11-2017

Novel Statistical Models for Quantitative Shape-Gene Association Selection

Xiaotian Dai
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd
Part of the Mathematics Commons, and the Statistical Models Commons

Recommended Citation
https://digitalcommons.usu.edu/etd/6856

This Dissertation is brought to you for free and open access by the Graduate Studies at DigitalCommons@USU. It has been accepted for inclusion in All Graduate Theses and Dissertations by an authorized administrator of DigitalCommons@USU. For more information, please contact dylan.burns@usu.edu.
NOVEL STATISTICAL MODELS FOR QUANTITATIVE SHAPE-GENE ASSOCIATION SELECTION

by

Xiaotian Dai

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

Approved:

Guifang Fu, Ph.D. Adele Cutler, Ph.D.
Major Professor Committee Member

Chris Corcoran, Ph.D. Daniel Coster, Ph.D.
Committee Member Committee Member

Xiaojun Qi, Ph.D. Mark R. McLellan, Ph.D.
Committee Member Vice President for Research and Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2017
ABSTRACT

Novel Statistical Models for Quantitative Shape-Gene Association Selection

by

Xiaotian Dai, Doctor of Philosophy
Utah State University, 2017

Major Professor: Guifang Fu, Ph.D.
Department: Mathematics and Statistics

The primary goal of this dissertation is to develop novel statistical models for genome-wide association studies (GWAS) with non-single-valued biological phenotypes. In particular, we focus on quantitative shape-gene association selection. Most of the current GWAS research only investigated the associations between one individual biological phenotype and a large number of genes, while we see more and more non-single-valued biological phenotypes that can only be quantified using multiple or even high-dimensional biological phenotypes. Shape is a typical example of non-single-valued biological phenotypes because its complexity can never be effectively represented by single-valued biological trait. As a result, other literature developed multivariate or functional shape descriptors to quantify biological shapes accurately. In this dissertation, we propose three statistical models to handle multivariate, functional, and multilevel functional phenotypes, with applications to biological shape data using different shape descriptors. To the best of our knowledge, there is no statistical model developed for multilevel functional phenotypes. Even though multivariate regressions have been well-explored and these approaches can be applied to genetic studies, we show that the model proposed in this dissertation can outperform other alternatives regarding variable selection and prediction through simulation examples and real data examples. Although
motivated ultimately by GWAS research, the proposed models aim to have a broader impact on large-scale machine learning problems with multivariate, functional, and multilevel functional responses.
PUBLIC ABSTRACT

Novel Statistical Models for Quantitative Shape-Gene Association Selection

Xiaotian Dai

Other research reported that genetic mechanism plays a major role in the development process of biological shapes. The primary goal of this dissertation is to develop novel statistical models to investigate the quantitative relationships between biological shapes and genetic variants. However, these problems can be extremely challenging to traditional statistical models for a number of reasons: 1) the biological phenotypes cannot be effectively represented by single-valued traits, while traditional regression only handles one dependent variable; 2) in real-life genetic data, the number of candidate genes to be investigated is extremely large, and the signal-to-noise ratio of candidate genes is expected to be very high. In order to address these challenges, we propose three statistical models to handle multivariate, functional, and multilevel functional phenotypes, with applications to biological shape data using different shape descriptors. To the best of our knowledge, there is no statistical model developed for multilevel functional phenotypes. Even though multivariate regressions have been well-explored and these approaches can be applied to genetic studies, we show that the model proposed in this dissertation can outperform other alternatives regarding variable selection and prediction through simulation examples and real data examples. Although motivated ultimately by genetic research, the proposed models can be used as general-purpose machine learning algorithms with far-reaching applications.
This work is dedicated to my beloved parents.
ACKNOWLEDGMENTS

This research was supported by National Science Foundation grant DMS-1413366 to Dr. Guifang Fu.
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>PUBLIC ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2 BAYESIAN VARIABLE SELECTION FOR MULTIVARIATE REGRESSION</td>
<td>3</td>
</tr>
<tr>
<td>2.1 INTRODUCTION</td>
<td>3</td>
</tr>
<tr>
<td>2.2 METHODOLOGY</td>
<td>6</td>
</tr>
<tr>
<td>2.2.1 Prior distributions and hyperparameters</td>
<td>6</td>
</tr>
<tr>
<td>2.2.2 Posterior distributions</td>
<td>8</td>
</tr>
<tr>
<td>2.2.3 Estimation</td>
<td>10</td>
</tr>
<tr>
<td>2.3 GENETIC APPLICATION</td>
<td>10</td>
</tr>
<tr>
<td>2.4 SIMULATION EXAMPLES</td>
<td>13</td>
</tr>
<tr>
<td>2.4.1 Simulation example 1</td>
<td>14</td>
</tr>
<tr>
<td>2.4.2 Simulation example 2</td>
<td>17</td>
</tr>
<tr>
<td>2.5 REAL DATA ANALYSIS</td>
<td>19</td>
</tr>
<tr>
<td>2.5.1 Rice shape data</td>
<td>19</td>
</tr>
<tr>
<td>2.5.2 Flowering time data</td>
<td>22</td>
</tr>
<tr>
<td>2.6 Discussion</td>
<td>23</td>
</tr>
<tr>
<td>3 FUNCTIONAL RANDOM FORESTS FOR CURVE RESPONSE</td>
<td>25</td>
</tr>
<tr>
<td>3.1 INTRODUCTION</td>
<td>25</td>
</tr>
<tr>
<td>3.2 METHODOLOGY</td>
<td>29</td>
</tr>
<tr>
<td>3.2.1 Functional data analysis</td>
<td>30</td>
</tr>
<tr>
<td>3.2.2 Functional regression tree</td>
<td>34</td>
</tr>
<tr>
<td>3.2.3 Random forests</td>
<td>38</td>
</tr>
<tr>
<td>3.2.4 Tuning parameters</td>
<td>40</td>
</tr>
<tr>
<td>3.3 SIMULATION STUDIES</td>
<td>40</td>
</tr>
<tr>
<td>3.3.1 Simulation 1: a single-causative-predictor design</td>
<td>41</td>
</tr>
<tr>
<td>3.3.2 Simulations 2-4: a two-interactive-causative-predictors design</td>
<td>42</td>
</tr>
<tr>
<td>3.3.3 Simulations 5: a three-interactive-causative-predictors design</td>
<td>44</td>
</tr>
<tr>
<td>3.3.4 Simulations 6: a two-correlated-causative-predictors design</td>
<td>44</td>
</tr>
<tr>
<td>3.3.5 Simulation results</td>
<td>45</td>
</tr>
<tr>
<td>3.4 REAL DATA ANALYSIS</td>
<td>49</td>
</tr>
<tr>
<td>3.5 DISCUSSION</td>
<td>54</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Simulation results for Setting 1 of Example 1.</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Simulation results for Setting 2 of Example 1.</td>
<td>15</td>
</tr>
<tr>
<td>2.3 Simulation results for Setting 3 of Example 1.</td>
<td>16</td>
</tr>
<tr>
<td>2.4 Simulation results for Example 2.</td>
<td>19</td>
</tr>
<tr>
<td>2.5 Results of real data analysis on rice shape data</td>
<td>22</td>
</tr>
<tr>
<td>2.6 Results of real data analysis on flowering time data using five-fold cross-validation</td>
<td>22</td>
</tr>
<tr>
<td>3.1 The odds ratio table for Epistasis design in simulations 2, 3, and 4</td>
<td>43</td>
</tr>
<tr>
<td>3.2 Results of six simulation designs</td>
<td>46</td>
</tr>
<tr>
<td>3.3 The MISE results of five fold cross-validation under possible values chosen for two tuning parameters</td>
<td>51</td>
</tr>
<tr>
<td>4.1 Coefficients and functions used in Simulation I and Simulation II</td>
<td>76</td>
</tr>
<tr>
<td>4.2 Variable selection performance on simulated datasets</td>
<td>77</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2.1</td>
<td>The five template shapes used in Example 2</td>
</tr>
<tr>
<td>2.2</td>
<td>The Euclidean distances of 272 observations in the flowering time data are approximated in a two dimensional space, and these observations are grouped based on nine SNPs selected by the BMVS method, with black dots representing genotype aa and red dots representing genotype AA.</td>
</tr>
<tr>
<td>3.1</td>
<td>Three mean shapes serving as the ‘truth’ for simulation studies</td>
</tr>
<tr>
<td>3.2</td>
<td>Illustration of the prediction for one replication in Simulation 2: The simulated response samples (transparent green lines on the first row) and predicted response curves (transparent green lines on the the second row) output from the FunFor approach. Each column corresponds to a different genotype group. The black lines demonstrate the three true shapes used for producing Simulation 2.</td>
</tr>
<tr>
<td>3.3</td>
<td>Same information as the Figure 3.2, but converting curves into shapes</td>
</tr>
<tr>
<td>3.4</td>
<td>The shape description process: (A), (B), &amp; (C) are to recognize a shape from an image; (D) &amp; (E) are to transform a shape to a curve.</td>
</tr>
<tr>
<td>3.5</td>
<td>Choice of the number of PC kept: the solid line indicates the percentage of variation explained by each PC; The dashed line indicates the cumulative percentage of variation explained by the first few PCs.</td>
</tr>
<tr>
<td>3.6</td>
<td>Prediction variable importance measures of 104 genetic markers. The highest VIMs are highlighted with red dots.</td>
</tr>
<tr>
<td>3.10</td>
<td>Prediction of new shapes when the genetic predictors are given. The solid lines are the original observed shape, and the dot lines are the predicted shape output from the FunFor approach.</td>
</tr>
<tr>
<td>3.7</td>
<td>The mean predicted shape curves output from FunFor approach across all individuals belong to each genotype of the three most important markers ORPM_190, GCPM_1812, and U50206. Black curve for AA, red curve for Aa, and blue curve for aa.</td>
</tr>
<tr>
<td>3.8</td>
<td>The mean predicted shape curves output from FunFor approach across all individuals belong to each genotype of the three least important markers Pe_5, Pe_8, and GCPM_1941. Black curve for AA, red curve for Aa, and blue curve for aa.</td>
</tr>
</tbody>
</table>
3.9 The first two Eigenfunctions output from FunFor approach across all individuals belong to each genotype of the three most important markers $ORPM_{190}$, $GCPM_{1812}$, and $U50206$. Black curve for AA, red curve for Aa, and blue curve for aa. Solid line for the first Eigenfunction $\hat{v}_1(t)$ and the dash line for the second Eigenfunction $\hat{v}_2(t)$.

4.1 Sample images of $Populus euphratica$ leaves

4.2 Additive effects of selected markers

4.3 Dominant effects of selected markers

4.4 Visualizing subject effects of causative markers

4.5 Visualizing subject effects of non-causative markers
CHAPTER 1
INTRODUCTION

This dissertation is motivated by genome-wide association studies (GWAS) with non-single-valued biological phenotypes, which is supported by a National Science Foundation grant to investigate the associations between biological shapes and genetic variants. Three novel statistical models are proposed to handle challenging genetic association or selection problems with applications to different shape descriptors in three articles, which serve as three relatively independent chapters in this dissertation.

In Chapter 2 (the first article) we propose a novel Bayesian approach to perform variable selections in multivariate regression settings that consider the relationship between a multivariate response vector like a multivariate shape descriptor and a large set of predictors ($p >> n$ or $p = e^{o(n)}$) like genetic variants covering the entire genome. Quantitative genetic studies investigate associations between multivariate phenotypes and large sets of single-nucleotide polymorphisms (SNPs) [1–3]. Multivariate regression is the most traditional statistical approach used to model the relationship between a multivariate response and a set of predictors. But it fails to work when the number of predictors is larger than the sample size ($p > n$). When the number of candidate predictors is much larger than the sample size (e.g., [2,4]) and variable selection is not performed, overfitting and low prediction accuracy may become problematic. Therefore, we aim to develop the Bayesian multivariate variable selection (BMVS) method to identify causative genes in a genetic association study with multivariate shape descriptors or other multivariate phenotypes.

In Chapter 3 (the second article) we propose a new approach called functional random forests (FunFor), which facilitates an extension of the traditional random forests methodology (accommodating a single univariate response) to provide repeated measures, functional, longitudinal, or curve response settings. The inputs of the FunFor method will be a response curve that repeatedly measures the same variable at multiple time or location points
and a large number of multiple predictors for a given individual. FunFor outputs the predicted response curve for each individual along with the importance rank of each predictor according to its association strength with the response curve. FunFor keeps many of the agreeable properties of traditional random forests: effectively modeling both the linear and complex nonlinear relationships between the curve response and the predictor; producing a joint model rather than a marginal model by modeling multiple predictors simultaneously; capturing the intricate higher-order interactions and also accounting for correlations among predictors; demonstrating feasibility for binary, continuous, and categorical predictors; embracing a nonparametric approach without assuming any specific model structure, distribution, or data type; and requiring few tuning parameters. In this chapter, we illustrate the performance of FunFor approach through six novel simulation designs and one real data analysis from the genetic shape applications. To the best of our knowledge, we are the first to invent a series of novel simulation designs to connect shape response with gene and environmental factors from different scenarios.

In Chapter 4 (the third article) we introduce the multilevel functional GWAS (mfGWAS) method, a novel statistical approach, to handle both multilevel functional responses and high-dimensional predictor space. The mfGWAS method borrows the strength from the statistical methods proposed by Di et al. (2009) [5] and Crainiceanu et al. (2009) [6]. This approach models multilevel functional data through MFPCA, a combination of functional principal component analysis (FPCA) [7–10] and the multilevel functional mixed effects model [11]. The parameters used in the mfGWAS method can be estimated through Markov chain Monte Carlo (MCMC) sampling. This new method can serve not only as a domain-specific method for GWAS problems but also a standard approach to variable selection in regressions with multilevel functional responses. In the article, we test our new method through simulation studies. We also apply the new method to a real multilevel functional data example. The data presented contain leaf shape measurements of *Populus euphratica*. 
CHAPTER 2
BAYESIAN VARIABLE SELECTION FOR MULTIVARIATE REGRESSION

2.1 INTRODUCTION

Multivariate response variables have been attracting a lot of attention in various disciplines, including ecology, geology, psychology, genetics, and others. For example, De’Ath (2002) [12] modeled the relationship between multi-species and environmental factors. Tsitsika et al. (2009) [13] analyzed the important predictors affecting internet access characteristics and addiction score for adolescents. Quantitative genetic studies investigate associations between multiple biological traits and large sets of single-nucleotide polymorphisms (SNPs) [1–3]. Multivariate regression is the most traditional statistical approach used to model the relationship between a multivariate response and a set of predictors. But it fails to work when the number of predictors is larger than the sample size ($p > n$). When the number of candidate predictors are much larger than the sample size (e.g., [2, 4]) and variable selection is not performed, overfitting and low prediction accuracy may become problematic.

Variable selection approaches have been well-established in univariate regression settings where there is only one response variable. The traditional stepwise variable selection approaches have proven to be computationally intensive and unstable [14, 15]. The well-known penalization approaches have been widely used for variable section purposes, having the assumption that the model is sparse and the active predictors should have nonzero coefficients. These include the bridge regression [16], the least absolute shrinkage and selection operator (LASSO, [17]), adaptive LASSO [18], and the smoothly clipped absolute deviation (SCAD, [15]), to name a few. Tibshirani (1996) [17] proposed to use a $L_1$ penalty function $p_\lambda(|\beta|) = \lambda|\beta|$ on the least square estimates of regression coefficients $\beta$’s (LASSO), where $\lambda$ is a penalization parameter. Li et al. (2011) [19] applied Bayesian LASSO to the genome-
wide association studies (GWAS) field. Rothman et al. (2010) [20] used LASSO penalty to reduce model size while accounting for correlated errors. Despite LASSO’s popularity, its theoretical properties are controversial. Some research claimed that the selection consistency of LASSO cannot be established and its performance for high-dimensional predictors is not well explored [15,18,21]. Frank and Friedman (1993) [16] proposed to use a $L_q$ penalty function $p_\lambda(|\beta|) = \lambda |\beta|^q$ (bridge regression), but it fails to produce a sparse solution when $q > 1$ [15].

Consider a univariate regression model here: $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$. Bayesian methods usually introduce a latent variable vector $(Z_i)$ for each predictor, in which $Z_i = 0$ or $1$, $i = 1,\ldots,p$, to indicate whether or not the $i^{th}$ predictor should be included into the final model. The final model can be selected based on the posterior probability of the $Z_i$’s after generating chains of Gibbs samplers. Mitchell and Beauchamp (1988) [22] proposed the use of spike and slab priors on regression coefficients $\beta$’s. When $Z_i = 0$, the $i^{th}$ predictor is considered inactive and $\beta_i$ will have a prior distribution with concentrated probability mass around zero, which is referred to as the spike prior. When $Z_i = 1$, the $i^{th}$ predictor is considered active in the model and $\beta_i$ will have a prior distribution with diffusing probability density, which is referred to as the slab prior. George and McCulloch (1993) [23] proposed the stochastic search variable selection (SSVS) approach to select “promising” subsets of predictors. Their framework used a normal distribution with zero mean and a small fixed variance as the spike prior, and another normal distribution with zero mean and a large fixed variance as the slab prior. Guan and Stephens (2011) [24] applied the spike and slab priors to large-scale genetic selection problems. Most recently, Narisetty and He (2014) [21] questioned the theoretical selection consistency property for all these approaches because they used fixed hyperparameters for spike and slab priors. Specifically, Narisetty and He (2014) [21] showed the SSVS does not guarantee selection consistency since it uses normal distributions with fixed variances for the spike and slab priors. Narisetty and He (2014) [21] presented the first work changing the fixed variances of the spike and slab priors into sample-size-dependent shrinking and diffusing priors. The strong selection consistency
and the asymptotic connections with $L_0$ penalty functions were proven to be satisfied in their new variable selection approach. Additionally, these desirable properties hold for ultrahigh dimensional cases, including situations where the number of predictors is in the exponential form of the sample size (i.e., $p = e^{o(n)}$) [21].

Compared to the aforementioned univariate variable selection approaches, variable selection approaches for multivariate regressions are undeveloped. Bedrick and Tsai (1994) [25] pioneered model selection criteria for multivariate regression. Breiman and Friedman (1997) [26] proposed the Curds and Whey algorithm to enhance the prediction accuracy of multivariate regressions by taking advantages of correlations between response variables. Their algorithm fitted a full model using $L_2$ shrinkage penalty, which cannot shrink unimportant predictors to produce a sparse model [15]. Brown et al. (1998) [27] developed a Bayesian method for multivariate variable selection and prediction using a latent vector $Z_{p \times 1}$ to indicate the inclusions and exclusions of $p$ predictors. The final model can be selected by finding the latent vector with the highest posterior probability. A drawback of the algorithm, however, is that it is not computationally feasible when $p$ is greater than 20 since it involves $2^p$ choices [27].

In this article, we propose a novel Bayesian multivariate variable selection (BMVS) approach to perform variable selections in multivariate regression settings that consider the relationship between a multivariate response vector and a large set of high-dimension predictors ($p \gg n$ or $p = e^{o(n)}$). BMVS is a natural extension of the work of Narisetty and He (2014) [21], with the univariate response of the latter being replaced by a multivariate response and the covariance structures of multiple response variables accurately estimated in one integrated framework. All unknown parameters, including the covariance structure of multiple response variables and predictor coefficients, are estimated through a fast-updating Markov chain Monte Carlo (MCMC) algorithm. We demonstrate that the proposed BMVS method performed well in two different simulation designs. In the first simulation, we generated continuous predictors and a continuous multivariate response vector by using the standard statistical multivariate regression approach. In the second simulation, we designed
a shape response to connect with a set of single-nucleotide polymorphism (SNP) genotypes. Based on these two simulations, we verify that the proposed BMVS method works well for both categorical and continuous predictors. In this paper, we only consider multivariate continuous response. We also apply the proposed statistical method to two real world genetic association problems in order to demonstrate its variable selection and prediction accuracy. We also compare the BMVS model to other approaches that have been widely used in multivariate regression literature.

2.2 METHODOLOGY

2.2.1 Prior distributions and hyperparameters

Consider the multivariate regression model:

\[ Y = X\beta + E, \]

where \( Y \in \mathbb{R}^{n \times q} \) is the response matrix, \( X \in \mathbb{R}^{n \times p} \) is the predictor matrix, \( \beta \in \mathbb{R}^{p \times q} \) is the parameter coefficient matrix, and \( E \in \mathbb{R}^{n \times q} \) is the error matrix with each row following an independent and identical distribution \( N_q(0, \Sigma_Y) \). Here \( p \) denotes the number of predictors, \( q \) denotes the number of response variables, and \( n \) denotes the number of observations. Let \( Y_m \) and \( X_m \) denote the \( m^{th} \) row vector in \( Y \) and \( X \), respectively, where \( m = 1, \ldots, n \). The \( i^{th} \) row vector of \( \beta \), \( \beta_i \), having a \( 1 \times q \) dimension, corresponds to the coefficient vector of the \( i^{th} \) predictor, where \( i = 1, \ldots, p \). For selection purpose, we assume that only active predictors have non-zero coefficients and that the true model is sparse (i.e., active/non-active ratio is small). Both the regression coefficients \( \beta \) and the covariance matrix of response variables \( \Sigma_Y \) are unknowns that need to be estimated.

For each of the predictors, we introduce a latent binary variable \( Z_i \) to indicate whether or not it should be included in the model. An active predictor corresponds to \( Z_i = 1 \) and a non-active predictor corresponds to \( Z_i = 0 \), with distribution \( p(Z_i = 1) = 1 - p(Z_i = 0) = \phi \). Intuitively, the priors of \( \phi \) should be given according to the signal-to-noise ratio of the
candidate predictors. However, this ratio is unknown in real datasets. We empirically verify through simulations that the choice of priors for $\phi$ has trivial impacts on the final model.

