STATISTICAL METHODS FOR BRAIN IMAGING GENOMICS

Yue Shan

A proposal submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2021

Approved by:

Hongtu Zhu

Yun Li

Kinh Truong

Yuchao Jiang

Jason Stein

ABSTRACT

Yue Shan: Statistical Methods for Brain Imaging Genomics
(Under the direction of Hongtu Zhu and Yun Li)

Brain Imaging genetic studies examine genetic basis of brain images to better understand the genetic impact on behavior and disease phenotypes. Methods for identifying genetic associations with voxelwise brain imaging data have evolved from parallel analysis on each voxel to incorporating spatial smoothness and correlation to increase statistical detection power. Challenges still exist on the joint analysis of imaging data and genetic data, including imperfect alignment of affected regions and registration error, low signal to noise ratio in high-dimensional data, complex relationships, high computation complexity, and between-study heterogeneity. To address these issues, the following methods are proposed.First, to deal with imperfect alignment and registration error in brain imaging data, we proposed a region-based functional genome-wide association detection method, which also reduces computation burden as compared to standard voxelwise methods. The method summarizes regional voxelwise measurements into density curves. The non-parametric ball covariance test is then used to detect association between the log-quantile transformed regional densities and genetic markers. We compared the ball covariance test with other state-of-the-art methods on simulated datasets and demonstrate good sensitivity and specificity of our method. Second, we combined functional partial least squares with distance correlation to reduce computation burden of high dimensional data and allow flexible characterization of the imaging-genetic relationship. Third, given imaging-genetic data from more than one studies, we theoretically compared the ensembled learner and merged learner in the prediction problem, where learners are trained using the multivariate varying coefficient model and multi-study data are assumed to come from a mixed model, where the mixed effect represents inter-study heterogeneity.

To my mentor, parents and friends, I couldn't have done this without you. Thank you for all of your support along the way.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: LITERATURE REVIEW 3

2.1 Statistical Methods for Association Detection 3

2.1.1 Association Testing Methods Used in Genomics 4

2.1.2 Functional Frameworks 10

2.2 Covariance & High-Dimensional Projection 14

2.2.1 Covariance and Correlation Measurements 14

2.2.2 Weight Projection 21

2.3 Between-Study Heterogeneity 23

2.3.1 Meta-Analysis 24

2.3.2 Refined Meta-Analysis 25

2.3.3 Meta-Analysis for Functional Models 26

2.3.4 Comparison with Mega-Analysis 26

CHAPTER 3: REGION-BASED FUNCTIONAL METHOD 29

3.1 Introduction 29

3.2 Method 30

3.2.1 Regional Distributions and Transformation 31

3.2.2 Region-based Statistical Test 33

3.2.3 Global Screening 37

3.3 Simulation 40

3.3.1 Single voxel/region comparison 40

3.3.2 Comparison for genetic data 44

3.3.3 Registration error consideration 47

3.3.4 Global Screening 49

3.4 Real Data Application 55

3.4.1 Comparison with FGWAS 59

3.5 Discussion 59

CHAPTER 4: FPLS WITH DISTANCE CORRELATION 67

4.1 Introduction 67

4.2 Methods 68

4.2.1 The APLS algorithm 69

4.2.2 The FPLS-DC algorithm 70

4.2.3 Hypothesis Testing 73

4.2.4 Two step estimation algorithm 76

4.3 Simulation 76

4.3.1 Data Generation 76

4.3.2 Screening 77

4.3.3 FPLS-DC and Statistical Tests 79

4.4 ADNI Application 82

4.4.1 Data Description 82

4.4.2 Analysis Procedure 83

4.4.3 Results and Interpretations 85

4.5 Discussion 88

CHAPTER 5: ENSEMBLE- VS MERGE-BASED LEARNERS 92

5.1 Introduction 92

5.2 Method 94

5.2.1 MVCM for One Study 95

5.2.2 MVCM for Multiple Studies 97

5.2.3 Theoretical Conclusions 99
5.3 Simulation 102
5.4 Imaging Genomic Application 106
5.4.1 Data Description 107
5.4.2 Study Design and Implementation 109
5.4.3 Result 112
5.5 Discussion 113
CHAPTER 6: DISCUSSION 115
6.1 Future Research for Project 1 115
6.2 Future Research for Project 2 116
6.3 Future Research for Project 3 117
APPENDIX 1: METHOD 1 - ACCELERATED BCOR 118
APPENDIX 2: METHOD 1 - FGWAS RESULT DETAILS 128
APPENDIX 3: METHOD 2 - EVALUATION OF ESTIMATION 130
APPENDIX 4: METHOD 2 - VOXELWISE TYPE I ERROR 132
APPENDIX 5: METHOD 2 - NON-THRESHOLDED TEST 133
APPENDIX 6: METHOD 3 - THEORETICAL DETAILS 135
APPENDIX 7: METHOD 3 - MORE ON TRANSITION POINT 145
REFERENCES 147

LIST OF TABLES

2.1 Summary of methods reviewed in Section 2.1.1 15

2.2 Summary of methods reviewed in Section 2.1.2 15

2.3 Summary of methods reviewed in Section 2.2. 23

2.4 Summary of methods reviewed in Section 2.3. 28

3.1 Dimensions of variables and applicable methods 42

3.2 Rejection rates of 10 simulation settings 43

3.3 Type I error rate of BCov 44

3.4 Power comparison for smaller effect size 45

3.5 Power comparison under nonlinear relationship 46

3.6 Summary of 13 settings 49

3.7 Significant SNPs on the left hippocampus surface (a) 57

3.8 Significant SNPs on the left hippocampus surface (b) 58

3.9 Significant SNPs on the right hippocampus surface (a) 60

3.10 Significant SNPs on the right hippocampus surface (b) 61

3.11 Computation time (3.3.1) 66

3.12 Computation time (3.3.3) 66

4.1 Inclusion rates of causal SNPs 78

4.2 Type I Error for Local and Global Tests 81

4.3 Power for Local and Global Tests 82

6.1 Raw voxelwise type I error rates for non-causal SNPs 132

6.2 Raw voxelwise type I error rates for causal SNPs 132

6.3 Type I Error and Power 134

LIST OF FIGURES

3.1 Steps in method 1 31

3.2 Type I errors using BCov and GBJ 50

3.3 Power of BCov 51

3.4 Power of GBJ 52

3.5 Power comparison between methods 53

3.6 Inclusion rates 56

3.7 Density curves of region 47 by dosage levels 62

3.8 Comparison with FGWAS (top SNP from our method) 63

3.9 Comparison with FGWAS (top SNP from FWAS) 64

4.1 Top SNP Ranking: Screening VS opt-dCor 86

4.2 P-values for screened SNPs 86

4.3 P-values for significant regions 89

4.4 Subregions of the hippocampus surface. 90

5.1 Boxplots of means of $\log \frac{MPE_e^2(\mathcal{S})}{MPE_m^2(\mathcal{S})}$ over 100 replicates 105

5.2 Boxplots of 100 bootstrap MPE's 112

6.1 Boxplots of $\hat{\boldsymbol{\gamma}}$ vs $\boldsymbol{\gamma}$ 131

6.2 Values of terms in τ_2 for varied sample size. 146

CHAPTER 1: INTRODUCTION

Brain Imaging genetic studies reveal genetic basis of brain images. Methods for identifying genetic associations with voxelwise brain imaging data have evolved from parallel analysis on each voxel to incorporating spatial smoothness and correlation to make better use of the information in voxelwise brain-imaging data and increase statistical detection power. Meanwhile, there are still issues to account for in imaging genetic data analysis, such as imperfect alignment of affected regions and registration error, low signal to noise ratio in high-dimension data, high computation complexity due to high-dimension in both imaging and genetic data, and between-study heterogeneity. To address the above issues, we propose the following methods in my dissertation building on current work in the field.

First, to deal with imperfect alignment and registration error in brain imaging data, we propose a region-based functional genome-wide association detection (rfGWAD) method, which also reduces computation burden at the same time as compared to standard voxelwise methods. In particular, our method first summarizes voxelwise measurements in the brain imaging in each small region into a density curve for each subject, while the region slides across the entire brain image in a sliding-window scheme as similar to that applied in genetic analysis. The density curves are then transformed using log-quantile-density transformation so as to apply commonly used functional methods that are designed for analyzing functions in the Hilbert space. The non-parametric ball covariance test is used to detect association between the transformed regional densities and genetic markers. We compare the ball covariance test with other state-of-the-art methods on simulated datasets and demonstrate good sensitivity and specificity of our method. We also apply the rfGWAD method to the hippocampal surface data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

Second, we proposed a weighted distance covariance method to extract SNP-related signal from the image data. In particular, dimension reduction is first performed on the image data w.r.t. each SNP, where correlation with the dimension-reduced image is maximized. Then, the distance covariance with the SNP is maximized to find an optimal linear combination of the reduced dimensions of the image. A voxelwise local test, as well as a global test is proposed. For genome-wide analysis, a screening procedure is proposed to reduce computational burden. A simulation study is performed on each step of the proposed method. Then, the proposed method is applied to the genome-wide SNPs and the hippocampus surface radial distance data from the ADNI study.

Third, in order to train a prediction model with data available from more than one studies, one can either obtain the final trained learner (i.e. prediction model) by combining learners trained from different studies or train a single learner based on the merged data of all studies. The question of which strategy would achieve better prediction accuracy is discussed in a recent work by Guan et al. (2019), where learners were trained using modeles such linear regression and ridge regression. The theorems derived therein provide a useful guideline for the decision making. We extended their idea to the brain imaging setting and derived similar guidelines. While the response variable for linear regression models are one dimensional, brain imaging data is usually high-dimensional and modeled as functional responses. One commonly used model for brain imaging data is the multivariate varying coefficient model (MVCM) (Zhu et al., 2012), which takes into consideration the smoothness feature of the voxelwise brain imaging data. In this chapter, we derived the training strategy choosing guideline for learners modeled by MVCM, validated the theorems through simulations, and applied the derived guideline to neuroimaging datasets.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we review existing statistical methods on the three topics in the dissertation. In Section 2.1, we review popular genome-wide association methods (GWAMs) testing one marker at a time or a set of markers at a time, as well as functional frameworks for analyzing imaging-genetic data. In Section 2.2, we review nonparametric association measurements and methods for extracting signal in the presence of vast majority of the dimensions being noise through high-dimensional projection. In Section 2.3, we review methods related to meta-analysis.

2.1 Statistical Methods for Association Detection

We review popular GWAMs and functional framework for imaging-genetic data analysis in this section. Boosting the development of statistical methods in brain imaging genomics, GWAMs are borrowed into brain imaging genetic frameworks by duplicating the analysis on each dimension of the imaging response and accounting for intrinsic nature of brain images, such as spatial smoothness and correlation. For example, GWAS is implemented on each voxel of brain imaging phenotype in vGWAS (Stein, 2010) and FVGWAS (Huang, 2015), SKAT (Liu, 2007; Wu, 2011) is embedded in Ge's (2012) and FMEM (Lin, 2014), and GEMMA (Zhou and Stephens, 2012) and simplified REML (Lippert, 2014) are implemented in Ganjgahi's framework (2018). In our proposed rfGWAD framework, GWAMs or general association testing methods can be easily embedded into the pipeline. In section 2.1.1, we review the popular GWAMs that have been implemented in imaging genetic methods or can be embedded in imaging genetic frameworks in the current or future projects; in section 2.1.2, we review functional framework for imaging-genetic data analysis.

2.1.1 Association Testing Methods Used in Genomics

In GWAMs, the simplest case is where we test the association between one phenotype and one marker or a set of markers, assuming independence between subjects. The single-marker strategy genome-wide association study (GWAS) implements a simple linear regression model for each genetic marker throughout the genome. This is usually implemented for common variants, since rare variants tend to yield low power in GWAS. Marker-set methods are proposed to analyze rare variants for a power boost or integrate the collective contribution of a set of markers. Markers are usually grouped according to their locations into a gene set. The set of markers are then tested together in a linear regression model, which can be either a fixed effect model or a linear mixed model (LMM) (e.g. SKAT (Wu, 2011)). With a fixed effect model, burden test (Asimit, 2012; Morgenthaler, 2007; Li, 2008; Morris, 2010; Madsen, 2009) collapses a set of variants into genetic scores. One can perform the burden test with the set of markers selected based on data-adaptive weights or thresholds, termed adaptive burden tests (Han, 2010; Hoffmann, 2010; Lin, 2011; Price, 2010; Liu, 2010; Ionita-Laza, 2011). With fixed effect model, one can also regress the phenotype of interest on polygenic risk scores (PRS) (Dudbridge, 2013; Dima, 2015; Chasioti, 2019), which is often calculated as the sum of dosages of the markers weighted by their effect sizes on a phenotype (e.g. case-control status). Fixed-effects models have higher power as compared to LMM when the effects of the grouped markers are in the same direction. Using LMM models, variance-component tests (Wu, 2011; Pan, 2009; Neale, 2011) examine the variance of genetic effects and reveal nonlinear effects introduced by the interaction among SNPs. The model in sequence kernel association test (SKAT) (Wu, 2011) accommodates the least-square kernel machine (LSKM) (Liu et al., 2007) provides a general framework that can accommodate other variance-component tests:

$$y_i = \boldsymbol{z}_i^T \boldsymbol{\alpha} + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{2.1}$$

where $\boldsymbol{z}_i$ is a $q \times 1$ vector of covariates that can include a column of 1's for the intercept, $\boldsymbol{\alpha}$ is a $q \times 1$ vector of coefficients for the covariates, $\boldsymbol{x}_i$ is a $p \times 1$ vector of genotypes, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of random effects where each β_j follows an arbitrary distribution with a mean of zero and a variance of $w_j \tau$ under $H_0 : \boldsymbol{\beta} = \mathbf{0}$ (τ is a variance component and w_j is a prespecified weight for variant j), and ϵ_i is an error term with mean 0 and variance σ^2. Testing H_0 is thus equivalent to testing $H_0 : \tau = 0$. The proposed score statistic is thus

$$T_{SKAT} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T K (\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \tag{2.2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $K = XWX^T$, and $\hat{\boldsymbol{\mu}} = Z^T \hat{\boldsymbol{\alpha}}$ is the predicted mean of $\boldsymbol{y}$ under H_0. There are also methods that combine burden and variance-component tests (Lee, 2012; Derkach, 2013; Sun, 2013). Lee et. al. (2014) reviewed a list of popular methods with their pros and cons summarized .

There are a range of methods for genetic trait mapping that account for related individuals (RI) in the sample. Traditional linkage methods examines the flow of mutants from ancestors down the pedigree. Linkage analysis can be conducted using Sequential Oligogenic Linkage Analysis Routines (SOLAR) (https://hpc.nih.gov/docs/solar-8.1.1/) (Almasy, 1998), which is now a part of SOLAR-Eclipse (An Imaging Genetics Analyses Software, available at http://solar-eclipse-genetics.org), based on identity by descent (IBD) matrices. Multipoint engine for rapid likelihood inference (Merlin) (Abecasis, 2002) is widely used to perform a combined linkage and association analysis based on pedigree data, where large pedigree can be broken down using external algorithm (Scuteri, 2007). As association screening such as GWAS becomes more popular, RI methods are developed under the association screening framework by modifying the simple linear regression into LMM or generalized estimating equations (GEE). SUGEN (Genetic Association Analysis Under Complex Survey Sampling, available at http://dlin.web.unc.edu/software/sugen/) (Lin, 2014) accounts for relatedness between subjects by grouping related subjects into families and is implemented through GEE.

LMM methods account for sample relatedness based on a genetic relationship matrix (GRM) or kinship matrix, where relationship between subjects are measured by a value between 0 and 1 (in GRM) or a number between 0 and 0.5 (for kinship coefficient). The LMM model to account for RI is as follows (Kang, 2008):

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \epsilon, \tag{2.3}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of observed phenotypes, X is an $n \times q$ matrix of fixed effects including mean, SNPs, and other confounding variables. $\boldsymbol{\beta}$ is a $q \times 1$ vector representing coefficients of the fixed effects. Z is an $n \times r$ incidence matrix mapping each observed phenotype to one of r inbred strains. $\boldsymbol{u}$ is the random effect of the mixed model with $Var(\boldsymbol{u}) = \sigma_g^2 K$ where K is the $r \times r$ kinship matrix inferred from genotypes, and ϵ is an $n \times 1$ vector of residual effect such that $Var(\epsilon) = \sigma_\epsilon^2 I$. The overall phenotypic variance-covariance matrix can be represented as equation $V = \sigma_g^2 Z K^T Z^T + \sigma_\epsilon^2 I$. Algorithms are developed to accelerate LMM in GWAS, such as fastGWA (Jiang, 2019) in GCTA (https://cnsgenomics.com/software/gcta), genome-wide efficient mixed model association (GEMMA) (Zhou and Stephens, 2012), EMMA (Kang, 2008), and EMMAX (Kang, 2010). Other software that can be used to perform LMM in association studies to account for RI are EPACTs with EMMAX option (https://genome.sph.umich.edu/wiki/EPACTS), BOLT-LMM (Loh, 2015), the bioconductor package GENESIS (https://www.bioconductor.org/packages/release/bioc/html/GENESIS.html), and R packages such as cpgen (https://cran.r-project.org/web/packages/cpgen/index.html).

Association testing methods are also developed for multivariate response. Dutta et. al. (2014) extended the SKAT method (Wu, 2011) to the multivariate response scenario, named multi-SKAT. The generalized Berk-Jones (GBJ) test (Sun, 2019) can also work with multi-dimensional response by combining a set of p-values and their correlation matrix and to obtain an overall p-value for the global null hypothesis. The ball covariance (BCov) test (Pan, 2019) is a nonparametric rank-based test of associations, which allows multi-dimensional variables on both sides of the association relationship. The ball covariance equals to 0 if and only if the two variables are unrelated. Ball covariance is described in more details in section 2.2 in the literature review. Methods for multivariate response also include RdCov (Deb and Sen, 2019), adaptive Mantel test (Pluta, 2018), and sKPCR (Gong, 2018). Those methods are compared in the simulation study of project 1 in section 3.3, thus we give more details on them as follows.

The adaptive Mantel (AdaMant) test (Pluta, 2018) considers a set of linear model score tests in a unified Mantel test framework, where each type of model is represented by a distinct similarity metric. The Mantel test statistic is in the following form:

$$T(X,Y) = tr(HK) = \sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}K_{ij}, \tag{2.4}$$

where X is $n \times p$, Y is $n \times M$, $H = YY^T$, and $K = XWX^T$ for a weight matrix W. Pluta et. al. derived that the score test statistics for fixed effects linear regression model T_F, variance components model T_V, and ridge regression T_λ can all be written in the form of 2.4 with different W. In particular,

$$\begin{aligned} &T_F = tr(HK_F),\ K_F = XW_FX^T,\ W_F = I \qquad (2.5)\\ &T_V = tr(HK_V),\ K_V = XW_VX^T,\ W_V = (X^TX)^{-1}\\ &T_\lambda = tr(HK_\lambda),\ K_\lambda = XW_\lambda X^T,\ W_\lambda = (X^TX + \lambda I)^{-1} \end{aligned}$$

For high-dimensional Y, tests can also be calculated to with weight matrix W similar to K. After calculating J individual test statistics based on different similarity metrics, the AdaMant test statistic is defined as the minimum p-value of the J tests

$$T_{AdaMant} := \min_{m \in 1,\dots,J} p_m. \tag{2.6}$$

The p-value of the AdaMant test is then calculated through permutation. AdaMant is implemented in the AdaMant R package available at https://github.com/dspluta/adamant. Currently, the AdaMant R package supports multi-dimensional X and one-dimensional Y.

Data with multi-dimensional response is also analyzed using principal-component-based methods, such as probablistic PCA (Tipping and Bishop, 1999; Bishop, 2006) and kernel principal component analysis (kPCA) (Schölkopf, 1997). Building on those methods, Gong et. al. (2019) proposed a structured kernel principal component regression (sKPCR) that allows user-defined covariance structure between features and samples and accommodates nonlinear dimension reduction through kPCA. For the purpose of notation consistency, let Y denote the $n \times M$ high-dimensional imaging data matrix, and $\boldsymbol{x}$ is a $n \times 1$ numeric variable that we are interested to test the association against Y. The first step of sKPCR is a structured principal component analysis (sPCA), which extracts the first L principal components of Y. The sPCA extends probablistic PCA by allowing covariance structures between rows and columns of Y:

$$Y = W^T U + E, \quad \text{where } E \sim MN_{M,n}(0, R, Q), \tag{2.7}$$

where R is the covariance matrix between columns of E and Q is the covariance matrix between rows of E. Adopting the maximum likelihood estimation (MLE) approach in probablistic PCA, sPCA is solved by maximizing the following log-likelihood:

$$\begin{aligned}\log P(Y, U|W) &= -\frac{1}{2} tr\left[Q^{-1}(Y - W^T U)^T R^{-1}(Y - W^T U)\right] - \frac{1}{2} tr(U^T U) + Const \qquad (2.8)\\ &= -\frac{1}{2} tr\left[(\tilde{R}Y\tilde{Q} - \tilde{R}W^T U\tilde{Q})^T(\tilde{R}Y\tilde{Q} - \tilde{R}W^T U\tilde{Q})\right] - \frac{1}{2} tr(U^T U) + Const\\ &= -\frac{1}{2} tr\left[(\tilde{Y} - \tilde{W}^T\tilde{U})^T(\tilde{Y} - \tilde{W}^T\tilde{U})\right] - \frac{1}{2} tr(U^T U) + Const,\end{aligned}$$

where $Q^{-1} = \tilde{Q}\tilde{Q}^T$, $R^{-1} = \tilde{R}\tilde{R}^T$, $\tilde{Y} = \tilde{R}Y\tilde{Q}$, $\tilde{W} = W\tilde{R}$, and $\tilde{U} = U\tilde{Q}$. Hence, the solution of sPCA is equivalent to the standard PCA or probabilistic PCA problems using the 'weighted' data matrix $\tilde{Y}$. After extracting the first L PCs through sPCA, the association between first k PCs and $\boldsymbol{x}$ is measured by $S_k = \sum_{i=1}^{k} r_i^2$, where r_i is the correlation between the ith PC and Y and $k = 1, \dots, L$. Then, the p-value of S_k, denoted as p_k, is calculated through permutation. Lastly, the test statistic for sKPCR is defined as

$$T_{sKPCR} = \min_{m \in 1, \dots, L} p_m, \tag{2.9}$$

and the p-value for T_{sKPCR} is obtained through permutation.

Deb and Sen (2019) proposed a framework that includes testing for mutual independence between random vectors by constructing rank-based measurements such as rank-based distance covariance (RdCov), which builds on distance covariance (dCov) (Székely, 2007). These measurements are based on multivariate ranks and thus distribution-free, i.e. the null distributions of the test statistics are free of the underlying data generating distributions. The use of ranks to perform distribution-free statistical inference is ubiquitous in one-dimensional problems in nonparametric statistics, such as Spearman's rank-correlation, Wilcoxon signed-rank test, Mann-Whitney rank-sum test, and Kruskal-Wallis test. In particular, one-dimensional ranks can be interpreted as

$$\hat{\sigma} := \underset{\sigma = (\sigma(1), \dots, \sigma(n)) \in S_n}{\arg\max} \sum_{i=1}^{n} \left| y_i - \frac{\sigma(i)}{n} \right|^2, \tag{2.10}$$

where S_n is the set of all permutations of $\{1, 2, ..., n\}$ and $y_i \in \mathbb{R}$. Extending it to multivariate setting where $\boldsymbol{y}_i \in \mathbb{R}^M$, replace the discrete uniform numbers $\{i/n : 1 \le i \le n\}$ by the set of multivariate rank vectors $\{\boldsymbol{c}_1, \dots, \boldsymbol{c}_n\} \subset [0,1]^M$ – a sequence of "uniform-like" points in $[0,1]^M$ (a recommended choice is the Halton sequences). The multivariate rank is then defined as

$$\hat{\sigma} := \underset{\sigma = (\sigma(1), \dots, \sigma(n)) \in S_n}{\arg\max} \sum_{i=1}^{n} ||\boldsymbol{y}_i - \boldsymbol{c}_{\sigma(i)}||^2, \tag{2.11}$$

where $|| \cdot ||$ denotes the Euclidean norm in $\mathbb{R}^M$. This optimization problem can be viewed as an assignment problem, which can be solved by the Hungarian algorithm. Then, to calculate a rank-based statistic, such as RdCov, we calculate dCov based on the multivariate ranks inferred from data. Deb and Sen derived the asymptotic null distributions of the proposed test statistics and showed that their proposed tests are consistent against all fixed alternatives, i.e. the probability of rejecting the null, calculated under the alternative, converges to 1 as the sample size increases. Moreover, the proposed tests are tuning-free, computationally feasible and are well-defined under minimal assumptions on the underlying distributions (e.g. no moment assumptions is needed). The proposed framework for multivariate distribution-free nonparametric testing is general and is also applicable for comparing two distributions and beyond. The algorithm can be easily implemented using a combination of existing tools (example R scripts available at https://github.com/NabarunD/MultiDistFree).

Other genetic methods that can be carried into imaging-genetic setting include transcriptome wide association study (TWAS) (Gusev, 2016) and linkage disequilibrium score regression (LDSC) (Bulik-Sullivan, 2015). A comprehensive review of recent statistical methods in brain imaging genomics are given by Shen and Thompson (2019). The review focuses on association testing methods in brain imaging genomics, including voxelwise GWAS, multi-marker methods, multivariate/functional methods, and sparse canonical correlation analysis (SCCA) methods.

2.1.2 Functional Frameworks

Analysis frameworks for detecting the association between voxelwise brain images and genetic markers throughout the whole genome has been evolving in recent years. Voxelwise genome-wide association study (vGWAS) was first carried out by Stein et. al. (2010) to perform single-voxel-single-marker analysis across all voxels and genome-wide genetic markers. This idea has been carried out in fast voxelwise GWAS (FVGWAS) (Huang, 2015), which improved computation efficiency in a pipeline combining a heteroscedastic linear model, global sure independent screening (GSIS) (Fan and Lv, 2008), and a detection procedure based on wild bootstrap. Functional frameworks are then developed to account for intrinsic spacial structure and correlation in brain imaging data. Suppose that $y_i(s)$ is the functional imaging response for subject i ($i = 1, \ldots, n$) and $s \in \mathcal{S}$ denote a point on the image domain, e.g. a voxel in a 3D image (or a pixel in a 2D image). s is continuous in nature but discrete in the observed data. Suppose M is the total number of pixels (or voxels in 2D image) in the observed image, then we have the observed image domain $\mathcal{S} = \{s_1, \ldots, s_M\}$. Therefore, the observed functional response $y_i(s)$ can also be viewed as a multivariate response y_{ij} ($j = 1, \ldots, M$), with $y_{ij} = y_i(s_j)$. The single-voxel-single-marker strategy analyzes y_{ij} one voxel at a time, i.e. for each j parallelly, while the functional methods treats $y_i(\cdot)$ as a single functional response on $\mathcal{S}$.

Fan and Gijbels (1996) proposed the local polynomial model (LPM) to analyze functional response that accounts for the smooth nature in the functional coefficient $\beta(s)$ such as in the following simplest functional linear regression model:

$$y_i(s) = x_i \beta(s) + \epsilon(s), \tag{2.12}$$

where $\hat{\beta}(s)$ is estimated through local polynomial kernel (LPK) smoothing by integrating nearby information through a kernel on distances in the functional domain. This gives a smooth $\hat{\beta}(s)$ on $\mathcal{S}$, which coincides with the smoothness assumption of functional effect in brain images. Although we assume smoothness in the true imaging response $\tilde{y}_i(\cdot)$, the observed values $y_i(\cdot)$ are usually not smooth due to noise:

$$y_i(s) = \tilde{y}_i(s) + \epsilon(s). \tag{2.13}$$

A usual conduct in imaging data analysis is to first obtain a smoothed representation of $y_i(\cdot)$, denoted as $f_i(\cdot)$, and then perform further statistical inference on $f_i(\cdot)$. Comparing to directly analyzing $y_i(\cdot)$, Zhang and Chen (2007) proved that, under some mild conditions, smoothing the imaging first with LPK has an asymptotically ignorable effect on the statistical inference that follows.

Zhu et. al. (2012) proposed the multivariate varying coefficient model (MVCM) building on LPM to incorporate multiple modes of measurements on the imaging domain:

$$y_{ij}(s) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j(s) + \eta_{ij}(s) + \epsilon_{ij}(s) \quad \text{for } j = 1, \ldots, J, \tag{2.14}$$

where $\boldsymbol{x}_i$ is a $p \times 1$ covariate vector, $\boldsymbol{\beta}_j(s)$ is the $p \times 1$ functional coefficients characterizing covariates effect, $j \in \{1, \ldots, J\}$ indicates the jth imaging measurement. $\epsilon_{ij}(s)$ represents the measurement error, and $\eta_{ij}(s)$ characterizes individual curve deviation from $\boldsymbol{x}_i^T \boldsymbol{\beta}_j(s)$ and within-curve dependencies. Let $\boldsymbol{\epsilon}_i(s) = (\epsilon_{i1}, \ldots, \epsilon_{iJ})^T$ and $\boldsymbol{\eta}_i(s) = (\eta_{i1}, \ldots, \eta_{iJ})^T$; $\boldsymbol{\epsilon}_i(s)$ and $\boldsymbol{\eta}_i(s)$ are independent and identical copies of $SP(\mathbf{0}, \Sigma_\epsilon)$ and $SP(\mathbf{0}, \Sigma_\eta)$, respectively, where $SP(\mu, \Sigma)$ denotes a stochastic process vector with mean function $\mu(s)$ and covariance function $\Sigma(s,t)$. $\Sigma_\epsilon(s,t) = S_\epsilon(s) I(s = t)$, where $S_\epsilon(s) = (r_{\epsilon,jj'}(s))$ is a $J \times J$ matrix of functions of s and $I(\cdot)$ is an indicator function.

Huang (2017) developed a functional GWAS (FGWAS) pipeline that employs the MVCM (Zhu, 2012) and techniques in FVGWAS (Huang, 2015). In particular, the MVCM model accounts for spatial smoothness, spatial correlation, and low-dimensional representation of the voxelwise brain imaging phenotype. Spatial smoothness is assumed on both $\boldsymbol{\beta}_j(s)$ and $\eta_{ij}(s)$ in (2.14), estimated using local linear regression (Fan and Gijbels, 1996). In local linear regression, local information is synthesized through a weighted sum, where "local" and "weights" are defined by bandwidth and kernel function of spatial distance, respectively. Spatial correlation is accounted for through $\eta_{ij}(s)$ in (2.14). Low-dimensional representation is achieved by functional principal analysis on $\eta_{ij}(s)$ through spectral decomposition of $\Sigma_\eta(s,t)$ and approximating $\eta_{ij}(s)$ by a low number of principal components. The GSIS procedure based on global test statistic is used to select promising SNPs from genotype data and thus reduce computation complexity by focusing only on the set of promising SNPs. In particular, GSIS selects the top-ranking SNPs based on global test statistic or the corresponding p-value. Lastly, cluster-size inference (Ge, 2012) with wild-bootstrap is used to identify region associated with a SNP, which reduces computation complexity and boosts power at the same time. A parallel algorithm is developed to reduce computation time. FGWAS is suitable for analyzing 1D curves, 2D surfaces, or 3D images.

Marker-set analysis is implemented through least-squares kernel machines (Liu, 2007; Wu, 2011) for high-dimensional imaging response. Ge et. al. (2012) developed a pipeline to identify imaging regions associated with genetic markers by applying the marker-set model on each voxel. Cluster size inference is then performed based on random field theory, making use of the spatial smooth nature of brain images. On the other hand, Lin et. al. (2014) extended the least-squares kernel machine (Liu, 2007; Wu, 2011) into a functional mixed effects model (FMEM) to test the association between the imaging response and a set of genetic markers. The FMEM is as follows: for $i = 1, \dots, n$,

$$y_i(s) = \boldsymbol{z}_i^T \boldsymbol{\alpha}(s) + \boldsymbol{x}_i^T \boldsymbol{\beta}(s) + \epsilon(s), \tag{2.15}$$

where $\boldsymbol{z}_i$ is a $q \times 1$ vector for covariates as fixed effects, $\boldsymbol{\alpha}(s)$ is the $q \times 1$ vector of functional coefficients, $\boldsymbol{x}_i$ is a $p \times 1$ vector the random genetic effect, $\boldsymbol{\beta}(s)$ is the $p \times 1$ vector of coefficients that follows $N(0, \sigma_\beta^2(s)\boldsymbol{I}_p)$, and $\epsilon_i(s) \sim N(0, \sigma_\epsilon^2(s))$ is independent across i and independent of $\beta(s)$ for all s. Lin et. al. (2014) also implemented a jumping surface model on the variance components of the genetic random effects and fixed effects as piecewise smooth functions of the voxels.

In order to account for RI and population structure for the high-dimension imaging response setting, a simple conduct is to apply LMM for each genetic marker and on each voxel in the image. However, the computation complexity of this conduct becomes a challenge. In order to reduce computation burden, Ganjgahi et. al. (2018) applied transformation of the data and model, such as in GEMMA (Zhou and Stephens, 2012) and simplified REML (Lippert, 2014), and proposed a non-iterative variance component estimator. This largely reduces the computation and makes permutation tests feasible, which allows inference on powerful spatial tests like the cluster size statistic.

Another functional framework, named low-rank linear regression model (L2RM), is proposed by Kong et. al. (2019) to detect association between high-dimensional response (i.e. imaging response) and high-dimensional covariates (e.g. genetic markers) when coefficient matrices have low-rank structures. L2RM consists of two steps: screening and estimation. The screening procedure is developed to select promising covariates when the number of covariates is extremely large, such as the number of SNPs in the genome. The screening is based on the spectral norm of each coefficient matrix, which is fast and efficient. The estimation procedure is based on trace norm regularization, which explicitly imposes a low rank structure on the coefficient matrices. The sure independence screening property of the screening procedure is investigated when the dimension of the response and the dimension of the covariates diverge at the exponential order of the sample size. Asymptotic properties of the estimation procedure are investigated, such as estimation consistency, rank consistency, and nonasymptotic error bound under. For the two-step screening and estimation procedure, a theoretical guarantee for the overall solution is established.

Method is also proposed (Zhang, 2018) to extract signal from noisy high-dimensional brain imaging data through functional linear regression model. In this method, Zhang optimizes the weight projection of the imaging response to achieve maximum power of the association test. Details of this method is discussed in section 2.2 in the literature review.

2.2 Covariance Measurements and High-Dimensional Projection

2.2.1 Covariance and Correlation Measurements

Correlation coefficients have been widely developed to measure statistical dependence between objects in Hilbert spaces. Pearson correlation (Pearson, 1895) is commonly used to detect monotonic dependence between two random variables. Distance correlation (Székely, 2007) and projection correlation (Zhu, 2017) have been proposed to detect nonlinear and nonmonotonic dependence between two random vectors of arbitrary dimension, with the independence-zero equivalence property for random vectors in metric spaces of strong negative

Categories		**Name**	**Method**
Single-marker		GWAS	Simple linear regression
Marker-set	Fixed model	Burden test	Collapse a set of variants into genetic scores
		PRS	Sum of dosages of markers weighted by effect sizes
	LMM	SKAT	$y_i = \boldsymbol{z}_i^T \boldsymbol{\alpha} + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$ $T_{SKAT} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T, K(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$
RI	Linkage	SOLAR	Sequential Oligogenic Linkage Analysis Routines
		Merlin	Multipoint engine for rapid likelihood inference
	Association	SUGEN	GEE grouping related subjects into families
		LMM	$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \epsilon$
Multivariate response		Multi-SKAT	Multivariate extension of SKAT
		GBJ	Combination a set of p-values and their correlation
		BCov	nonparametric rank-based test
		RdCov	Rank-based dCov
		AdaMant	$T(X,Y) = tr(HK) = \sum_{i=1}^n \sum_{j=1}^n H_{ij}K_{ij}$
		sKPCR	$Y = W^T U + E, \quad \text{where } E \sim MN_{M,n}(0, R, Q)$

Table 2.1: Summary of methods reviewed in Section 2.1.1

Categories		**Name**	**Method**
Voxelwise	Single-marker	vGWAS	Voxelwise GWAS
		FVGWAS	Fast Voxelwise GWAS
	Marker-set	Ge et al 2012	Voxelwise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures
Functional	Single-marker	LPM	$y_i(s) = x_i\beta(s) + \epsilon(s)$
		Smoothing first	$y_i(s) = \tilde{y}_i(s) + \epsilon(s)$
		MVCM	$y_{ij}(s) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j(s) + \eta_{ij}(s) + \epsilon_{ij}(s)$
		FGWAS	Functional GWAS, with MVCM, GSIS, and cluster-size inference through wild-bootstrap
		Zhang 2018	Optimal weight projection of the imaging response to achieve maximum power of the association test
	Marker-set	FMEM	$y_i(s) = \boldsymbol{z}_i^T, \alpha(s) + \boldsymbol{x}_i^T \boldsymbol{\beta}(s) + \epsilon(s)$
		L2RM	Spectral-norm-based screening and trace-norm-based regularization
	RI	Ganjgahi et al 2018	Transformation (e.g. GEMMA, simplified REML) and non-iterative variance component estimator

Table 2.2: Summary of methods reviewed in Section 2.1.2

type. Since complex objects (e.g., different brain subcortical structures) often reside in Banach spaces, Pan et. al. (2019) developed ball covariance as a generic nonparametric and model-free measure of dependence for Banach spaces as well as metric spaces. It is nonnegative and independence-zero equivalent. The Heller-Heller-Gorfine (HHG) measure (Heller, Heller, and Gorfine 2013) is a special case of ball covariance by choosing a proper weight.

Pearson correlation (Pearson 1895), or the product moment correlation coefficient, is a measure of the linear correlation between two variables X and Y. It has a value between -1 and 1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. The population Pearson's correlation coefficient is

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \tag{2.16}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively, and μ_X and μ_Y are the corresponding means. The sample Pearson correlation correlation is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{2.17}$$

where n is the sample size and $\bar{x}$ and $\bar{y}$ are the sample means.

The Spearman's ρ (Spearman 1904) is a nonparametric measure of rank correlation, which measures monotonic relationship and ranges from -1 to 1. It equals to the Pearson's correlation coefficient applied to the ranks of the variables $r_s = \rho_{r_X, r_Y}$,where r_X and r_Y are the ranks of X and Y respectively. The Spearman's rank correlation coefficient is appropriate for both continuous and discrete ordinal variables.

