association between apolipoprotein CI HpaI polymorphism and sporadic Alzheimer's disease in Chinese". In: *Acta Neurologica Scandinavica* 109.2, pp. 140–145.

Speed, Doug and David J Balding (2014). "MultiBLUP: improved SNP-based prediction for complex traits". In: *Genome Research* 24.9, pp. 1550–1557.

Strittmatter, Warren J et al. (1993). "Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease". In: *Proceedings of the National Academy of Sciences* 90.5, pp. 1977–1981.

Subramanian, Indhupriya et al. (2020). "Multi-omics data integration, interpretation, and its application". In: *Bioinformatics and Biology Insights* 14, p. 1177932219899051.

Sun, Hokeun and Shuang Wang (2012). "Penalized logistic regression for high-dimensional DNA methylation data with case-control studies". In: *Bioinformatics* 28.10, pp. 1368–1375.

Swallow, William H and John F Monahan (1984). "Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components". In: *Technometrics* 26.1, pp. 47–57.

Tang, Ming-Xin et al. (1998). "The APOE ϵ4 allele and the risk of Alzheimer disease among African Americans, Whites, and Hispanics". In: *The Journal of the American Medical Association* 279.10, pp. 751–755.

The 1000 Genomes Project Consortium (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking).

Tibshirani, Robert et al. (2012). "Strong rules for discarding predictors in Lasso-type problems". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.2, pp. 245–266.

Tycko, Benjamin et al. (2004). "APOE and APOC1 promoter polymorphisms and the risk of Alzheimer disease in African American and Caribbean Hispanic individuals". In: *Archives of Neurology* 61.9, pp. 1434–1439.

VanRaden, Paul M (2008). "Efficient methods to compute genomic predictions". In: *Journal of Dairy Science* 91.11, pp. 4414–4423.

Vaske, Charles J et al. (2010). "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM". In: *Bioinformatics* 26.12, pp. i237–i245.

Wang, Bo et al. (2014a). "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature Methods* 11.3, p. 333.

Wang, Jie et al. (2015). "Lasso screening rules via dual polytope projection". In: *Journal of Machine Learning Research* 16.1, pp. 1063–1101.

Wang, Quan et al. (2019). "A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data". In: *Nature Neuroscience* 22.5, p. 691.

Wang, Wenting et al. (2012). "iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data". In: *Bioinformatics* 29.2, pp. 149–159.

Wang, Xiaoyue et al. (2014b). "Widespread genetic epistasis among cancer genes". In: *Nature Communications* 5.1, pp. 1–10.

Wang, Xiaqiong and Yalu Wen (2021). "A penalized linear mixed model with generalized method of moments for complex phenotype prediction". In: *bioRxiv*.

Weissbrod, Omer et al. (2016). "Multikernel linear mixed models for complex phenotype prediction". In: *Genome Research* 26.7, pp. 969–979.

Wen, Yalu and Qing Lu (2020). "Multikernel linear mixed model with adaptive Lasso for complex phenotype prediction". In: *Statistics in Medicine* 39.9, pp. 1311–1327.

Wen, Yalu et al. (2016). "Risk prediction modeling of sequencing data using a forward random field method". In: *Scientific Reports* 6, p. 21120.

Wheeler, Heather E et al. (2014). "Poly-omic prediction of complex traits: OmicKriging". In: *Genetic Epidemiology* 38.5, pp. 402–415.

Wu, Lan et al. (2014). "Nonnegative-Lasso and application in index tracking". In: *Computational Statistics & Data Analysis* 70, pp. 116–126.

Wu, Tong Tong et al. (2009). "Genome-wide association analysis by Lasso penalized logistic regression". In: *Bioinformatics* 25.6, pp. 714–721.

Xiang, Zhen James and Peter J Ramadge (2012). "Fast Lasso screening tests based on correlations". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2137–2140.

Xiang, Zhen James et al. (2016). "Screening tests for Lasso problems". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.5, pp. 1008–1027.

Xu, Jing et al. (2019a). "A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data". In: *BMC Bioinformatics* 20.1, pp. 1–11.

Xu, Wanxue et al. (2019b). "Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers". In: *Signal Transduction and Targeted Therapy* 4.1, pp. 1–11.

Yang, Jian et al. (2010). "Common SNPs explain a large proportion of the heritability for human height". In: *Nature Genetics* 42.7, p. 565.

Yang, Sheng and Xiang Zhou (2020). "Accurate and scalable construction of polygenic scores in large biobank data sets". In: *The American Journal of Human Genetics* 106.5, pp. 679–693.

Yuan, Yinyin et al. (2011). "Patient-specific data fusion defines prognostic cancer subtypes". In: *PLoS Computational Biology* 7.10, e1002227.

Zannis, Vassilis I et al. (1993). "Genetic mutations affecting human lipoproteins, their receptors, and their enzymes". In: *Advances in Human Genetics 21*, pp. 145–319.

Zeng, Irene Sui Lan and Thomas Lumley (2018). "Review of statistical learning methods in integrated omics studies (an integrated information science)". In: *Bioinformatics and Biology Insights* 12, p. 1177932218759292.

Zeng, Ping and Xiang Zhou (2017). "Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models". In: *Nature Communications* 8.1, pp. 1–11.

Zeng, Yaohui et al. (2021). "Hybrid safe–strong rules for efficient optimization in Lasso-type problems". In: *Computational Statistics & Data Analysis* 153, p. 107063.

Zhang, Shihua et al. (2012). "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data". In: *Nucleic Acids Research* 40.19, pp. 9379–9391.

Zhao, Bo et al. (2012). "A Bayesian approach to discovering truth from conflicting sources for data integration". In: *arXiv preprint arXiv:1203.0058*.

Zhao, Sihai Dave and Yi Li (2012). "Principled sure independence screening for Cox models with ultra-high-dimensional covariates". In: *Journal of Multivariate Analysis* 105.1, pp. 397–411.

Zhao, Yihua et al. (2006). "General design Bayesian generalized linear mixed models". In: *Statistical Science*, pp. 35–51.

Zhou, Guangyan et al. (2020). "Network-based approaches for multi-omics integration". In: *Computational Methods and Data Analysis for Metabolomics*. Springer, pp. 469–487.

Zhou, Qin et al. (2014a). "APOE and APOC1 gene polymorphisms are associated with cognitive impairment progression in Chinese patients with late-onset Alzheimer's disease". In: *Neural Regeneration Research* 9.6, p. 653.

Zhou, Qin et al. (2014b). "Association between APOC1 polymorphism and Alzheimer's disease: a case-control study and meta-analysis". In: *PLoS One* 9.1, e87017.

Zhou, Xiang (2017). "A unified framework for variance component estimation with summary statistics in genome-wide association studies". In: *The Annals of Applied Statistics* 11.4, p. 2027.

Zhou, Xiang et al. (2013). "Polygenic modeling with Bayesian sparse linear mixed models". In: *PLoS Genetics* 9.2, e1003264.

Zhu, Jun (1995). "Analysis of conditional genetic effects and variance components in developmental genetics". In: *Genetics* 141.4, pp. 1633–1639.

Zhu, Jun and Bruce S Weir (1996). "Mixed model approaches for diallel analysis based on a bio-model". In: *Genetics Research* 68.3, pp. 233–240.

Zou, Hui and Runze Li (2008). "One-step sparse estimates in nonconcave penalized likelihood models". In: *The Annals of Statistics* 36.4, p. 1509.