12-1-2020

Bayesian Variable Selection Methods for Genome-Wide Association Studies with Categorical Phenotypes

Benazir Rowe

BAYESIAN VARIABLE SELECTION METHODS FOR GENOME-WIDE ASSOCIATION STUDIES WITH CATEGORICAL PHENOTYPES

By

Benazir Rowe

Bachelor of Science - Mathematical Economics
Kyrgyz Russian Slavic University, Kyrgyzstan
2012

Master of Science - International Economics and Finance
Otto von Guericke University Magdeburg, Germany
2014

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy - Mathematical Sciences

Department of Mathematical Sciences
College of Sciences
The Graduate College

University of Nevada, Las Vegas
December 2020

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

November 9, 2020

This dissertation prepared by

Benazir Rowe

entitled

Bayesian Variable Selection Methods for Genome-Wide Association Studies with Categorical Phenotypes

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Mathematical Sciences
Department of Mathematical Sciences

Amei Amei, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Dean

Malwane Ananda, Ph.D.
Examination Committee Member

Kaushik Ghosh, Ph.D.
Examination Committee Member

Guogen Shan, Ph.D.
Graduate College Faculty Representative

ABSTRACT

Bayesian Variable Selection Methods for Genome-wide Association Studies with Categorical Phenotypes

by

Benazir Rowe

Dr. Amei Amei, Examination Committee Chair
Associate Professor of Mathematics
University of Nevada, Las Vegas, USA

Genome-wide association studies (GWAS) attempt to find the associations between genetic markers and studied traits (phenotypes). The problem of GWAS is complex and various methods have been developed to approach it. One of such methods is Bayesian variable selection (BVS). We describe the BVS methods in detail and demonstrate the ability of BVS method Posterior Inference via Model Averaging and Subset Selection (piMASS) to improve the power of detecting phenotype-associated genetic loci, potentially leading to new discoveries from existing data without increasing the sample size.

We present several ways to improve and extend the applicability of piMASS for GWAS. The first method incorporates non-genetic covariates in the BVS process for GWAS with continuous phenotype, therefore making it possible to account for population stratification. Next, we extend the method mentioned above to work with binary phenotype. Finally,

we extend the piMASS method to work with ordinal phenotype. The presented methods allow the existing BVS methods reach wider applicability and higher quality of detected associations. We conduct simulation studies and compare the results to the original method piMASS to show their efficacy. We also apply two of the methods to the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI1) dataset containing data on Alzheimer's patients with categorical phenotype and demonstrate the method's ease of use and applicability. Finally, we discuss the potential of the methods in GWAS and possible directions for further research.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr Amei Amei for her expertise, patience and determination. Working together strengthened not only my knowledge, but also my character and resilience. I also want to thank Dr Xiangning Chen for introducing me to psychiatric genetics and genetic data as well as proposing the topic for this dissertation. Big thanks to Chong Cheng and Ron Young for their help with supercomputing. I want to thank my professors Dr Kaushik Ghosh and Dr Hokwon Cho who contributed to my knowledge and development as a statistician.

I want to thank my husband Nathan for his unequivocal support and encouragement of my PhD endeavors. I want to thank my son Daniel who keeps me motivated and brings joy and positivity to my every day. I want to thank my mom Aigul for giving me love for math from an early age and my dad Erkin for giving me the idea to do a PhD in mathematics and my sister Asel, who encouraged me to achieve the highest heights I can. Also thanks to all my family for their support. Big thanks to my math department friends Jessica, Katlyn, EJ, Shar, Marvin, Daniel, Edward, Julie and others, your friendship certainly made my stay at UNLV.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

In genetics, genome-wide association study (GWAS) is a method that aims at detecting genomic loci associated with complex traits in the population, such as heart disease, diabetes, auto-immune diseases and psychiatric disorders [Visscher et al., 2012]. Typical GWAS measures hundreds of thousands or millions of genetic variants in thousands or tens of thousands of individuals. In situations where most of the regressors have no relationship to the response, including all regressors in a regression model can lead to poor performance [Hoff, 2009]. Therefore, only variables with evidence for association should be included in the model, which gives not only a more concise and practical model, but also enhances model prediction. Hence, the problem of GWAS is a variable selection problem.

Variable selection is one of the key aspects of the regression modelling, however, common approaches such as exhaustive search and stepwise procedures are prohibitive in GWAS context due to large number of potential covariates. Bayesian approach offers a practical solution to this issue: any collection of models with different sets of variables can be compared using

their Bayes factors. Bayes factor is a ratio of two likelihoods representing any two statistical models and provides the summary of evidence provided by the data in favor of one scientific theory, represented by a statistical model, as opposed to another [Kass and Raftery, 1995]. Bayesian variable selection (BVS) regression is a way to approach variable selection problem by specifying prior distributions on model parameters and calculating Bayes factors of various models, which contain relevant information to perform the selection. In situations where the number of variables is large, the space of possible models can be explored using Markov chain Monte Carlo (MCMC) methods such as Metropolis-Hastings algorithm.

Another important feature of the GWAS is that effect sizes of genetic variants are often very small and the total proportion of the variance of the response variable explained by covariates is also small. Those features together require modification of the existing BVS methodology to the problem in question. BVS is especially advantageous in GWAS context, since it can handle large number of parameters and provides easily interpretable measures of confidence, specifically posterior probability of a coefficient being nonzero for each variable included in the model. It also allows modelling observable outcomes conditional on a set of parameters, which themselves can be given a probabilistic specification in terms of further parameters, known as hyperparameters [Gelman et al., 2013]. Data characteristics and dimensionality play an important role in the specification of the BVS model.

In this chapter we first describe the existing methods in BVS. Next we focus on the posterior inference via Model Averaging and Subset Selection (piMASS) [Guan and Stephens, 2011], a BVS method with novel hyperparameter specifications tailored to the GWAS setting. In

presenting these methods we highlight their uses and motivate the models and methods proposed in the upcoming chapters. At the end of the chapter, we provide an outline of the dissertation.

1.2 Review of Bayesian variable selection

Below we present a general overview of BVS models. The goal of BVS is to choose relevant covariates from a larger set of potential covariates, which is a subset selection problem. Subset selection is often formulated in the context of multiple linear regression models. Thus, the individual's response y_i is modelled using standard linear regression:

$$y_i = \mu + x_{i1}\beta_1 + x_{i2}\beta_2 + \ ... \ + x_{ip}\beta_p + \epsilon_i, \tag{1.1}$$

where y_i is the response for i-th individual, μ is an intercept, $x_{i1}, ... \ , x_{ip}$ are covariates measured on the ith individual, $\beta_1, ... \ , \beta_p$ are the corresponding regression coefficients, and ϵ_i is an error term. We assume that ϵ_is are independent and identically distributed with $\epsilon_i \sim N(0, \sigma^2)$. The goal of this setup is to identify the relevant covariates among $x_{i1}, ... \ , x_{ip}$ based on the evidence for association with the response y_i.

Therefore, normal linear regression model may be used to describe the relationship between the response $\boldsymbol{y}$ and potential covariates $\boldsymbol{X_1}, \boldsymbol{X_2}, \ ... \ , \boldsymbol{X_p}$, that is:

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma) \sim N_n(\boldsymbol{\mu} + \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}), \tag{1.2}$$

where $\boldsymbol{y}$ is a response vector of size $n \times 1$, $\boldsymbol{X} = [\boldsymbol{X_1}, \ ... \ , \boldsymbol{X_p}]$ is an $n \times p$ matrix, where $\boldsymbol{X_j} = (x_{1j}, \ ... \ , x_{nj})'$ is a column vector containing the observed values of the jth covariate at the n individuals, $\boldsymbol{\mu} = \mu \times \mathbb{1}$, where $\mathbb{1}$ is a $n \times 1$ vector of ones, $\boldsymbol{\beta}$ is a vector of regression coefficients of size $p \times 1$ and σ is positive scalar.

The problem of variable selection arises when one is uncertain about which predictors are relevant to the outcome. Each of those possible 2^p subsets is denoted using vector $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} = (\gamma_1, ... \ , \gamma_p)' \in \{0, 1\}^p, \tag{1.3}$$

where γ_i takes two values, 1 and 0, meaning covariate i is included in the model or not. The size of the subset can be represented as $d_{\boldsymbol{\gamma}} = \boldsymbol{\gamma}'\mathbb{1}$, where $\mathbb{1}$ is a $p \times 1$ vector of ones. Then the possible models describing the relationship between response and the selected subset of covariates are:

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma) \sim N_n(\boldsymbol{\mu} + \boldsymbol{X_\gamma}\boldsymbol{\beta_\gamma}, \ \sigma^2\boldsymbol{I}), \tag{1.4}$$

where $\boldsymbol{X_\gamma}$ is an $n \times d_\gamma$ matrix whose columns correspond to the design matrix $\boldsymbol{X}$ restricted to the columns j for which $\gamma_j = 1$ and $\boldsymbol{\beta_\gamma}$ denotes the corresponding regression coefficients. Methods capable of avoiding the calculation of 2^p possible models are essential in high dimensional variable selection.

The joint prior on model parameters is [George and McCulloch, 1993]:

$$\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma})\pi(\sigma^2|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}), \tag{1.5}$$

where every part is further specified below.

First, we assume that $\boldsymbol{\gamma}$ components are independent, which means we expect that inclusion of any variable in the model has no impact on whether any other variable should be included. Therefore, $\boldsymbol{\gamma}$ is assigned a prior distribution

$$\pi(\boldsymbol{\gamma}) = \prod_{i=1}^{p} \omega_i^{\gamma_i} (1 - \omega_i)^{(1-\gamma_i)}, \tag{1.6}$$

which provides computational simplicity. Here $\pi(\gamma_i = 1) = 1 - \pi(\gamma_i = 0) = \omega_i$ can be interpreted as the prior probability that X_i is included in the model. Setting ω_i to be small will put increased weight on parsimonious models.

Second, variance of the residuals σ^2 is assigned an inverse gamma prior distribution:

$$\pi(\sigma^2 \mid \boldsymbol{\gamma}) = IG(\lambda/2,\ k/2). \tag{1.7}$$

While setting k to be constant leads to reasonable results [George and McCulloch, 1997], it can be beneficial to have k decrease with the number of selected variables $d_{\boldsymbol{\gamma}}$. When prior information about σ^2 is scarce, it is recommended to choose k and λ so that prior assigns substantial probability to the interval between the least squares estimate based on a saturated model and the sample variance of $\boldsymbol{y}$.

Finally, the vector of regression coefficients $\boldsymbol{\beta}$ is assigned a multivariate normal prior [George and McCulloch, 1993]:

$$\pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma}) = N_p(\mathbf{0}, \boldsymbol{W}_{(\sigma,\gamma)}), \tag{1.8}$$

where $\mathbf{0}$ is an p-vector of zeros and $\boldsymbol{W}_{(\sigma,\boldsymbol{\gamma})}$ is the p-dimensional covariance matrix where the ith diagonal element is appropriately set to be large or small based on whether $\gamma_i = 1$ or 0.

Specification of $\boldsymbol{W}_{(\sigma,\boldsymbol{\gamma})}$ determines the key properties of the specified hierarchical prior. The effect of this specification for the variable selection are of particular importance and should reflect the properties of the specific context in which the selection is performed.

For example, consider the following conjugate specification of $\boldsymbol{W}_{(\sigma,\boldsymbol{\gamma})}$:

$$\pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma}) = N_p(\mathbf{0}, \sigma^2 \boldsymbol{D}_{\boldsymbol{\gamma}} \boldsymbol{R}_{\boldsymbol{\gamma}} \boldsymbol{D}_{\boldsymbol{\gamma}}), \tag{1.9}$$

where σ^2 is the variance of the residuals, $\boldsymbol{D}_{\boldsymbol{\gamma}}$ is a diagonal matrix, where each diagonal element of $\boldsymbol{D}_{\boldsymbol{\gamma}}$ contains the corresponding standard deviation of the corresponding component of vector $\boldsymbol{\beta}$, i.e. $(\boldsymbol{D}_{\boldsymbol{\gamma}})_{ii} = sd(\beta_i)$, $i = 1, \dots, n$ and $\boldsymbol{R}_{\boldsymbol{\gamma}}$ is a correlation matrix of $\boldsymbol{\beta}$. One simple choice for the correlation matrix is $\boldsymbol{R}_{\boldsymbol{\gamma}} = \boldsymbol{I}$, making components of $\boldsymbol{\beta}$ uncorrelated. Other possible choice is $\boldsymbol{R}_{\boldsymbol{\gamma}} = (\boldsymbol{X}'_{\boldsymbol{\gamma}} \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$, replicating the correlation structure of least squares estimate.

The ith diagonal element of $\boldsymbol{D}^2_{\boldsymbol{\gamma}}$ representing the variance of $\boldsymbol{\beta}$ coefficients is denoted by

$$(\boldsymbol{D}^2_{\boldsymbol{\gamma}})_{ii} = \begin{cases} \nu_0, & \text{if } \gamma_i = 0, \\ \nu_1, & \text{if } \gamma_i = 1. \end{cases} \tag{1.10}$$

Under (1.10) each component of $\boldsymbol{\beta}$ is assumed to be a weighted sum of two normal distributions, sometimes called a scaled mixture of normals:

$$\pi(\beta_i \mid \sigma, \boldsymbol{\gamma}) = (1 - \gamma_i) N(0, \sigma^2 \nu_0) + \gamma_i N(0, \sigma^2 \nu_1). \tag{1.11}$$

It is therefore convenient to present the conditional prior for $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \sigma^2, \boldsymbol{\gamma} \sim N_{d_{\boldsymbol{\gamma}}}(0, \sigma^2 \nu_1 \boldsymbol{I}_{\boldsymbol{d_\gamma}}), \tag{1.12}$$

$$\boldsymbol{\beta}_{-\boldsymbol{\gamma}} \mid \boldsymbol{\gamma} \sim N_{n-d_{\boldsymbol{\gamma}}}(0, \sigma^2 \nu_0 \boldsymbol{I}_{\boldsymbol{n-d_\gamma}}), \tag{1.13}$$

where $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denotes a vector of regression coefficients corresponding to elements of $\boldsymbol{X}$ for which $\gamma_j = 1$ and $\boldsymbol{\beta}_{\boldsymbol{-\gamma}}$ denotes the vector of $\boldsymbol{\beta}$ coefficients for which $\gamma_j = 0$.

Since conditional distribution of $\boldsymbol{\beta}$ and σ given $\boldsymbol{\gamma}$ is conjugate for (1.4)
[George and McCulloch, 1993], the resulting hierarchical mixture prior can be called conjugate prior, so $\boldsymbol{\beta}$ and σ can be eliminated by routine integration from the full posterior $\pi(\boldsymbol{\beta}, \sigma, \boldsymbol{\gamma}|\boldsymbol{y})$. This gives convenient computational methods for posterior evaluation.

When $\nu_0 \equiv 0$ combining the likelihood from (1.4) with the priors (1.6)-(1.9) and (1.12)-(1.13) gives the joint posterior:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma|\boldsymbol{y}) \propto \sigma^{(n+d_{\boldsymbol{\gamma}}+\lambda+1)/2} \Big| \boldsymbol{D}_{\boldsymbol{\gamma}} \boldsymbol{R}_{\boldsymbol{\gamma}} \boldsymbol{D}_{\boldsymbol{\gamma}} \Big|^{-1/2} \exp\Big(-\frac{1}{2\sigma^2} \Big| \widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}_{\boldsymbol{\gamma}}} \boldsymbol{\beta}_{\boldsymbol{\gamma}} \Big|^2 \Big) \exp\Big(-\frac{k}{2\sigma^2} \Big) \pi(\boldsymbol{\gamma}), \tag{1.14}$$

where

$$\widetilde{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix} \text{ and } \widetilde{\boldsymbol{X}_{\boldsymbol{\gamma}}} = \begin{bmatrix} \boldsymbol{X}_{\boldsymbol{\gamma}} \\ (\boldsymbol{D}_{\boldsymbol{\gamma}} \boldsymbol{R}_{\boldsymbol{\gamma}} \boldsymbol{D}_{\boldsymbol{\gamma}})^{-1/2} \end{bmatrix}.$$

With the setup provided above, the marginal posterior distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ carries the information to conduct the variable selection. Based on response data $\boldsymbol{y}$, the posterior distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ identifies the models better supported by the prior distributions and the data. The ability of the above setup to correctly select the associated variables in a reasonable amount of time depends on a few factors. One, parameters of the prior must be set in a way that posterior distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ will assign higher probability to the subsets of covariates that are of interest for a given problem [George and McCulloch, 1997]. Next, since $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ is not always available in closed form, it is necessary to be able to compute $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ at least to the

extent it is possible to differentiate between $\gamma_i = 0$ and $\gamma_i = 1$ for each γ_i. Hyperparameters allow to set important characteristics of the model, such as model sparsity and proportion of the variance in the response variable explained by the covariates. The next section introduces a BVS model with priors on hyperparameters tailored to the GWAS.

1.3 piMASS and its novel prior on variance component

The ability to specify priors is an important feature of Bayesian approach, since it helps to focus on a particular subset of parameters and saves important computation time exploring regions of parameter space that are less plausible. Such priors have been proposed for continuous and binary response variables in a linear model in the method piMASS [Guan and Stephens, 2011]. The novel prior on the variance of regression coefficients is what sets piMASS apart from the previous work done in the area of BVS. We chose piMASS as our base BVS method because of its GWAS -tailored prior structure and feasible computational properties.

In particular, piMASS provides a BVS setup in a linear setting:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \tag{1.15}$$

where y_i is a response of the i-th subject, μ is an intercept, x_{ij} is a j-th covariate of i-th subject, β_js are the corresponding regression coefficients and ϵ_i is an error term. ϵ_i are

independent and identically distributed $\epsilon_i \sim N(0, 1/\tau)$, where τ denotes the inverse of the variance, often referred to as precision.

Thus the distribution of the response is:

$$\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \tau, \boldsymbol{\beta}, \boldsymbol{X} \sim N_n(\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I_n}), \tag{1.16}$$

relating the response variable $\boldsymbol{y}$ to covariates $\boldsymbol{X}$. In the context of GWAS the response $\boldsymbol{y}$ is the phenotype (observable and measurable characteristic) under investigation for each of the n study subjects. Next, $\boldsymbol{X}$ is an $n \times p$ matrix of covariates , where x_{ij} is the genotype of subject i at the genetic variant or single nucleotide polymorphism (SNP) j and $x_{ij} = 0, 1$ or 2 depending on whether the subject has 0, 1 or 2 copies of the minor allele. In model (1.16) $\boldsymbol{\mu}$ is an n-vector with each component equal to the same scalar μ, $\boldsymbol{\beta}$ is a p-vector of regression coefficients and τ is the inverse variance of the residual errors.

The binary indicators denoted as $\boldsymbol{\gamma}$ and defined in (1.3) show which elements of $\boldsymbol{\beta}$ are nonzero. Then the model becomes:

$$\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \tau, \boldsymbol{\beta}, \boldsymbol{X} \sim N_n(\boldsymbol{\mu} + \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau^{-1}\boldsymbol{I_n}), \tag{1.17}$$

where $\boldsymbol{X}_{\boldsymbol{\gamma}}$ denotes the design matrix $\boldsymbol{X}$ restricted to those columns j for which $\gamma_j = 1$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denotes a corresponding vector of regression coefficients. The setup above follows the standard setup for the hierarchical mixture model for variable selection and allows to account

for sparsity that is inherent to GWAS. Next, the priors on model parameters:

$$\tau \sim Gamma(\lambda/2, k/2) \tag{1.18}$$

$$\mu|\tau \sim N(0, \sigma_\mu^2/\tau) \tag{1.19}$$

$$\gamma_j \sim Bernoulli(\pi) \tag{1.20}$$

$$\boldsymbol{\beta_\gamma}|\tau, \boldsymbol{\gamma} \sim N_{d_\gamma}(0\ , (\sigma_a^2/\tau)\ \boldsymbol{I}_{d_\gamma}) \tag{1.21}$$

$$\boldsymbol{\beta_{-\gamma}}|\boldsymbol{\gamma} \sim \delta_0, \tag{1.22}$$

where δ_0 is a point mass at 0 and $\lambda, k, \sigma_\mu, \pi$ and σ_a are hyperparameters. Hyperparameters π, which reflects the model sparsity and σ_a, which reflects the typical size of a nonzero regression coefficients are playing important roles in tailoring model to GWAS. Therefore, instead of being assigned fixed values, in piMASS they get their own prior distributions. Other hyperparameters are less critical and in practice $\lambda, k \to 0$ and both response vector $\boldsymbol{y}$ and covariates $\boldsymbol{X_1}, \ldots, \boldsymbol{X_p}$ are centered to have mean 0, setting $\mu = 0$.

Equations (1.21) and (1.22) give

$$\pi(\beta_i \mid \gamma_i) = (1 - \gamma_i)\delta_0 + \gamma_i N(0, (\sigma_a^2/\tau). \tag{1.23}$$

Since (1.23) contains the mixture of a normal distribution and a point mass at 0, any $\beta_i \neq 0$ will be included in the model. This is different compared to (1.11), where the mixture prior for $\boldsymbol{\beta}$ contains two normal distributions with different variances, in which case covariates with larger β_i are selected. In contrast, (1.23) will lead to selection based on how different β_i is from 0 rather than the absolute size of the β_i. This property is desirable for GWAS since effect sizes of SNPs are often small, yet still valuable.

Additionally, the specification of $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ requires the choice of a prior correlation matrix $\boldsymbol{R}_{\boldsymbol{\gamma}}$. Here $\boldsymbol{R}_{\boldsymbol{\gamma}} = \boldsymbol{I}$, setting the components of $\boldsymbol{\beta}$ to be independent, which allows for rapid update of the posterior if $\boldsymbol{\gamma}$ is changed one component at a time [George and McCulloch, 1993].

The prior on π is specified in the following way:

$$log(\pi) \sim U(a, b) \tag{1.24}$$

where $a = log(1/p)$ and $b = log(M/p)$, so the lower and upper limits of π correspond to the models which are expected to include between 1 and M variables.

The novelty of piMASS is in nonstandard prior on variance component σ_a. The motivation behind the formulation is that in piMASS, priors on variance component translate assumptions about the the expected proportion of the variance in $\boldsymbol{y}$ explained by $\boldsymbol{X}_{\boldsymbol{\gamma}}$ (PVE). The main idea here is to choose a prior on $\boldsymbol{\beta}$ given τ so that induced prior on PVE is approximately uniform on $(0, 1)$.