The Kendall's τ coefficient (Kendall 1938) is also rank-based, which is defined as follows:

$$\tau = \frac{\sum_{i,j=1}^{n} I\{(x_i-x_j)(y_i-y_j)>0\} - \sum_{i,j=1}^{n} I\{(x_i-x_j)(y_i-y_j)<0\}}{\binom{n}{2}}. \tag{2.18}$$

two special types of correlation, two Cramér-von Mises criterion dependence measures based on the empirical distribution function

Cramér-von Mises criterion for comparing two distribution functions $F_1(x)$ and $F_2(s)$:

$$\omega^2 = \int_{-\infty}^{\infty} [F_1(x) - F_2(x)]^2 dF_1(x) \tag{2.19}$$

Hoeffding's dependence measure (**?**) and the one propsed by Blum, Kiefer, and Rosenblatt (1961) can be used to measure nonlinear dependence without moment conditions. In particular, Hoeffding's dependence measure (**?**) is defined as follows:

$$H = \int [F_{xy}(x, y) - F_x(x)F_y(y)]^2 dF(x, y); \tag{2.20}$$

and the dependence measurement introduced by Blum, Kiefer, and Rosenblatt (1961) is as follows: for $\boldsymbol{x} = (x_i, \dots, x_m) \in R^m$,

$$B_n = \int [F_n(\boldsymbol{x}) - \prod_{j=1}^{m} F_{nj}(x_j)]^2 dF_n(\boldsymbol{x}). \tag{2.21}$$

To measure the dependence of X and Y of arbitraty dimensions p and q, respectively, the empirical distance covariance (Szekely et al., 2007) is proposed as follows: let A and B be the $n \times n$ double-centered Euclidean distance matrices of X and Y, respectively. In particular, define pairwise distance between subject i and subject j based on X and Y, respectively, as follows:

$$a_{ij} = \|X_i - X_j\|_p, \quad b_{ij} = \|Y_i - Y_j\|_q, \tag{2.22}$$

for $i, j = 1, \dots, n$, where $\|\cdot\|_p$ and $\|\cdot\|_q$ stand for Euclidean norms. Let $\bar{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}$, $\bar{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$, $\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}$, and $\bar{b}_{i\cdot}, \bar{b}_{\cdot j}$ and $\bar{b}_{\cdot cdot}$ defined analogously, then

$$A_{ij} = \bar{a}_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, \quad B_{ij} = \bar{b}_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot\cdot}. \tag{2.23}$$

The empirical distance covariance is thus

$$dCov_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}, \tag{2.24}$$

which is based on the agreement of X and Y in terms of pairwise distance between subjects (i,j), while ball covariance is based on the empirical distribution differences of joint distribution and product of marginal distributions at all points for all resolutions, which will be defined later. (The intuition of this empirical measurement can be seen from Theorem 1 of (Szekely et al., 2007) for how it is associated with its theoretical definition based on the squared difference between characteristic functions $||\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)||^2$.)

The distance covariance (and correlation) and a projection correlation to be propsed by Zhu et. al. (2017) are powerful in detecting nonlinear and nonmonotonic dependence between two random vectors of arbitrary dimension. The projection correlation (Zhu et al. 2017) "first projects the multivariate random vectors into a series of univariate random variables, then detects nonlinear dependence by calculating the Pearson correlation between the dichotomized univariate random variables." It is defined as follows:

$$PC(X,Y) = \int\int\int [F_{U,V}(u,v) - F_U(u)F_V(v)]^2 dF_{U,V}(u,v)d\alpha d\beta, \tag{2.25}$$

where $U = \alpha^T X$, $V = \beta^T Y$, and F's are the corresponding distribution functions. We can see that the projection covariance is also based on the previously mentioned Cramér-von Mises criterion. Mathematical simplification of the above definition of the the projection covariance is applied.

The Heller-Heller-Gorfine (HHG) measure (Heller, Heller, and Gorfine 2013) of dependence is based on the ranks of a_{ij}'s and b_{ij}'s as defined in the distance covariance. The motivation of HHG is very similar to that of ball covariance: "if X and Y are dependent and have a continuous joint density, then there exists a point (x0, y0) in the sample space of (X, Y) and radii Rx0 and Ry0 around x0 and y0, respectively, such that the joint distribution of X and

Y differs from the product of the marginal distributions in the Cartesian product of balls around (x0, y0). " For $S(i,j)$ being the Pearson's test for the following 2×2 contingency table

	$d(y_i,\cdot) \leq d(y_i,y_j)$	$d(y_i,\cdot) > d(y_i,y_j)$	
$d(x_i,\cdot) \leq d(x_i,x_j)$	$A_{11}(i,j)$	$A_{12}(i,j)$	$A_{1\cdot}(i,j)$
$d(x_i,\cdot) > d(x_i,x_j)$	$A_{21}(i,j)$	$A_{22}(i,j)$	$A_{2\cdot}(i,j)$
	$A_{\cdot 1}(i,j)$	$A_{\cdot 2}(i,j)$	$N-2$

the HHG measure is defined as follows:

$$T = \sum_{i \neq j} S(i,j). \tag{2.26}$$

The HHG measure can be derived from Ball Covariance by choosing a proper weight (Zhu et al. 2019).

Hilbert-Schmidt covariance (aka Hilbert-Schmidt independence criterion [HSIC] [Gretton et al. (2008)]) is defined as the squared Hilbert-Schmidt norm (the sum of squared singular values) of the cross-covariance operator. The test based on HSIC costs $O(n^2)$, where n is the sample size.

Ball covariance is defined by using the projection-type method and Hoeffding's dependence measure (Hoeffding, 1948) on the corresponding one-dimensional space of radial distance: for random variables X and Y,

$$\begin{aligned} BCov^2_w(X,Y) := \int (\theta - \mu \otimes \nu)^2 [\bar{B}_\rho(x_1,x_2) \times \bar{B}_\zeta(y_1,y_2)] \\ w_1(x_1,x_2) w_2(y_1,y_2) \theta(dx_1,dy_1) \theta(dx_2,dy_2), \end{aligned} \tag{2.27}$$

where $(\mathscr{X},\rho)$ and $(\mathscr{Y},\zeta)$ are two Banach spaces, where the norms ρ and ζ also represent their induced distances. θ is a Borel probability measure on $\mathscr{X} \times \mathscr{Y}$, μ and ν are two Borel probability measures on $\mathscr{X}$ and $\mathscr{Y}$, respectively, and (X,Y) is a B-valued random variable defined on a probability space such that $(X,Y) \sim \theta$, $X \sim \mu$, and $Y \sim \nu$. $\bar{B}(x_1,x_2)$ is the closed ball with the center x_1 and the radius $\rho(x_1, x_2)$ in $\mathscr{X}$ and $\bar{B}(y_1, y_2)$ is the closed ball with the center y_1 and the radius $\zeta(y_1, y_2)$ in $\mathscr{Y}$. $\{W_i = (X_i, Y_i), i = 1, 2, \ldots\}$ is an infinite sequence of i.i.d. samples of (X, Y), and $\omega = (\omega_1, \omega_2)$ is the positive weight function on the support set of θ. $[\theta - \mu \otimes \nu]^2(A \times B) := [\theta(A \times B) - \mu(A)\nu(B)]^2$ for $A \in \mathscr{X}$ and $B \in \mathscr{Y}$. Positive weights w_1 amd w_2 allow flexibility and connection with HHG.

The empirical ball covariance is defined as

$$BCov^2_{w,n}(X, Y) := \frac{1}{n^2} \sum_{i,j=1}^{n} (\Delta^{XY}_{ij,n} - \Delta^{X}_{ij,n}\Delta^{Y}_{ij,n})^2 \hat{w}_1(X_i, X_j)\hat{w}_2(Y_i, Y_j), \tag{2.28}$$

where

$$\Delta^{XY}_{ij,n} = \frac{1}{n} \sum_{k=1}^{n} \delta^X_{ij,k}\delta^Y_{ij,k}, \quad \Delta^{X}_{ij,n} = \frac{1}{n} \sum_{k=1}^{n} \delta^X_{ij,k}, \quad \Delta^{Y}_{ij,n} = \frac{1}{n} \sum_{k=1}^{n} \delta^Y_{ij,k}, \tag{2.29}$$

where $\delta^X_{ij,k} := I\{X_k \in \bar{B}_\rho(X_i, X_j)\}$ and $\delta^Y_{ij,k} := I\{Y_k \in \bar{B}_\zeta(Y_i, Y_j)\}$. $B_\rho(X_i, X_j)$ denotes the ball with center X_i and radius $d(X_i, X_j)$. $B_\zeta(Y_i, Y_j)$ denotes the ball with center Y_i and radius $d(Y_i, Y_j)$. Besides measuring dependence, empirical ball covariance is also used as a test statistic of independence. Its asymptotic distributions under both null and alternative hypotheses are derived (Pan, 2019). Several choices of weights w_k $(k = 1, 2)$ are proposed by Zhu et. al. (2019). For example, for $k = 1$, the probability weight $\hat{w}_1(X_i, X_j) = [\Delta^X_{ij,n}]^{-1}$ and the Chi-square weight $\hat{w}_1(X_i, X_j) = [\Delta^X_{ij,n}(1 - \Delta^X_{ij,n})]^{-1}$; the corresponding ball covariance are denoted as $BCov^2_{\Delta,n}$ and $BCov^2_{\chi^2,n}$, respectively. $BCov^2_{\Delta,n}$ focuses on smaller balls, while $BCov^2_{\chi^2,n}$ standardizes $(\Delta^{XY}_{ij,n} - \Delta^X_{ij,n}\Delta^Y_{ij,n})^2$ by the variance of $\delta^X_{ij,k}$. Furthermore, $BCov^2_{\chi^2,n}$ is asymptotically equivalent to HHG (Pan et al. 2018). Denote $BCov^2_{w,n}(X, Y)$ when $w_1 = w_2 = 1$ as $BCov^2_n(X, Y)$. The performance of those three types of weighted ball covariance differ under different scenarios as discussed in (Zhu et al. 2019).

2.2.2 Weight Projection for High-Dimensional Imaging Response

Brain imaging data is high-dimensional: there are 15,000 vertices on the left or right hippocampus surface from the ADNI study or $208 \times 256 \times 256 \approx 13$ million voxels in the UK Biobank T1 brain image. It is thus hard to detect the association between brain image and genetic markers, if the association signal exists only in a small part of the image while the vast majority of the image is noise. Therefore, it is of interest to assign higher weights to voxels with signals during the association detection procedure.

Zhang (2018) carried out this idea in (2.14) for $J = 1$. For testing

$$H_0 : \boldsymbol{C}\boldsymbol{\beta}(s) = b_0(s) \ \forall s \in \mathcal{S} \ \ v.s. \ \ H_1 : \boldsymbol{C}\boldsymbol{\beta}(s) \neq b_0(s) \ \exists s \in \mathcal{S}, \tag{2.30}$$

Zhang emphasizes signal over noise by maximizing the power of the global test. In particular, a weight function $\omega(s)$ is introduced to project the image to a scalar: $y_{\omega,i} = \int_{\mathcal{S}} y(s)\omega(s)ds$. The MVCM in (2.14) then becomes a scalar model

$$y_{\omega,i} = \boldsymbol{x}_i^T \boldsymbol{\beta}_\omega + \eta_{\omega,i}, \tag{2.31}$$

where $\boldsymbol{\beta}_\omega = \int_{\mathcal{S}} \boldsymbol{\beta}(s)\omega(s)ds$ and $\eta_\omega = \int_{\mathcal{S}} \eta(s)\omega(s)ds$. $\epsilon_{\omega,i} = \int_{\mathcal{S}} \epsilon_i(s)\omega(s)ds$ vanishes because it converges to 0 in probability under local kernel smoothing. After projection, a standard wald-type statistic is thus

$$T_n(\omega) = \frac{\hat{\boldsymbol{\beta}}_\omega^T \boldsymbol{C}^T [\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1} \boldsymbol{C}\hat{\boldsymbol{\beta}}_\omega}{\iint \hat{\Sigma}_\eta(s,s')\omega(s)\omega(s')dsds'}. \tag{2.32}$$

The test statistic $T_n(\omega)$ is also the signal-to-noise ratio that dominates the asymptotic power and thus is set as the objective function to optimize ω. However, maximizing $T_n(\omega)$ is an ill-conditioned problem, because the eigenvalues of $\hat{\Sigma}_\eta(s,s')$ usually decrease to zero very fast, which yields a ∞ value of $T_n(\hat{\omega})$. To address this issue, a ridge penalty term is added to the objective function

$$L(\omega) = \frac{\hat{\boldsymbol{\beta}}_\omega^T \boldsymbol{C}^T[\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}_\omega}{\iint \hat{\Sigma}_\eta(s,s')\omega(s)\omega(s')dsds' + \lambda\|\omega(s)\|_2^2}, \tag{2.33}$$

where $\|\omega(s)\|_2^2 = \int_S \omega^2(s)ds$. A closed-form solution of $\hat{\omega}$ is derived and the global test statistic is based on the optimal projection $\hat{\omega}$. The p-value of the test statistic is obtained based on wild-bootstrap.

Hu et. al. (2019) proposed distance canonical correlation analysis (DCCA), which maximizes the distance correlation (Székely, 2007) between two high-dimensional vectors $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times M}$ by solving the optimization problem

$$(\hat{v}_1, \hat{v}_2) = \underset{v_1,v_2}{\arg\max}\left[\frac{v_1' k(X,Y) v_2}{\sqrt{v_1' k(X,X) v_1}\sqrt{v_2' k(Y,Y) v_2}}\right]^{\frac{1}{2}}, \tag{2.34}$$

where $k(X,Y)$ is the $(p \times M)$ distance kernel matrix with the (i,j)th element

$$k(X,Y)_{i,j} = k(x_i, y_j) := \sum_{k,l=1}^{n} |x_{i,k} - x_{i,l}||y_{j,k} - y_{j,l}|, \tag{2.35}$$

where $x_i \in \mathbb{R}^{n\times 1}$ is a single feature from data X, $y_j \in \mathbb{R}^{n\times 1}$ is a single feature from data Y, $x_{i,k}$ is the kth element of x_i, and $y_{j,l}$ is the lth element of y_j. $\hat{v}_1$ and $\hat{v}_2$ can be solved by applying the solution of canonical correlation analysis (CCA):

$$\begin{aligned}
\hat{v}_1 &= k(X,X)^{-\frac{1}{2}} k(X,Y) k(Y,Y)^{-1} k(Y,X) k(X,X)^{-\frac{1}{2}} \qquad (2.36)\\
\hat{v}_2 &= k(Y,Y)^{-\frac{1}{2}} k(Y,X) k(X,X)^{-1} k(X,Y) k(Y,Y)^{-\frac{1}{2}}.
\end{aligned}$$

The maximized association measurement through weight projection between image and genetic markers can be viewed as the amount of available information (Wang, 2020) in the image, with regard to the genetic markers of interest. Ball covariance is a generic measurement of associations with good performance. We therefore aim to develop a ball

	Categories	**Name**	**Method**
2.2.1	Hilbert space	dCov	$dCov_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^n A_{ij}B_{ij}$ where A and B are $n \times n$ double-centered Euclidean distance matrices of X and Y, respectively
		HHG	A special case of ball covariance
	Banach space	BCov	$BCov_{w,n}^2(X,Y) =$ $\frac{1}{n^2}\sum_{i,j=1}^n(\Delta_{ij,n}^{XY} - \Delta_{ij,n}^{X}\Delta_{ij,n}^{Y})^2\hat{w}_1(X_i,X_j)\hat{w}_2(Y_i,Y_j)$
2.2.2	MVCM	Zhang 2018	$y_{\omega,i} = \boldsymbol{x}_i^T\boldsymbol{\beta}_\omega + \eta_{\omega,i}$ $L(\omega) = \frac{\hat{\boldsymbol{\beta}}_\omega^T\boldsymbol{C}^T[\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T]^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}_\omega}{\iint\hat{\Sigma}_\eta(s,s')\omega(s)\omega(s')dsds'+\lambda\lVert\omega(s)\rVert_2^2}$
	dCov	DCCA	$(\hat{v}_1,\hat{v}_2) = \underset{v_1,v_2}{\arg\max}\left[\frac{v_1'k(X,Y)v_2}{\sqrt{v_1'k(X,X)v_1}\sqrt{v_2'k(Y,Y)v_2}}\right]^{\frac{1}{2}}$

Table 2.3: Summary of methods reviewed in Section 2.2.

canonical correlation analysis based on ball correlation (Pan, 2019) with similar strategies from DCCA.

2.3 Statistical Methods for Between-Study Heterogeneity in NeuroImaging Genetic Studies

There is an increasing need to combine the analysis of multiple brain imaging genetic datasets as the number of such studies increase. Brain imaging genetic cohorts with large sample sizes are available in UK Biobank (Sudlow, 2015), ADNI (Weiner, 2017), Philadelphia Neurodevelopmental Cohort (PNC) (Satterthwaite, 2016), the Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository (Jernigan, 2017), the Human Connectome Project (HCP) (Glasser, 2016), the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), and the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium (Thompson, 2014; Thompson, 2019). The ENIGMA consortium took the lead in meta-analysis of neuroimaging genetics cohorts. Studies (Hibar, 2015; Satizabal, 2019) combined the analysis of ENIGMA cohorts using fixed-effect methods from METAL (Willer, 2010), through inverse-variance-weighted model (Hibar, 2015) or sample-size-weighted model (Satizabal, 2019). Although, the neuroimaging responses considered are volumes of particular regions of interest (ROI) in the brain that are summarized from voxelwise neuroimage data. Our goal for project 3 is to apply meta-analysis on high-dimensional (e.g.

voxelwise) neuroimaging data and account for heterogeneity between studies. Therefore, we review the existing meta-analytic approaches in section 2.3.1 and discuss approaches beyond meta-analysis to account for between-study heterogeneity in section 2.3.2.

2.3.1 Meta-Analysis

Meta-analysis is widely used to integrate analysis from multiple studies. It is particularly useful when it is inappropriate to pool data from different studies for a mega-analysis because of constraints in sharing of individual level data or differences in study design, modeling method, or phenotype distribution. Meta-analysis results in little or no loss of efficiency compared to mega-analysis (Lin, 2010). There are two types of models for meta-analysis: the fixed-effects models and the random-effects models (DerSimonian and Laird, 1986). Suppose there are K studies, from which we would like to summarize the analysis results. For $k = 1, \dots, K$, let β_k be the parameter of interest, e.g. treatment or genetic effects. In fixed-effects models, we assume homogeneous predictor effect across studies: $\beta_k = \beta$ for all $k = 1, \dots, K$. In random-effects models, we assume that the predictor effect varies across models: $\beta_k = \beta + \xi_k$, where $\xi_k \sim N(0, \tau^2)$.

Random-effects models address the between-study heterogeneity, which may be due to population structure or other measured or unmeasured variables that are not accounted for in the model. Work has been done to further investigate random-effects meta-analysis models. Han and Eskin (2011) pointed out that previous random-effects models have been using $\hat{\tau}^2_{MA}$ under the null hypothesis $H_0 : \beta = 0$, which gave rise to the phenomenon that fixed-effects models are almost always more powerful than random-effects models, even with sufficient heterogeneity between studies. They found that previous mixed-effects models conducts a test equivalent to a likelihood ratio test using $\hat{\tau}^2_{MA}$ in the null model. Therefore, they made a correction by assuming $\tau^2 = 0$ under the null, i.e. using a joint null hypothesis: $H_0 : \beta = \tau^2 = 0$. This idea is taken by Tang and Lin (2014), where random-effects meta-analysis models are developed for commonly used gene-level (i.e. multi-marker) association tests, including burden test, variable threshold test, and variance-component test. They demonstrated that their proposed random-effects tests under the corrected null hypothesis are substantially more powerful than the fixed-effects tests in the presence of moderate and high between-study heterogeneity and achieve similar power to the latter when the heterogeneity is low. Their methods are implemented in the software MASS (Tang, 2013).

In order to address between-study heterogeneity arising from trait-dependent sampling (TDS) in sequencing studies, which is usually adopted for economic feasibility, Lin, Zeng, and Tang (2013) proposed a maximum likelihood estimation (MLE) approach to test the genetic association with the primary trait and with the secondary trait that is usually associated with the primary trait. This work corrects the inflated type I error and possibly decreased power caused by ignoring the trait-dependent sampling. A meta-analysis strategy is proposed to properly combine the association results from multiple studies with measurements of a trait, on which the sampling is based. Their proposed meta-analysis is substantially more powerful than the analysis of any single study, and the meta-analysis of results from standard linear regression (ignoring trait-dependent sampling) can be less powerful than the analysis of a single study. The method supports score test, Wald test, and likelihood ratio test, and the score test, referred to as SCORE-SeqTDS, is implemented in the software MASS (Tang, 2013; Tang, 2017).

Besides MASS (Tang, 2013), other commonly used meta-analysis software for sequencing studies (accounting for multiple markers in each model and its test) include RAREMETAL (Liu, 2014) and MetaSKAT (Lee, 2013). RAREMETAL implements methods under fixed-effects models and MetaSKAT implements methods under both fixed-effects models and random-effects models. One of the mostly-used software for association scans is METAL (Willer, 2010). It combines across studies either the p-values or the test statistics and the corresponding standard errors, where sample sizes and directions of effect can be taken into account.

2.3.2 Refined Meta-Analysis

DerSimonian and Laird (2015) reviewed meta-analysis in clinical trials and mentioned

Han and Eskin (2011)'s work on the alternative null hypothesis $H_0 : \beta = \tau = 0$. For the refined method, it recommended the robust variance estimate (Sidik and Jonkman, 2005): instead of assuming $var(Y_i) = (w_i)^{-1}$, $Var(Y_i)$ is replaced by $(Y_i - \hat{\mu})^2$, i.e.

$$var(\hat{\mu}) = \sum \hat{w}_i (Y_i - \hat{\mu})^2 / \sum \hat{w}_i (K - 1) \tag{2.37}$$

Jackson and Riley (2014) extended this robust variance estimate from univariate case to multivariate meta-analysis and meta-regression. They showed through simulation that this refined method performed better than standard method when the number of studies $n = 2, 3$. Ito and Sugasawa (2019) addresses the problem of underestimating standard error using standard method due to the small number of studies. The proposed method does not depend on bootstrap, has a relatively simple expression, and is theoretically justified.

2.3.3 Meta-Analysis for Functional Models

On sequencing meta-analysis, Chiu et al. (2017) proposed a method to combine multivariate response (Wang et al., 2015) with a functional model for meta-analysis (Fan et al., 2015). In particular, the genetic effect from sequencing data is modeled through a fixed-effect functional model. Aimed for neuroimaging data modeling, Sørensen et al. (2020) developed meta-analysis for generalized additive models (GAM) (Hastie and Tibshirani, 1986) $g(\mu) = \beta_0 + \sum_{s=1}^{S} f_s(\mathcal{X})$ for both fixed effect meta-analysis and random effect meta-analysis in a pointwise manor through standard approaches. Software implementation of this method is available as an R package metagam.

2.3.4 Comparison with Mega-Analysis

Mega-analysis is an alternative approach to integrate information across studies by pooling data from multiple studies. Although imaging-genetic studies may collect similar types of data such as T1 images, functional MRIs, and genotypes for each individual, the protocol of data collection and processing usually differ across studies. In addition, the population composition could also vary across studies, such as age, disease status, and ethnicity. Mega-analytic approaches have been applied to address between-study heterogeneity, such as analysis of large clusters (Bellamy, 2005), direct surrogate variable analysis (dSVA) (Lee, 2017), and the confounder adjusted testing and estimation (CATE) approach (Wang, 2017). These methods are also extended to the high-dimensional space. The massive-univariate analysis of neuroimaging data (Guillaume, 2018) modeled the unknown covariates via adopting and modifying the CATE, and Huang (2019) embedded surrogate variable analysis (or latent effect adjustment, confounder adjustment) in a functional regression model, treating images as functional responses.

Lin and Zeng (2010) compared the efficiency of meta-analysis versus mega-analysis and found that for all commonly used parametric and semiparametric models, there is no asymptotic efficiency gain by analyzing original data if the parameter of main interest has a common value across studies, the nuisance parameters have distinct values among studies, and the summary statistics are based on maximum likelihood. Zeng and Lin (2015) derived asymptotic properties of the estimated parameters $(\hat{\beta}_{MA}, \hat{\tau}^2_{MA})$ under random-effects meta-analysis model and compared their efficiency with MLE estimators $(\hat{\beta}_{MLE}, \hat{\tau}^2_{MLE})$ from mega-analysis using analysis of large clusters (Bellamy, 2005). The surprising finding from Zeng and Lin (2015) is that the former is always at least as efficient as the latter.

As for prediction instead of hypothesis testing, Guan et. al. (2019) compared ensembling (i.e. cross-study learning or meta-analysis) and merging (i.e. mega-analysis) in terms of prediction accuracy, measured by mean squared prediction error (MSPE). The comparison is considered in the linear regression setting using random-effects meta-analysis model, where coefficients are assumed to vary across studies. They show analytically and confirm via simulation that merging yields lower prediction error than cross-study learning when the predictor-outcome relationships are relatively homogeneous across studies; as heterogeneity increases, there exists a transition point beyond which cross-study learning (i.e. meta-analysis) outperforms merging. The above discussion is considered in two scenarios: (1) least squares (LS) modeling when the number of predictors is less than the sample size; (2) ridge regression

	Categories		**Name**	**Method**
2.3.1	fixed-effects model			$\beta_k = \beta$
	random-effects models	Original	DerSimonian & Laird, 1986	$\beta_k = \beta + \xi_k$
		Corrected	Han & Eskin 2011	$H_0 : \beta = \tau^2 = 0$
			Tang & Lin 2014	For multi-marker association tests: burden test, variable threshold test, variance-component test
	TDS		Lin et al 2013	MLE approach for TDS in sequencing studies
2.3.2	Mega-analysis	Analysis of large clusters	Bellamy et al 2005	Analysis of large clusters
		dSVA	Lee et al 2017	Direct surrogate variable analysis
		CATE	Wang et al 2017	Confounder adjusted testing and estimation
		Functional	Huang 2019	Surrogate variable analysis in functional framework
	Meta vs mega	Hypothesis testing	Lin & Zeng 2010 Zeng & Lin 2015	No asymptotic efficiency gain by analyzing original data
		Prediction	Guan et al 2019	Mega-analysis yields lower MSPE than meta-analysis when $Var(\xi_k)$ is small

Table 2.4: Summary of methods reviewed in Section 2.3.

modeling when the number of predictors is larger than the sample size. Under each scenario, the author derived the corresponding transition points for equal and unequal variances of elements in the random coefficient vector, respectively.

Methods for ensembling are similar to those for meta-analysis in the way that they both integrate results from each study through a weighted average. Therefore, theories from ensembling can potentially be extended to statistical testing meta-analysis. The difference between the two is that, prediction accuracy (e.g. MSPE) is used to compare ensembling versus merging, while statistical efficiency (e.g. variance of the coefficient of interest) is usually examined when comparing meta-analysis against mega-analysis.

CHAPTER 3: REGION-BASED FUNCTIONAL METHOD

3.1 Introduction

As technology advances in data acquisition for brain images and genetic data, the number of brain imaging genetic studies are growing with increasing sample sizes. Such studies build on our knowledge in health-related mechanism and contribute in disease prevention and treatments. Therefore, efficient approaches are in need to learn important information from imaging-genetic data. It is of wide interest to detect association between brain images and genetic markers. In particular, one is interested in identifying region in the brain image that is associated with a genetic marker or a set of genetic markers from a pool of thousands or millions of genetic markers. This association detection was first handled by performing analysis for each voxel (or pixel for 2D image) in the image and for each marker or marker set throughout the whole genome (Stein, 2010). Recent approaches improved detection power through making use of intrinsic spatial smoothness and correlation in brain images (Huang, 2017; Lin, 2014; Zhang, 2018; Kong, 2019).

There are still problems unsolved that motivate our proposed approach: in brain images, the region associated with a predictor may not be well aligned across subjects and tend to be heterogeneous across subjects (Liu, 2018; Huang, 2019); in addition, registration errors also arise from the image preprocessing, while most existing approaches assume perfect alignment and registration across subjects. We propose a region-based functional genome-wide association detection (rfGWAS) approach to address the above problems. In particular, we move a small region (i.e. a 2D or 3D window) across the image domain and perform functional genome-wide association detection on the small region at each location as it moves. We consider the probability distribution of measurements on all voxels in the region as the functional response in rfGWAS. The association is then tested between the distribution function and genetic markers. This approach can then be used to detect regions that contain genetically-associated regions (GARs) that vary in location and size.

Compared to traditional voxel-based methods, treating a small region as the finest unit reduces computation complexity from $N_g \times M$ to $N_g \times m$, where N_g is the number of markers (or marker sets), M is the number of voxels in the entire image, and m is the number of small regions (i.e. window locations) on which we test the association. The high computation complexity of voxel-based methods, including single-voxel methods and functional methods, may not be necessary, because one is not necessarily interested in whether a single voxel is associated with a predictor. The detected association is regarded interesting and meaningful only if it is larger than a certain size. Although several methods (Ge et al., 2012; Gong et al., 2018) use cluster-size analysis with random field theory to summarize voxelwise results into clusters (i.e. regions) in the image, those approaches are still based on computation on each voxel.

In the method section below, we elaborate on how rfGWAS works. In the simulation section, we tested the method in several scenario with multiple choices of the association testing method. We then apply rfGWAS5r to the hippocampus surface data from the ADNI study (Weiner et al., 2017) and present the detected signals. Lastly, we discuss pros and cons of rfGWAS, as well as future directions to pursue.

3.2 Method

We propose the region-based functional genome-wide association (rfGWAS) approach: first cover the entire image with small regions, then summarize the voxelwise values in each region by its estimated distribution function and transform it to a function in a Hilbert space, and lastly test the association of the transformed distribution function with the predictor of interest.

Consider a sample of n individuals, where each individual i $(i = 1, \dots, n)$ has an image $\{y_i(s) : s \in \mathcal{D}\}$, $\mathcal{D} = \{s_1, \dots, s_M\}$, and a $(p \times 1)$ vector of predictors of interest (e.g. genetic markers) $\boldsymbol{x}_i$.

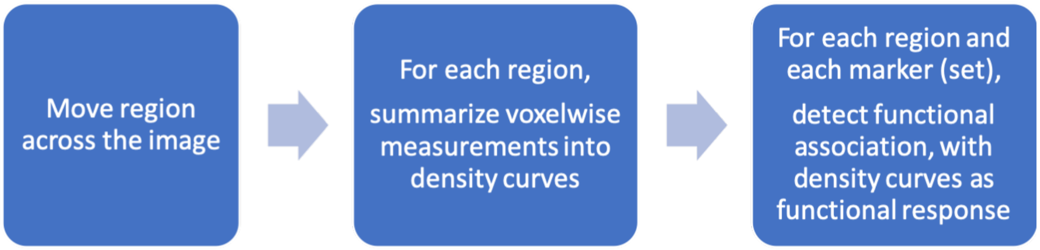

Move region across the image
For each region, summarize voxelwise measurements into density curves
For each region and each marker (set), detect functional association, with density curves as functional response

Figure 3.1: Steps in the region-based functional genome-wide association (rfGWAS) pipeline.

The following subsections describes in details the components of our pipeline: (i) summarize the image using regional distributions and the log-quantile transformation; and (ii) pin down the predictor (e.g. genetic marker) - subregion pairs as signals identified in the association test.

3.2.1 Regional Distributions and Transformation

The rfGWAS approach is inspired by the sliding-window strategy (Hudson et. al., 1988) in genetic data analysis, where a window of a certain length slide along the sequence of markers in disjoint steps with a certain step size. Each analysis is done with all markers in the window included. In genetic analysis, markers lie in a linear sequence and the window is one dimensional (1D). In rfGWAS, we take a similar approach to slide a 2D window across the entire 2D image or a 3D window across the entire 3D image with a particular step size, so that the entire image is fully covered by all the possible window locations and there are some overlap between adjacent window locations. In this chapter, we refer each window location as a "region" in the image, and the entire imaging domain $\mathcal{D}$ is thus covered by multiple overlapping regions $\{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$: $\mathcal{D} = \cup_{k=1,\ldots,m}\mathcal{D}_k$. From now on, we assume the image is 3D in our following discussion for simple wording; the 2D scenario will be similar.

After the division, we analyze each region for each marker (set) parallelly. For each marker, we treat each region as a smallest/finest distinguishable unit and do not distinguish voxels within this unit region. In particular, we focus on the distribution of the measurement on all voxels within this region and estimate the distribution function $f_{ij}(y)$ for the region in the image $\{y_i(s) : s \in \mathcal{D}_j\}$ of each subject i. We are then interested to test the association between $f_{ij}(y)$ and marker (set) $\boldsymbol{x}_i$ across $i = 1, \ldots, n$.

This conduct of treating a small region on the image as a finest unit instead of distinguishing each unit in the image can also deal with registration imperfectness. In voxelwise-treating scheme, each voxel is assumed to be aligned perfectly across subjects, while in reality, the registration of a voxel across subjects would jitter around the perfect-registration locus. Specifically, given a region $D \in \mathcal{D}$, the corresponding region measured in subject i is $D_i = \phi_i(D_0)$, where the map $\phi_i(\cdot)$ is random. Then, the voxelwise method performed on voxel s_k is in reality performing the analysis on $\phi_i^{-1}(s_k)$ in the perfect registration scenario. This is not clinically rigorous if we assume biological significance vary across voxels.

Now let us discuss the scheme of treating a region as the finest distinguishable unit in the analysis. We would like to detect a signal at D_0. Suppose the sliding region at this area $\mathcal{D}_k$ is large enough s.t. it covers $D_0 \cup (\cup_{i=1,\ldots,n} D_{0,i}) \in \mathcal{D}_k$. Then, the distribution $f_{ik}(\cdot)$ would solely rely on the signal with some random noise, and the registration error would not affect $f_{ik}(\cdot)$. Therefore, we continue with our signal-detection pipeline based on the regional distribution $f_{ik}(\cdot)$.

As addressed by Petersen and Müller (2016), the density histogram integrates to 1 and and thus does not lie on the Hilbert space where the common statistical methods are developed; thus, they propose to transform the density function f_{ij} to the Hilbert space using the log-quantile density (LQD) transformation:

$$g_{ij}(t) = -log\{f_{ij}(Q_{ij}(t))\}, \quad \text{where } t \in [0,1], \tag{3.1}$$

where $Q_{ij} = F_{ij}^{-1}$ is the quantile function, and $F_{ij}(y) = \int_{-\infty}^{y} f_{ij}(u)du$ is the cumulative distribution function of $\{y_i(s) : s \in D_k\}$. We then transform the problem to testing the association between $g_{ij}(t)$ and $\boldsymbol{x}_i$.

When dealing with functional data, there are two general approaches (Liu, 2018): one approach is to treat it as the sum of weighted base functions, and the other approach is to see it as a set of interpolated values of the function, in our case, $\{g_{ij}(t) : t \in T\}$, where $T = \{t_1, \ldots, t_r\}$. Here, we adopt the second approach, where r can be a relatively small integer, such as 9 and an example of T is $\{0.1, 0.2, \ldots, 0.9\}$. Then, $q_{ijk} = Q_{ij}(t_k)$ is the $(100 \times t_k)$th percentile (or the kth quantile). Therefore, the LQD is also a function of the quantiles q_{ijk}, $k = 1, \ldots, r$:

$$\tilde{g}_{ij}(q) := -log\{f_{ij}(q)\}, \quad \text{where } q \in \mathcal{R}. \tag{3.2}$$

We choose constants a and b s.t. $a < 0$ and $aC + b > 0$, where $C = \max(q)$. This gives

$$\frac{\partial \tilde{g}}{\partial f} = -\frac{1}{af + b} \in [-\frac{1}{aC + b}, -\frac{1}{b}], \tag{3.3}$$

s.t. $\tilde{g}$ does not explode for values of $f(\cdot)$ close to 0, and $\Delta g(\cdot)$ is more sensitive to changes when the values of $f(\cdot)$ are large.

3.2.2 Region-based Statistical Test

As a result, the problem is now testing the association between multivariate variable $\boldsymbol{g}_{ij} := (g_{ij}(t_1), \ldots, g_{ij}(t_r))^T$ and single- or multi-variate variable $\boldsymbol{x}_i$. There are a full class of statistical methods to deal with this problem. In particular, we will choose the ball covariance test (Pan, 2019), which is based on the comparison between the two inter-subject similarity/distance matrices looking at the two variables respectively, regardless the whether the dimension of each of the variable equals to 1. This nonparametric method stand out from the commonly used linear-regression-based models in two ways: firstly, this ball covariance measurement does not restrict the relationship to be measured between the two variables to be linear; secondly, this ball covariance test directly addresses the interest of association detection and does not force one variable to be the response and the other to be a predictor, thus avoid the hassle to decide between an image-on-scalar model and a scalar-on-image model, where we are not interested in prediction using some estimated coefficients.

$$\begin{aligned} &H_0 : \boldsymbol{g}_{ij} \text{ is not associated with } \boldsymbol{x}_i. \\ &H_1 : \boldsymbol{g}_{ij} \text{ is associated with } \boldsymbol{x}_i. \end{aligned} \tag{3.4}$$

GBJ Test The input of the GBJ test are: (1) a test on each $t \in T$; (2) the correlation between t's. Using the GBJ function of the GBJ R package, we obtain the output, which is the global test across $t \in T$.

The GBJ package is designed for correlated tests in GWAS, where the correlation between tests are due to correlated predictors (genetic markers) in each linear model. We use the GBJ test slightly different in the way that our correlation between tests are due to the correlation between response variables, denoted as Y_j and $Y_{j'}$.

Now we derive the correlation between test statistics at t_j and $t_{j'}$. Suppose a t-test is used.

In a linear model, the estimated effect

$$\hat{\beta}_j = (X^T X)^{-1} X^T Y_j, \tag{3.5}$$

and its variance is

$$Var(\hat{\beta}_j) = Var((X^T X)^{-1} X^T Y_j) = \sigma_j^2 (X^T X)^{-1}. \tag{3.6}$$

The t-test statistic at t_j is

$$T_j = \frac{\hat{\beta}_j}{\widehat{se}(\hat{\beta}_j)} = \frac{\sqrt{(X^T X)^{-1}} X^T Y_j}{\sigma_j}, \tag{3.7}$$

where $sigma_j$ is the standard error of the error term in the linear model at t_j.

Therefore, the correlation between the t-test statistics

$$\begin{aligned} Corr(T_j, T_{j'}) &= Cov(T_j, T_{j'}) \\ &= Cov(\frac{\sqrt{(X^TX)^{-1}}X^TY_j}{\sigma_j}, \frac{\sqrt{(X^TX)^{-1}}X^TY_{j'}}{\sigma_{j'}}) \\ &= Cov(\frac{Y_j}{\sigma_j}, \frac{Y_{j'}}{\sigma_{j'}}) \end{aligned} \tag{3.8}$$

Ball Covariance Test For a genetic marker (group) $\boldsymbol{x}_i$, let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ denote the matrix of markers (or marker groups), i.e. predictors, for all individuals and $\boldsymbol{G}_j = (\boldsymbol{g}_{1j}, \ldots, \boldsymbol{g}_{nj})^T$ the matrix of transformed regional density of the image for all individuals. Using the ball covariance (BCov) test, H_0 is

$$BCov_{w,n}(\boldsymbol{G}_j, \boldsymbol{X}) = 0, \tag{3.9}$$

where $BCor_{w,n}$ is a measure of association introduced by Pan, et al. (2019) and it equals to 0 if and only if there is no association between $\boldsymbol{x}_i$ and $\boldsymbol{G}_j$. Ball covariance is defined as follows:

For random variables X and Y,

$$\begin{aligned} BCov_w^2(X,Y) := \int (\theta - \mu \otimes \nu)^2 [\bar{B}_\rho(x_1, x_2) \times \bar{B}_\zeta(y_1, y_2)] \\ w_1(x_1, x_2) w_2(y_1, y_2) \theta(dx_1, dy_1) \theta(dx_2, dy_2), \end{aligned} \tag{3.10}$$

where ρ and ζ are distance measures for X and Y, respectively; $(X,Y) \sim \theta$, $X \sim \mu$, and $Y \sim \nu$. Here, we use the Euclidean distance for the distance measures.