Based on this, the priors of the regression coefficients $\beta$ can be given as

$$
\beta_i | (\sigma_{\beta}^2, Z_i = 1) \sim N_q(0, \sigma_{\beta}^2 \tau_1^2 I_q),
$$

$$
\beta_i | (\sigma_{\beta}^2, Z_i = 0) \sim N_q(0, \sigma_{\beta}^2 \tau_0^2 I_q),
$$

(2.2)

$$
\sigma_{\beta}^2 \sim IG(\alpha_1, \alpha_2),
$$

where $I_q$ is a $q$ by $q$ identity matrix, and $\sigma_{\beta}^2$ is a scalar parameter following an inverse gamma distribution with a shape parameter $\alpha_1$ and a scale parameter $\alpha_2$. $\alpha_1$ and $\alpha_2$ are given small values so that the priors are essentially noninformative.

The sample-size-dependent parameters $\tau_1^2$ and $\tau_0^2$ control the priors of $\beta_i$ to make it either a slab prior or a spike prior. If $Z_i = 0$, $\beta_i$ will have a spike prior distribution with concentrated probability mass around zero. If $Z_i = 1$, $\beta_i$ will have a slab prior distribution with diffusing probability density. Inspired by Narisetty and He (2014), we design the prior hyperparameters of $\tau_0^2$ and $\tau_1^2$ as

$$
\tau_0^2 = \frac{\hat{\sigma}_Y^2}{10n},
$$

$$
\tau_1^2 = \hat{\sigma}_Y^2 \max \left( \frac{(pq)^2}{100n}, \log n \right),
$$

(2.3)

where $\hat{\sigma}_Y^2$ is the average sample variances of response variables $\hat{\sigma}_Y^2 = \sum_{k=1}^q \hat{\sigma}_{Y_k}^2 / q$. It asymptotically satisfies the conditions that $\tau_1^2 \to \infty$ and $\tau_0^2 \to 0$ as $n \to \infty$.

A natural choice for the prior distribution of covariance matrix $\Sigma_Y$ would be a probability distribution defined on positive-definite matrices with the conjugate property. According to Dawid (1981) [28], directly modeling the matrix variate has the advantage of preserving the matrix structures without breaking the matrices down into multiple row or column vectors that dramatically increase complexity and computation costs. As such, we choose the inverse Wishart distribution:

$$
\Sigma_Y \sim IW(\nu, \Lambda),
$$

(2.4)
where $\nu$ is the degrees of freedom, and $\Lambda$ is a positive-definite scale matrix. The default values for these prior hyperparameters are $\nu = q + 1$ and $\Lambda = I_q$ [29]. Due to the conjugacy of the inverse Wishart distribution, $\Sigma_Y$ can be conveniently estimated along with other unknown parameters.

### 2.2.2 Posterior distributions

The joint posterior distribution containing all unknown parameters can be derived as:

$$f(\beta_i, Z_i, \sigma_\beta^2, \Sigma_Y | Y)$$

$$\propto f(Y | \beta_i, Z_i, \sigma_\beta^2, \Sigma_Y) \times \prod_{i=1}^{p} f(\beta_i) \times \prod_{i=1}^{p} f(Z_i) \times f(\sigma_\beta^2) \times f(\Sigma_Y)$$

$$\propto \left(\frac{1}{\sqrt{|\Sigma_Y|}}\right)^n \exp\left(-\frac{1}{2} \sum_{m=1}^{n} (Y_m - X_m \beta) \Sigma_Y^{-1} (Y_m - X_m \beta)'\right)$$

$$\times \prod_{i=1}^{p} \left((1 - \phi) \frac{1}{\sqrt{|\sigma_\beta^2 \tau_1 I_q|}} \exp(-\frac{1}{2} \beta_i (\sigma_\beta^2 \tau_0 I_q)^{-1} \beta_i')\right)^{1-Z_i}$$

$$\times \sigma_\beta^{-2(\alpha_1+1)} \exp\left(-\frac{\alpha_2}{\sigma_\beta^2}\right) \times |\Sigma_Y|^{-(\nu+q+1)/2} \exp\left(tr(\Lambda \Sigma_Y^{-1})\right).$$

The posterior distribution of $\beta_i$, the coefficients of the $i$th predictor, is

$$f(\beta_i | Z_i, \sigma_\beta^2, \Sigma_Y, Y)$$

$$\propto f(Y | \beta_i, Z_i, \sigma_\beta^2, \Sigma_Y) \times f(\beta_i)$$

$$\propto \exp\left(-\frac{1}{2} \sum_{m=1}^{n} (Y_m - \mu_m(-\beta_i) - X_{im} \beta_i) \Sigma_Y^{-1} (Y_m - \mu_m(-\beta_i) - X_{im} \beta_i)' - \frac{1}{2} \beta_i (\sigma_\beta^2 \tau_1 I_q)^{-1} \beta_i'\right)$$

$$\propto \exp\left(\beta_i ((\sigma_\beta^2 \tau_1 I_q)^{-1} + \sum_{m=1}^{n} X_{mi}^2 \Sigma_Y^{-1}) \beta_i' - 2 \sum_{m=1}^{n} (Y_m - \mu_m(-\beta_i) \Sigma_Y^{-1} (X_{mi} \beta_i))\right)$$

$$\propto N_q(\mu_{\beta_i}, \Sigma_{\beta_i}),$$
where \( \mu_{(-\beta_i)} = X\beta - X_i\beta_i \). We also have

\[
\mu_{\beta_i} = \left( \sigma^2_{\beta_i} I_q \right)^{-1} + \sum_{m=1}^{n} X_m^2 \Sigma_Y^{-1} \left( \sum_{m=1}^{n} (Y_m - \mu_{m(-\beta_i)}) \Sigma_Y^{-1} X_{mi} \right),
\]

\[
\Sigma_{\beta_i} = \left( \sigma^2_{\beta_i} I_q \right)^{-1} + \sum_{m=1}^{n} X_m^2 \Sigma_Y^{-1},
\]

where \( X_{mi} \) is the \( m \)th observation of the \( i \)th predictor.

The posterior distribution of \( \sigma^2_\beta \) is

\[
f(\sigma^2_\beta | \beta, \Sigma_Y, Y) \\
\propto f(Y | \beta_i, Z_i, \sigma^2_\beta, \Sigma_Y) \times f(\sigma^2_\beta) \\
\propto \sigma^{-2(\alpha_1+1)} \exp \left( -\frac{\alpha_2}{\sigma^2_\beta} \right) \\
\times \prod_{i=1}^{p} \frac{1}{\sqrt{|\sigma^2_\beta \tau^2_i I_q|}} \exp \left( -\frac{1}{2} \beta_i (\sigma^2_{\beta_i} I_q)^{-1} \beta_i' \right)^{1-Z_i} \left( \frac{1}{\sqrt{|\sigma^2_\beta \tau^2_i I_q|}} \exp \left( -\frac{1}{2} \beta_i (\sigma^2_{\beta_i} I_q)^{-1} \beta_i' \right) \right)^{Z_i} \\
\propto IG\left( \alpha_1 + pq, \alpha_2 + \sum_{i=1}^{p} \beta_i (\tau^2_i I_q)^{-1} \beta_i' \right).
\]

Then \( Z_i \) can be estimated using

\[
P(Z_i | \beta_i, \sigma^2_\beta) \propto \int \left\{ \exp \left( -\frac{n}{2\sigma^2_\beta} (b - \beta_i)^2 \right) \\
\times \left( (1 - \phi) \frac{1}{\sqrt{|\sigma^2_\beta \tau^2_i I_q|}} \exp \left( -\frac{1}{2} b (\sigma^2_{\beta_i} I_q)^{-1} b' \right) \right)^{1-Z_i} \left( \phi \frac{1}{\sqrt{|\sigma^2_\beta \tau^2_i I_q|}} \exp \left( -\frac{1}{2} b (\sigma^2_{\beta_i} I_q)^{-1} b' \right) \right)^{Z_i} \right\} \text{d}b.
\]

Therefore, we have

\[
P(Z_i = 1 | \beta_i, \sigma^2_\beta) = \frac{\phi_{q}(\beta_i; \mathbf{0}, \sigma^2_{\beta_i} \tau^2_i I_q)}{\phi_{q}(\beta_i; \mathbf{0}, \sigma^2_{\beta_i} \tau^2_i I_q) + (1 - \phi) \theta_{q}(\beta_i; \mathbf{0}, \sigma^2_{\beta_i} \tau^2_0 I_q)},
\]

where \( \theta_{q}(\cdot) \) is the probability density function of a \( q \)-dimensional multivariate normal distribution.

Finally, the posterior distribution of \( \Sigma_Y \) is
\[ f(\Sigma_{Y}|\beta, \sigma_{\beta}^{2}, Y) \]
\[ \propto f(\Sigma_{Y}, \beta_{i}, Z_{i}, \sigma_{\beta}^{2}|Y) \times f(\Sigma_{Y}) \]
\[ \propto \left( \frac{1}{\sqrt{|\Sigma_{Y}|}} \right)^{n} \exp\left( -\frac{1}{2} \sum_{m=1}^{n} (Y_{m} - X_{m}\beta)\Sigma_{Y}^{-1}(Y_{m} - X_{m}\beta)' \right) \]
\[ \times |\Sigma_{Y}|^{-(\nu+q+1)/2}\exp\left( tr(\Lambda_{\Sigma}Y^{-1}) \right) \]
\[ \propto IW(n + \nu, \Lambda + \sum_{m=1}^{n} (Y_{m} - X_{m}\beta)(Y_{m} - X_{m}\beta)'). \]

2.2.3 Estimation

Since all posterior distributions have standard forms, we use Gibbs samplers to simulate and estimate them. The selection of the final model is based on the marginal posterior probabilities of \( Z_i \)'s. A higher posterior probability \( p(Z_i = 1|\beta_i, \sigma_{\beta}^{2}) \) indicates stronger associations between the corresponding predictor and the multivariate response. Therefore, we rank predictors by their posterior probabilities \( P(Z_i = 1|Y, X) \). To determine how many predictors to keep and identify the optimal model, we use the multivariate corrected Akaike information criterion (AICc) [25]. The multivariate AICc puts a higher penalty on model size than the conventional AIC, and is more suitable for high-dimensional data with sparse structures; see Bedrick and Tsai (1994) [25].

2.3 GENETIC APPLICATION

In genetic association or genetic selection applications, it is common to have a large set of single-nucleotide polymorphisms (SNPs; \( p \gg n \)) and multiple response variables (\( q > 1 \)). Similar to many other categorical features, the genotypes of active SNPs may not have ordinal effects on response variables. To accurately model the association between SNPs and the response vector, we code the original genotypes into two indicator matrices \( X_{a}^{n \times p} \) (additive effect) and \( X_{d}^{n \times p} \) (dominant effect).
The additive effect is the mean difference on phenotypic values when substituting one allele for the other one, and the dominant effect is the mean difference on phenotypic values between heterozygote and homozygotes.

Equation (1) can be rewritten as

\[ Y = \mathbf{1}\beta_0 + \mathbf{X}^a\beta^a + \mathbf{X}^d\beta^d + \mathbf{E}, \]

where \( \mathbf{1} \) is an \( n \times 1 \) matrix of 1’s, and \( \beta_0 \in \mathbb{R}^{1 \times q} \) is the baseline effect or intercept. \( \beta^a \) and \( \beta^d \) are the regression coefficients of the additive and dominant parameters, respectively. A genotype is treated as active if either its additive or dominant coefficient (or both) is important. All other notations and structures are the same as Equation (1).

We introduce latent variables \( Z^a_i \) and \( Z^d_i \) for \( \beta^a_i \) and \( \beta^d_i \), respectively. Similarly, the posterior distributions for all unknown parameters contained in Equation (5) can be derived
as

\[ f(\beta_0|\beta^a, \beta^d, \sigma^2_\beta, \Sigma_Y, Y) \propto N_q(\mu_{\beta_0}, \Sigma_{\beta_0}), \]

\[ f(\beta_i^a|Z_i^a, \sigma^2_\beta, \Sigma_Y, Y) \propto N_q(\mu_{\beta_i^a}, \Sigma_{\beta_i^a}), \]

\[ P(Z_i^a = 1|\beta_i^a, \sigma^2_\beta) = \frac{\phi \theta_q(\beta_i^a; 0, \sigma^2_\beta^2 I_q)}{\phi \theta_q(\beta_i^a; 0, \sigma^2_\beta^2 I_q) + (1 - \phi) \theta_q(\beta_i^a; 0, \sigma^2_\beta^2 I_q)}, \]

\[ f(\beta_i^d|Z_i^d, \sigma^2_\beta, \Sigma_Y, Y) \propto N_q(\mu_{\beta_i^d}, \Sigma_{\beta_i^d}), \]

\[ P(Z_i^d = 1|\beta_i^d, \sigma^2_\beta) = \frac{\phi \theta_q(\beta_i^d; 0, \sigma^2_\beta^2 I_q)}{\phi \theta_q(\beta_i^d; 0, \sigma^2_\beta^2 I_q) + (1 - \phi) \theta_q(\beta_i^d; 0, \sigma^2_\beta^2 I_q)}, \]

\[ f(\sigma^2_\beta|\beta^a, \beta^d, \Sigma_Y, Y) \propto IG\left(\alpha_1 + (2p + 1)q, \right. \]

\[ \left. \alpha_2 + \sum_{m=1}^p (\beta^a_i^2 \tau^2_i I_q)^{-1} \beta^{a^t}_i + \sum_{m=1}^p (\beta^d_i^2 \tau^2_i I_q)^{-1} \beta^{d^t}_i + \beta_0 (\tau^2 I_q)^{-1} \beta_0 \right), \]

\[ f(\Sigma_Y|\beta^a, \beta^d, \sigma^2_\beta, Y) \propto IW\left(n + \nu, \right. \]

\[ \Lambda + \sum_{m=1}^n (Y_m - \beta_0 - X_m \beta^a - X_m \beta^d)(Y_m - \beta_0 - X_m \beta^a - X_m \beta^d)' \), \]

where

\[ \mu_{\beta_0} = \left( \Sigma_{\beta_0}^{-1} + \sum_{m=1}^n X_m^2 \Sigma_Y^{-1} \right)^{-1} \left( \sum_{m=1}^n (Y_m - \mu_{m(-\beta_0)}) \Sigma_Y^{-1} X_m \right)', \]

\[ \Sigma_{\beta_i} = \left( \Sigma_{\beta_0}^{-1} + \sum_{m=1}^n X_m^2 \Sigma_Y^{-1} \right)^{-1}, \]

\[ \mu_{\beta_0} = \left( (\sigma^2_\beta^2 \tau^2_i I_q)^{-1} + \sum_{m=1}^n X_m^2 \Sigma_Y^{-1} \right)^{-1} \left( \sum_{m=1}^n (Y_m - \mu_{m(-\beta_0)}) \Sigma_Y^{-1} X_m \right)', \]

\[ \Sigma_{\beta_i} = \left( (\sigma^2_\beta^2 \tau^2_i I_q)^{-1} + \sum_{m=1}^n X_m^2 \Sigma_Y^{-1} \right)^{-1}, \]

\[ \mu_{\beta_0} = \left( (\sigma^2_\beta^2 \tau^2_i I_q)^{-1} + \sum_{m=1}^n X_m^2 \Sigma_Y^{-1} \right)^{-1} \left( \sum_{m=1}^n (Y_m - \mu_{m(-\beta_0)}) \Sigma_Y^{-1} X_m \right)', \]

\[ \Sigma_{\beta_i} = \left( (\sigma^2_\beta^2 \tau^2_i I_q)^{-1} + \sum_{m=1}^n X_m^2 \Sigma_Y^{-1} \right)^{-1}, \]
and $\Sigma_{\beta_0} = \sigma_\beta^2 \tau_1^2 I_q$. The choices of all hyperparameters stay the same as in Section 2.1, except that $\tau_1^2 = \sigma_\tau^2 \max \left( \frac{(2pq)^{q+1}}{100n}, \log n \right)$ because the dimensionality of the design matrix has doubled.

The genetic association or selection is performed in the following steps:

- Step 1: Rank all SNPs by the maximum posterior probabilities of the latent variables $Z_{ia}$ and $Z_{id}$

$$\max \left( P(Z_{ia} = 1|Y, X), P(Z_{id} = 1|Y, X) \right).$$

- Step 2: Use the multivariate corrected Akaike information criterion (AICc) to choose the optimal model size. A SNP’s dominant effects are included in the model by default if its additive effects are already included, and vice versa.

In addition to genetic data, Equation (5) is ready to be flexibly extended to more general cases containing many other continuous or categorical covariates. If genome-wide association studies (GWAS) with millions of SNPs were under consideration, we would first perform an initial screening on the candidate SNPs using the distance correlation (DC) sure independence feature screening proposed by Li et al. (2012) [30]. After the initial screening, we would keep a subset of candidate predictors of size $d = 200$ (those with the highest DC scores) to further analyze using the BMVS method.

### 2.4 SIMULATION EXAMPLES

We examine the performance of the BMVS method using two completely different simulation designs, and also compare the BMVS with other relevant multivariate variable selection methods. We fix sample size to be $n = 200$ and varies two levels of the number of predictors ($p = 500$ and $p = 1,000$). All the simulation results are calculated based on 100 replications. In each replication, we perform 1000 burn-in iterations followed by 2000 update iterations to estimate unknown parameters. After varying the initial values for $\phi$, we find that simulation results remained stable. So we set $\phi = 0.1$ for all simulations.
2.4.1 Simulation example 1

In Example 1 we perform three different simulation settings based on the standard statistical multivariate regression model described in Equation (1). The predictors in Example 1 are continuous variables.

- **Setting 1**: Generate each predictor independently from a standard normal distribution: \( X_{mi} \sim N(0,1) \). Set the first ten predictors to be active and generate their corresponding coefficients from a uniform distribution \( \beta \sim \text{Uniform}(1,3) \) (having moderate effects). Set the coefficients of all other non-causative predictors to be zero. Let the \( kj^{th} \) component of \( \Sigma_Y \) be \( \sigma_{kj} = 0.5|j-k| \), where \( 1 \leq k \leq q \) and \( 1 \leq j \leq q \). Finally, generate \( Y \) using Equation (1). Set the number of response variables to be \( q = 5 \).

- **Setting 2**: Maintain all other settings from Setting 1, but increase the difficulty level by setting the number of response variables at \( q = 30 \).

- **Setting 3**: Increase the difficulty level of Setting 1 through two adjustments: 1) introduce correlations among predictors (instead of the independence outlined in Setting 1). The predictors are generated from a multivariate normal distribution: \( X_m \sim N_p(0, \Sigma_X) \). The \( kj^{th} \) component of \( \Sigma_X \) is \( \sigma_{kj} = 0.5|j-k| \), where \( 1 \leq k \leq p \) and \( 1 \leq j \leq p \). 2) Generate the coefficients of active predictors from a uniform distribution \( \beta \sim \text{Uniform}(0.5,0.8) \), which represents very weak signals that are not easily differentiated from the non-active candidates. Maintain all other settings from Setting 1.

We compare the BMVS method with three other variable selection methods: 1) canonical correlation analysis (CCA) coupled with Wilk’s Lambda test to test the significance of correlation coefficients; 2) multivariate LASSO (M-LASSO); 3) multivariate Elastic Net (M-EN), with parameters tuned by the multivariate AICc. We implement M-LASSO and M-EN using the R package *glmnet* [31] and CCA using the R package *CCA* [32].

The results of the three settings are summarized in Tables 2.1, 2.2, and 2.3, respectively. The true model is denoted by \( T \), and the selected model is denoted by \( \hat{T} \). We used five criteria to assess the performance of these approaches: the average of posterior probabilities of all
Table 2.1: Simulation results for Setting 1 of Example 1.

<table>
<thead>
<tr>
<th>mpp _1</th>
<th>mpp _0</th>
<th>P(T = T)</th>
<th>P(T \supset T)</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMVS</td>
<td>0.999</td>
<td>5.234 \times 10^{-14}</td>
<td>0.990</td>
<td>0.990</td>
</tr>
<tr>
<td>CCA</td>
<td>0.390</td>
<td>0.960</td>
<td>0.085</td>
<td></td>
</tr>
<tr>
<td>M-LASSO</td>
<td>0.000</td>
<td>1.000</td>
<td>0.949</td>
<td></td>
</tr>
<tr>
<td>M-EN</td>
<td>0.000</td>
<td>1.000</td>
<td>0.948</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Simulation results for Setting 2 of Example 1.

<table>
<thead>
<tr>
<th>mpp _1</th>
<th>mpp _0</th>
<th>P(T = T)</th>
<th>P(T \supset T)</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMVS</td>
<td>1.000</td>
<td>1.215 \times 10^{-15}</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>CCA</td>
<td>0.350</td>
<td>0.960</td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td>M-LASSO</td>
<td>0.000</td>
<td>1.000</td>
<td>0.950</td>
<td></td>
</tr>
<tr>
<td>M-EN</td>
<td>0.000</td>
<td>1.000</td>
<td>0.948</td>
<td></td>
</tr>
</tbody>
</table>

active and non-active predictors (mpp\_1 and mpp\_0, respectively), the power that the exact true model is selected (P(\hat{T} = T)), the power that the true model is contained in the selected model (P(\hat{T} \supset T)), and the false discovery rate (FDR).

Despite the fact that all approaches have equally high powers in \( p(\hat{T} \supset T) \), the BMVS method significantly outperforms all other approaches because it has the highest powers in selecting the exact true model while also attaining the lowest false discovery rates. These statements are uniformly true for all Tables 2.1 - 2.3. In particular, when the M-EN and M-LASSO approaches fail to pick the exact model (\( P(\hat{T} = T) = 0 \)), the BMVS achieve astonishingly high powers of nearly 100% (see Tables 2.1 & 2.2). Also, in order to guarantee that the true model would be contained in their selected sets, the M-EN and M-LASSO
Table 2.3: Simulation results for Setting 3 of Example 1.