In more detail, let $V(\boldsymbol{\beta}, \tau)$ denote the empirical variance of $\boldsymbol{X\beta}$ relative to the residual variance τ^{-1}:

$$V(\boldsymbol{\beta}, \tau) = \frac{Var(\boldsymbol{X\beta})}{\tau^{-1}} = \frac{\frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{X\beta})_i - \mu]^2}{\tau^{-1}} = \frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{X\beta})_i]^2\tau. \tag{1.25}$$

Equation (1.25) assumes covariates $\boldsymbol{X}$ have been centered, setting $\mu = 0$. Next, let $PVE(\boldsymbol{\beta}, \tau)$ be the total proportion of variance in $\boldsymbol{y}$ explained by $\boldsymbol{X}$ if the true values of regression co-

efficients are $\boldsymbol{\beta}$:

$$\begin{aligned} PVE(\boldsymbol{\beta},\tau) &:= \frac{Var(\boldsymbol{X}\boldsymbol{\beta})}{Var(\boldsymbol{y})} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{X}\boldsymbol{\beta})_i\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{X}\boldsymbol{\beta})_i\right]^2 + \ \tau^{-1}} \\ &= \frac{V(\boldsymbol{\beta},\tau)}{V(\boldsymbol{\beta},\tau)+1}. \end{aligned} \tag{1.26}$$

Define $v(\boldsymbol{\gamma},\sigma_a) = E(V(\boldsymbol{\beta},\tau))$. Then σ_a is connected to $V(\boldsymbol{\beta},\tau)$ through its expectation $E(V(\boldsymbol{\beta},\tau))$ in the following way:

$$\begin{aligned} v(\boldsymbol{\gamma},\sigma_a) &= E(V(\boldsymbol{\beta},\tau)|\boldsymbol{\gamma},\sigma_a,\tau) \\ &= E\Big(\frac{1}{n}\sum_{i=1}^{n}[(X\boldsymbol{\beta})_i]^2\tau|\boldsymbol{\gamma},\sigma_a,\tau\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2E(\beta_i^2)\tau \\ &= \frac{1}{n}\sum_{j:\gamma_j=1}x_j^2(\sigma_a^2/\tau)\cdot\tau + \frac{1}{n}\sum_{k:\gamma_k=0}x_k^2\cdot 0\cdot\tau \\ &= \sigma_a^2\sum_{j:\gamma_j=1}s_j, \end{aligned} \tag{1.27}$$

where $s_j = \frac{1}{n}\sum_{i=1}^{n}x_{ij}^2$.

In piMASS, h is defined as:

$$h(\boldsymbol{\gamma},\sigma_a) = \frac{v(\boldsymbol{\gamma},\sigma_a)}{v(\boldsymbol{\gamma},\sigma_a)+1}, \tag{1.28}$$

where h denotes the approximation to the expectation of the $PVE(\boldsymbol{\beta},\boldsymbol{\tau})$ for given values of $\boldsymbol{\gamma}$ and σ_a:

$$E(PVE(\boldsymbol{\beta},\tau)) = E\Big(\frac{V(\boldsymbol{\beta},\tau)}{V(\boldsymbol{\beta},\tau)+1}\Big) \approx \frac{E(V(\boldsymbol{\beta},\tau))}{E(V(\boldsymbol{\beta},\tau)+1)} = \frac{v(\gamma,\sigma_a)}{v(\gamma,\sigma_a)+1}. \tag{1.29}$$

Here h is not exactly the expectation of the $PVE(\boldsymbol{\beta},\boldsymbol{\tau})$, but instead a ratio of expectations of its numerator and denominator.

Combining (1.27) and (1.28) and specifying a uniform prior on h, independent of $\boldsymbol{\gamma}$, an induced prior on σ_a given γ is obtained using the relationship

$$\sigma_a^2 = \frac{h}{1-h}\frac{1}{\sum_{j:\gamma_j=1} s_j}, \tag{1.30}$$

which is true since

$$\begin{aligned}
\frac{h(\boldsymbol{\gamma},\sigma_a)}{1-h(\boldsymbol{\gamma},\sigma_a)} &= \frac{\frac{v(\boldsymbol{\gamma},\sigma_a)}{v(\boldsymbol{\gamma},\sigma_a)+1}}{1-\frac{v(\boldsymbol{\gamma},\sigma_a)}{v(\boldsymbol{\gamma},\sigma_a)+1}} \\
&= \frac{\frac{v(\boldsymbol{\gamma},\sigma_a)}{v(\boldsymbol{\gamma},\sigma_a)+1}}{\frac{v(\boldsymbol{\gamma},\sigma_a)+1}{v(\boldsymbol{\gamma},\sigma_a)+1}-\frac{v(\boldsymbol{\gamma},\sigma_a)}{v(\boldsymbol{\gamma},\sigma_a)+1}} \\
&= v(\boldsymbol{\gamma},\sigma_a) \\
&= \sigma_a^2 \sum_{j:\gamma_j=1} s_j.
\end{aligned} \tag{1.31}$$

Here the priors are specified and the inference is performed in terms of $(h, \boldsymbol{\gamma})$ rather than $(\sigma_a, \boldsymbol{\gamma})$. Note that induced prior on σ_a^2 is noninformative.

piMASS follows the linear model described in the previous section, except it offers a novel treatment of the prior on hyperparameter σ_a by connecting it to the PVE. The uniform prior on h suggests that the number of relevant variables is no longer necessarily positively correlated to the proportion of variance explained by them. Such prior structure is proper for the situation where there are many relevant variables, with each having tiny effect and overall proportion of variability explained by relevant covariates is small. This property of piMASS matches the genetic architecture of many common diseases, where many loci, each with a very small effect, collectively contribute to the disease.

To successfully perform variable selection in GWAS context, it is potentially beneficial to include variables accounting for population structure [Kärkkäinen and Sillanpää, 2012], which

do not necessarily follow the same distribution as the one specified above for $\boldsymbol{X}$. Including population structure covariates is especially important since they often contribute to spurious associations between genetic variants and the phenotype, which can lead to the selection of irrelevant covariates. While for piMASS, it is possible to account for population structure covariates when phenotype is continuous, when phenotype is categorical there is no obvious way to do it. Therefore, it is potentially beneficial to develop a BVS method that both includes population structure variables and works for categorical data. In the next chapters we focus on extending piMASS to include population structure variables first for continuous, then for binary and ordinal categorical phenotypes.

1.3.1 MCMC scheme for piMASS

The likelihood function of the model in (1.17) is:

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau) \propto \tau^{n/2} \exp\big(-\frac{\tau}{2}(\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^T(\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})\big). \tag{1.32}$$

MCMC is used to sample from the joint posterior distribution of $(h, \pi, \boldsymbol{\gamma})$, which is given by

$$
\begin{aligned}
p(h,\pi,\boldsymbol{\gamma}|\boldsymbol{y}) &= \frac{p(h,\pi,\boldsymbol{\gamma},\boldsymbol{y})}{p(\boldsymbol{y})} \\
&= \frac{p(\boldsymbol{y}|h,\pi,\boldsymbol{\gamma})p(h,\pi,\boldsymbol{\gamma})}{p(\boldsymbol{y})} \\
&\propto p(\boldsymbol{y}|h,\pi,\boldsymbol{\gamma})p(h,\pi,\boldsymbol{\gamma}) \\
&= p(\boldsymbol{y}|h,\boldsymbol{\gamma})p(h,\pi,\boldsymbol{\gamma}) \\
&= p(\boldsymbol{y}|h,\boldsymbol{\gamma})p(h)p(\pi,\boldsymbol{\gamma}) \\
&= p(\boldsymbol{y}|h,\boldsymbol{\gamma})p(h)p(\boldsymbol{\gamma}|\pi)p(\pi).
\end{aligned}
\tag{1.33}
$$

Here $p(\boldsymbol{y}|h,\pi,\boldsymbol{\gamma}) = p(\boldsymbol{y}|h,\boldsymbol{\gamma})$ since hyperparameter π affects $\boldsymbol{y}$ only through $\boldsymbol{\gamma}$, and $p(h,|\pi,\boldsymbol{\gamma}) = p(h)$ since $\pi,\boldsymbol{\gamma}$ do not affect h. Parameters $\boldsymbol{\beta},\tau$ can be integrated out to compute the marginal likelihood $p(\boldsymbol{y}|h,\boldsymbol{\gamma})$.

Since piMASS follows conjugate hierarchical prior setup, it allows $\boldsymbol{\beta}$ and σ to be integrated out analytically to compute $p(\boldsymbol{y}|h,\boldsymbol{\gamma})$ [Servin and Stephens, 2007]. The Bayes factor is:

$$
\frac{p(\boldsymbol{y} \mid h,\boldsymbol{\gamma})}{p(\boldsymbol{y} \mid h,\boldsymbol{\gamma}=\mathbf{0})} = n^{1/2} \mid \boldsymbol{\Omega} \mid^{1/2} \frac{1}{\sigma_a(h,\boldsymbol{\gamma})^{|\boldsymbol{\gamma}|}} \Big(\frac{\boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\Omega}\boldsymbol{X}_{\boldsymbol{\gamma}}^t\boldsymbol{y}}{\boldsymbol{y}^t\boldsymbol{y} - n\bar{\boldsymbol{y}}^2} \Big)^{-n/2}, \tag{1.34}
$$

where $\boldsymbol{\Omega} = \big(\sigma_a(h,\boldsymbol{\gamma})^{-2}\boldsymbol{I}_{|\boldsymbol{\gamma}|} + \boldsymbol{X}_{\boldsymbol{\gamma}}^t\boldsymbol{X}_{\boldsymbol{\gamma}}\big)^{-1}$ and $\mathbf{0}$ denotes a p-vector of all zeros. For each sampled value of $h,\boldsymbol{\gamma}$ from the posterior, $\boldsymbol{\beta}$ and τ are obtained by sampling from their conditional distributions given $\boldsymbol{y},\boldsymbol{\gamma},h$:

$$
\tau|y,h,\boldsymbol{\gamma} \sim \Gamma\big(n/2,\ \ 2/\big(\boldsymbol{y}^t\boldsymbol{y} - \boldsymbol{y}^t\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\Omega}\boldsymbol{X}_{\boldsymbol{\gamma}}^t\boldsymbol{y}\big)\big), \tag{1.35}
$$

$$
\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\tau,h,\boldsymbol{\gamma} \sim N(\boldsymbol{\Omega}\boldsymbol{X}_{\boldsymbol{\gamma}}^t\boldsymbol{y}, (1/\tau)\ \boldsymbol{\Omega}), \tag{1.36}
$$

$$\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\tau, \boldsymbol{y}, h, \gamma \sim \delta_0. \tag{1.37}$$

MCMC algorithm is based on Metropolis-Hastings algorithm, using a simple local proposal to jointly update $h, \pi, \boldsymbol{\gamma}$ with more details in [Guan and Stephens, 2011].

1.4 Outline of the dissertation

While piMASS is a powerful BVS method for GWAS, we identified several ways to improve it, with the novel methods for GWAS analysis in Bayesian framework described in the following chapters. The dissertation is organized as follows.

Chapter 1 contains the review of BVS methods. In particular, we focus on BVS method piMASS, developed specifically for genetic data, which is central to the current dissertation and methodology proposed in later chapters.

Chapter 2 contains the application of the material presented in Chapter 1 to binary phenotype data, which resulted in a publication [Rowe et al., 2019] in the journal *npj Schizophrenia* on November 19, 2019. The work was co-authored with Xiangning Chen (Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, NV, USA), Zuoheng Wang (Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA), Jingchun Chen (Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, NV, USA) and Amei Amei (Department of Mathematical Sciences, University of Nevada, Las

Vegas, NV, USA).

We applied piMASS to the molecular genetics of schizophrenia (MGS) data set with $5,334$ subjects and compared the results with the previous univariate analysis of the MGS data set. The results showed that piMASS can improve the power of detecting schizophrenia-associated SNPs, potentially leading to new discoveries from existing data without increasing the sample size.

In Chapter 3 we consider a Bayesian approach to incorporate non-genetic covariates that account for confounding factors such as population structure for continuous phenotypes. We include such covariates in the model by specifying priors on them and using them in variable selection process. We run a simulation study to investigate the performance of the proposed model by comparing the extended model to the model without population structure correction. In our simulations, we use two sets of genotype data. First set of genotype data is simulated using package COSI2 [Shlyakhter et al., 2014]. Second set of genotype data is a subset of the publicly available MGS data set. We investigate the properties of the extended model using both sets of genotype data. In conclusion, we discuss the accomplishments and limitations of the proposed method.

In Chapter 4, we present a model accounting for population structure covariates with binary phenotypes. We conduct simulations assessing the model's performance. We discuss the model's performance and applicability, as well as its limitations. In addition, we conduct real data analysis and discuss its results.

In Chapter 5 we provide the extension of piMASS to work with ordinal phenotype data with more than two categories, which can significantly expand its applicability to a wider range of existing data sets. Theoretical foundations, practical application as well as real data analysis are provided in this chapter.

Chapter 6 discusses the BVS and its applications presented in Chapters 1 and 2 and summarizes the methods proposed in Chapters 3, 4 and 5 for the Bayesian analysis of GWAS and discusses the directions for future research.

CHAPTER 2

GENOME-WIDE ASSOCIATION STUDY OF SCHIZOPHRENIA USING BAYESIAN VARIABLE SELECTION

2.1 Introduction

In this chapter we present the GWAS of schizophrenia we conducted using BVS method piMASS. The material of this chapter is based on the publication "Biological and practical implications of genome-wide association study of schizophrenia using Bayesian variable selection" [Rowe et al., 2019] in the journal *npj Schizophrenia* on November 19, 2019. The work was co-authored with Xiangning Chen (Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, NV, USA), Zuoheng Wang (Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA), Jingchun Chen (Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, NV, USA) and Amei Amei (Department of Mathematical Sciences, University of Nevada, Las Vegas, NV, USA).

To date GWAS have identified over 100 loci associated with schizophrenia. Most of these studies test genetic variants for association one at a time. In this chapter, we show GWAS of the MGS dataset with $5,334$ subjects using multivariate BVS method piMASS and com-

parison of our results with the previous univariate analysis of the MGS dataset. We showed that piMASS can improve the power of detecting schizophrenia-associated SNPs, potentially leading to new discoveries from existing data without increasing the sample size. We tested SNPs in groups to allow for local additive effects and used permutation test to determine statistical significance in order to compare our results with univariate method. The previous univariate analysis of the MGS dataset revealed no genome-wide significant loci. Using the same dataset, we identified a single region that exceeded the genome-wide significance. The result was replicated using an independent Swedish Schizophrenia Case-Control Study (SSCCS) dataset. Based on the SchiZophrenia Gene Resource database 2.0 (SZGR 2.0), we found 63 SNPs from the best performing regions that are mapped to 27 genes known to be associated with schizophrenia. Overall, we demonstrated that piMASS could discover association signals that otherwise would need a much larger sample size. Our study has important implication that reanalyzing published datasets with BVS methods like piMASS might have more power to discover new risk variants for many diseases without new sample collection, ascertainment, and genotyping.

2.2 Background

Schizophrenia is a severe psychiatric disorder with an estimated global lifetime prevalence of $0.4 - 0.75\%$ with no significant differences across urban, rural, and mixed sites or gen-

ders [Saha et al., 2005], [Moreno-Küstner et al., 2018]. While being a low prevalence disorder, it has substantial societal burden [Charlson et al., 2018]. The estimated heritability of schizophrenia ranges from 70 to 90% [Sullivan et al., 2003]. The common susceptibility variants of such disease are typically identified by association studies, such as genome-wide association studies. In these studies, SNPs are often tested one at a time. In recent years, genetic studies of schizophrenia have made substantial progress. Since the report of the major histocompatibility complex (MHC) locus on chromosome 6 in 2009 [Shi et al., 2009], the number of schizophrenia-associated genetic loci has risen to 5 loci in 2011 [Ripke et al., 2011] and to 108 loci in 2014 [Ripke et al., 2014]. This increase in the number of significant loci could be partially explained by the increase in the sample size of the studies that led to improvement in the statistical power of the association tests. However, these studies also suggest that common variants usually have small to medium effects that makes them hard to reach the typical GWAS significance threshold ($P = 5 \times 10^{-8}$). The application of regression methods on set of genetic variants with appropriate prior specification may have the potential to uncover the largely hidden heritability.

It is well known that the single-SNP approach has its advantage in its simplicity of use, well-established pipeline and low computational burden. However, one of the major drawbacks is that it may miss some potential additive effects derived from sets of SNPs or genes. Methods like Bayesian variable selection take these considerations into account and analyze multiple loci simultaneously. For diseases with complex genetic architecture, such as schizophrenia, it is possible that BVS combined with powerful computing resources might be supe-

rior to single-SNP approach. Indeed, Bayesian methods have demonstrated their abilities in search for genetic risk factors in schizophrenia and other complex disorders [Hall et al., 2007], [Baragatti et al., 2011], [Carbonetto et al., 2012]. Since then, much have been developed in the area of Bayesian GWAS [Servin and Stephens, 2007], [Logsdon et al., 2010]. The piMASS algorithm is one of such examples [Guan and Stephens, 2011]. It offers a BVS procedure that is designed for continuous phenotypes with an extension to binary phenotypes using a probit link function. By considering a set of genetic variants, piMASS extracts more information beyond the marginal associations in standard single-SNP analyses while maintaining reasonable computation time [Guan and Stephens, 2011]. Therefore, piMASS has potential to uncover more associations through reanalysis of existing GWAS datasets.

In this study, we chose a dataset with a moderate sample size, MGS that has previously been analyzed using univariate methods [Shi et al., 2009] and reanalyzed the dataset using piMASS. We hypothesize that piMASS can discover more associations when applied to MGS dataset compared with the single-SNP methodology used in [Shi et al., 2009]. Specifically, we use piMASS to evaluate associations of a set of genetic variants in a moderate sample size of $5,334$ subjects ($2,681$ cases and $2,653$ controls) with binary phenotype using Posterior Inclusion Probabilities (PIPs) - measures of confidence that individual variants have nonzero effects, no interaction effects considered. Such direct comparison with one of the most common GWAS methods can shed the light on the utility of piMASS in analysis of moderate size datasets. We used permutation test to validate our findings with an independent schizophrenia case-control dataset of similar size ($2,895$ cases and $3,836$ controls). Our

results indicate that compared with single-SNP approaches, BVS method, such as piMASS, could discover association signals with a relatively small sample size that might have been undetectable by single-SNP approaches.

2.3 Methods

2.3.1 GWAS datasets

In discovery, we performed association analysis using the MGS study ($n = 5,334$) that consists of $2,681$ schizophrenia cases and $2,653$ healthy controls of European ancestry. Details of the dataset have previously been described [Shi et al., 2009]. In validation, we chose batches 5 and 6 of the SSCCS dataset ($n = 6,731$), including $2,895$ cases and $3,836$ controls [Ripke et al., 2011].

Both MGS and SSCCS GWAS datasets were downloaded from NIMH Genetic Repository and Resource (`https://www.nimhgenetics.org/`) upon approval. The genotypes were downloaded from NIMH without further quality check because the genotypes from NIMH were checked and met the standard requirement of NIMH. The two datasets were genotyped using different platforms: the MGS was genotyped using the Affymetrix 6.0 chip that includes $638,937$ SNPs, and the batches 5 and 6 of the SSCCS were genotyped using the Illumina OmniExpress chip that includes $646,699$ SNPs. In this work, we did not do any SNP annotation as both datasets are using GRCh37/hg19 as the human reference genome. While

the SSCCS dataset has more subjects, its order of magnitude is approximately the same as of MGS ($5,334$ subjects with $638,937$ SNPs in MGS vs. $6,731$ subjects with $646,699$ SNPs in SSCCS) in a sense that we expect both datasets to have similar power in detecting the associations.

The genotype of an individual is coded as 0, 1, or 2 whether the subject has 0, 1, or 2 copies of the minor allele. Missing genotypes were replaced by the sample average of the genotypes at the position where the genotype is missing. Phenotypes were recorded as a binary variable indicating presence or absence of a schizophrenia diagnosis.

2.3.2 Study design

The objective of the current study is to evaluate whether piMASS can improve the detection of association signals as compared with standard univariate procedure. To this end, we chose a dataset with a moderate sample size that has previously been analyzed using univariate methods [Shi et al., 2009]. We did not perform genotype imputation and used the same set of markers as in the study [Shi et al., 2009]. Guan and Stephens mention that BVS regression tends to spread the association signal (the PIPs) among the correlated SNPs [Guan and Stephens, 2011]. Therefore, to apply piMASS, we partitioned the genome-wide data into smaller regions to capture additive effects of neighboring SNPs. Based on our computational resources, we set $1,000$ SNPs as the region in a single run. Given that we did not impute genotypes, it was more practical to proceed with regions containing equal

number of SNPs. Since piMASS searches for various model configurations by proposing to add, remove, and switch covariates in the model, we used a sliding window approach in which each chromosome was cut into regions of $1,000$ SNPs with the overlap of 500 SNPs. This ensures that piMASS has the opportunity to explore models containing all nearby SNPs. This approach produced $1,266$ regions in the MGS dataset. A typical region spans around 37 million base pairs.

We tested each region for association with phenotype using piMASS by performing MCMC runs with one million iterations each. The convergence of the 2 parallel MCMC runs was confirmed by the Gelman-Rubin statistic being < 1.04. We did not include covariates to correct for potential population stratification since it has been noted in the literature that Bayesian regression models simultaneously fitting multiple SNPs are robust for population stratification [Kärkkäinen and Sillanpää, 2012]. To distinguish SNPs with the strongest evidence of association, we use PIP for each SNP. Since nearby SNPs are usually correlated and the PIPs can spread around correlated SNPs, it is possible that none of the single SNPs in the region have high PIP but the sum of PIPs would be high indicating the posterior probability of at least one of the SNPs should be included in the model. The design of combining multiple loci into a region helps better explore all possible models. Given that the regions are defined in terms of fixed number of SNPs, we used the sum of PIPs as the main measurement of association following analysis [Guan and Stephens, 2011].

piMASS allows user to input parameters (priors) appropriate for the specific question in a GWAS and therefore utilize the existing domain knowledge. From the latest GWASs, it is

known that more than 100 genetic loci are associated with schizophrenia, but each locus has very small effect [Shi et al., 2009], [Ripke et al., 2011], [Ripke et al., 2014]. This knowledge can be utilized to specify the ranges of the prior parameters. In the model, we specified the following two parameters: the proportion of the phenotypic variance explained by relevant variants and the proportion of SNPs that we expect to be relevant to the phenotype. The first parameter represents the estimate of overall signal in the genotype data, e.g., how much variation in the diagnosis can be explained by the SNP data. We set a prior on this quantity to be uniformly distributed from 0.01% to 1% according to the two previous schizophrenia studies [Ripke et al., 2014], [Lee et al., 2012]. These two studies suggest that the variants across the genome collectively explain 18-23% of phenotype variation [Ripke et al., 2014], [Lee et al., 2012]. But for each variant, the variation explained is very small. Ripke et al. [Ripke et al., 2011] found 108 variants, the expected proportion explained by a single variant would be $0.184/108 = 0.0017$, which falls in the range we used as a priori information in the model. The prior on the second parameter was set in such a way that the expected number of SNPs that are relevant to the phenotype ranges from 1 to 5 loci in each region containing a group of $1,000$ SNPs. One could also set restriction on the total number of SNPs to be allowed in the model. In this study, we set it to 5 SNPs due to computation time considerations.

The key feature of prior setting in piMASS is that the number of relevant variables is no longer necessarily positively correlated to the proportion of variance explained by them. Such prior structure is proper for the situation where there are many relevant variables, each has

tiny effect and overall proportion of variability explained by relevant covariates is still small. This feature of piMASS matches the genetic architecture of schizophrenia where many loci, each with a very small effect, collectively contribute to the disease.