The empirical ball covariance is defined as

$$BCov_{w,n}^2(X,Y) := \frac{1}{n^2} \sum_{i,j=1}^{n} (\Delta_{ij,n}^{XY} - \Delta_{ij,n}^{X} \Delta_{ij,n}^{Y})^2 \hat{w}_1(X_i, X_j) \hat{w}_2(Y_i, Y_j), \tag{3.11}$$

where

$$\Delta_{ij,n}^{XY} = \frac{1}{n}\sum_{k=1}^{n}\delta_{ij,k}^{X}\delta_{ij,k}^{Y}, \quad \Delta_{ij,n}^{X} = \frac{1}{n}\sum_{k=1}^{n}\delta_{ij,k}^{X}, \quad \Delta_{ij,n}^{Y} = \frac{1}{n}\sum_{k=1}^{n}\delta_{ij,k}^{Y}, \tag{3.12}$$

where $\delta_{ij,k}^{X} := I\{X_k \in \bar{B}_\rho(X_i, X_j)\}$ and $\delta_{ij,k}^{Y} := I\{Y_k \in \bar{B}_\zeta(Y_i, Y_j)\}$. $B_\rho(X_i, X_j)$ denotes the ball with center X_i and radius $d(X_i, X_j)$. $B_\zeta(Y_i, Y_j)$ denotes the ball with center Y_i and radius $d(Y_i, Y_j)$.

In our case, we can use Euclidean distance when calculating the distance between two subjects in $\boldsymbol{X}$ and in $\boldsymbol{G}$:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{i'k})^2} \tag{3.13}$$

$$d(\boldsymbol{g}_{ij}, \boldsymbol{g}_{i'j}) = \sqrt{\sum_{k=1}^{r}[g_{ij}(t_k) - g_{i'j}(t_k)]^2} \tag{3.14}$$

A PC-based Euclidean distance can be used alternatively to account for correlation between dimensions in X or G.

Under H_0,

$$nBCov_{w,n}(G_j, X) \xrightarrow[d]{n \longrightarrow \infty} \sum_{v=1}^{\infty}\lambda_v Z_v^2,$$

where Z_v's are independent standard normal random variables, and λ_v's are non-negative constants that depend on the distribution of (G_j, X). P-value is calculated through parametrically as in (Pan, et. al., 2019). The test is conducted using the "ballgamma" package in R.

Cluster Size Test Besides regional tests that favors strong region-marker association that passes the multiple-adjustment threshold, another type of test that favors a moderate strength of region-marker association but across a significantly large area also provides meaningful insights in the phenotype-marker relationship. The latter is typically examined using the so-called "cluster-size analysis", which is a common conduct in imaging genomics (Ge et al., 2012; Huang et al., 2017).

We implement the cluster-size test here to provide insight on whether a large area in the image is associated with a marker, where each (small) region in the area has association with the marker that passes a user-defined threshold α_C.

For marker g and a region j, let $p_j(g)$ be the p-value of the regional test between region j and marker g. Then, obtain $S(g) = \{j : p_j(g) < \alpha\}$, which is the set of regions with p-value $p_j(g)$ passing the threshold α. The regions in $S(g)$ can be isolated or connected (termed "clustered" in the cluster-size analysis). Group the connected regions in $S(g)$ into "mega-regions": $S_m(g) = \{\boldsymbol{A}_k : \boldsymbol{A}_k = \{A_j \ (j = 1, \ldots, m_{gk}) : A_j \text{ and } A_{j'} \text{ are connected}\}\}$, where $\boldsymbol{A}_m$ can be a single region A_j if A_j is isolated (not connected to any other region $A_{j'}$), i.e. we allow $m_{gk} = 1$. Let $A(g, \alpha)$ be the area of the largest mega-region in $S_m(g)$. Then, the p-value associated with $A(g, \alpha)$ can be obtained through wild bootstrap, which controls for the FWE (Huang et al., 2017, 2015).

3.2.3 Global Screening

In order to reduce computation, we propose a global screening procedure to subset the large number of markers before formal statistical tests on each region and each marker.

Subsample-based Screening For the purpose of screening to reduce computation, we implement an approximated T_j based on subsampling. Given the fact that T_j based on ball correlation requires a computational complexity $O(n^2 \log n)$ under our multivariate case, the most efficient way to reduce computation complexity is to reduce n. We have considered another possibility of simplifying the algorithm based on the fact that the response is unchanged when testing against all markers, although the computation reduction using relevant technique does not change the order of the computation complexity and only result in a minor reduction in computation. Besides the fact that computation of T_j is a high order of n, the major reason we consider using the subsampling scheme to reduce computation resides in the nature of genotype data – most markers have low minor allele frequencies (MAFs). Given a marker, let p be its MAF and $q = 1 - p$, and let A and B be its major and minor allele, respectively. Under Hardy-Weinberg equilibrium, the number of individuals with two mutation alleles, i.e. with genotype BB, is expected to be np^2; the number of individuals with one mutation allele, i.e. with genotype AB (or BA), is expected to be npq; and the number of individuals with both alleles non-mutated, i.e. with genotype AA, is expected to be nq^2.

Let's look at a real example: for the UK Biobank imaging cohort, currently there are nearly 40,000 individuals with images available. For a SNP with MAF 0.1, i.e. $p = 0.1$, the number of individuals in the three genotype groups BB, AB, and AA are 400, 7,200, and 32,400, respectively. Statistical methods model the relationship between the genotype and the response of interest, denoted as y, to investigate the difference between these three groups in terms of y. Given the large difference in the sample sizes between the three groups, our intuition is that the power of a statistical test would not suffer much if we reduce the sample size in the genotype-AA group (or also in the genotype-AB group). For a most conservative sample size reduction – reducing the sample size of genotype-AA group to 7,200, our total sample size $\tilde{n}$ becomes 14,800. With a computation complexity to the order of $n^2 \log(n)$, the computation is reduced by 8 times as a result. If we further reduce the sample size in each group to 400, the resulting total sample size of 1,200 would give a computation reduction of 1,660 times.

In genomic analyses, a large sample size enables signal detection given a large number of markers for multiple comparison adjustment; but for the purpose of screening, we include a conservative large pool markers to allow the computation complexity for the next-step analysis. Therefore, a smaller sample size would not be a problem for the use of screening, and a smaller sample size should not largely affect the rank of the test statistics. Therefore, we propose a screening procedure using a simplified test statistic based on a subset of individuals.

In particular, we conduct the subsampling by setting a total subsample size $\tilde{n}$. Suppose we have x in integer dosage values 0, 1, and 2. Let $\tilde{n}_0, \tilde{n}_1, \tilde{n}_2$ be the corresponding sample size in each dosage group in the subsample (and let n_0, n_1, n_2 be the corresponding sample sizes in the entire sample with size n). We propose a simple subsampling strategy for the tentative testing stage, which can potentially be improved and optimized in later work. The ideal case of testing the difference between the group for optimized power would to have balanced sample size $\tilde{n}_0 = \tilde{n}_1 = \tilde{n}_2 = \tilde{n}/3$, but given the fact of Hardy-Weinberg equilibrium, we expect $\tilde{n}_2 \leq \tilde{n}_1 \leq \tilde{n}_0$, and it is even likely that $tilden/3 > n_2$, which is the total number of individuals with dosage level $x = 2$ available in the entire cohort. Therefore, we take $\tilde{n}_2 = \min(n_2, \tilde{n}/3)$, and then from the remaining sample size $\tilde{n} - \tilde{n}_2$, subsample $\tilde{n}_1 = \min(n_1, (\tilde{n} - \tilde{n}_2)/2)$ and $\tilde{n}_0 = \tilde{n} - \tilde{n}_2 - \tilde{n}_1$. Samples are randomly drawn in each dosage group with the designated subsample size (i.e. $\tilde{n}_k$ samples randomly drawn from n_k subjects in dosage group $x = k$, for $k = 0, 1, 2$). This subsample technique pushes the subsample sizes to the optimal case of $\tilde{n}_0 = \tilde{n}_1 = \tilde{n}_2 = \tilde{n}/3$, and when the sample sizes in dosage groups are strongly unbalanced that this optimal scheme cannot be fulfilled, the above proposed subsampling strategy results in $\tilde{n}_0, \tilde{n}_1, \tilde{n}_2$ tilted towards the distribution of sample size in dosage groups, s.t. the resulting statistic tilded towards the statistic based on the entire sample, given that the value of the statistic could be influenced by the sample size distribution (n_0, n_1, n_2) in the dosage groups.

A Global Statistic Let $\boldsymbol{T} = (T_1, \ldots, T_m)^T$ be the vector of test statistics for regions $1, \ldots, m$. Let $e_j = I\{T_j > \alpha_G\}$, where $j = 1, \ldots, m$ and α_G is a global-screening threshold; let $J_{\alpha_G} = (e_1, \ldots, e_m)$. Construct a global statistic as the average of large elements in $\boldsymbol{T}$ that pass the global-screening threshold α_G:

$$T_G = \frac{J_{\alpha_G}\boldsymbol{T}}{J_{\alpha_G}J_{\alpha_G}^T} = \frac{1}{\tilde{m}} \sum_{j:T_j > \alpha_G} T_j, \tag{3.15}$$

where $\tilde{m} = J_{\alpha_G}J_{\alpha_G}^T$ is the number of regions with $T_j > \alpha_G$. The distribution of T_G is not derived for its complexity when T_j's are ball correlations and also for the reason that in global screening, markers with top-ranked T_G's are selected without the need to calculate the corresponding p-values.

The Screening Procedure Sort T_G's for all markers and select the n_0-top-ranking markers. A simplified calculation of BCor for SNP screening is presented in the Appendix.

3.3 Simulation

3.3.1 Compare association testing methods on one region or one voxel

We simulated 10 settings as in section 4.1 of Pan et. al. (2018) to compare the statistical testing methods on one region or one voxel. The first 4 settings are generated under the null hypothesis, where X and Y are independent. In settings 6 to 10, X and Y are related. The 10 settings are described as follows:

1. X, Y are independent from the standard normal distribution $N(0,1)$.

2. X, Y are independent from the binomial distribution $B(10, 0.5)$.

3. Z_1, Z_2 are independent from the binomial distribution $B(10, 0.5)$, and

$$X = Z_1^2 + Z_1,\ Y = Z_2^2 + Z_2.$$

4. (X, Y) are from the 10-dimensional multivariate normal distribution with $\mu = 0$, $cov(X_i, X_i) = cov(Y_i, Y_i) = 1,\ i = 1, \ldots, 5,\ cov(X_i, X_j) = cov(Y_i, Y_j) = 0.2,\ 1 \leq i < j \leq 5$ and $cov(X_i, Y_j) = 0, i, j = 1, \ldots, 5$.

5. $Z_1, \ldots, Z_4$ are independent from the binomial distribution $B(10, 0.5)$, $Z_5 \sim B(20, 0.5)$,

$$Y = sin(Z_1)(Z_2 + Z_5)^2/(Z_3 + 1) + log(Z_4 + Z_5 + 1),$$

where $X = (Z_1, \ldots, Z_4)$.

6. Z_1, Z_2, Z_3, Z_4 are independent from the standard normal distribution $N(0,1)$, and

$$Y_1 = (4 - Z_1^2)Z_3Z_4 + Z_2Z_4^2,$$
$$Y_2 = (4 - Z_2^2)Z_3Z_4 + Z_1Z_4^2,$$

where $X = (Z_1, Z_2, Z_3)$, $Y = (Y_1, Y_2)$.

7. X, Y are from the 10-dimensional multivariate normal distribution with mean $\mu = 0, cov(X_i, X_i) = cov(Y_i, Y_i) = 1$, $i = 1, \dots, 5$, $cov(X_i, X_j) = cov(Y_i, Y_j) = 0.2$, $1 \le i < j \le 5$ and $cov(X_i, Y_j) = 0.2$, $i, j = 1, \dots, 5$.

8. $Z_1, \dots, Z_6$ are independent from the standard normal distribution $N(0,1)$,

$$Y_1 = \sqrt{|Z_1 + Z_4^2|} + Z_5^2,$$
$$Y_2 = Z_2Z_5 + \tanh(Z_6),$$

where $X = (Z_1, Z_2, Z_3, Z_4, Z_6)$, $Y = (Y_1, Y_2)$ and $\tanh(\cdot)$ is the hyperbolic tangent function.

9. $Z_1, \dots, Z_5$ are independent from the t distribution with 3 degrees of freedom and

$$Y_1 = (Z_1 + Z_2)^3 + Z_5^2,$$
$$Y_2 = Z_3Z_4 + \tanh(Z_5),$$
$$Y_3 = Z_1^2 + Z_5,$$

where $X = (Z_1, Z_5)$, $Y = (Y_1, Y_2, Y_3)$ and $\tanh(\cdot)$ is the hyperbolic tangent function.

Setting	Dim of X	Dim of Y	BCov	RdCov	sKPRC	AdaMant	lm
1	1	1	x	x		x	x
2	1	1	x	x		x	x
3	1	1	x	x		x	x
4	5	5	x	x	x		
5	4	1	x	x		x	x
6	3	2	x	x	x		
7	5	5	x	x	x		
8	6	2	x	x	x		
9	2	3	x	x	x		
10	7	2	x	x	x		
SNP	1	1	x	x		x	x

Table 3.1: Dimensions of variables and applicable statistical testing methods.

10. $Z_1, \ldots, Z_7$ are independent from the t distribution with 1 degree of freedom and

$$Y_1 = \tanh(Z_1 Z_2 Z_6),$$

$$Y_2 = \cos[I(Z_4 > 0)Y_1 + I(Z_4 < 0)Z_2],$$

where $X = (Z_1, \ldots, Z_7)$, $Y = (Y_1, Y_2)$ and $I(\cdot)$ is the indicator function.

The dimensions of X and Y of the 10 settings, as well as a setting to mimic the dosage values of genetic markers in X (detailed in the next subsection), are summarized in table and methods applicable are summarized in Table 3.1.

For each of the 10 settings, we simulated 1,000 random datasets with sample size 30, 50, 70, 100, respectively. Then we applied all applicable methods in each setting to test the association between X and Y. The null hypothesis is rejected at level of 0.05. Rejection rates among the 1000 repeats are listed in Table 3.2. Settings 1 to 4 reflects the type I errors under different scenario.

As can be seen from the above table, BCov with p-values calculated using parametric methods has large type I errors. We think that this is due to the small sample sizes ($n \leq 100$). In order to test this, we increased the sample size to 300, 500, 700, and 1,000, and generated 1,000 random samples for each setting and each sample size. The rejection rates (type I

s	n	BCov						Rd-	sK-	Ada-	lm
		const.	prob.	chisq	sw	hbe	lpb4	Cov	PCR	Mant	
1	30	0.049	0.061	0.051	0.158	0.149	0.153	0.045		0.049	0.055
	50	0.05	0.043	0.045	0.112	0.097	0.1	0.04		0.046	0.049
	70	0.056	0.053	0.059	0.115	0.101	0.105	0.048		0.048	0.049
	100	0.056	0.048	0.05	0.085	0.072	0.078	0.052		0.051	0.049
2	30	0.051	0.057	0.047	0.07	0.067	0.068	0.045		0.043	0.043
	50	0.049	0.053	0.041	0.067	0.063	0.065	0.042		0.059	0.054
	70	0.039	0.042	0.039	0.052	0.048	0.05	0.045		0.039	0.042
	100	0.041	0.031	0.046	0.051	0.044	0.047	0.055		0.051	0.057
3	30	0.046	0.052	0.043	0.076	0.071	0.072	0.045		0.038	0.045
	50	0.054	0.057	0.052	0.077	0.07	0.072	0.042		0.05	0.05
	70	0.037	0.042	0.039	0.059	0.053	0.053	0.045		0.042	0.049
	100	0.04	0.038	0.046	0.064	0.053	0.056	0.055		0.047	0.051
4	30	0.05	0.055	0.058	0.243	0.237	0.24	0.049	0.08		
	50	0.055	0.049	0.059	0.172	0.16	0.166	0.049	0.08		
	70	0.053	0.053	0.058	0.124	0.12	0.122	0.055	0.078		
	100	0.059	0.061	0.06	0.118	0.109	0.112	0.051	0.062		
5	30	0.866	0.934	0.935	0.975	0.97	0.972	0.187		0.217	0.221
	50	0.991	0.999	0.999	0.999	0.998	0.999	0.43		0.3	0.308
	70	1	1	1	1	1	1	0.69		0.352	0.36
	100	1	1	1	1	1	1	0.939		0.474	0.485
6	30	0.506	0.631	0.574	0.792	0.777	0.787	0.55	0.149		
	50	0.745	0.909	0.845	0.879	0.867	0.872	0.857	0.156		
	70	0.873	0.991	0.961	0.945	0.935	0.938	0.97	0.15		
	100	0.962	1	0.994	0.98	0.979	0.98	0.993	0.192		
7	30	0.448	0.355	0.457	0.725	0.715	0.717	0.1	0.089		
	50	0.712	0.569	0.723	0.849	0.837	0.845	0.15	0.06		
	70	0.866	0.739	0.876	0.936	0.931	0.931	0.289	0.081		
	100	0.97	0.913	0.977	0.99	0.988	0.988	0.49	0.074		
8	30	0.725	0.746	0.738	0.913	0.91	0.912	0.326	0.116		
	50	0.955	0.967	0.963	0.983	0.982	0.983	0.672	0.148		
	70	0.984	0.997	0.992	0.996	0.995	0.995	0.943	0.181		
	100	0.999	1	1	0.999	0.999	0.999	0.997	0.216		
9	30	0.993	1	0.998	0.999	0.999	0.999	0.949	0.628		
	50	1	1	1	1	1	1	1	0.782		
	70	1	1	1	1	1	1	1	0.858		
	100	1	1	1	1	1	1	1	0.889		
10	30	0.377	0.447	0.394	0.639	0.609	0.627	0.082	0.098		
	50	0.634	0.757	0.656	0.774	0.755	0.765	0.104	0.093		
	70	0.77	0.914	0.832	0.862	0.845	0.852	0.171	0.103		
	100	0.913	0.983	0.943	0.952	0.936	0.949	0.181	0.097		

Table 3.2: Rejection rates of 10 simulation settings to compare different methods on only one region or voxel. The first three BCov methods are based on bootstrap.

s	n	p.sw	p.hbe	p.lpb4
1	200	0.07	0.059	0.062
	300	0.07	0.055	0.065
	500	0.065	0.058	0.061
	700	0.063	0.057	0.06
	1000	0.051	0.045	0.048
2	200	0.073	0.068	0.071
	300	0.06	0.052	0.054
	500	0.075	0.069	0.071
	700	0.063	0.054	0.057
	1000	0.038	0.036	0.038
3	200	0.08	0.073	0.075
	300	0.054	0.052	0.053
	500	0.067	0.063	0.065
	700	0.061	0.059	0.06
	1000	0.043	0.038	0.04
4	200	0.073	0.061	0.066
	300	0.079	0.07	0.073
	500	0.069	0.06	0.064
	700	0.063	0.052	0.055
	1000	0.071	0.056	0.062

Table 3.3: Type I error rate of BCov, with p-values calculated using parametric methods, for larger sample sizes.

errors) for settings 1 to 4 for larger sample sizes are shown in Table 3.3.

3.3.2 Compare association testing methods on SNP dosages for different MAFs

Furthermore, we also tested the performance of the methods when X takes values 0, 1, or 2 to mimic the dosages of genotype data. We varied MAFs from 0.05, 0.1, 0.25, to 0.5. 1,000 repeated datasets are generated for each MAF level. The model used to generate the data is

$$Y_i = 0.4X_i + \epsilon_i, \tag{3.16}$$

where $\epsilon_i \sim N(0,1)$ and $X_i \sim Binom(2,p)$. For $p = 0.05, 0.1, 0.25$, and 0.5, the variation explained by the marker is $1.5\%, 2.8\%, 5.7\%$, and 7.4%, respectively. The statistical testing methods applicable in this setting is listed in Table 3.1, and rejection rates (powers) are shown in Table 3.4. Number of valid datasets simulated are also listed for each setting and

MAF	n	#Repeats	BCov.sw	BCov.hbe	BCov.lpb4	RdCov	AdaMant	lm
0.05	100	1000	0.10	0.10	0.10	0.09	0.23	0.22
0.05	200	1000	0.23	0.22	0.23	0.15	0.37	0.40
0.05	500	1000	0.55	0.53	0.54	0.25	0.77	0.78
0.05	750	1000	0.76	0.75	0.76	0.39	0.91	0.92
0.05	1000	1000	0.87	0.87	0.87	0.50	0.97	0.97
0.1	100	1000	0.20	0.20	0.20	0.19	0.35	0.37
0.1	200	1000	0.43	0.42	0.42	0.33	0.62	0.63
0.1	500	1000	0.83	0.82	0.82	0.69	0.96	0.96
0.1	750	1000	0.95	0.95	0.95	0.86	0.99	0.99
0.1	1000	1000	0.99	0.99	0.99	0.96	1.00	1.00
0.25	100	1000	0.41	0.40	0.41	0.51	0.64	0.67
0.25	200	1000	0.71	0.69	0.70	0.80	0.92	0.93
0.25	500	1000	0.98	0.98	0.98	1.00	1.00	1.00
0.25	750	999	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1000	1000	1.00	1.00	1.00	1.00	1.00	1.00
0.5	100	1000	0.42	0.40	0.41	0.67	0.80	0.81
0.5	200	1000	0.73	0.72	0.72	0.94	0.97	0.98
0.5	500	1000	0.99	0.99	0.99	1.00	1.00	1.00
0.5	750	1000	1.00	1.00	1.00	1.00	1.00	1.00
0.5	1000	1000	1.00	1.00	1.00	1.00	1.00	1.00

Table 3.4: Compare power of methods for smaller effect size and including larger sample size.

sample size, because when MAF is low and when the sample size is small, there are cases where all subjects are assigned dosage value 0. We excluded those cases in the association testing using different statistical methods and calculation of rejection rates. The numbers of valid repeats ($\leq 1,000$) are listed in column 3 of Table 3.4.

For a non-linear model

$$Y_i = -0.4(X_i - 1.2)^2 + \epsilon_i, \tag{3.17}$$

where $\epsilon_i \sim N(0,1)$ and $X_i \sim Binom(2,p)$. For $p = 0.05, 0.1, 0.25$, and 0.5, the variation explained by the marker is $2.6\%, 4.5\%, 6.6\%$, and 5.0%, respectively. The resulting powers of different methods are compared in Table 3.5.

MAF	n	#Repeats	BCov.sw	BCov.hbe	BCov.lpb4	RdCov	AdaMant	lm
0.05	100	1000	0.16	0.16	0.16	0.13	0.33	0.35
0.05	200	1000	0.37	0.36	0.37	0.21	0.57	0.59
0.05	500	1000	0.81	0.80	0.80	0.42	0.93	0.94
0.05	750	1000	0.95	0.95	0.95	0.64	0.99	0.99
0.05	1000	1000	0.99	0.99	0.99	0.78	1.00	1.00
0.1	100	1000	0.35	0.34	0.35	0.27	0.50	0.51
0.1	200	1000	0.63	0.62	0.63	0.51	0.79	0.81
0.1	500	1000	0.98	0.98	0.98	0.91	1.00	1.00
0.1	750	1000	1.00	1.00	1.00	0.98	1.00	1.00
0.1	1000	1000	1.00	1.00	1.00	1.00	1.00	1.00
0.25	100	1000	0.56	0.54	0.55	0.56	0.55	0.56
0.25	200	1000	0.85	0.84	0.84	0.86	0.86	0.87
0.25	500	1000	1.00	1.00	1.00	1.00	1.00	1.00
0.25	750	1000	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1000	1000	1.00	1.00	1.00	1.00	1.00	1.00
0.5	100	1000	0.38	0.37	0.37	0.25	0.22	0.22
0.5	200	1000	0.68	0.66	0.67	0.46	0.35	0.35
0.5	500	1000	0.97	0.97	0.97	0.87	0.66	0.69
0.5	750	1000	1.00	1.00	1.00	0.99	0.83	0.86
0.5	1000	1000	1.00	1.00	1.00	1.00	0.92	0.92

Table 3.5: Power comparison between methods for nonlinear dosage relationship.

3.3.3 Compare association testing methods on multiple regions with registration error

We simulate the imaging response with the effect from a single (genetic) predictor of interest. The imaging response is simulated to represent the left hippocampus surface data from ADNI with a total number of $M = 15,000$ grid points, with sample size $n = 100$, 200, or 500, which produces the response matrix with dimension $n \times M$. The coordinates of $M = 15,000$ voxels of the hippocampus surface is mapped to a 150×100 2-D rectangle. We thus work on this 2-D rectangle for subregion division and characterization of the connected affected region. We suppose the mechanism of how the predictor (genetic marker) affect a subregion of the response (image) as follows: there is a region A in the brain image that is affected by the genetic marker. In order to mimic the inaccurate registration of the images, we allow A to vary across subjects and let A_i be the affected region in subject i. Simulate A_i to be a round area with center (x_j, y_i) and radius r_i, and let $x_i \sim N(x_0, \sigma_x^2)$, $y_i \sim N(y_0, \sigma_y^2)$, and $r_i \sim N(10, \sigma_r^2)$. Let $(x_0, y_0) = (50, 50)$, $\sigma_x^2 = \sigma_y^2 \in \{1, 2, 2.5, 3, 4, 5\}$, and $\sigma_r^2 \in \{0, 1, 2, 3\}$. This yields a region with 317 voxels on average, which is approximately 2% of the entire rectangle (representing the hippocampus surface) area.

At each grid point d in the imaging response, simulate a $y(d)$ as the observed value of the functional response, which follows the model

$$y_i(d) = z_i \gamma(d) + x_i \beta_i(d) + \epsilon_i(d) \tag{3.18}$$

and let $\beta_i(d) = a\beta_0(d)$, where $\beta_{0i}(d) = 1$ for $d \in A_i$ and equals to 0 otherwise, and $a = \{0, 0.25, 0.5\}$, which allows us to evaluate the type I error and power at different effect sizes through the rejection rate among the 1000 simulating runs under each of these three scenario.

In this simulation study, we generate the response as a smooth function of d, so as to avoid the task of adding noise in the simulated data and then smooth the data in the preprocessing step. Therefore, $\epsilon_i(d)$ is simulated as

$$\epsilon_i(d) = \xi_{1,i}\psi_1(d) + \xi_{2,i}\psi_2(d), \tag{3.19}$$

where

$$\psi_1(d) = \sqrt{0.5}\sin(\frac{2\pi x_d}{150})$$
$$\psi_2(d) = \sqrt{0.5}\cos(\frac{2\pi y_d}{100})$$

and

$$\xi_{1,i} = \lambda_1 Z_{1,i}, \quad \xi_{2,i} = \lambda_2 Z_{2,i}; \ with$$
$$\lambda_1 = \sqrt{1.2}, \quad \lambda_2 = \sqrt{1.0},$$

where $Z_{1,i}$ and $Z_{2,i}$ are independent standard normal random variables; $i = 1, \dots, n$. This yields $\epsilon_i(d)$ to range from $(-6.29, 6.29)$, with sample variance 2.03, which is close to $2.2 = \lambda_1^2 + \lambda_2^2$. We assume that the response is already adjusted for the covariates and that x_i is a scalar that takes values 0, 1, and 2. Thus, x_i's are simulated from $Binom(2, 0.2)$. In summary, we simulate the data in the 13 settings as listed in Table 3.5.

We analyzed the simulated datasets using the rbGWAD pipeline. Firstly, cover the entire image with overlapping regions. Set $b = 10$ for the analysis in all the 13 settings, then each squared region with side length $2b$ would include $(2b)^2 = 400$ grid points. In the case of $M = 15,000$ and $b = 10$, the number of regions $m = 126 = (\frac{150}{b} - 1)(\frac{100}{b} - 1)$. Among which, we expect around 9 subregions to detect significant effect. Secondly, for region j $(j = 1, \dots, 126)$ of subject i $(i = 1, \dots, n)$, summarize the measurements of all voxels into the LQD function $g_{ij}(t)$ $(t \in T = [0, 1])$. In the analysis, we set T to be $r = 21$ equivalently spaced points on $[0, 1]$ to numerically represent $g_{ij}(t)$: $T = \{0, 0.05, 0.10, 0.15, \dots, 0.95, 1\}$. Lastly, test the association between $g_{ij}(t)$ $(t \in T)$ and x_i for $j = 1, \dots, 126$.

	β	n	$\sigma_x^2 = \sigma_y^2$	σ_r^2
1	0	200	2	0
2	0.25	200	2	0
3	0.5	200	2	0
4	0.25	100	2	0
5	0.25	500	2	0
6	0.25	200	1	0
7	0.25	200	2.5	0
8	0.25	200	3	0
9	0.25	200	4	0
10	0.25	200	5	0
11	0.25	200	3	1
12	0.25	200	3	2
13	0.25	200	3	3

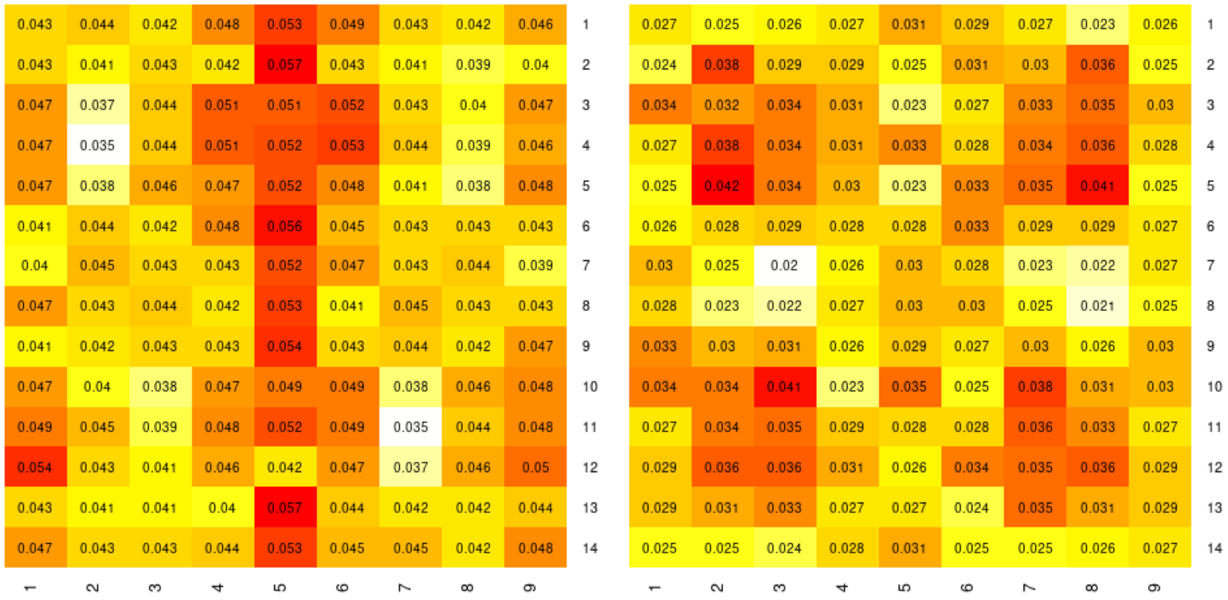
Table 3.6: A summary of the 13 settings for the simulated hippocampus surface data.

We mainly compared the BCov test and GBJ test on all 13 settings, and the rejection rates on each region are listed in heatmaps. Figure 3.1 compared the type I errors of BCov test and GBJ test, Figure 3.2 lists the rejection rates in settings 2-13 by BCov, and Figure 3.3 lists the rejection rates in settings 2-13 by GBJ. As can be seen from Figure 3.1, type I errors are well controlled for both BCov test and GBJ test, although GBJ may have a deflated type I error. Comparing Figure 3.2 and Figure 3.3, we can see that BCov tend to give a more accurate detected region.

We also compared BCov and GBJ with other testing methods, including AdaMant, RdCov, and sKPCR, on a setting with registration errors in terms of both location shift and size variation of the affected region. Those rfGWAS implementations are also compared with the voxelwise method using simple linear regression. The rejection rates of those methods are compared in heatmaps in Figure 3.4.

3.3.4 Global Screening

The SNP data is simulated using COSI (Schaffner et al. 2005). First, 9880 haplotypes are simulated using COSI for 100 1Mb regions. Then, the SNPs are ascertained/thinned to match HapMapII SNP frequency distribution. Those haplotypes are randomly combined in pairs



	1	2	3	4	5	6	7	8	9
1	0.043	0.044	0.042	0.048	0.053	0.049	0.043	0.042	0.046
2	0.043	0.041	0.043	0.042	0.057	0.043	0.041	0.039	0.04
3	0.047	0.037	0.044	0.051	0.051	0.052	0.043	0.04	0.047
4	0.047	0.035	0.044	0.051	0.052	0.053	0.044	0.039	0.046
5	0.047	0.038	0.046	0.047	0.052	0.048	0.041	0.038	0.048
6	0.041	0.044	0.042	0.048	0.056	0.045	0.043	0.043	0.043
7	0.04	0.045	0.043	0.043	0.052	0.047	0.043	0.044	0.039
8	0.047	0.043	0.044	0.042	0.053	0.041	0.045	0.043	0.043
9	0.041	0.042	0.043	0.043	0.054	0.043	0.044	0.042	0.047
10	0.047	0.04	0.038	0.047	0.049	0.049	0.038	0.046	0.048
11	0.049	0.045	0.039	0.048	0.052	0.049	0.035	0.044	0.048
12	0.054	0.043	0.041	0.046	0.042	0.047	0.037	0.046	0.05
13	0.043	0.041	0.041	0.04	0.057	0.044	0.042	0.042	0.044
14	0.047	0.043	0.043	0.044	0.053	0.045	0.045	0.042	0.048

	1	2	3	4	5	6	7	8	9
1	0.027	0.025	0.026	0.027	0.031	0.029	0.027	0.023	0.026
2	0.024	0.038	0.029	0.029	0.025	0.031	0.03	0.036	0.025
3	0.034	0.032	0.034	0.031	0.023	0.027	0.033	0.035	0.03
4	0.027	0.038	0.034	0.031	0.033	0.028	0.034	0.036	0.028
5	0.025	0.042	0.034	0.03	0.023	0.033	0.035	0.041	0.025
6	0.026	0.028	0.029	0.028	0.028	0.033	0.029	0.029	0.027
7	0.03	0.025	0.02	0.026	0.03	0.028	0.023	0.022	0.027
8	0.028	0.023	0.022	0.027	0.03	0.03	0.025	0.021	0.025
9	0.033	0.03	0.031	0.026	0.029	0.027	0.03	0.026	0.03
10	0.034	0.034	0.041	0.023	0.035	0.025	0.038	0.031	0.03
11	0.027	0.034	0.035	0.029	0.028	0.028	0.036	0.033	0.027
12	0.029	0.036	0.036	0.031	0.026	0.034	0.035	0.036	0.029
13	0.029	0.031	0.033	0.027	0.027	0.024	0.035	0.031	0.029
14	0.025	0.025	0.024	0.028	0.031	0.025	0.025	0.026	0.027

Figure 3.2: The type I errors using BCov (left) and GBJ (right).

to generate $9880/2 = 4940$ individuals and 1000 individuals. Then, in each cohort, SNPs are filtered for MAF>0.05. We end up with 77,603 SNPs for the cohort with sample size N=4940 and 77,100 SNPs with the cohort with sample size N=1000. Randomly select 100 causal SNPs from the two cohorts, respectively, and generate y based on their additive effect $y_i(d) = \sum_{g=1}^{100} x_i^{(g)} \beta^{(g)}(d) + \epsilon_i$ similar to the generating procedure in the third simulation setting. The SNP data X remains the same across the 100 repeats and all the testing scenarios. Use $a = 0.25$ as the effect size, which is the medium effect size in simulation setting 3 (0, 0.25, or 0.5) and about half of the dosage simulation effect size (0.4).

For the subsample-based screening, the subsample sizes for N=4940 are n=2000, 1000, and 500, and the subsample sizes for N=1000 are n=1000, 500, and 200. The inclusion set size from 100 to 2,000 (100, 300, 600, 900, 1200, 1500, 1800, 2000 as in FGWAS).

For the 100 causal SNPs ($g = 1, \dots, 100$), simulate the affected regions as in section ?? (add section number for simulation 3). In particular, denote the affected region of marker g in individual i as $A_i^{(g)}$, which is a round area with center $(x_i^{(g)}, y_i^{(g)})$ and radius $r_i^{(g)}$, where $x_i^{(g)} \sim N(x_0^{(g)}, \sigma_x^2)$, $y_i^{(g)} \sim N(y_0^{(g)}, \sigma_y^2)$, and $r_i^{(g)} \sim N(10, \sigma_r^2)$. Use the values of the parameters

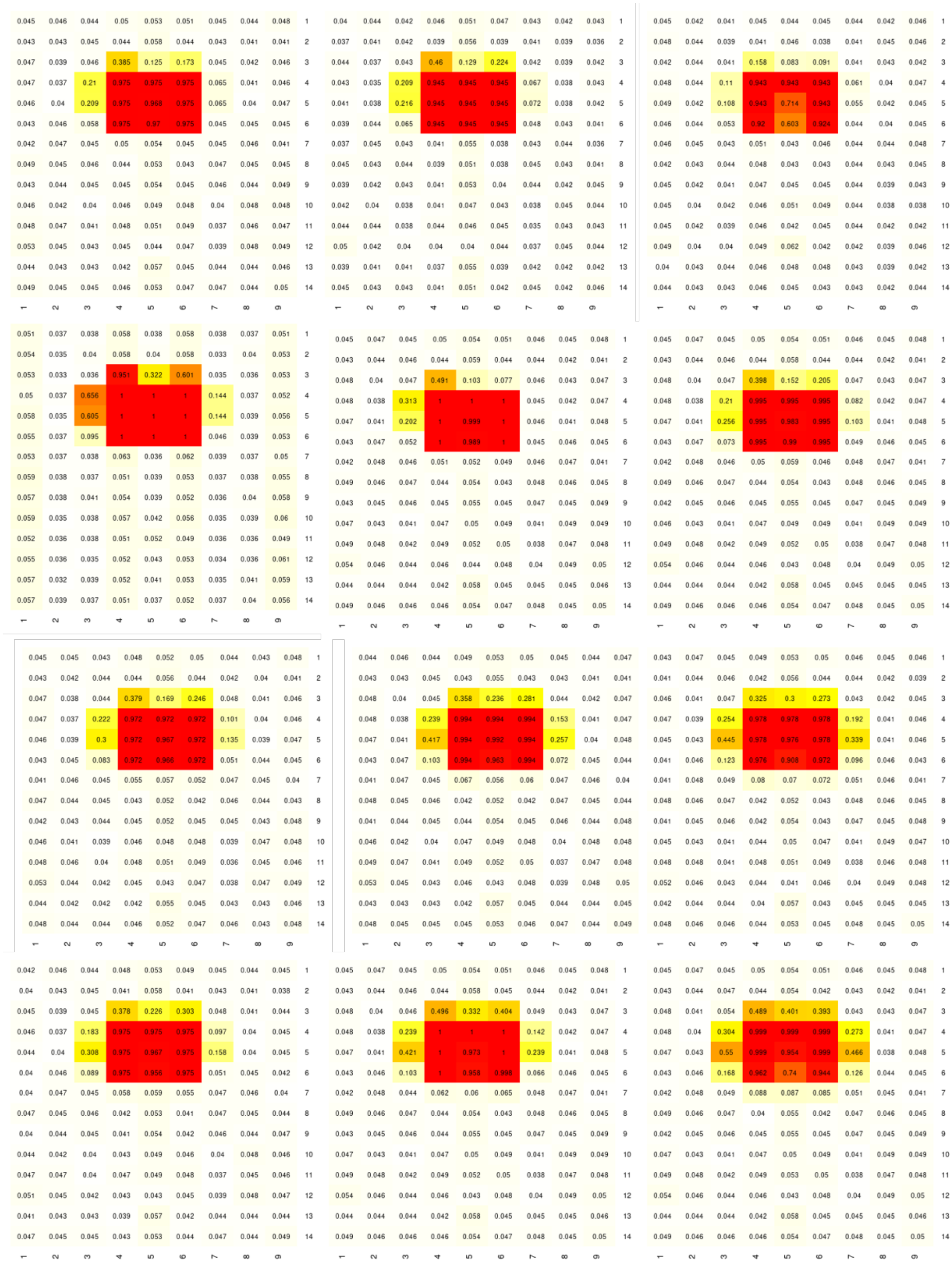


Figure 3.3: Rejection rates using BCov for settings 2-13 (plots ordered by row).