<table>
<thead>
<tr>
<th></th>
<th>$m_{pp1}$</th>
<th>$m_{pp0}$</th>
<th>$P(\hat{T} = T)$</th>
<th>$P(\hat{T} \supset T)$</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 500$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMVS</td>
<td>0.994</td>
<td>0.001</td>
<td>0.750</td>
<td>0.940</td>
<td>0.022</td>
</tr>
<tr>
<td>DC+BMVS</td>
<td>0.880</td>
<td>0.950</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCA</td>
<td>0.300</td>
<td>0.990</td>
<td>0.116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-LASSO</td>
<td>0.000</td>
<td>1.000</td>
<td>0.949</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-EN</td>
<td>0.000</td>
<td>1.000</td>
<td>0.948</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 1000$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMVS</td>
<td>0.988</td>
<td>7.070 x 10^{-5}</td>
<td>0.820</td>
<td>0.880</td>
<td>0.006</td>
</tr>
<tr>
<td>DC+BMVS</td>
<td>0.950</td>
<td>0.970</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCA</td>
<td>0.270</td>
<td>1.000</td>
<td>0.134</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-LASSO</td>
<td>0.000</td>
<td>1.000</td>
<td>0.950</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-EN</td>
<td>0.000</td>
<td>1.000</td>
<td>0.948</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Approaches are forced aggressively expand their model sizes while overfitting, scoring a FDR of 95%. The BMVS, on the other hand, scores FDRs of less than 1% (see Tables 2.1 - 2.3). The CCA approach has a low FDR, but its power in $p(\hat{T} = T)$ is only about one third of that of the BMVS approach. Intuitively speaking, we know that it is much harder to achieve $\{\hat{T} = T\}$ than $\{\hat{T} \supset T\}$, and the difference between $\{\hat{T} = T\}$ and $\{\hat{T} \supset T\}$ is set by false discoveries. Setting 3 represents the most difficult case because its active signal is very weak and the correlations between predictors add further confounding factors. As a result, it is to be expected that the results of Setting 3 are a little worse than those of the first two settings. We demonstrate that the initial screening, before applying the BMVS (DC+BMVS), is able to slightly improve the results. The superior performances of the BMVS approach are robust when the number of predictors are doubled from $p = 500$ to $p = 1000$. This also further demonstrates that the choice of $\phi = 0.1$ works well for different signal-to-noise ratios. In addition, the results are very similar when we increase the number of response variables from $q = 5$ to $q = 30$. At minimum, this simulation demonstrates the potential of the BMVS approach to simultaneously fit 30-dimension multivariate response vectors that are large enough for real-life applications.
2.4.2 Simulation example 2

In Example 2, we design a new simulation to connect morphological shapes with SNPs in order to illustrate the power of the BMVS method in genetic association studies for multivariate response vectors. There are no such layouts in existing literature; we are the first to invent the design. We choose five bottle shapes as the true response template (shown in Figure 2.1). We use Elliptic Fourier descriptors (EFDs), proposed by Kuhl and Giardina (1982) [33], to describe the shape. Specifically, the outlines of these five bottles are extracted as a sequence of \(x\) and \(y\) boundary pixel coordinates. Then the \(x\)-\(y\) coordinates are approximated by Fourier transform and four coefficient vectors (denoted by \(a_N, b_N, c_N,\) and \(d_N\)) are estimated as follows:

\[
x(t) = a_0 + \sum_{n=1}^{N} \left( a_n \cos \frac{2n\pi t}{T} + b_n \sin \frac{2n\pi t}{T} \right)
\]

and

\[
y(t) = c_0 + \sum_{n=1}^{N} \left( c_n \cos \frac{2n\pi t}{T} + d_n \sin \frac{2n\pi t}{T} \right)
\]

where \(N\) is the number of harmonics to keep. In this example, we keep 16 harmonics and concatenate the Fourier coefficients into one long vector with a length of 64. Following the standard procedures of other shape analysis literature, we apply principal component analysis (PCA) to reduce dimensionality and keep the first five principal components (PCs) to represent each bottle shape [34] (More than 98% of the variation in Fourier coefficients can be explained by the first five PCs). The five PCs of five template bottle shapes are denoted by \(S_{5\times5} = [s_1, s_2, s_3, s_4, s_5]'\), and these multivariate vectors represent the shapes demonstrated in Figure 2.1, from left to right. For more details about EFD and its algorithm, see Kuhl and Giardina (1982) [33]. The EFD algorithm was implemented in the R package Momocs [35].
The simulation process is designed as follows: Generate a $5 \times 5$ matrix $\mathbf{u}$ with each component independently simulated from a normal distribution $u \sim N(0, 0.1)$, then compute the $5 \times 5$ coefficient matrix $\mathbf{\beta}_{\text{active}}$ by solving $\mathbf{s} = \mathbf{u} \mathbf{\beta}_{\text{active}}$; 2) Generate a $n \times p$ intermediate predictor matrix $\mathbf{U}$ following the same pattern as $\mathbf{u}$ in that each component is simulated from a normal distribution $U_{im} \sim N(0, 0.1)$; 3) Derive the multivariate response matrix $\mathbf{Y}$ by $\mathbf{Y}_{n \times q} = \mathbf{U}_{n \times p} \mathbf{\beta}_{p \times q}$, where the first five rows of $\mathbf{\beta}_{p \times q}$ are $\mathbf{\beta}_{\text{active}}$ (active) and the rest of the entries are zero (non-active); 4) Generate the SNPs (the predictor matrix in categorical type) $\mathbf{X}_{n \times p}$ from

$$X_{mi} = \begin{cases} AA, & U_{im} > c_2 \\ Aa, & c_1 \leq U_{im} \leq c_2 \\ aa, & U_{im} < c_1, \end{cases}$$

where $c_1$ is the first quartile and $c_2$ is the third quartile of normal distribution $N(0, 0.1)$. This guarantees the general rule of genetic literature by having proportions of AA and aa at 1/4 and that of Aa at 1/2. In this case, the first five SNPs truly affect the shape response; that is, the true model size $|T| = 5$.

Since the predictors are categorical, we compare the BMVS approach with three relevant variable selection methods: 1) MANOVA Wilk’s Lambda test with p-values adjusted by the Benjamini-Hochberg false discovery rate controlling procedure; 2) M-LASSO; 3) M-EN. As shown in Table 2.4, the false discovery rate of the BMVS method is drastically lower.
Table 2.4: Simulation results for Example 2.

<table>
<thead>
<tr>
<th></th>
<th>mpp1</th>
<th>mpp0</th>
<th>( P(T = T) )</th>
<th>( P(T \supset T) )</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>p = 500</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC+BMVS</td>
<td>0.860</td>
<td>1.000</td>
<td>0.033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MANOVA</td>
<td>0.520</td>
<td>1.000</td>
<td>0.098</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-LASSO</td>
<td>0.010</td>
<td>1.000</td>
<td>0.837</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-EN</td>
<td>0.000</td>
<td>1.000</td>
<td>0.974</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p = 1000</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC+BMVS</td>
<td>0.840</td>
<td>1.000</td>
<td>0.043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MANOVA</td>
<td>0.520</td>
<td>1.000</td>
<td>0.096</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-LASSO</td>
<td>0.000</td>
<td>1.000</td>
<td>0.963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-EN</td>
<td>0.000</td>
<td>1.000</td>
<td>0.974</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(with 3% versus 96%) than those of the M-LASSO and M-EN approaches. The MANOVA approach after applying the Benjamini-Hochberg correction also controls the false discovery rate very well (9%), but its power in finding the exact model is only half that of the BMVS's (52% versus 86%). We also notice that the BMVS's performance is stable when the number of predictors increase from 500 to 1000. We are not surprised to see that the results of Example 2 are inferior to those of Example 1 because here the response is connected with the intermediate predictor matrix \( U \) instead of the categorical matrix \( X \). Also, the data are not simulated directly from Equation (5) to be used for analysis, but instead go through a few indirect steps in order to connect multiple shapes with multiple SNPs.

2.5 REAL DATA ANALYSIS

In this section we apply our multivariate variable selection method to two real data examples in genome-wide association studies, each containing multiple response variables.

2.5.1 Rice shape data

In this example, we analyze a dataset related to shape of rice (\textit{Oryza sativa}). Iwata et al. (2015) [36] selected 179 rice accessions as representatives of the rice germplasm at the National Institute of Agrobiological Sciences Genebank. The rice accessions were genotyped with 3,254 SNPs, with missing genotypes imputed by Iwata et al. (2015) [36]. The SNPs do
not contain heterozygous genotype due to the inbreeding nature of *Oryza sativa* [37], which means the SNPs only have genotypes AA and aa. As a result, we only detect additive effects for the candidate SNPs. Though we use the same data, the study of Iwata et al. (2015) [36] mainly focused on shape prediction, which is different from our variable selection focus.

We directly use the shape descriptor, EFD coefficients, reported by Iwata et al. (2015) [36] as inputs for the shape response. Then we perform PCA and keep the first six PCs, which explain about 99% of the total variation in the shape descriptor coefficients (the first six PCs explain 94.66%, 1.85%, 1.00%, 0.57%, 0.52%, and 0.26% of the total variation respectively). In our analysis, we exclude the SNPs with minor allele frequencies of less than 5%, as per common practice in genetic literature. We perform a five-fold cross-validation on the entire observed data and calculate the following numerical assessments: 1) average size of selected models (|T|); 2) the Frobenius matrix norm of the difference between predicted phenotypes $\hat{Y}$ and observed phenotypes $Y$ ($\|Y - \hat{Y}\|_F$); 3) multivariate AICc of selected models (AICc). The BMVS method is again compared with M-LASSO and M-EN.

Table 2.5 shows the results of the five-fold cross-validation obtained using these three methods. Of the three methods, the BMVS method produces the lowest values of $|T|$ (8.40), $\|Y - \hat{Y}\|_F$ (0.8844), and AICc (776.5670). The M-LASSO approach selects an average model size of 139, which is 15 times larger than that of the BMVS, dramatically inflating both its prediction error and its AICc. In addition to comparing BMVS with standard multivariate variable selection approaches, we also compare it with its univariate counterpart to assess whether a multivariate approach performs better than a univariate approach when handling multiple responses. "Univariate counterpart" refers to the process of fitting each component of the multivariate response using univariate analysis and repeating the process six times. To make comparison fair and also minimize all other irrelevant factors, we used the Bayesian variable selection method proposed by Narisetty and He (2014) [21] from which the BMVS is extended. These "univariate counterpart" yield a predictor error $\|Y - \hat{Y}\|_F$ of 7.0076, which is almost ten times larger than that yielded by BMVS. This finding empirically confirms the claims of Breiman and Friedman (1997) [26].
Fig. 2.2: The Euclidean distances of 272 observations in the flowering time data are approximated in a two dimensional space, and these observations are grouped based on nine SNPs selected by the BMVS method, with black dots representing genotype aa and red dots representing genotype AA.
Table 2.5: Results of real data analysis on rice shape data

|           | $|T|$  | $||\hat{Y} - \tilde{Y}||_F$ | AICc      |
|-----------|-------|-----------------------------|-----------|
| BMVS      | 8.40  | 0.8844                      | 776.5670  |
| M-LASSO   | 139.60| 4.4521                      | 9915.9830 |
| M-EN      | 101.60| 1.7921                      | 3716.6670 |

Table 2.6: Results of real data analysis on flowering time data using five-fold cross-validation

|           | $|T|$  | $||\hat{Y} - \tilde{Y}||_F$ | AICc      |
|-----------|-------|-----------------------------|-----------|
| $\hat{T}_{BMVS}$ | 9     | 16.6450                      | -179.2434 |
| $\hat{T}_1$    | 70    | 64.7921                      | 227.8176  |
| $\hat{T}_2$    | 215   | 997.0443                     | 5294.179  |

2.5.2 Flowering time data

In this example, we consider the genome-wide association studies for three flowering time related phenotypes of 272 rice *Oryza sativa* accessions: flowering time (FT) at Aberdeen, FT ratio of Arkansas/Aberdeen, and FT ratio of Faridpur/Aberdeen. The dataset is publicly available at [http://ricediversity.org/](http://ricediversity.org/). These three phenotypes of the same rice accession are likely to be correlated, sharing genetic basis in some way. The joint multivariate model BMVS is therefore practically preferred over separate univariate models. The genotype data includes 36,901 SNPs that have been genotyped by Zhao et al. (2011) [37] and imputed by Iwata et al. (2015) [36]. For this high-dimensional setting, as in the simulation examples, we perform an initial screening on the candidate SNPs using the DC sure independence feature screening. After the initial screening, we keep a subset containing 200 candidate SNPs and then perform variable selection using the proposed BMVS method.

Zhao et al. (2011) [37] conducted genome-wide association studies for the same dataset but used linear mixed model to fit each individual phenotype separately. They reported significant SNPs ($p$-value $< 10^{-4}$) for each individual phenotype. Here, we compare three sets of results as follows: 1) the SNPs selected by the BMVS method, denoted by $\hat{T}_{BMVS}$; 2) the SNPs that are simultaneously significant for all three phenotypes as reported by Zhao et al. (2011) [37], denoted by $\hat{T}_1$ (the intersection set); 3) the SNPs that are significant for at least one phenotype as reported by Zhao et al. (2011) [37], denoted by $\hat{T}_2$ (the union set).
To perform a fair comparison, we apply multivariate linear regression model to perform five-fold cross-validation for these three sets of results (see Table 2.6). The BMVS method selects 9 SNPs, the intersection set $\hat{T}_1$ contains 70 SNPs, and the union set $\hat{T}_2$ contains 215 SNPs. Compared to $\hat{T}_{BMVS}$, $\hat{T}_2$ is astonishingly large, which partially illustrates the drawbacks of the separate univariate models [26]. BMVS outperform the other methods by using the smallest model size and achieving the best results. Specifically, it achieves both the smallest prediction error (16 versus 64 or 997) and the smallest AICc (-179 versus 227 or 5294). The negative sign of the AICc obtained by BMVS may cause some confusion, but the AICc rules indicate smaller results are always more desirable than larger ones. In addition, we further compare the sets of results by directly fitting a multivariate regression model that use all observations (without doing cross validation) and all corresponding SNPs contained in each set of result. 8 out of 9 selected SNPs are significant at the 0.05 level for $\hat{T}_{BMVS}$, while for $\hat{T}_1$ only 10 out of 70 selected SNPs are significant.

To visualize the associations between the 3-dimensional response vector and each of the nine SNPs selected by BMVS (i.e., set $\hat{T}_{BMVS}$), we use the multidimensional scaling approach to demonstrate the Euclidean distances of 272 observations in a two-dimensional space (see Figure 2.2) [38]. Each subplot represents a different genotype effect corresponding to each of the nine SNPs, with black dots for genotype aa and red dots for genotype AA. The two genotypes of each SNPs show quite different patterns for the variation of the response vector, as shown in Figure 2.2.

2.6 Discussion

Even though this article primarily focuses on variable selection, the real data examples have also shown that the BMVS method is very competitive in terms of prediction accuracy. If prediction is the ultimate goal, one can use other criteria such as cross-validation or bootstrap to substitute AICc during the variable selection process. In general, Bayesian methods are more computationally expensive than the frequentist counterparts due to their long iteratively updating process. However, the introduction of initial screening methods such as DC sure independence screening (i.e., DC+BMVS) dramatically reduces the com-
putational cost of the BMVS method and makes its use feasible for ultrahigh-dimensional data.

This characteristic of the BMVS method would be very useful for genome-wide association studies, where only a very small subset of genes are truly active and the signal-to-noise ratio is very small. In addition to genetic data situations, the BMVS method has potential to achieve a broad impact because it can be flexibly applied to far-reaching fields that involve multiple response variables [12,13,39]. Another appealing characteristic of the BMVS method is that it skips the mundane process of tuning parameter selections, as required by penalized regression approaches. All simulation and real data examples use the recommended values for hyperparameters with no additional tunings required.

Many studies modeled a multivariate response by separately fitting each component of the response vector using univariate models [21,40–42]. Compared to the BMVS approach, which handles the multivariate response as one unit, the multiple separate univariate regression methods have some notable disadvantages: 1) They fail to take advantage of the correlations between multiple response variables [26]; correlated variables share information in a way that is important for discovering additional signals [2]. 2) It is noted that one single trait cannot adequately represent information described by multiple biological traits. For example, a shape is more accurately described by one vector in its entirety rather than by multiple univariate variables [41–43]. Hence, separately modeling single traits results in a loss of information [44]. 3) When a series of models are fitted, deriving a generalized interpretation for all the response variables is difficult [41,42]. It is more appropriate to group multiple correlated response variables together for practical purposes, whereas separating them could produce incomplete or misleading results. Breiman and Friedman (1997) [26] provided theoretical proofs demonstrating that prediction accuracy can be improved when modeling multiple correlated response variables by using a multivariate regression model rather than using multiple separate univariate regression models.
CHAPTER 3
FUNCTIONAL RANDOM FORESTS FOR CURVE RESPONSE

3.1 INTRODUCTION

A common interest in scientific studies is the repeated measurement of the same continuous response across time or location on the same study unit, which is typically termed “repeated measure”, “longitudinal”, “functional”, or “curve” data. One typical example is the growth of children, using a somatic variable measured regularly from birth to adulthood. Analysis of curve responses has become a very active research topic as technological measurement devices have become more sophisticated (e.g., new sensors and electrodes, easy storage). Its applications so far have included global climate change; functional near infrared spectroscopy (fNIRS); functional magnetic resonance imaging (fMRI); electroencephalography (EEG); positron emission tomography (PET); diffusion tensor imaging (DTI); biological shape or contour studies; speech recognition; population kinetics of plasma folate, and more [45–47]. Despite the fact the data are measured in the form of multiple discrete points in these applications, an underlying trajectory is clearly manifested as a discretized smooth random function or curve. Therefore, modeling the response curve as a continuum function and capturing the overall dynamic trajectories of the function over time is demanding, in particular in situations where the data are recorded by machines over a dense grid of time points.

In addition to the aforementioned response curves, the motivating data also have a large number of predictors. Scientists are particularly interested in delineating which of the predictors are associated with the variation of these response curves’ dynamic trajectories, or in making predictions regarding patterns or time-to-event for new observations. Some predictors (such as BMI) are continuous, and some (such as genotype) are categorical; and the intricate and complicated correlations and interactions among these predictors enhance
analytical challenges.

Substantial work has been done in the functional data analysis field by estimating the overall trend of curves using a mean function and estimating dominant trends of autocovariance using eigenfunctions [46, 48]. However, modeling the relationship between a curve response and a large set of predictors is rare. Even rarer are studies exploring interactions and nonlinear structures despite the fact that such scenarios exist widely in practices. The majority of current functional data analysis literature has focused on modeling the time-dependent response curve itself without considering its relation with any predictor [45, 48–52]. Müller and Yao (2006) [9] regressed longitudinal response trajectories on only one predictor, and Müller (2005) [46] studied the association when both predictor and response are functional curves having the same number of time points. A few other methods were developed for classifying functional response curves, which were equivalent to the model having a functional curve as the predictor (instead of the response) and a categorical response variable. This article is motivated by a desire to identify which of a model’s many predictors are related to functional curved responses and how they are related. This is an undeveloped area in all fields of current literature.

The classification and regression tree (CART) methods have obtained widespread popularity since the mid-1980s because they are well suited for analyses of complex relationships between a set of predictors and a univariate response, including nonlinear structure, high-order complex interactions, lack of balance, and missing values, and others [53, 54]. Breiman et al. (1984) [53] set a classic example of building a tree model for one univariate response. The earliest attempt to relate tree-based models to longitudinal data seems to have been performed by Segal (1992) [55], who assumed three well-known covariance structures: independence, compound symmetry, and a first-order autoregressive model. As pointed out by Zhang and Singer (2013) [56], Segal’s studies were restricted to longitudinal data having a regular structure, with an equal number of observations and being measured via a regular and consistent schedule. Abdolell et al. (2002) [57] extended CART from univariate continuous response to longitudinal continuous response but only in cases that considered
one categorical predictor. Yu and Lambert (1999) [58] attempted a regression tree model with a longitudinal curve as the response. But they first reduced the dimensionality of the curve and then fitted a standard multivariate tree model by using principal component scores or estimating low-order coefficients of spline basis functions as the response (in the form of a multivariate vector rather than a curve), without directly fitting the longitudinal curve. Zhang and Singer (2013) [56] advocated the use of multivariate adaptive regression splines to generate regression trees for longitudinal data. These works represented breakthroughs in extending standard CART models from one univariate response variable to a curve response, but they mostly followed the CART approach and consequently inherited the following weaknesses of the CART method [59]: 1) Selecting variables using CART for splits results in a bias toward certain types of variables [59]. Specifically, this bias works in favor of categorical variables with more levels [60]. 2) CART models may not provide high predictive accuracy. 3) CART models are known to be unstable; small changes in the data may lead to large changes in the results.