In the initial application of piMASS to the MGS dataset, sum of PIPs for the $1,266$ regions in the MGS dataset did not show clear separations (Fig. 2.1). The sum of the PIP for each region varies, yet there are no outstanding spikes that can be seen from the graph. Therefore, it was not clear which regions contain SNPs that are associated with the trait. We borrowed frequentist permutation test to determine an appropriate significance threshold, which would make our results comparable with those reported previously [Shi et al., 2009]. We used empirical method based on the Fishers concept of a permutation test. Given that 1266 overlapping regions were tested simultaneously, it was necessary to correct the significance threshold for multiple comparisons. We chose Bonferroni correction for this purpose. Since our focus is on gene discovery, we set $\alpha = 0.1$, being more liberal than the traditional 5% level to compensate for the well-known conservativeness of Bonferroni multiple correction. Although the larger may produce more false positives, we expect to eliminate them in the validation step of the study. Each region that survived the significance threshold of $\frac{p}{n} = \frac{0.1}{1266} \approx 7.9 \times 10^{-5}$ was further validated using an independent dataset. The target number of permutations was set to $100,000$. To save computation time, we stopped permutations for each region once the regions empirical p-value exceeded the significance threshold.

.

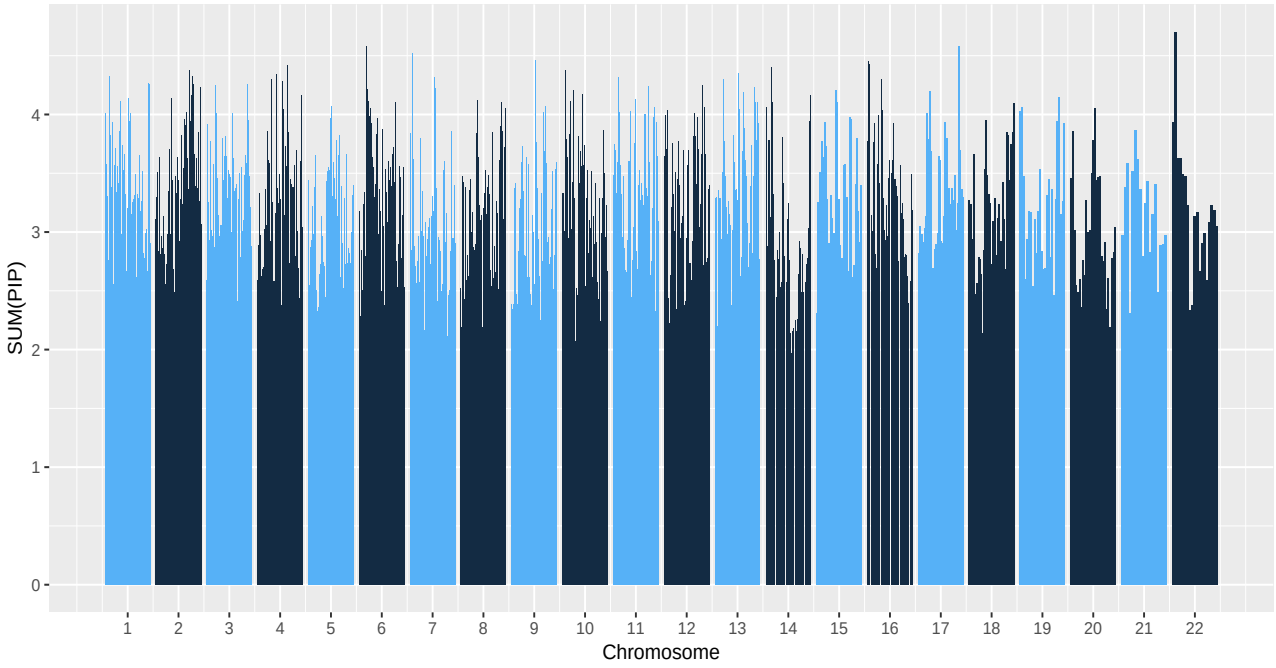

Figure 2.1: piMASS genome-wide region-based performance of the MGS dataset. The sum of posterior inclusion probabilities (PIPs) for each of the $1,266$ overlapping regions spanning 22 chromosomes of the MGS dataset

2.3.3 Validation

The main goal of validation is to check if the results obtained using piMASS could be replicated in an independent dataset so that the method could be applied to discover new loci in general. Several regions in the discovery dataset that performed best in terms of the empirical p-values from the permutation were tested and verified using an independently collected dataset (SSCCS dataset) with similar sample size. As mentioned above, the two datasets were genotyped using different platforms with different sets of SNPs. As we only

included the SNPs that belong to the region with the same position interval, the number of SNPs in each validation region is not exactly $1,000$. Accordingly, the mean of PIPs is used as a measure of association to account for the variable number of markers in the discovery and the corresponding validation regions. We use $\alpha = 0.05$ and correct for multiple comparisons based on the number of regions undergoing the validation. This yields empirical p-value based on $1,000$ permutations for each validation region.

Since the best performing regions based on the permutation test are in the top 2% of all regions according to the initial run using the MGS dataset, we also conducted the separate piMASS analysis of SSCCS dataset with no permutation test, followed by an overlap analysis of both datasets based on piMASS results without permutation test. In detail, top 5% of the best performing regions based on sum of PIPs were selected. For each such region, top 1% of the best performing SNPs based on PIPs were selected, comprising a table of SNPs with the highest PIP among regions with the highest sum of PIPs for each dataset. We checked every SNP in each table if it is within $100,000$ base pairs (bp) distance of SNPs from the other dataset. Those SNPs comprise overlap set between the two datasets. Next, for each pair in the overlap set we checked LD between the SNPs in the pair. If two SNPs were in LD, they were considered in the consensus set. The larger size of the consensus set points toward consistency of the piMASS method.

2.4 Results

In the discovery dataset MGS, region 29 on chromosome 15 containing SNPs between $83,907,801$ and $86,887,657$ reached genome-wide significance after Bonferroni correction (Rank 1, $P_{disc} = 1.43 \times 10^{-5}$, $P_{vali} = 0.001$). rs16940789, rs16941261, rs4887364, rs991728, rs2114252, and rs994068 are among the SNPs with top 1% highest PIP and mapped to gene *NTRK3* in the SchiZophrenia Gene Resource database, SZGR 2.0 (`https://bioinfo.uth.edu/SZGR/`), a comprehensive database of variants and genes reported to have an association with schizophrenia [Jia et al., 2017]. *NTRK3* has been shown to be associated with bipolar and other psychiatric disorders [Nurnberger et al., 2014], [Forstner et al., 2014], [Verma et al., 2008]. The gene encodes a member of the neurotrophic tyrosine receptor kinase (*NTRK*) family, which is involved in the nervous system. rs16940789 was also mapped to gene LINC00052 (an RNA gene that is affiliated with the noncoding RNA). The locus had not been previously reported in refs [Ripke et al., 2014], [Ruderfer et al., 2019].

Although only one locus surpassed Bonferroni correction ($P_{disc} < 7.9 \times 10^{-5}$), some regions with the empirical p-value (P_{disc}) that are close to the cutoff might still be of interest because Bonferroni correction is known to be too conservative. We used a design of sliding window with overlapping SNPs. Table 1 lists 12 regions with the best association metric (P_{disc}) based on $100,000$ permutations using the MGS dataset, as well as their corresponding empirical p-values based on 1000 permutations using the SSCCS dataset (P_{vali}). For each of the 12 regions reported in Table 1, we ranked the 1000 SNPs in terms of the PIP generated from the initial run using the MGS dataset and listed the top ten SNPs based on PIP within each

region. The Manhattan plot of the PIP of individual SNP from the initial run using the MGS dataset is shown in Fig. 2.2 using $-\log_{10}(1-\text{PIP})$ as the y-axis.

Chr	Region [a]	Start position[b]	End position[b]	Rank[c]	P_{disc}[d]	P_{vali}[e]
15	29	83,907,801	86,887,657	1	1.43E-05	0.001
19	5	15,724,023	22,638,628	2	1.67E-04	<0.001
14	33	86,399,092	90,573,122	3	2.20E-04	<0.001
9	24	34,905,605	70,379,322	4	2.50E-04	0.002
14	6	29,288,170	33,177,081	5	3.00E-04	<0.001
8	30	53,113,091	57,376,926	5	3.00E-04	0.001
20	1	9795	2,715,620	7	3.33E-04	<0.001
18	15	28,642,588	33,646,071	8	5.00E-04	<0.001
15	28	80,260,648	85,190,202	9	5.50E-04	<0.001
1	36	81,955,643	85,727,849	10	5.67E-04	<0.001
13	41	98,395,342	101,641,945	11	6.00E-04	<0.001
3	15	22,010,347	25,354,138	12	7.50E-04	0.001

Table 2.1: Regions with best association metrics (P_{disc}) based on permutation test

[a] Regions were assigned separately to each chromosome starting from 1
[b] Start position reflects the position of the first SNP included in the region, end position reflects the position of the last SNP included in the region
[c] Rank is based on empirical p-value calculated from permutation test using the MGS dataset
[d] Empirical p-value based on $100,000$ or less permutations using the discovery dataset (MGS)
[e] Empirical p-value based on $1,000$ permutations using the validation dataset (SSCCS)

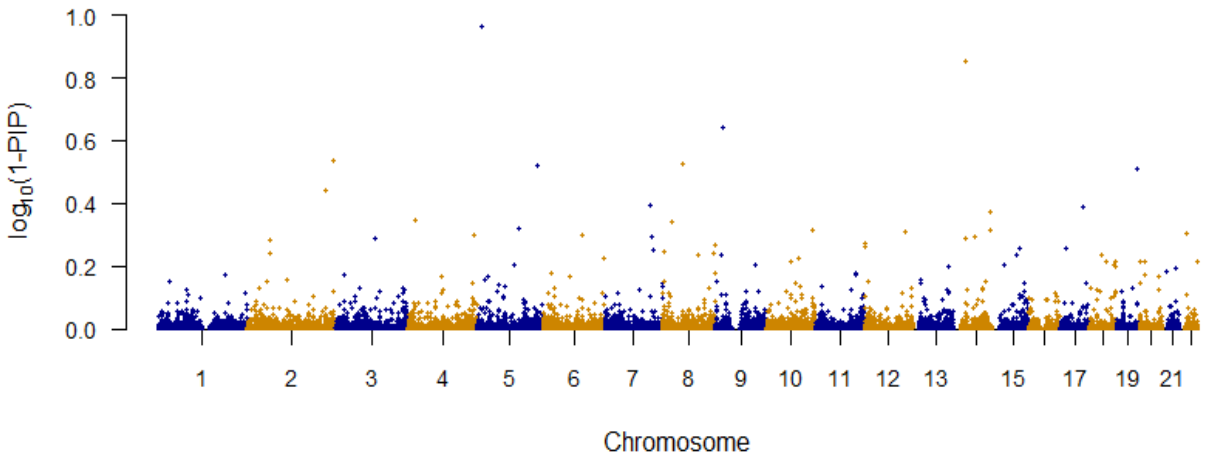

Figure 2.2: Manhattan plot of 1-PIP for the MGS dataset

Among the top 1% of the SNPs of the validated best performing regions in permutation test, there are 5 SNPs that have been mapped to genes associated with schizophrenia in the GWAS Catalog [MacArthur et al., 2017] (Table 2). Average C-scores based on Combined Annotation-Dependent Depletion (CADD) method are also listed in Table 2 [Kircher et al., 2014]. SNPs rs993804 and rs4858697, located at 3p24.2, are in Linkage Disequilibrium (LD) ($R^2 = 0.45$) and are mapped to the gene AC092422.1 (*RARB*) (chr3:24687919-25174305). AC092422.1 (*RARB*) had been reported to be associated with schizophrenia and bipolar disorder in a meta-analysis for genome-wide association data using European-American samples [Wang et al., 2010]. SNP rs2044117, located at 13q32.3, is mapped to genes *NALCN-AS1* and *NALCN*. *NALCN-AS1* was reported to be associated with schizophrenia and bipolar disorder [Wang et al., 2010]. The *NALCN* was reported to be associated with multiple traits including bipolar disorder, eating disorder, schizophre-

nia, adolescent idiopathic scoliosis, HIV-associated dementia, psychosis, recurrent major depressive disorder, etc. [Hall et al., 2018], [Kanazawa et al., 2013], [Levine et al., 2012], [Liu et al., 2018], [Liu et al., 2016]. SNP rs9554752 is also mapped to *NALCN* and it is in LD with rs2044117 ($R^2 = 0.11$). SNP rs915071, located at 14q12, is mapped to genes AL352984.2: LOC105370439 and LOC105370440 that were reported to be associated with schizophrenia and bipolar disorder [Wang et al., 2010]. The CADD scores for SNPs in Table 2 range from 0.898 to 9.444. Resulting CADD scores point to the fact that piMASS alone cannot discover causal variants. The main reason is that piMASS is a tool for association testing, and hence the SNPs discovered are not necessarily causal. Moreover, other study characteristics like region-based design and unimputed dataset add to the fact that additional steps may be necessary to investigate the pinpointed regions for causal SNPs.

Chr	Gene Region	SNP	Position[a]	MAF[b]	PIP	C-score[c]
3	AC092422.1 (RARB)	rs993804	25,070,680	0.27	0.059	5.917
3	AC092422.1 (RARB)	rs4858697	25,075,091	0.46	0.044	2.97
13	NALCN, NALCN-AS1	rs2044117	101,055,958	0.13	0.124	9.444
13	NALCN	rs9554752	101,073,961	0.35	0.040	1.960
14	LOC105370439,LOC105370440	rs915071	31,964,652	0.40	0.738	0.898

[a] Position is referred to NHGRI-EBI GWAS Catalog
[b] Minor allele frequency (MAF) in the 1,000 Genomes Phase 3 combined population
[c] Average C-score based on Combined AnnotationDependent Depletion (CADD) method

Table 2.2: SNPs with their mapped genes

Based on the permutation test on the discovery dataset, region 5 on chromosome 19 has the second smallest empirical p-value (Rank 2, $P_{disc} = 1.67 \times 10^{-4}$, $P_{vali} \leq 0.001$. Among

the SNPs having the highest 1% PIPs within this region, there are six SNPs (rs2965189, rs2916074, rs4808200, rs4808203, rs4808964, and rs10419912) and they are in high LD ($R^2 \geq 0.93$) with rs2905426 located at 19p13.11. The SNP rs2905426 is a variant belonging to a regulatory region of genes *GATAD2A* and *MAU2*. This variant was previously reported to be associated with schizophrenia from the Psychiatric Genomic Consortium (PGC) study, where 128 independent associations with 108 conservatively defined loci were identified in a GWAS of up to $36,989$ cases and $113,075$ controls [Ripke et al., 2014] (see also refs [Lieberman et al., 2005], [Hindorff et al., 2009], [Sullivan et al., 2012]). Based on SZGR 2.0, we found 63 SNPs out of the 120 SNPs that are mapped to 27 genes in the SZGR 2.0 database and shown to be associated with schizophrenia.

We also conducted an overlap analysis between MGS and SSCCS datasets based on a single run of piMASS without the permutation test. The results of the piMASS analysis of the SSCCS dataset based on sum of PIPs for each region are presented in Figure 2.3. Figure 2.4 shows the Manhattan plot based on individual PIPs using $-\log_{10}(1 - PIP)$ as the y-axis. The overlap analysis suggests that 46.3% of the pairs were in LD with each other in the $100,000$ bp overlap among top 1% of the SNPs of the top 5% of the regions. rs7746199 chr6:27261324 belongs to extended MHC region and has been previously implicated in association with schizophrenia [Shi et al., 2009]. rs7746199 as well as rs2747421, rs2535238, rs375984, rs2747421, rs2535238, rs375984 SNPs are in the consensus set and are in LD with rs1153229265 of chr6 reported by PGC7.

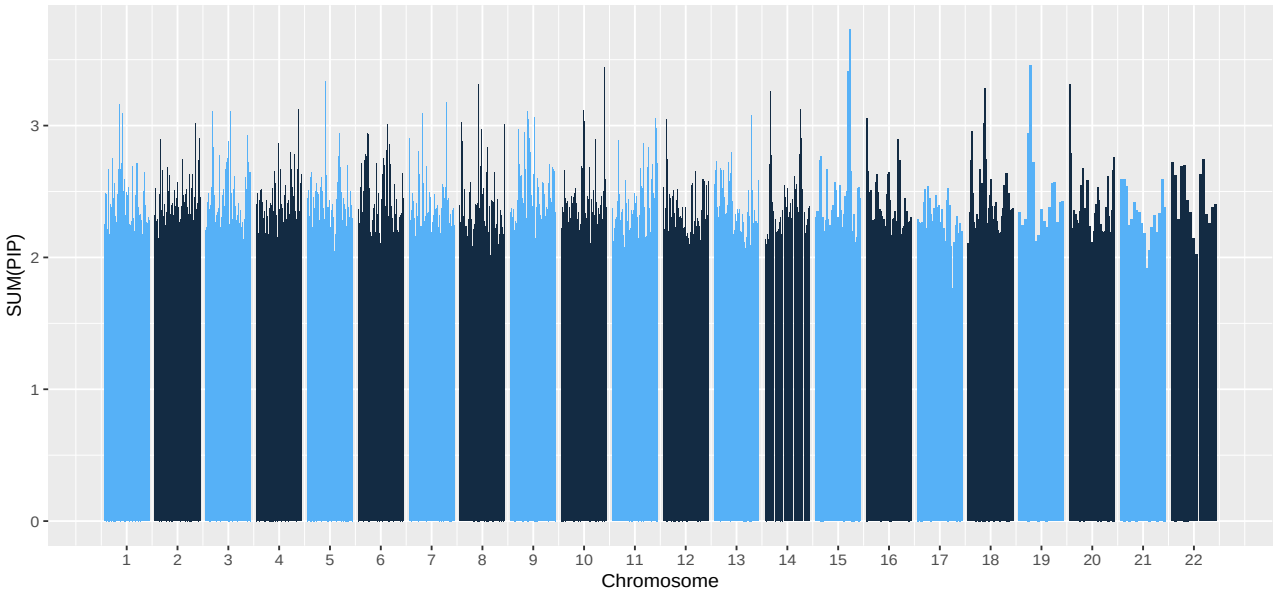

Figure 2.3: piMASS genome-wide region-based performance of the SSCCS dataset. The sum of posterior inclusion probabilities (PIPs) for each of the 1244 overlapping regions spanning 22 chromosomes of the SSCCS dataset

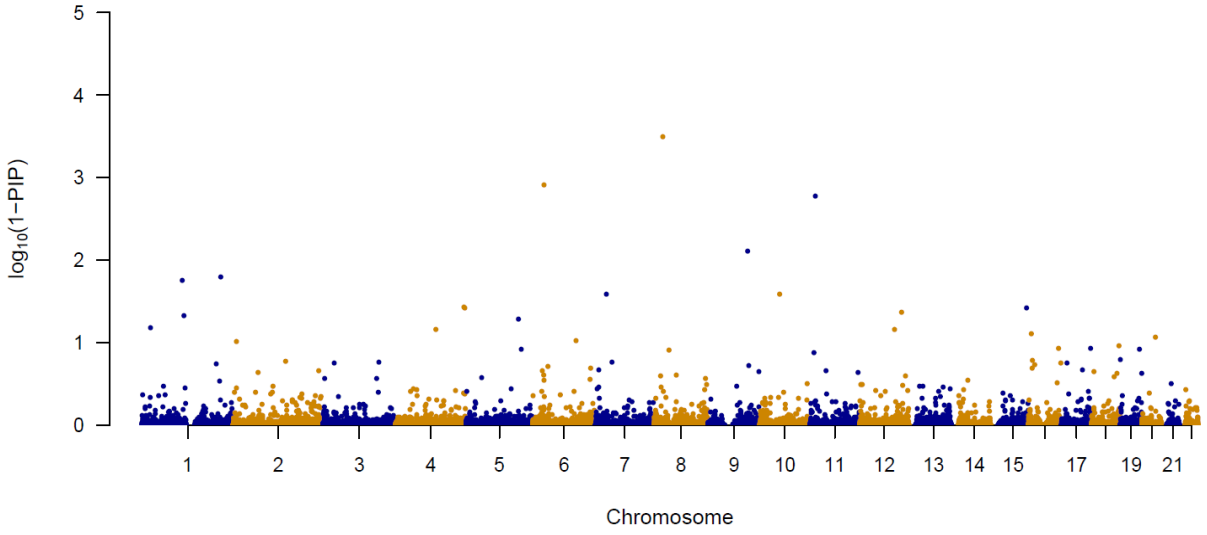

Figure 2.4: Manhattan plot of 1-PIP for the SSCCS dataset

2.5 Discussion

Guan and Stephens developed a BVS regression model for large-scale datasets primarily focusing on analysis of quantitative traits [Guan and Stephens, 2011]. In this study, we used the BVS regression model to conduct a case-control GWAS of schizophrenia with binary phenotype for datasets with moderate sample sizes (cases/controls for MGS and SSCCS datasets are $2,681/2,653$ and $2,895/3,836$, respectively). We have demonstrated that the BVS methods can discover association signals that otherwise would need a much larger sample size to discover.

Application of BVS to the MGS European ancestry case-control sample produced 17 regions having $P_{disc} < 8 \times 10^{-4}$ based on 100,000 permutations. Among them, 12 regions were validated using the SSCCS dataset with $P_{vali} \leq 0.002$ based on $1,000$ permutations. The region with the smallest p-value, which belongs to chromosome 15 reached genome-wide significance. SNPs with the highest PIP from this region are mapping to gene *NTRK3* that encodes a member of the *NTRK* family and has been reported to be associated with bipolar and other psychiatric disorders. Five SNPs among the twelve validated regions are mapped to genes that are known to be associated with schizophrenia and other mental disorders, such as bipolar disorder, eating disorder, adolescent idiopathic scoliosis, psychosis, recurrent major depressive disorder [Wang et al., 2010], [Liu et al., 2016], [Kanazawa et al., 2013], [Levine et al., 2012], [Liu et al., 2018], [Hall et al., 2018]. A cluster of six SNPs on chromosome 19 are found to be in high LD with rs2905426, which is mapped to the regulatory region of *GATAD2A* and *MAU2*, genes that are known to be associated

with schizophrenia and bipolar disorder [Ripke et al., 2014].

Our BVS analysis of the MGS European ancestry case-control dataset identified one region that reached genome-wide significance, while the original GWAS of the MGS case-control dataset that used single-SNP approach did not find any significant signal [Shi et al., 2009]. This result indicates that piMASS method has the potential to uncover more associations even in moderate sample size setting, as compared with the single-SNP approach. Our results suggest that BVS methods, such as piMASS, can be used to reanalyze published datasets to discover new risk variants for many diseases without new sample collection, ascertainment, and genotyping.

Our analysis produced a single region that achieved genome-wide significance that has not been reported in large-scale schizophrenia GWAS. The region is on chromosome 15 containing SNPs between $83,907,801$ and $86,887,657$. Among SNPs with top 1% highest PIP in this region is rs16940789 in the genes *LINC00052* and *NTRK3*.

While we were able to demonstrate the potential superiority of piMASS over standard GWAS, there are a few ways it could be further improved. First, performing genotype imputation for both the discovery and validation datasets could provide more precise comparison between the two datasets. Imputation enables direct comparison between datasets, which could be beneficial to the understanding of the piMASS performance. Second, given that population stratification and cryptic relatedness are among the confounding factors in genetic association studies [Astle et al., 2009], [Price et al., 2010], a strategy that accounts for population structure could improve the accuracy of association discovery and extend the application to

datasets with less homogeneous population structure. Third, comparison of piMASS with other methods beyond single-SNP approach can help placing it in a hierarchy of other GWAS tools for real data. Another potential direction of further research is the extension of the model to handle categorical response data, thus allowing to analyze phenotypes with polychotomous scale such as addiction and other diseases. The classical approach to multinomial response data is to fit a categorical response regression using maximum likelihood and make inference about the model based on the associated asymptotic theory. It has been pointed out that the inference based on the classical approach is questionable for small sample sizes and Bayesian methods provided an attractive alternative [Albert and Chib, 1993]. Having reached the conclusion that it is possible to uncover more associations using single dataset with moderate sample size, it now makes sense to move on to apply piMASS to larger, more heterogeneous datasets, imputed datasets and ultimately perform meta-analysis of the results of the BVS analysis of multiple datasets.