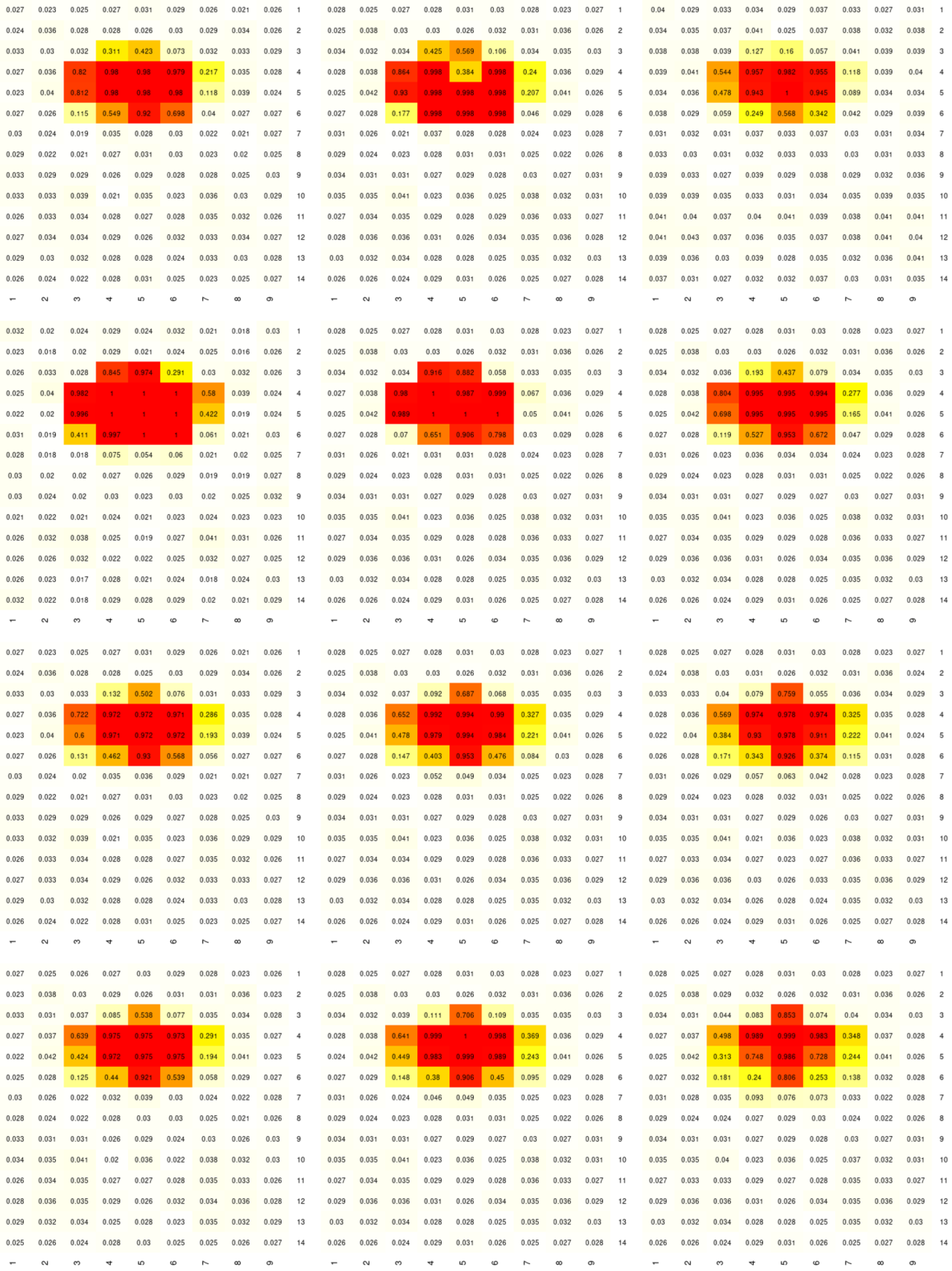
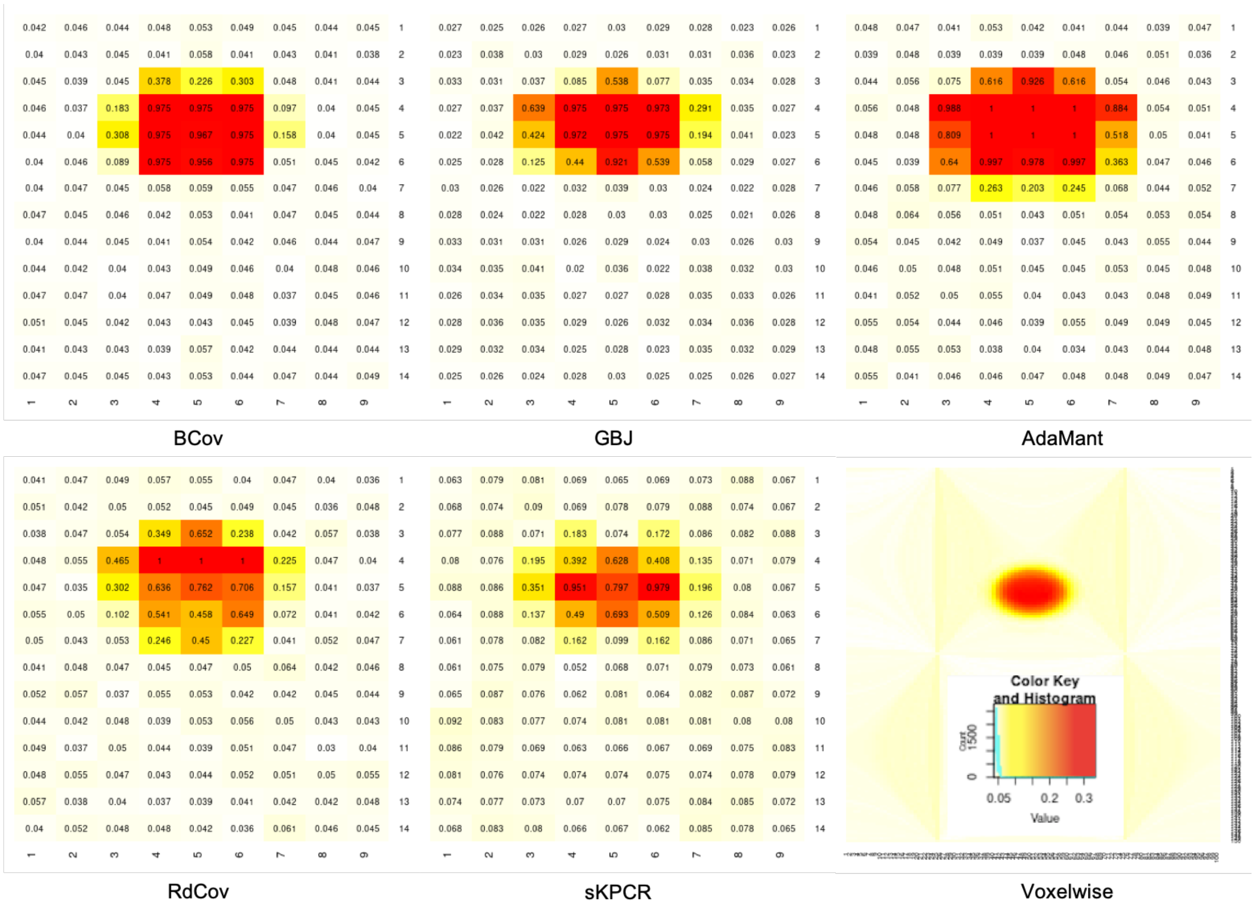


Figure 3.4: Rejection rates using GBJ for settings 2-13 (plots ordered by row).



BCov

1	2	3	4	5	6	7	8	9	
0.042	0.046	0.044	0.048	0.053	0.049	0.045	0.044	0.045	1
0.04	0.043	0.045	0.041	0.058	0.041	0.043	0.041	0.038	2
0.045	0.039	0.045	0.378	0.226	0.303	0.048	0.041	0.044	3
0.046	0.037	0.183	0.975	0.975	0.975	0.097	0.04	0.045	4
0.044	0.04	0.308	0.975	0.967	0.975	0.158	0.04	0.045	5
0.04	0.046	0.089	0.975	0.956	0.975	0.051	0.045	0.042	6
0.04	0.047	0.045	0.058	0.059	0.055	0.047	0.046	0.04	7
0.047	0.045	0.046	0.042	0.053	0.041	0.047	0.045	0.044	8
0.04	0.044	0.045	0.041	0.054	0.042	0.046	0.044	0.047	9
0.044	0.042	0.04	0.043	0.049	0.046	0.04	0.048	0.046	10
0.047	0.047	0.04	0.047	0.049	0.048	0.037	0.045	0.046	11
0.051	0.045	0.042	0.043	0.043	0.045	0.039	0.048	0.047	12
0.041	0.043	0.043	0.039	0.057	0.042	0.044	0.044	0.044	13
0.047	0.045	0.045	0.043	0.053	0.044	0.047	0.044	0.049	14

GBJ

1	2	3	4	5	6	7	8	9	
0.027	0.025	0.026	0.027	0.03	0.029	0.028	0.023	0.026	1
0.023	0.038	0.03	0.029	0.026	0.031	0.031	0.036	0.023	2
0.033	0.031	0.037	0.085	0.538	0.077	0.035	0.034	0.028	3
0.027	0.037	0.639	0.975	0.975	0.973	0.291	0.035	0.027	4
0.022	0.042	0.424	0.972	0.975	0.975	0.194	0.041	0.023	5
0.025	0.028	0.125	0.44	0.921	0.539	0.058	0.029	0.027	6
0.03	0.026	0.022	0.032	0.039	0.03	0.024	0.022	0.028	7
0.028	0.024	0.022	0.028	0.03	0.03	0.025	0.021	0.026	8
0.033	0.031	0.031	0.026	0.029	0.024	0.03	0.026	0.03	9
0.034	0.035	0.041	0.02	0.036	0.022	0.038	0.032	0.03	10
0.026	0.034	0.035	0.027	0.027	0.028	0.035	0.033	0.026	11
0.028	0.036	0.035	0.029	0.026	0.032	0.034	0.036	0.028	12
0.029	0.032	0.034	0.025	0.028	0.023	0.035	0.032	0.029	13
0.025	0.026	0.024	0.028	0.03	0.025	0.025	0.026	0.027	14

AdaMant

1	2	3	4	5	6	7	8	9	
0.048	0.047	0.041	0.053	0.042	0.041	0.044	0.039	0.047	1
0.039	0.048	0.039	0.039	0.039	0.048	0.046	0.051	0.036	2
0.044	0.056	0.075	0.616	0.926	0.616	0.054	0.046	0.043	3
0.056	0.048	0.988	1	1	1	0.884	0.054	0.051	4
0.048	0.048	0.809	1	1	1	0.518	0.05	0.041	5
0.045	0.039	0.64	0.997	0.978	0.997	0.363	0.047	0.046	6
0.046	0.058	0.077	0.263	0.203	0.245	0.068	0.044	0.052	7
0.048	0.064	0.056	0.051	0.043	0.051	0.054	0.053	0.054	8
0.054	0.045	0.042	0.049	0.037	0.045	0.043	0.055	0.044	9
0.046	0.05	0.048	0.051	0.045	0.045	0.053	0.045	0.048	10
0.041	0.052	0.05	0.055	0.04	0.043	0.043	0.048	0.049	11
0.055	0.054	0.044	0.046	0.039	0.055	0.049	0.049	0.045	12
0.048	0.055	0.053	0.038	0.04	0.034	0.043	0.044	0.048	13
0.055	0.041	0.046	0.046	0.047	0.048	0.048	0.049	0.047	14

RdCov

1	2	3	4	5	6	7	8	9	
0.041	0.047	0.049	0.057	0.055	0.04	0.047	0.04	0.036	1
0.051	0.042	0.05	0.052	0.045	0.049	0.045	0.036	0.048	2
0.038	0.047	0.054	0.349	0.652	0.238	0.042	0.057	0.038	3
0.048	0.055	0.465	1	1	1	0.225	0.047	0.04	4
0.047	0.035	0.302	0.636	0.762	0.706	0.157	0.041	0.037	5
0.055	0.05	0.102	0.541	0.458	0.649	0.072	0.041	0.042	6
0.05	0.043	0.053	0.246	0.45	0.227	0.041	0.052	0.047	7
0.041	0.048	0.047	0.045	0.047	0.05	0.064	0.042	0.046	8
0.052	0.057	0.037	0.055	0.053	0.042	0.042	0.045	0.044	9
0.044	0.042	0.048	0.039	0.053	0.056	0.05	0.043	0.043	10
0.049	0.037	0.05	0.044	0.039	0.051	0.047	0.03	0.04	11
0.048	0.055	0.047	0.043	0.044	0.052	0.051	0.05	0.055	12
0.057	0.038	0.04	0.037	0.039	0.041	0.042	0.042	0.048	13
0.04	0.052	0.048	0.048	0.042	0.036	0.061	0.046	0.045	14

sKPCR

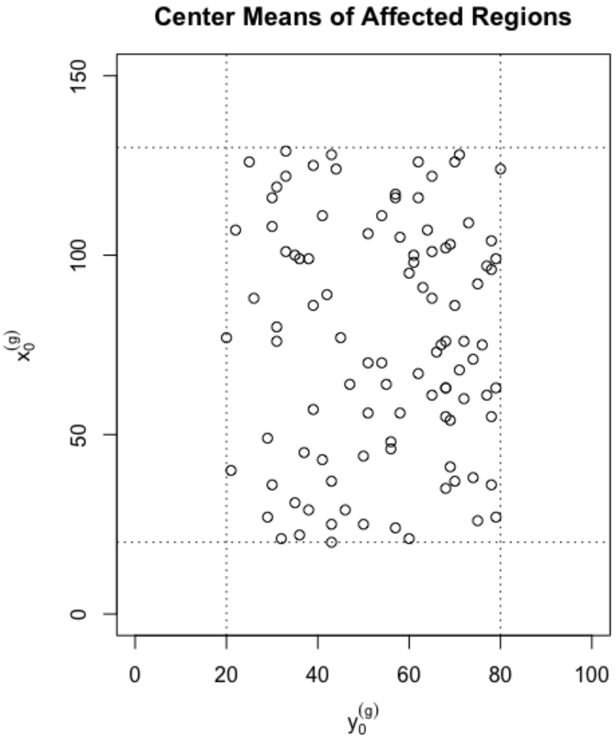
1	2	3	4	5	6	7	8	9	
0.063	0.079	0.081	0.069	0.065	0.069	0.073	0.088	0.067	1
0.068	0.074	0.09	0.069	0.078	0.079	0.088	0.074	0.067	2
0.077	0.088	0.071	0.183	0.074	0.172	0.086	0.082	0.088	3
0.08	0.076	0.195	0.392	0.628	0.408	0.135	0.071	0.079	4
0.088	0.086	0.351	0.951	0.797	0.979	0.196	0.08	0.067	5
0.064	0.088	0.137	0.49	0.693	0.509	0.126	0.084	0.063	6
0.061	0.078	0.082	0.162	0.099	0.162	0.086	0.071	0.065	7
0.061	0.075	0.079	0.052	0.068	0.071	0.079	0.073	0.061	8
0.065	0.087	0.076	0.062	0.081	0.064	0.082	0.087	0.072	9
0.092	0.083	0.077	0.074	0.081	0.081	0.081	0.08	0.08	10
0.086	0.079	0.069	0.063	0.066	0.067	0.069	0.075	0.083	11
0.081	0.076	0.074	0.074	0.074	0.075	0.074	0.078	0.079	12
0.074	0.077	0.073	0.07	0.07	0.075	0.084	0.085	0.072	13
0.068	0.083	0.08	0.066	0.067	0.062	0.085	0.078	0.065	14

Figure 3.5: Comparison of the rejection rates using different methods to analyze simulated hippocampus surface data under setting 11.

as in setting 12 of simulation 3, i.e. effect size $a = 0.25$, $\sigma_x^2 = \sigma_y^2 = 3$, and $\sigma_r^2 = 2$. Allow $(x_0^{(g)}, y_0^{(g)})$ to be different across $g = 1, \ldots, 100$ so that marker effects do not lay over the same region, which could add a layer of detection difficulty; different $(x_0^{(g)}, y_0^{(g)})$'s also mimic the real scenario that different markers affect different regions. The entire image is simulated by a 150×100 rectangle. If $A_i^{(g)}$ touches the boarder, let $A_i^{(g)}$ be the part that is within the image. If $(x_0^{(g)}, y_0^{(g)})$ is near the edge of the 150×100 image s.t. $A_i^{(g)}$ crosses the boarder, the resulting size of $A_i^{(g)}$ would shrink and the effect of marker g is thus different from other markers due to the location of $(x_0^{(g)}, y_0^{(g)})$, and thus resulting an unfair comparison with other markers in the global screening. Therefore, restrict $(x_0^{(g)}, y_0^{(g)})$ to stay away from the image boarder s.t. given the variation in $(x_i^{(g)}, y_i^{(g)})$ and $r_i^{(g)}$, $A_i^{(g)}$ would still largely remain in the 150×100 image. Generally, we expect only few realizations of a normal random variable to be more than 2 times standard errors away from its mean. So the most distant points in $A_i^{(g)}$ should stay within $2\sigma_x + (10 + 2\sigma_r) = 2\sigma_y + (10 + 2\sigma_r) \approx 16.3$ to $(x_0^{(g)}, y_0^{(g)})$. Extend this "safe range" to 20 for ease of specification and for a safer range. Therefore, restrict $x_0^{(g)} \in [20, 130]$ and $y_0^{(g)} \in [20, 80]$. Randomly scatter $(x_0^{(g)}, y_0^{(g)})$ in $[20, 130] \times [20, 80]$ for $g = 1, \ldots, 100$ (Figure "Center Means of Affected Regions").

Use 100 Monte Carlo realizations to calculate the average causal SNP inclusion rate for each setting, and summarize the result in Figure 3.6, where the causal SNP inclusion rate is calculated as the ratio of number of causal SNPs included in the top G_0 SNPs over the total number of 100 causal SNPs.

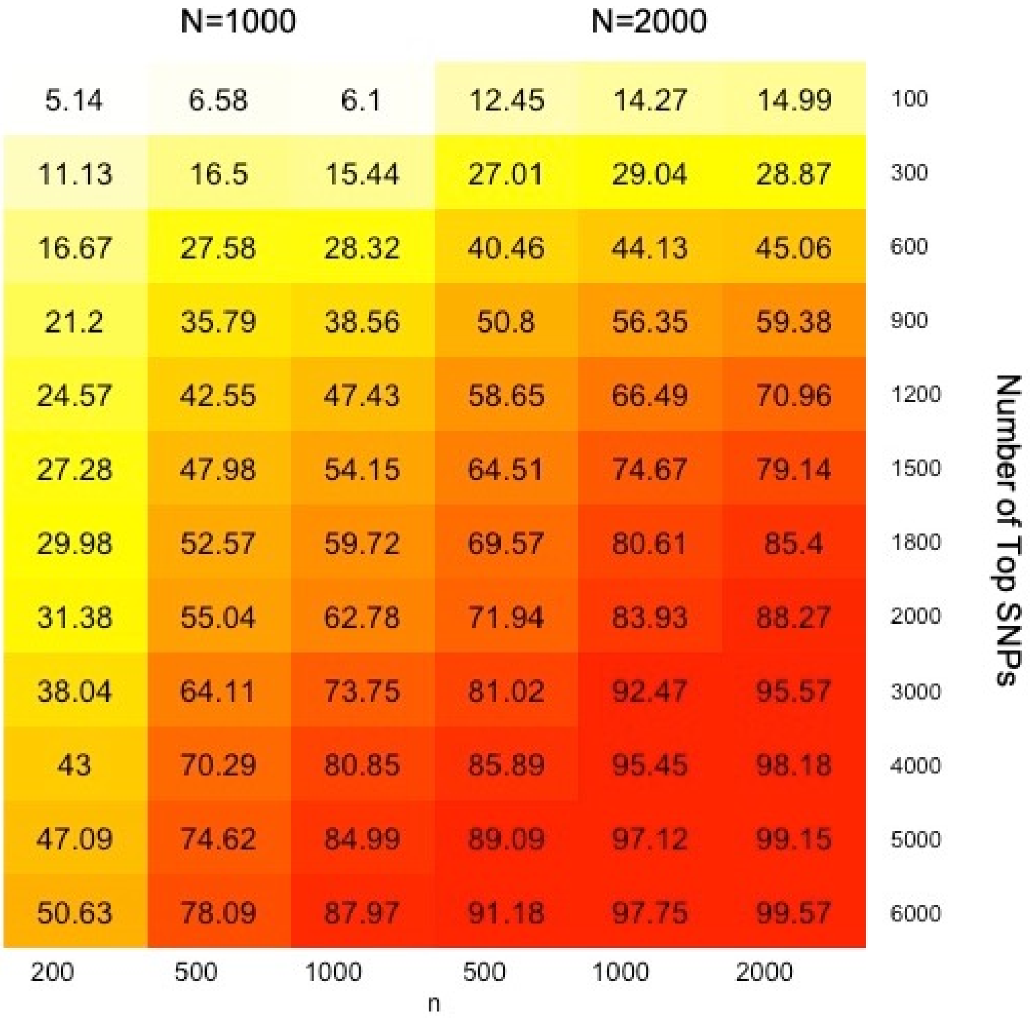
From Figure 3.6, we can see that for the same total sample size N, the inclusion rate increases with the subsample size n and the number of top SNPs included for screening. Also, for the same subsample size n, the inclusion rate increases as the total sample size N increases. This is because when the total sample size increases, the number of subjects with alternative alleles also increases. These subjects are likely to be included into the subsample and create a larger effective sample size for the evaluation. Therefore, a true signal becomes more likely to be detected in such cases.



3.4 Real Data Application

The brain imaging genetic data used in this analysis is from ADNI. We used the hippocampus surface data as described in (Huang, 2017). Number of SNPs in this analysis is 503,892. In this analysis, we included 730 individuals. In either the left or right hippocampus surface data, each individual has 15,000 data points, which was covered by 126 overlapping regions in the same manner as described in section 3.3.3 in the simulation. The p-value threshold under Bonferroni correction is 7.875e-10 $= 0.05/126/503{,}893$. The analysis using rfGWAS is the same as described in section 3.3.3, with BCov test used in the association testing step. We discovered 47 signals as shown in Tables 3.7-3.8, with p-values sorted from smallest to largest.

Figure 3.7 displays the density curves of region 47 on the left hipopcampus surface by dosage levels (0, 1, or 2, representing genotypes CC, TC or CT, and TT, respectively) of SNP rs2736372 (at chromosome 8, position 11143451) with 30 samples randomly selected for each dosage level. This region-SNP combination is selected because it is the most significant



	N=1000			N=2000		
Number of Top SNPs	200	500	1000	500	1000	2000
100	5.14	6.58	6.1	12.45	14.27	14.99
300	11.13	16.5	15.44	27.01	29.04	28.87
600	16.67	27.58	28.32	40.46	44.13	45.06
900	21.2	35.79	38.56	50.8	56.35	59.38
1200	24.57	42.55	47.43	58.65	66.49	70.96
1500	27.28	47.98	54.15	64.51	74.67	79.14
1800	29.98	52.57	59.72	69.57	80.61	85.4
2000	31.38	55.04	62.78	71.94	83.93	88.27
3000	38.04	64.11	73.75	81.02	92.47	95.57
4000	43	70.29	80.85	85.89	95.45	98.18
5000	47.09	74.62	84.99	89.09	97.12	99.15
6000	50.63	78.09	87.97	91.18	97.75	99.57

Figure 3.6: Inclusion rates

Chr	SNP	Pos	Reg.	P-value
8	rs2736372	11143451	47	3.78E-13
2	rs2218495	193450323	77	7.09E-13
3	rs7625888	55886177	3	3.00E-12
3	rs212016	59956565	68	5.49E-12
5	rs6875999	95496745	36	1.45E-11
13	rs1361576	103206029	117	1.55E-11
1	rs6663218	94176826	82	1.84E-11
9	rs9407756	16321304	18	3.05E-11
9	rs10780800	88699264	55	3.35E-11
9	rs10780800	88699264	66	5.05E-11
15	rs11629621	75867792	54	5.48E-11
20	rs1077577	56520675	37	5.90E-11
1	rs7525764	50617284	114	9.41E-11
1	rs6424278	231284033	29	1.16E-10
11	rs1901843	5922991	100	1.18E-10
8	rs958648	11141305	47	1.26E-10
4	rs1902859	81376727	49	1.36E-10
16	rs4580141	8555210	87	1.56E-10
4	rs13152105	77962544	122	1.62E-10
5	rs16771	103605382	51	1.87E-10
11	rs10790680	123820969	62	2.10E-10
3	rs1320288	21777300	45	2.17E-10
11	rs12789966	98547209	124	2.22E-10
4	rs10029313	40044904	113	2.57E-10

Table 3.7: 47 significant SNPs identified by rfGWAS on the left hippocampus surface (a).

Chr	SNP	Pos	Reg.	P-value
15	rs1869258	41590913	124	2.78E-10
14	rs6572493	22042795	118	2.80E-10
6	rs9457760	157889784	90	2.87E-10
8	rs4840969	8385081	23	2.99E-10
8	rs12676771	126603006	108	3.11E-10
4	rs1365488	100889738	47	3.17E-10
6	rs16900289	157926844	26	3.22E-10
8	rs10505101	108353121	5	3.25E-10
20	rs4812716	41569181	59	3.43E-10
15	rs11852746	53771557	76	3.68E-10
9	rs10961291	13866562	45	3.90E-10
2	rs1975583	200934147	85	4.16E-10
10	rs183827	83098389	88	4.20E-10
17	rs181246	53561087	85	4.39E-10
8	rs7816304	142558176	65	4.41E-10
15	rs2118157	32769236	8	4.86E-10
21	rs233232	44792782	75	5.35E-10
6	rs17053280	128837580	77	5.35E-10
17	rs202581	10732458	77	5.82E-10
12	rs12312219	50837457	105	7.07E-10
18	rs163221	22763991	67	7.37E-10
10	rs11257127	11570878	66	7.47E-10
4	rs13101823	26784600	89	7.87E-10

Table 3.8: 47 significant SNPs identified by rfGWAS on the left hippocampus surface (b).

association detected as can be seen from Table 3.6.

73 signals are discovered for the right hippocampus surface as listed in Tables 3.9-3.10.

3.4.1 Comparison with FGWAS

We then applied FGWAS (Huang et al., 2017) to the same data, i.e. the same set of SNPs on the hippocampus surface of the same imaging modality (radius) from ADNI with the same set of subjects and same set of covariates adjusted.

For the top SNP for each side from our method, extract and plot their local signals of adjusted p-values, as well as the signal pattern from our method as comparison (Figure 3.8).

We also plotted the top SNP for each side from FGWAS in Figure 3.9.

A detailed description of the FGWAS result can be found in Appendix 2.

3.5 Discussion

In this chapter, a regional functional GWAS (rfGWAS) framework is developed for imaging-genetic data to detect associations between genetic markers and brain regions, while taking into consideration registration errors across individual images. This purpose is accomplished by sliding a window across the entire brain and examining the strength of association between genetic markers and the brain region at each location of the window. The regions are overlapped and cover the entire brain, each summarized by a selected set of quantiles from the smoothed density curve of the distribution of the voxelwise values in the region. The association between each target region and the genetic markers is examined in a GWAS manner, where the genetic markers can either be examined one by one, or in groups, depending on the association detection method of choice. One can select methods from the pool of both linear and nonlinear method to examine the association strength between genetic markers and the region of interest. The method we propose is the ball covariance (Pan et al., 2019), which can capture nonlinear associations between two arbitrarily-dimensional variables, with good power and type I error control. We then examined our method by using simulation studies and applied it to the hippocampus surface radial distance data from the ADNI cohort. Significant SNP-region pairs are identified and reported, and the results from our method are

Chr	SNP	Pos	Reg.	P-value
6	rs9381225	42987267	38	9.95E-14
11	rs4603268	95454068	68	1.75E-13
11	rs7944077	11488123	65	2.77E-13
21	rs2834215	33718756	98	1.55E-12
15	rs2036349	95944662	84	2.88E-12
22	rs878734	41993678	100	3.06E-12
13	rs719103	49894661	76	8.60E-12
14	rs9323434	62777402	80	8.87E-12
8	rs406800	88604177	52	1.07E-11
11	rs7108440	1630212	64	1.63E-11
8	rs445448	88604096	52	2.15E-11
22	rs739079	47977447	100	2.51E-11
14	rs10150375	62761025	80	3.48E-11
21	rs1059293	33731563	98	3.62E-11
11	rs7109505	36245328	66	4.49E-11
13	rs9568431	49953648	76	4.69E-11
4	rs1382658	164256751	29	4.73E-11
3	rs4077972	72079519	20	4.88E-11
19	rs4802867	56991697	95	5.07E-11
1	rs12132073	110965216	4	5.21E-11
1	rs1055565	209672320	7	5.56E-11
1	rs2131562	110968762	4	6.36E-11
19	rs17305227	58986294	95	6.96E-11
12	rs4764764	99787007	73	7.09E-11
21	rs1044213	33743561	98	7.89E-11
5	rs17088025	97451359	33	8.16E-11
11	rs7107470	11506226	65	8.85E-11
22	rs4821512	35343640	100	8.98E-11
11	rs12293070	131668714	69	9.94E-11
13	rs1924373	49843707	76	1.12E-10
2	rs7597214	109311672	12	1.14E-10
12	rs7313843	15102438	70	1.20E-10
9	rs2477518	27589746	56	1.23E-10
5	rs12188891	163535330	35	1.42E-10
12	rs10860647	99766229	73	1.44E-10
22	rs4821513	35343765	100	1.46E-10
3	rs7643459	7979828	17	1.51E-10

Table 3.9: 73 significant SNPs identified by rfGWAS on the right hippocampus surface (a).

Chr	SNP	Pos	Reg.	P-value
3	rs9808938	61524608	19	1.57E-10
4	rs6449435	18517616	24	1.83E-10
13	rs9568430	49950822	76	1.91E-10
4	rs2314137	162485246	29	2.06E-10
11	rs522616	102220258	68	2.07E-10
8	rs391916	88581402	52	2.17E-10
11	rs7944077	11488123	65	2.27E-10
1	rs497870	213178768	7	2.38E-10
8	rs12386970	22741098	50	2.69E-10
5	rs6555554	8611215	30	2.81E-10
16	rs251108	67608709	86	3.00E-10
3	rs2695642	22105396	18	3.17E-10
8	rs452325	88574482	52	3.21E-10
11	rs1944936	74096811	67	3.28E-10
14	rs12435524	21204085	78	3.43E-10
6	rs1132643	123088715	41	3.73E-10
10	rs2084882	33905831	61	3.83E-10
20	rs1927333	20206395	96	3.89E-10
8	rs12676238	117640909	53	3.90E-10
14	rs6571976	21182434	78	4.00E-10
17	rs16848	73174851	90	4.23E-10
10	rs6481857	33903912	61	4.44E-10
2	rs999890	208898877	15	4.87E-10
2	rs13407268	208922184	15	4.87E-10
2	rs10177810	208926434	15	4.87E-10
4	rs1154968	66956730	26	5.37E-10
14	rs1997916	62821120	80	5.78E-10
11	rs10500662	6734213	65	5.86E-10
13	rs9529822	70394210	76	6.02E-10
3	rs6789391	1827696	17	6.61E-10
13	rs9542535	70391150	76	6.65E-10
9	rs10733310	16454980	55	6.89E-10
2	rs10190458	208919295	15	7.20E-10
4	rs10034869	167770230	29	7.33E-10
2	rs1077583	6868481	9	7.54E-10
1	rs10888541	151476212	5	7.63E-10

Table 3.10: 73 significant SNPs identified by rfGWAS on the right hippocampus surface (b).

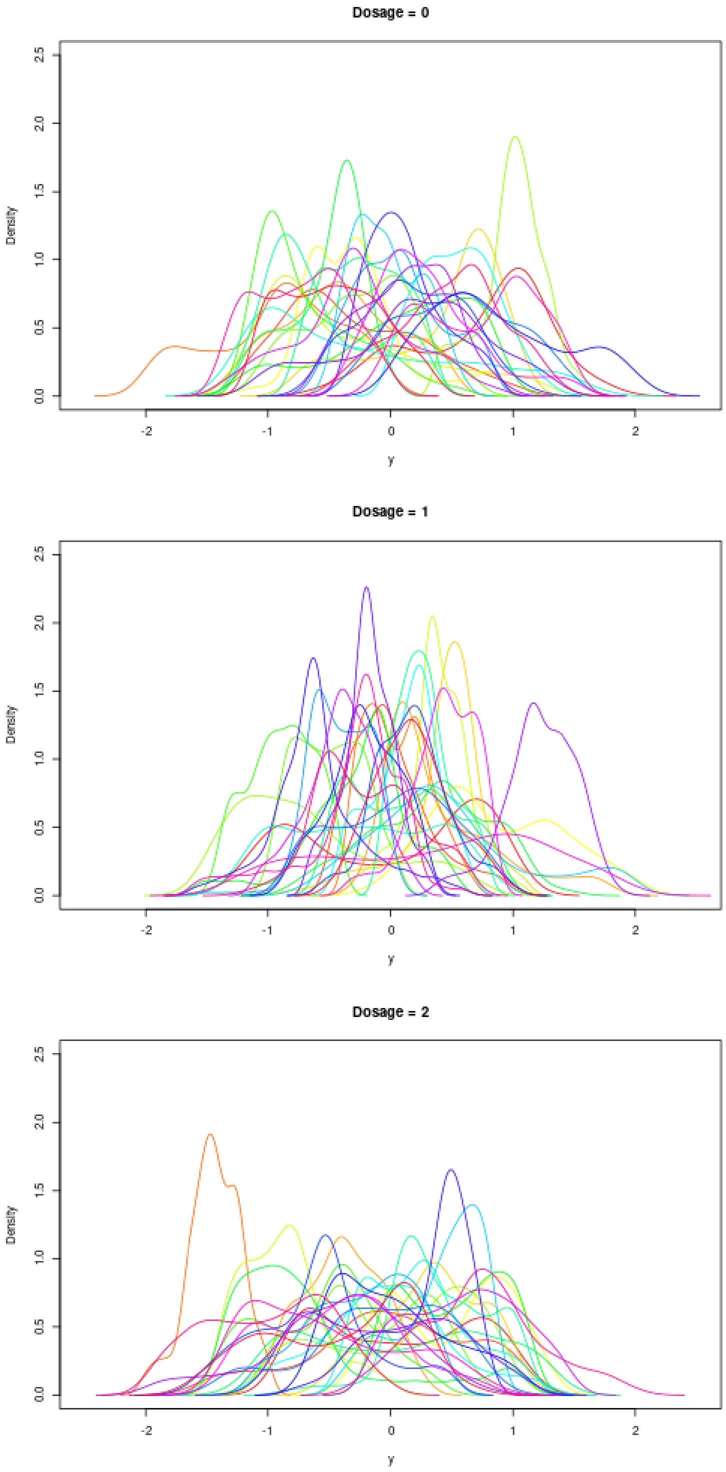


Figure 3.7: Density curves of region 47 by dosage levels (0, 1, or 2, representing genotypes CC, TC or CT, and TT, respectively) of SNP rs2736372 (at chromosome 8, position 11143451) with 30 samples randomly selected for each dosage level.

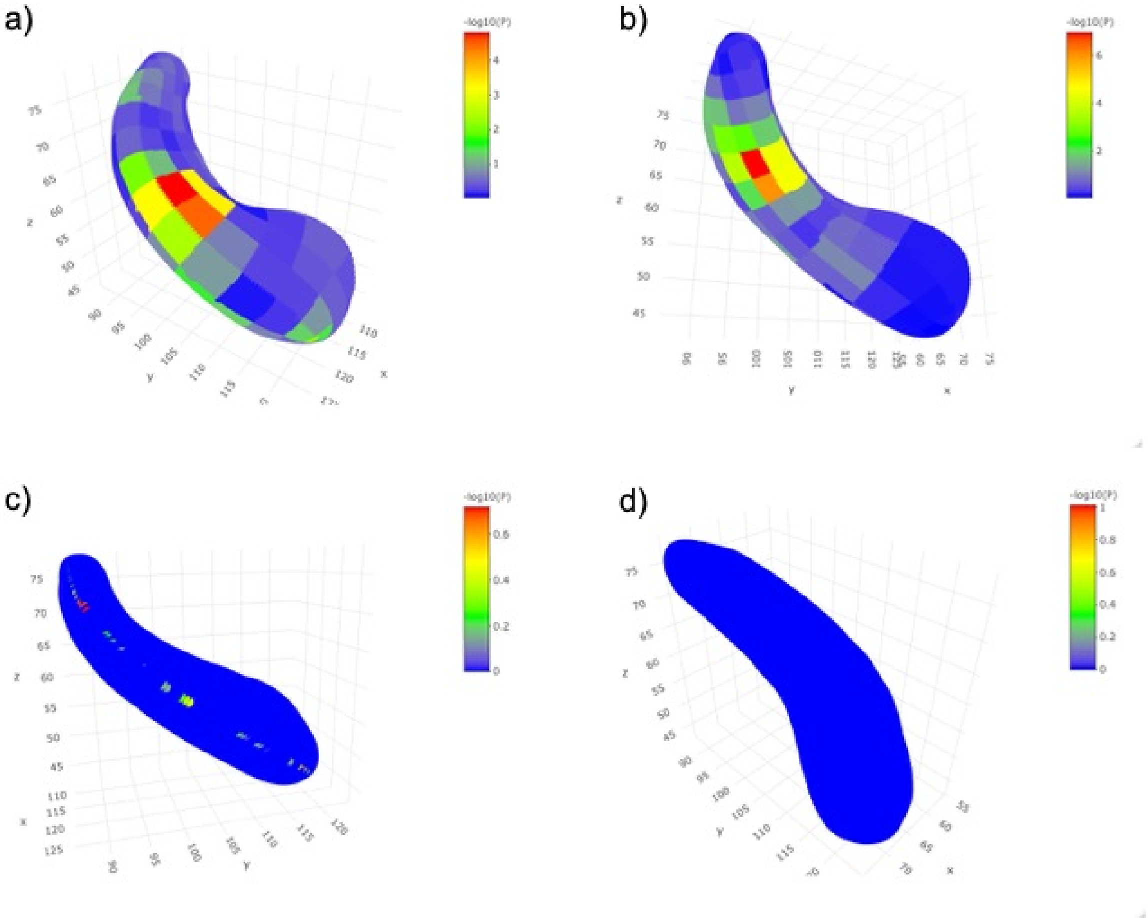


Figure 3.8: Comparison with FGWAS local testing results for the top SNP on each side of hippocampus from our result: a) Left hippocampus surface heatmap of $-\log_{10}$(p-value) for SNP rs2736372 (chr8:1,114,345) from our BCov-based method; b) right hippocampus surface heatmap of $-\log_{10}$(p-value) for SNP rs9381225 (chr6:4,298,726) from our BCov-based method; Voxelwise p-values at each voxel is the average of its corresponding p-values in all subregions that cover this voxel. c) left hippocampus surface heatmap of $-\log_{10}$(p-value) for SNP rs2736372 (chr8:1,114,345) from FGWAS (sprinkled highlights on the cutting line due to lack of smoothing on the edge); and d) right hippocampus surface heatmap of $-\log_{10}$(p-value) for SNP rs9381225 (chr6:4,298,726) from FGWAS (all blue). The hippocampus in the plots are rotated so that highlighted regions can be visible. The coordinates are the same across plots from our method and FGWAS.