Efforts to overcome the limitations of the CART method led to the invention of random forests, proposed by Breiman (2001) [61]. Although the random forests (RF) has gained widespread popularity and emerged as a versatile and highly-accurate methodology in various disciplines, it was designed for only one univariate response (i.e., a single variable with one dimension). It is not reasonable to construct a set of separate RF models for each individual component of the curve measurement when curve or longitudinal responses are considered, because the multiple components collected from the same individual are not independent. Segal and Xiao (2011) [62] were the first (and only) ones to extend random forests from one univariate response to one multivariate response. However, a multivariate response and a curve response are dramatically different in several ways: 1) One essential feature of a curve data is the existence of autocorrelation induced by repeatedly making multiple observations on the same response variable and the same individual at different times. A large proportion of autocorrelation in curve data actually comes from the time variation [46]. Therefore, ignoring the autocorrelation will produce inaccurate model estimates.
and misleading analytic results [55]. 2) A curve response can be thought of as a discretized function showing clear underlying trajectories, but the multivariate modeling does not utilize the time course of the data and hence can not describe the overall time trends of the response in the dynamic or functional sense. Yu and Lambert (1999) [58] gave an example in which naively applying a multivariate regression tree model to a curve response was not successful. 3) A multivariate response takes the form of random vectors, but a longitudinal data takes the form of random functions or curves when time dynamics or spatial dynamics are major components and the data are actually treated as infinite-dimensional functions. According to Muller, extending from one-dimension to infinite-dimension is a bigger leap compared to extension from one-dimensional to finite-dimensional multivariate response, both conceptually and theoretically speaking [46,52]. A typical example for a curve setting is the human growth curve, where events such as the pubertal growth spurt occur at different times for different individuals [46,63–66]. A typical example for a multivariate setting is the measurement of multiple health indicators at a fixed time (e.g., BMI, height, weight, heart beat, blood pressure, etc), where the health conditions vary for different individuals.

In this article, we propose a new approach called functional random forests (FunFor), which facilitates an extension from the traditional random forests methodology (accommodating a single univariate response) to accommodate repeated measures, functional, longitudinal, or curve response settings. The inputs of the FunFor method will be a response curve that repeatedly measures the same variable at multiple time or location points and a large number of multiple predictors for a given individual. FunFor outputs the predicted response curve for each individual along with the importance rank of each predictor according to its association strength with the response curve. FunFor keeps many of the agreeable properties of traditional random forests: effectively modeling both the linear and complex nonlinear relationships between the curve response and the predictor; producing a joint model rather than a marginal model by modeling multiple predictors simultaneously; capturing the intricate higher-order interactions and also accounting for correlations among predictors; demonstrating feasibility for binary, continuous, and categorical predic-
tors; embracing a nonparametric approach without assuming any specific model structure, distribution, or data type; and requiring few tuning parameters. Throughout this article, we concentrate on functional regression (in which the response is continuous, but the predictors can be continuous or categorical) rather than on classification problems.

Quantitative genetic studies on shape trait have been attracting a lot of attentions [67–73], which represent an interesting application for our proposed FunFor approach. The aims of these studies is to detect important genetic and/or environmental factors associated with shape trait, where the shape that is accurately described by a high dimensional response curve together with a large scale of single-nucleotide polymorphisms (SNPs) and/or environmental factors. In particular, the Gene × Gene (also named Epistasis) and Gene × Environment interactions have been raised as a critical but challenging topics in genetic literature. These demanding in quantitative genetic shape research exactly matches with the capability of our proposed FunFor model. Therefore, we illustrate the performance of FunFor approach through six novel simulation designs and one real data analysis from the genetic shape applications. To the best of our knowledge, we are the first to invent a series of novel simulation designs to connect shape response with gene and environmental factors from different scenarios.

3.2 METHODOLOGY

Let $Y_{ik}, i = 1, \ldots, n; k = 1, \ldots, K$, denote the originally observed longitudinal or curve response for the $i^{th}$ individual measured at the $k^{th}$ time (or location). Here $K$ is the length of the response curve (i.e., total number of time points repeatedly measured for each individual), and $n$ is the number of observations, i.e., the sample size. Let $X \in \mathbb{R}^{n \times p}$ be the predictor matrix. Let $X_{j}, j = 1, \ldots, p$ be the $j^{th}$ predictor, where $p$ is the number of predictors. Let $Y_{i} = [Y_{i1}, \ldots, Y_{iK}]$ be the response curve of the $i^{th}$ individual. Let $(X_{i}, Y_{i})$ denote the full data of the $i^{th}$ individual, including all predictors and the entire response curve.
3.2.1 Functional data analysis

The dense but discretely recorded, noisy response curve, $Y_i$, can be modeled as an underlying function plus noise [74]:

$$Y_i = f_i(t) + \varepsilon_i, \ i = 1, \ldots, n; \ t \in [0,1],$$  \hspace{1cm} (3.1)

where $f_1(t), \ldots, f_n(t)$ are a collection of independent realizations of a random functional process $f(t)$ defined on $L^2([0,1])$. The $\varepsilon_i$ are the experimental error vector and are assumed to be independent, with $E(\varepsilon_{ik}) = 0$ and $Var(\varepsilon_{ik}) = \sigma^2_k, k=1,\ldots,K$.

The mean function of $f(t)$ is $E(f(t)) = \mu(t)$, a smooth function of $t \in [0,1]$. And the auto-covariance function of $f(t)$ is $G(s,t) = cov\{f(s), f(t)\} = E\{[f(s) - \mu(s)][f(t) - \mu(t)]\}$ in $L^2([0,1]^2)$, a bivariate positive definite smooth function of $s, t \in [0,1]$. The term “smooth” refers to functions that are twice continuously differentiable.

Functional principal component analysis

In order to model the auto-covariance function, functional principal component analysis (FPCA) interprets $G(s,t)$ as the kernel of a linear integral operator on the space of square-integrable functions on $[0,1]$, mapping $f \in L^2([0,1])$ to $A_G f \in L^2([0,1])$ defined by

$$(A_G f)(t) = \int_T f(s)G(s,t)ds.$$  \hspace{1cm} (3.2)

An eigenfunction $v$ of the auto-covariance operator $A_G$ is a solution to the equation $(A_G v)(t) = \lambda v(t)$, with eigenvalue $\lambda$. We assume that the operator $A_G$ has a sequence of smooth orthonormal eigenfunctions $v_l(t) \in L^2([0,1])$ satisfying $\int_T v_k(t)v_l(t)dt = \delta_{kl}$ (here $\delta_{kl}$ is the Kronecker symbol), with ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.

By Mercer’s Theorem, applying a spectral decomposition on the function $G(s,t)$ yields

$$G(s,t) = \sum_{l=1}^{\infty} \lambda_l v_l(s)v_l(t).$$  \hspace{1cm} (3.3)
The generalized Fourier expansion (Karhunen–Loeve Theorem [75] or functional principal component expansion) decompose the random functions \( \{ f_i(t) \} \) as [8–10]

\[
f_i(t) = \mu(t) + \sum_{l=1}^{\infty} \zeta_{il} v_l(t),
\]

where the sum is defined in the sense of \( L_2 \) convergence. And

\[
\zeta_{il} = \langle f_i - \mu, v_l \rangle = \int_0^1 (f_i(t) - \mu(t)) v_l(t) dt
\]

are uncorrelated random variables with \( E(\zeta_l) = 0 \), and \( \text{var}(\zeta_l) = \lambda_l \), subject to the \( L_2 \) convergence. \( \zeta_l \) are frequently referred to as the \( l \)th functional principal component scores (PC) or the \( l \)th dominant modes of random effects.

Equation (3.4) indicates that the dynamic trends of random function \( f_i(t) \) can be modeled by the mean trend function \( \mu(t) \), the eigenfunctions \( v_l(t) \), and the distribution of functional principal component scores \( \zeta_{il} \). Combining Equations (3.1) and (3.4), the curve response can be estimated as

\[
Y_i = \hat{f}_i(t) = \hat{\mu}(t) + \sum_{l=1}^{L} \hat{\zeta}_{il} \hat{v}_l(t), \quad i = 1, \ldots, n; \quad k = 1, \ldots, K; \quad t \in [0, 1].
\]

Here \( \hat{\mu}(t) \) is the estimated overall mean function, and the \( \hat{v}_l(t) \) and \( \hat{\zeta}_{il} \) are the estimated eigenfunctions and functional principal component scores of the estimated auto-covariance function. \( L \) is the number of PCs to be retained, which is pre-specified according to the proportion of total variation of the response curve explained by the first few PCs. The linear combination of the first \( L \) eigenfunctions and the mean function is used to effectively approximate the trajectories of random function \( f_i(t) \) [76]. The estimated curves \( \hat{Y}_{ik} \) or \( \hat{f}_i(t) \) should be less noisy than the originally observed measurement \( Y_{ik} \).

\section*{Smoothing and Estimating}

Since the real data in practice are contaminated with measurement errors, the sample
mean vector and eigenvectors of the sample covariance matrix $\hat{G}$ or $\Sigma_Y$ of the raw data tend to be noise, which affect the accuracy of the results. Therefore, we need to estimate a smooth version of the mean function and the covariance function. In this section we describe the details related to the estimating and smoothing processes for the mean function $\hat{\mu}(t)$, the covariance function $\hat{G}(s, t)$, and the eigenfunction $\hat{v}_l(t)$, all of which are involved in Equation (3.6). Specifically, we applied the univariate P-spline smoother approach to estimate the mean function $\hat{\mu}(t)$. To estimate and smooth the covariance function $\hat{G}(s, t)$, we applied the sandwich smoother proposed by Xiao et al. (2013a) [77], and also the fast covariance estimation (FACE) described by Xiao et al. (2013b) [78].

P-spline is a combination of B-spline and difference penalties on the estimated coefficients. After giving the optimal number of knots and the degree of smoothing [79], B-splines can be easily constructed by joining multiple pieces of polynomial fittings between knots. However, determining the optimal number of knots can be difficult given that the use of too many knots leads to overfitting and the use of too few leads to a loss of accuracy. To solve this difficulty, the P-spline starts with a relative large number of knots and then uses a second difference penalty on the coefficients of adjacent B-splines to restrict the flexibility of the fitted curve [79]. The second difference penalty numerically approximate the integrated square of the second derivative of the fitted curve. Compared to other smoothing splines or thin plate splines approaches, the P-spline uses fewer knots and less computation to handle high-dimensional curves, represents a more straightforward extension of linear regression models, and better conserves moments (means, variances) of the data.

The sandwich smoother is a fast P-spline method for bivariate smoothing that uses a tensor product structure to simplify an asymptotic analysis and speed up computation. Sandwich smoother applies univariate P-spline smoothers simultaneously along both the row and column coordinates on the sample covariance matrix. Situations where $K > 500$ raise challenges for covariance matrix smoothing in terms of vasty computation and storage burden for high-dimensional matrices [78]. FACE, a fast implementation of the sandwich
smoother [78], was purposefully invented to solve this problem by addressing the high-dimensional covariance function smoothing in functional data analysis.

Let $\bar{G}$ denote the $K \times K$ sample covariance matrix of the response curve. The smooth covariance function estimated by FACE as

$$\hat{G} = S\bar{G}S.$$  \hspace{1cm} (3.7)

Here $S$ is a $K \times K$ symmetric smoother matrix, which can be constructed by the P-splines approaches as $S = B(B^TB + \lambda P)^{-1}B^T$ [78, 79]. Here $c$ is the number of equally spaced knots plus the order of smoothing; $\lambda$ is the smoothing parameter; $B$ is the $K \times c$ matrix with each component $\{B_w(t_k); k = 1, \ldots, K; w = 1, \ldots, c\}$ representing a B-spline basis function; and $P_{c \times c}$ is a symmetric penalty matrix.

Since $\bar{G}$ is low rank (with a small sample size but a high-dimensional curve), Xiao et al. (2013b) [78] proved that the dimension of $\hat{G}$ is at most $c$. Their proposition leads to reduce a computation of big matrix $K \times K$ to a small matrix $c \times c$, allowing spectral decomposition of the smooth estimator of the covariance without calculating or storing the empirical covariance operator. Agreeably, $\hat{G}$ is directly derived after a sequence of several spectral decompositions and matrix performances involving $B$, $P$, $Y$, $\lambda$ etc. $\hat{G}$ is able to keep the symmetric and positive semi-definite property of $\hat{G}$. During the smoothing and estimating processes, the tuning parameter $\lambda$ is selected by minimizing the pooled generalized cross validation (PGCV), a functional extension of the generalized cross validation [78, 80].

Once the covariance function $\hat{G}(s,t)$ is smoothed and estimated, we diagonalize it and its eigendecomposition provides the eigenfunctions estimates [48, 81–84]. The principal component scores are estimated from Equation (3.5). The FACE is implemented by an R package named refund [78, 85], which directly outputs smooth estimators for the covariance function, eigenfunctions, and principal component scores.
3.2.2 Functional regression tree

A regression tree is constructed by successively splitting the predictor space into mutually exclusive subregions and then repeatedly partitioning individuals into those subregions [53]. This tree-growing procedure is employed in a top-down direction, starting from the root node (top of the tree where all observations belong to a single region), generating a sequence of nodes, and stopping at the terminal nodes or “leaves” (bottom of the tree where observations stop being further divided). Each binary split, indicated via two new descendant or child nodes further down the tree, is made by assigning observations with predictor values less than a cutoff point to the left branch, or, to the right branch if their predictor values are higher than the cutoff.

Let \( s \) denote a possible cutoff point value of a given predictor, \( X_j \). Let \( R \) denote a parent node under consideration. For each \( j \) and each \( s \), the pair of left and right descendant child nodes are defined as, \( R_L(j, s) = \{(X_i, Y_i) | X_{ij} < s\} \), and \( R_R(j, s) = \{(X_i, Y_i) | X_{ij} \geq s\} \). At each splitting step, all predictors and their possible outpoint values are evaluated and then the best predictor and cutoff point combination (the one that maximizes the splitting function) is ultimately chosen to produce a split. We will later describe the splitting functions in great detail. We integrate the traditional regression tree model with functional data analysis to invent the new approach presented in this paper.

Before detailing the splitting function for a functional regression tree, we give a brief overview of the principal concepts of the univariate and multivariate splitting functions because they also form the basis for that of the functional regression tree. From this overview, important modification details necessary to extend the application from a univariate to a curve response will be made clear.

**Univariate splitting function**

Let \( y \) denote the univariate response variable and \( y_i \) be the response value of the \( i^{th} \) individual. Let \( \bar{y}_R = \frac{\sum_{i \in R} y_i}{n_R} \) be the average response value of the observations that are divided into a node \( R \), where \( n_R \) is the number of individuals that belong to this node. The
The residual sum of squares (RSS) for node $R$ is computed as

$$RSS(R) = \sum_{i \in R} (y_i - \hat{y}_R)^2.$$  

(3.8)

The most popular splitting criterion for a univariate regression tree was defined [55] as

$$\phi_1(j, s, R) = RSS(R) - RSS(R_L(j, s)) - RSS(R_R(j, s)),$$  

(3.9)

which measured the deviance around the subgroup mean or the within-node homogeneity. Maximizing $\phi_1(j, s, R)$, achieving the greatest possible reduction in the residual sum of squares, was equivalent to minimizing

$$RSS(R_L(j, s)) + RSS(R_R(j, s)).$$

Another splitting criterion was defined [55] as

$$\phi_2(j, s, R) = |\hat{y}_{R_L(j, s)} - \hat{y}_{R_R(j, s)}|,$$  

(3.10)

which measured the between-node separation. Both splitting criteria reduce heterogeneity in the response distribution, or the dissimilarity of individuals within the same leaf node.

**Multivariate splitting function**

Let $\vec{y} = (y_1, \ldots, y_K)^T$ be a multivariate response, and $\vec{y}_i = (y_{i1}, \ldots, y_{iK})^T$ be the multivariate response value of the $i^{th}$ individual. Let $\hat{y}_R = \sum_{i \in R} \frac{\vec{y}_i}{n_R}$ be the average value of multivariate response obtained by the observations that are divided into a node $R$.

Defining a multivariate version of residual sum of squares, a critical step making it feasible to extend from a univariate response to a multivariate response was based on the Mahalanobis Distance [55]

$$RSS_m(R) = \sum_{i: \vec{y}_i \in R} (\vec{y}_i - \hat{y}_R)^T \hat{\Sigma}(\theta, R)^{-1} (\vec{y}_i - \hat{y}_R),$$  

(3.11)
where $\hat{\Sigma}(\theta, R)$ was the estimated covariance matrix of the multivariate response obtained by the observations belonging to node $R$, depending on the parameter $\theta$. The splitting functions for the multivariate response were very similar to those outlined in Equations (3.9) and (3.10), substituting $RSS(R)$ with $RSS_m(R)$ and $\hat{y}_R$ with $\hat{\vec{y}}$. Segal (1992) [55] also proposed another multivariate splitting function using the likelihood ratio test of covariance matrices, the maximization of which separated the two child nodes in terms of covariance structure. However, the proposed splitting function requires the covariance structure to be known (e.g., autoregressive, compound symmetry, or independence), which is impossible for general practical applications.

**Functional splitting function**

Inspired by the extension of the splitting criteria from a univariate response to a multivariate response, we design innovative splitting criteria for a curve response. In this step we input the response data as the smooth function $\hat{f}_i(t)$, instead of the original raw data $Y_{ik}$, as the output of the functional data analysis from Equation (3.6). This is because $\hat{f}_i(t)$ is an accurate estimator of the original raw data $Y_{ik}$ after extra noise are largely removed and after a smooth function is extracted from the original discretely collected points [8]. For a functional response, we define the average function of all smooth functions that are divided into node $R$ as $\hat{f}_R(t) = \sum_{i \in R} \hat{f}_i(t)/n_R$. Accordingly, the new residual sum of squares for the functional curve of node $R$ will be computed based on the integrated squared error (ISE)

$$RSS_f(R) = \sum_{i \in R} ||\hat{f}_i(t) - \hat{f}_R(t)||^2 = \sum_{i \in R} \int_0^1 [\hat{f}_i(t) - \hat{f}_R(t)]^2 dt,$$

where the integral can be approximated by Riemann sum.

The new splitting criteria extended from $\phi_1(j, s, R)$ is

$$\phi_{f1}(j, s, R) = RSS_f(R) - RSS_f(R_L(j, s)) - RSS_f(R_R(j, s)).$$  \hspace{1cm} (3.12)
And the new splitting criteria extended from $\phi_2(j, s, R)$ can be designed as

$$\phi_{f2}(j, s, R) = ||\hat{f}_{RL}(j, s)(t) - \hat{f}_{RR}(j, s)(t)||^2 = \int_{0}^{1} [\hat{f}_{RL}(j, s)(t) - \hat{f}_{RR}(j, s)(t)]^2 dt. \quad (3.13)$$

All predictors and all of their possible cutoff point values of each of the predictors are evaluated for each split and then the best predictor and cutoff point combination (the one that maximizes $\phi_1(i, j, s, R)$ or $\phi_2(i, j, s, R)$) is ultimately chosen to produce a split. Since $\phi_1(i, j, s, R)$ is computationally expensive, we will use $\phi_2(i, j, s, R)$ in all simulation and real data analysis studies. The binary splitting process for each node stops when the splitting criterion of that node (defined in Equation (3.12) or (3.13)) is less than a pre-specified value, when the number of observations within that node is less than a pre-specified threshold, or when the size of the tree is less than a pre-specified value. During the tree-growing process, the estimated response curve $\hat{f}_i(t)$ in Equation (3.6) for each individual will keep updating at each candidate split of each node. This is because a split causes an observation regrouping, so the individual pool (and also the mean and covariance functions) in the parent node differs from those of its descendant nodes.

At the end of the functional regression tree’s fitting process, we estimate each individual’s response curve by the mean response curve of all individuals that fall into each terminal node, i.e., $\hat{Y}_i = \hat{f}_i(t) = \hat{f}_R(t)$. This rule is feasible not only to estimate for existing observations but also to predict for new observations by simply following the entire predictor splitting procedures of the tree. When a tree grows deeper, the homogeneity of the response in each leaf node will increase. A fully-grown functional regression tree may be too complex, likely overfitting the data and leading to poor prediction accuracy. We first grow the tree to the maximum depth, and record the sequence of how the splits are made. We then prune the tree back in reverse order of the sequence [86]. After a split is pruned, we calculate the mean squared prediction error (MSPE) of the corresponding subtree though five-fold cross-validation. Finally, we select the optimal tree size by using the subtree with the lowest MSPE.
3.2.3 Random forests

To overcome the limitations of using a single regression tree (described in the Introduction section), Breiman (2001) [61] invented the random forest (RF) approach, which creates an ensemble of trees to provide a consensus vote.

Randomness

Contrasted with the single tree approach, random forests introduce two instances of randomness during the tree-growing process. First, instead of using all original observations, the data used to grow each tree come from one bootstrap sample of the original observations, which is randomly drawn with replacement (training set). Aggregating the predictions over hundreds of trees can significantly reduce the variance and increase both prediction accuracy and stability when compared to a single tree. The individuals left out of the bootstrap sample and not used in the construction of each tree are called out-of-bag (OOB) observations, which comprise about one third of the original observations. Due to the fact that the OOB samples are not used for model fitting, OOB samples can naturally be chosen as test data and used to estimate the prediction error [87]. Second, only a subset of predictors randomly drawn (instead of evaluating the full list of all predictors) are used to determine the best split for each node of each tree to further gain prediction accuracy. By forcing each split to only consider a subset of the predictors, the RF model gives each variable a similar chance to be considered and prevents the variables having the strongest marginal effects from always dominating, and prevents confounding influences that always exist for those predictors with strong collinearity. In addition, when the number of predictors is very large, the consideration of only a randomly-selected subset of variables for each split leads to better computational efficiency. At the end of the random forest fitting, we estimate the response curve of the \( i^{th} \) individual by the average of the response curve estimators (i.e., \( \hat{f}_R(t) \)) obtained from all trees that select the \( i^{th} \) individual through bootstrap sampling (training set), i.e., \( \bar{Y}_i = \hat{f}_i(t) = \bar{f}_R(t) \), where \( \bar{f}_R(t) \) is the average across multiple trees. Similarly, the prediction of the \( i^{th} \) individual should be computed across all trees that does not select the \( i^{th} \) individual through bootstrap sampling (OOB sample).
Prediction variable importance measure

In addition to the estimated or predicted response curve, the prediction variable importance measure (PVIM) of each predictor is another output of the FunFor approach. The PVIM of each predictor \( X_j \) is obtained via the difference in average prediction error before and after randomly permuting \( X_j \) while keeping other predictors untouched [87]. Let \( B_q \) denote the set of OOB sample of the \( q^{th} \) tree and ‘ntree’ denote the total number of trees, then the PVIM of predictor \( X_j \) can be defined as

\[
PVIM(X_j) = \frac{1}{\text{ntree}} \left( \sum_{q=1}^{\text{ntree}} \frac{1}{|B_q|} \sum_{i \in B_q} \left[ ||Y_{ik} - \hat{Y}_{ik}^p||^2 - ||Y_{ik} - \hat{Y}_{ik}||^2 \right] \right), \tag{3.14}
\]

where \( \hat{Y}_{ik} \) is still the predicted response curve obtained from the functional random forest method before permuting \( X_j \), and \( \hat{Y}_{ik}^p \) is the predicted response curve after permuting \( X_j \), and \( |B_q| \) denotes the size of the set.