2.6 Data availability

The molecular genetics of schizophrenia (MGS) data and batches 5 and 6 of the Swedish Schizophrenia CaseControl Study (SSCCS) data that support the findings of this study are available in the NIMH Genetic Repository and Resource (`https://www.nimhgenetics.org/`) upon approval of NIMH.

2.7 Code availability

The code used to generate the output has been uploaded to GitHub public repository (`https://github.com/rowedata/sczGWAS`) under MIT license and includes the description of the input data format.

CHAPTER 3

BAYESIAN VARIABLE SELECTION METHOD WITH POPULATION STRATIFICATION CORRECTION FOR GWAS WITH CONTINUOUS PHENOTYPE

3.1 Background

When performing GWAS, the unknown covariance structure stemming from ignored origin of individuals from multiple populations (population stratification) or their relatedness can lead to false association signals, not related to the association of tested genetic marker and phenotype. Current pooled large scale genetic data sets can contain distantly related subjects. Such genetic relatedness prevents standard association studies from correctly identifying the causal variants and leads to many false positive associations [Martin and Eskin, 2017]. Some studies suggest that modest amounts of stratification can exist even in well-designed studies [Freedman et al., 2004]. Therefore, controlling for those confounding factors is important for precise identification of genetic associations.

There is evidence that in multi locus BVS models, genetic relationship between individuals can be captured by genetic markers themselves [Kärkkäinen and Sillanpää, 2012]. Other

studies suggest that inclusion of additional covariates should significantly improve the power of BVS models [Banerjee et al., 2018]. It is possible that some applications of Bayesian association models can benefit from incorporating population stratification covariates [Iwata et al., 2009]. In this chapter we explore such solution for BVS method piMASS for GWAS with continuous phenotype.

Various methods have been developed to correct for population stratification. Those methods could be divided into two broad categories. In the first category, there are methods such as genomic control, which assume that unobserved genealogy is creating dependence among individuals, thus inflating the regular test statistic and one can directly estimate dependence structure and use it to correct the test statistic [Devlin and Roeder, 1999]. In the second category, population membership or proportion of ancestry from different populations can be thought of as unmeasured covariates [Wu et al., 2011]. One such method is the method of structured associations, which uses a set of unlinked genetic markers to estimate the ancestry of sampled individuals to use in the association test [Pritchard et al., 2000]. Another such method is the method of principal components, which uses top eigenvectors of the sample covariance matrix as covariates in a regression setting [Price et al., 2006]. In this chapter, we want to include the non-genetic covariates accounting for population stratification in the BVS multiple regression model. This makes the suggested approach similar to the methods from the second category.

In section 2 we propose a BVS model with additional covariates that can control for population stratification for continuous phenotypes. In section 3 we provide the description and

results of simulation studies evaluating the performance of the proposed model. In section 4 we discuss the results and point out directions for further research.

3.2 Methods

3.2.1 Priors for non-genetic covariates

Following the approaches that include population stratification covariates in the regression model, we want to incorporate the confounding variables into piMASS to account for their effects.

Consider a multiple linear regression model for subject i:

$$y_i = \mu + \alpha_1 x_{i1} + ... + \alpha_q x_{iq} + \beta_1 g_{i1} + ... + \beta_p g_{ip} + \epsilon_i, \tag{3.1}$$

where y_i is the phenotype of subject i, μ is an intercept, $x_{i1}, ... , x_{iq}$ are q non-genetic covariates that can be used to account for population stratification and $\alpha_1, ... , \alpha_q$ are their corresponding regression coefficients, $g_{i1}, ... , g_{ip}$ are p genetic variants measured at ith individual with their corresponding regression coefficients $\beta_1, ... , \beta_p$ and ϵ_i is an error term. We assume that ϵ_is are independent and identically distributed with $\epsilon_i \sim N(0, \tau^{-1})$.

The relationship between the phenotype vector $\boldsymbol{y}$, genotype matrix $\boldsymbol{G}$ and covariate matrix $\boldsymbol{X}$ can be expressed as:

$$f(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \tau) = N_n(\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{G}\boldsymbol{\beta},\ \tau^{-1}\boldsymbol{I_n}), \tag{3.2}$$

where $\boldsymbol{y}$ is an n-vector of phenotypes for n individuals, $\boldsymbol{X}$ is an $n \times q$ matrix of non-genetic covariates, $\boldsymbol{\alpha}$ is a q-vector of the corresponding regression coefficients, $\boldsymbol{G} = (\boldsymbol{g_1}, \dots, \boldsymbol{g_p})$ is an $n \times p$ matrix of p genetic variants. Here $\boldsymbol{g_j} = (g_{1j}, \dots, g_{nj})'$ is a column vector containing the observed values of jth genetic variant or SNP at the n individuals. Effects of the genetic variants are of primary interest and are represented by a p-vector $\boldsymbol{\beta}$.

Following the framework laid out in Chapter 1, we use a vector of binary indicators defined in (1.3) to represent possible models describing the relationship between response and covariates:

$$\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{G}, \boldsymbol{X}, \tau \sim N_n\big(\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{G_\gamma}\boldsymbol{\beta_\gamma},\ \tau^{-1}\boldsymbol{I_n}\big), \tag{3.3}$$

where $\boldsymbol{G_\gamma}$ denotes the matrix $\boldsymbol{G}$ restricted to those columns j for which $\gamma_j = 1$ and $\boldsymbol{\beta_\gamma}$ is a corresponding vector of regression coefficients. Here, selection is performed only for genetic covariates and all members of $\boldsymbol{X}$ are always included in the model.

The priors on the model parameters $\tau, \mu, \boldsymbol{\gamma}, \boldsymbol{\beta}$ and their corresponding hyperparameters are the same as in piMASS and are given by equations (5.4)-(5.8). Next, we specify the prior for the set of parameters $\boldsymbol{\alpha}$. When choosing a prior for $\boldsymbol{\alpha}$ we consider its two main properties: it should reflect our prior knowledge and be reasonable in terms of computational complexity.

In cases when we do not have enough information about parameters, which in our case is true for $\boldsymbol{\alpha}$, we can instead construct a prior that is minimally informative to reflect the state of our prior knowledge [Hoff, 2009]. Therefore, for the set of parameters $\boldsymbol{\alpha}$ we propose a unit information prior [Kass and Raftery, 1995].

The unit information prior is derived in the following way. In (3.3), the precision of $\boldsymbol{\alpha}$ is its inverse variance, or $\tau(\boldsymbol{X}'\boldsymbol{X})$. Since it can be viewed as information contained in n observations, the amount of information in one observation should be $\frac{1}{n}\tau(\boldsymbol{X}'\boldsymbol{X})$. Thus, we propose the following prior on the coefficients of the non-genetic covariates:

$$\boldsymbol{\alpha}|\tau \sim N_q\Big(0, \frac{n}{\tau}\big(\boldsymbol{X'X}\big)^{-1}\Big), \tag{3.4}$$

where q is the number of non-genetic covariates included in the model. One advantage of such prior is that the estimation is invariant to the change in the scale of the covariates [Hoff, 2009]. The proposed prior distribution cannot be considered a real prior distribution, as it requires knowledge of $\boldsymbol{y}$ to be constructed. Nevertheless, the amount of information used is small and can be thought of as the prior distribution of a person with unbiased but weak prior information [Hoff, 2009].

3.2.2 MCMC scheme for additional covariates

Our MCMC setting preserves most of the parameter updates of piMASS. This allows us to perform the genetic covariate selection while accounting for the effects of additional covariates, such as the ones accounting for population stratification.

The priors for parameters as well as hyperparameters are listed below in detail:

$$
\begin{aligned}
\tau &\sim \Gamma(\lambda/2, k/2), \\
\mu|\tau &\sim N(0, \sigma_\mu^2/\tau), \\
\gamma_j &\sim Bernoulli(\pi), \\
\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\tau, \boldsymbol{\gamma} &\sim N_{d_\gamma}\big(\mathbf{0}\ , (\sigma_a^2/\tau)\ \boldsymbol{I}_{d_\gamma}\big), \\
\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma} &\sim \delta_0, \\
h &\sim U(0,1), \\
\log(\pi) &\sim U\big(log(1/p), \log(M/p)\big), \\
\boldsymbol{\alpha}|\tau &\sim N_m\Big(\mathbf{0}, \frac{n}{\tau}(\boldsymbol{X}'\boldsymbol{X})^{-1}\Big).
\end{aligned}
\tag{3.5}
$$

where $\lambda, k, \sigma_\mu, M, p, n$ are given constants, $U(x|a,b) = \frac{1}{b-a}$ is the continuous uniform density function, $\Gamma(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$ is the gamma density function, $N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is the probability density function of a Normal distribution, $Bernoulli(x|\pi) = \pi^x(1-\pi)^{1-x}$ is the Bernoulli probability mass function and $N_k(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-k/2}|\boldsymbol{\Sigma}|^{-1/2}e^{(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu}))}$ is the probability density of a multivariate normal distribution.

There are several practical solutions that were implemented to facilitate computational simplicity. First, by centering the response variable and the covariates (genetic and non-genetic) to have mean 0, we set $\mu = 0$ in the model. Additionally, the maximum number of SNPs considered for association with the phenotype, M, is set to 300 to limit the computational burden.

We initialize the model parameters by sampling values from their respective prior distribu-

tions or setting them to previously specified values. We use $\phi^{(0)}$ to denote the set of initial values of the model parameters:

$$\phi^{(0)} = \{h^{(0)}, \pi^{(0)}, \boldsymbol{\gamma}^{(0)}, \tau^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}\}. \tag{3.6}$$

Under the conditions described in (3.5) and denoting $pr(h, \pi, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau)$ as a joint prior on model parameters, the posterior distribution is proportional to the following expression:

$$\begin{aligned}
p(h, \pi, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau | \boldsymbol{y}) &\propto L(\boldsymbol{y} | h, \pi, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau) \times pr(h, \pi, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau) \\
&= L(\boldsymbol{y} | h, \pi, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau) \times p(h) \times p(\boldsymbol{\gamma} | \pi) \times p(\pi) \\
&\times \quad p(\boldsymbol{\alpha} | \tau, \sigma_a) \times p(\boldsymbol{\beta} | \tau, h) \times p(\tau) \\
&= \prod_{j=1}^{n} N_k(\boldsymbol{y_j} | \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau^{-1} \boldsymbol{I}_{\boldsymbol{d_\gamma}}) \times U(h | 0, 1) \times Bernoulli(\gamma | \pi) \\
&\times U\big(\log(\pi) | \log(1/p), \log(M/p)\big) \times N_q\Big(\boldsymbol{\alpha} | \mathbf{0}, \frac{n}{\tau}\big(\boldsymbol{X'X}\big)^{-1}\Big) \\
&\times N_p\big(\boldsymbol{\beta}_{\boldsymbol{\gamma}} |\ \mathbf{0}, \tau^{-1} \boldsymbol{I}_{\boldsymbol{d\gamma}}\big) \times \Gamma(\tau | \lambda, k).
\end{aligned} \tag{3.7}$$

In Bayesian inference, we first require the joint posterior distribution of all unknowns, and then we integrate this distribution over the unknowns that are not of immediate interest to obtain the desired marginal distribution of the parameters of interest [Gelman et al., 2013]. Here, explicitly computing the normalizing constant for our model and providing the full posterior distribution in closed form is problematic due to the high dimensionality of the parameter space. Therefore, we introduce MCMC sampling scheme needed to perform the estimation of the proposed model.

Let

$$\boldsymbol{Z} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}. \tag{3.8}$$

Then $\boldsymbol{Z}$ contains the variation in $\boldsymbol{y}$ left after subtracting the current estimates of the effects $\boldsymbol{\alpha}$ of the confounding variables $\boldsymbol{X}$. Therefore, the distribution of $\boldsymbol{Z}$ is

$$\boldsymbol{Z}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{G}, \tau \sim N_n(\boldsymbol{G}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau^{-1}\boldsymbol{I}_{\boldsymbol{n}}), \tag{3.9}$$

which is the distribution of the phenotype assumed in piMASS. Here, the conditional distribution of $\boldsymbol{Z}$ follows the model originally proposed in [Guan and Stephens, 2011]. Since the effects of non-genetic covariates have been accounted for, we can apply the MCMC scheme proposed in piMASS to $\boldsymbol{Z}$. Next, we examine the posterior conditional distribution of the parameters (3.7) to determine the detailed updating strategies.

Parameters $h, \pi, \boldsymbol{\gamma}$

In (3.7) we pull out the densities for $h, \pi, \boldsymbol{\gamma}$ to obtaint samples from their joint posterior distribution on the product space $(0,1) \times (0,1) \times \{0,1\}^p$, which is given by:

$$\begin{aligned} p(h, \pi, \boldsymbol{\gamma}|\boldsymbol{Z}) &\propto p(h) \times p(\boldsymbol{\gamma}|\pi) \times p(\pi) \times p(\boldsymbol{Z}|h, \pi, \boldsymbol{\gamma}) \\ &= U(h|0,1) \times Bernoulli(\gamma|\pi) \times U\big(\log(\pi)\big|\log(1/p), \log(M/p)\big) \\ &\times p(\boldsymbol{Z}|h, \boldsymbol{\gamma}), \end{aligned} \tag{3.10}$$

where $p(\boldsymbol{Z}|h, \boldsymbol{\gamma})$ can be computed by integrating out $\boldsymbol{\beta}$ and τ analytically [Guan and Stephens, 2011]. Since the full conditional distributions for $h, \pi, \boldsymbol{\gamma}$ are not avail-

able in closed form, we use Metropolis-Hastings algorithm. Metropolis-Hastings algorithm requires neither proposal distribution to be symmetric nor posterior conditional distribution to be in a closed form. Below we describe the joint Metropolis-Hastings update for $h, \pi, \boldsymbol{\gamma}$ in detail.

First, the proposal distributions for $h, \pi, \boldsymbol{\gamma}$ are the following.

Propose $\boldsymbol{\gamma}^*$ using piMASS rank-based proposal [Guan and Stephens, 2011]. The proposal adds, removes or switches a single γ_i component. The proposal to remove as well as switch (add and remove) a covariate is uniform. The proposal to add a covariate is a mixture of uniform and rank-based selection. Rank-based selection chooses among the covariates with largest single Bayes factors (3.11):

$$\frac{p(\boldsymbol{Z}|h, \boldsymbol{\gamma} = \boldsymbol{\gamma}_i)}{p(\boldsymbol{Z}|h, \boldsymbol{\gamma} = \mathbf{0})} = n^{1/2}|\boldsymbol{\Omega}|^{1/2}\frac{1}{\sigma_a(h, \boldsymbol{\gamma})^{|\boldsymbol{\gamma}|}}\Big(\frac{\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{Z}'\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\Omega}\boldsymbol{X}'_{\boldsymbol{\gamma}}\boldsymbol{Z}}{\boldsymbol{Z}'\boldsymbol{Z} - n\bar{\boldsymbol{Z}}^2}\Big)^{-n/2}, \tag{3.11}$$

where $|\boldsymbol{\gamma}| := \sum_j \gamma_j$, $\boldsymbol{\gamma}_i$ is a vector with all its components equal to zero, except the i-th component, $\boldsymbol{\Omega}$ is defined in (1.34) and σ_a is defined in (1.30).

Here, we choose the proposal to add a covariate to be 30% rank-based and 70% uniform (vs. 70% rank-based and 30% uniform in the original piMASS). Since we expect confounding variables to have an effect that is not captured by genetic covariates, we prefer to rely less on the marginal associations and allow more random exploration of the covariates when adding a covariate. Such setting is desirable in a situation with confounding variables. More details on the proposal distribution for $\boldsymbol{\gamma}$ can be found in [Guan and Stephens, 2011].

Proposal distribution for π is beta distribution, which uses the proposed $\boldsymbol{\gamma}^*$ obtained in the

previous step:

$$\pi^* \sim Beta\big(|\boldsymbol{\gamma}^*|,\ p - |\boldsymbol{\gamma}^*| + 1\big). \tag{3.12}$$

Proposal distribution for h is a random walk

$$h^* = h^{(s)} + U(-0.1, 0.1), \tag{3.13}$$

where $h^{(s)}$ is the current value of h at s-th iteration.

Let $\boldsymbol{\theta} = (h, \pi, \boldsymbol{\gamma})$, then $\boldsymbol{\theta}^{(s)}$ denotes the current value of $\boldsymbol{\theta}$ at the s the iteration and $\boldsymbol{\theta}^*$ denotes the proposed value. We will update $\boldsymbol{\theta}^{(s)}$ using Metropolis-Hastings algorithm.

The Metropolis-Hastings ratio is:

$$\begin{aligned} r_{MH} &= \frac{p(\boldsymbol{\theta}^*|\boldsymbol{Z})}{p(\boldsymbol{\theta}^{(s)}|\boldsymbol{Z})} \times \frac{J_s(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^*)}{J_s(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)})} \\ &= \frac{p(\boldsymbol{Z}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\boldsymbol{Z}|\boldsymbol{\theta}^{(s)})p(\boldsymbol{\theta}^{(s)})} \times \frac{J_s(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^*)}{J_s(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)})} \\ &= \frac{p(\boldsymbol{Z}|h^*, \boldsymbol{\gamma}^*)p(\boldsymbol{\gamma}^*|\pi^*)p(\pi^*)}{p(\boldsymbol{Z}|h^{(s)}, \boldsymbol{\gamma}^{(s)})p(\boldsymbol{\gamma}^{(s)}|\pi^{(s)})p(\pi^{(s)})} \times \frac{J_s(\boldsymbol{\theta}^{(s)}|\boldsymbol{\theta}^*)}{J_s(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)})}, \end{aligned} \tag{3.14}$$

where $p(\boldsymbol{\theta}|\boldsymbol{Z})$ is the target probability distribution given in (3.10) and J_s is a joint proposal distribution for $\boldsymbol{\theta}^{(s)}$. Sample r from $U(0,1)$ and set $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^*$ if $r < r_{MH}$, otherwise let $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)}$. This finishes the joint update of $h, \pi, \boldsymbol{\gamma}$ using Metropolis-Hastings algorithm.

Parameters $\boldsymbol{\beta}, \tau$

Since posterior conditional distributions are known explicitly, we use Gibbs sampling to sample $\boldsymbol{\beta}$ and τ. For each value of $h, \boldsymbol{\gamma}$ sampled from their posterior, we obtain samples from the posterior distribution of $\boldsymbol{\beta}$ and τ by sampling from their conditional distributions given $\boldsymbol{Z}, \boldsymbol{\gamma}, h$, which are available analytically.

Parameter τ has gamma posterior conditional distribution given $\boldsymbol{Z}, \boldsymbol{\gamma}, h$:

$$\tau | \boldsymbol{Z}, h, \boldsymbol{\gamma} \sim \Gamma(n/2, 2/(\boldsymbol{Z}'\boldsymbol{Z} - \boldsymbol{Z}'\boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\Omega}\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{Z})). \tag{3.15}$$

Parameters $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ have multivariate Gaussian posterior conditional distribution given $\boldsymbol{Z}, h, \boldsymbol{\gamma}, \tau$:

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{Z}, h, \boldsymbol{\gamma}, \tau \sim N(\boldsymbol{\Omega}\boldsymbol{X}_{\boldsymbol{\gamma}}'\boldsymbol{Z}, (1/\tau)\boldsymbol{\Omega}). \tag{3.16}$$

Parameters $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ are sampled from

$$\boldsymbol{\beta}_{-\boldsymbol{\gamma}} | \boldsymbol{Z}, h, \boldsymbol{\gamma}, \tau \sim \delta_0, \tag{3.17}$$

where δ_0 denotes a point mass on 0.

Parameters $\boldsymbol{\alpha}$

Now we introduce the variable $\boldsymbol{Q}$, which contains the variation left after accounting for current estimated effects of genetic covariates:

$$\boldsymbol{Q} = \boldsymbol{y} - \boldsymbol{G}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}. \tag{3.18}$$

Here $\boldsymbol{Q}$ contains the variation due to non-genetic covariates in the phenotype. Its distribution is

$$\boldsymbol{Q} \sim N_n(\boldsymbol{X}\boldsymbol{\alpha},\ \ \tau^{-1}\boldsymbol{I_n}), \tag{3.19}$$

and its likelihood is

$$p(\boldsymbol{Q}|\boldsymbol{X},\boldsymbol{\alpha},\tau) \propto \exp\Big(-\frac{\tau}{2}[\boldsymbol{Q}'\boldsymbol{Q} - 2\boldsymbol{\alpha}'\boldsymbol{X}'\boldsymbol{Q} + \boldsymbol{\alpha}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\alpha}]\Big). \tag{3.20}$$

Using $\boldsymbol{Q}$, we sample the coefficients of non-genetic covariates $\boldsymbol{\alpha}$ from their posterior distribution. Under the prior (3.4) and likelihood (3.20) the posterior conditional distribution of $\boldsymbol{\alpha}$ given $\boldsymbol{y}, \boldsymbol{X}, \tau$ is multivariate normal

$$\boldsymbol{\alpha}|\ \boldsymbol{y},\boldsymbol{X},\tau \sim N_m\left(\frac{n}{n+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Q},\ \ \frac{n}{n+1}\frac{1}{\tau}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right), \tag{3.21}$$

where the value of τ is obtained from (3.15).

3.3 Simulations

We perform simulation studies to investigate how well our method identifies associations with causal variants and compare the results to the model piMASS. We conducted simulations on two types of genetic data: computer simulated and real genotypes, and we use Receiver Operating Characteristic (ROC) curves to assess the performance of the proposed method.

3.3.1 Simulated data

To test the model described above we simulate data consisting of genetic variants, non-genetic covariates and phenotype. Below is the description of the simulation of each necessary part.

1. Genetic variants. We conducted simulation studies using two types of genetic variants: independent genetic variants simulated using package COSI2 [Shlyakhter et al., 2014] and a subset of real genotype data from the MGS data set described in [Shi et al., 2009].

Using package COSI2 we simulated 1000 uncorrelated SNPs. The package produces simulated data that closely resemble empirical data in allele frequency, linkage disequilibrium, and population differentiation [Schaffner et al., 2005]. We chose to simulate SNPs of European ancestry. The simulated SNPs comprise matrix $\boldsymbol{G}$ with dimensions $n \times p$ with $n = 500$ subjects and $p = 1,000$ SNPs. For real genotypes, we select a subset of $1,000$ SNPs and 500 subjects from a larger MGS data set containing $5,334$ subjects and $638,937$ SNPs. Both data sets are of the same size in terms of number of subjects and SNPs.