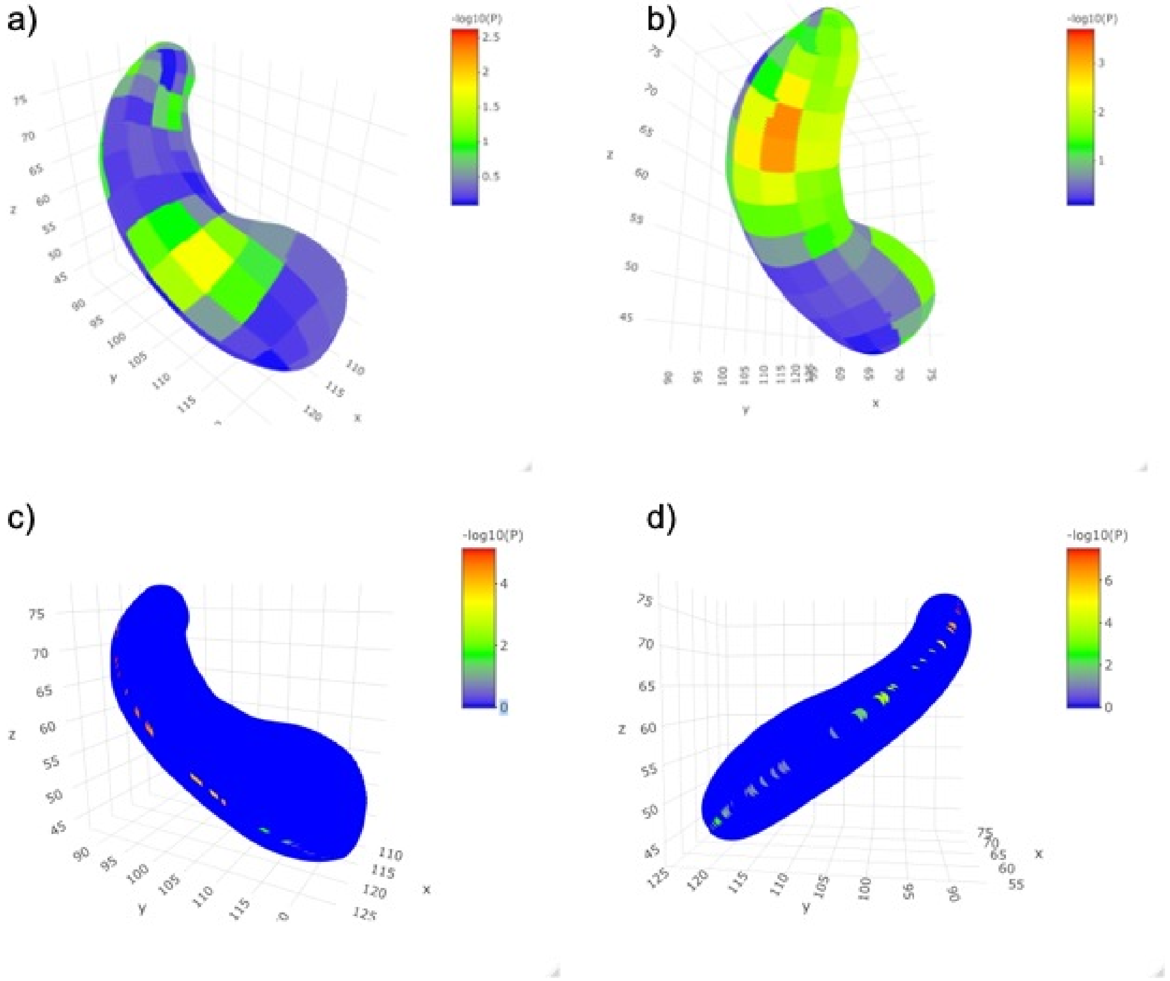


Figure 3.9: Comparison with FGWAS local testing results for the top SNP on each side of hippocampus from our result: a) Left hippocampus surface heatmap of $-\log_{10}$(p-value) for SNP rs592629 (chr18:74,186,483) from our BCov-based method; b) right hippocampus surface heatmap of $-\log_{10}$(p-value) for SNP rs4681527 (chr3:145,483,129) from our BCov-based method; Voxelwise p-values at each voxel is the average of its corresponding p-values in all subregions that cover this voxel. c) left hippocampus surface heatmap of $-\log_{10}$(p-value) for SNP rs592629 (chr18:74,186,483) from FGWAS (almost all p-value equal to 0, of which the $-\log_{10}$(p-value) were set to 165); and d) right hippocampus surface heatmap of $-\log_{10}$(p-value) for SNP rs4681527 (chr3:145,483,129) from FGWAS (almost all p-value equal to 0, of which the $-\log_{10}$(p-value) were set to 165). The hippocampus in the plots are rotated so that highlighted regions can be visible. The coordinates are the same across plots from our method and FGWAS.

compared with that from FGWAS (Huang et al., 2017).

The regional analysis strategy not only reduces computation cost, but also detects regional signal allowing registration error in the data. These two aspects advance the voxelwise methods, such as FGWAS (Huang et al., 2017). rfGWAS is a flexible framework that can incorporate both linear and nonlinear association testing methods. The association test can accommodate either a single-marker test or a marker-set test, based on the statistical test selected. The user-defined window size gives another layer of flexibility, as well as space for future research. One draw back of using the Ball Covariance test as compared to the linear-regression-based test is that we are not informed of how the genetic markers influence target regions, e.g. the effect sizes. The ball covariance only indicates the strength of the association. Covariates can be adjusted by regressing them out on the original voxelwise data or the transformed density curves. In the real-data analysis for ADNI, we regressed out the covariates on the voxelwise data so that the result is comparable with that from FGWAS (Huang et al., 2017).

The simulation study took a reasonable amount of computation time. For simulation implemented in Section 3.3.1, the computation time is summarized in Table 3.11. As for the parametric BCov methods, it took 3 hours to analyze 20k simulated datasets with sample sizes 200,300,500,700, and 1000, which is 540 seconds / 1k runs, or 0.54 seconds per run. For the simulation implemented in section 3.3.2, it took 33.5 minutes to run the 16k simulated datasets using all four methods (BCov, RdCov, AdaMant, and simple linear regression). Computation time for each of the 1,000 simulated datasets on all 126 regions implemented in section 3.3.3 is summarized in Table 3.12.

In future research, theoretical discussion on registration error can be done, including forming the theoretical assumption that describes the registration error and how the method in our proposed framework can improve power in capturing signals with registration error in the given data. The size of the sliding region can also be discussed.

	BCov	BCov_{bs}	GBJ	AdaMant	RdCov	sKPRC	lm
Time (minutes)	6.7	21	NA	2.4	68	10	31
# simulated datasets	40k	40k	24k	16k	40k	24k	16k
Seconds / 1k runs	10.05	31.5	NA	9	102	25	116.25

Table 3.11: Computation time (in minutes) for all simulated datasets (n=30,50,70,100), for simulation implemented in section 3.3.1. BCov_{bs} is the BCov test with p-values calculated using bootstrap. lm represents the voxelwise method with simple linear regression model applied on each voxel.

	BCov	GBJ	AdaMant	RdCov	sKPRC	Voxelwise (lm)
Time (s)	8~9	30~40	3~4	180~330	4.2	18~47

Table 3.12: Computation time (in seconds) for each of the 1,000 replicates on all 126 regions as implemented in section 3.3.3.

CHAPTER 4: FPLS WITH DISTANCE CORRELATION

4.1 Introduction

Alzheimer's Disease (AD) is one of the most common dementia that affects memory, cognitive abilities and behavior. The AD patient gradually loses the body functions and ultimately goes to death, which withdraws from family and society. Thus, the Alzheimer's Disease Neuroimaging Initiative (ADNI) aims to improve clinical trials for the prevention and treatment of AD. The ADNI was launched in 2003 by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, Food and Drug Administration, private pharmaceutical companies and non-profit organizations as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological biomarkers can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessening the time and cost of clinical trials.

As of 2012, Apolipoprotein E4 (apoE4) was found to be the most prevalent genetic risk factor of Alzheimer's Disease in a variety of ethnic group (Sadigh-Eteghad et al., 2012). Thus, it is great interest to examine the dependence between the brain structure and the gene. The challenge is that the high dimensionality of imaging data in detecting the association. A functional linear model (FLM) and its variations as the popular prediction models based on functional predictor have gained extensive attention recently. Many estimation methods have been developed to estimate the coefficient function of FLM. The most common method is the functional principal component (FPCA). As an alternative to FPCA, functional partial least squares (FPLS) was introduced to estimate the coefficient function (Preda et al., 2007). FPLS can explicitly incorporates the information from response thus achieving better prediction accuracy. However, computational and theoretical complexity of FPLS make it difficult to implement. Thus, alternative partial least squares (APLS) was proposed by Delaigle and Hall (2012) to overcome this difficulty. The APLS algorithm gives an explicit formulae to estimate the basis function which can reduce the intensity of computation.

Inspired by the satisfactory performance of APLS , we consider the distance correlation, first proposed by Székely et al. (2007), to extend the APLS algorithm. Compared to the Pearson's correlation, distance correlation has three major advantages. First, distance correlation is zero if and only if two random vectors are independent. Second, distance correlation can deal with random vectors of arbitrary dimension. Third, distance correlation does not rely any model assumption, thus it is robust to the model misspecification. In this chapter, we generalise the APLS algorithm and propose a Non-parametric functional partial least squares through distance correlation (FPLS-DC). The major contributions of this chapter are threefold: (a) We extend APLS to a complex and general FPLS-DC framework; (b) we develop a computationally efficient algorithm based on Sequential Quadratic Programming to estimate the coefficients; (c) we show by the simulation and real data analysis that FPLS-DC can yield more accurate and robust results than APLS in different scenarios, especially the non-linear relationship.

The rest of the chapter is organized as follows. Section 2 introduces the APLS algorithm and FPLS-DC algorithm. Section 3 presents the Monte Carlo simulation results. Section 4 provides a through real data analysis of ADNI. Section 5 discussed the extensions and possible future research.

4.2 Methods

Let $Y(\cdot)$ be a random functional data defined on compact space $\mathcal{S} \subset \mathbb{R}^K$, where K is a positive integer, and X is a scalar variable of interest. For $i = 1, \dots, n$, the independent observation $\{(x_i, \boldsymbol{y}_i){:}1, \dots, n\}$ is given such that $\boldsymbol{y}_i = (y_i(s) : s \in S)$ and x_i's are the independent realizations of $Y(\cdot)$ and X, respectively.

4.2.1 The APLS algorithm

The functional linear model is usually used to access the association between $Y(\cdot)$ and X, that is,

$$X = a_0 + \int_{\mathcal{S}} Y(s)b(s)ds + \epsilon, \tag{4.1}$$

where ϵ is a scalar random variable with $E(\epsilon \mid X) = 0$, a_0 is a scalar parameter, and $b(\cdot)$ is an unknown coefficient function on $\mathcal{S}$. FPLS algorithm can be used to find the orthogonal basis function $\{\psi_j(s)\}_{j\geq 1}$ to approximate the functional coefficient $b(s)$. However, we introduce APLS algorithm, due to the difficulty of theoretical justification and computational implementation of FPLS algorithm.

Let $K(s,t) = Cov(Y(s), Y(t))$ and

$$K(b)(t) = \int_S K(s,t)b(s)ds$$

be the functional operator, then we obtain the first p non-orthogonal basis functions

$$\psi_j(s) = K^j(b)(s) = \int_S K^{j-1}(b)(t)K(s,t)dt$$

for $j = 2, \dots, p$. $b(s)$ can be approximated by the linear combination of the first p bases $\psi_j(s)$, that is $\sum_{j=1}^p \gamma_j \psi_j(s)$, where $\gamma_j = \int_S b(s)\psi_j(s)ds$. Model (4.1) is equivalent to

$$X = a_0 + \sum_{j=1}^{p} \gamma_j \int_{\mathcal{S}} Y(s)\psi_j(s)ds + \epsilon,$$

Thus, the estimation of $b(s)$ can be approximated in two steps. We can estimate $\hat{\psi}_j(s)$ and obtain $\hat{\gamma}_j$ through linear regression. An efficient APLS algorithm is summarised in Algorithm 1.

Actually, the estimator $\hat{\gamma}_j$ in Algorithm 1 can also be obtained by maximizing the

Algorithm 1 APLS algorithm

```
Input: Given the sample $(x_i, y_i(\cdot))_{i=1,\dots,n}$.
Output: $\hat{b}(s)$.
$\hat{K}(s,t) = n^{-1}\sum_{i=1}^{n}[y_i(s) - \bar{y}(s)][y_i(t) - \bar{y}(t)]$, where $\bar{y}(s) = n^{-1}\sum_{i=1}^{n} y_i(s)$ and $\tilde{\psi}_1(s) = \widehat{K^1(b)}(s) = n^{-1}\sum_{i=1}^{n}[y_i(s) - \bar{y}(s)][x_i - \bar{x}]$;
for $j = 2, \cdots, p$ do
    $\tilde{\psi}_j(s) = \widehat{K^j(b)}(s) = \int_S \widehat{K^{j-1}(b)}(t)\hat{K}(s,t)dt$;
Orthogonalize $\tilde{\psi}_j(s)$ to obtain $\hat{\psi}_j(s)$;
for $i = 1, \dots, n$ do
    for $j = 1, \dots, p$ do
        $z_{ij} = \int_{\mathcal{S}} y_i(s)\hat{\psi}_j(s)ds$;
for $j = 1, \dots, p$ do
    Estimate $\hat{\gamma}_j$ through linear regression $x_i = a_0 + \sum_{j=1}^{p}\gamma_j z_{ij} + \epsilon_i$, i.e. $\hat{\boldsymbol{\gamma}} = (Z^T Z)^{-1} Z^T \hat{X}$,
where $(Z)_{ij} = z_{ij}$, $\hat{X} = (x_1, \dots, x_n)^T$, and $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_p)^T$;
return $\hat{b}(s)$ is approximated by $\sum_{j=1}^{p}\hat{\gamma}_j\hat{\psi}_j(s)$.
```

correlation function, and the sample form of this model can be obtained as follows

$$\max_{\{\gamma_j\}_{j=1,\dots,p}} Cor_n\left(X, \sum_{j=1}^{p}\gamma_j \int_{\mathcal{S}} Y(s)\hat{\psi}_j(s)ds\right). \tag{4.2}$$

However, correlation cannot fully capture the nonlinear relationship between X and $\int_{\mathcal{S}} Y(s)\hat{\psi}_j(s)ds$, which means APLS algorithm can not be applied to estimate $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$ directly. Therefore, we propose a nonlinear estimation method based on the APLS algorithm to substitute the linear coefficient estimation method, which is described in the next subsection.

4.2.2 The FPLS-DC algorithm

In this subsection, we introduce distance correlation and its unbiased empirical version before presenting the FPLS-DC algorithm.

Definition 4.2.1. *The distance covariance (dCov) between random vectors X and Y with finite first moments is the nonnegative number $dCov(X,Y)$ defined by*

$$dCov^2(X,Y) = \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}} dtds,$$

where f_X and f_Y are the characteristic functions of X and Y, and $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$.

Definition 4.2.2. *The distance correlation (dCor) between random vectors X and Y with finite first moments is the nonnegative number dCor(X,Y) defined by*

$$dCor^2(X,Y) = \begin{cases} \frac{dCov^2(X,Y)}{\sqrt{dCov^2(X,X)dCov^2(Y,Y)}}, & dCov^2(X,X)dCov^2(Y,Y) > 0, \\ 0, & dCov^2(X,X)dCov^2(Y,Y) = 0. \end{cases}$$

The unbiased version of the squared sample distance covariance and distance correlation is defined as follows.

Definition 4.2.3. *Let (x_i, y_i), $i = 1, \dots, n$ denote a sample of observations from the joint distribution (X,Y) of random vectors X and Y. Let A $=(a_{ij})$ be the Euclidean distance matrix of the sample $x_1, \dots, x_n$, and B= (b_{ij}) be the Euclidean distance matrix of the sample $y_1, \dots, y_n$. Then if E($|X| + |Y|$) $< \infty$, for $n > 3$, the following*

$$dCov_n^2(X,Y) = \frac{1}{n(n-3)} \sum_{i \neq j} \tilde{A}_{ij} \widetilde{B}_{ij}$$

is an unbiased estimator of squared population distance covariance $dCov^2(X,Y)$, where

$$\widetilde{A}_{ij} = \begin{cases} a_{ij} - \frac{1}{n-2} \sum_{\ell=1}^n a_{i\ell} - \frac{1}{n-2} \sum_{k=1}^n a_{kj} + \frac{1}{(n-1)(n-2)} \sum_{k,\ell=1}^n a_{k\ell}, & i \neq j \\ 0, & i = j \end{cases}$$

and the form of $\widetilde{B}_{ij}$ is similar to that of $\widetilde{A}_{ij}$.

Definition 4.2.4. *The empirical distance correlation $dCor_n(X,Y)$ is the square root of*

$$dCor_n^2(X,Y) = \begin{cases} \frac{dCov_n^2(X,Y)}{\sqrt{dCov_n^2(X,X)dCov_n^2(Y,Y)}}, & dCov_n^2(X,X)dCov_n^2(Y,Y) > 0, \\ 0, & dCov_n^2(X,X)dCov_n^2(Y,Y) = 0. \end{cases}$$

Distance correlation has an encouraging performance at detecting the nonlinear relationship between X and $\sum_{j=1}^p \gamma_j \int_{\mathcal{S}} Y(s) \hat{\psi}_j(s) ds$. Thus, we can substitute correlation with distance correlation and obtain the following model

$$\max_{\hat{b}} dCor_n^2(X, \sum_{j=1}^{p} \gamma_j \int_{\mathcal{S}} Y(s)\hat{\psi}_j(s)ds). \tag{4.3}$$

The estimation of $b(s)$ can be divided two parts. First, we obtain p basis functions $\hat{\psi}_j(s)$ ($j = 1, \ldots, p$) through the APLS algorithm described in the previous subsection. Second, we estimate γ_j for $j = 1, \ldots, p$ by maximizing $dCor_n(X, \sum_{j=1}^{p} \gamma_j \int_{\mathcal{S}} Y(s)\hat{\psi}_j(s)ds)$.

For convenience , we can simplify the form of model (4.3). Denote the projected image of $Y(\cdot)$ along the jth direction $\hat{\psi}_j(s)$ as

$$Z_j = \int_{\mathcal{S}} Y(s)\hat{\psi}_j(s)ds$$

for $j = 1, \ldots, p$. Then the projected image $\tilde{Y}_{\tilde{b}}$ of $Y(\cdot)$ along $\tilde{b}(\cdot)$ is approximated by

$$\int_{\mathcal{S}} Y(s) \sum_{j=1}^{p} \gamma_j \hat{\psi}_j(s)ds = \sum_{j=1}^{p} \gamma_j Z_j.$$

Let $Z = (Z_1, \ldots, Z_p)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$, then $\tilde{Y}_{\tilde{b}}$ is approximated by $Z\boldsymbol{\gamma} = \sum_{j=1}^{p} \gamma_j Z_j$. Following above discussion, we can maximize the squared empirical distance correlation

$$dCor_n^2(X, Z\boldsymbol{\gamma}) = \frac{dCov_n^2(X, Z\boldsymbol{\gamma})}{\sqrt{dCov_n^2(X, X) dCov_n^2(Z\boldsymbol{\gamma}, Z\boldsymbol{\gamma})}}. \tag{4.4}$$

to obtain $\hat{\boldsymbol{\gamma}}$.

The maximization of (4.4) is equivalent to maximize the squared $dCov_n^2(X, Z\boldsymbol{\gamma})$ under a constraint of $\boldsymbol{\gamma}$, that is

$$\max_{\boldsymbol{\gamma}} dCov_n^2(X, Z\boldsymbol{\gamma}) \quad \text{s.t.} \quad \boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1. \tag{4.5}$$

Denote $\boldsymbol{a}_{i,j} = Z_i - Z_j$ and $b_{i,j} = |x_i - x_j|$, the lagrange function of (4.5) can be written as

$$L(\boldsymbol{\gamma}, \lambda) = \sum_{i \neq j} |\boldsymbol{a}_{i,j}^T \boldsymbol{\gamma}| \left[\frac{b_{i,j}}{n(n-3)} - \frac{2\sum_{k \neq i} b_{i,k}}{n(n-2)(n-3)} + \frac{\sum_{k \neq l} b_{k,l}}{n(n-1)(n-2)(n-3)} \right] - \lambda(\boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1).g$$

Algorithm 2 FPLS-DC algorithm

```
Input: Given the sample $(x_i, y_i(\cdot))_{i=1,\dots,n}$,1000 unit vectors that follow a normal distribution.
Output: $\hat{b}(\cdot)$.
$\hat{K}(s,t) = n^{-1}\sum_{i=1}^{n}[y_i(s) - \bar{y}(s)][y_i(t) - \bar{y}(t)]$, $\bar{y}(s) = n^{-1}\sum_{i=1}^{n} y_i(s)$ and $\hat{\psi}_1(s) = \widehat{K^1(b)}(s) =$
$n^{-1}\sum_{i=1}^{n}[x_i(s) - \bar{x}(s)][y_i - \bar{y}]$;
for $j = 2, \dots, p$ do
    $\hat{\psi}_j(s) = \widehat{K^j(b)}(s) = \int_S \widehat{K^{j-1}(b)}(t)\hat{K}(s,t)dt$;
Orthogonalize $\tilde{\psi}_j(s)$ to obtain $\hat{\psi}_j(s)$;
for $i = 1, \dots, n$ do
    for $j = 1, \dots, p$ do
        $z_{ij} = \int_{\mathcal{S}} y_i(s)\hat{\psi}_j(s)ds$;
Select the vector, which corresponds to the largest $dCov^2$, as the starting point $\gamma^{(0)}$ of
the SQP algorithm;
while $\boldsymbol{\gamma}^{(k+1)}$ is not close to $\boldsymbol{\gamma}^{(k)}$ do
    Searching the iterative direction $d_\gamma$ given the current iterate $(\boldsymbol{\gamma}^{(k)}, \lambda^{(k)})$, it can be
obtained by solving Quadratic Programming (QP) subproblem which is

    $$\underset{d_\gamma}{\text{minimize}} - \left(\nabla\mathcal{L}\left(\boldsymbol{\gamma}^{(k)}, \lambda^{(k)}\right)^T d_\gamma + \frac{1}{2}d_\gamma^T H^{(k)} d_\gamma\right)$$
    $$\text{subject to } \nabla l\left(\boldsymbol{\gamma}^{(k)}\right)^T d_\gamma + l\left(\boldsymbol{\gamma}^{(k)}\right) = 0;$$

The step length $\alpha$ should be determined to update $\boldsymbol{\gamma}^k$ in the next iteration that is
$\boldsymbol{\gamma}^{(k+1)} = \boldsymbol{\gamma}^{(k)} + \alpha d_\gamma$, $\lambda^{(k)}$ should be updated too;
Choosing the convergent value of $\boldsymbol{\gamma}^{(k)}$ as the estimator $\hat{\boldsymbol{\gamma}}$;
return $\hat{b}(s)$ is approximated by $\sum_{j=1}^{p}\hat{\gamma}_j\hat{\psi}_j(s)$.
```

Sequential Quadratic Programming can be used to maximize $L(\boldsymbol{\gamma}, \lambda)$. Let's give some definitions, $\gamma_u^{(k)}$ is the value of γ_u in step k, and $\boldsymbol{\gamma}^{(k)} = (\gamma_1^{(k)}, \dots, \gamma_p^{(k)})^T$ is the value of $\boldsymbol{\gamma}$ in step k. $d_\gamma = \boldsymbol{\gamma} - \boldsymbol{\gamma}^{(k)}$, $l(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T\boldsymbol{\gamma} - 1$. $\nabla l(\boldsymbol{\gamma}^{(k)})$ is the gradient of l at $\boldsymbol{\gamma}^{(k)}$ and $\nabla\mathcal{L}\left(\boldsymbol{\gamma}^{(k)}, \lambda^{(k)}\right) = \left(\frac{\partial L(\boldsymbol{\gamma}^k)}{\partial\gamma_u^{(k)}}\right)_{p\times 1}$ is the gradient of $\mathcal{L}\left(\boldsymbol{\gamma}^{(k)}, \lambda^{(k)}\right)$ at $\boldsymbol{\gamma}^{(k)}$. $H^{(k)}$ is the Hessian of $\mathcal{L}\left(\boldsymbol{\gamma}^{(k)}, \lambda^{(k)}\right)$ at $\boldsymbol{\gamma}^{(k)}$. FPLS-DC algorithm is summarized in Algorithm 2.

4.2.3 Hypothesis Testing

The global test of the whole image with regard to each SNP can be implemented through the dCov test between the SNP and the projected image with respect to this SNP. Under the hypothesis that the X and the projected image $Y(\cdot)$ along the direction $\hat{b}(s)$ is unrelated, then we get the hypothesis testing

$$H_0 : X \perp\!\!\!\perp Y(\cdot) \quad \text{vs} \quad H_1 : X \not\!\perp\!\!\!\perp Y(\cdot),$$

we would expect $\text{dCov}(X, \int_{\mathcal{S}} Y(s)\hat{b}(s)ds)$ to be zero at all $s \in \mathcal{S}$ under the hypothesis, the p-value of a dCov test is usually obtained through permutation. In order to reach the significant p-value after Bonferroni correction, one needs at least $\frac{1}{\frac{0.05}{n_G}} = 20\ n_G$ permutations, where n_G is the total number of top SNPs obtained from screening. If $n_G = 4000$, then we need at least 8×10^4 permutations. Increase the number of permutations to 3 times, we would need 2.4×10^5 permutations to calculate the p-value of each SNP. Therefore, the number of permutations is not computationally affordable. Actually, a gamma function can be used to approximate the null distribution, and the mean and variance of the gamma function can be estimated using a smaller number of permutations. We use dcov.gamma() function in the 'kpcalg' R package to approximate dCov test, which is applied directly to the SNP and the projected image along a thresholded $\tilde{b}(s) = \hat{b}(s)\ I\{|T(s)| > \Phi^{-1}(1-\alpha)\}$ that filters out the noise at non-related voxels, where $T(s)$ is the local test statistic at s voxel to be introduced next, Φ^{-1} is the inverse cdf of the standard normal distribution, and α is the significance level adjusted for the number of voxels in the image, e.g. $\alpha = \frac{0.05}{m}$, where m is the number of voxels. Therefore, the p-value of dCov test can be approximated through dcov.gamma() function.

Actually, the Type I Error can be influenced by p, Type I Error is high for large p, the reason may come from two sources: 1) the base functions $\hat{\psi}_j(s)$ extracted as correlated to x using the APLS algorithm; and 2) the coefficients $\hat{\boldsymbol{\gamma}}$ estimated as most correlated to x through maximized distance covariance. Both of these two procedures move the projected image to the direction to be related to x. Therefore, the test of independence between the projected image and x is likely to be rejected. By decreasing p, we should expect a lowered Type I Error. However, the choice of p may depend on the data, and in the real scenario when the truth is unknown, it is hard to select a best p and therefore this direct dCov test is not recommended.The diminished power in the quadratic case is likely due to using the linear method in base-function extraction through APLS.

The local test at each voxel can be obtained through $\hat{b}(s_m)$ for $m = 1, \ldots, M$, where M is the number of voxel. Since $\hat{b}(\cdot)$ maximizes the sample distance covariance between X and the projected image $\int_{\mathcal{S}} Y(s)\hat{b}(s)ds$, and puts a weight at each voxel s of the image $Y(s)$, the weighted sum across $\mathcal{S}$ is most closely related to X. Therefore, $\hat{b}(\cdot)$ represents the signal pattern across $\mathcal{S}$ that is related to X. Under the null hypothesis of independence between X and $Y(\cdot)$, $\hat{b}(\cdot)$ is close to zero . Then we can get the hypothesis testing

$$H_0 : \hat{b}(s_m) = 0 \quad \text{vs} \quad H_1 : \hat{b}(s_m) \neq 0, \qquad \text{for} \;\; m = 1, \ldots, M.$$

Since we did not assume any parametric model for the data generation of X and $Y(\cdot)$ related to $b(\cdot)$, in the freedom without any other assumptions, we assume that $\hat{b}(s_m)$ follows a zero-mean normal distribution. We can use bootstrap to estimate $Var[\hat{b}(s_m)]$ at each voxel, which can then be used to construct the test statistic, under the assumption of signal independence between any two voxels in $\mathcal{S}$. The bootstrap procedure is proposed as follows,

Step 1. Randomly choose n samples $\{x_i^*, y_i^*(\cdot)\}_{i=1,\ldots,n}$ from the observation $\{x_i, y_i(\cdot)\}_{i=1,\ldots,n}$ with replacement.

Step 2. Obtain the estimator $\hat{b}^*(s_m)$ based on the sample $\{x_i^*, y_i^*(\cdot)\}_{i=1,\ldots,n}$.

Step 3. Repeat Step1 and Step2 K times to get K estimates $\{\hat{b}_k^*(s_m)\}_{k=1,\ldots,K}$, for $m = 1, \ldots, M$.

Step 4. Calculate the sample variance of $\{\hat{b}_k^*(s_m)\}_{k=1,\ldots,K}$ as the estimate of $Var[\hat{b}(s_m)]$.

After obtaining $\widehat{Var}[\hat{b}(s_m)]$, the test statistic at each voxel is

$$T(s_m) = \frac{\hat{b}(s_m)}{\sqrt{\widehat{Var}[\hat{b}(s_m)]}},$$

which follows the standard normal distribution under the null. The p-value for $T(s_m)$ is thus calculated as

$$2(1 - \Phi(|T(s_m)|)),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

4.2.4 Two step estimation algorithm

The dimension of SNPs and image is high so that the computational cost of FPLS-DC algorithm is unaffordable. In order to relieve the computational intensity, we propose a two step algorithm to estimate the functional coefficient. Firstly, We select the first FPLS-derived base to decrease the dimension of image, then we use APLS algorithm to decrease the dimensional of SNPs based on the selected base. In the estimation algorithm, SNP is the response variable and the projected image along the first direction is the explanatory variable, when Bonferroni-adjusted p-value is less 0.05, SNPs can be selected for further investigation. Secondly, the FPLS-DC algorithm is used to estimate $\hat{b}(\cdot)$ corresponding to each of the screened SNPs after screening the SNPs.

4.3 Simulation

4.3.1 Data Generation

In this section, we examine our proposed method by some numerical studies. We simulate SNP data through coalescent simulation (COSI) (Schaffner et al. 2005). 2000 haplotypes are simulated using COSI for 100 1Mb regions. Then, the SNPs are ascertained and thinned to match HapMapII SNP frequency distribution. Those haplotypes are randomly combined in pairs to form the genotype data of 1000 individuals. Then, the SNPs are filtered with minor allele frequency (MAF) bigger than 0.05. After filtering, we end up with 77,100 SNPs.

We randomly select 100 SNPs as causal and generate the imaging data, which is a 150×100 rectangle for each subject i ($i = 1, \ldots, n$), based on their additive effect

$$y_i(s) = \sum_{g=1}^{100} f(x_i^{(g)})\beta^{(g)}(s) + \epsilon_i(s),$$

where $g = 1, \dots, 100$ is the SNP indicator, $f(x) = x$ or $(x - 0.4)^2$, $s = (s_x, s_y)$, and $\beta^{(g)}(s) = 0.25 \cdot I\{(s_x - c_x^{(g)})^2 + (s_y - c_y^{(g)})^2 \leq 10^2\}$. Let $(c_x^{(g)}, c_y^{(g)})$ vary across $g = 1, \dots, 100$, so that marker effects do not lay over the same location. Restrict $c_x^{(g)} \in [20, 130]$ and $c_y^{(g)} \in [20, 80]$ such that the affected region of each SNP is entirely contained within the image. $\epsilon_i(s)$ is simulated as $\epsilon_i(s) = \xi_{1,i}\phi_1(s) + \xi_{2,i}\phi_2(s)$, where $\phi_1(s) = \sqrt{0.5}\sin(\frac{2\pi s_x}{150})$, $\phi_2(s) = \sqrt{0.5}\cos(\frac{2\pi s_y}{100})$, $\xi_{1,i} = \lambda_1 Z_{1,i}$, $\xi_{2,i} = \lambda_2 Z_{2,i}$, $\lambda_1 = \sqrt{1.2}$, $\lambda_2 = \sqrt{1.0}$, and $Z_{1,i}$ and $Z_{2,i}$ are independent standard normal random variables; $i = 1, \dots, n$. This yields $\epsilon_i(s)$ to range from $(-6.29, 6.29)$, with sample variance 2.03, which is close to $2.2 = \lambda_1^2 + \lambda_2^2$.

4.3.2 Screening

Global screening is performed for $p = 1, \dots, 10$, p is the number of bases function. All SNPs are sorted according to the p-value in the linear regression model, where the SNP dosage is the response variable and the image projected along each base are the explanatory variables. Causal SNP inclusion rate was averaged over the 100 replicates and summarized in Table 3.6, where the causal SNP inclusion rate was calculated as the percentage of the number of causal SNPs included in the top SNPs among the total number of 100 causal SNPs.

In each section of $f(x)$ in Table 4.1, as n_{g_s} increases, the inclusion rate increases given the same p. This is what we would expect, because the more SNPs included, the more causal SNPs are likely to be covered. As p increases, the inclusion rate increases for $n_{g_s} \geq 1000$; but for smaller n_{g_s}'s, the causal SNP inclusion rate drops to 0 when p is large enough. This informs us the distribution of causal SNP rankings with different values of n_{g_s} and p. The causal SNPs may not rank at the very top in the screening. For each p, the distribution interval of the cuasal SNP rankings is between the n_{g_s}'s corresponding to 0 inclusion rate and 100% inclusion rate. We can see that as p increases, the distribution interval shrinks. For $p = 7, \dots, 10$, the distribution interval is constrained within $n_{g_s} = 500$ to 3000; for $p = 6$, the interval expanded to $(100, 4000)$; and for $p = 5$ and 4, the interval further expanded to $(0, 5000)$. For p shrinks down from 3 to 1, the upper bound of the interval goes beyond 5000,

$f(x)$	p	Number of Top SNPs (n_{g_s})							
		100	250	500	1000	2000	3000	4000	5000
x	1	0.38	0.93	1.66	2.93	5.40	7.51	9.70	11.67
	2	9.42	16.54	26.60	39.83	53.92	61.70	66.81	69.42
	3	10.83	21.82	37.20	58.42	79.57	87.80	91.63	93.69
	4	9.99	26.22	42.99	68.25	87.16	95.34	100.00	100.00
	5	10.00	28.97	52.00	72.05	89.00	98.00	99.00	100.00
	6	0.00	39.19	61.87	79.15	94.06	99.00	100.00	100.00
	7	0.00	0.00	0.00	84.58	97.00	100.00	100.00	100.00
	8	0.00	0.00	0.00	89.00	98.00	100.00	100.00	100.00
	9	0.00	0.00	0.00	89.05	98.00	100.00	100.00	100.00
	10	0.00	0.00	0.00	89.19	98.01	100.00	100.00	100.00
$(x-0.4)^2$	1	0.58	1.10	2.01	3.42	5.95	8.30	10.53	12.65
	2	6.78	16.59	24.61	35.64	44.47	47.59	49.77	51.30
	3	5.03	14.63	24.75	40.44	51.25	56.96	60.60	63.36
	4	11.01	20.01	32.78	46.10	64.09	72.01	77.21	80.05
	5	13.00	23.00	35.98	52.97	69.00	76.36	79.24	83.04
	6	12.00	25.00	41.76	56.41	72.61	79.08	81.11	88.25
	7	16.00	33.00	45.10	59.15	78.00	82.01	86.49	90.54
	8	0.00	33.55	47.00	65.19	78.17	86.09	89.72	93.00
	9	0.00	33.19	47.27	65.79	80.38	87.81	90.61	94.14
	10	0.00	33.26	47.44	66.38	81.34	88.14	91.02	94.53

Table 4.1: Inclusion rates of causal SNPs: the percentage of the number of causal SNPs included in the top SNPs within the total number of 100 causal SNPs, for $p = 1, \ldots, 10$.

where we did not examine in the simulation. As p increases to 6, we can see that the top 100 SNPs from screening do not include any causal SNPs, the reason is that all of the top 100 SNPs are non-causal due to noise introduced from the additional bases, or due to those top-ranking non-causal SNPs being correlated with the causal SNPs.

If we use 80% inclusion rate as the criteria for good screening performance, $n_{g_s} = 3000$ and $p = 6$ can give good screening results according to Table 4.1 to detect SNPs with roughly linear effect. Using only the top base $\hat{\psi}_1(s)$ with $n_{g_s} = 2000$ for screening, although computationally affordable, may include only a small proportion of true causal SNPs. Using the University cluster by running 200 jobs simultaneously, the screening procedure with $p = 6$ is estimated to take 21 days to screen all 503,892 SNPs for both the left side and right side of the hippocampus surface radial distance data ($m = 15,000$), while with $p = 1$ it only took 40 to 60 minutes with 50 simultaneous jobs, which is 2419 times faster. Users can choose their p and n_{g_s} to balance the computation cost and the true causal-SNP inclusion rate referencing Table 4.1. The computation time can be calculated by running a few SNPs on a machine of choice by the user.

4.3.3 FPLS-DC and Statistical Tests

A causal SNP and a non-causal SNP were selected for this simulation. The non-causal SNP is selected so that it is not correlated to any of the 100 causal SNPs, where a significant correlation is indicated by a p-value smaller than 0.05 in a Pearson's correlation test performed in R using the cor.test() function. The FPLS-DC algorithm was implemented with $p = 10$. Then, the local test was implemented on $\hat{b}(s)$, with the $Var[\hat{b}(s)]$ estimated using the sample variance of $\hat{b}(s)$ over the 100 replicates from the non-causal SNP, which is used to calculate the test statistics for both the causal and non-causal SNPs. The null hypothesis is rejected if the local p-value is less than $3.33 \times 10^{-6} = \frac{0.05}{15000}$.