The larger a PVIM is, the more significant the corresponding predictor is in predicting the response curve. A predictor with a negative or close-to-zero PVIM value is interpreted as unimportant [61]. The idea underlying the PVIM is that if a predictor is strongly associated with the response curve, then a permutation breaking this association will yield a substantially high prediction error compared to that before the permutation. In contrast, the prediction error before and after a permutation will be close to zero if a predictor is not that important in predicting the response curve. The PVIM ranks the predictors from most important to least important, based on the strength of their associations with the response curve. In addition, the PVIM assesses each predictor’s overall impact by conditioning on all other predictors in the same model, without requiring explicitly putting interactive terms or any other formal structures to be inserted into the model. The PVIM ranks all the predictors based on both their strong marginal associations and their strong interactive associations. Thus, predictors that ranked in the top are most strongly associated with the curve response, some due to strong marginal effects, and others due to weak marginal but strong interactive effects.
3.2.4 Tuning parameters

There are two tuning parameters involved in the FunFor approach. The size of this subset of the predictors, designated ‘mtry’, is the primary tuning parameter for the random forest procedure [88]. For regression, it has been suggested that ‘mtry’ should be approximately one third of the total number of predictors \( p \). As stated by Breiman (2001) [61], RF achieves exceptional prediction accuracy that is stable under a wide range of values of ‘mtry’. Another tuning parameter is the value of \( L \) described in Equation (3.6). We propose two options to determine its value:

1. Option 1: Determine a value of \( L \) using all observations at the root node. Then use this fixed value of \( L \) for all splits during the tree growth process without considering that the variation mode changes as observations change.

2. Option 2: Specify the minimum percentage of total variation explained by the data (e.g., 90%). Adaptively and flexibly determine the value of \( L \) at each split according to different observations obtained at the parent and child nodes. Keep updating the value of \( L \) during the tree growth process.

The proposed FunFor method comprises the integration of functional data analysis (Section 2.1), functional regression trees (Section 2.2), random forests (Section 2.3), and tuning parameters (Section 2.4). The FunFor approach uses the same idea of random forests in introducing two instances of randomnesses and also growing an ensemble of functional regression trees to yield a consensus vote. Similar to the standard RF model, the output of the FunFor is the estimated response curve for each individual \( \hat{Y}_{ik} \) along with the permutation variable importance measures for each predictor based on their association strength with the response curve.

3.3 SIMULATION STUDIES

We invent six novel simulation designs to explore the performance of the FunFor approach in detecting influential predictors under different levels of difficulty. Each simulation
design is performed with 100 time replications. In each replication, we simulate 100 individuals ($n = 100$) and 100 SNPs ($p = 100$). The SNP data are coded as 0 for $aa$ or $bb$, 1 for $Aa$ or $Bb$, and 2 for $AA$ or $BB$. We set three 360-dimensional curves as the mean responses under each of the three genotype groups, $\mu_1$, $\mu_2$, and $\mu_3$. These three curves represent three true poplar leaf shapes varying from lanceolate to ovate, as shown in Figure 3.1. See Fu et al. (2013) [89] for the shape analysis details used to transform an image to a curve.

### 3.3.1 Simulation 1: a single-causative-predictor design

In this design, we consider a scenario in which only one causative predictor contributes to the response. As a categorical variable, each SNP is generated using binomial distribution with a random minor allele frequency ($\text{MAF} \sim \text{Uniform}(0.1, 0.5)$). Among all generated SNPs, we randomly choose one to truly contribute to the response (call it $X^*_1$), and all the rest are noise. Then the shape curve samples for each replication are generated based on this truly causative SNP and the three mean shape curves are

$$Y^T_i = [\mu_1^T, \mu_2^T, \mu_3^T] \cdot I_{X^*_1} + \epsilon_{360 \times 1}, i = 1, \ldots, 100,$$

where $\epsilon_{360 \times 1} \sim \mathcal{N}(0_{360 \times 1}, \Sigma_{360 \times 360})$, and $\Sigma$ is the sample covariance matrix of the three mean shape curves. $I_{X^*_1}$ is an indicator function that is defined as
\[ I_{X_1^*} = \begin{cases} 
(1, 0, 0)^T, & \text{if the genotype of } X_1^* \text{ is AA}, \\
(0, 1, 0)^T, & \text{if the genotype of } X_1^* \text{ is Aa}, \\
(0, 0, 1)^T, & \text{if the genotype of } X_1^* \text{ is aa}. 
\end{cases} \]

### 3.3.2 Simulations 2-4: a two-interactive-causative-predictors design

In simulations 2-4, we consider the scenarios in which two causative predictors jointly and interactively contribute to the response curve. Marchini et al. (2005) [90] proposed three polygenetic simulation models in human genetics literature and connected the disease status ($Y = 0$ or 1) with two SNPs through the odds ratio of $3 \times 3$ combinations of two SNP genotypes (see Table 4.1). They designed three tables to vary the difficulty level of the interaction effects, respectively. However, Marchini et al. (2005)'s simulation models cannot be directly implemented in our case for the following reasons: 1) Their response only has two groups ($Y = 1$ for disease and $Y = 0$ for non-disease) and hence the response samples can be generated easily once the odds ratio is given. However, our simulation considers three response groups (demonstrated in Figure 3.1), and the process of generating response samples coming from three groups based on the odds ratio is not straightforward. 2) Their response handles only one numeric value but our simulation considers a high-dimensional curve. To take advantage of the Gene×Gene interactive designs of Marchini et al. (2005) [90], we propose a few modifications: 1) We extend the traditional logistic model that Marchini et al. (2005) [90] used for binary response to multinomial logistic model for three groups. 2) We still use Table 4.1, but interpret the components of this Table 4.1 as the probabilities of choosing the first shape versus choosing the second shape ($p_1/p_2$) instead of the odds ratio. Similarly, the probabilities of choosing the first shape versus choosing the third shape ($p_1/p_3$) can also be computed by using the same Table 4.1 and changing a different set of parameters. That is, we use two sets of $\alpha$'s and $\theta$'s values ($\alpha_1 = 0.8, \theta_1 = 0.5, \alpha_2 = 0.9, \theta_2 = 0.4$). These values are chosen because they are close to those suggested by Marchini et al. (2005) [90]. Once these two probabilities of 'choosing ratios' are computed, we are able to derive the probabilities of choosing each shape because ($p_1 + p_2 + p_3 = 1$).
Table 3.1: The odds ratio table for Epistasis design in simulations 2, 3, and 4

<table>
<thead>
<tr>
<th>Simulation</th>
<th>II (Model 1)</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>α</td>
<td>α(1+θ)</td>
<td>α(1+θ)^2</td>
<td>α(1+θ)^3</td>
</tr>
<tr>
<td>Bb</td>
<td>α(1+θ)</td>
<td>α(1+θ)^2</td>
<td>α(1+θ)^3</td>
<td>α(1+θ)^4</td>
</tr>
<tr>
<td>bb</td>
<td>α(1+θ)^2</td>
<td>α(1+θ)^3</td>
<td>α(1+θ)^4</td>
<td>α(1+θ)^4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation</th>
<th>III (Model 2)</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>Bb</td>
<td>α</td>
<td>α(1+θ)</td>
<td>α(1+θ)^2</td>
<td>α(1+θ)^3</td>
</tr>
<tr>
<td>bb</td>
<td>α</td>
<td>α(1+θ)^2</td>
<td>α(1+θ)^3</td>
<td>α(1+θ)^4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation</th>
<th>IV (Model 3)</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>Bb</td>
<td>α</td>
<td>α(1+θ)</td>
<td>α(1+θ)</td>
<td></td>
</tr>
<tr>
<td>bb</td>
<td>α</td>
<td>α(1+θ)</td>
<td>α(1+θ)</td>
<td></td>
</tr>
</tbody>
</table>

We generate the two causative SNPs following a Binomial distribution with a fixed MAF=0.25 and all other noise SNPs with varying MAF ∼ Uniform(0.1,0.5). Then we randomly choose two SNPs to truly contribute to the response (call it $X_1^*$ and $X_2^*$). Based on the genotype combination of $X_1^*$ and $X_2^*$ after reading Table 4.1 using aforementioned modifications, we compute $p(Y_i = \mu_1)$, $p(Y_i = \mu_2)$, and $p(Y_i = \mu_3)$. In this design, we jointly connect three mean shape curves with nine combinations of two SNPs and then simulate all response curve samples accordingly though a multinomial logistic regression model.

**Simulation 2:** The ‘choosing ratio’ baseline is α when the genotype combination is $AABB$, and then it increases multiplicatively $(1+\theta)$ whenever at least one copy of minor allele $a$ or $b$ appears (see the first panel of Table 4.1). Since the effects show up in both within and between loci, this model has strong marginal additive effects that are the easiest to be detected.

**Simulation 3:** The ‘choosing ratio’ baseline is α when the genotype combination is $AABB$, and then it increases multiplicatively $(1+\theta)$ whenever at least one copy of minor alleles $a$&$b$ simultaneously appears, additionally it increase another multiplicatively $(1+\theta)$ whenever one extra minor allele $a$ or $b$ appears (see the second panel of Table 4.1). Since the effect of each allele still occurs on some certain conditions (when at least one copy of $a$&$b$ shows up), this model is harder to be detected compared to Simulation 2.

**Simulation 4:** The ‘choosing ratio’ baseline is α when the genotype combination is...
AABB, and then it increases multiplicatively \((1 + \theta)\) when both minor alleles \(a\&b\) simultaneously appears, but does not further increase when additional copies of minor allele appears (see the third panel of Table 4.1). Compared to Simulation 3, the power of \((1 + \theta)\) stays the same for four combinations when \(a\&b\) shows up simultaneously. Since the effect of each individual allele has no role in changing the probability, this model is also harder to be detected than Simulation 2.

3.3.3 Simulations 5: a three-interactive-causative-predictors design

In this design, we consider the joint effects of three causative predictors (two SNPs and one environmental factor). This represents an example containing both categorical and continuous type in predictor. We generate 100 SNPs following exactly the same design as Simulation 1. Then three environmental factors are generated from Light \(\sim\) Uniform\((10,000, 100,000)\)(lx), Temperature \(\sim\) Uniform\((50,100)\)(\(^\circ\)F), & and Relative Humidity \(\sim\) Uniform\((10,90)\)(%).

Then we set the Light and two randomly chosen SNPs to be three predictors that truly contribute to the response. We connect these three causative predictors with the shape curve by the same table and same design as those used in Simulation 2. All of the other 98 SNPs (with varying MAF \(\sim\) Uniform\((0.1, 0.5)\)) and the Temperature and Relative Humidity are set to be noise. Due to the variation between these predictors, we standardize the three environmental factors before conducting analyses.

3.3.4 Simulations 6: a two-correlated-causative-predictors design

In this design, we consider the joint effects of two causative predictors between which there exists a strong collinearity between these two predictors. The collinearity in statistical term is manifested via linkage disequilibrium (LD; a genetics term), which is defined as the non-random association between genetic alleles. We simulate the first causative SNP, \(X^*_{1}\), following a Binomial distribution with MAF=0.3. Then we simulate the second causative SNP, \(X^*_{2}\), with MAF=0.5 based on the conditional probability \(p(X^*_{2} | X^*_{1})\). Refer Foulkes (2009) [91] for the detailed conditional table for all \(3 \times 3\) combinations. The LD parameter is set to be \(D = 0.1\). After two highly correlated causative SNPs are ready, we connect the
shape curve with these two SNPs using the same Equation (3.15) as Simulation 1. But the indicator function need to be extended from one SNP to two SNPs as follows:

\[ I_{X_1^*, X_2^*} = \begin{cases} 
(1, 0, 0)^T, & \text{if the genotype of } X_1^* \text{ is } AA \text{ and the genotype of } X_2^* \text{ is } BB, \\
(0, 1, 0)^T, & \text{if the genotype of } X_1^* \text{ is } Aa \text{ and the genotype of } X_2^* \text{ is } Bb, \\
(0, 0, 1)^T, & \text{otherwise.}
\] 

All other noise SNPs are generated by Binomial distribution with varying MAF \( \sim \text{Uniform}(0.1, 0.5) \).

### 3.3.5 Simulation results

For simplicity, we use the same parameter values \((L = 1, \text{mtry} = 10, \text{ntree} = 100)\) when applying the FunFor approach to fit each of the six simulation designs. Then we use the PVIM to rank all predictors for each of the 100 replications of each simulation design. Three criteria, \(R\), \(p\), and \(M\), are used to assess the performance of the FunFor approach:

1. \(R\) is defined as the average rank of each causative predictor across 100 replications. The smaller the \(R\) value is, the better the approach performs.

2. \(M = \max(R_j)\) is defined as the minimum selection size, i.e., the minimum number of predictors that is required to include all of the causative predictors in each replication. Therefore, a \(M\) value close to the number of causative predictors indicates good performance. We also calculate the 5%, 25%, 50%, 75%, and 95% quantiles of \(M\) among the 100 replications.

3. To determine the power of successful selection, we need to define a threshold \(d\) such that all predictors whose ranks are above \(d\) are thrown away, leaving only predictors whose ranks are below \(d\) as a selected subset. Liu et al. (2014) [92] suggested that the threshold be computed as \(d = \left[ p^{4/5} / \log(p^{4/5}) \right] \). After a selected subset is obtained for each replication, we compute the powers as the proportion of truly causative predictors that are successfully selected. Specifically, the individual power \(P_{w}\), is defined as the proportion of each predictor that is being successfully selected within the threshold \(d\) across 100 replications. The overall power \(P_a\), is defined as the proportion of all
Table 3.2: Results of six simulation designs

<table>
<thead>
<tr>
<th>Average rank</th>
<th>( R_{X_1^*} )</th>
<th>( R_{X_2^*} )</th>
<th>( R_{\text{Light}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation 1</td>
<td>1.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation 2</td>
<td>1.43 1.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation 3</td>
<td>2.77 2.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation 4</td>
<td>2.1 3.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation 5</td>
<td>4.28 5.61 5.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation 6</td>
<td>1.08 1.96</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Minimum model size</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantiles</td>
<td>5%</td>
</tr>
<tr>
<td>Simulation 1</td>
<td>1</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>2</td>
</tr>
<tr>
<td>Simulation 3</td>
<td>2</td>
</tr>
<tr>
<td>Simulation 4</td>
<td>2</td>
</tr>
<tr>
<td>Simulation 5</td>
<td>3</td>
</tr>
<tr>
<td>Simulation 6</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Powers</th>
<th>( P_{X_1^*} )</th>
<th>( P_{X_2^*} )</th>
<th>( P_{\text{Light}} )</th>
<th>( P_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation 1</td>
<td>100%</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>100% 100%</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Simulation 3</td>
<td>97% 95%</td>
<td></td>
<td></td>
<td>90%</td>
</tr>
<tr>
<td>Simulation 4</td>
<td>97% 96%</td>
<td></td>
<td></td>
<td>92%</td>
</tr>
<tr>
<td>Simulation 5</td>
<td>93% 94% 91%</td>
<td></td>
<td></td>
<td>79%</td>
</tr>
<tr>
<td>Simulation 6</td>
<td>100% 100%</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

causative predictors that are simultaneously selected within the threshold \( d \) across 100 replications. The larger the \( P_a \) and \( P_w \) are, the better the model performs.

As shown in Table 4.2, the results of Simulations 1, 2, and 6 achieve a level of perfection because they rank the two truly causative predictors as the top two predictors on average, the ranks of which are definitely higher than all other noise predictors in terms of the prediction variable importance measure (top panel of Table 4.2); Additionally, there is not much diversity among all 100 replications with all quantiles of \( M \) being exactly equaling to the number of truly causative predictors (middle panel of Table 4.2). It indicates that all replications, without exception, locate the best results. Also, both individual powers and overall power are all 100% (bottom panel of Table 4.2), which means that all of the causative
predictors are successfully selected. We recognize that the results of Simulations 3 and 4 drop down a little bit compared to those of Simulations 1, 2, and 6 (see Table 4.2). Since we made quite a few modifications on the three simulation models proposed by Marchini et al. (2005) [90], the response curve and the causative predictors are not connected in a direct and straightforward way, so they should be a little more difficult to detect. In addition, Marchini et al. (2005) [90] also reported something similar: the simulation results of his models 2 and 3 (corresponding to our simulation 3 & 4) were worse than those of model 1 (corresponding to our simulation 2).

We are not surprised to see that the results of Simulation 5 are the worst among the six simulation designs, because Simulation 5 represents the hardest setting. The other five simulations only consider pure SNP data (categorical), but in Simulation 5 we include both SNP and environmental (continuous) data. The diverse ranges of these predictors with \{0, 1, 2\}, (10,000, 100,000), (50,100), and (10,90) increase the level of difficulty. Moreover, three causative predictors itself are harder to be detected than two causative predictors. Despite these anticipated difficulties, the results of Simulation 5 are still good based on the facts that the three causative predictors are all ranked, on average, around the top five and that the individual powers for all three causative predictors are greater than 90% (see Table 4.2). We acknowledge that the overall power of 79% is indeed not high, but this low power is caused by an inappropriate choice of threshold $d$. For a more difficult scenario, the threshold should be correspondingly larger. After choosing a better $d$ that can flexibly accommodate the harder scenario [93], we believe that the overall power $P_a$ for Simulation 5 will increase.
Fig. 3.2: Illustration of the prediction for one replication in Simulation 2: The simulated response samples (transparent green lines on the first row) and predicted response curves (transparent green lines on the second row) output from the FunFor approach. Each column corresponds to a different genotype group. The black lines demonstrate the three true shapes used for producing Simulation 2.

Fig. 3.3: Same information as the Figure 3.2, but converting curves into shapes
In addition to Table 4.2, we also demonstrate the prediction performance of the FunFor approach by visualizing one replication from Simulation 1 (see Figures 3.2 and 3.3). Figure 3.3 displays information that is similar to that in Figure 3.2, but converts the curves back to shapes. The 100 simulated samples are divided into three groups based on the genotype information of the single causative SNP described in Simulation 1. Each separate column of Figures 3.2 and 3.3 represents a different genotype. The first row of Figures 3.2 and 3.3 stands for the original simulated response samples \( Y_i \), and the second row is the predicted response output from the FunFor model, i.e., \( \hat{Y}_i = \hat{f}_{R}(t) \). The truth (i.e., each mean leaf used to generate data) illustrating in Figure 3.1 is now visualized by a solid black line. The original simulated and finally predicted response samples are demonstrated by transparent green lines. As shown in the second row of Figures 3.2 and 3.3, the predicted response curves of genotype group AA and Aa are very close to the truth, with predictions capturing not only global trends but also subtle local details of the mean curves used to simulate the sample. The prediction of the genotype group aa (the third column) is not perfect because some details near the tips and spaced petiolar sinus of the leaf do not overlay, but the global overall trends still match well. Despite the fact that the simulated response samples are noisy (see the first row of Figures 3.2 and 3.3) and the majority of SNPs are noise who may confound the true signal, the prediction results are great.

3.4 REAL DATA ANALYSIS

In this section, we analyze the leaf shapes of a natural population of 421 Populus euphratica (also named Euphrates Poplar or Desert Poplar) plants, which naturally inhabit along river valleys in arid regions of the Xin Jiang province of China. Twenty five leaves were randomly collected from each plant. The leaves of Populus euphratica are polymorphic with complex and irregular details on their boundaries, so a small set of loose landmark points will be incapable of accurately describing them. We use a 910 \times 1 dimensional curve to describe each shape based on the directional radii method [94]. After fixing a starting point and an end point, the x-y coordinates of all points on the boundary of each shape are recorded one by one assuming the boundary is a closed one-pixel wide curve. Then
Fig. 3.4: The shape description process: (A), (B), & (C) are to recognize a shape from an image; (D) & (E) are to transform a shape to a curve.

the radii connecting each point on the boundary to the centroid are computed and forming a radii curve (invariant to translation) without information loss. Normalization is applied to the radii curve to filter the scale effect. To make it invariant to rotation, a directional radii curve is constructed by rearranging the normalized radii curve based on the pair of the longest and shortest radii. Finally, a high-dimensional directional radii curve is ready to accurately represent each leaf shape after variation caused by pose (translation, scale, and rotation) are aligned and the length of each curve is normalized by transformations [94]. As observed in Figure 3.4, the conversion between the shape and the curve is accurate, with sharp, complex, and irregular teeth on the boundary well-maintained.

Since the 25 leaves collected from the same tree are not independent and they share the same genetic information, we use the average of 25 shape curves as the responses curve. The averaging process may cause shape curves to lose some subtle information, such as the small zig-zags on the leaf teeth. However, the global trends of leaf shapes, which are mainly
affected by genetics, are more accurate when an averaging process is used. For each of the 421 plants, 104 markers were also genotyped. We apply the FunFor approach to detect important markers associated with the variation of shape curves.