2. Non-genetic covariates. Our goal is to consider the situation where there is a correlation between genetic variants and additional covariates. First, we randomly selected 5 genetic variants $\boldsymbol{g_k}, k = 1, \dots, 5$ from $1,000$ SNPs. Each new covariate is generated from $\boldsymbol{g_k}$ in the following way:

$$\boldsymbol{X_k} = \boldsymbol{g_k} + \boldsymbol{\epsilon_k}, \tag{3.22}$$

where $\boldsymbol{\epsilon_k} = \epsilon_{1k}, \dots, \epsilon_{nk}$ and each $\epsilon_{ik} \sim U\left(-\frac{3}{5}d,\ \frac{3}{5}d\right)$, where d is the smallest difference between adjacent unique g_{ij} values [Chambers, 2018]. Next, we standardize $\boldsymbol{X_k}$ to set $\boldsymbol{\mu} = \boldsymbol{0}$ in (3.2).

Non-genetic covariates comprise matrix $\boldsymbol{X}$ of dimensions $n \times q$, $q = 5$. Each covariate $\boldsymbol{X_i}$ is correlated with the corresponding genetic variant. Average correlation among 5 pairs of genetic and derived non-genetic covariate is 0.75. Situation, where non-genetic covariates, such as Principal Components (PC), are correlated with certain SNPs is desirable, since PC-correlated SNPs can be used to successfully predict structure and ancestry proportions [Paschou et al., 2007].

3. Phenotype. We simulate phenotype for subject i using the following normal model:

$$y_i = \sum_{j \in C} g_{ij}\beta_j + \sum_{k=1}^{5} x_{ik}\alpha_k + \epsilon_i, \tag{3.23}$$

where C is a set of 3 causal variants randomly selected from the available $1,000$ genetic variants, effects of genetic variants are $\beta_j \sim N(0,1)$, effects of non-genetic covariates are $\alpha_k \sim N\big(0, n\big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\big)$ and error term $\epsilon_i \sim N(0,1)$.

We generate two complete datasets. First dataset consists of independent genetic variants, non-genetic covariates derived from it and 100 phenotypes generated using the combination of genetic and non-genetic data. Second dataset consists of real genetic variants, non-genetic covariates and 100 phenotypes generated using (3.23).

3.3.2 Simulation results for independent COSI genotype

We applied both piMASS and the proposed method to the simulated data set with independent genotypes with $50,000$ MCMC iterations. To lessen auto-correlation among the sampled values, we sampled values every 10 iteration, thinning the MCMC output to a total of $5,000$ samples. We did not detect any lack of convergence by looking at trace plots and Gelman-Rubin (GR) convergence diagnostics. Both methods produce the posterior inclusion probability (PIP) for each SNP. The PIP is given by:

$$\text{PIP}_i = \frac{\text{number of iterations with } i\text{-th SNP included in the model}}{\text{total number of iterations}}, \quad i = 1, \dots, n. \tag{3.24}$$

Overall, larger PIP signals stronger observed association with the phenotype under the proposed model. We use PIPs and ROC curves to evaluate the performance of the two methods. The proposed model can be viewed as a classification model with continuous output (PIP), which estimates class membership probability (associated or not associated with phenotype) [Fawcett, 2006]. To produce an ROC curve, different thresholds are applied to predict class membership. For any given threshold, we consider all causal SNPs which PIPs exceed the threshold to be true positives and all other SNPs with PIPs exceeding the threshold to be false positives. Each point on the graph represents a threshold with the x-coordinate being a false positive rate (FPR) or 1-specificity, which can be calculated as:

$$\text{FPR} = \frac{\text{number of false positives}}{\text{total number of negative cases in the data}} \tag{3.25}$$

and the y-coordinate representing true positive rate (TPR) or sensitivity:

$$\text{TPR} = \frac{\text{number of true positives}}{\text{total number of positive cases in the data}}. \tag{3.26}$$

Figure 3.1 contains ROC curves of the piMASS and proposed model.

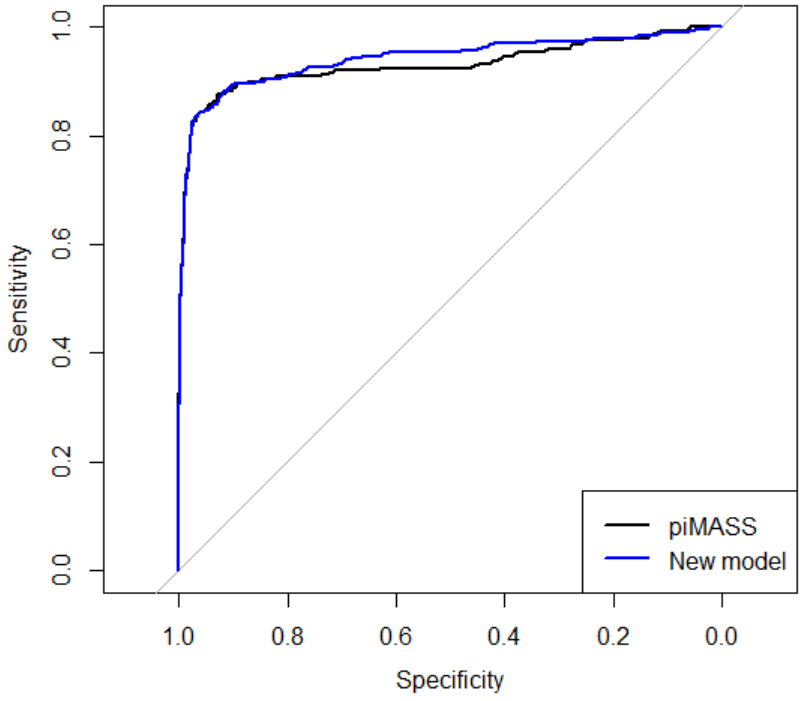

Figure 3.1: ROC curves for independent genetic variants for the continuous phenotype method

The new model's ROC curve is in blue and it either coincides or is visibly above the ROC curve generated by piMASS. To quantify this visual difference we use Area Under the Curve

(AUC) measure. Larger AUC means better prediction ability of a classifier. By definition, random classification of case and control subjects provides an AUC of 50% while perfect classification provides an AUC of 100%. AUCs for the two methods are presented in the Table 3.1.

Method	AUC
piMASS	0.9300449
new Method	0.9417234

Table 3.1: AUC measurements for independent genetic variants for the continuous phenotype method.

The AUCs for the two methods are different, but pretty close together, so it is not clear whether the difference in AUCs is random. To that end, we perform a significance test for two ROC curves provided in R package pROC [Robin et al., 2011], DeLong's test for two correlated ROC curves [DeLong et al., 1988]. It reports p-value$_{AUC} = 0.01477$ under the alternative hypothesis: true difference in AUC is not equal to 0. Since p-value is less than standard significance threshold of 5%, we can conclude that it is unlikely that the observed difference has occurred by chance.

Since AUC summarizes the entire ROC curve, a potential drawback of this metric comes from including regions with low levels of specificity that frequently are not practically relevant [Ma et al., 2013]. Thus, we turn our attention to Partial Area Under the Curve (PAUC) measurement presented in Table 3.2, which measures AUC for chosen acceptable levels of

FPR. Here we restrict the AUC measurement to FPR between 0 and 0.2. This measurement is of more interest since the performance of the method is particularly important in the area where the FPR is low.

Method	PAUC
piMASS	0.1338045
new method	0.1454622

Table 3.2: PAUC measurements for independent genetic variants for the continuous phenotype method.

To see if the observed difference in PAUCs is random, we conduct the bootstrap test for two correlated ROC curves. We obtain p-value$_{PAUC} = 0.01249$, yielding similar conclusions as AUC metric that the observed difference is unlikely to have occurred by chance.

3.3.3 Simulation results for real genotypes

Now we present the simulation results that assess the performance of our method using real genotype data. We ran both piMASS and proposed method for $50,000$ iterations. We sampled parameter values every 10 iteration, thinning the MCMC output to a total of $5,000$ samples. We did not detect any lack of convergence while running standard diagnostic procedures. We summarize the results below.

The ROC curve for the two methods are shown in Figure 3.2.

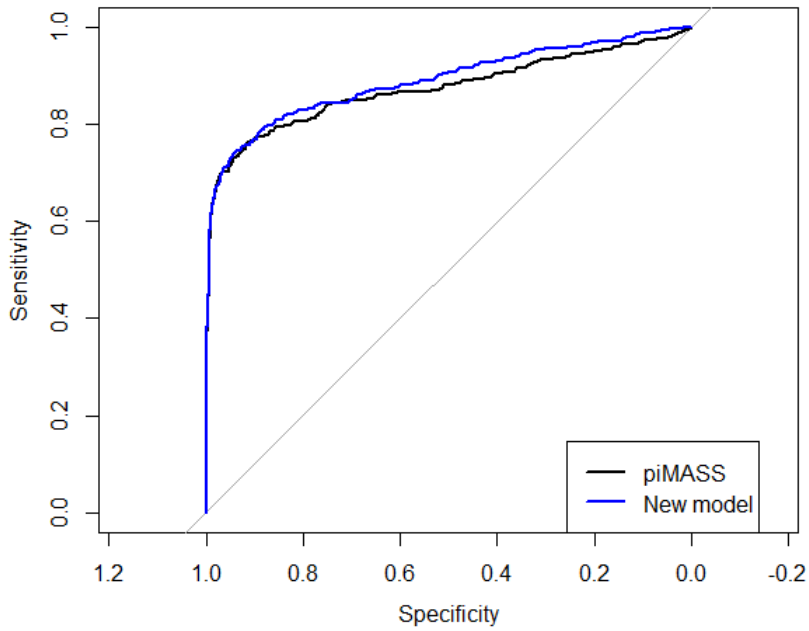

Figure 3.2: ROC curves for real genetic variants for the continuous phenotype method

The new model's ROC curve is in blue and is visibly above the black curve for the benchmark model piMASS, however, it is not clear by how much. To quantify this visual difference, we calculate AUCs for each model, which are presented in the Table 3.3 below.

Method	AUC
piMASS	0.8724
new Method	0.8887

Table 3.3: AUC measurements for real genetic variants for the continuous phenotype method.

To test the significance of the observed difference in AUCs, we perform a DeLong's significance test for two ROC curves provided in pROC package R, which yields $p\text{-value}_{AUC} =$

0.002198. This leads to the conclusion that the observed difference in AUCs is unlikely to have occurred by chance.

To focus on the practically relevant FPR between 0 and 0.2, we calculate PAUCs for each method, presented in the Table 3.4:

Method	PAUC
piMASS	0.0848
new Method	0.1

Table 3.4: PAUC measurements for real genetic variants for the continuous phenotype method.

The bootstrap test for two correlated partial ROC curves with 2000 bootstrap iterations yields $p\text{-value}_{PAUC} = 0.002575$. Given the obtained p-value, we can conclude that the observed difference in the performance of two models is unlikely to have occurred by chance.

Both AUC and PAUC results for both simulated and real genetic variants show that the difference between the proposed method and piMASS is unlikely to be random. We also notice that the proposed method showed higher AUC for independent genetic variants 0.9417 vs 0.8887 for real genetic variants (Tables 3.1, 3.3). The same holds true for PAUC for independents variants of 0.1454622 vs 0.1 for real genetic variants (Tables 3.2, 3.4). These differences can be attributed to the fact that simulated genetic variants are independent and uncorrelated, whereas the real genetic variants have natural correlation structure that might be interfering with the new method's ability to detect true associations.

3.4 Conclusion

Herein, we have developed Bayesian multivariate linear regression method that measures associations between genetic variants and phenotypes, while taking into account non-genetic covariates that account for population stratification. We carried out the simulation studies using both real and simulated genetic variants and compared the performance of the proposed method to the piMASS.

Simulation results provided evidence that there is a nonrandom difference between the two classifiers, with the new method performing better in terms of both AUC and PAUC for data sets with both correlated and uncorrelated genetic variants. The observed difference in the performance of the proposed method between the two data sets hints at a possible role of the correlation structure of the genetic variants. Thus, taking into account the correlation structure of the genetic variants is one possible direction for further research.

CHAPTER 4

BAYESIAN VARIABLE SELECTION METHOD WITH POPULATION STRATIFICATION CORRECTION FOR GWAS WITH BINARY PHENOTYPE

4.1 Background

In GWAS, many complex disease phenotypes are often recorded as a binary variable. In such studies, the goal is to compare the frequency distribution of genotypes between cases and controls [Wu et al., 2011]. A difference in the allele frequency of an SNP between cases and controls points at possible causal role of the SNP on the phenotype. However, presence of undetected population stratification in large scale genetic studies can lead to both false positives and failure to detect genuine associations [Marchini et al., 2004].

One way to account for population stratification is to use logistic mixed models (LMMs) and include the variables accounting for spurious associations. Methods like PC-select [Tucker et al., 2014] use LMM to account for population stratification in GWAS. EIGENSTRAT is another computationally efficient method that is widely used in single-SNP testing setting [Price et al., 2006] . However, neither of the methods mentioned above is directly

applicable to BVS.

Therefore, there is a need for a method that allows for population stratification correction in Bayesian setting, since BVS is the method that can achieve the maximum power among other GWAS multiple regression methods [Banerjee et al., 2018]. One way to account for additional covariates in Bayesian setting is by modelling binary response using the idea of data augmentation [Albert and Chib, 1993]. A population stratification correction method that can work with BVS in case-control settings has the potential to yield results superior to the existing methods. We introduce such a method in the sections below.

The rest of the Chapter is organized as follows. Section 2 provides the theoretical framework for including additional covariates in piMASS when the trait of interest is binary. Section 3 contains simulation studies using an independent set of genotype data generated using COSI2 and a subset of the MGS genotype data. The studies compare the performance of the proposed method to piMASS. Section 4 contains real data analysis using the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI1) dataset.

4.2 Methods

4.2.1 Data augmentation approach

A standard data augmentation approach [Albert and Chib, 1993] is to create n latent variables $Z_1, \ldots, Z_n$, where Z_i are independent

$$Z_i \sim N(\mu + \alpha_1 x_{i1} + \ldots + \alpha_q x_{iq} + \beta_1 g_{i1} + \ldots + \beta_p g_{ip},\ 1), \tag{4.1}$$

where τ is set to a fixed value of 1 to avoid the improper posterior on τ and $\boldsymbol{Z}$ [Albert and Chib, 1993]. Then $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ is assumed to follow a linear regression model:

$$\boldsymbol{Z} \mid \mu, \tau, \boldsymbol{\beta}, X, G \sim N_n(\boldsymbol{\mu} + X\boldsymbol{\alpha} + G_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, I_n), \tag{4.2}$$

which is the model for continuous outcome introduced in Chapter 3.

The relationship to the observed outcomes y_i is defined in the following way:

$$y_i = \begin{cases} 0, & \text{if } Z_i < 0, \\ 1, & \text{if } Z_i \geq 0. \end{cases}$$

Then, the observed variables y_i are independent Bernoulli random variables

$$y_i \sim Bernoulli(p_i), \tag{4.3}$$

where p_i is the probability of an outcome being a case.

Then the probit binary regression model is

$$p_i = P(y_i = 1 \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\mu}) = \Phi(\mu + \alpha_1 x_{i1} + \ldots + \alpha_q x_{iq} + \beta_1 g_{i1} + \ldots + \beta_p g_{ip}), \tag{4.4}$$

where Φ is the Gaussian cumulative distribution function connecting the probabilities p_i to the explanatory variables.

Z_i are unknown, however, given the observed y_i, Z_i follows a truncated normal distribution [Albert and Chib, 1993]

$$Z_i|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma} \sim N(\boldsymbol{\mu} + X\boldsymbol{\alpha} + G_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}},\ 1) \text{ truncated at the left by 0 if } y_i = 1 \tag{4.5}$$

$$Z_i|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma} \sim N(\boldsymbol{\mu} + X\boldsymbol{\alpha} + G_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}},\ 1) \text{ truncated at the right by 0 if } y_i = 0 \tag{4.6}$$

This approach connects the probit binary regression model on the y_i with a normal linear regression model on the latent data Z_i. Posterior inference requires one additional update of the $\boldsymbol{Z}$ variables compared to the quantitative trait.

4.2.2 MCMC scheme for binary phenotypes

MCMC setting for binary phenotype preserves most of the parameter updates described in Chapter 3.

The approach described above allows usage of the same priors as for continuous data, only these priors now relate to the unobserved latent continuous variables and not to the observed binary outcomes.

A few practical constraints were suggested by Guan & Stephens [Guan and Stephens, 2011] to ensure identifiability and improve mixing. First, an identifiability constraint was im-

posed on $\boldsymbol{Z}$ to have empirical variance 1. Second, we center all variables and set $\boldsymbol{\mu} = 0$. Third, to improve mixing, we make an assumption that marginal distribution of $\boldsymbol{Z}$ is normal [Guan and Stephens, 2011]. In particular $Z_i, ..., Z_n|\boldsymbol{y}$ are restricted to take fixed set of values, which are n equally spaced quantiles of standard normal distribution with the values corresponding to n_0 individuals with $y_i = 0$ being constrained to the first n_0 of those quantiles. This is a reasonable assumption if we there are no large values of $\boldsymbol{\beta}$.

Local Metropolis-Hastings proposals for $\boldsymbol{Z}$ involve randomly picking a pair of individuals (m, n) with the same binary phenotype and propose to swap values Z_m and Z_n. The Metropolis-Hastings ratio is calculated using (3.14) while using $\boldsymbol{Z}$ as an outcome variable and the decision to accept or reject the proposed $\boldsymbol{Z}$ as well as other proposed parameters is based on it.

4.3 Simulations

We conducted simulation studies to see how well our method identifies associations with causal variants when the phenotype is binary and compare the results to piMASS. We performed simulations on two types of genetic data: computer simulated and real genotypes, and we use ROC curves to assess the performance of the methods.

4.3.1 Simulated data

To test the model, we generate 3 necessary components: genetic variants, non-genetic covariates and the phenotype.

1. Genetic variants. We conducted simulation using two types of genetic variants: independent genetic variants simulated using the COSI2 package [Shlyakhter et al., 2014], and a subset of real genotype data from the MGS data set described in [Shi et al., 2009]. The independent genetic variants we used are the same as in Chapter 3 and comprise genotype matrix $\boldsymbol{G}$ with dimensions $n \times p$ with $n = 500$ subjects and $p = 1,000$ SNPs. The real genetic variants are a subset of $1,000$ subjects and $9,306$ SNPs of chromosome 21 from a larger MGS dataset containing $5,334$ subjects and $638,937$ SNPs.

2. Non-genetic covariates. As in Chapter 3, our goal is to consider the situation where there is a correlation between genetic variants and additional covariates. First, we randomly selected 5 genetic variants $\boldsymbol{g_k}$, $k = 1, \dots, 5$ from $9,306$ SNPs. Each new covariate is generated from $\boldsymbol{g_k}$ in the following way:

$$\boldsymbol{X_k} = \boldsymbol{g_k} + \boldsymbol{\epsilon_k}, \tag{4.7}$$

where $\boldsymbol{\epsilon_k} = (\epsilon_{1k}, \dots, \epsilon_{nk})$ and each $\epsilon_{ik} \sim U\big(-\frac{3}{5}d,\ \frac{3}{5}d\big)$, where d is the smallest difference between adjacent unique g_{ij} values [Chambers, 2018]. Next, we standardize $\boldsymbol{X_k}$. Non-genetic covariates comprise matrix $\boldsymbol{X}$ of dimensions $n \times q$, $q = 5$. Each covariate $\boldsymbol{X_i}$ is correlated with the corresponding genetic variant. Average correlation among 5 pairs of genetic and derived non-genetic covariate is 0.79. Top eigenvectors or principal components (PCs) of a

genetic relatedness matrix are often used as additional covariates in population stratification correction [Price et al., 2006]. Our simulated high correlation between SNPs and additional covariates is desirable since in some situations, SNPs, correlated with the PCs of the genetic relatedness matrix, can be used to successfully account for population structure and ancestry proportions [Paschou et al., 2007].

3. Phenotype. We simulate the latent phenotype for subject i using the following model:

$$Z_i = \sum_{j \in C} g_{ij}\beta_j + \sum_{k=1}^{5} x_{ik}\alpha_k + \epsilon_i, \tag{4.8}$$

where for COSI genotypes C is a set of 3 variants randomly selected from the available $1,000$ genetic variants, and for real genotypes C is a set of 10 variants randomly selected from $9,306$ genetic variants. The effects of genetic variants are $\beta_j \sim N(0,1)$, the effects of non-genetic covariates are $\alpha_k \sim N\big(0, n\big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\big)$, and the error term $\epsilon_i \sim N(0,1)$. Next, we convert these n quantitative phenotypes to n binary phenotypes by mapping the smallest $n/2$ values to $y = 0$ and the rest to $y = 1$. We generate two complete datasets. Both datasets consists of genetic variants, non-genetic covariates and 100 phenotypes generated using (4.8).

4.3.2 Simulation results for independent COSI genotypes

We applied both piMASS and the proposed method to the simulated data set with independent genotypes. For the results provided below, we ran $50,000$ MCMC iterations.

To reduce auto-correlation among the sampled values, we sampled parameter values every 10 iterations, leaving a total of $5,000$ samples. We did not detect lack of convergence by looking at trace plots and Gelman-Rubin (GR) convergence diagnostics. Both methods produced the PIP for each SNP. We use ROC curves to evaluate the performance of the two methods. Figure 4.1 contains ROC curves of the piMASS and proposed model.

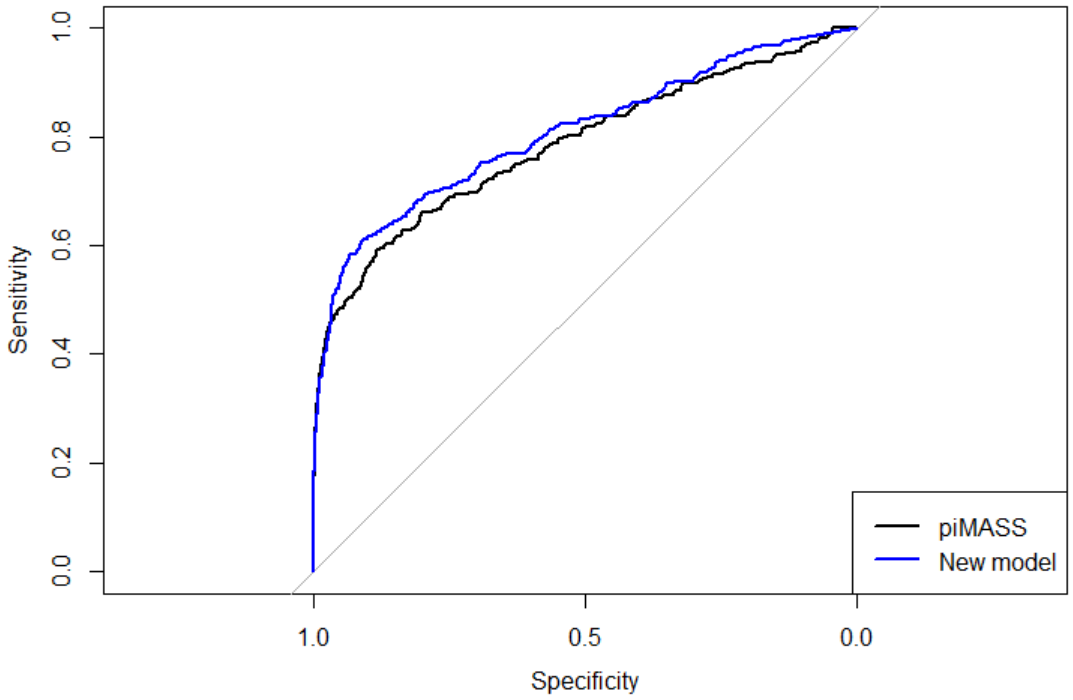

Figure 4.1: ROC curves for independent genetic variants for the binary phenotype method

The new model's ROC curve is in blue and it is visibly above the ROC curve generated by piMASS. To quantify this visual difference we use AUC measure. AUCs for the two methods are presented in the Table 4.1.