In order to evaluate the effectiveness of the FPLS-DC method, We compare the linear regression method (FPLS) , the modified dCor-based method (wdCor) (Wen et al., 2020), and PC-based method for $\boldsymbol{\gamma}$ estimation with our proposed method. The original wdCor (Wen et al., 2020) estimates $\boldsymbol{\gamma}$ by optimizing the exponential of distance correlation, since we are aiming for a linear combination of $\hat{\psi}_j(s)$'s eventually, therefore modify dCor-based method to estimate $\boldsymbol{\gamma}$: for $j = 1, \ldots, p$,

$$\hat{\gamma}_j = w_j^{opt} = \frac{\beta_j^{r_{opt}}}{||\boldsymbol{\beta}^{r_{opt}}||_2}$$

$$\text{with} \qquad \beta_j = dCor(X, Z_j), \quad w_j = \frac{\beta_j^r}{||\boldsymbol{\beta}^r||_2}, \quad \text{and}$$

$$r_{opt} = \underset{r \in \{1,3,\ldots,15\}}{\arg\max}\, dCor\big(X, \sum_{j=1}^{p} w_j Z_j\big).$$

where $||\cdot||_2$ stands for the Euclidean norm and $Z_j = \int_{\mathcal{S}} Y(s)\hat{\psi}_j(s)ds$.

For the above three comparable methods, after obtaining $\hat{\boldsymbol{\gamma}}$ and thus $\hat{b}(s)$, the same local statistical test as proposed in section 4.2.3 is conducted at each voxel. Then, the type I errors and power are calculated and summarized in Table 4.3. The Type I Errors for local tests are family-wise error rates for the entire image across 15,000 voxels, estimated using average rejection rates over the 100 replicates, where the family-wise rejection is made if any of the voxelwise local test is rejected for a p-value $< 3.33 \times 10^{-6} \approx \frac{0.05}{15000}$ according to the Bonferroni correction for multiple testings. The theoretical family-wise error rate in this case is $0.049 \approx 1 - (1 - \frac{0.05}{15000})^{15000}$.

From Table 4.3, we can see that our proposed method FPLS-DC gives good control of local Type I Error when tested on the non-causal SNP under $f(x) = x$, while under $f(x) = (x - 0.4)^2$, the local Type I Error is inflated. For all methods, the Type I Error in the unaffected region of the causal SNP is larger than their corresponding Type I Error tested with the non-causal SNP and likely to be inflated. The Type I Error for global tests are well controlled, and power is mostly suffered in the quadratic case.

The high Type I Error for large p may come from two sources: 1) the base functions $\hat{\psi}_j(s)$ extracted as correlated to x using the APLS algorithm; and 2) the coefficients $\hat{\boldsymbol{\gamma}}$ estimated as most correlated to x through maximized distance covariance. Both of these two procedures move the projected image to the direction to be related to x. Therefore, the test

Type I Error	f	Method	p									
			1	2	3	4	5	6	7	8	9	10
Local Test: Non-Causal SNP	f_1	FPLS-DC	1.00	<0.01	<0.01	<0.01	0.01	<0.01	<0.01	<0.01	0.02	0.04
		FPLS	1.00	<0.01	0.03	<0.01	<0.01	<0.01	<0.01	<0.01	0.19	0.23
		wdCor	1.00	<0.01	<0.01	<0.01	<0.01	0.01	<0.01	<0.01	0.12	0.19
		PC	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
	f_2	FPLS-DC	1.00	<0.01	<0.01	<0.01	<0.01	<0.01	0.02	<0.01	0.03	0.02
		FPLS	1.00	<0.01	0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.15	0.12
		wdCor	1.00	<0.01	<0.01	<0.01	<0.01	0.02	0.06	0.11	0.24	0.29
		PC	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Local Test: Causal SNP	f_1	FPLS-DC	1.00	<0.01	<0.01	<0.01	<0.01	0.04	0.20	0.35	0.48	0.45
		FPLS	1.00	0.12	0.71	0.88	1.00	1.00	1.00	1.00	1.00	1.00
		wdCor	1.00	<0.01	<0.01	<0.01	0.02	<0.01	0.03	0.02	0.09	0.12
		PC	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.11	0.05	0.02
	f_2	FPLS-DC	1.00	<0.01	<0.01	<0.01	0.03	<0.01	<0.01	0.02	0.12	0.07
		FPLS	1.00	<0.01	0.13	0.91	1.00	1.00	1.00	1.00	1.00	1.00
		wdCor	1.00	0.01	<0.01	<0.01	<0.01	<0.01	0.01	0.01	0.23	0.31
		PC	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Global Test	f_1	FPLS-DC	0.02	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
		FPLS	0.02	<0.01	0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.02	<0.01
		wdCor	0.02	<0.01	<0.01	<0.01	<0.01	0.01	<0.01	<0.01	0.02	0.04
		PC	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
	f_2	FPLS-DC	0.02	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
		FPLS	0.02	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
		wdCor	0.02	<0.01	<0.01	<0.01	<0.01	0.01	<0.01	<0.01	0.03	0.03
		PC	0.26	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Table 4.2: Type I Error for Local and Global Tests. $f_1(x) = x$ and $f_2(x) = (x - 0.4)^2$.

of independence between the projected image and x is likely to be rejected. The number of base functions used here is $p = 10$. By decreasing p, we should expect a lowered Type I Error. However, the choice of p may depend on the data, and in the real scenario when the truth is unknown, it is hard to select a best p and therefore this direct dCov test is not recommended. The diminished power in the quadratic case is likely due to using the linear method in base-function extraction through APLS.

Power	f	Method	p									
			1	2	3	4	5	6	7	8	9	10
Local Test	f_1	FPLS-DC	1.00	<0.01	0.41	0.45	0.28	0.29	0.23	0.18	0.16	0.16
		FPLS	1.00	0.30	0.80	1.00	1.00	0.97	0.94	0.91	0.89	0.87
		wdCor	1.00	<0.01	0.34	0.31	0.27	0.28	0.31	0.30	0.29	0.29
		PC	1.00	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	0.02	0.01
	f_2	FPLS-DC	0.98	<0.01	0.06	0.10	0.09	0.09	0.06	0.05	0.04	0.05
		FPLS	0.97	<0.01	0.13	0.46	0.68	0.78	0.81	0.82	0.79	0.79
		wdCor	0.98	<0.01	0.02	0.04	0.05	0.04	0.05	0.07	0.06	0.05
		PC	0.53	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Global Test	f_1	FPLS-DC	0.05	<0.01	0.69	0.77	0.71	0.75	0.69	0.80	0.80	0.77
		FPLS	0.05	0.81	0.90	0.94	1.00	1.00	1.00	1.00	1.00	1.00
		wdCor	0.05	<0.01	0.73	0.70	0.68	0.68	0.66	0.63	0.64	0.65
		PC	1.00	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.07	0.15	0.12
	f_2	FPLS-DC	0.04	<0.01	0.08	0.06	0.05	0.04	0.03	0.02	0.02	0.05
		FPLS	0.04	0.03	0.14	0.16	0.14	0.22	0.86	1.00	1.00	1.00
		wdCor	0.04	0.01	0.08	0.09	0.08	0.08	0.09	0.09	0.11	0.11
		PC	1.00	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Table 4.3: Power for Local and Global Tests. $f_1(x) = x$ and $f_2(x) = (x - 0.4)^2$.

4.4 ADNI Application

4.4.1 Data Description

In this section, we apply our pipeline to the hippocampus surface dara from the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). AD is known to affect the hippocampus, which plays a major role in memory and learning. The purpose of the real data analysis is to examine the genetic effect of SNPs on hippocampus. The initial ADNI study (ADNI-1) were followed by ADNI-GO and ADNI-2 under different image acquisition protocols. Our analysis focuses on data acquired from ADNI-1, which includes 818 healthy, AD, and mild cognitive impaired (MCI) subjects with genotype data.

We used the hippocampus surface radial distance data from the baseline image measurement of ADNI-1 that are processed by Huang et al. (2017). Radial distance, the distance from the medial core to each surface point, informs us the deformation along the surface normal direction (Pizer et al., 1999; Thompson et al., 2004; Styner et al., 2004). The ADNI-1 images are obtained from 1.5 T MRI scanners with a 256 x 256 x 170 acquisition matrix, where each voxel is of size $1.25 \times 1 : 26 \times 1.2$ mm^3. The raw MRI images are pre-processed using standard pipelines as described by Huang et al. (2017), including alignment correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation, and registration. Then, all voxels in the brain image are labelled by regions of interest (ROI). As one of the 93 ROIs, the hippocampus is extracted and its surface registered (Huang et al., 2017; Shi et al., 2013) prior to obtaining the surface statistics, including the radial distance that is used in our analysis.

The genotype data from ADNI-1 were acquired using the Illumina Human 610-Quad BeadChip platform and includes 620,901 SNPs. We restricted the analysis to only Caucasians. After quality control and SNP filtering for miner allele frequency (MAF) > 0.05, the pre-processed imaging-genetic data for our analysis contains 503,892 SNPs and 730 subjects. Those subjects age from 54 to 91, with 431 males and 299 females. To remove potential effects from covariate, we regressed out age, gender, APOE-ϵ4 (the strongest known genetic risk factor for AD), and the top five genetic principal components from the processed radial distance hippocampus data prior to performing the analyses proposed in this chapter.

4.4.2 Analysis Procedure

Using the imaging-genetic data described above, we first implemented a global screening to filter promising SNPs. For this smaller set of SNP, APLS and dCov optimization are applied to estimate $\hat{b}(s)$. Then, we demonstrate the application of voxelwise local test and global test on the screened SNP that yields that largest distance correlation between the SNP and the projected image along $\hat{b}(s)$, for each side of the hippocampus.

The global screening was performed using only one base ($p = 1$), i.e. the first base function $\hat{\psi}_1(s)$ extracted from the FPLS algorithm, for the left hippocampus and the right hippocampus, respectively. In particular, for each of the 503,892 SNPs, first, calculate $\hat{\psi}_1(s)$ and project the image of each subject along $\hat{\psi}_1(s)$ to get the projected image $\tilde{y}_i = \int_{\mathcal{S}} y_i(s)\hat{\psi}_1(s)ds$ for each subject i; then, fit a linear regression SNP ~ $\tilde{Y}$ and obtain the model p-value. After obtaining the p-values corresponding to all the SNPs, select the SNPs with p-value $< 4.96 \times 10^{-8} = \frac{0.05}{503{,}892 \times 2}$ as the screened SNPs to proceed into the next steps, the refined analyses using more bases and enhanced coefficient estimation algorithm, which more computationally intensive.

After the screening step, the screened SNPs are analyzed as follows for the left hippocampus and right hippocampus, respectively. For each of the screened SNPs for the corresponding side of the hippocampus, use FPLS to extract the first 10 base functions $\hat{\psi}_j(s)$ for $j = 1, \ldots, 10$. Next, use the SQP to estimate $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{10})^T$ that maximizes dCov under the constraint $|\boldsymbol{\gamma}| = 1$, as described in the method section. Then, estimate $\hat{b}(s) = \sum_{j=1}^{10} \gamma_j \hat{\psi}_j(s)$ for this SNP.

After computing $\hat{b}(s)$ for each of the screened SNPs corresponding to the left and right hippocampus, respectively, perform the local test at each voxel s for the targeted SNPs. As a demonstration, we perform the statistical tests on the SNP that yields the maximum distance correlation with the projected image, for the left hippocampus and right hippocampus, respectively. First, compute the local test statistics $T(s) = \frac{\hat{b}(s)}{\sqrt{\widehat{Var}[\hat{b}(s)]}}$, where $\widehat{Var}[\hat{b}(s)]$ is estimated using the sample variance of 50 bootstrap samples. For each bootstrap sample, $\hat{b}(s)$ is estimated. Permutation is added in each bootstrap to mimic the distribution of $\hat{b}(s)$ under the null. After obtaining the local test statistics $T(s)$ at each voxel $s \in \mathcal{S}$, where $|\mathcal{S}| = 15,000$, the p-value at each voxel is calculated as $1 - 2\Phi\big(|T(s_m)|\big)$ as proposed in the method section. Two significance levels are considered: $t_1 = \frac{0.05}{15000 \times n_{g_s}}$, where n_{g_s} is the total number of screened SNPs of the left and right hippocampus, and a more relaxed $t_2 = \frac{0.05}{15000} = 3.33 \times 10^{-6}$.

After applying the local test, the global test is applied based on the local test results. In particular, a dCov test with gamma approximation for the dCov distribution (Szekely et al., 2007) is performed to test the independence between the target SNP and the projected image along the thresholded $\hat{b}(s)$: $\tilde{b}(s) = \hat{b}(s)\ I\{|T(s)| > \Phi^{-1}(1 - \alpha)\}$, where $T(s)$ is the local test statistic at s, Φ^{-1} is the inverse cdf of the standard normal distribution, and $\alpha = \frac{0.05}{15000}$. This dCov test was implemented using the dcov.gamma() function from the 'kpcalg' R package. The minimum non-zero p-value that dcov.gamma() can give is 1.11×10^{-16}; p-values smaller than that are output as 0.

The global test depends on the voxelwise local test p-values, as well as $\hat{b}(s)$, and therefore requires performing the local test prior to performing the global test. The local test, which includes a bootstrap procedure to estimate the variance of $\hat{b}(s)$ at each voxel s, took around four hours on the University cluster. Therefore, to perform the refined analysis, including $\hat{b}(s)$ estimation and the local and global tests, the estimated run time on 2000 screened SNPs for each side of the hippocampus is about one week, by submitting 100 parallel jobs to the University cluster. More computation resource should be required in order to perform the formal bootstrap-based global test to obtain a legal p-value as proposed at the beginning of section 4.2.3.

4.4.3 Results and Interpretations

The analysis results are described as follows. After screening the 503,892 SNPs for the left and right hippocampus, respectively, 1358 genome-wide significant SNPs are included for the left hippocampus and 2083 for the right, which gives $n_{g_s} = 3441$ screened SNPs in total. After estimating $\hat{b}(s)$, $-\log_1 0$(p-value)'s from screening and the distance correlations between each SNP and its corresponding projected image along $\hat{b}(s)$ are compared in the scatter plots in Figure 4.1. From this figure, We can see that most of the screened SNPs have low dCor values. Among the 1358 screened SNPs for the left hippocampus, rs5996980 (chr22:25,819,296 in GRCh38.p13) yields the maximum distance correlation with the projected image long its corresponding $\hat{b}(s)$. Among the 2083 screened SNPs for the right hippocampus, rs8108292 (chr19:57,201,927 in GRCh38.p13) yields the maximum distance correlation with the projected image long its corresponding $\hat{b}(s)$. Local tests are performed on these two SNPs w.r.t. their corresponding projected images. The voxelwise p-values are calculated and plotted in Figure 4.2.

SNP rs5996980 has minimum voxelwise p-value 1.21×10^{-10}. Out of the 15,000 voxels on the left hippocampus surface, there are only 2 voxels at the bottom of the hippocampus with significant p-values using the more stringent threshold t_1. If using threshold t_2, the

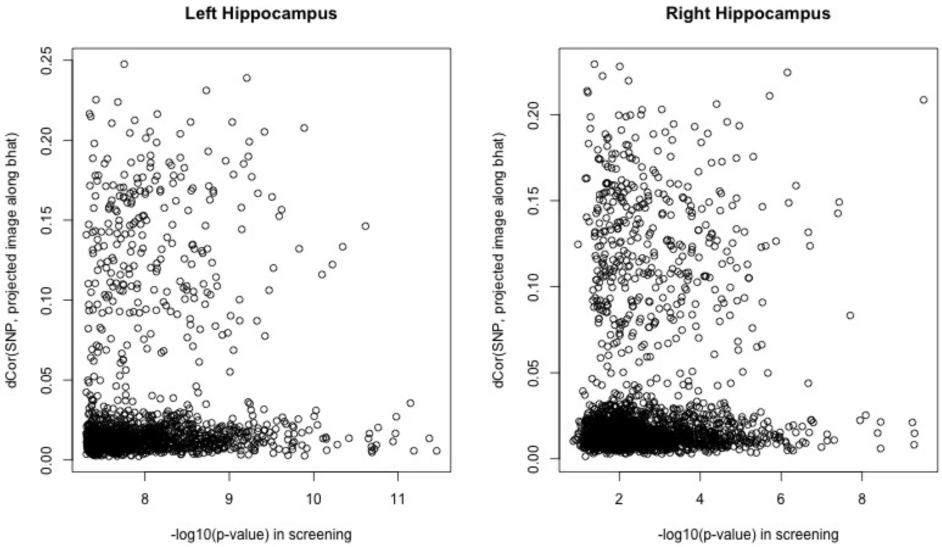


Figure 4.1: Top SNP Ranking: Screening VS Opt. dCor. $-\log_1 0$(p-value) obtained from the model in screening is plotted against the distance correlation between each SNP and the projected image along $\hat{b}(s)$. $p = 1$ is used in screening. There are 1358 screened SNPs on the left and 2083 on the right.

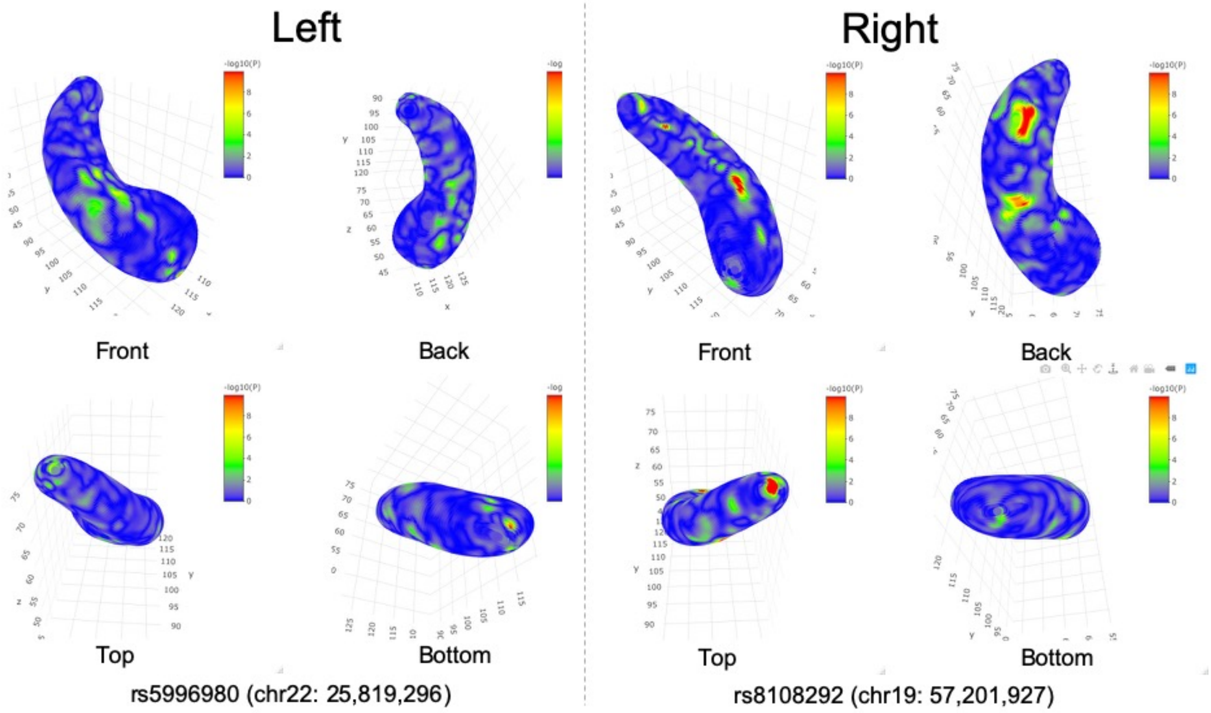


Figure 4.2: $-\log_{10}$(p-value) for the top screened SNP for each side of the hippocampus.

number of significant voxels increase to 19, including a total of 6 voxels at the bottom tip of the hippocampus, 3 voxels near the top tip, and 10 voxels on the Fimbria in two connected clusters of 5 voxels. An anatomy of the hippocampus can be found in Figure 4.4, which is used in (Huang et al., 2017). The global test was also performed on rs5996980 and the left hippocampus, yielding a p-value of 0.257. The large p-value is probably due to only a small number of voxels (19 out of 15,000) being locally significant to pass the threshold t_2. rs5996980 is an intron variant located inside gene MYO18B. Although intron variants are not transcribed and translated to form proteins, they may play a regulatory role. This SNP is not reported in the GWAS Catalog. In NCBI, it is indicated as not reported in ClinVar for clinical significance and no citations in publications. MYO18B, myosin XVIIIB, is a protein-coding gene. In the GWAS Catalog, it is reported to include SNPs associated with brain functions (schizophrenia, bipolar disorder and depression combined, cognitive performance, mathematical ability, and neurofibrillary tangles), heart functions (atrial fibrillation, left ventricular mass to end-diastolic volume ratio, left ventricle wall thickness, and llectrocardiogram morphology), metabolism (trans fatty acid levels, urinary calcium excretion, and urate levels in lean individuals) , and other traits (adolescent idiopathic scoliosis, gut microbiome measurement, eye morphology measurement, uterine fibroids, and acute myeloid leukemia). These functions of MYO18B is related to the function of the hippocampus such as learning and memory, as well as the role related to neurological and psychiatric disorders. In particular, the fimbria-fornix volume is associated with spatial memory and olfactory identification in humans (Dahmani et al., 2020).

SNP rs8108292 has minimum voxelwise p-value 3.65×10^{-97}. Out of the 15,000 voxels on the left hippocampus surface, there are 160 voxels (shown in the top row of Figure 4.3) with significant p-values using the more stringent threshold t_1. If using threshold t_2, the number of significant voxels increase to 340 (shown in the second row of Figure 4.3). The global test was also performed on rs8108292 and the right hippocampus, yielding a p-value of 2.79×10^{-4}. This relatively small p-value is probably contributed by the relatively large number of voxels

(340 out of 15,000) being locally significant to pass the threshold t_2. rs8108292 is an intron variant located within gene ZNF264. This SNP is not reported in the GWAS Catalog. In NCBI, it is indicated as not reported in ClinVar for clinical significance and no citations in publications. ZNF264, zinc finger protein 264, is a protein-coding gene. In the GWAS Catalog, it is reported to include SNPs associated with body mass index through an intergenic SNP rs11670527 (chr19:57,182,570 in GRCh38.p13) upstream of gene ZNF264 (Kusic et al., 2020). According to the review paper on zink finger proteins (ZNFs) by Cassandri et al. (2017), the wide varieties of ZNFs are abundant in the human body and able to interact with DNA, RNA, PAR (poly-ADP-ribose) and other proteins, and thus involved in regulation processes such as transcriptional regulation, ubiquitin-mediated protein degradation, signal transduction, actin targeting, DNA repair, and cell migration.

4.5 Discussion

In this article, we developed a new FPLS-DC algorithm for FLM to estimate the coefficient function. FPLS-DC algorithm has great performance on nonlinear model and high-dimensional data. We propose the test statistic to evaluate the effectiveness of FPLS-DC algorithm through Monte Carlo method in the simulation and to apply it on the real data analysis of ADNI. The result in simulation and real data analysis shows that FPLS-DC algorithm plays a great role in functional data analysis.

FPLS-DC is proposed under the assumption that the moment exists, when this assumption can not hold, ball covariance, which is constructed by indicator function, can be used to estimate the coefficient for each direction, similar to maximizing the distance covariance. Local test is obtained under the assumption of independence between two voxels in $\mathcal{S}$, the power may be influenced by interaction between two SNPs, and unobserved casual SNPs. These influence factors should be considered in the test. The comparison between FPLS-DC and APLS is not considered in simulation and real data analysis, in terms of computational efficiency and the statistical performance, so we can elaborate the difference in the two models. The idea of generating bases in FPLS-DC is based on APLS algorithm, and the bases

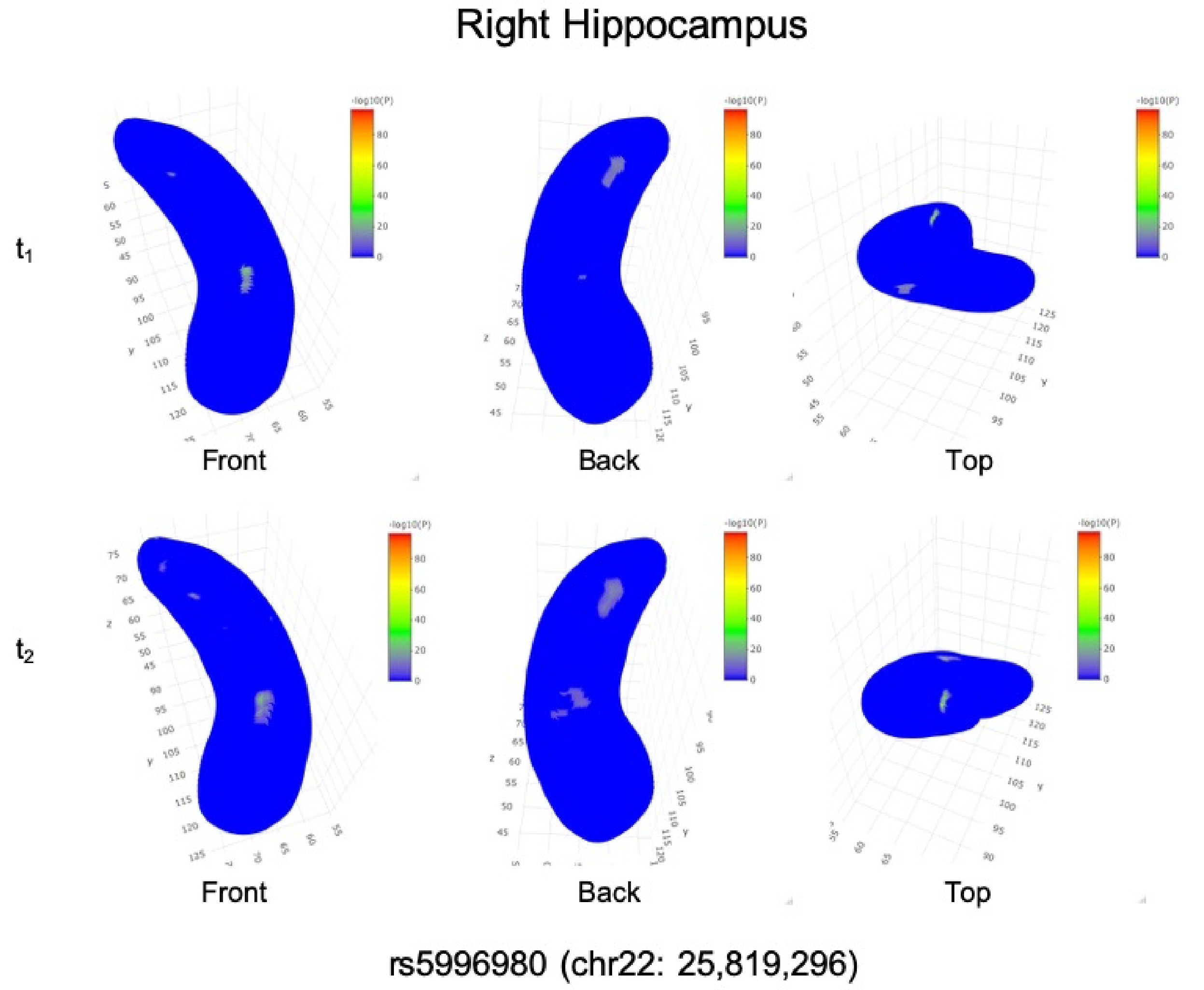


Figure 4.3: $-\log_{10}$(p-value) for significant regions (w.r.t t_1 and t_2, respectively) for the SNP with maximum dCor with the right hippocampus.

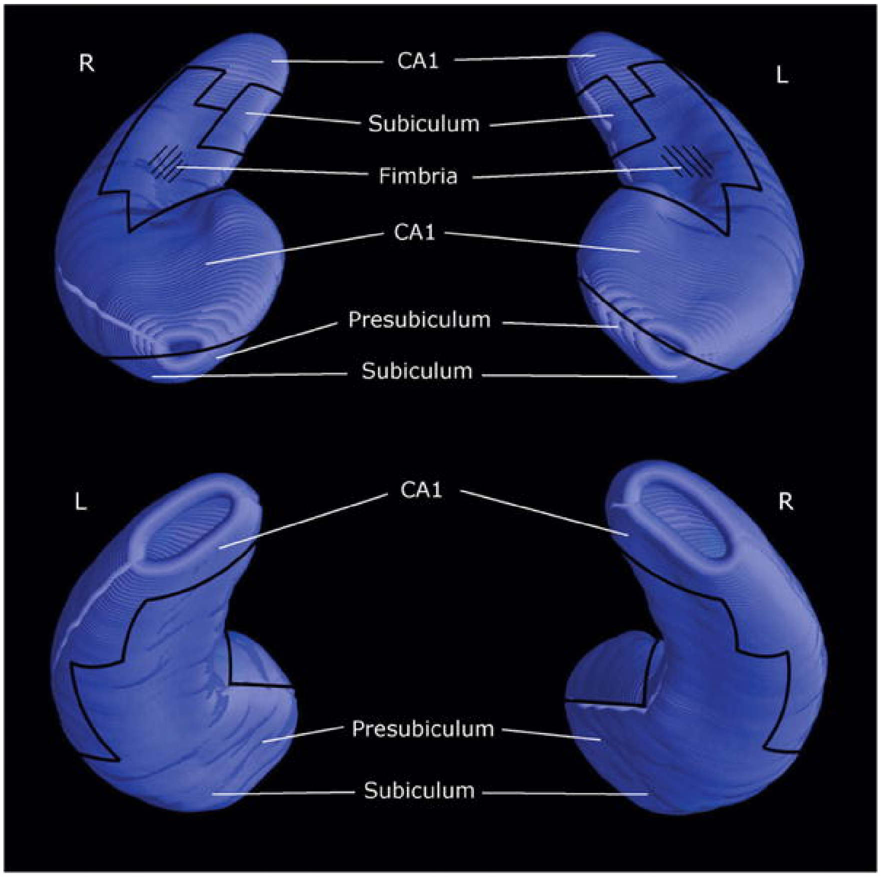


Figure 4.4: Subregions of the hippocampus surface.

constructed in APLS is nonorthogonal, we can find a novel method of generating orthogonal bases directly in the future research. We also can explore a more effective algorithm in comparison of FPLS-DC in the future.

CHAPTER 5: ENSEMBLE- VS MERGE-BASED LEARNERS

5.1 Introduction

Efforts in utilizing multi-study neuroimaging data extends from the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium (Thompson et al., 2014; Thompson and et. al., 2019) to the combination of recent large-scale studies such as UK Biobank (UKB) (Sudlow et al., 2015), the Adolescent Brain Cognitive Development (ABCD) Study (Casey et al., 2018), the Alzheimer's Disease Neuroimaging Initiative(ADNI) (Weiner et al., 2017), the Human Connectome Project (HCP) (Somerville et al., 2018), Philadelphia Neurodevelopmental Cohort (PNC) (Satterthwaite et al., 2016), and the Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository (Jernigan et al., 2016). These studies collect not only brain images but also genetic and other health-related information, providing valuable data for researchers to study brain structures and functions. Utilizing data from more than one studies can not only gain us power and prediction accuracy from increased sample size, but also help us identify more reliable factors and reach more general conclusions by accounting for inter-study heterogeneity. Inter-study heterogeneity may come from various sources such as differences in data collection centers, study environment, population composition (e.g., race), study design, data collection and processing protocol, and other study-specific factors (Zhang et al., 2020; Fortin et al., 2017; Leek and Storey, 2007; Mirzaalian et al., 2016). For the purpose of prediction, one can train the prediction model based on data from multiple studies using one of the two following strategies: 1) obtain the final trained learner (i.e. prediction model) by combining the learners each trained from one of the studies, or 2) train a single learner based on the merged data from all the studies.

In the first strategy, the final prediction is usually calculated as a weighted average of predictions made using learners trained on different studies, which is equivalent to making the final prediction using a learner, of which the parameters are weighted average of parameters from the different learners, if a linear regression model is used. Call this final learner *an ensemble-based learner*. This multi-study ensemble strategy was proposed by Patil and Parmigiani (2018) to leverage information from multiple studies using ensemble learning methods Dietterich (2000), which combine predictions from multiple models. This is similar to the combination of results for statistical inference, where meta-analysis average summary statistics from different studies by applying different weights (Cochran, 1954; DerSimonian and Laird, 2015; Jackson and Riley, 2014; Tang and Lin, 2014), and fusion learning (Cheng et al., 2017; Cai et al., 2020) combines the confidence distributions, rather than point estimates, for the parameter of interest from different studies.

The second strategy merge data from all studies and train a single learner based on the merged data. Call this learner *a merge-based learner*. Fixed-effect or random-effect models can be used to train the learner, where inter-study heterogeneity can be accounted for using methods such as analysis of large clusters (Bellamy et al., 2005), principal components analysis (Price et al., 2006), confounder adjusted testing and estimation (CATE) (Wang et al., 2017), and direct surrogate variable analysis (dSVA) (Lee et al., 2017). For neuroimaging data, Zhu et al. (2012); Huang et al. (2017) developed a fixed effects multivariate varying coefficient model (MVCM) to account for the smooth property of brain imaging data, Lin et al. (2014) developed a functional mixed effects model inspired by MVCM, Guillaume et al. (2018) modified CATE to better suit the analysis of neuroimaging data, and Huang et al. (2021) developed a functional hybrid factor regression model based on MVCM and modified CATE to account for inter-study heterogeneity.

Deciding between the two strategies is a critical decision for the final learner to achieve better prediction accuracy. The merge-based learner can sometimes perform better (Xu et al., 2008; Taminau et al., 2014; Kosch and Jung, 2019), and the ensemble-based learner can perform better in other cases (Bravata and Olkin, 2001), e.g. when the studies are heterogeneous (Patil and Parmigiani, 2018), while Lagani et al. (2016) found the two strategies comparable in reconstruction of gene interaction networks. The question of which strategy is expected to achieve better prediction accuracy is discussed by Guan et al. (2019), where learners were trained by modeles such as linear regression and ridge regression. Theorems derived therein provide a useful guideline for deciding which strategy to choose. A mixed effects model is assumed to be the data-generating model for heterogeneous studies. It is shown that merge-based learner yields lower prediction error than ensemble-based learner when heterogeneity is low; but as heterogeneity increases, there exists a transition point beyond which the ensemble-based learner outperforms the merge-based learner. The transition point was characterized analytically, and optimal ensemble weights are derived. The two strategies were also compared in the problem of hypothesis testing, where meta-analysis was compared with mega-analysis (Lin and Zeng, 2010a,b; Zeng and Lin, 2015).

In this chapter, we extended the guideline by Guan et al. (2019) to the brain imaging setting and derived similar guidelines. While the response variable for models in Guan et al. (2019) are real valued, brain imaging data is usually voxelwise and modeled as functional responses. We therefore use the model MVCM to train both the ensemble-based learner and the merge-based learner with data from more than one neuroimaging studies and compare their performance w.r.t. prediction accuracy. We derived the strategy-decision guideline for the MVCM learners in the method section (section 2), validated the theorems through simulation studies (section 3), and demonstrated our theoretical conclusions in neuroimaging studies (section 4).

5.2 Method

We use the following notations: denote numbers with lower-case letters, denote vectors with bold lower-case letters, and denote matrices with upper-case letters. " I_N is the $N \times N$ identity matrix, $0_{N\times M}$ is an $N \times M$ matrix of 0's, $\mathbf{0}_N$ is a vector of 0's with length N, $\mathbf{1}_N$ is a vector of 1's with length N, $tr(A)$ is the trace of matrix A, $diag(\boldsymbol{u})$ is a diagonal matrix with $\boldsymbol{u}$ along its diagonal, $(A)_{ij}$ is the entry in row i and column j of matrix A (Guan et al. 2019)", $A \otimes B$ is the Kronecker product of two matrices A and B, $||A||_F = \sqrt{\sum_{i,j} a_{ij}^2}$

the Frobenius norm of a matrix A, and $vec(A)$ is the vectorization of matrix A, i.e. for an $N \times M$ matrix $A = (a_{ij})$, $vec(A) = (a_{11}, \dots, a_{M1}, \dots, a_{1N}, \dots, a_{MN})^T$ is the stack of columns of A. $SP(\boldsymbol{\mu}(\cdot), \Sigma(\cdot, \cdot))$ denotes a stochastic process vector with mean function $\boldsymbol{\mu}(\cdot)$ and covariance function $\Sigma(\cdot, \cdot)$, $I\{\cdot\}$ is an indicator function, $B \circ C$ denotes Hadamard product of equal-dimension matrices B and C, i.e. $(B \circ C)_{ij} = (B)_{ij}(C)_{ij}$, and the bias of an estimator $\hat{\theta}$ of parameter θ is defined as $bias\{\hat{\theta}\} = E[\hat{\theta}] - \theta$. Notations used throughout this chapter are summarized at the end of this chapter in section 6.

First, I will present the MVCM model for one general study, as well as its varying-coefficient estimation and prediction (section 2.1). Then, we will discuss the scenario when we have multiple studies and lay out the underlying model that generates the multi-study data, with parameter estimation and prediction accuracy comparison introduced (section 2.2). Section 2.3 presents the theoretical conclusions for prediction error comparisons under different assumptions on the variances of the random effects, as well as the optimal ensemble weight. The theoretical details and proofs are relegated to the Appendix.

5.2.1 MVCM for One Study

For $i = 1, \dots, n$, $j = 1, \dots, J$, and $\boldsymbol{s} \in \mathcal{S}_0 \subset \mathrm{I\!R}^d$, where n is the sample size, J is the number of responses, and $\mathcal{S}_0$ is the common domain of all functional responses,

$$y_{ij}(\boldsymbol{s}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j(\boldsymbol{s}) + \eta_{ij}(\boldsymbol{s}) + \epsilon_{ij}(\boldsymbol{s}) \tag{5.1}$$

where $y_{ij}(\boldsymbol{s})$ is the jth functional response for subject i, $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^T$ are the covariates, and $\boldsymbol{\beta}_j(\boldsymbol{s}) = \big(\beta_{1j}(\boldsymbol{s}), \dots, \beta_{pj}(\boldsymbol{s})\big)^T$ are the functional coefficients. Denote $\boldsymbol{\eta}_i(\boldsymbol{s}) = \big(\eta_{i1}(\boldsymbol{s}), \dots, \eta_{iJ}(\boldsymbol{s})\big)^T$ and $\boldsymbol{\epsilon}_i(\boldsymbol{s}) = \big(\epsilon_{i1}(\boldsymbol{s}), \dots, \epsilon_{iJ}(\boldsymbol{s})\big)^T$, $\boldsymbol{\eta}_i(\cdot) \overset{i.i.d.}{\sim} SP(\mathbf{0}_J(\cdot), \Sigma_\eta(\cdot, \cdot))$ and $\boldsymbol{\epsilon}_i(\cdot) \overset{i.i.d.}{\sim} SP(\mathbf{0}_J(\cdot), \Sigma_\epsilon(\cdot, \cdot))$, where $\Sigma_\epsilon(\boldsymbol{s}, \boldsymbol{t}) = S_\epsilon(\boldsymbol{s}) I\{\boldsymbol{s} = \boldsymbol{t}\}$. $\eta_i(\cdot)$ and $\epsilon_{i'}(\cdot)$ are independent $\forall i, i' \in \{1, \dots, n\}$.

The above model can also be written in matrix form:

$$Y(\boldsymbol{s}) = XB(\boldsymbol{s}) + H(\boldsymbol{s}) + E(\boldsymbol{s}), \tag{5.2}$$

where $(Y(\boldsymbol{s}))_{ij} = y_{ij}(\boldsymbol{s})$, $(X)_{il} = x_{il}$, $(B(\boldsymbol{s}))_{lj} = \beta_{lj}(\boldsymbol{s})$, $(H(\boldsymbol{s}))_{ij} = \eta_{ij}(\boldsymbol{s})$, and $(E(\boldsymbol{s}))_{ij} = \epsilon_{ij}(\boldsymbol{s})$, for $i = 1, \ldots, n$, $j = 1, \ldots, J$, and $l = 1, \ldots, p$.