During this real data analysis, we also empirically explore whether the FunFor approach is robust over different choices of tuning parameters, as mentioned in the Methodology Section (2.4). Since ‘mtry’ is suggested to be $p/3 \approx 35$, we employ three mtry values: 10, 20, and 35. In addition, we also try the two options (fix version and adaptive version) of choosing $L$ as described in Section (2.4). For the fixed setting, we choose $L = 4$ because the first four PCs can explain about 90% of the total variation in the response curve using all individuals in the root node. The performance of FunFor approach is assessed by evaluating the mean integrated squared error (MISE) between the observed and the predicted response curve using five fold cross-validation. The results shown in Table 3.3 indicate that the combination of the adaptive $L$ setting and $mtry = 20$ yields the best prediction results (118.9). Therefore, we use these two settings for the remaining analyses. However, the differences in MISEs are small among the six different combinations, which empirically verifies that the FunFor method is robust over a reasonable choice ranges of the two tuning parameters.

The PVIMs of each markers are demonstrated in Figure 3.6. The majority of markers have PVIMs close to zero, which matches with the sparsity phenomenon of the genetic dataset. The three markers with the highest PVIMs (their PVIMs of more than 200 dramatically standing out) are $ORPM\_190$, $GCPM\_1812$, and $U50206$, highlighted in red dots. In order to demonstrate the genetic effects of these astonishingly important markers, Figure 3.7

---

<table>
<thead>
<tr>
<th>$L$</th>
<th>mtry</th>
<th>10</th>
<th>20</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive L</td>
<td></td>
<td>119.5</td>
<td>118.9</td>
<td>119.4</td>
</tr>
<tr>
<td>Fixed $L = 4$</td>
<td></td>
<td>119.7</td>
<td>119.3</td>
<td>118.9</td>
</tr>
</tbody>
</table>
Fig. 3.5: Choice of the number of PC kept: the solid line indicates the percentage of variation explained by each PC; The dashed line indicates the cumulative percentage of variation explained by the first few PCs.

visualizes the average of the predicted shape curves’ $\tilde{Y}_i = \tilde{f}_R(t)$ FunFor outputs across all individuals under each genotype group. The three genotype groups exhibit different variability around their basic growth curve shapes, As shown in Figure 3.7, the shape curves of different genotypes differ at multiple points, which indicate the association between the markers and the shape curve. The marker $ORPM_190$ only has two genotypes ($AA \& aa$), therefore its red line (corresponding to Aa) does not show up. Its two lines differ mainly on the intervals from [100, 250] and [650,800]. For marker $GCPM_1812$, the entire trends of aa (blue line) and Aa (red line) are very similar, but they are dramatically different from the curve of AA (black line), which represents a very striking codominant genetic effect in the genetic literature. For marker $U50206$, the entire trends of AA (black line) and Aa (red line) are very similar, but they are different from the curve of aa (blue line), which represents a standard recessive genetic effect in the genetic literature. Compared to the large global variation difference showing in marker $GCPM_1812$, the variation of marker $U50206$ mainly occurs in local, subtle details. As a comparison, the genotypes of three genetic markers
Fig. 3.6: Prediction variable importance measures of 104 genetic markers. The highest VIMs are highlighted with red dots.

(Pe_5, Pe_8, GCPM_1941) with low PVIMs are also visualized in Figure 3.8. Compared to Figure 3.7, Figure 3.8 shows that the average shape curves of the three different genotypes almost overlap each other. These results reconfirms that the markers with low PVIMs are unlikely to be associated with the responses curve. In addition to the mean shape curve, we also demonstrate the first two eigenfunctions for these three markers (see Figure 3.9) and notice even more dramatic differences between the three genotypes, which confirms the in accuracy resulting from assuming independence structure or neglecting the covariance matrix.

As a general-purpose machine learning approach does, the FunFor can predict the response curve for brand new individuals by following the same tree building and ensemble process once these individuals’ predictors are given. To visualize the prediction capability of FunFor for new observations, we divide our data into training and test sets by five-fold cross-validation, and then use the training set to fit the FunFor approach and use the test
set to predict the response curve. Figure 3.10 visualizes four shape examples. We notice that the predicted shapes (dotted lines) and the original shapes (solid lines) match with great consistency.

Fig. 3.10: Prediction of new shapes when the genetic predictors are given. The solid lines are the original observed shape, and the dot lines are the predicted shape output from the FunFor approach.

3.5 DISCUSSION

The performance of the tree-based models over traditional regression models (e.g., GLM, GAM) in terms of prediction accuracy has been widely noticed [95]. In addition, the tree-based models are particularly useful for categorical predictors. A categorical predictor is usually treated as a continuous variable or coded by dummy variables in standard regression models. However, the former misleadingly imposes an order on a nominal predictor, and the linear combination of an ordinal predictor does not really have natural interpretations. The latter dramatically increases the number of total predictors, which increases the burden for a high-dimensional data.
Several attempts have been made to extend the tree-based models from an univariate response to a longitudinal or curve response. One important step enabling this extension is the replacement of the conventional goodness-of-split criteria with new rules so as to accommodate the curve response. Some limitations in similar works of existing literature are as follows: Using the likelihood ratio statistic to evaluate the goodness-of-split requires the distributions and model structures of the data to be known, but assuming a parametric form may be misleading if the underlying data do not satisfy the main assumptions. In addition, the mean and covariance structure of the data is keep changing after the individuals are recursively divided into two child nodes. Therefore, a nonparametric modeling will be more accurate and flexible; A splitting rule based solely on deviations around subgroup mean vectors may neglect the roles of the covariance structure. Analyses that ignore the covariance structure are known to produce incorrect variance estimates [8,96]. Instead the framework that the FunFor approach used to estimate the curves from Equation (3.6) considers both the mean and the covariance structure. Moreover, all tree-based models handling a longitudinal or curve response in current literature are restricted to a single regression tree, suffering from the limitations as discussed in the Introduction section. Another method addressing tree-based models for longitudinal data is proposed by Loh and Zheng (2013) [59]. Their proposed algorithm conducted chi-square tests of the residual curve patterns to select a variable to split each node of the tree. However, the dimensionality of the frequency table may cause the chi-square tests to lose power, because there will be $2^K$ different residual patterns and the computational cost will be high if the length of the curve $K$ is large. In addition, Loh and Zheng (2013)’s work also deals with single regression trees rather than random forests.

The FunFor approach explores more underlying data structures than the traditional RF approach. It describes the trajectory of a curve that changes over time and also shows subject-specific events that different individuals may experience. It not only makes response predictions for new individuals when only predictors are available (as traditional RF approach does), but it also predicts the time-to-event in the future or response values un-
der new time points for both observed individuals and brand new individuals. For a curve response, the heterogeneity can pertain to both the mean and the covariance structure [55]. After the FunFor approach goes through a serious of recursive splitting processes, the observations belonging to the same terminal node are more homogeneous in terms of both the mean and the covariance structure of the response curve.

Like the traditional random forest approach does, the PVIM output provided by the FunFor cannot gives a p-value or significance test result. But it does give a relative importance rank for all predictors based on their association strength with the response curve. And this association strength can be nonlinear, complex, and/or interactive, considering all other predictors simultaneously in the same model (i.e., a joint model instead of an individual model). If a threshold or cutoff is needed for the variable selection purposes, we suggest using the adaptive threshold determination approach as proposed by Diaz-Uriarte and De Andres (2006) [93].

Compared to the traditional RF approach, the FunFor requires a higher computational cost because handling a high-dimensional curve itself is much more time consuming than handling only a single value. The smoothing and estimation process need to be called at each split candidate of each node for each tree, and repeated multiple times for different multiple split candidates, multiple nodes, and multiple trees. Therefore, it is critical for us to choose the FACE algorithm to smooth the covariance matrix because this represents the most time-consuming step. Xiao et al. (2013b) [78] claim that, by using this method, their storage cost is only $O(Kc)$ instead of $K \times K$, and they are provided with instantaneous smoothing for matrices of dimension $K < 10,000$.

We took care to use only two tuning parameters, $mtry$, and $L$. Breiman (2001) [61] claimed that the generalization error for an ensemble of trees converges asymptotically to a limit as the value of $ntree$ increases, so the value of $ntree$ does not impact the prediction if it is large enough. The simulation and real data analyses are restricted to data observed at equally-spaced, fixed time points with a regular and consistent grid and no missing values. The curve may pose a practical problem by being measured at an irregular grid with various
time or spatial location points, and being corrupted with missing values in the response variables. Using time-warping adjustments [97] or treating the time points as an extra random variable [9,84,98] (e.g., the PACE algorithm proposed by Yao et al. (2005) [84]) will easily make our FunFor approach still works for fixed or random time points and for regular or irregular measuring scales.

Since shape represents a good example of high-dimensional curves, we work on a few quantitative genetics shape application samples. In current shape literature, principal components analysis has been widely used to decrease the dimensions of the original shape curve [72,89,99,100], then each PC is analyzed individually using various univariate statistical models. The FunFor approach introduces a very new direction for the shape applications in that it models the entire shape curve instead of modeling the PCs one by one. The FunFor approach has a great potential to be applied to a wide range of practical problems in which analysts want to delineate which of the predictors are associated with the variation of dynamic trajectory of these response curves, or in which they want to predict the curve response using a set of predictors.
Fig. 3.7: The mean predicted shape curves output from FunFor approach across all individuals belong to each genotype of the three most important markers ORPM_190, GCPM_1812, and U50206. Black curve for AA, red curve for Aa, and blue curve for aa.
Fig. 3.8: The mean predicted shape curves output from FunFor approach across all individuals belong to each genotype of the three least important markers $Pe_5$, $Pe_8$, and $GCPM_{1941}$. Black curve for AA, red curve for Aa, and blue curve for aa.
Fig. 3.9: The first two Eigenfunctions output from FunFor approach across all individuals belong to each genotype of the three most important markers ORPM\_190, GCPM\_1812, and U50206. Black curve for AA, red curve for Aa, and blue curve for aa. Solid line for the first Eigenfunction $\hat{v}_1(t)$ and the dash line for the second Eigenfunction $\hat{v}_2(t)$
CHAPTER 4
SHAPE MODELING FOR MULTIPLE LEAVES COLLECTED FROM THE SAME TREE

4.1 INTRODUCTION

Recent advancements in technologies have facilitated the collection and computation of high-dimensional functional data. Functional data analysis (FDA) has now become an important component in many research fields, such as quantitative genetics [42, 101–103], medical science [104–106], computer vision and image analysis [107–109], and so on. Ramsay (2006) [7] provides a broad overview of current methodologies and applications in the functional data analysis area. In addition to traditional single-level functional data, many current functional data contain measurements on the same subject at multiple visits. For example, the shape of a leaf can be measured several times to assess its growth, or shape curves of multiple leaves from one single tree can be measured to find out causative genetic and environmental factors. Also, Di et al. (2009) [5] and Nieto et al. (1997) [110] introduced a large dataset from the Sleep Heart Health Study in order to examine the potential associations between sleep-disordered breathing and health outcomes, in which functional data of each subject/participant are observed at multiple levels/visits. Chen and Müller (2012) [111] studied the curves of mortality rates repeatedly measured for each of the calendar years from 1960 to 2006 across 32 countries.

Functional regression analysis, a generalization of regression methodology to functional responses and/or functional predictors, has become an increasingly popular subfield of FDA [6, 7, 96, 98, 103, 112–114]. Research has been conducted on developing regression models for functional responses [7, 113, 114], and also on incorporating functional phenotypes into genome-wide association studies (GWAS) [103]. However, these methods do not consider functional measurements on the same subject at multiple visits, or say, multilevel functional
phenotypes.

Some other research also extended FDA to multilevel functional data studies [5,111,115–118]. However, they focused on modeling mean structure and variance/covariance structure of multilevel functional data instead of functional regression. Crainiceanu et al. (2009) [6] proposed a generalized multilevel functional linear model (GMFLM) to incorporate multilevel functional data into the linear regression model. However, in the regression model for GWAS, phenotypes are usually considered as responses. GMFLM only works with functional predictors instead of functional responses and does not include variable selection features in the case of high-dimensional predictor space ($p \gg n$), which makes the model not directly applicable to problems with multilevel functional phenotypes and a large number of genetic variants such as single-nucleotide polymorphisms (SNPs).

GWAS are currently under intense methodological and computational research to unveil and characterize the genetic control of complex biological phenotypes. Most of the GWAS research only involve univariate response. However, some biological phenotypes can always be better described by trajectories or curves for various reasons such that the phenotype undergoes a development process [43,119] or the phenotype (i.e., the shape of an organism) cannot be effectively represented by one single value [41–43]. Functional regression analysis, coupled with regularization techniques, can be used to handle both high-dimensional phenotypes and high-dimensional predictor space [103]. However, due to advancement in technology and computation, these high-dimensional biological phenotypes can easily be measured on the same subject at multiple visits. Although it may encounter unprecedented statistical challenges to tackle GWAS problems with multilevel functional phenotypes, we would like to develop a new statistical approach that incorporates these new features, which can eventually have a broad impact on both functional data analysis and genetic association research fields.

Here we introduce the multilevel functional GWAS (mfGWAS) method, a novel statistical approach, to handle both multilevel functional responses and high-dimensional predictor space. The mfGWAS method borrows the strength from the statistical methods proposed
by Di et al. (2009) [5] and Crainiceanu et al. (2009) [6]. This approach models multilevel functional data through MFPCA, a combination of functional principal component analysis (FPCA) [7–10] and the multilevel functional mixed effects model [11]. FPCA is a commonly used statistical method to represent random functions in eigenbasis, which is an orthonormal basis of the Hilbert space $L^2$ consisting of the eigenfunctions that describe the major "Modes of Variation" of the data [7–10,120]. Similarly, the mfGWAS method decomposes multilevel functional phenotypes into a functional mixed effects model and uses FPCA to represent the variance components included in the mixed effects model.

On the other hand, when we want to select a group of causative predictors from a high-dimensional predictor space using regression models, regularization techniques are often favored because they can identify non-zero coefficients, enhance model predictability, and avoid overfitting [121,122]. Yuan and Lin (2006) [123] proposed group lasso, which can select significant groups of predictors by encouraging sparsity at the level of groups of predictors. Park and Casella (2008) [124] provided a Bayesian framework for estimating parameters in a lasso regression model. Li et al. (2015) [103] used Bayesian group lasso to select causative SNPs in a functional response model. Our proposed method incorporates Bayesian group lasso to identify causative genes and/or other covariates regulating functional biological phenotypes. The mfGWAS method also has the flexibility on including non-regularized control variables to correct confounding effects such as those of population structure, sample relatedness, and so on. We will discuss these merits of the proposed method in detail later.

The parameters used in the mfGWAS method can be estimated through Markov chain Monte Carlo (MCMC) sampling. This new method can serve not only as a domain-specific method for GWAS problems but also a standard approach to variable selection in regressions with multilevel functional responses. In section 2, we introduce the framework of the mfGWAS method. Section 3 provides a thorough discussion about the Bayesian inference and MCMC sampling of parameters in the mfGWAS method. The details about Bayesian group lasso regularization and choices of prior distributions are also discussed in section 3. Section 4 tests our new method through simulation studies. In section 5, we apply the mfGWAS
method to a real multilevel functional data example. The data presented contain leaf shape measurements of *Populus euphratica*, a species of poplar tree found along river valleys in arid regions ranging from North Africa to northwestern China. Section 6 summarizes our conclusions about the new method.

### 4.2 METHODOLOGY

#### 4.2.1 The mfGWAS method

We consider GWAS problems with multilevel functional phenotypes. Suppose the biological phenotypes are measured on $I$ subjects with $J$ visits/curves per subject. As mentioned before, the repeatedly measured curves on the same subject can be decomposed into a multilevel functional mixed effects model [5,6]. Let $\{Y_{ij}(t), t \in [0,1]\}$ denote the observed curve of $i^{th}$ subject and $j^{th}$ visit, where $i = 1, \ldots, I$, and $j = 1, \ldots, J$. $Y_{ij}(t)$, a function of time or spatial units, can be represented by the following mixed model:

$$Y_{ij}(t) = \mu(t) + V_j(t) + S_i(t) + U_{ij}(t) + \epsilon_{ij}(t), \quad (4.1)$$

where $\epsilon_{ij}(t)$ is an i.i.d. noise with a constant variance $\sigma^2_{\epsilon}$. $V_j(t)$, $S_i(t)$, $U_{ij}(t)$, and $\epsilon_{ij}(t)$ are assumed to be uncorrelated. Equation (1) can be reduced to a two-way ANOVA model if there is no time variable $t$.

$\mu(t)$ is the overall mean function, and $V_j(t)$ is the mean shift from the overall mean function at visit $j$. Given that the sample size is large enough, $\mu(t)$ can be estimated by $\bar{Y}_\cdot(t)$, where $\bar{Y}_\cdot(t)$ is the estimated overall mean function and is obtained by averaging over all subjects and visits. Assuming $\sum_j V_j(t) = 0$, $V_j(t)$ can be estimated by $\bar{Y}_j(t) - \bar{Y}_\cdot(t)$, where $\bar{Y}_j(t)$ is obtained by averaging over all subjects at visit $j$. In many applications, $V_j(t)$ could be set to zero when the repeatedly measured curves are exchangeable within the same subjects and equation (1) is reduced to a one-way functional ANOVA [5,6]. As $\mu(t)$ and $V_j(t)$ are considered as fixed effects, $\bar{Y}_\cdot(t)$ and $\bar{Y}_j(t) - \bar{Y}_\cdot(t)$ should provide a good empirical estimates of these effects in practice. Therefore, we can subtract the estimates of $\mu(t)$ and
$V_j(t)$ from $\{Y_{ij}(t)\}$ and let $y_{ij}(t) = S_i(t) + U_{ij}(t) + \epsilon_{ij}(t)$.

$U_{ij}(t)$ is the visit and subject specific residual from the fixed effects $S_i(t)$ above at visit $j$. $U_{ij}(t)$ is considered a random effect term and can be decomposed by the Karhunen-Loève (KL) expansion: $U_{ij}(t) = \sum_{l=1}^{\infty} \zeta_{ijl} \phi_l(t)$ \([75, 125]\), where \(\{\phi_l(t)\} = \{\phi_1(t), \phi_2(t), \ldots\}\) are eigenfunctions in an orthonormal basis of $L^2[0, 1]$, $\{\zeta_{ij1}, \zeta_{ij2}, \ldots\}$ are principal component scores with $E(\zeta_{ijl}) = 0$ and $\text{Var}(\zeta_{ijl}) = \lambda_l$, and $\{\lambda_l\}$ are the corresponding eigenvalues of $\{\phi_l(t)\}$ with $\lambda_1 \geq \lambda_2 \geq \cdots$. For practical problems, the infinite dimensionality of KL expansion should be truncated to a finite number $L$ based on the proportion of variance explained by the first $L$ principal components. There are many possible approaches to the selection of $L$. For example, Staicu et al. (2010) \([118]\) used a sequence of hypothesis testing for the null hypothesis that a particular eigenvalue is equal to zero to decide the appropriate number of principal components. Some other research also used model selection criterion such as Bayesian information criterion (BIC) to choose optimal $L$ \([96]\). Di et al. (2009) \([5]\) proposed a simple rule for choosing $L$, which is always preferred in practice. According to this rule, an optimal $L$ needs to satisfy the following two conditions:

1. the proportion of variance explained by the first $L$ principal components is larger than a threshold (i.e., 90%);
2. the proportion of variance explained by any additional principal component is less than another threshold (i.e., 1%).

This rule was proved to work well in simulation studies and practical problems by Di et al. (2009) \([5]\).

In $y_{ij}(t)$, $U_{ij}(t)$ is a random functional effect term representing the visit specific residual within each subject, so $\{\phi_l(t)\}$ can be obtained from diagonalizing the empirical estimates of within-subjects covariance matrix:

$$\text{cov}\{U_{ij}(t_1), U_{ij}(t_2)\} = \frac{\sum_i \sum_j y_{ij}(t_1)y_{ij}(t_2)}{IJ} - \frac{\sum_i \sum_{j_1<j_2} y_{ij_1}(t_1)y_{ij_2}(t_2)}{IJ(J-1)}, \quad (4.2)$$
where the first term calculates the total covariance matrix, the second term calculates the between-subjects covariance matrix, and their difference is the within-subjects covariance matrix. It may be necessary to smooth the covariance matrix if it is contaminated with noise. A bivariate thin-plate spline smoother is used to smooth the covariance matrix and obtain an empirical estimate of the covariance operator, as suggested by Yao and Lee (2006) [126] and Di et al. (2009) [5]. In our model, \( \{\zeta_{ijl}\} \) and \( \{\lambda_l\} \) are treated as random values and will be estimated in a joint model with other parameters via Markov chain Monte Carlo (MCMC) sampling.