Method	AUC
piMASS	0.7842681
new Method	0.8068356

Table 4.1: AUC measurements for independent genetic variants for the binary phenotype method.

The AUCs for the two methods are different, however we still conduct the DeLong's test for two correlated ROC curves to see if the difference in AUCs is random. The test reports p-value$_{AUC} = 0.09434$ under the alternative hypothesis: true difference in AUC is not equal to 0. Although the p-value is larger than standard significance threshold of 5%, it is still unlikely that the observed difference has occurred by chance. It is possible that two methods have similar power to detect associations on the provided dataset.

To investigate the practically relevant areas of specificity, we provide PAUC measurement presented in Table 4.2. PAUC measures AUC for chosen acceptable levels of FPR. Here we restrict the AUC measurement to FPR between 0 and 0.2.

Method	PAUC
piMASS	0.05676018
new Method	0.06501258

Table 4.2: PAUC measurements for independent genetic variants for the binary phenotype method.

To see if the observed difference in PAUCs is random, we conduct a bootstrap test for two correlated ROC curves. We obtain p-value$_{PAUC} = 0.2718$, which means it is likely that the

performance of the two methods is similar in the selected range of specificity for a given dataset.

4.3.3 Simulation results for real genotypes

Now we present the simulation results that assess the performance of our method for binary phenotypes using real genotype data of the size $n \times p$, $n = 1,000$ individuals, $p = 9,306$ SNPs, which is a subset of the MGS dataset described in [Shi et al., 2009] and is availablein the NIMH Genetic Repository and Resource (`https://www.nimhgenetics.org/`) upon approval of NIMH.

We ran both piMASS and proposed method for $80,000$ iterations. To reduce auto-correlation in the samples, we sampled parameter values every 10 iterations leaving a total of $8,000$ samples. We did not see any evidence for the lack of convergence while running standard diagnostic procedures such as trace plots. We run both piMASS and the new method on the real genotype data with 100 binary phenotypes each and summarized the results below.

Figure 4.2 contains ROC curves of the piMASS and proposed model for real genetic data with binary phenotypes.

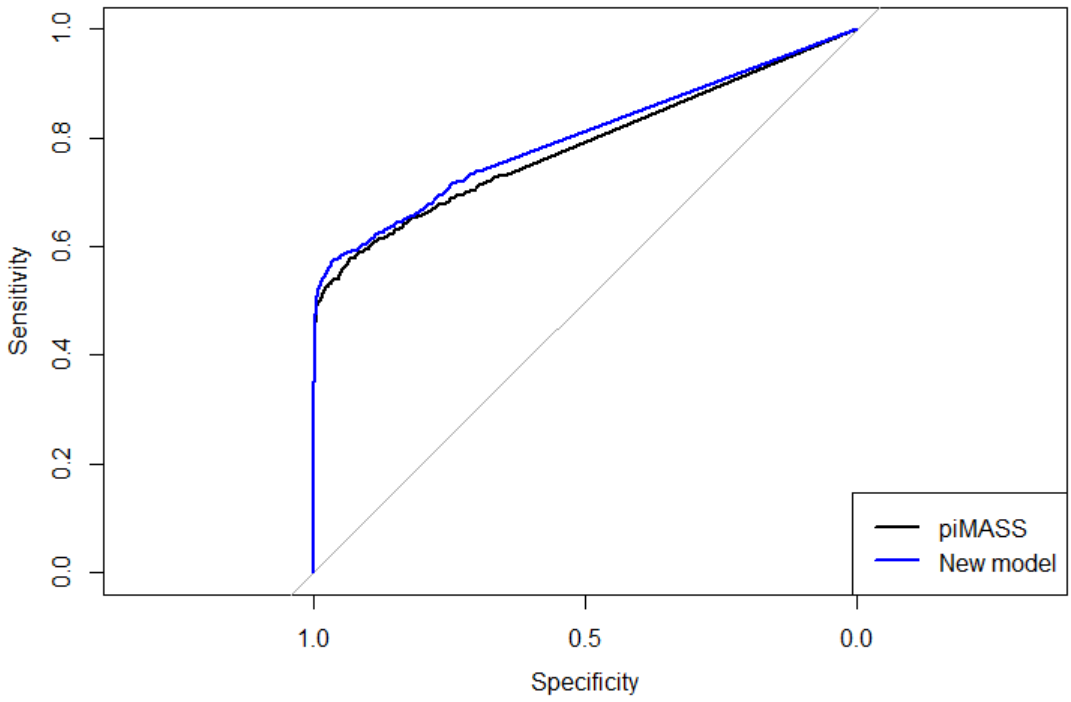

Figure 4.2: ROC curves for real genetic variants for the binary phenotype method

The new model's ROC curve is in blue and is visibly above the black curve for the benchmark model piMASS for most values of specificity, however, it is not clear by how much. To quantify this visual difference, we calculate AUCs for each model, which are presented in the Table 4.3 below.

Method	AUC
piMASS	0.7845306
new Method	0.7998063

Table 4.3: AUC measurements for real genetic variants for the binary phenotype method.

To test the significance of the observed difference in AUCs, we perform a DeLong's significance test for two ROC curves provided in R package pROC , which yields p-value$_{AUC}$ = 0.07877. As with the previous datasets, this means that it is still unlikely that the observed difference in AUCs have occurred by chance.

Next we focus on the practically relevant FPR between 0 and 0.2 and calculate PAUCs for each method, presented in the Table 4.4:

Method	PAUC
piMASS	0.04823724
new Method	0.05347307

Table 4.4: PAUC measurements for real genetic variants for the binary phenotype method.

The bootstrap test for two correlated partial ROC curves with 2000 bootstrap iterations yields p-value$_{PAUC}$ = 0.07692. We can say that the observed difference in the performance of two models is unlikely to have occurred by chance. It is possible that the models perform similarly for a given dataset.

Since p-values do not measure the probability that the studied hypothesis is true [Wasserstein and Lazar, 2016], we cannot conclude that the two methods are the same. In this case the usefulness of the population stratification correction should be evaluated on a case by case basis.

4.4 Real data analysis

Alzheimer disease (AD) is characterized by progressive cognitive decline usually with short term memory impairment and inevitably affecting all intellectual functions, leading to complete dependence and premature death [Mayeux and Stern, 2012]. In 2019, 5.8 million Americans were living with Alzheimer's dementia. Out of them 5.6 million age 65 or older, which means 1 in 10 people age 65 or older has AD [Hebert et al., 2013]. Genetic factors play a major role in determining a person's risk to develop AD [Bertram and Tanzi, 2012]. AD heritability varies from 58-79% depending on age at onset; however, only a portion of the likely substantial genetic contribution to this disease has been determined [Reitz et al., 2011].

We use dataset ADNI1 provided by Alzheimer's Disease Neuroimaging Initiative (ADNI) (`http://adni.loni.usc.edu`). ADNI seeks to develop biomarkers of the disease and advance the understanding of AD pathophysiology, improve diagnostic methods for early detection of AD and improve clinical trial design. Additional goals are examining the rate of progress for both mild cognitive impairment and Alzheimer's disease, as well as building a large repository of clinical and imaging data.

ADNI1 GWAS data set contains $620,901$ SNP and copy number variant (CNV) markers for 757 subjects (449 males, 308 females) genotyped using Illumina Human610-Quad BeadChip platform with 3 types of Patient Diagnosis groups: Cognitively Normal (CN), Mild Cognitive Impairment (MCI) and Alzheimer's Disease.

Since we are interested in applying the new method to the data set with the binary pheno-

types, we combined MCI and AD diagnosis patients to be cases in our analysis. We did it for the following reasons. At present time, there is increasing recognition and understanding of the MCI entity as a stage which is a frequent precursor and harbinger of subsequently manifest AD [Reisberg and Gauthier, 2008]. MCI, as a stage in the evolution of subsequently manifest AD, has been estimated to have a duration of approximately seven years [Reisberg, 1986]. Since the SNP data of the person diagnosed with MCI will not change in seven years, we can use existing genetic data for MCI patients to enhance our AD analysis. Neurotypical Controls serve as controls in our analysis.

Some methods apply the principal components analysis to genotype data to infer continuous axes of variation, which reduce the data to a small number of dimensions, describing as much variability as possible [Price et al., 2006]. Such axes of variation, also sometimes referred to as principal components are often used as additional covariates in population stratification correction [Price et al., 2006]. We selected top 5 eigenvectors of a genotype covariance matrix to use as additional covariates. The PCs were obtained from ADNI1 genotype data using package PLINK available at `http://pngu.mgh.harvard.edu/purcell/plink/` [Purcell et al., 2007].

We ran the new method described above for $100,000$ MCMC iterations. We sampled every 10 iterations to reduce auto-correlation, which yielded $10,000$ samples. The standard diagnostic procedures did not indicate any lack of convergence.

The Manhattan plot of the PIP of individual SNP using the ADNI1 dataset is shown in Figure 4.3 using $\log_{10}(1 - \text{PIP})$ as the y-axis.

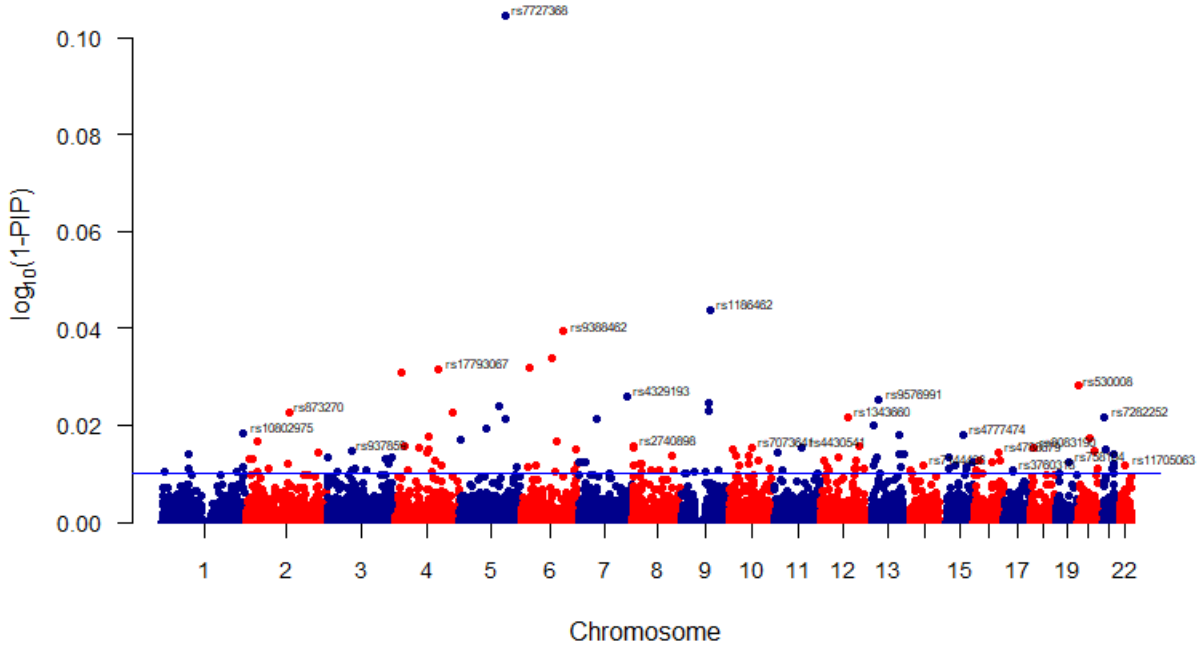

Figure 4.3: Manhattan plot of 1-PIP for ADNI1 data set for the binary phenotype method

Below we present a table containing top 30 SNPs that had the best association metric (PIP) in our analysis.

Chr	rsID	Position	Rank	PIP
5	rs7727368	133028407	1	0.21388
9	rs1186462	82876046	2	0.09573
6	rs9388462	126318495	3	0.08653
6	rs9345231	93201586	4	0.07474
6	rs3846829	24411174	5	0.07091
4	rs17793067	126391075	6	0.06987
4	rs6449272	16583116	7	0.06847
20	rs530008	862420	8	0.06263
7	rs4329193	141589735	9	0.05821
13	rs9576991	39806690	10	0.05651
9	rs2807302	81387555	11	0.05508
5	rs7724537	115331803	12	0.05369
9	rs7870046	77803715	13	0.05157
4	rs11132004	169215891	14	0.05101
2	rs873270	127575888	15	0.05089
12	rs1343660	80215275	16	0.04869
21	rs7282252	18188480	17	0.04834
5	rs2867327	133029144	18	0.04821
7	rs923823	53034446	19	0.04779
13	rs17079354	23269225	20	0.0451
5	rs10061143	80630102	21	0.04339
1	rs10802975	239880394	22	0.04168
13	rs9513661	99232340	23	0.04088
15	rs4777474	70005090	24	0.04055
4	rs7682461	98516722	25	0.03965
20	rs1739646	36470998	26	0.03937
5	rs828311	3619024	27	0.0384
2	rs2699150	36315894	28	0.03741
6	rs12203747	104803799	29	0.03736
8	rs2740898	3891496	30	0.03576

Table 4.5: Top 30 SNPs with largest PIP for the binary phenotype method

We used GWAS catalog available at `https://www.ebi.ac.uk/gwas` to match SNPs with high PIPs to the genes associated with Alzheimer's disease. The top signal is rs7727368, an intergenic SNP located at chr5:133000508 which has not been detected in previous studies. However, rs7870046 (rank 13) is located at chr9:78613895 and is a part of the protein coding gene PCSK5 which plays role in Alzheimer's disease progression score [Scelsi et al., 2018]. Additionally, rs2740898 (rank 30) is located at chr8:3904088 and is a part of protein coding gene CSMD1 which influences Alzheimer's disease cognitive decline [Sherva et al., 2014] and Alzheimer's disease with visuospatial domain impairment [Mukherjee et al., 2018]. Next, rs3846829 (rank 5) is located at chr6:24303195 is a part of the DCDC2 gene influencing information processing speed [Luciano et al., 2011] and intelligence [Davies et al., 2018].

rs2807302 (rank 11) is located at chr9:82197735 is a part of the TLE4 gene and is associated with bipolar disorder [Jiang and Zhang, 2011] and mathematical ability [Lee et al., 2003]. rs1343660 (rank 16) is located at chr12:81691144 and is a part of the protein coding gene PPFIA2 which is implicated in self reported educational attainment [Rietveld et al., 2014] and schizophrenia [Levinson et al., 2012]. rs17079354 (rank 20) is located at chr13:24371225 and is a part of the protein coding gene MIPEP which is implicated in smoking behavior [Park et al., 2015]. rs10802975 (rank 22) is located at chr1:241813771 is a part of the OPN3 gene influencing response to antidepressants [Fabbri et al., 2019] and gut microbiome measurement [Hughes et al., 2020]. rs9513661 (rank 23) is located at chr13:100434339 and is a part of the protein coding gene CLYBL which is implicated in smoking behavior [Wootton et al., 2019]. rs7682461 (rank 25) is located at chr4:98297699 is a part of pro-

tein coding gene STPG2 which is implicated in intelligence [Savage et al., 2018] and alcohol consumption measurement [Brazel et al., 2019].

There are several reasons why we did not find the signals discovered in large Alzheimer's GWAS [Jansen et al., 2019], [Kunkle et al., 2019], [Waring and Rosenberg, 2008], [Bertram and Tanzi, 2009] in the current analysis. First, the signal tends to be spread among correlated SNPs and any single SNP may not get a large value of PIP to stand out. Next, since here we are using unimputed data, we can only find associations with the markers that are present in the ADNI1 data set. Finally, in this particular application, we used relatively small number of iterations compared to the number of genetic variants in the data set. This means that there was less opportunity to explore the large amount of SNPs. Due to all the factors mentioned above, the results received are expected. For a more thorough analysis, additional efforts should be applied. However, we were able to demonstrate the ease of use and applicability of the proposed method.

4.5 Conclusion

In this chapter we provided a BVS method for GWAS with binary phenotypes which provides the measure of association between genotypes and phenotypes while taking into account additional covariates that can account for population stratification.

We constructed the method using the data augmentation approach [Albert and Chib, 1993]

and implemented a simplified version of the method by imposing additional identifiability constraints.

We carried out the simulation studies using both independent and real genetic variants. We compared the performance of the proposed method to the piMASS using ROC curves. Simulation results provided evidence that there is a difference between the two classifiers that is unlikely to have ocurred by chance, however it is possible that performance of the two methods for binary data with both correlated and uncorrelated genetic variants is similar.

Next we applied the proposed method to real genetic data with binary phenotypes from the ADNI project whose goal is to develop biomarkers of the Alzheimer's disease and advance the understanding of its pathophysiology. The analysis showed some SNPs having strong associations with the phenotype including several previously unknown markers. Overall, we demonstrated the applicability of the method to the real data.

CHAPTER 5

BAYESIAN VARIABLE SELECTION METHOD FOR GWAS WITH ORDINAL PHENOTYPE

5.1 Background

Many GWAS phenotypes are often measured on continuous or binary scale. The tools to treat binary and continuous data have been described in previous chapter. However, some phenotypes naturally take ordered, discrete values. Examples include (a) subtypes defined from multiple sources of clinical information and (b) derived phenotypes generated by specific phenotyping algorithms for electronic health records (EHR) [German et al., 2020]. For many complex diseases such as substance use disorders or progression of Alzheimer's disease, phenotypes are often measured on ordinal scale. In that case, GWAS frameworks developed for other data types are not directly applicable and require further adaptation.

Treating ordinal data as binary, continuous or multinomial leads to misleading inference, loss of power and inconsistent results [German et al., 2020]. Therefore, the development of methods that can utilize the information contained in ordinal phenotypes and still use the advantages of a Bayesian approach is of potential benefit.

Fitting ordinal regression using maximum likelihood is possible for single-SNP approaches, but not applicable to BVS. There are several ways to connect the ordinal outcome to BVS. One useful solution is the data augmentation framework developed for Bayesian modelling with ordinal response [Albert and Chib, 1993]. There have been several methods proposed in treating ordinal and categorical outcome for BVS [Sha et al., 2004], [Kwon et al., 2007]. However, those variable selection approaches were implemented for models not tailored to GWAS. Here, we provide an approach that extends piMASS to work with ordinal phenotype data. Below we provide the theoretical foundation for BVS with ordinal phenotype.

5.2 Extension to ordinal categorical phenotypes

Following our previous notation, we let $\boldsymbol{X}, \boldsymbol{y}$ indicate the observed data. Here $\boldsymbol{X} = (\boldsymbol{x_1},\ \dots\ , \boldsymbol{x_n})'$ is an $n \times p$ design matrix containing the set of p potential predictors, where $\boldsymbol{x_i} = (x_{i1},\ \dots\ , x_{ip})'$ is a column vector containing the observed covariates of the ith individual and $\boldsymbol{y} = (y_1, y_2, \dots\ , y_n)'$ is a response vector, where y_i takes one of J ordered categories, $1,\ \dots\ ,\ J$, where $J > 2$.

Each outcome y_i is associated with a vector $(p_{i1}, p_{i2},\ \dots\ ,\ p_{iJ})$, where $p_{ij} = P(y_i = j)$ is the probability that individual i falls into jth category. Now define the cumulative probabilities

$$\eta_{ij} = P(y_i \leq j) = \sum_{k=1}^{j} p_{ik},\ i = 1, ..., n,\ j = 1, ..., J-1 \text{ and } \eta_{iJ} = 1. \tag{5.1}$$

Following data augmentation approach [Albert and Chib, 1993], we introduce n latent variables $\boldsymbol{Z} = (Z_1, ..., Z_n)$ into the problem. The $\boldsymbol{Z}$ is assumed to be an underlying continuous variable that determines the value of $\boldsymbol{y}$. The correspondence between Z_i and y_i is:

$$y_i = j \quad \text{if} \quad s_{j-1} < Z_i \leq s_j, i = 1, ..., n, \ j = 1, ..., J, \tag{5.2}$$

where the boundaries $s_1, ..., s_{J-1}$ are unknown, and $-\infty = s_0 < s_1 < \ ... \ < s_{J-1} < s_J = \infty$. For the purposes of parameter identifiability, we set $s_1 = 0$.

Now we use standard normal linear regression model to describe the relationship between observed potential predictors $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$. The probabilities η_{ij} can be related to the linear predictor in the following way:

$$\eta_{ij} = \Phi(s_j - \mu - \boldsymbol{x_i'}\boldsymbol{\beta}), \ \ i = 1, \ ... \ , n, \ \ j = 1, \ ... \ , J - 1, \tag{5.3}$$

where Φ is the Gaussian continuous distribution function. This model is motivated by the assumption that there exists a latent continuous random variable Z_i distributed $N(\mu + \boldsymbol{x_i'}\boldsymbol{\beta}, 1)$, where we set $\tau = 1$ to avoid the improper posterior on $\tau, \boldsymbol{Z}$ [Albert and Chib, 1993]. Now the problem is reframed as a normal regression problem where the response is in the form of the grouped data. In the model above, the vector of regression coefficients $\boldsymbol{\beta}$ and the bin boundaries $s_2, ... \ , s_{J-1}$ are unknown.

Incorporating variable selection into the model

In a GWAS context, most predictors provide no information about the phenotype. To identify relevant predictors, we introduce a binary inclusion/exclusion vector $\boldsymbol{\gamma}$ that induces mixture prior on regression coefficients. Next, we specify the priors on model parameters:

$$\tau \sim Gamma(\lambda/2, k/2) \tag{5.4}$$

$$\mu|\tau \sim N(0, \sigma_\mu^2/\tau) \tag{5.5}$$

$$\gamma_j \sim Bernoulli(\pi) \tag{5.6}$$

$$\boldsymbol{\beta_\gamma}|\tau, \boldsymbol{\gamma} \sim N_{d_\gamma}(0\ , (\sigma_a^2/\tau)\ \boldsymbol{I}_{d_\gamma}) \tag{5.7}$$

$$\boldsymbol{\beta_{-\gamma}}|\boldsymbol{\gamma} \sim \delta_0, \tag{5.8}$$

where δ_0 is a point mass on 0 and $\lambda, k, \sigma_\mu, \pi$ and σ_a are hyperparameters. Hyperparameters π, which reflects the model sparsity and σ_a, which reflects the typical size of a nonzero regression coefficients are playing important roles in tailoring model to GWAS. The priors on hyperparameters are also the same as in piMASS and are listed in equations (1.24) - (1.30). Similar to [Guan and Stephens, 2011] we set $\mu = 0$ and $\tau = 1$. We also denote $\boldsymbol{x_{i\gamma}}$ as a vector whose entries correspond to the vector $\boldsymbol{x_i}$ restricted to the values j for which $\gamma_j = 1$ and $\boldsymbol{\beta_\gamma}$ is the vector of corresponding regression coefficients.

For the remaining boundaries, we assign diffuse priors following [Kwon et al., 2007] that express no prior belief by setting the boundaries s_j to be uniformly distributed on (s_{j-1}, s_{j+1}).