$\forall \boldsymbol{s} \in \mathcal{S}_0$, $\boldsymbol{\beta}_j(\boldsymbol{s})$ is estimated using the idea of least squares estimation, with smoothing done by a weighted sum of local information near $\boldsymbol{s}$ through a kernel function $K(|\boldsymbol{s} - \boldsymbol{s}'|)$, derived based on Taylor expansion of $\boldsymbol{\beta}_j(\cdot)$ near $\boldsymbol{s}$ (MVCM).

In collected data, $\mathcal{S}_0$ contains finite number of elements. Denote the number of elements in $\mathcal{S}_0$ as V; $\forall v \in \{1, \ldots, V\}$, denote $\boldsymbol{s}_v \in \mathcal{S}_0$. Define $R = X^T X$, the scaled kernel function for distance d and bandwidth h ($d, h \in \mathbb{R}^+$) being $K_h(d) = \frac{1}{h}K(\frac{d}{h})$, the 2×2 matrix $D_h(\boldsymbol{s}) = \sum_{v=1}^{V} K_h(|\boldsymbol{s}_v - \boldsymbol{s}|)\left(1, \frac{|\boldsymbol{s}_v - \boldsymbol{s}|}{h}\right)^T \left(1, \frac{|\boldsymbol{s}_v - \boldsymbol{s}|}{h}\right)$, the first-order smoothing coefficient $c_h(\boldsymbol{s}, \boldsymbol{s}_v) = K_h(|\boldsymbol{s}_v - \boldsymbol{s}|)(1, 0)D_h(s)^{-1}\left(1, \frac{|\boldsymbol{s}_v - \boldsymbol{s}|}{h}\right)^T$, the second-order smoothing coefficient $a_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) = c_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}) \cdot c_h(\boldsymbol{s}, \boldsymbol{s}_{v_2})$, and smoothed response $\tilde{Y}(\boldsymbol{s}) = \sum_{v=1}^{V} c_h(\boldsymbol{s}, \boldsymbol{s}_v) Y(\boldsymbol{s}_v)$. The estimator of the varying coefficient is as follows:

$$\hat{B}(\boldsymbol{s}) = R^{-1} X^T \tilde{Y}(\boldsymbol{s}). \tag{5.3}$$

Note that the original MVCM (ref), with equations laid out for $d = 1$, allows a unique bandwidth h_j for each response j. For simplicity, here we assume a common $h = h_j$ across all responses $j = 1, \ldots, J$. While FGWAS (ref) also extends the original MVCM (ref) to $d > 1$, they consider unique bandwidths $\{h_{jk}\}$, not only for each response j, but also for each dimension $k \in \{1, \ldots, d\}$, which constitute their bandwidth matrix H.

Let X_0 denote a design matrix, we then make prediction of its corresponding functional response $Y_0(\boldsymbol{s})$ as

$$\hat{Y}_0(\boldsymbol{s}) = X_0 \hat{B}(\boldsymbol{s}). \tag{5.4}$$

If we have both X_0 and $Y_0(\boldsymbol{s})$ available, then $\big(X_0, Y_0(\boldsymbol{s})\big)$ can serve as testing data to evaluate how well $\hat{Y}_0(\boldsymbol{s}) = X_0 \hat{B}(\boldsymbol{s})$ predicts $Y_0(\boldsymbol{s})$.

In order to measure the prediction accuracy, define the mean prediction error (MPE) on

$\mathcal{S} \subset \mathcal{S}_0$ as the square root of the mean squared errors:

$$\text{MPE}(\mathcal{S}) = \sqrt{\frac{\sum_{\boldsymbol{s}\in\mathcal{S}} \left\|Y_0(\boldsymbol{s}) - \hat{Y}_0(\boldsymbol{s})\right\|_F^2}{n_0 J|\mathcal{S}|}} \qquad (5.5)$$

$$= \sqrt{\frac{\sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{i=1}^{n_0} \sum_{j=1}^{J} \left[y_{0ij}(\boldsymbol{s}) - \hat{y}_{0ij}(\boldsymbol{s})\right]^2}{n_0 J|\mathcal{S}|}},$$

where $y_{0ij}(\boldsymbol{s}) = \left(Y_0(\boldsymbol{s})\right)_{ij}$, $\boldsymbol{x}_{0i}$ is the ith column of X_0^T, and $\hat{y}_{0ij}(\boldsymbol{s}) = \left(\hat{Y}_0(\boldsymbol{s})\right)_{ij} = \boldsymbol{x}_{0i}^T\hat{\boldsymbol{\beta}}_j(\boldsymbol{s})$. A higher MPE indicates lower prediction accuracy.

5.2.2 MVCM for Multiple Studies

Consider K comparable independent studies that measure the same functional outcomes and the same p predictors, and the datasets have been harmonized so that emasurements across studies are on the same scale (Guan et al. 2019). For study $k \in \{1,\dots,K\}$, let n_k denote the number of observations, $Y_k(\cdot)$ the functional response in study k, and X_k the design matrix. Assume the data is generated from the following mixed effects model

$$Y_k(\boldsymbol{s}) = X_k B(\boldsymbol{s}) + Z_k\Gamma_k(\boldsymbol{s}) + H_k(\boldsymbol{s}) + E_k(\boldsymbol{s}), \qquad (5.6)$$

where $B(\cdot)$ is the common varying coefficient of predictor effect shared among studies, $Z_k = X_k U \in \mathbb{R}^{n_k\times q}$, where $(U)_{ij} = I\{\text{the } i\text{th predictor is the } j\text{th random effect}\}$, consists of subcolumns of X_k, and $\Gamma_k(\cdot)$ is the study-specific effect that is assumed to be random for the q predictors in Z_k ($q \le p$): $\Gamma_k(\cdot) \sim SP(0_{q\times J}, G(\cdot,\cdot))$, where $0_{q\times J}$ is a $q \times J$ matrix of 0's, and $G(\boldsymbol{s},\boldsymbol{t})$ is a four dimensional tensor with $\left(G(\boldsymbol{s},\boldsymbol{t})\right)_{ll'jj'} = \text{Cov}\left[\left(\Gamma_k(\boldsymbol{s})\right)_{lj}, \left(\Gamma_k(\boldsymbol{t})\right)_{l'j'}\right]$ $\left(l, l' \in \{1,\dots,q\}; j, j' \in \{1,\dots,J\}\right)$. The noise terms $H_k(\boldsymbol{s})$ and $E_k(\boldsymbol{s})$ are assumed to inherit their assumptions from the original MVCM and independent between the K studies.

The above model in matrix format can be equivalently written as follows:

$$y_{kij}(\boldsymbol{s}) = \boldsymbol{x}_{ki}^T\boldsymbol{\beta}_j(\boldsymbol{s}) + \boldsymbol{z}_{ki}^T\gamma_{kj}(\boldsymbol{s}) + \eta_{kij}(\boldsymbol{s}) + \epsilon_{kij}(\boldsymbol{s}), \qquad (5.7)$$

where $\boldsymbol{x}_{ki}$ and $\boldsymbol{z}_{ki}$ are the ith rows of X_k and Z_k, respectively; $\boldsymbol{\beta}_j(\boldsymbol{s})$ and $\boldsymbol{\gamma}_{kj}(\boldsymbol{s})$ are the jth columns of $B(\boldsymbol{s})$ and $\Gamma_k(\boldsymbol{s})$, respectively; $y_{kij}(\boldsymbol{s}) = \big(Y_k(\boldsymbol{s})\big)_{ij}$, $\eta_{kij}(\boldsymbol{s}) = \big(H_k(\boldsymbol{s})\big)_{ij}$, and $\epsilon_{kij}(\boldsymbol{s}) = \big(E_k(\boldsymbol{s})\big)_{ij}$. Denote the $q \times q$ matrix $G_j(\boldsymbol{s},\boldsymbol{t}) = \mathrm{Cov}\big[\boldsymbol{\gamma}_{kj}(\boldsymbol{s}),\boldsymbol{\gamma}_{kj}(\boldsymbol{t})\big]$. Assume $G_j(\boldsymbol{s},\boldsymbol{t})$ is diagonal: $G_j(\boldsymbol{s},\boldsymbol{t}) = diag(\sigma^2_{j1}(\boldsymbol{s},\boldsymbol{t}),\ldots,\sigma^2_{jq}(\boldsymbol{s},\boldsymbol{t}))$.

Given data from multiple studies $\big(X_k, Y_k(\cdot)\big)$ $(k = 1,\ldots,K)$, the estimation of $B(\cdot)$ is commonly pursued through one of the two competing schemes:

1. **Ensemble**: estimate $B(\cdot)$ based on data from each study k:

$$\hat{B}_k(\boldsymbol{s}) = R_k^{-1} X_k^T \tilde{Y}_k(\boldsymbol{s}), \tag{5.8}$$

where $R_k = X_k^T X_k$, and combine all the K estimators

$$\hat{B}_e(\boldsymbol{s}) = \sum_{k=1}^{K} w_k \hat{B}_k(\boldsymbol{s}) \tag{5.9}$$

with $w_k \in (0,1), \forall k \in \{1,\ldots,K\}$ and $\sum_{k=1}^{K} w_k = 1$.

2. **Merge**: estimate $B(\cdot)$ based on merged data from the K studies:

$$\hat{B}_m(\boldsymbol{s}) = R_m^{-1} X_m^T \tilde{Y}_m(\boldsymbol{s}), \tag{5.10}$$

where $X_m = (X_1^T,\ldots,X_K^T)^T$, $R_m = X_m^T X_m$, and $Y_m(\boldsymbol{s}) = (Y_1(\boldsymbol{s})^T,\ldots,Y_K(\boldsymbol{s})^T)^T$;

The two estimation schemes above can each define a learner for prediction:

$$\hat{Y}_{0,e}(\boldsymbol{s}) = X_0 \hat{B}_e(\boldsymbol{s}), \tag{5.11}$$

$$\hat{Y}_{0,m}(\boldsymbol{s}) = X_0 \hat{B}_m(\boldsymbol{s}). \tag{5.12}$$

In order to compare the prediction accuracy of the learners, introduce a separate set of data, $\big(X_0, Y_0(\boldsymbol{s})\big)$ (for $\boldsymbol{s} \in \mathcal{S}_0$), as the testing set. For each learner, calculate the prediction accuracy in terms of MPE as defined in (5.5):

$$\mathrm{MPE}_e(\mathcal{S}) = \sqrt{\frac{\sum_{\boldsymbol{s}\in\mathcal{S}} \left\|Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s})\right\|_F^2}{n_0 J|\mathcal{S}|}}, \tag{5.13}$$

$$\mathrm{MPE}_m(\mathcal{S}) = \sqrt{\frac{\sum_{\boldsymbol{s}\in\mathcal{S}} \left\|Y_0(\boldsymbol{s}) - \hat{Y}_{0,m}(\boldsymbol{s})\right\|_F^2}{n_0 J|\mathcal{S}|}}. \tag{5.14}$$

Then, we can compare the prediction accuracy of the two learners $\hat{B}_m(\cdot)$ and $\hat{B}_e(\cdot)$ by comparing $\mathrm{MPE}_m(\mathcal{S})$ and $\mathrm{MPE}_e(\mathcal{S})$, which is equivalent to comparing

$$\sum_{\boldsymbol{s}\in\mathcal{S}} \left\|Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s})\right\|_F^2 \qquad \text{VS} \qquad \sum_{\boldsymbol{s}\in\mathcal{S}} \left\|Y_0(\boldsymbol{s}) - \hat{Y}_{0,m}(\boldsymbol{s})\right\|_F^2. \tag{5.15}$$

Therefore, it is of interest to investigate: under what condition does ensembling asymptotically give better prediction than merging:

$$E\left[\sum_{\boldsymbol{s}\in\mathcal{S}} \left\|Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s})\right\|_F^2\right] \quad \leq \quad E\left[\sum_{\boldsymbol{s}\in\mathcal{S}} \left\|Y_0(\boldsymbol{s}) - \hat{Y}_{0,m}(\boldsymbol{s})\right\|_F^2\right], \tag{5.16}$$

where $Y_0(\cdot)$ is treated as given and fixed and $Y_k(\cdot)$ as random, for $k = 0, 1, \ldots, K$.

5.2.3 Theoretical Conclusions

Denote the sum of smoothed residual errors $\sigma^2_{r,\mathcal{S}} = \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{v_1,v_2} a_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) tr\big(\Sigma_\eta(\boldsymbol{s},\boldsymbol{t}) + \Sigma_\epsilon(\boldsymbol{s},\boldsymbol{t})\big)$, the sum of smoothed variation for the lth ($l \in \{1, \ldots, q\}$) random effect $\sigma^2_{\cdot l,\mathcal{S}} = \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{v_1,v_2} a_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) \sum_{j=1}^J \sigma^2_{jl}(\boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2})$, and the mean variation of the q random effects (smoothed and summed over $\boldsymbol{s} \in \mathcal{S}$) $\sigma^2_{\mathcal{S}} = \frac{1}{q}\sum_{l=1}^q \sigma^2_{\cdot l,\mathcal{S}}$. Then, the comparison in (5.16) can be represented by the relationship between the two sources of variation in the estimated varying coefficient: the mean variation of the random effect $\sigma^2_{\mathcal{S}}$ and the variation of the residual terms $\sigma^2_{r,\mathcal{S}}$, as presented in the following two theorems.

For the simplicity of notations, we define the following terms: for $l = 1, \dots, q$,

$$
\begin{aligned}
b_0 &= \sum_{k=1}^{K} w_k^2 \; tr\Big(R_k^{-1} R_0\Big) - tr\big(R^{-1} R_0\big), \\
b_{1l} &= tr\Big(R_m^{-1} \sum_{k=1}^{K} X_k^T Z_k \boldsymbol{u}_l \boldsymbol{u}_l^T Z_k^T X_k R_m^{-1} R_0\Big) - \Big(\sum_{k=1}^{K} w_k^2\Big) \cdot \boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l, \\
b_2 &= tr\Big(R^{-1} \sum_{k=1}^{K} X_k^T Z_k Z_k^T X_k R^{-1} R_0\Big) - \left(\sum_{k=1}^{K} w_k^2\right) tr\big(Z_0^T Z_0\big), \\
\tau_{1a} &= \frac{b_0}{q \cdot \min_l b_{1l}}, \qquad \tau_{1b} = \frac{b_0}{q \cdot \max_l b_{1l}}, \qquad \text{and} \qquad \tau_2 = \frac{b_0}{b_2}.
\end{aligned}
$$

The comparison of prediction accuracy of the learners from ensembling and merging has the property in the theorems below. It is assumed that $b_{1l} > 0$ $(l = 1, \dots, q)$ for theorem 1 and $b_2 > 0$ for theorem 2. Conditions and proofs of b's > 0 are delegated to the Appendix.

Theorem 1 (a) Suppose $\min_l b_{1l} > 0$. A sufficient condition for

$$
E\left[\sum_{\boldsymbol{s} \in \mathcal{S}} \left\| Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s}) \right\|_F^2\right] \quad \leq \quad E\left[\sum_{\boldsymbol{s} \in \mathcal{S}} \left\| Y_0(\boldsymbol{s}) - \hat{Y}_{0,m}(\boldsymbol{s}) \right\|_F^2\right]
$$

is

$$
\sigma_{\mathcal{S}}^2 \geq \sigma_{r,\mathcal{S}}^2 \cdot \tau_{1a}. \tag{5.17}
$$

(b) Suppose $\max_l b_{1l} > 0$. A sufficient condition for

$$
E\left[\sum_{\boldsymbol{s} \in \mathcal{S}} \left\| Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s}) \right\|_F^2\right] \quad \geq \quad E\left[\sum_{\boldsymbol{s} \in \mathcal{S}} \left\| Y_0(\boldsymbol{s}) - \hat{Y}_{0,m}(\boldsymbol{s}) \right\|_F^2\right]
$$

is

$$
\sigma_{\mathcal{S}}^2 \leq \sigma_{r,\mathcal{S}}^2 \cdot \tau_{1b}. \tag{5.18}
$$

We therefore have the following asymptotic conclusion when equal ensemble weights are used:

Corollary 1: Suppose $w_k = \frac{1}{K}$ and there exist positive definite matrices $A_1, A_2, A_{(l)} \in \mathbb{R}^{p \times p}$ such that as $K \to \infty$,

1. $\frac{1}{K}\sum_{k=1}^{K} R_k \to A_1$

2. $\frac{1}{K}\sum_{k=1}^{K} R_k^{-1} \to A_2$

3. $\frac{1}{K}\sum_{k=1}^{K} X_k^T Z_k \boldsymbol{u}_l \boldsymbol{u}_l^T Z_k^T X_k \to A_{(l)}$ for $l = 1, \ldots, q$.

(a) If $\min_l \left\{ tr(A_1^{-1} A_{(l)} A_1^{-1} R_0) - \boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l \right\} > 0$, then

$$\tau_{1a} \to \frac{tr(A_2 R_0) - tr(A_1^{-1} R_0)}{\min_l \left\{ tr(A_1^{-1} A_{(l)} A_1^{-1} R_0) - \boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l \right\} q}. \tag{5.19}$$

(b) If $\max_l \left\{ tr(A_1^{-1} A_{(l)} A_1^{-1} R_0) - \boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l \right\} > 0$, then

$$\tau_{1b} \to \frac{tr(A_2 R_0) - tr(A_1^{-1} R_0)}{\max_l \left\{ tr(A_1^{-1} A_{(l)} A_1^{-1} R_0) - \boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l \right\} q}. \tag{5.20}$$

If the variance of the random effect $\gamma_{kjl}(\boldsymbol{s})$ remains the same across $l = 1, \ldots, q$, where $\gamma_{kjl}(\boldsymbol{s})$ is the lth element in the $q \times 1$ vector $\boldsymbol{\gamma}_{kj}(\boldsymbol{s})$, then $\sigma^2_{\cdot l, \mathcal{S}} = \sigma^2_{\mathcal{S}}$ for $l = 1, \ldots, q$.

Theorem 2: Suppose $\sigma^2_{\cdot l, \mathcal{S}} = \sigma^2_{\mathcal{S}}$ (for $l = 1, \ldots, q$) and $b_2 > 0$, then

$$E\left[\sum_{\boldsymbol{s} \in \mathcal{S}} \left\| Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s}) \right\|_F^2 \right] \quad \leq \quad E\left[\sum_{\boldsymbol{s} \in \mathcal{S}} \left\| Y_0(\boldsymbol{s}) - \hat{Y}_{0,m}(\boldsymbol{s}) \right\|_F^2 \right]$$

is equivalent to

$$\sigma^2_{\mathcal{S}} \geq \sigma^2_{r, \mathcal{S}} \cdot \tau_2. \tag{5.21}$$

Similar to Corollary 1, we have the following conclusion following from the above theorem:

Corollary 2: Suppose the random effects have equal variances $\sigma^2_{\cdot l, \mathcal{S}} = \sigma^2_{\mathcal{S}}$ (for $l \in \{1, \ldots, q\}$) and there exist positive definite matrices $A_1, A_2, A_3 \in \mathbb{R}^{p \times p}$ such that as $K \to \infty$,

1. $\frac{1}{K}\sum_{k=1}^{K} R_k \to A_1$

2. $\frac{1}{K}\sum_{k=1}^{K} R_k^{-1} \to A_2$

3. $\frac{1}{K}\sum_{k=1}^{K} X_k^T Z_k Z_k^T X_k \to A_3$

4. $tr(A_1^{-1}A_3A_1^{-1}R_0) - tr(Z_0^TZ_0) > 0$

where → denotes almost sure convergence. If we set $w_k = \frac{1}{K}$, then

$$\tau_2 \to \frac{tr(A_2R_0) - tr(A_1^{-1}R_0)}{tr(A_1^{-1}A_3A_1^{-1}R_0) - tr(Z_0^TZ_0)}. \tag{5.22}$$

Optimal Ensemble Weights: With the ensembled learner $\hat{B}_e(\boldsymbol{s}) = \sum_{k=1}^K w_k\hat{B}_k(\boldsymbol{s})$, it is of interest to derive the ensembling weights w_k $(k = 1,\ldots,K)$ that achieve the minimum expected prediction error

$$E\left[\sum_{\boldsymbol{s}\in\mathcal{S}} \left\|Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s})\right\|_F^2\right].$$

The optimal ensemble weights

$$w_{k,opt} = \frac{a_k}{\sum_{k=1}^K a_k}, \tag{5.23}$$

where

$$a_k = \left[tr(G_{\mathcal{S}}Z_0^TZ_0) + \sigma_{r,\mathcal{S}}^2 tr(R_k^{-1}R_0)\right]^{-1}. \tag{5.24}$$

5.3 Simulation

Simulation study is conducted to verify the theoretical transition points in Theorem 1 and Theorem 2, by simulating data with different levels of random effect variance $\sigma_{\mathcal{S}}^2$ and comparing $MPE_e(\mathcal{S})$ vs $MPE_m(\mathcal{S})$. We consider two scenarios: the general scenario corresponding to Theorem 1 and the homogeneous scenario corresponding to Theorem 2. 5000 replicated datasets are generated for each scenario with 10 levels of the variance for the random effects, which include the theoretical transition points and 0.

In particular, let the domain of the functional response $\mathcal{S}_0$ be a set of 50 equally-spaced real numbers from 0.02 to 1 by 0.02, and the region of interest $\mathcal{S} = \mathcal{S}_0$. Let $K(s) = \frac{3}{4}(1-s^2)I\{s \leq 1\}$ (the the Epanechnikov kernel) and bandwidth $h = 0.07$, which will include 3 points to the left and 3 points to the right of the target point for smoothing. The number of training studies $K = 4$, with $n_k = 100$ for $k = 1, 2, 3, 4$ and $n_0 = 400$. The number of predictors $p = 3$, with all elements in the first column of the design matrix X_k equal to 1. The number of random effects $q = 2$, with the second and third column of the design matrix corresponding to random effects, i.e. $U = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$. We use the same X_k's across all replicates, scenarios, and settings. Since the theoretical transition points are based on X_k, $\sigma^2_{r,\mathcal{S}}$, and h, the theoretical transition point will remain the same within each scenario. This will give us a guideline for choosing the values of $\sigma^2_{\mathcal{S}}$ that remain unchanged across replicates within each scenario. For the purpose of simplicity, we also use the same X_k's for both scenarios. Therefore, a total number of five sets of X_k's are simulated, $k = 0, 1, 2, 3, 4$. Since the first column of X_k's are 1's, we only need to generate the second and third columns of X_k's. We randomlly sample columns 2 and 3 from a bivariate normal distribution, with mean $(0, 0)^T$ and covariance matrix $\begin{pmatrix} 1 & r_k \\ r_k & 1 \end{pmatrix}$, for $k = 1, \dots, 4$. To introduce variation in the design matrix structure between training studies, let $r_1 = -0.75, r_2 = -0.25, r_3 = 0.25$, and $r_4 = 0.75$. To generate the testing study X_0, generate X_k's ($k = 1, 2, 3, 4$) one more time, and combine them as X_0.

For $s \in \mathcal{S}_0$, $Y_k(s)$'s are generated according to model (5.6). Therefore, in order to generate $Y_k(s)$, $B(s)$, $\Gamma_k(s)$, $H_k(s)$ and $E_k(s)$ are generated first. Let the number of responses $J = 1$ for the purpose of simplicity, then the dimensions of $G(s,t)$, $\Sigma_\eta(s,t)$ and $\Sigma_\epsilon(s,t)$ are now 2×2 , 1×1, and 1×1, respectively. For replicates $b = 1, \dots, 5000$, let $B_b(s)$ denote the varying coefficient $B(s)$ in replicate b. Allow $B_b(s)$ to vary across b. In particular, for $l = 0, 1, 2$, let $\beta_{lb}(s) = \big(B_b(s)\big)_{l1} = c_{lb} \cdot \beta_{l0}(s)$, where $c_{0b} \sim N(1, 0.05^2)$, $c_{1b} \sim N(1, 0.05^2)$, $c_{2b} \sim N(0.4, 0.02^2)$, and $\beta_{00}(s) = s^2$, $\beta_{10}(s) = (1-s)^2$, $\beta_{20}(s) = 4s(1-s)$. Denote $\eta_i(s) = \big(H_k(s)\big)_{i1}$. Let $\eta_i(s) = \xi_{i1}\psi_1(s) + \xi_{i2}\psi_2(s)$, where $\xi_{i1} \sim N(0, \lambda_1)$ and $\xi_{i2} \sim N(0, \lambda_2)$, with $\lambda_1 = 0.4$, $\lambda_2 = 0.2$, $\psi_1(s) = \sqrt{2}\sin(2\pi s)$, and $\psi_2(s) = \sqrt{2}\cos(2\pi s)$. Therefore, $\Sigma_\eta(s,t) = Cov\big(\eta_i(s), \eta_i(t)\big) = \lambda_1 2\sin(2\pi s)\sin(2\pi t) + \lambda_2 2\cos(2\pi s)\cos(2\pi t)$. Let $\epsilon_i(s) \sim N(0, 0.2)$. Therefore, $\sigma^2_\epsilon(s,t) = +0.2I\{s = t\}$. Then, $\sigma^2_{r,\mathcal{S}} = \sum_{s \in \mathcal{S}} \sum_{v_1, v_2} a_h(s, s_{v_1}, s_{v_2})\big(\Sigma_\eta(s,t) + \Sigma_\epsilon(s,t)\big)$ can be calculated accordingly. Let the random effect $\Gamma_k(s) = \Gamma_k$ remain unchanged across s. Therefore, the variance of the random effects becomes a constant matrix $G(s,t) = G = diag(\sigma_1^2, \sigma_2^2)$. For the homogeneous scenario, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. For the heterogeneous case, let $\sigma_1^2 = \frac{2}{3}\sigma^2$ and $\sigma_2^2 = \frac{4}{3}\sigma^2$ s.t. the mean random effect variance is still σ^2. For the homogeneous scenario, calculate the theoretical transition point $\sigma_t^2 = \sigma_{r,\mathcal{S}}^2 \cdot \tau_2$ according to Theorem 2, and set the 10 levels of σ^2 values as $(0,1,\ldots,9) \times \sigma_t^2/4$. For the heterogeneous scenario, two theoretical transition points σ_{ta}^2 and σ_{tb}^2 are calculated according to part (a) and part (b) of Theorem 1, respectively, with $\sigma_{tb}^2 \le \sigma_{ta}^2$; the 10 levels of $\sigma^2 = \frac{1}{2}(\sigma_1^2 + \sigma_2^2)$ values are set as $(0,1,\ldots,3) \times \sigma_{t1}^2/3$, $(\sigma_{t1}^2 + \sigma_{t2}^2)/2$, σ_{t2}^2, and $(1,\ldots,4) \times \sigma_{t1}^2/3 + \sigma_{t2}^2$, which include both transition points and 0. For studies $k = 1,\ldots,4$ and $b = 1,\ldots,5000$, generate $\boldsymbol{\gamma}_{bk}(s) = (\gamma_{1bk}(s), \gamma_{2bk}(s))^T$, where $\gamma_{1bk}(s) = \gamma_{1bk} \sim N(0,\sigma_1^2)$ and $\gamma_{2bk}(s) = \gamma_{2bk} \sim N(0,\sigma_2^2)$. Then, $Y_{bk}(s)$ is generated as

$$Y_{bk}(s) = X_k\beta_b(s) + Z_k\gamma_{bk}(s) + \eta_{bk}(s) + \epsilon_{bk}(s). \tag{5.25}$$

The above procedure is repeated to generate another four sets of Y for $k = 1,\ldots,4$, which are combined as the response for the test set.

For $b = 1,\ldots,5000$ under each of the 10 levels of σ^2 in each scenario (homogeneous and heterogeneous), $\hat{B}_{e,b}(s)$ and $\hat{B}_{m,b}(s)$ are estimated based on $\big(X_k, Y_{bk}(s)\big)$ (for $k = 1,\ldots,4$) according to (5.8), (5.9), and (5.10), with equal weights $w_k = \frac{1}{4}$ for $k = 1,2,3,4$. Then, prediction is made for the test set according to (5.11) and (5.12) for each estimated $\hat{B}_{e,b}(s)$ and $\hat{B}_{m,b}(s)$, with prediction errors $MPE_e(\mathcal{S})$ and $MPE_m(\mathcal{S})$ calculated following (5.13) and (5.14) accordingly. For each replicate b under each σ^2 in each scenario, $\log \frac{MPE_e^2(\mathcal{S})}{MPE_m^2(\mathcal{S})}$ is calculated. Then, a sample means of $\log \frac{MPE_e^2(\mathcal{S})}{MPE_m^2(\mathcal{S})}$ over every 100 replicates is calculated to estimate the expected MPE's in the theorem, and the boxplot of the 50 averaged $\log \frac{MPE_e^2(\mathcal{S})}{MPE_m^2(\mathcal{S})}$ is plotted for each σ^2 in each scenario in Figure 1. The theoretical transition points on σ^2, calculated as $\frac{\sigma_{r,\mathcal{S}}^2 \cdot \tau}{\sum_{s,s_{v_1},s_{v_2}} a(s,s_{v_1},s_{v_2})}$, where $\tau = \tau_{1a}, \tau_{1b}$, or τ_2, are indicated by the red dashed lines.

We can see from Figure 1 that, in both the plot for homogeneous scenario and the plot for heterogeneous scenario, the averaged $\log \frac{MPE_e^2(\mathcal{S})}{MPE_m^2(\mathcal{S})}$ decreases as σ^2 increases. Since a smaller

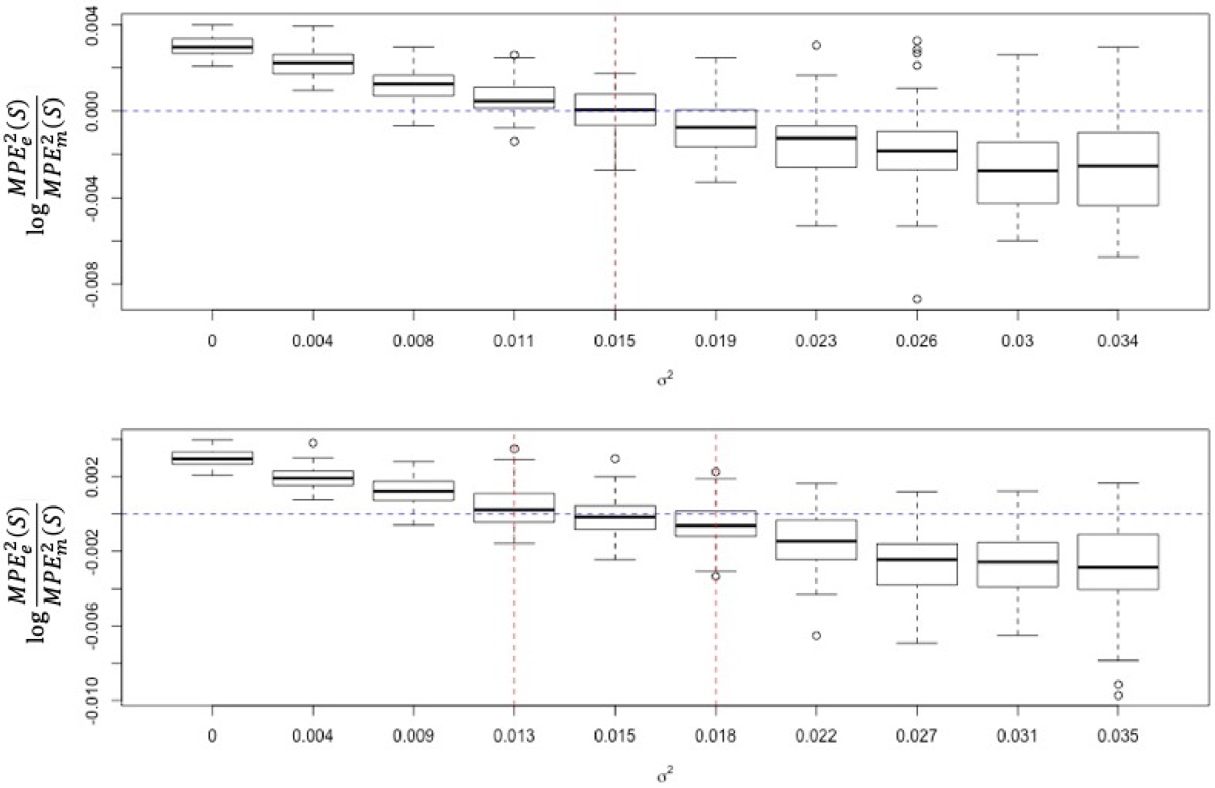


Figure 5.1: Boxplots of means of $\log \frac{MPE_e^2(\mathcal{S})}{MPE_m^2(\mathcal{S})}$ over 100 replicates for homogeneous scenario (top) and heterogeneous scenario (bottom). The theoretical transition points are indicated by the red dashed lines.

$MPE(\mathcal{S})$ indicates better prediction, this means that the ensemble-based estimator $\hat{B}_e(s)$ gradually over-performs the merge-based estimator $\hat{B}_e(s)$ in prediction accuracy as the random effect variance σ^2 increases. The empirical transition point, indicated by $\log \frac{MPE_e^2(\mathcal{S})}{MPE_m^2(\mathcal{S})} = 0$, i.e. $MPE_e(\mathcal{S}) = MPE_m(\mathcal{S})$, coincides with the theoretical transition point indicated by the red dashed line in the homogeneous scenario (top plot in Figure 1). While in the heterogeneous scenario, transition from merge-based learner gives better performance to ensemble-based learner gives better performance is to happen in the interval $\left[\sigma_{r,\mathcal{S}}^2 \cdot \tau_{1b}, \sigma_{r,\mathcal{S}}^2 \cdot \tau_{1a}\right]$ according to Theorem 1. This theoretical transition point coincides with the empirical transition point indicated by the boxplot with $\log \frac{MPE_e^2(\mathcal{S})}{MPE_m^2(\mathcal{S})}$ near 0. Hence, this simulation study verifies our theoretical conclusions derived in Theorem 1 and Theorem 2.

5.4 Imaging Genomic Application

In this section, we illustrate a practical example with brain imaging genetic datasets used for learners training and testing, where genetic information are used as predictors and voxelwise imaging feature as the functional response. The prediction accuracy of merge-based and ensemble-based learners are compared both by calculating the actual MPE's on the testing data, and through comparing their expected MPE's utilizing the general-case conclusion in Theorem 1, by estimating related terms therein from data. This allows us to take a glance at whether the theory-based guidance indicates a better multi-study learner (ensemble v.s. merging) in a real-world example. To compare the performance of the learners in the presence of potential between-study variation, we consider the following two scenarios:

A. train the learners using different subsets of the same study and test the learners on a held out subset;

B. train the learners using different studies and test them on an independent study.

After training and testing the learners, the prediction accuracy is compared between merge- and ensemble-based learners, with the theoretical transition point estimated and referenced.

5.4.1 Data Description

We consider the following three studies for this illustration: UK Biobank (UKB) (Sudlow et al., 2015), which includes elder individuals aged from 45 to 80, Adolescent Brain Cognitive Development (ABCD) (Casey et al., 2018), which includes adolescents aged from 9 to 11, and the Human Connectome Project (HCP) (Somerville et al., 2018), which includes young adults aged from 22 to 35. Besides age range, the variation between the three studies may also come from their differences in image quality, protocol implementations, etc.

The functional imaging response data used in this application is introduced as follows. In recent brain imaging-genetic studies (Zhao et al., 2019, 2020, 2019, 2020, 2019), brain white matter connectivity is shown to present more signals of significant association with genetic markers (Zhao et al., 2020) as compared to other commonly considered brain imaging features such as cortical volume in subregions of the brain. Therefore, we target on diffusion magnetic resonance images (dMRIs), which capture brain white matter microstructural connectivity that can be quantified by diffusion tensor imaging (DTI) models (Basser et al., 1994). Among all the DTI-derived parameters, the fractional anisotropy (FA), highly sensitive to general connectivity changes, is a feature of interest in many studies (Grieve et al., 2007). A recent recent study investigated the genome-wide association with the top five imaging PCs of multiple DTI-derived parameters in 21 pre-defined brain white matter tracts (Zhao et al., 2020) generated from the ENIGMA-DTI pipeline (Jahanshad et al., 2013; Kochunov et al., 2014). Among all the parameters and white matter tracts investigated, FA in genu of corpus callosum (GCC) is significantly linked to the maximum number of SNPs. Therefore, we target on GCC FA as our imaging functional response variable in this application. We used the GCC FA data from Zhao et al. (2020) for the three studies introduced above, which is measured on the functional domain $\mathcal{S}_0$ that contains $V = 1834$ voxels in a 3D space ($\mathcal{S}_0 \subset \mathbb{R}^3$). The number of responses $J = 1$, and functional response values $Y_{ki}(\boldsymbol{s}) := Y_{ki1}(\boldsymbol{s}) \in (0,1) \subset \mathbb{R}$, for $\boldsymbol{s} \in \mathcal{S}_0$, $i = 1, \dots, n_k$, and $k = 0, 1, \dots, K$.

The predictor data is described as follows: SNPs rs12653308 and rs2237077 are selected as predictors because they are significantly associated with FA through both GWAS with the top five GCC FA PCs and voxelwise regression using MVCM with voxelwise GCC FA, besides being mutually uncorrelated and available in all the three studies introduced above. The SNPs are modeled as a quantitative variable, with value equal to its number of alternative alleles (0, 1, or 2) (i.e. the additive model is implemented w.r.t. each SNP). The number of alternative alleles are extracted from imputed genotype data, with ABCD and HCP genotype data imputed (Zhao et al., 2020) using the 1000 Genomes reference panel and UKB genotype data imputed using Haplotype Reference Consortium and UK10K + 1000 Genomes reference panels (Bycroft et al., 2018). Age and gender were initially considered as predictors, as well, but finally dropped because of their estimated MVCM varying coefficients being close to zero across $\boldsymbol{s} \in \mathcal{S}_0$.

Individual filtering in the three study cohorts are implemented for the following aspects. For all the three studies mentioned above, their original cohorts include siblings or families. Therefore we filter for only independent individuals to be included into our analysis. In particular, ABCD is know to have twins in the cohort, therefore, a random individual from each pair of twins is excluded. The HCP cohort include both twins and non-twin siblings, and the mother and father IDs are available for each individual. Therefore, individuals are filtered so that non of the individuals included share the same mother or father. General family structure is possible in the UKB cohort, and genetically-derived kinship coefficient is released with the genotype data for each pair of individuals that are related up to the third degree. Therefore, we group individuals related up to the third degree into families and selected a largest set of unrelated individuals in each family for inclusion using the igraph R package (Csardi and Nepusz, 2006) following the procedure by Bycroft et al. (2018). Furthermore, we also filtered for individuals based on their self-reported race. For HCP and ABCD, only individuals with European ancestry are included, and only British individuals are included for UKB. After further excluding the individuals with missing data, the final number of individuals included for HCP, ABCD and UKB are 298, $5,088$ and $16,703$, respectively.

5.4.2 Study Design and Implementation

Using data described above, we apply the following procedure to each of the two scenarios:

1) estimate the variances of random effect and noise terms using linear mixed model, and calculate the optimal weights and the terms in Theorem 1 accordingly;

2) train the merge-based and ensemble-based learners to predict the functional response in the testing data, and calculate the MPE corresponding to each learner;

3) repeat step 2) on 100 bootstrap samples.