\( S_i(t) \) is the mean deviation from the overall mean function and the visit mean function for subject \( i \). The effect of each individual subject in an experiment, \( S_i(t) \), can be decomposed into their own genetic effects and environmental factors within subjects. As mentioned above, \( U_{ij}(t) \) is considered a variance component representing random functional effects, while the effects of each individual subject, \( S_i(t) \), are defined as fixed functional effects. We decompose \( S_i(t) \) into the additive and dominant effects of causative genetic markers and into the influences of environmental factors which are included as covariates in our model. After eigenfunctions and truncation lag \( L \) are determined, \( y_{ij}(t) \) can then be rewritten into a linear mixed model. For simplicity and conciseness of presentation, we introduce this model in the framework of a general time point \( t \):

\[
\begin{align*}
y_{ij}(t) &= C(t)X_i^c + A(t)X_i^a + D(t)X_i^d + \sum_{l=1}^{L} \zeta_{ijl} \phi_l(t) + \epsilon_{ij}(t); \\
\zeta_{ijl} &\sim N(0, \lambda_l); \quad \epsilon_{ij}(t) \sim N(0, \sigma^2_\epsilon).
\end{align*}
\] (4.3)

In equation (3), \( C(t) = (C_1(t), ..., C_q(t)) \) is a \( q \)-dimensional vector of the coefficients of covariates’ effects at time point \( t \), \( X_i^c = (X_{i,1}^c, ..., X_{i,q}^c)^T \) is a \( q \)-dimensional column vector of observed covariate values of subject \( i \), and \( q \) is the number of covariates. \( A(t) = (A_1(t), ..., A_p(t)) \) and \( D(t) = (D_1(t), ..., D_p(t)) \) are \( p \)-dimensional vectors of the coefficients of genetic markers’ additive and dominant effects, respectively, at time point \( t \). \( X_i^a = (X_{i,1}^a, ..., X_{i,p}^a)^T \) and \( X_i^d = (X_{i,1}^d, ..., X_{i,p}^d)^T \) are \( p \)-dimensional column vectors of additive and dominant indicators of genotyped genetic markers of subject \( i \), and \( p \) is the total
number of genetic markers. The additive and dominant indicators are defined as follows:

\[
X_{i,m}^{a} = \begin{cases} 
1, & \text{if the genotype of marker } m \text{ is AA}, \\
0, & \text{if the genotype of marker } m \text{ is Aa}, \\
-1, & \text{if the genotype of marker } m \text{ is aa}, 
\end{cases}
\]

\[
X_{i,m}^{d} = \begin{cases} 
1, & \text{if the genotype of marker } m \text{ is Aa}, \\
0, & \text{if the genotype of marker } m \text{ is AA or aa}, 
\end{cases}
\]

\[i = 1, \ldots, I; \ m = 1, \ldots, p.\]

\(X_{i}^{a}\) and \(X_{i}^{d}\) are defined according to the biological meanings of the additive effects \(A(t)\) and the dominant effects \(D(t)\). For the \(m^{th}\) genetic marker, \(A_{m}(t)\) is the expected difference in phenotypic values when substituting one allele for the other one, and \(D_{m}(t)\) is the expected difference in phenotypic values between heterozygote and homozygotes.

In order to approximate the functional coefficients in equation (3), we incorporate the non-parametric basis spline (B-spline) method to quantify the infinite-dimensional functional data. This approximation is mandatory in a longitudinal data setting such that observations are made on irregular grids. Many other methods have been proposed in order to estimate functional coefficients and approximate functional data, such as local polynomial methods and smoothing spline methods. Li et al. (2015) [103] used Legendre polynomials to approximate functional coefficients by a small number of expansion coefficients. Among these optional methods, B-spline is a widely used non-parametric approximation approach in biological and genetic research; see, for example, functional gene expression data analysis in Luan and Li (2003) [127] and Daub et al. (2004) [128]. Akin to Legendre polynomials, functional data can be approximated by a number of expansion coefficients and basis functions using B-spline algorithm. There are some other properties of B-spline that make it powerful in our model. For example:

- B-spline has more flexibility when a local modification is needed on the basis system such as increasing the number of knots [129]. This is especially useful when the measurements are more fluctuant than others on some parts of the time or spatial grids.
Even though we only consider equidistant knots in this paper, one can easily change the B-spline basis functions without modifying other parts of the mfGWAS method.

- A B-spline of degree $d$ is differentiable at the joining points up to the order of $d - 1$. For example, a B-spline of degree 1 is not differentiable at the joining points. This could be helpful when the functional data approximated are not differentiable at some points.

- Once a B-spline system is set up, its application is no more difficult than a polynomial regression [79], as discussed later.

According to the B-spline algorithm, the functional coefficients of $k^{th}$ covariate, $C_k(t)$, can be approximated by a basis expansion system with $v$ basis functions:

$$C_k(t) = \Phi^T(t) c_k, \quad k = 1, ..., q, \quad (4.4)$$

where $c_k = (c_{k1}, ..., c_{kv})^T$ is a $v$ dimensional column vector of expansion coefficients for $k^{th}$ covariate, $\Phi(t) = (\Phi_1(t), ..., \Phi_v(t))^T$ is a $v$ dimensional column vector of basis functions' values at time $t$. Similarly, the additive and dominant effects of $m^{th}$ genetic marker can be rewritten as:

$$A_m(t) = \Phi^T(t) a_m, \quad m = 1, ..., p, \quad (4.5)$$

and

$$D_m(t) = \Phi^T(t) d_m, \quad m = 1, ..., p, \quad (4.6)$$

where $a_m = (a_{m1}, ..., a_{mv})^T$ and $d_m = (d_{m1}, ..., d_{mv})^T$ are $v$ dimensional column vectors of expansion coefficients. After we have set up a B-spline system, any other spline functions of the same degree can be expressed as a liner combination of the B-spline functions. Also, local smoothness and fluctuation can be achieved through B-spline knots and expansion coefficients’ regularization discussed in the next section. As a result, all the functional effects are approximated by the same B-spline system, as shown in equation (4), (5), and (6), which also significantly reduce the model complexity and computational costs.
Finally, let $T_{ij}$ be the number of measurements of subject $i$ at visit $j$, equation (3) becomes

$$y_{ij}(t_\omega) = (\Phi^T(t_\omega)c_1, \ldots, \Phi^T(t_\omega)c_q)X^c_i + (\Phi^T(t_\omega)a_1, \ldots, \Phi^T(t_\omega)a_p)X^a_i$$

$$+ (\Phi^T(t_\omega)d_1, \ldots, \Phi^T(t_\omega)d_p)X^d_i + \sum_{l=1}^{L} \zeta_{ijl}(t_\omega) + \epsilon_{ij}(t_\omega), \quad (4.7)$$

$i = 1, \ldots, I, j = 1, \ldots, J, \omega = 1, \ldots, T_{ij}$.

### 4.2.2 Bayesian inference and MCMC sampling

Equation (7) involves a large number of unknown parameters in the case of ultra-high dimensional predictor space in GWAS and multilevel functional measurements on the same subject. In a high-dimensional regression model, parameters of predictors should be regularized in order to prohibit overfitting and achieve better prediction powers on new data. A popular regularization technique is the lasso of Tibshirani (1996) [17], in which parameters are constrained by $L_1$ least squares. Moreover, the group lasso of Yuan and Lin (2006) [123] partitions predictors into subgroups and imposes a constraint on the sum of $L_2$ norms of the subgroup of parameters. As a result, significant subgroups of parameters can be selected by encouraging sparsity at the group level. In our case, the subgroups of predictors are the expansion coefficient vectors $\|a_m\|$’s and $\|d_m\|$’s, and non-zero expansion coefficient vectors are selected through group lasso regularization. For example, let $\|a_m\|$ be the $L_2$ norm of vector $a_m$. The additive effect of the $m^{th}$ genetic marker is identically zero if and only if $\|a_m\| = 0$. This rule holds for all the functional effects. For equation (7), additive and dominant effects of genetic markers can be partitioned into $2 \times p$ groups of $v$-dimensional vectors and we want to minimize the following penalized least squares:

$$\arg\min_{a_m, d_m} \left( \frac{1}{2} \|y - \hat{y}\|^2 + \lambda_R \sum_{m=1}^{p} \|a_m\| + \lambda^*_R \sum_{m=1}^{p} \|d_m\| \right), \quad (4.8)$$

where $y$ contains $I \times J$ vectors of measurements, $\hat{y}$ contains $I \times J$ vectors of predicted values, and $\lambda_R$ and $\lambda^*_R$ are regularization parameters. Equation (8) does not impose penalty on the expansion coefficients $c_k$’s since the covariates are usually treated as control variables.
in genome-wide association studies. However, one can easily modify the prior distributions discussed later to perform variable selection on the covariates as well.

For traditional lasso regression, $\lambda_R$ and $\lambda_R^*$ are usually selected by cross-validation or using a criterion such as the $C_p$ [123]. Park and Casella (2008) [124] proposed a Bayesian framework so that all the parameters in a lasso regression can be estimated through MCMC sampling. Tibshirani (1996) [17] suggested Laplace priors on the regression parameters regularized. As a direct extension, Bayesian group lasso uses multivariate Laplace priors on subgroups of parameters in a regression model. A potential advantage of the multivariate Laplace prior is that it can pull the parameters of non-causative predictors to 0 faster than multivariate normal or Student-$t$ priors [124]. Therefore, we assume a $v$-dimensional multivariate Laplace prior over each vector of additive effects:

$$
\text{M-Laplace}(a_m | 0, (v\lambda^2_R/\sigma^2_\epsilon)^{-\frac{1}{2}}) = (v\lambda^2_R/\sigma^2_\epsilon)^{\frac{v}{2}} \exp\left(-\left(v\lambda^2_R/\sigma^2_\epsilon\right)^{\frac{1}{2}} ||a_m||\right),
$$

and also a $v$-dimensional multivariate Laplace prior over each vector of dominant effects:

$$
\text{M-Laplace}(d_m | 0, (v\lambda^2_R/\sigma^2_\epsilon)^{-\frac{1}{2}}) = (v\lambda^2_R/\sigma^2_\epsilon)^{\frac{v}{2}} \exp\left(-\left(v\lambda^2_R/\sigma^2_\epsilon\right)^{\frac{1}{2}} ||d_m||\right).
$$

However, the corresponding posterior distribution of Laplace prior does not have a standard form. For efficient sampling, Park and Casella (2008) [124] rewrote Laplace prior as a scale mixture of normal and exponential distributions. Similarly, multivariate Laplace distribution can be rewritten as a scale mixture of multivariate normal and Gamma distributions:

$$
\text{M-Laplace}(a_m | 0, (v\lambda^2_R/\sigma^2_\epsilon)^{-\frac{1}{2}}) \propto \int_0^{\infty} \text{MVN}(a_m | 0, \sigma^2_\epsilon \tau^2_m) \text{Gamma}(\tau^2_m | \frac{v + 1}{2}, \frac{2}{\sigma^2_\epsilon v\lambda^2_R}), (4.9)
$$

where $\frac{v + 1}{2}$ is the shape parameter of Gamma distribution, and $\frac{2}{\sigma^2_\epsilon v\lambda^2_R}$ is the scale parameter of Gamma distribution; see Raman et al. (2009) [130] for the derivation of above equation.

Therefore, we can rewrite the priors of regularized parameters by a hierarchical expansion of multivariate Laplace distribution:

$$
a_m | \sigma^2_\epsilon, \tau^2_m \sim \text{MVN}(a_m | 0, \sigma^2_\epsilon \tau^2_m),
$$
\[ \tau_m^2 | \lambda^2_R \sim \text{Gamma}(\frac{v+1}{2}, \frac{2}{v\lambda^2_R}), \]
\[ d_m | \sigma^2_\epsilon, \tau_m^2 \sim \text{MVN}(d_m | 0, \sigma^2_\epsilon \tau_m^2), \]
\[ \tau_m^2 | \lambda^2_R \sim \text{Gamma}(\frac{v+1}{2}, \frac{2}{v\lambda^2_R}), \]
\[ \sigma^2_\epsilon \sim 1/\sigma^2_\epsilon. \]

We also assume Gamma priors on regularization parameters:
\[ \lambda^2_R \sim \text{Gamma}(\alpha_R, \beta_R), \]
and \[ \lambda^2_R * \sim \text{Gamma}(\alpha_R^*, \beta_R^*). \]

Other than the regularized parameters, Bayesian inference about other parameters is routine:
\[ c_k \sim \text{MVN}(0, \Sigma_{c_k}), \]
\[ \zeta_{ijl} \sim \text{N}(0, \lambda_l), \quad \lambda_l \sim \text{IG}(\alpha_l, \beta_l), \]
where \( \Sigma_{c_k} \) is the covariance matrix of parameter vector \( c_k \), and \( \text{IG}(\alpha, \beta) \) is an inverse gamma distribution with shape parameter \( \alpha \) and scale parameter \( \beta \). The values of \( \alpha \)'s and \( \beta \)'s are small so that these Gamma and inverse Gamma priors are essentially noninformative.

These conjugate priors can guarantee standard-form posterior distributions. We then can use Gibbs sampling [131] to simulate these unknown parameters from their posterior distributions. Assuming \( a_m, \tau_m^2, d_m, \tau_m^2, \zeta_{ijl} \) are conditionally independent and \( \lambda^2_R, \lambda^2_R^*, \lambda_l, \sigma^2_\epsilon \) are independent, the joint posterior distribution of all the parameters can be expressed as:

\[
f(c_k, a_m, \tau_m^2, \lambda^2_R, d_m, \tau_m^2, \lambda^2_R^*, \zeta_{ijl}, \lambda_l, \sigma^2_\epsilon | y) \\
\propto f(y|\text{others}) f(\sigma^2_\epsilon) \prod_{k=1}^q f(c_k) \\
\times \prod_{m=1}^p f(a_m | \tau_m^2) f(\tau_m^2 | \lambda^2_R) f(\lambda^2_R) \prod_{m=1}^p f(d_m | \tau_m^2) f(\tau_m^2 | \lambda^2_R^*) f(\lambda^2_R^*) \\
\times \prod_{i=1}^I \prod_{j=1}^J \prod_{l=1}^L f(\zeta_{ijl} | \lambda_l) \prod_{l=1}^L f(\lambda_l).\]

The posterior distributions of \( a_m, \tau_m^2, \lambda_R^2 \) can be derived as follows:

\[
\begin{align*}
\text{f}(a_m | \text{others}) & \propto f(y | \text{others}) \cdot f(a_m | \tau_m^2) \\
& \propto \exp\left[-\frac{1}{2\sigma^2_{\epsilon}} \left( \frac{a_m^T a_m}{\tau_m^2} + \sum_i \sum_j (y_{ij} - \hat{y}_{ij})^T (y_{ij} - \hat{y}_{ij}) \right) \right] \\
& \propto \exp\left[-\frac{1}{2\sigma^2_{\epsilon} \tau_m^2} \left( a_m^T (\tilde{I} + IJ \tau_m^2 \Phi^T \Phi) a_m - 2\tau_m^2 \sum_i \sum_j (y_{ij} - \hat{y}_{ij(-a_m)}) \Phi a_m \right) \right] \\
& \propto \text{MVN}(\mu_{a_m}, \Sigma_{a_m}),
\end{align*}
\]

where

\[
\begin{align*}
\mu_{a_m} &= (\tilde{I} + IJ \tau_m^2 \Phi^T \Phi)^{-1} \left[ \tau_m^2 \sum_i \sum_j (y_{ij} - \hat{y}_{ij(-a_m)}) \Phi \right]^T, \\
\Sigma_{a_m} &= \sigma^2_{\epsilon} \tau_m^2 (\tilde{I} + IJ \tau_m^2 \Phi^T \Phi)^{-1},
\end{align*}
\]

and \( \tilde{I} \) is a \( v \) by \( v \) identity matrix.

\[
\begin{align*}
\text{f}(\tau_m^2 | \text{others}) & \propto f(a_m | \tau_m^2) \cdot f(\tau_m^2 | \lambda_R^2) \\
& \propto (\tau_m^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \left( \frac{a_m^T a_m}{\tau_m^2 \sigma^2_{\epsilon}} + v \lambda_R^2 \tau_m^2 \right) \right],
\end{align*}
\]

and thus \( f(1/\tau_m^2 | \text{others}) \propto \text{IG}(v \lambda_R^2; \sqrt{\frac{v \lambda_R^2 \sigma^2_{\epsilon}}{a_m^T a_m}}) \).

\[
\begin{align*}
\text{f}(\lambda_R^2 | \text{others}) & \propto \prod_{m=1}^{p} p f(\tau_m^2 | \lambda_R^2) \cdot f(\lambda_R^2) \\
& \propto (\lambda_R^2)^{\alpha - 1} \exp(-\beta_R \lambda_R^2) \times \left( \frac{v \lambda_R^2}{2} \right)^{\frac{vp + p}{2}} \exp\left[-\lambda_R^2 \left( \frac{v \sum_m \tau_m^2}{2} \right) \right] \\
& \propto \text{Gamma}(\alpha_R + \frac{vp + p}{2}, \beta_R + \frac{v \sum_m \tau_m^2}{2}).
\end{align*}
\]

Likewise, the posterior distributions of \( d_m \) is \( f(d_m | \text{others}) \propto \text{MVN}(\mu_{d_m}, \Sigma_{d_m}) \),

where
\[ \mu_{dm} = (I + \tau_m^{*2}\Phi^T\Phi)^{-1}\sum_i \sum_j (y_{ij} - \hat{y}_{ij(-d_m)})\Phi_i^T, \]

\[ \Sigma_{dm} = \sigma^2_{r_m^{*2}}(I + \tau_m^{*2}\Phi^T\Phi)^{-1}. \]

Also, we have \( f(1/\tau_m^{*2}|\text{others}) \propto IG(v\lambda_R^{*2}, \sqrt{v\lambda_R^{*2}\sigma^2_{\tau_m^{*2}}/d_m}) \) and \( f(\lambda_R^{*2}|\text{others}) \propto \Gamma(a_R + \frac{vp + p}{2}, \frac{\sum_m \tau_m^{*2}}{2}). \) The posterior distribution of unregularized parameter \( c_k \) is \( f(c_k|\text{others}) \propto \text{MVN}_v(\mu_{c_k}, \Sigma_{c_k}), \)

where

\[ \mu_{c_k} = (\Sigma_{c_k}^{-1} + J(X_{i,k}^c)^2\Phi^T\Phi)^{-1}\sum_j \sum_i (y_{ij} - \hat{y}_{ij(-c_k)})(X_{i,k}^c \Phi_i^T, \]

\[ \Sigma_{c_k} = \sigma^2_{r_m^{*2}}(\Sigma_{c_k}^{-1} + J(X_{i,k}^c)^2\Phi^T\Phi)^{-1}. \]

Let \( \zeta_{ij} = (\zeta_{ij1}, ..., \zeta_{ijL})^T \) be a \( L \)-dimensional column vector of principal component scores, \( \Lambda = \text{diag}(\lambda_1, ..., \lambda_L) \) be a diagonal matrix of eigenvalues, and \( \Psi = (\phi_1^T, ..., \phi_L^T) \) be a matrix of eigenfunctions. The posterior distribution of \( \zeta_{ij} \) and \( \lambda_l \) are:

\[ f(\zeta_{ij}|\text{others}) \]

\[ \propto f(y_{ij}|\text{others})f(\zeta_{ij}|\Lambda) \]

\[ \propto \exp\left[-\frac{(y_{ij} - \hat{y}_{ij})^T(y_{ij} - \hat{y}_{ij})}{2\sigma^2_{\zeta}} - \frac{1}{2} \zeta_{ij}^T(\Lambda)^{-1}\zeta_{ij}\right] \]

\[ \propto \exp\left[-\frac{1}{2\sigma^2_{\zeta}}(\zeta_{ij}^T(\Psi^T\Psi + \sigma^2_{\zeta}(\Lambda)^{-1})\zeta_{ij} - 2(y_{ij} - \hat{y}_{ij(-c_k)})(\Psi^T\zeta_{ij}))\right] \]

\[ \propto \text{MVN}_L(\mu_{\zeta_{ij}}, \Sigma_{\zeta_{ij}}), \]

where

\[ \mu_{\zeta_{ij}} = (\Psi^T\Psi + \sigma^2_{\zeta}(\Lambda)^{-1})^{-1}[(y_{ij} - \hat{y}_{ij(-c_k)})\Psi_i^T, \]

\[ \Sigma_{\zeta_{ij}} = \sigma^2_{\zeta}(\Psi^T\Psi + \sigma^2_{\zeta}(\Lambda)^{-1})^{-1}. \]
and

\[
f(\lambda|\text{others}) = f(\lambda_1) \prod_{i=1}^{I} \prod_{j=1}^{J} f(\zeta_{ij}|\lambda_i) \\
\propto \lambda_i^{-(\alpha_l + \frac{1}{2}IJ+1)} \exp(-\frac{1}{2} \sum_i \sum_j \zeta_{ij}^2 + \beta_l) \\
\propto IG\left(\frac{1}{2}IJ + \alpha_l, \frac{1}{2} \sum_i \sum_j \zeta_{ij}^2 + \beta_l\right).
\]

Last, the posterior distribution for \(\sigma^2_\epsilon\) is:

\[
f(\sigma^2_\epsilon|\text{others}) = f(y|\text{others}) f(\sigma^2_\epsilon) \\
\propto \exp\left(-\frac{1}{2\sigma^2_\epsilon} \sum_i \sum_j (y_{ij} - \hat{y})^T (y_{ij} - \hat{y})\right) \times \left(\sigma^2_\epsilon\right)^{-\frac{1}{2} \sum_i \sum_j T_{ij} + 1} \\
\propto \text{Scale-inv-}\chi^2\left(\sum_i \sum_j T_{ij}, \frac{\sum_i \sum_j (y_{ij} - \hat{y})^T (y_{ij} - \hat{y})}{\sum_i \sum_j T_{ij}}\right),
\]

where \text{Scale-inv-}\chi^2(A, B) is an scaled inverse chi-squared distribution with degrees of freedom parameter \(A\) and scale parameter \(B\).