5.2.1 MCMC algorithm

Next we provide the updating steps for the new parameters: the latent variables $\boldsymbol{Z}$ and the boundary parameters $\boldsymbol{s}$:

1. Update the vector of latent variables $\boldsymbol{Z}$ from its posterior distribution given $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{s}, \boldsymbol{y}$, which is a truncated normal density under the constraints defined in (5.2):

$$Z_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{s},\ \boldsymbol{y}\ \sim\ N(\boldsymbol{x}'_{\boldsymbol{i\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}},\ 1), \tag{5.9}$$

truncated on the left by s_{j-1} and on the right by s_j.

2. Update the boundary parameters s_j from their posterior densities given $\boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{s}_{-j}$ where $\boldsymbol{s}_{-j}$ is the vector $\boldsymbol{s}$ without the j-th element.

This conditional distribution is uniform on the interval

$$s_j|\boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{s}_{-j} \sim U(\big[\max\{\max\{Z_i : y_i = j\}, s_{j-1}\}\}, \min\{\min\{Z_i : y_i = j+1\}, s_{j+1}\}\big]), \tag{5.10}$$

as described in [Albert and Chib, 1993].

5.2.2 Implementation

To implement a MCMC setting for ordinal phenotypes, we extend the Guan and Stephens' approach to handling binary phenotypes to ordinal categorical phenotypes, which preserves

most of the parameter updates of piMASS. We implement the same practical constraints that were suggested by Guan and Stephens [Guan and Stephens, 2011].

First, an identifiability constraint was imposed on $\boldsymbol{Z}$ to have empirical variance 1. Second, we center all variables and set $\mu = 0$. Third, to improve mixing, we make an assumption that the marginal distribution of $\boldsymbol{Z}$ is normal [Guan and Stephens, 2011]. In particular, $Z_1, \ldots, Z_n \mid \boldsymbol{y}$ are restricted to take a fixed set of values, which are set to be n equally spaced quantiles of the standard normal distribution. The number of subjects in each of the categories is n_1, n_2, ..., n_J, where $n = n_1 + n_2 + \ldots + n_J$. We assign the first n_1 of these quantiles to the n_1 individuals with $y_i = 1$, the next n_2 of these quantiles to the n_2 individuals with $y_i = 2$ and so on with the last n_J of these quantiles assigned to the n_J individuals with $y_i = J$. This way we create a $\boldsymbol{Z}$ vector with larger values of $\boldsymbol{Z}$ corresponding to larger value of the observed phenotype. This is a reasonable assumption if there are no large values of $\boldsymbol{\beta}$. Let $s_2 = \max_{Z_i}\{y_i = 2\}$, ... , $s_{J-1} = \max_{Z_i}\{y_i = J-1\}$ be the initial values for $\boldsymbol{s}$.

Local Metropolis-Hastings proposals for $\boldsymbol{Z}$ involve randomly picking a pair of individuals (m, n) with the same ordinal phenotype (both 0, 1 or 2) and propose to swap their latent response values Z_m and Z_n. The Metropolis-Hastings ratio is calculated using (3.14). The decision to accept or reject the proposed $\boldsymbol{Z}$ and other proposed parameters is based on (3.14).

5.3 Real data analysis

We use dataset ADNI1 provided by Alzheimer's Disease Neuroimaging Initiative (ADNI) (`http://adni.loni.usc.edu`). ADNI1 GWAS data set contains $620,901$ SNP and CNV markers for 757 subjects (449 males, 308 females) genotyped using Illumina Human610-Quad BeadChip platform with 3 types of Patient Diagnosis groups: 214 Cognitively Normal, 363 Mild Cognitive Impairment, 180 Alzheimer's Disease patients. Here we code the phenotype as follows: 0-Cognitively Normal, 1-Mild Cognitive Impairment, 2-Alzheimer's disease. We use this phenotype with 3 categories in our analysis.

We applied the method proposed in this chapter to the data described above with $1,000,000$ MCMC iterations. We sampled every 10 iterations, therefore reducing MCMC output to a total of $100,000$ samples to reduce autocorrelation. The standard diagnostic procedures did not indicate lack of convergence.

The results of the analysis are presented below. The Manhattan-like plot of the PIPs of individual SNPs using the ADNI1 dataset is shown in Figure 5.1, with PIP on the y-axis.

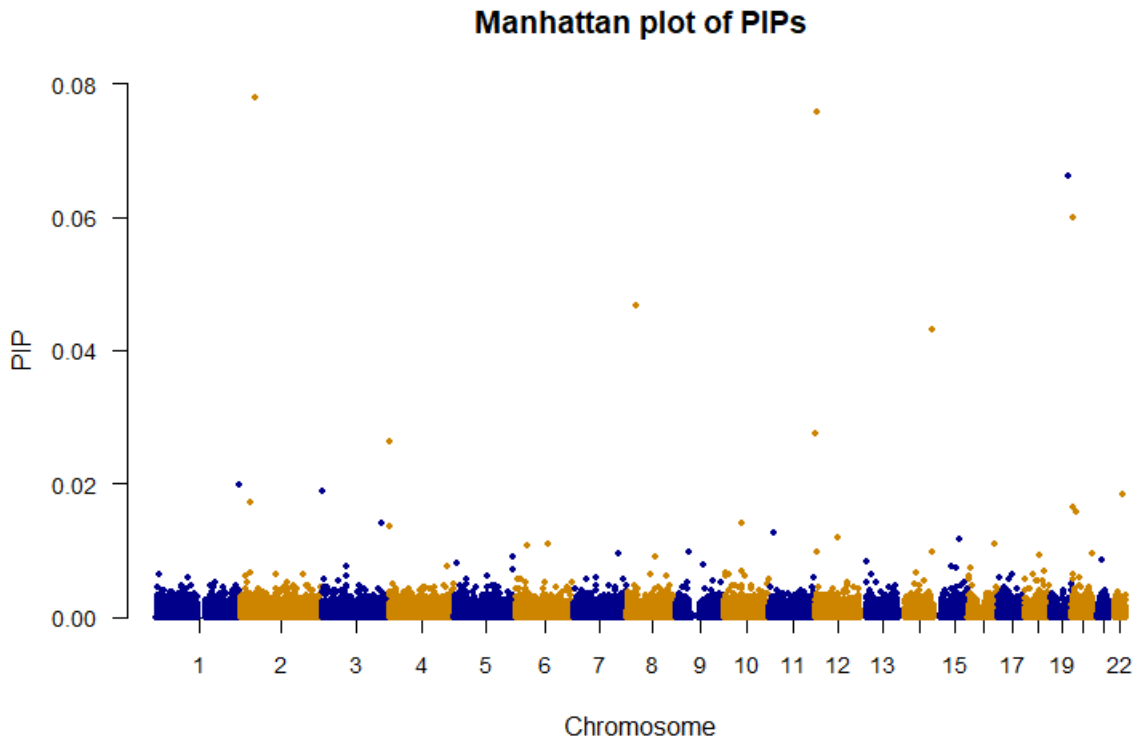

Figure 5.1: Plot of PIP for ADNI1 data set for the categorical phenotype method

To investigate the SNPs that visibly separate from the rest, below we present a table containing top 30 SNPs with the best association metric (PIP).

Chr	rsID	Position	Rank	PIP
2	rs4953672	42953942	1	0.07794
12	rs12822144	4176072	2	0.07572
19	rs2075650	50087459	3	0.0661
20	rs6116375	4399364	4	0.05996
8	rs2055195	26696942	5	0.04668
14	rs10133989	100949306	6	0.04305
12	rs2906109	1142624	7	0.02757
4	rs2306245	852156	8	0.02635
1	rs10924809	244929211	9	0.01981
3	rs11706690	398018	10	0.0189
22	rs17365991	40120125	11	0.01839
2	rs6731612	31199704	12	0.01711
20	rs530652	878560	13	0.01652
20	rs362584	10202475	14	0.01571
3	rs10804857	176728521	15	0.01402
10	rs6480572	53339650	16	0.01401
4	rs2279186	865887	17	0.01353
11	rs274492	11056714	18	0.01266
12	rs10506562	67295983	19	0.0118
15	rs734854	72068901	20	0.01159
6	rs9390855	96659997	21	0.01088
16	rs3751834	77573909	22	0.01087
6	rs86715	33588413	23	0.01062
12	rs4149577	6317783	24	0.00975
9	rs10973041	36713143	25	0.00974
14	rs17613673	98338710	26	0.00969
20	rs11906462	60569397	27	0.00945
7	rs9641952	132953992	28	0.00943
18	rs883218	45400797	29	0.00918
8	rs1382617	83474224	30	0.00908

Table 5.1: Top 30 SNPs with largest PIP for the categorical phenotype method

First, we focus on the SNPs with the highest obtained PIP and investigate their role and whether they have been implicated in previous studies. We used GWAS catalog available at `https://www.ebi.ac.uk/gwas` to match SNPs with high PIPs to the genes associated with Alzheimer's disease as well as mental disorder related diseases.

The top signal is rs4953672, an intergenic SNP located at chr2:43100438 which has not been detected in previous studies. The second best signal rs12822144 an SNP located at chr12:4305811 within the intergenic region between *RPL18P9* and *CCND2* and has been found to be associated with *APOE* ϵ-4 non-carriers of Alzheimer's Disease [Jiang et al., 2015]. The third best signal rs2075650 is located at chr19:45395619 in *TOMM40* gene. rs2075650 has confirmed highest association with the apolipoprotein E (*APOE*) locus ($P = 1.8{\times}10^{-157}$) [Harold et al., 2009]. rs2075650 was also found to be the top SNP associated with late-onset Alzheimer's dementia [Heinzen et al., 2010]. Other studies of late onset Alzheimer's disease also confirmed *APOE*'s risk effect ($P = 1 \times 10^{-36}$) [Naj et al., 2010]. Additionally, the locus rs2075650 reached genome-wide significant evidence for association with AD at the *APOE* ($P = 1.04 \times 10^{-295}$) [Seshadri et al., 2010]. The fact that our analysis with MCI patients picked up this signal points at its potential role in the development of MCI.

Next, rs6116375 (rank 4), an intergenic SNP located at chr20:4451364 is known to be associated with Alzheimer's disease in whole-genome analyses of whole-brain data [Huang et al., 2015]. rs2055195 (rank 5) is a part of the *ADRA1A* gene located at chr8:26641025. The *ADRA1A* gene has been associated with grey matter volume measurement [Alliey-Rodriguez et al., 2019]. rs2906109 (rank 7) is a part of *ERC1* gene located at

chr12:1272363 is in linkage disequilibrium (LD) with rs61913097 ($R^2 = 0.5981$), which was found to be associated with attention deficit hyperactivity disorder [Kweon et al., 2018]. rs2306245 (rank 8) is a part of the *GAK* gene and located at chr4:862156 is in LD with rs11248051 ($R^2 = 0.2215$), which plays a role in Parkinson's disease [Hamza et al., 2010]. rs10924809 (rank 9) is located near *SCCPDH* at chr1:246862588. rs10924809 is in LD with rs6426328 ($R^2 = 0.1048$), which is known to be associated with apolipoprotein B measurement [Richardson et al., 2020]. rs10924809 is also in LD with rs3007305 ($R^2 = 0.1863$), which is associated with schizophrenia, bipolar disorder, attempted suicide [Mullins et al., 2019].

Some SNPs that have not been mentioned before were not found to be in LD with known markers, but map to genes associated with Alzheimer's disease and other brain disorders. For example, rs11706690 (rank 10) is a part of *CHL1* gene and is located at chr3:423018. *CHL1* gene is associated with Parkinsonism in frontotemporal lobe dementia [Pottier et al., 2018]. rs17365991 (rank 11) is located at chr22:41790179 and a part of the *TEF* gene. *TEF* gene is associated with mood instability measurement [Nagel et al., 2018] and multiple sclerosis [Beecham et al., 2013]. rs530652 (rank 13) is located at chr20:930560 and belongs to *RSPO4* gene. *RSPO4* is associated with logical memory (immediate recall) in Alzheimer's disease dementia [Chung et al., 2018]. rs10804857 (rank 15) is located at chr3:175245827 and is mapped to *NAALADL2* gene. *NAALADL2* gene is associated with late-onset Alzheimer's disease [Mez et al., 2017].

Several markers were found to be in LD with markers highlighted in previous association studies. There was a group of markers that were connected to the SNPs associated with

various phenotypes reflecting the cognitive function. First, rs362584 (rank 14) is located at chr20:10254475 in *SNAP25* gene. rs36258 is in LD with rs362987 ($R^2 = 0.3957$), which is associated with self reported educational attainment [Rietveld et al., 2014]. Moreover, rs9390855 (rank 21) is located at chr6:96553276 and is mapped to *FUT9* gene. rs9390855 is in LD with rs7763181 ($R^2 = 0.1399$), which is associated with self reported educational attainment [Lee et al., 2018]. Additionally, rs6480572 (rank 16) is located at chr10:53669644 and is a part of *PRKG1* gene. rs6480572 is in LD with rs10823860 ($R^2 = 1$), which is associated with mathematical ability [Lee et al., 2018]. Also, rs9641952 (rank 28) is located at chr7:133303452 and is mapped to gene *EXOC4*. rs9641952 is in LD with rs4728302 ($R^2 = 0.1312$), which is associated with intelligence [Sniekers et al., 2017]. The variety of highlighted SNPs points at links with various markers of person's general cognitive performance to MCI and AD.

Several associations were found to be connected to some other brain-related diseases like multiple sclerosis and Parkinson's disease. rs2279186 (rank 17) is located at chr4:875887 and is mapped to *GAK* gene. rs2279186 is in LD with rs873786 ($R^2 = 0.1952$), which is associated with Parkinson's disease [Nalls et al., 2019]. rs4149577 (rank 24) located at chr12:6447522 is a part of *TNFRSF1A* and is in LD with rs1800693 ($R^2 = 0.5496$), which is associated with multiple sclerosis [Andlauer et al., 2016]. rs883218 (rank 29) is also in LD with rs28512338 ($R^2 = 0.1746$), which is associated with multiple sclerosis [Consortium et al., 2019].

We also found a group of markers that are associated with various physical illnesses not directly connected to the brain and its physiology. For example, rs6480572 (rank 16) is

also in LD with rs10823893 ($R^2 = 0.2468$), which is associated with body mass index [Pulit et al., 2019]. rs734854 (rank 20) is located at chr15:74281848 and is mapped to *STOML1* gene. rs734854 is in LD with rs2507 ($R^2 = 0.2232$), which is associated with coronary artery disease [van der Harst and Verweij, 2018]. rs883218 (rank 29) is located at chr18:47146799 is mapped to gene *LIPG* and is in LD with rs1105654 ($R^2 = 0.2359$). rs1105654 is associated with metabolic syndrome [Lind, 2019].

Since majority (63%) of the SNPs among the top 30 SNPs were mapped to genes with known association with AD or other mental disorders, the top ranking SNP rs4953672 (rank 1) has potential to be a novel marker associated with AD and the further investigation of rs4953672 is worth conducting.

5.4 Conclusion

Herein, we broadened the BVS method piMASS to work with ordinal categorical phenotype. The approach involved extension of the method suggested by [Guan and Stephens, 2011] for binary data to ordinal categorical response variable.

We applied the new method to the ADNI1 data set provided by Alzheimer's Disease Neuroimaging Initiative, which contains $620,901$ SNP and CNV markers for 757 subjects and contains 3 groups of patients: AD, MCI and controls. The analysis showed strong overlap with many markers previously discovered in other association studies, including some

previously unknown markers.

Overall, we demonstrated the practical value of the method and its simplicity of use in analyzing ordinal outcomes. Given relative sparsity of the methods dealing with GWAS for ordinal outcome, the method has a lot of potential in analyzing new traits.

CHAPTER 6

CONCLUSION

In this work, we have presented various ways to improve and extend the applicability of BVS methods for GWAS. Since the problem of GWAS is complex and depends on a lot of factors, incorporating more features of the association between genotypes and phenotype can make the selection of relevant genotypes more precise. We attempted to extend the genome wide studies in two main directions. First, we want to incorporate the effects of additional covariates into the association analysis, which can alleviate the problems such as population stratification which are often present in large scale samples. Second, we want to extend the applicability of BVS methods for GWAS to work with phenotypes expressed on various scales: continuous, binary and ordinal. Those two general directions allow the existing BVS methods reach wider applicability and potentially higher precision.

In Chapter 1, we presented the BVS methods in a context of GWAS. We discussed typical prior setup and discussed choices that one faces when tailoring the variable selection method to a particular problem. In particular, specifying the form of the variance of the regression coefficients can have significant effect on the way variable selection is performed. We then focused on the theoretical foundation of the BVS method piMASS [Guan and Stephens, 2011],

which was developed with GWAS studies in mind and has prior structure reflecting many constraints and assumptions genome wide association studies usually face. Particular attention was given to the prior on the variance of the effect size, allowing the inclusion of many variables with small effects, which is often the case in most GWAS. We discussed the MCMC scheme of the piMASS, which will serve as the base for our developments in later chapters. At the end of the chapter we provide the outline of the dissertation.

In Chapter 2, we performed the full GWAS of the MGS dataset containing genetic data on schizophrenia patients and controls with 5334 subjects using multivariate BVS method piMASS. We contrasted our results with the previous univariate analysis of the MGS dataset [Shi et al., 2009]. We showed that piMASS can improve the power of detecting SNPs associated with schizophrenia, potentially leading to new discoveries from existing data sets without increasing the sample size. To allow for local additive effects, we tested SNPs in groups. We used permutation test to determine statistical significance to compare our results with univariate method. While the previous univariate analysis of the MGS dataset revealed no genome-wide significant loci, using the same dataset we identified a single region that exceeded the genome-wide significance. The result was replicated using an independent Swedish Schizophrenia CaseControl Study (SSCCS) dataset. The study demonstrated the capability of piMASS to analyze real data sets and get meaningful conclusions from the analysis. The results of this chapter explains our motivation for using the piMASS as a base methods for extensions provided in the next chapters.

In Chapter 3 we presented the first extension of the piMASS, which involved incorporating

additional covariates, potentially accounting for population stratification in the variable selection process for GWAS with continuous phenotype. Since current pooled genetic datasets can contain distantly related subjects, it is possible that their relatedness can lead to false positive associations. The ability to account for population stratification allows for more precise association identification by decreasing the effects of possible population stratification often present in the sample. We implemented the extension by including additional covariates in the regression model and specifying priors on them. When identifying the associations, their effect is subtracted from the continuous phenotype, leaving the part that is assumed to vary due to genetic effects. We provided full theoretical treatment of the problem and specified a detailed MCMC algorithm. We performed simulation studies using both simulated and real genetic data and generated 100 phenotypes for each data type. ROC analysis of the data sets containing simulated genetic data showed slight increase in the AUC for the proposed method. This difference was found to be statistically significant at 5% level (p-value = 0.01477) according to the DeLong's test for two correlated ROC curves, which tests if the difference in AUCs is random. The PAUC metric showed similar results with the new method showing slight increase and statistically significant difference (p-value= 0.01249). The ROC analysis of the datasets containing real genotypes showed similar results for AUC (p-value= 0.002198) and PAUC (p-value= 0.002575). Both AUC and PAUC results for both simulated and real genetic variants show that the difference between the proposed method and piMASS is unlikely to be random. Overall, simulation studies confirmed the efficacy of the proposed method.

In Chapter 4 we presented the second extension, which included incorporating additional covariates in the BVS model for GWAS with binary phenotype. Binary phenotype requires additional layer of analysis accounting for it, which poses additional challenges. The goal of the extension is to increase the precision of the association analysis by decreasing the effects of population stratification. We provided theoretical framework for including additional covariates when the phenotype is binary, which is based on the data augmentation approach [Albert and Chib, 1993] and involved introducing latent continuous variables that serve as an underlying response. We implemented the method and conducted simulation studies to check its efficacy. The ROC analysis of the studies performed using data simulated using package COSI showed the significant visible dominance of the new method. However, according to the DeLongs test for two correlated ROC curves values of AUC showed the p-values of 9.4%, which is above the 5% significance threshold. The results for PAUC are similar and showed p-value of 27.18% which shows that for a specified range of specificity the performance of the two methods is similar. Performance in simulations using real data was slightly better with p-values for AUC and PAUC being 7.8% and 7.7%, saying that the difference is unlikely to have occurred by chance. In this chapter we also provide real data analysis using real dataset ADNI1 containing the 620,901 SNP and copy number variant (CNV) markers for 757 subjects (449 males, 308 females) with 3 types of Patient Diagnosis groups: Neurotypical Controls, Minor Cognitive Impairment, Alzheimers Disease. We combined the MCI and AD together to get binary phenotype. We successfully applied the new method to this data and discussed the results. Overall, we demonstrated the ease of use and applicability of the

proposed method.

Chapter 5 contains the third extension which included the BVS method that can work for GWAS with ordinal phenotypes. Many association studies can be performed on traits that are recorded on ordinal scale, however, there are not many methods that easily permit that, especially among BVS methods. Therefore, creating a BVS methods capable of finding associations with an ordinal outcome is useful. Here we extended the piMASS to work with ordinal phenotype by extending the simplified data augmentation approach [Albert and Chib, 1993] offered in [Guan and Stephens, 2011]. We then performed a real data analysis of the data set containing genetic data and ordinal phenotype. We applied the method to the ADNI1 dataset containing the phenotype with 3 categories: 0-Neurotypical Controls, 1-MCI and 2-Alzheimer's disease.

We found various previously discovered associations as well as some novel associations. The top signal rs4953672 has not been detected in the previous studies. The third best signal rs2075650 has been found to have significant association with Alzheimer's disease from numerous large scale GWAS and has confirmed highest association with the apolipoprotein E (*APOE*) locus. Overall, several markers were found in previous Alzheimer's GWAS studies. We also found a group of markers (rs36258, rs939085, rs6480572, rs9641952) that are connected to the SNPs associated with various phenotypes reflecting the cognitive function. Such phenotypes included self reported educational attainment, mathematical ability and intelligence. Moreover, we found a group of markers (rs6480572, rs73485, rs883218) that were in LD with the markers associated with various physical illnesses, not directly connected to

the brain and its physiology. Such phenotypes included body mass index, coronary artery disease and metabolic syndrome. Additionally, some markers (rs227918, rs4149577, rs88321) were in LD with markers associated with other brain-related diseases such as Parkinsons disease and multiple sclerosis. Significant overlap with previous research suggests the method works well and presence of new associations point in the direction of potential improvement over existing methods.

There are several ways current work could be expanded and improved. One possible extension is to develop a combination method that both accounts for additional covariates related to population stratification and works for ordinal phenotypes. In Chapter 4 we offered a method to account for population stratification for binary phenotypes and it is straightforward to extend it to work with ordinary phenotypes using the approach presented in Chapter 5.

Next, accounting for correlation structure among covariates is a potential source of significant improvement. However, computational factors should be considered when implementing it since repetitive large scale matrix operations may serve as a limiting factor for the applicability of such method. Additionally, extending BVS method to work with unordered categorical phenotype is another possible extension that can increase the applicability of the method. Moreover, in this work we adopted simplified view of population stratification that assumes that several derived covariates can account for population structure complexity. It is possible there is a more detailed way to incorporate this into BVS.

As field of GWAS continues to develop and grow, there is more need for dedicated statistical tools. We addressed some of the issues that arise, in particular for neuropsychiatric and

other phenotypes, which are expressed as a categorical variable. Ultimately, we would like to see the development of the methods that can use the information obtained from GWAS and translate it into clinical therapies and allow for more personalized decision making for patients.