All the unknown parameters can be estimated by drawing random samples from their corresponding posterior distributions. We do not need to reject any samples because all of these posterior distributions have a standard form. After the MCMC chains have converged, we can select causative markers according to their corresponding parameter estimates. As mentioned before, the assumption is that the effects of the \(m^{th}\) genetic marker is identically zero if and only if \(||a_m|| = 0\) and \(||d_m|| = 0\). In practice, we propose to rank the genetic markers based on the \(L_2\) norms of their expansion coefficients, \(||a_m||\) and \(||d_m||\), and then select an optimal model size using the Bayesian information criterion (BIC) [132]. The performance of the proposed algorithm is tested in a variety of simulation settings.
4.3 SIMULATION STUDIES

In this section, the mfGWAS method is tested on simulated data. We use the mfGWAS method and criterion described above to select causative genetic markers. The functional effects of one covariate is also included in our simulations assuming that there is a need for a control variable, but its parameters are not regularized and not used for selection purpose. We simulate datasets with the number of subjects $I = 100$ or $300$, the number of visits per subject $J = 5$, the number of measurements per visit $T = 50$, the number of genetic markers $p = 3000$, and an additional continuous covariate.

The genotypes of all the genetic markers are generated independently from a binomial distribution with a random minor allele frequency ($X_m \sim \text{Binomial}(2, p_m)$), where $p_m$ is the minor allele frequency of the $m^{th}$ marker, and $p_m$ is simulated from a uniform distribution ($p_m \sim \text{Uniform}(0, 1)$). The additive and dominant indicator matrices are then derived based on their genotypes. The covariate is assumed to be standardized and is generated from a standard normal distribution $N(0, 1)$.

There are two methods used to simulate the functional coefficients/effects: 1) the first method is to generate phenotypic values using equation (7) by setting parameters’ values such as the expansion coefficients of B-spline functions; 2) the second method is to set the functional effects as general functions of time, which is not model specific and a good way to test the power of our model in general cases. For consistency, we assume the functional effects can be well approximated by five basis functions ($v = 5$) in our simulation studies, in which cubic splines and default knots at 0, 0.5 and 1 are used. In practical problems, there are more systematic approaches to set up a B-spline basis system; see, for example, de Boor (1978) [133] and Schumaker (2007) [134]. The choices of spline orders and knots depend on the global complexity and local fluctuation of curves. Among 3000 candidate genetic markers, we assume five of them are truly causative, for which their additive and/or dominant effects are not identically zero. Specifically, we assign values to $c_1, a_1, a_2, a_3, d_3, a_4, a_5$ in Simulation I, and set the functional effects $C_1(t), A_1(t), A_2(t), A_3(t), D_3(t), A_4(t), A_5(t)$ as functions of time in Simulation II. Among the five causative markers, only one of them has both non-
Table 4.1: Coefficients and functions used in Simulation I and Simulation II

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Parameter</th>
<th>Expansion Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$c_1$</td>
<td>1 0 -4 -1 4</td>
</tr>
<tr>
<td></td>
<td>$a_1$</td>
<td>4 1 3 4 2</td>
</tr>
<tr>
<td></td>
<td>$a_2$</td>
<td>3 3 0 -4 3</td>
</tr>
<tr>
<td></td>
<td>$a_3$</td>
<td>2 5 0 0 0</td>
</tr>
<tr>
<td></td>
<td>$d_3$</td>
<td>1 1 5 1 1</td>
</tr>
<tr>
<td></td>
<td>$a_4$</td>
<td>4 3 0 1 3</td>
</tr>
<tr>
<td></td>
<td>$a_5$</td>
<td>1 1 1 1 -5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Parameter</th>
<th>Functions of Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$C_1(t)$</td>
<td>$t/10$</td>
</tr>
<tr>
<td></td>
<td>$A_1(t)$</td>
<td>$10\sqrt{t}$</td>
</tr>
<tr>
<td></td>
<td>$A_2(t)$</td>
<td>$\exp(2t)$</td>
</tr>
<tr>
<td></td>
<td>$A_3(t)$</td>
<td>$\log(0.05t)$</td>
</tr>
<tr>
<td></td>
<td>$D_3(t)$</td>
<td>$5$</td>
</tr>
<tr>
<td></td>
<td>$A_4(t)$</td>
<td>$1/t$</td>
</tr>
<tr>
<td></td>
<td>$A_5(t)$</td>
<td>$10t$</td>
</tr>
</tbody>
</table>

zero additive effects and non-zero dominant effects. All the other functional effects are set to identically zero. The expansion coefficients and functions used in these two simulation designs are listed in Table 4.1.

For the subject/visit specific residual $U_{ij}(t)$, we set the true $L$ to be 4 and the eigenvalues \{\lambda_l\} to be 1, 0.9, 0.6, and 0.5. Then the subject/visit specific principal component scores \{\zeta_{ij}\} are generated from their corresponding eigenvalues: $\zeta_{ij} \sim N(0, \lambda_l)$. The corresponding eigenfunctions are assumed to be mutually orthogonal and are set as follows:

$$\{\phi_1(t), \phi_2(t), \phi_3(t), \phi_4(t)\} = \{\sqrt{2}\sin(4\pi t), \sqrt{2}\cos(4\pi t), \sqrt{2}\sin(8\pi t), \sqrt{2}\cos(8\pi t)\}.$$ 

Last, the variance of the i.i.d. noise $\epsilon_{ij}(t)$, $\sigma^2_\epsilon$, is set to be 0 (no noise) or 1 (noisy). In the simulation studies, we assume all of these information are unknown to researchers, so the eigenvectors are extracted from empirical covariance matrix and the number of principal components $L$ are decided using the rule proposed by Di et al. (2009) [5].

In order to assess the variable selection performance of the mfGWAS method, two types of criteria are used:
Table 4.2: Variable selection performance on simulated datasets

<table>
<thead>
<tr>
<th>Simulation</th>
<th>(\sigma_\epsilon)</th>
<th>(I)</th>
<th>Overall Performance</th>
<th>Individual Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>Type I Error</td>
<td>(p_1)</td>
<td>(p_2)</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>100</td>
<td>0.96</td>
<td>0.0056</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>300</td>
<td>1.00</td>
<td>0.0028</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>100</td>
<td>0.86</td>
<td>0.0004</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>300</td>
<td>1.00</td>
<td>0.0003</td>
</tr>
<tr>
<td>II</td>
<td>0</td>
<td>100</td>
<td>0.78</td>
<td>0.0005</td>
</tr>
<tr>
<td>II</td>
<td>0</td>
<td>300</td>
<td>0.92</td>
<td>0.0003</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>100</td>
<td>0.46</td>
<td>0.0013</td>
</tr>
<tr>
<td>II</td>
<td>1</td>
<td>300</td>
<td>0.88</td>
<td>0.0030</td>
</tr>
</tbody>
</table>

- **Power**: the average proportion of successfully identified causative markers among all simulation replicates;
- **Type I error**: the average proportion of mistakenly identified non-causative markers among all simulation replicates.

Besides the overall performance, we also calculate the power of successfully selecting each individual causative markers, denoted by \(p_1\), \(p_2\), \(p_3\), \(p_4\), & \(p_5\). The simulation results are shown in Table 4.2.

Clearly, the number of subjects \(I\), or, say, sample size, plays a critical role in how well the mfGWAS method can identify truly causative genetic markers. This makes perfect sense in a high-dimensional data setting. The simulation results indicate that the selection power can be equal to or close to 100% and the type I error rate can be very low if the sample size is large enough. The noise level, \(\sigma_\epsilon\), and the simulation designs play less important roles in the variable selection performance of the mfGWAS method. A higher noise level is likely to affect the selection power and type I error rate in the simulation studies. However, for a reasonably large sample size, the variable selection power of the mfGWAS method is still promising even if the observations are noisy. Interestingly, in simulation I, the simulation setting with added noise (\(\sigma_\epsilon = 1\)) produces slightly less false positives than the simulation setting without added noise (\(\sigma_\epsilon = 0\)) does, which suggests that the impact of increased noise level can be ignored and the results are likely to be incurred by randomness in data.
simulation as other parameters in the simulation settings are exactly the same. According to
the results of simulation II, the impact of different simulation designs is also limited when the
sample size is large enough ($I = 300$), which suggests that the mfGWAS method can handle
general functional phenotypes other than the equation (7) specific expansion coefficients as
used in simulation I.

The estimated $a_m$ and $d_m$ of non-causative markers are not exactly zeros due to the
random sampling errors of MCMC chains, but, ideally, they should be very close to zeros
and can be discarded as noises in the final model. The usage of model size selection criterion
BIC avoids the necessity of specifying a threshold for the magnitude of functional effects
of a particular genetic marker, which is often ambiguous in practice. According to the
simulation results, BIC model selection can select an optimal model size and effectively
reduce the number of false positives identified.

4.4 REAL DATA ANALYSIS

In this section, we present a real data example using the mfGWAS method. The
data presented contain leaf shape measurements of *Populus euphratica*, and have not been
published in other literature. The *Populus euphratica*, also known as the Euphrates Poplar
or Desert Poplar, are usually found along river valleys in arid regions. Its natural habitat
ranges from North Africa to northwestern China. The leaves of *Populus euphratica* are
highly polymorphic and irregular, as shown in Figure 4.1. A shape descriptor based on
centroid-radii model and wavelet transform [94] is used to quantify leaf shape.

For each leaf image, the coordinates of all points on the leaf margin are recorded one
by one assuming the margin is a closed one-pixel wide curve, and the coordinate of centroid
is the mean coordinates of all points on the margin. Then a vector of radii between margin
pixels and centroid is calculated. This radii vector is invariant to positions or translations
and is used as a raw shape vector for further analysis. After the preprocessing step, all
the shape curves need to be aligned so that they are invariant to scales and rotations.
In order to remove scale effects, all the shape curves are normalized by dividing them by
their corresponding Euclidean norms. It is more complicated to remove the rotation effects.
According to Kong et al. (2007) [94], we can create a directional radii vector by making the minimum radius point the starting point of the vector and directing the rest of the vector towards the maximum radius point. This rearrangement of shape vector can remove rotation effects because the maximum and the minimum radii are invariant to different rotations of an object. Once all the images have gone through the above process, a Haar wavelet transform \([135,136]\) is applied to all the shape curves so that they have a common length. In our data analysis, all the leaf margins are transformed into \(910 \times 1\) dimensional shape vectors, and the shape curves are scaled to fit into the time interval \([0, 1]\). After several transformations, the values in a shape curve range from 0.3 to 1, and no longer represent the original scale of a leaf image.

There are shape measurements of 421 \textit{Populus euphratica} trees contained in our data, and each tree has 25 leaves sampled and 104 genetic markers genotyped \((I = 421, J = 25, p = 104)\). The 25 leaves of each tree were sampled together and there is no specific order of sampling, so the visit effect \(V_j(t)\) is set to zero. Then we apply the mfGWAS method to select significant additive and dominant effects of genetic markers, where all effects are likely to be functions of time or spatial units. In our real data example, the
Fig. 4.2: Additive effects of selected markers

The number of subjects $I$ is larger than the number of predictors $p$, which can guarantee a perfect statistical power according to the results of simulation studies. For the visit and subject-specific residual, $U_{ij}(t)$, the number of principal components $L$ is decided based on the following two conditions: 1) the proportion of variance explained by the first $L$ principal components is larger than 90%; 2) the proportion of variance explained by any additional principal component is less than $1/910 = 0.1\%$. The functional effects are approximated by seven basis functions ($v = 7$), in which cubic splines and 5 equidistant knots (including the knot at time point 0) are used.

In order to estimate the unknown parameters in the mfGWAS method, we perform 1,000 burn-in iterations followed by 2,000 iterations for the MCMC chains. There are eight genetic markers identified as causative by the mfGWAS method: $ORPM_{137}$, $GCPM_{1158}$, $ORPM_{29}$, $GCPM_{2658}$, $GCPM_{51}$, $ORPM_{127}$, $U78791$, and $U536$. We then reconstruct their additive and dominant effects from their estimated expansion coefficients. These functional effects are shown in Figure 4.2 and Figure 4.3, respectively, where their 95% confidence intervals are also estimated according to Jones et al. (2006) [137] and plotted. Recall
that the additive effect of a marker stands for the average difference in phenotypic value when substituting one allele for the other allele, and the dominant effect of a marker stands for the average difference in phenotypic value between heterozygote and homozygotes. In order to visualize the relationship between functional effects of genetic markers and subject effects of shape curves, we average 25 shape curves from each tree to obtain empirical estimates of subject effects and group the 421 averaged curves by their corresponding genotypes to explore the effects of a genetic marker on a tree’s subject effect. As shown in Figure 4.4 and Figure 4.5, the 421 curves according to three genotypes of four example markers were enveloped in their corresponding colored bands. In order to achieve a better visualization, the bands are drawn with a narrower range from the 25th to the 75th percentile of radii values at each measurement location, which can remove some extreme observations. The three bands corresponding to different genotypes of causative markers U78791 and U536 are shown in Figure 4.4. According to Figure 4.2 and Figure 4.3, the dominant effect of U78791 is negative at around 0.5, and its additive effect is positive at around 0.5. This can be verified by Figure 4.4A, in which the curves of genotype AA tend to have higher
radii at around 0.5 than those of \textit{Aa} and \textit{AA}, and the bands of \textit{Aa} and \textit{AA} mostly overlap. The other marker in Figure 4.4, \textit{U536}, indicates even stronger visual effects than those of \textit{U78791} in terms of bands’ separation, probably due to the smaller variation of curves within genotypes. Both of its additive and dominant effects are positive at around 0.5. As a result, the curves of heterozygote genotype \textit{Aa} tend to have lower radii at around 0.5 than those of the genotype \textit{AA} and \textit{aa}, and the bands of \textit{AA} and \textit{aa} mostly overlap. The additive and dominant effects of other causative markers can be interpreted in a similar way. Compared to Figure 4.4, the three bands of non-causative markers \textit{Pe2} and \textit{Pe8} mostly overlap in Figure 4.5, which indicates that their additive and dominant effects are not significant. The eight selected markers have non-zero additive and dominant effects on different aspects of the leaf shapes of \textit{Populus euphratica}, as summarized in Figure 4.2 and Figure 4.3. A major advantage of the mfGWAS method over traditional single-SNP based models is that it can fit various effects and parameters jointly, and can also successfully identify polygenetic effects and weaker associations, which could be challenging for single-SNP based GWAS models [138].

4.5 DISCUSSION

The multilevel functional GWAS method discussed in this paper is motivated by current data containing multilevel functional phenotypes and a high-dimensional predictor space. The parameters of the mfGWAS are easy to generate in the proposed Bayesian framework and require minimal fine tuning. This new method is not only applicable to genetic association studies, but also to any other statistical problems with similar data settings.

The mfGWAS method uses a joint mixed effects model and decomposes the repeatedly measured functional data into an overall mean effect, a visit effect, a subject effect, a visit and subject-specific residual, and white noise with constant variance. Intuitively, the effect of each individual subject can be explained by potential genetic and covariates’ effects. For GWAS problems, the number of predictors is usually much larger than the sample size, which renders simple regression models unable to identify causative predictors. Regularization regression methods are commonly used in practice to enhance model simplicity and
predictability and prevent overfitting. The Bayesian regularized regression procedure employed in the mfGWAS method is a direct extension of the Bayesian lasso [124], the group lasso [123], and Bayesian group lasso for non-parametric functional response models [103]. In particular, we propose to use a Bayesian analysis via MCMC sampling to estimate parameters in the mfGWAS method simultaneously. The prior distributions used all guarantee conjugacy and the posterior distributions all have closed forms so that the efficient Gibbs sampling algorithm can be used. For model robustness, the researcher can choose which parameters to be regularized and only include a final model of particular interest. Also, inspired by Di et al. (2009) [5] and current studies on functional principal component analysis (i.e., [7–10]), we use Karhunen-Loève (KL) expansion to model the visit and subject-specific residual, which is a natural decomposition of functional variability into different principal directions of variation.

In section 4, we test the power of the mfGWAS method on identifying significant genetic markers in simulation studies. The simulation results indicate that when the sample size is large enough, the variable selection performance of mfGWAS method is promising regardless of simulation designs and noise level of observations. In a real data analysis example, our method selects eight causative genetic markers from 104 candidate markers. The causative markers identified are \textit{ORPM\_137}, \textit{GCPM\_1158}, \textit{ORPM\_29}, \textit{GCPM\_2658}, \textit{GCPM\_51}, \textit{ORPM\_127}, U78791, and U536. Their additive and dominant effects are estimated associated with 95% confidence interval, and these effects can be well verified through the analysis of the average shape of each tree, which is an empirical estimate of each tree’s subject effect.

The mfGWAS method involves a large number of parameters, especially when there are tens of thousands of genes or thousands of subjects with multiple visits/curves. Although MCMC algorithm has theoretical advantages on solving problems involving a large number of unknown parameters and handling parameter uncertainty, the computational costs of Bayesian approaches are usually higher than those of frequentist methods. On the other hand, the simulation studies have proved that a large sample size can significantly improve the statistical power. It would be promising to improve statistical power and reduce com-
putational cost at the same time. Some methodologies have been proposed to do feature screening on a ultrahigh-dimensional predictor space, such as Li et al. (2012) [30] and Fan and Lv (2008) [139]. The extension of these methods for functional data can be used as a preprocessing step for our mfGWAS method. If the mfGWAS method can focus on a smaller set of pre-selected candidate genes, we can not only gain statistical power but also conquer the limitation of computational power. As a result, more potential predictors such as plasticity effects, the responses of phenotypes to environmental factors [140], can be well-explored without concerning the model complexity. Also, high-order interactions between genetic markers and other covariates can be incorporated into the model, which can provide a better understanding of the underlying mechanism of complex biological phenotypes.
Fig. 4.4: Visualizing subject effects of causative markers
Fig. 4.5: Visualizing subject effects of non-causative markers
In this dissertation three novel statistical models are proposed to handle challenging gene-shape association selection problems.

The first article demonstrates that the proposed BMVS method significantly outperforms other state-of-the-art multivariate variable selection approaches because it has the highest powers in selecting the exact true model while also attaining the lowest false discovery rates throughout different simulation examples. Even though this article primarily focuses on variable selection, the real data examples have also shown that the BMVS method is very competitive regarding prediction accuracy. If prediction is the ultimate goal, one can use other criteria such as cross-validation or bootstrap to substitute AICc during the variable selection process. In general, Bayesian methods are more computationally expensive than the frequentist counterparts due to their long iteratively updating process. However, the introduction of initial screening methods such as DC sure independence screening (i.e., DC+BMVS) dramatically reduces the computational cost of the BMVS method and makes its use feasible for ultrahigh-dimensional data. This characteristic of the BMVS method would be useful for genome-wide association studies, where only a small subset of genes are truly active and the signal-to-noise ratio is very small.

The second article aims to detect important genetic and/or environmental factors associated with shape trait, where the shape that is accurately described by a high-dimensional response curve together with a large scale of single-nucleotide polymorphisms (SNPs) and/or environmental factors. The proposed FunFor approach introduces a very new direction for the shape applications in that it models the entire shape curve instead of modeling the PCs one by one. It describes the trajectory of a curve that changes over time (or spatial units) and also shows subject-specific events that different individuals may experience. It not only makes response predictions for new individuals when only predictors are available (as tra-
ditional RF approach does), but it also predicts the time-to-event in the future or response values under new time points for both observed individuals and brand new individuals. The FunFor approach has a great potential to be applied to a wide range of practical problems in which analysts want to delineate which of the predictors are associated with the variation of dynamic trajectory of these response curves, or in which they want to predict the curve response using a set of predictors.

The third article shows that the mfGWAS method uses a joint mixed effects model and decomposes the repeatedly measured functional data into an overall mean effect, a visit effect, a subject effect, a visit and subject-specific residual, and white noise with constant variance. Intuitively, the effect of each individual subject can be explained by potential genetic and covariates’ effects. For GWAS problems, the number of predictors is usually much larger than the sample size, which renders simple regression models unable to identify causative predictors. In particular, we propose to use a Bayesian analysis via MCMC sampling to estimate parameters in the mfGWAS method simultaneously. The prior distributions used all guarantee conjugacy, and the posterior distributions all have closed forms so that the relatively efficient Gibbs sampling algorithm can be used. We test the power of the mfGWAS method on identifying significant genetic markers in simulation studies. The simulation results indicate that when the sample size is large enough, the variable selection performance of mfGWAS method is promising regardless of simulation designs and noise level of observations. Although MCMC algorithm has theoretical advantages on solving problems involving a large number of unknown parameters and handling parameter uncertainty, the computational costs of Bayesian approaches are usually higher than those of frequentist methods. On the other hand, the simulation studies have proved that a large sample size can significantly improve the statistical power. It would be promising to improve statistical power and reduce computational cost at the same time. Some methodologies have been proposed to do feature screening on an ultrahigh-dimensional predictor space, such as Li et al. (2012) [30] and Fan and Lv (2008) [139]. The extension of these methods for functional data can be used as a preprocessing step for our mfGWAS method. If the mfG-
WAS method can focus on a smaller set of pre-selected candidate genes, we can not only gain statistical power but also conquer the limitation of computational power. As a result, more potential predictors such as plasticity effects, the responses of phenotypes to environmental factors [140], can be well-explored without concerning the model complexity. Also, high-order interactions between genetic markers and other covariates can be incorporated into the model, which can provide a better understanding of the underlying mechanism of complex biological phenotypes.
REFERENCES


CURRICULUM VITAE

Xiaotian Dai
( November 2017)

CAREER OBJECTIVE

Obtain a career opportunity as a data scientist / statistician; bring strong Mathematics, analytical and programming skills.

EDUCATION & HONOR

• B.S. in Statistics, Utah State University, Logan, UT.

• Ph.D. in Statistics, Utah State University, Logan, UT.

• Merit in Research Accomplishment (2017), Department of Mathematics and Statistics, Utah State University, Logan, UT.

EXPERIENCE

• Enterprise Data Scientist, Micron Technology Inc., Boise, ID (July, 2017 - Present)

• Summer Intern (Data Science), IM Flash Technologies (An Intel & Micron Joint Venture), Lehi, UT (May, 2016 - August, 2016)

• Graduate Research Assistant / Teaching Assistant, Utah State University, Logan, UT (September, 2011 - May, 2017)