BIBLIOGRAPHY

[Albert and Chib, 1993] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

[Alliey-Rodriguez et al., 2019] Alliey-Rodriguez, N., Grey, T. A., Shafee, R., Asif, H., Lutz, O., Bolo, N. R., Padmanabhan, J., Tandon, N., Klinger, M., Reis, K., et al. (2019). NRXN1 is associated with enlargement of the temporal horns of the lateral ventricles in psychosis. *Translational Psychiatry*, 9(1):1–7.

[Andlauer et al., 2016] Andlauer, T. F., Buck, D., Antony, G., Bayas, A., Bechmann, L., Berthele, A., Chan, A., Gasperi, C., Gold, R., Graetz, C., et al. (2016). Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Science Advances*, 2(6):e1501678.

[Astle et al., 2009] Astle, W., Balding, D. J., et al. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471.

[Banerjee et al., 2018] Banerjee, S., Zeng, L., Schunkert, H., and Söding, J. (2018). Bayesian multiple logistic regression for case-control GWAS. *PLoS Genetics*, 14(12):e1007856.

[Baragatti et al., 2011] Baragatti, M. et al. (2011). Bayesian variable selection for probit mixed models applied to gene selection. *Bayesian Analysis*, 6(2):209–229.

[Beecham et al., 2013] Beecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., Cotsapas, C., Shah, T. S., Spencer, C., Booth, D., Goris, A., et al. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, 45(11):1353–1360.

[Bertram and Tanzi, 2009] Bertram, L. and Tanzi, R. E. (2009). Genome-wide association studies in Alzheimer's disease. *Human Molecular Genetics*, 18(R2):R137–R145.

[Bertram and Tanzi, 2012] Bertram, L. and Tanzi, R. E. (2012). The genetics of Alzheimer's disease. In *Progress in Molecular Biology and Translational Science*, volume 107, pages 79–100. Elsevier.

[Brazel et al., 2019] Brazel, D. M., Jiang, Y., Hughey, J. M., Turcot, V., Zhan, X., Gong, J., Batini, C., Weissenkampen, J. D., Liu, M., Surendran, P., et al. (2019). Exome chip meta-analysis fine maps causal variants and elucidates the genetic architecture of rare coding variants in smoking and alcohol use. *Biological Psychiatry*, 85(11):946–955.

[Carbonetto et al., 2012] Carbonetto, P., Stephens, M., et al. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

[Chambers, 2018] Chambers, J. M. (2018). *Graphical Methods for Data Analysis*. CRC Press.

[Charlson et al., 2018] Charlson, F. J., Ferrari, A. J., Santomauro, D. F., Diminic, S., Stockings, E., Scott, J. G., McGrath, J. J., and Whiteford, H. A. (2018). Global epidemiology and burden of schizophrenia: findings from the global burden of disease study 2016. *Schizophrenia Bulletin*, 44(6):1195–1203.

[Chung et al., 2018] Chung, J., Wang, X., Maruyama, T., Ma, Y., Zhang, X., Mez, J., Sherva, R., Takeyama, H., Lunetta, K. L., Farrer, L. A., et al. (2018). Genome-wide association study of Alzheimer's disease endophenotypes at prediagnosis stages. *Alzheimer's & Dementia*, 14(5):623–633.

[Consortium et al., 2019] Consortium, I. M. S. G., ANZgene, IIBDGC, and WTCCC2 (2019). Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*, 365(6460):eaav7188.

[Davies et al., 2018] Davies, G., Lam, M., Harris, S. E., Trampush, J. W., Luciano, M., Hill, W. D., Hagenaars, S. P., Ritchie, S. J., Marioni, R. E., Fawns-Ritchie, C., et al. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature Communications*, 9(1):1–16.

[DeLong et al., 1988] DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845.

[Devlin and Roeder, 1999] Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

[Fabbri et al., 2019] Fabbri, C., Kasper, S., Kautzky, A., Bartova, L., Dold, M., Zohar, J., Souery, D., Montgomery, S., Albani, D., Raimondi, I., et al. (2019). Genome-wide association study of treatment-resistance in depression and meta-analysis of three independent samples. *The British Journal of Psychiatry*, 214(1):36–41.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

[Forstner et al., 2014] Forstner, A. J., Basmanav, F. B., Mattheisen, M., Böhmer, A. C., Hollegaard, M. V., Janson, E., Strengman, E., Priebe, L., Degenhardt, F., Hoffmann, P., et al. (2014). Investigation of the involvement of MIR185 and its target genes in the development of schizophrenia. *Journal of Psychiatry & Neuroscience: JPN*, 39(6):386.

[Freedman et al., 2004] Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., et al. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, 36(4):388–393.

[Gelman et al., 2013] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

[George and McCulloch, 1993] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

[George and McCulloch, 1997] George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373.

[German et al., 2020] German, C. A., Sinsheimer, J. S., Klimentidis, Y. C., Zhou, H., and Zhou, J. J. (2020). Ordered multinomial regression for genetic association analysis of ordinal

phenotypes at Biobank scale. *Genetic Epidemiology*, 44(3):248–260.

[Guan and Stephens, 2011] Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815.

[Hall et al., 2007] Hall, H., Lawyer, G., Sillen, A., Jönsson, E. G., Agartz, I., Terenius, L., and Arnborg, S. (2007). Potential genetic variants in schizophrenia: a Bayesian analysis. *The World Journal of Biological Psychiatry*, 8(1):12–22.

[Hall et al., 2018] Hall, L. S., Adams, M. J., Arnau-Soler, A., Clarke, T.-K., Howard, D. M., Zeng, Y., Davies, G., Hagenaars, S. P., Fernandez-Pujals, A. M., Gibson, J., et al. (2018). Genome-wide meta-analyses of stratified depression in Generation Scotland and UK Biobank. *Translational Psychiatry*, 8(1):1–12.

[Hamza et al., 2010] Hamza, T. H., Zabetian, C. P., Tenesa, A., Laederach, A., Montimurro, J., Yearout, D., Kay, D. M., Doheny, K. F., Paschall, J., Pugh, E., et al. (2010). Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nature Genetics*, 42(9):781.

[Harold et al., 2009] Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., Pahwa, J. S., Moskvina, V., Dowzell, K., Williams, A., et al. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics*, 41(10):1088.

[Hebert et al., 2013] Hebert, L. E., Weuve, J., Scherr, P. A., and Evans, D. A. (2013). Alzheimer's disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*, 80(19):1778–1783.

[Heinzen et al., 2010] Heinzen, E. L., Need, A. C., Hayden, K. M., Chiba-Falek, O., Roses, A. D., Strittmatter, W. J., Burke, J. R., Hulette, C. M., Welsh-Bohmer, K. A., and Goldstein, D. B. (2010). Genome-wide scan of copy number variation in late-onset Alzheimer's disease. *Journal of Alzheimer's Disease*, 19(1):69–77.

[Hindorff et al., 2009] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional impli-

cations of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.

[Hoff, 2009] Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer.

[Huang et al., 2015] Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., Zhu, H., Initiative, A. D. N., et al. (2015). FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage*, 118:613–627.

[Hughes et al., 2020] Hughes, D. A., Bacigalupe, R., Wang, J., Rühlemann, M. C., Tito, R. Y., Falony, G., Joossens, M., Vieira-Silva, S., Henckaerts, L., Rymenans, L., et al. (2020). Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nature Microbiology*, 5(9):1079–1087.

[Iwata et al., 2009] Iwata, H., Ebana, K., Fukuoka, S., Jannink, J.-L., and Hayashi, T. (2009). Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among Oryza sativa L. germplasms. *Theoretical and Applied Genetics*, 118(5):865–880.

[Jansen et al., 2019] Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics*, 51(3):404–413.

[Jia et al., 2017] Jia, P., Han, G., Zhao, J., Lu, P., and Zhao, Z. (2017). SZGR 2.0: a one-stop shop of schizophrenia candidate genes. *Nucleic Acids Research*, 45(D1):D915–D924.

[Jiang et al., 2015] Jiang, S., Yang, W., Qiu, Y., Chen, H.-Z., (ADNI, A. D. N. I., et al. (2015). Identification of novel quantitative traits-associated susceptibility loci for APOE ε 4 non-carriers of Alzheimer's disease. *Current Alzheimer's Research*, 12(3):218–227.

[Jiang and Zhang, 2011] Jiang, Y. and Zhang, H. (2011). Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. *Genetic Epidemiology*, 35(2):125–132.

[Kanazawa et al., 2013] Kanazawa, T., Ikeda, M., Glatt, S. J., Tsutsumi, A., Kikuyama, H., Kawamura, Y., Nishida, N., Miyagawa, T., Hashimoto, R., Takeda, M., et al. (2013). Genome-wide association study of atypical psychosis. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 162(7):679–686.

[Kärkkäinen and Sillanpää, 2012] Kärkkäinen, H. P. and Sillanpää, M. J. (2012). Robustness of Bayesian multilocus association models to cryptic relatedness. *Annals of Human Genetics*, 76(6):510–523.

[Kass and Raftery, 1995] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

[Kircher et al., 2014] Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315.

[Kunkle et al., 2019] Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya, M., Van Der Lee, S. J., Amlie-Wolf, A., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics*, 51(3):414–430.

[Kweon et al., 2018] Kweon, K., Shin, E.-S., Park, K. J., Lee, J.-K., Joo, Y., and Kim, H.-W. (2018). Genome-wide analysis reveals four novel loci for attention-deficit hyperactivity disorder in Korean youths. *Journal of the Korean Academy of Child and Adolescent Psychiatry*, 29(2):62–72.

[Kwon et al., 2007] Kwon, D., Tadesse, M. G., Sha, N., Pfeiffer, R. M., and Vannucci, M. (2007). Identifying biomarkers from mass spectrometry data with ordinal outcome. *Cancer Informatics*, 3:19–28.

[Lee et al., 2018] Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Linnér, R. K., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50(8):1112–1121.

[Lee et al., 2003] Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics*, 19(1):90–97.

[Lee et al., 2012] Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., Keller, M. C., Visscher, P. M., and Wray, N. R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, 44(3):247–250.

[Levine et al., 2012] Levine, A. J., Service, S., Miller, E. N., Reynolds, S. M., Singer, E. J., Shapshak, P., Martin, E. M., Sacktor, N., Becker, J. T., Jacobson, L. P., et al. (2012). Genome-wide association study of neurocognitive impairment and dementia in HIV-infected adults. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159(6):669–683.

[Levinson et al., 2012] Levinson, D. F., Shi, J., Wang, K., Oh, S., Riley, B., Pulver, A. E., Wildenauer, D. B., Laurent, C., Mowry, B. J., Gejman, P. V., et al. (2012). Genome-wide association study of multiplex schizophrenia pedigrees. *American Journal of Psychiatry*, 169(9):963–973.

[Lieberman et al., 2005] Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keefe, R. S., Davis, S. M., Davis, C. E., Lebowitz, B. D., et al. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine*, 353(12):1209–1223.

[Lind, 2019] Lind, L. (2019). Genome-wide association study of the metabolic syndrome in UK Biobank. *Metabolic Syndrome and Related Disorders*, 17(10):505–511.

[Liu et al., 2018] Liu, J., Zhou, Y., Liu, S., Song, X., Yang, X.-Z., Fan, Y., Chen, W., Akdemir, Z. C., Yan, Z., Zuo, Y., et al. (2018). The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Human Genetics*, 137(6-7):553–567.

[Liu et al., 2016] Liu, X., Study, B. G., Kelsoe, J. R., and Greenwood, T. A. (2016). A genome-wide association study of bipolar disorder with comorbid eating disorder replicates the SOX2-OT region. *Journal of Affective Disorders*, 189:141–149.

[Logsdon et al., 2010] Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, 11(1):58.

[Luciano et al., 2011] Luciano, M., Hansell, N. K., Lahti, J., Davies, G., Medland, S. E., Räikkönen, K., Tenesa, A., Widen, E., McGhee, K. A., Palotie, A., et al. (2011). Whole genome association scan for genetic polymorphisms influencing information processing speed. *Biological Psychology*, 86(3):193–202.

[Ma et al., 2013] Ma, H., Bandos, A. I., Rockette, H. E., and Gur, D. (2013). On use of partial area under the ROC curve for evaluation of diagnostic performance. *Statistics in Medicine*, 32(20):3449–3458.

[MacArthur et al., 2017] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901.

[Marchini et al., 2004] Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517.

[Martin and Eskin, 2017] Martin, L. S. and Eskin, E. (2017). Population Structure in Genetic Studies: Confounding Factors and Mixed Models. *bioRxiv*, page 092106.

[Mayeux and Stern, 2012] Mayeux, R. and Stern, Y. (2012). Epidemiology of Alzheimer's disease. *Cold Spring Harbor Perspectives in Medicine*, 2(8):a006239.

[Mez et al., 2017] Mez, J., Chung, J., Jun, G., Kriegel, J., Bourlas, A. P., Sherva, R., Logue, M. W., Barnes, L. L., Bennett, D. A., Buxbaum, J. D., et al. (2017). Two novel loci, COBL and SLC10A2, for Alzheimer's disease in African Americans. *Alzheimer's & Dementia*, 13(2):119–129.

[Moreno-Küstner et al., 2018] Moreno-Küstner, B., Martin, C., and Pastor, L. (2018). Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PloS One*, 13(4):e0195687.

[Mukherjee et al., 2018] Mukherjee, S., Mez, J., Trittschuh, E. H., Saykin, A. J., Gibbons, L. E., Fardo, D. W., Wessels, M., Bauman, J., Moore, M., Choi, S.-E., et al. (2018). Genetic data and cognitively defined late-onset Alzheimer's disease subgroups. *Molecular Psychiatry*, 25:2942–2951.

[Mullins et al., 2019] Mullins, N., Bigdeli, T. B., Børglum, A. D., Coleman, J. R., Demontis, D., Mehta, D., Power, R. A., Ripke, S., Stahl, E. A., Starnawska, A., et al. (2019). GWAS of suicide attempt in psychiatric disorders and association with major depression polygenic risk scores. *American Journal of Psychiatry*, 176(8):651–660.

[Nagel et al., 2018] Nagel, M., Watanabe, K., Stringer, S., Posthuma, D., and Van Der Sluis, S. (2018). Item-level analyses reveal genetic heterogeneity in neuroticism. *Nature Communications*, 9(1):1–10.

[Naj et al., 2010] Naj, A. C., Beecham, G. W., Martin, E. R., Gallins, P. J., Powell, E. H., Konidari, I., Whitehead, P. L., Cai, G., Haroutunian, V., Scott, W. K., et al. (2010). Dementia revealed: novel chromosome 6 locus for late-onset Alzheimer's disease provides genetic evidence for folate-pathway abnormalities. *PLoS Genetics*, 6(9):e1001130.

[Nalls et al., 2019] Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., Tan, M., Kia, D. A., Noyce, A. J., Xue, A., et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology*, 18(12):1091–1102.

[Nurnberger et al., 2014] Nurnberger, J. I., Koller, D. L., Jung, J., Edenberg, H. J., Foroud, T., Guella, I., Vawter, M. P., and Kelsoe, J. R. (2014). Identification of pathways for bipolar disorder: a meta-analysis. *JAMA Psychiatry*, 71(6):657–664.

[Park et al., 2015] Park, S. L., Carmella, S. G., Chen, M., Patel, Y., Stram, D. O., Haiman, C. A., Le Marchand, L., and Hecht, S. S. (2015). Mercapturic acids derived from the

toxicants acrolein and crotonaldehyde in the urine of cigarette smokers from five ethnic groups with differing risks for lung cancer. *PloS One*, 10(6):e0124841.

[Paschou et al., 2007] Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W., and Drineas, P. (2007). PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, 3(9):e160.

[Pottier et al., 2018] Pottier, C., Zhou, X., Perkerson III, R. B., Baker, M., Jenkins, G. D., Serie, D. J., Ghidoni, R., Benussi, L., Binetti, G., de Munain, A. L., et al. (2018). Potential genetic modifiers of disease risk and age at onset in patients with frontotemporal lobar degeneration and GRN mutations: a genome-wide association study. *The Lancet Neurology*, 17(6):548–558.

[Price et al., 2006] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904.

[Price et al., 2010] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463.

[Pritchard et al., 2000] Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181.

[Pulit et al., 2019] Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., et al. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Human Molecular Genetics*, 28(1):166–174.

[Purcell et al., 2007] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

[Reisberg, 1986] Reisberg, B. (1986). Dementia: A systematic approach to identifying reversible causes. *Geriatrics*, 41(4):30–46.

[Reisberg and Gauthier, 2008] Reisberg, B. and Gauthier, S. (2008). Current evidence for subjective cognitive impairment (SCI) as the pre-mild cognitive impairment (MCI) stage of subsequently manifest Alzheimer's disease. *International Psychogeriatrics*, 20(1):1–16.

[Reitz et al., 2011] Reitz, C., Brayne, C., and Mayeux, R. (2011). Epidemiology of Alzheimer's disease. *Nature Reviews Neurology*, 7(3):137–152.

[Richardson et al., 2020] Richardson, T. G., Sanderson, E., Palmer, T. M., Ala-Korpela, M., Ference, B. A., Davey Smith, G., and Holmes, M. V. (2020). Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Medicine*, 17(3):e1003062.

[Rietveld et al., 2014] Rietveld, C. A., Esko, T., Davies, G., Pers, T. H., Turley, P., Benyamin, B., Chabris, C. F., Emilsson, V., Johnson, A. D., Lee, J. J., et al. (2014). Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences*, 111(38):13790–13794.

[Ripke et al., 2014] Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427.

[Ripke et al., 2011] Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., Lin, D.-Y., Duan, J., Ophoff, R. A., Andreassen, O. A., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*, 43(10):969.

[Robin et al., 2011] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Mller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.

[Rowe et al., 2019] Rowe, B., Chen, X., Wang, Z., Chen, J., and Amei, A. (2019). Biological and practical implications of genome-wide association study of schizophrenia using Bayesian variable selection. *npj Schizophrenia*, 5(1):1–7.

[Ruderfer et al., 2019] Ruderfer, D., Group, P. G. C. B. W., et al. (2019). Genomic dissection of bipolar disorder and schizophrenia in 50K cases, 50K controls and 28 subphenotypes. *European Neuropsychopharmacology*, 29:S814–S815.

[Saha et al., 2005] Saha, S., Chant, D., Welham, J., and McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. *PLoS Medicine*, 2(5):e141.

[Savage et al., 2018] Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., De Leeuw, C. A., Nagel, M., Awasthi, S., Barr, P. B., Coleman, J. R., et al. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics*, 50(7):912–919.

[Scelsi et al., 2018] Scelsi, M. A., Khan, R. R., Lorenzi, M., Christopher, L., Greicius, M. D., Schott, J. M., Ourselin, S., and Altmann, A. (2018). Genetic study of multimodal imaging Alzheimer's disease progression score implicates novel loci. *Brain*, 141(7):2167–2180.

[Schaffner et al., 2005] Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11):1576–1583.

[Servin and Stephens, 2007] Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics*, 3(7).

[Seshadri et al., 2010] Seshadri, S., Fitzpatrick, A. L., Ikram, M. A., DeStefano, A. L., Gudnason, V., Boada, M., Bis, J. C., Smith, A. V., Carrasquillo, M. M., Lambert, J. C., et al. (2010). Genome-wide analysis of genetic loci associated with Alzheimer's disease. *JAMA*, 303(18):1832–1840.

[Sha et al., 2004] Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., et al. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 60(3):812–819.

[Sherva et al., 2014] Sherva, R., Tripodis, Y., Bennett, D. A., Chibnik, L. B., Crane, P. K., De Jager, P. L., Farrer, L. A., Saykin, A. J., Shulman, J. M., Naj, A., et al. (2014). Genome-

wide association study of the rate of cognitive decline in Alzheimer's disease. *Alzheimer's & Dementia*, 10(1):45–52.

[Shi et al., 2009] Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., PeEr, I., Dudbridge, F., Holmans, P. A., Whittemore, A. S., Mowry, B. J., et al. (2009). Common variants on chromosome 6p22. 1 are associated with schizophrenia. *Nature*, 460(7256):753–757.

[Shlyakhter et al., 2014] Shlyakhter, I., Sabeti, P. C., and Schaffner, S. F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429.

[Sniekers et al., 2017] Sniekers, S., Stringer, S., Watanabe, K., Jansen, P. R., Coleman, J. R., Krapohl, E., Taskesen, E., Hammerschlag, A. R., Okbay, A., Zabaneh, D., et al. (2017). Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature Genetics*, 49(7):1107–1112.

[Sullivan et al., 2012] Sullivan, P. F., Daly, M. J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*, 13(8):537–551.

[Sullivan et al., 2003] Sullivan, P. F., Kendler, K. S., and Neale, M. C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry*, 60(12):1187–1192.

[Tucker et al., 2014] Tucker, G., Price, A. L., and Berger, B. (2014). Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics*, 197(3):1045–1049.

[van der Harst and Verweij, 2018] van der Harst, P. and Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation Research*, 122(3):433–443.

[Verma et al., 2008] Verma, R., Holmans, P., Knowles, J. A., Grover, D., Evgrafov, O. V., Crowe, R. R., Scheftner, W. A., Weissman, M. M., DePaulo Jr, J. R., Potash, J. B.,

et al. (2008). Linkage disequilibrium mapping of a chromosome 15q25-26 major depression linkage region and sequencing of NTRK3. *Biological Psychiatry*, 63(12):1185–1189.

[Visscher et al., 2012] Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24.

[Wang et al., 2010] Wang, K.-S., Liu, X.-F., and Aragam, N. (2010). A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. *Schizophrenia Research*, 124(1-3):192–199.

[Waring and Rosenberg, 2008] Waring, S. C. and Rosenberg, R. N. (2008). Genome-wide association studies in Alzheimer's disease. *Archives of Neurology*, 65(3):329–334.

[Wasserstein and Lazar, 2016] Wasserstein, R. L. and Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose.

[Wootton et al., 2019] Wootton, R. E., Richmond, R. C., Stuijfzand, B. G., Lawn, R. B., Sallis, H. M., Taylor, G. M., Hemani, G., Jones, H. J., Zammit, S., Smith, G. D., et al. (2019). Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a mendelian randomisation study. *Psychological Medicine*, 50(14):1–9.

[Wu et al., 2011] Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A comparison of association methods correcting for population stratification in case–control studies. *Annals of Human Genetics*, 75(3):418–427.

CURRICULUM VITAE

Benazir Rowe

Degrees:

Bachelor of Science - Mathematical Economics, 2012
Kyrgyz Russian Slavic University, Kyrgyzstan

Master of Science - International Economics and Finance, 2014
Otto von Guericke University Magdeburg, Germany

Special Honors and Awards:

UNLV summer doctoral research fellow 2017

Publications:

Rowe, B., Chen, X., Wang, Z., Chen, J., Amei, A. (2019). Biological and practical implications of genome-wide association study of schizophrenia using Bayesian variable selection. *npj Schizophrenia*, 5(1), 1-7.

Presentations:

"Bayesian variable selection for genome-wide association studies", Nevada ASA Fall Symposium, 2018
"Genome-wide association study of schizophrenia using Bayesian variable selection methods", NIPM Annual Symposium, 2018

Dissertation Title:

Bayesian Variable Selection Methods for Genome-wide Association Studies with Categorical Phenotypes

Dissertation Examination Committee:

Chairperson, Amei Amei, Ph.D.
Committee Member, Kaushik Ghosh, Ph.D.
Committee Member, Malwane Ananda, Ph.D.
Graduate Faculty Representative, Guogen Shan, Ph.D.

Email:

benazir.rowe@gmail.com