Each of the above steps are elaborated as follows:

Step 1) makes use of both the training datasets and the testing dataset, which are designated as follows: for scenario A, randomly split the UKB cohort into five subsets with approximately equal sizes, with $K = 4$ sets of size $n_k = 3341$ $(k = 1,2,3,4)$ used for training and one set of size $n_0 = 3339$ for testing; for scenario B, use ABCD $(n_1 = 5,088)$ and UKB $(n_2 = 16,703)$ for training $(K = 2)$ and HCP for testing $(n_0 = 298)$. In this step, the training sets and testing set are treated equally within each scenario. Our goal is to estimate the variance of the random effects G and the variance of the noise terms $\Sigma_\eta + \Sigma_\epsilon$ in model (5.6), or equivalently (5.7), which are needed to calculate $\sigma^2_{\mathcal{S}}$, $\sigma^2_{r,\mathcal{S}}$, τ_{1a} and τ_{1b} in Theorem 1, which provides insights on whether ensemble or merging is better for prediction without putting homogeneous assumption on the random effect variance. For the ensemble-based learner, we would like to combine study-specific learners (for $k = 1,\ldots,K$) based on both equal weights $w_k = \frac{1}{K}$ and the optimal weights $w_{k,opt}$ as in (5.23) and (5.24). The calculation of the optimal ensemble weights $w_{k,opt}$ are also based on the estimated G and $\sigma^2_{r,\mathcal{S}}$.

We estimate $G(\boldsymbol{s},\boldsymbol{t})$, $\Sigma_\eta(\boldsymbol{s},\boldsymbol{t})$ and $\Sigma_\epsilon(\boldsymbol{s},\boldsymbol{t})$ for $\boldsymbol{s},\boldsymbol{t} \in \mathcal{S}_0$ using the following strategy. Since the MVCM coefficient estimator $\hat{B}(\boldsymbol{s}) = R^{-1}X^T\tilde{Y}(\boldsymbol{s})$ in (5.3) takes the form of the coefficient estimator in a linear regression model with the smoothed response $\tilde{Y}(\boldsymbol{s})$, defined above equation (2.3), as the response variable, we estimate $G(\boldsymbol{s},\boldsymbol{t})$ and $\Sigma_\eta(\boldsymbol{s},\boldsymbol{t})$ for model (5.6), or equivalently model (5.7), by fitting a mixed regression model at each voxel $\boldsymbol{s} \in \mathcal{S}_0$ with the response being $\tilde{Y}(\boldsymbol{s})$. The smoothed response $\tilde{Y}(\boldsymbol{s})$ is obtained using bandwidth h equal to the median of the Euclidean distances between each voxel and its closest voxel: $h = \text{median}_{\boldsymbol{s}_1} \min_{\boldsymbol{s}_2} ||\boldsymbol{s}_1 - \boldsymbol{s}_2||$, and the Epanechnikov kernel function $K_h(x) = \frac{1}{h}K(\frac{x}{h})$, $\forall x \in \mathbb{R}$, so that adjacent voxels of each target voxel is considered during the smoothing procedure. At each voxel $\boldsymbol{s}$, the following linear mixed model is fit using the "lmer" function in the *lme4* R package (Bates et al., 2015):

$$\tilde{y}_{ki}(\boldsymbol{s}) = \boldsymbol{x}_{ki}^T \boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{z}_{ki}^T \boldsymbol{\gamma}_k(\boldsymbol{s}) + \eta_{ki}(\boldsymbol{s}), \tag{5.26}$$

for $i = 1, \dots, n_k$ and $k = 1, \dots, K$, where the response indicator j is dropped as compared to model (5.7) in our case of $J = 1$, $\boldsymbol{x}_{ki} = \boldsymbol{z}_{ki}$ is a 3×1 vector ($p = q = 3$), with the first element equal to 1 for intercept and the second and third elements as the number of alternative alleles for SNPs rs12653308 and rs2237077, respectively, and $\eta_{ki}(\boldsymbol{s})$ is the individual-specific variation that are not explained by the predictors. The final measurement error term $\epsilon_{ki}(\boldsymbol{s})$ is not included in this model because of the cancellation of the measurement error by the smoothing procedure, under the assumption that the functional imaging response and the varying coefficients are smooth by nature. Therefore, estimate $\epsilon_{ki}(\boldsymbol{s})$ by $y_{ki}(\boldsymbol{s}) - \tilde{y}_{ki}(\boldsymbol{s})$ and estimate $\Sigma_\epsilon(\boldsymbol{s})$ as the sample variance of $\{\epsilon_{ki}(\boldsymbol{s})\}_{k,i}$. With this estimating strategy, $G(\boldsymbol{s}, \boldsymbol{t})$ is assumed to take the form $G(\boldsymbol{s}, \boldsymbol{s}) I\{\boldsymbol{s} = \boldsymbol{t}\}$, i.e. $Cov(\boldsymbol{\gamma}(\boldsymbol{s}), \boldsymbol{\gamma}(\boldsymbol{t})) = 0$. $\hat{G}(\boldsymbol{s}, \boldsymbol{s})$ and the residual $\hat{\eta}_{ki}$ are extracted from the "lmer" function output, then $\Sigma_\eta(\boldsymbol{s}, \boldsymbol{t})$ is estimated using the sample covariance of $\hat{\eta}_{ki}$'s.

After obtaining $\hat{G}(\boldsymbol{s}, \boldsymbol{s})$, we calculate $\hat{\Sigma}_\eta(\boldsymbol{s}, \boldsymbol{t})$ and $\hat{\Sigma}_\epsilon(\boldsymbol{s}, \boldsymbol{s})$ accordingly. Their values in each scenario are shown as follows:

		Scenario A	**Scenario B**
Training Data		4 UKB subsamples ($n_k = 3341$)	ABCD ($n_1 = 5088$) UKB ($n_2 = 16,703$)
Testing Data		a UKB subsample ($n_0 = 3339$)	HCP ($n_0 = 298$)
$\hat{\sigma}^2_{\mathcal{S}}$		1.31×10^{-3}	0.569
$\hat{\sigma}^2_{r,\mathcal{S}}$		4.101	4.364
$w_k = \frac{1}{K}$	$\hat{\sigma}^2_{r,\mathcal{S}} \cdot \tau_{1a}$	1.83×10^{-6}	4.13×10^{-4}
	$\hat{\sigma}^2_{r,\mathcal{S}} \cdot \tau_{1b}$	8.14×10^{-7}	2.16×10^{-4}
$w_{k,opt}$	$w_{k,opt}$	$w_{1,opt} = 0.250, w_{2,opt} = 0.249$ $w_{3,opt} = 0.250, w_{4,opt} = 0.251$	$w_{1,opt} = w_{2,opt} = 0.500$
	$\hat{\sigma}^2_{r,\mathcal{S}} \cdot \tau_{1a}$	1.77×10^{-6}	4.10×10^{-4}
	$\hat{\sigma}^2_{r,\mathcal{S}} \cdot \tau_{1b}$	7.84×10^{-7}	2.15×10^{-4}

From the above table, we can see that $\hat{\sigma}^2_{\mathcal{S}}$ is greater than $\hat{\sigma}^2_{r,\mathcal{S}} \cdot \tau_{1a}$ for both scenarios, no matter using equal ensemble weights or optimal ensemble weights. According to Theorem 1, ensemble-based learner yields a smaller expected MPE, i.e. better prediction accuracy to be expected given testing data of which the study-specific coefficient has expectation 0. Therefore, we move on to Step 2) to calculate the MPE's from the learners.

Step 2) - 3) For each scenario, estimate $\hat{B}_m(\boldsymbol{s})$ according to (5.10) using merged data across all training datasets and $\hat{B}_k(\boldsymbol{s})$ according to (5.8) for each training dataset $k = 1, \dots, K$. Then, according to (5.9), calculate $\hat{B}_e(\boldsymbol{s})$ using equal ensemble weights $w_k = \frac{1}{K}$ and $\hat{B}_{e,opt}(\boldsymbol{s})$ using optimal ensemble weights $w_{k,opt}$ that are obtained in step 1). Lastly, use the estimated varying coefficients $\hat{B}_m(\boldsymbol{s})$, $\hat{B}_e(\boldsymbol{s})$ and $\hat{B}_{e,opt}(\boldsymbol{s})$ to predict Y_0 based on predictor data X_0 in the testing set and calculate the corresponding MPE's following (5.13) and (5.14). Repeat step 2) for 100 bootstrap samples, with each bootstrap sample randomly drawn with replacement from the original training data. The testing data remain unchanged.

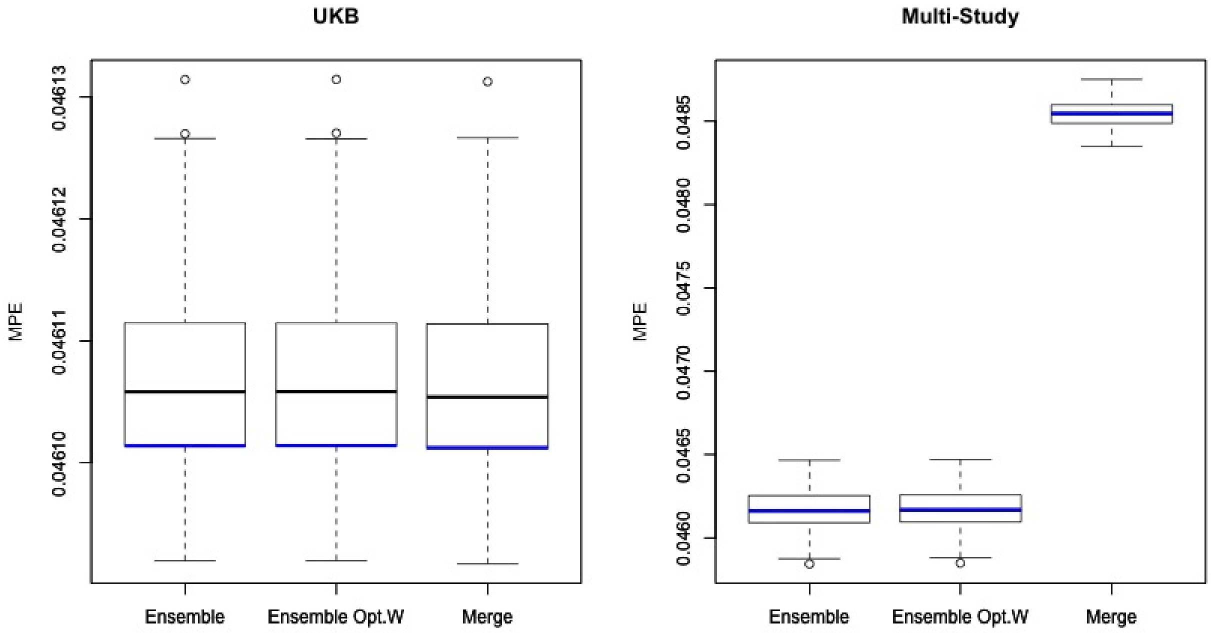


Figure 5.2: Boxplots of 100 bootstrap MPE's, with original MPE's indicated by the blue line segments. The boxplot on the left is for scenario A, and the boxplot on the right is for scenario B.

5.4.3 Result

For each scenario, the bootstrap MPE's calculated in step 3) are summarized into boxplots in Figure 2, with MPE's calculated in step 2) drawn as blue line segments on those boxplots. From Figure 2, we can see that for scenario A, we have the three similar boxplots of MPE's for ensemble-based learner using equal weights, ensemble-based learner using optimal weights and merge-based learners. The similarity between the three boxplots can be explained by the similarity of signals $B(\cdot)$ within the UKB cohort. This similarity can be explained by the small $\hat{\sigma}^2_{\mathcal{S}} = 1.31 \times 10^{-3}$ calculated for scenario A in step 1), as compared to $\hat{\sigma}^2_{\mathcal{S}} = 0.569$ for the three different studies in scenario B. For the comparison of expected MPE's for scenario A, although $\hat{\sigma}^2_{\mathcal{S}}$ is greater than $\hat{\sigma}^2_{r,\mathcal{S}} \cdot \tau_{1a}$, which guides the preference in ensemble-based learners as indicated by Theorem 1, we notice that both sides of the inequality (5.17) are very close to 0. For scenario B, we can see from Figure 2 that ensemble-based learners give better performance than merge-based learners. This aligns with the preference in the ensemble-based learners according to Theorem 1, since $\hat{\sigma}^2_{\mathcal{S}}$ is greater than $\hat{\sigma}^2_{r,\mathcal{S}} \cdot \tau_{1a}$. The similarity between the ensemble based learners with equal weights v.s. optimal weights can be explained by the optimal weights being very close to 0.5 (the $w_{k,opt}$ values listed in the table above are rounded to the third decimal place).

5.5 Discussion

We extended Guan et al. (2019) 's guideline of whether the ensembled learner or merged learner gives better prediction given data from multiple studies to the scenario of brain imaging data modeled by MVCM (Zhu et al., 2012). Transition points of whether ensembled learner or merged learner gives better prediction are derived allowing unequal variances for the random effects (Theorem 1) and assuming equal variances across all random effects (Theorem 2). We estimated the transition points based on simulated data under different variances of the random effects and observed that the MPE comparisons match the guideline provided by the theorems. As random effect variances increase, the ensembled learner gives increasingly better prediction. We then tried to apply the learner-choosing guideline to brain imaging data and examined guideline comparing to learner performances on a testing cohort, The comparison results align with our derived theoretical conclusions, demonstrating the usefulness of the learner-choosing guideline for imaging data using MVCM. When using the guideline given by the theoretical conclusions, please take caution that when the predictors in the model tend to have no signal or weak signals, the theoretical conclusion may not align with actual behavior due to noise.

Notice that the transition points τ_{1a}, τ_{1b} and τ_2 is not invariant with regard to sample size or the scale of the predictors in X_k. Therefore, one should take caution to harmonize the predictors to the same scale across the K cohorts. With regard to the transition point varying with sample size, one may understand the transition point as a comparison between ensembled learner and merged learner on two terms: A) variance of the random effect, i.e. between-study variation, and B) variance of the error terms, i.e. within-study variation. The above two terms are the source of variation in $hatB(\boldsymbol{s})$, and hence $\hat{Y}(\boldsymbol{s})$ and MPE. Given the same testing data, term A is invariant w.r.t. sample size and term B decreases as sample sizes increase. As a result, the transition point decreases with larger sample size in this case. An illustration can be seen in the top row of Figure 3. When the sample size is infinitely large, $\hat{B}(\boldsymbol{s}) \rightarrow B(\boldsymbol{s})$, and the comparison (5.16) reduces to the comparison of terms A_e versus A_m, i.e. the sign of d_{1l} or d_2, which are assumed to be positive. So, the assumption of d_{1l} or $d_2 > 0$ can be interpreted as the ensembled learner is always better than the merged learner given infinite sample size. Therefore, the comparison of the MPE's (5.16) is in essence the balance between the comparisons of terms A (for between-study variation) and terms B (for within-study variation), given finite sample size. In the appendix, a simple simulation study is conducted to observe the trend of the transition points as sample size increases.

In this chapter, the theorems give conclusions on the comparison of the expected MPE's. This is different from the expectation of the differences between the MPE's since MPE_e and MPE_m depend on data from the same studies or cohorts. If the difference between the MPE's are examined w.r.t. its distribution, e.g. expectation, variance, and the approximated distribution family, a statistical test would be possible to compare the prediction accuracy between the two families of learners. As this chapter focuses on the problem of prediction, given multiple studies, an equally interesting problem at need is statistical testing, or meta-analysis, in the realm of brain imaging data, which is a potential direction for future research.

CHAPTER 6: DISCUSSION

Future research with regard to each of the three projects is described as follows.

6.1 Future Research for Project 1

In future research, questions such as how to determine an optimal window (i.e. subregion) can be discussed, including the optimal size and shape of the subregion. On the other hand, we can explore how to define overlapping subregions on a non-regular closed surface, such as the hippocampus surface or surface of other ROIs in the brain. Data on the hippocampus surface in the real data application was reshaped to a 150×100 rectangle. The limitation is that the voxels on the two opposite sides of the rectangle do not have a chance to be included into the same subregion, and relative distances between voxels cannot be retained in the reshaped rectangle. For example, voxels on the top and bottom of the hippocampus may have to lay out in different subregions along the top and bottom edges of the rectangle. If we can design a strategy to move a circled or rectangular region across the surface of the ROI, not only those problems will be resolved, maintaining the relative position and distances between voxels, but the step of registering voxels on a non-regular closed shape to a rectangle on a flat surface can also be saved. Furthermore, the question rises as whether we can define subregions in the image based on the correlations between adjacent voxels, s.t. nearby voxels with their corresponding measurements closely related to each other are more likely to be grouped into the same subregion. This conduct gives the possibility to define more biologically meaningful subregions, and can potentially enhance computation and performance.

In order to further reduce computation, we can group nearby SNPs into blocks based on their correlations, and perform test of associations only on the representing SNP in each block. We can explore method to infer the image-SNP association of non-representing SNPs based on the inter-SNP correlation and the association between image and the representing SNP in the corresponding SNP block. Since tests such as dCov and BCov allow multivariate variables, there is also a possibility to screen the SNPs using multiple SNPs at a time, or even multiple, if not all, image subregions at one time. For example, for all the SNPs on each chromosome, split them into two halves, and perform a BCov test on each group of the SNPs. Assume that a BCov test on a group of SNPs would be significant if a significant SNP is included in the group of SNPs tested, then one can zoom into this group, e.g. split this group of SNPs into smaller groups until a fine-enough group (e.g. defined based on inter-SNP correlations) or a single SNP is detected significant enough for inclusion in the screening.

To fully address the problem of local alignment and registration error and how the proposed method and future research can address this issue, theoretical discussions can be pursued in future research.

6.2 Future Research for Project 2

Rather than using distance covariance for coefficient optimization and global test, other measures such as ball covariance can be used instead. If possible, we can consider extracting bases s.t. they maximize distance covariance or ball covariance, instead of maximizing Pearson's covariance as aimed by APLS. Bootstrap-based statistical tests will give more rigorous p-values, although more computationally intensive, and can be potentially used when computation power evolve. Otherwise, it is of interest to develop computationally efficient tests that are more rigorous. The cluster-size analysis as implemented by Huang et al. (2017) can also be implemented on $\hat{b}(s)$ to detect significantly large regions with signals above a certain threshold. Rather than doing linear combination of the bases, a multivariate test for dCov or BCov on all extracted bases can be implemented to investigate their association with the SNP. The multivariate tests such as dCov and BCov tests also allow the testing of multiple markers at the same time. Rather than using the dosage value of the SNP, other type of response values and other variables of interest can also fit in this framework, such as gene expression levels. Additionally, the subregion-based screening strategy can also be applied in this framework.

Other questions for future investigation are as follows: 1) Can registration error be accounted for similar to project 1? 2) How is it useful to reduce the image dimension w.r.t. each variable of interest through APLS? In genomic studies, genetic PCs can be used to adjust for population structure. Can the APLS-extracted bases also be used to adjust for structural effects that are related to images, such as in causal-inference models?

6.3 Future Research for Project 3

In future research, other models besides MVCM can be considered. Empirical conclusions can even be generated for deep learning methods. Besides prediction, statistical inference can also be discussed to compare meta- v.s. mega-analysis.

APPENDIX 1: METHOD 1 - ACCELERATED BCOR

Calculating the Ball Covariance (BCov) and the corresponding p-value can be computationally expensive. It took approximately one week to run BCov-based GWAS on one phenotype from UK Biobank (sample size $n \sim 10,000$) using the R package "ballgamma". This GWAS calculates a BCov-based p-value for each of the 10 million genetic markers. The run time is based on running the full-load number of parallel jobs (100+) on the Longleaf University cluster. The implementation of rfGWAS on the hippocampus surface includes $126 \times 2 = 252$ phenotypes. It would already be computationally challenging to implement GWAS on one region of interest in the brain, not to say the implementation of this algorithm to the entire brain and the full imaging sample size of 40,000. The main computation burden of rfGWAS comes from the repeated computation of BCov for each genetic marker, of which the computation complexity is $\mathrm{O}(n^2 \log n)$. This makes the implementation of rfGWAS non-realistic, especially on a large cohort such as UK Biobank. In the case of GWAS, where a genetic marker is recorded by the number of alternative alleles (0, 1, or 2) for an individual, and the phenotype of interest remains the same across all genetic markers, the computation of BCov has a great potential to be simplified. For the purpose of SNP screening, we only need to rank the BCov statistic, without needing to obtain the p-value corresponding to the BCov calculated for each SNP, which further decreased the computation burden of rfGWAS. In this section, we give the simplified calculation of BCov for SNP screening.

Empirical BCov

We use the empirical BCov (Pan et al. 2019) to estimate the relationship between two random variables X and Y. Define B-valued random variables (X, Y) on a probability space such that $(X, Y) \sim \theta$, $X \sim \mu$, and $Y \sim \nu$, where $(\mathscr{X}, \rho)$ and $(\mathscr{Y}, \zeta)$ are two Banach spaces with norms ρ and ζ representing their induced distances, θ is a Borel probability measure on $\mathscr{X} \times \mathscr{Y}$, μ and ν are two Borel probability measures on $\mathscr{X}$ and $\mathscr{Y}$, respectively. For $i = 1, \ldots, n$, let (X_i, Y_i) be a random realization of (X, Y). Then, the empirical BCov between

X and Y is defined as follows:

$$BCov^2_{w,n}(X,Y) := \frac{1}{n^2}\sum_{i,j=1}^{n}(\Delta^{XY}_{ij,n} - \Delta^{X}_{ij,n}\Delta^{Y}_{ij,n})^2\hat{w}_1(X_i,X_j)\hat{w}_2(Y_i,Y_j), \tag{6.1}$$

where

$$\Delta^{XY}_{ij,n} = \frac{1}{n}\sum_{k=1}^{n}\delta^X_{ij,k}\delta^Y_{ij,k}, \quad \Delta^{X}_{ij,n} = \frac{1}{n}\sum_{k=1}^{n}\delta^X_{ij,k}, \quad \Delta^{Y}_{ij,n} = \frac{1}{n}\sum_{k=1}^{n}\delta^Y_{ij,k},$$

with $\delta^X_{ij,k} := I\{X_k \in \bar{B}_\rho(X_i,X_j)\}$ and $\delta^Y_{ij,k} := I\{Y_k \in \bar{B}_\zeta(Y_i,Y_j)\}$. $\bar{B}_\rho(X_i,X_j)$ denotes the closed ball with center X_i and radius $\rho(X_i,X_j)$. $\bar{B}_\zeta(Y_i,Y_j)$ denotes the closed ball with center Y_i and radius $\zeta(Y_i,Y_j)$. When $\hat{w}_1 = \hat{w}_2 = 1$, the empirical BCov simplifies to

$$BCov^2_n(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n}(\Delta^{XY}_{ij,n} - \Delta^{X}_{ij,n}\Delta^{Y}_{ij,n})^2. \tag{6.2}$$

The above definition of the empirical BCov can be interpreted as follows: The Δ's describe the empirical density defined on the gradient defined by each pair of subjects (i,j), and the empirical BCov is the difference between the joint density and the product of marginal densities, which equals to zero when independence holds between X and Y; while the w's add weights on each pair of subjects for X and Y, respectively. In particular, $\Delta^X_{ij,n}$ is the *local* empirical density of X at X_i with resolution $\rho(X_i,X_j)$, because it equals to the proportion of subjects with X falling in $\bar{B}_\rho(X_i,X_j)$: $\Delta^X_{ij,n} = \frac{1}{n}\sum_{k=1}^{n}\delta^X_{ij,k}$, where $\delta^X_{ij,k} := I\{X_k \in \bar{B}_\rho(X_i,X_j)\}$ denotes whether subject k fall into $\bar{B}_\rho(X_i,X_j)$ in terms of X. The meanings of $\Delta^Y_{ij,n}$ and $\Delta^{XY}_{ij,n}$ follow similarly. Generally speaking, Δ_n's represent the local empirical densities at all data points $i = 1,\ldots,n$ across all possible resolutions (i,j) $(j = 1,\ldots,n)$, and the expression of $BCov^2_n(X,Y)$ is an average of all local empirical density differences (squared) corresponding to $[\theta(X,Y) - \mu(X)\nu(Y)]^2$ across all locations and resolutions defined by data points in the sample.

For the purpose of simplicity, we use the simplified empirical BCov, and therefore only need to calculate $\Delta^{XY}_{ij,n}$, $\Delta^{X}_{ij,n}$, and $\Delta^{Y}_{ij,n}$ for each pair of subjects (i,j). We use Euclidean distance (d) for ρ and ζ. Since in our case X_i is a number and $Y_i = (Y_{i1},\dots,Y_{ir})$ is an r-dimensional vector for $i = 1,\dots,n$, the distances between subjects $(i.j)$ w.r.t. X and Y, respectively, are thus

$$\rho(\boldsymbol{X}_i,\boldsymbol{X}_j) = d(\boldsymbol{X}_i,\boldsymbol{X}_j) = \sqrt{(X_i - X_j)^2} = |X_i - X_j|$$
$$\zeta(\boldsymbol{Y}_i,\boldsymbol{Y}_j) = d(\boldsymbol{Y}_i,\boldsymbol{Y}_j) = \sqrt{\sum_{k=1}^{r}(Y_{ik} - Y_{jk})^2}$$

on which all the δ's, thus Δ's and finally $BCov_n$, are based. As to be demonstrated later, $BCov_n$ is based on the rank of the distances between all pairs of subjects. Since d^2 preserves the rank of distances between subjects calculated under d, we use d^2 as the distance measure to save the calculation of taking the square root, which further simplifies the computation:

$$\rho(\boldsymbol{X}_i,\boldsymbol{X}_j) = d^2(\boldsymbol{X}_i,\boldsymbol{X}_j) = (X_i - X_j)^2$$
$$\zeta(\boldsymbol{Y}_i,\boldsymbol{Y}_j) = d^2(\boldsymbol{Y}_i,\boldsymbol{Y}_j) = \sum_{l=1}^{r}(Y_{il} - Y_{jl})^2$$

Simplified Calculation of BCov(X,Y)

Now, we discuss the calculation of $\delta_{ij,k}$ and $\Delta_{ij,n}$ and how it can be simplified for X, Y, and (X,Y) in the following three subsections, respectively.

Simplified Calculation of $\Delta^{X}_{ij,n}$ Using $\rho = d^2$,

$$\begin{aligned}\delta^{X}_{ij,k} &= I\{X_k \in \bar{B}_\rho(X_i,X_j)\}\\ &= I\{\rho(X_i,X_k) \le \rho(X_i,X_j)\}\\ &= I\{(X_k - X_i)^2 \le (X_j - X_i)^2\}\end{aligned}$$

For SNP screening, let X_i represent the dosage values of the target SNP for the ith subject ($i = 1, \ldots, n$), which takes values 0, 1, or 2. Therefore, pairwise distance between subjects (i, j) in terms of X has the following possible situations:

$\rho(X_i, X_j)$		X_j		
		0	1	2
X_i	0	0	1	4
	1	1	0	1
	2	4	1	0

Therefore, $\rho(X_i, X_j) = 0, 1$ or 4.

For $x = 0, 1, 2$, calculate the number and proportion of subjects with dosage value x, respectively: $s_x = \sum_{i=1}^n I\{X_i = x\}$ and $a_x = \frac{s_x}{n}$. Note that $s_0 + s_1 + s_2 = n$ and $a_0 + a_1 + a_2 = 1$. Then, given subject pair (i, j), $\Delta_{ij,n}^X = \frac{1}{n} \sum_{i=1}^n \delta_{ij,k}^X$ takes one of the following forms:

- if $\rho(X_i, X_j) = 0 \Longrightarrow \delta_{ij,k}^X = I\{X_k = X_i\} \Longrightarrow \Delta_{ij,n}^X = a_{X_i}$
- if $\rho(X_i, X_j) = 4 \Longrightarrow \delta_{ij,k}^X = 1 \Longrightarrow \Delta_{ij,n}^X = 1$
- if $\rho(X_i, X_j) = 1 \Longrightarrow \delta_{ij,k}^X = I\{X_k = X_i \text{ or } X_k = X_i \pm 1\}$, and therefore $\Delta_{ij,n}^X$ takes one of the following forms listed in the following table:

$\delta_{ij,k}^X$		X_k			$\Delta_{ij,n}^X = \frac{1}{n} \sum_{i=1}^n \delta_{ij,k}^X$
		0	1	2	
X_i	0	1	1	0	$a_0 + a_1 = 1 - a_2$
	1	1	1	1	1
	2	0	1	1	$a_1 + a_2 = 1 - a_0$

Therefore, based on the values of X_i and X_j, the value of $\Delta_{ij,n}^X$ are as follows:

$\Delta^X_{ij,n}$		X_j		
		0	1	2
X_i	0	a_0	$1 - a_2$	1
	1	1	a_1	1
	2	1	$1 - a_0$	a_2

For $i, j = 1, \ldots, n$, we expect the majority of $\Delta^X_{ij,n}$'s fall into this upper-left corner of the table, which represents non-mutants.

All the $\Delta^X_{ij,n}$'s can be stored in an $n \times n$ matrix Δ^X_n, with the (i, j)th element $(\Delta^X_n)_{i,j} = \Delta^X_{ij,n}$. One possible way to save computation resource is to save the $n \times n$ matrix Δ^X_n in the above collapsed 3×3 matrix, rather than saving an $n \times n$ matrix, while also saving the subject IDs in each dosage group: $g_0 = \{i : X_i = 0\}$, $g_1 = \{i : X_i = 1\}$, and $g_2 = \{i : X_i = 2\}$. g_0, g_1, and g_2 can be obtained when calculating s_0, s_1, and s_2.

Simplified Calculation of $\Delta^Y_{ij,n}$ Using $\zeta = d^2$,

$$\begin{aligned}
\delta^Y_{ij,k} &= I\{Y_k \in \bar{B}_\zeta(Y_i, Y_j)\} \\
&= I\{\zeta(Y_i, Y_k) \leq \zeta(Y_i, Y_j)\} \\
&= I\left\{\sum_{l=1}^{r}(Y_{kl} - Y_{il})^2 \leq \sum_{l=1}^{r}(Y_{jl} - Y_{il})^2\right\}
\end{aligned}$$

Our goal is to calculate $\Delta^Y_{ij,n} = \frac{1}{n}\sum_{k=1}^{n} \delta^Y_{ij,k}$ across all $i, j = 1, \ldots, n$, which is based on $\delta^Y_{ij,k}$ across all $i, j, k = 1, \ldots, n$. In order to make pairwise distance comparisons in $\delta^Y_{ij,k}$, we first need to calculate all pairwise distances $\zeta(Y_i, Y_j)$ for $i, j = 1, \ldots, n$, which can be saved in an $n \times n$ matrix Z.

Notice that $\delta^Y_{ij,k}$ means whether Y_k falls into ball $\bar{B}_\zeta(Y_i, Y_j)$, and thus $n\Delta^Y_{ij,n} = \sum_{k=1}^{n} \delta^Y_{ij,k}$ means the number of Y_k's that fall into ball $\bar{B}_\zeta(Y_i, Y_j)$. By sorting $\{\zeta(Y_i, Y_j) : j = 1, \ldots, n\}$ for indicator i, $\delta^Y_{ij,k}$ can be obtained for all $j, k = 1, \ldots, n$. Therefore, for each ball center Y_i $(i = 1, \ldots, n)$, obtain the ranks of $\{\zeta(Y_i, Y_1), \ldots, \zeta(Y_i, Y_n)\}$, denoted as $\{r_{i1}, \ldots, r_{in}\}$, where rank $r_{ij} = 1$ denotes the smallest element. Then, the number of Y_k's that fall into ball $\bar{B}_\zeta(Y_i, Y_j)$ equals to r_{ij}, i.e.

$$n\Delta^Y_{ij,n} = r_{ij}. \tag{6.3}$$

With the $n \times n$ distance matrix Z, this procedure is equivalent to sorting each row of Z. Then, the values of $\Delta^Y_{ij,n}$'s can be stored in an $n \times n$ matrix Δ^Y_n, with the (i,j)th element $(\Delta^Y_n)_{i,j} = \frac{r_{ij}}{n}$. Since Y remain unchanged for each SNP in the screening, we only need to calculate Δ^Y_n once for the screening of all SNPs. The distance matrix Z is also saved for calculating $\Delta^{XY}_{ij,n}$ at each SNP X.

Simplified Calculation of $\Delta^{XY}_{ij,n}$ Let Δ^{XY}_n be an $n \times n$ matrix with $(\Delta^{XY}_n)_{i,j} = \Delta^{XY}_{ij,n}$. We discuss the calculation of $n\Delta^{XY}_{ij,n} = \sum_{k=1}^n \delta^X_{ij,k}\delta^Y_{ij,k}$ in three scenarios: $\rho(X_i, X_j) = 0$, 1, and 4. The simplest case is when $\rho(X_i, X_j) = 4$, where $\delta_{ij,k} = 1$ for all $k = 1, \ldots, n$. Therefore,

$$n\Delta^{XY}_{ij,n} = \sum_{k=1}^n \delta^Y_{ij,k} = n\Delta^Y_{ij,n} = r_{ij} \tag{6.4}$$

according to equation (6.3). This happens when $\{i \in g_0 \text{ and } j \in g_2\}$ or $\{i \in g_2 \text{ and } j \in g_0\}$.

When i and j are from the same group g_x ($x = 0, 1$, or 2), we have $\rho(X_i, X_j) = 0$ and thus $\delta^X_{ij,k} = I\{X_k = x\}$. Therefore,

$$n\Delta^{XY}_{ij,n} = \sum_{k=1}^n I\{X_k = x\}\delta^Y_{ij,k} = \sum_{k \in g_x} \delta^Y_{ij,k}, \tag{6.5}$$

which is the number of subjects in g_x with distance to Y_i no larger than $\zeta(Y_i, Y_j)$. This equals to the rank of $\zeta(Y_i, Y_j)$ among $\{\zeta(Y_i, Y_j) : j \in g_x\}$ given i, denoted as $\tilde{r}_{x,ij}$. Therefore, equation (6.5) becomes

$$n\Delta^{XY}_{ij,n} = \sum_{k=1}^n I\{X_k = x\}\delta^Y_{ij,k} = \sum_{k \in g_x} \delta^Y_{ij,k} = \tilde{r}_{x,ij}. \tag{6.6}$$

$n\Delta^{XY}_{ij,n} = \tilde{r}_{x,ij}$ can be calculated for all $\{(i,j) : i \in g_x$ and $j \in g_x,$for $x = 0,1,2\} = \{(i,j) : X_i = X_j\}$ as follows: Given x ($x \in \{0,1,2\}$), extract the submatrix Z_{xx} from Z s.t. the row indicators and column indicators of Z_{xx} are both from g_x. Then, obtain the ranks $\tilde{r}_{ij}$ in each row of Z_{xx} as the (i,j)th element of $n\Delta^{XY}_n$.

If $\rho(X_i, X_j) = 1$, then $\delta^X_{ij,k} = I\{X_k = X_i \text{ or } X_k = X_i \pm 1\}$. The calculation of $n\Delta^{XY}_n = \sum_{k=1}^n \delta^X_{ij,k}\delta^Y_{ij,k}$ is discussed for $X_i = 0, 1,$ and 2, respectively, as follows:

- If $X_i = 1$, then $\delta^X_{ij,k} = 1$ for all $k = 1, \ldots, n$. Therefore, equation (6.4) holds and $(n\Delta^{XY}_n)_{i,j} = r_{ij}$.

- If $X_i = 0$, then $\delta^X_{ij,k} = I\{X_k = 0 \text{ or } 1\} = I\{X_k \neq 2\}$. Therefore,

$$n\Delta^{XY}_{ij,n} = \sum_{k=1}^{n} I\{X_k = 0 \text{ or } 1\}\delta^Y_{ij,k} = \tilde{r}_{01,ij} \tag{6.7}$$

  where $\tilde{r}_{01,ij}$ denotes the rank of element j in row i in the group $g_{01} = g_0 \cup g_1$. To calculate $\tilde{r}_{01,ij}$ in row i, extract elements with column indicator in g_{01} from row i and compute their ranks.

- If $X_i = 2$, then $\delta^X_{ij,k} = I\{X_k = 1 \text{ or } 2\} = I\{X_k \neq 0\}$. Therefore,

$$n\Delta^{XY}_{ij,n} = \sum_{k=1}^{n} I\{X_k = 1 \text{ or } 2\}\delta^Y_{ij,k} = \tilde{r}_{12,ij} \tag{6.8}$$

  where $\tilde{r}_{12,ij}$ denotes the rank is the rank of element j in row i in the group $g_{12} = g_1 \cup g_2$.

Summarize $n\Delta^{XY}_{ij,n}$ as follows:

$n\Delta^{XY}_{ij,n}$		X_j		
		0	1	2
X_i	0	$\tilde{r}_{0,ij}$	$\tilde{r}_{01,ij}$	r_{ij}
	1	r_{ij}	$\tilde{r}_{1,ij}$	r_{ij}
	2	r_{ij}	$\tilde{r}_{12,ij}$	$\tilde{r}_{2,ij}$

Compute $BCov_n$

According to equation (2),

$$n^2 BCov_n^2(X,Y) = \sum_{i,j=1}^{n} (\Delta_{ij,n}^{XY} - \Delta_{ij,n}^{X}\Delta_{ij,n}^{Y})^2. \tag{6.9}$$

Let $D_{ij}^{X_iX_j} = n\Delta_{ij,n}^{XY} - \Delta_{ij,n}^{X} \cdot n\Delta_{ij,n}^{Y}$ and the $n \times n$ matrix $D = n\Delta_n^{XY} - \Delta_n^{X} \cdot n\Delta_n^{Y} = \{D_{ij}^{X_iX_j}\}$, then

$$n^4 BCov_n^2(X,Y) = \mathbf{1}_n^T D^2 \ \mathbf{1}_n = \mathbf{1}_n^T (n\Delta_n^{XY} - \Delta_n^{X} \cdot n\Delta_n^{Y})^2 \ \mathbf{1}_n = \sum_{i,j=1}^{n} D_{ij}^2, \tag{6.10}$$

where $A^2 = \{a_{ij}\}$ for a matrix $A = \{a_{ij}\}$.

According to the table of $\Delta_{ij,n}^{X}$ and the table of $n\Delta_{ij,n}^{XY}$,

$$D_{ij}^{10} = D_{ij}^{20} = D_{ij}^{12} = D_{ij}^{02} = r_{ij} - 1 \times r_{ij} = 0 \tag{6.11}$$

$$D_{ij}^{xx} = \tilde{r}_{x,ij} - a_x r_{ij}, \text{ for } x =\in \{0,1,2\} \tag{6.12}$$

$$D_{ij}^{01} = \tilde{r}_{01,ij} - a_{01} r_{ij} \tag{6.13}$$

$$D_{ij}^{21} = \tilde{r}_{12,ij} - a_{12} r_{ij}, \tag{6.14}$$

where $a_{01} = a_0 + a_1$ and $a_{12} = a_1 + a_2$. Summarize $D_{ij}^{X_iX_j}$ as follows:

$D_{ij}^{X_iX_j}$		X_j		
		0	1	2
X_i	0	$\tilde{r}_{0,ij} - a_0 r_{ij}$	$\tilde{r}_{01,ij} - a_{01} r_{ij}$	0
	1	0	$\tilde{r}_{1,ij} - a_1 r_{ij}$	0
	2	0	$\tilde{r}_{12,ij} - a_{12} r_{ij}$	$\tilde{r}_{2,ij} - a_2 r_{ij}$

In the algorithm implementation, some ranks only need to be calculated for particular rows and columns. For example, to obtain $\tilde{r}_{01,ij}$'s for subject pair (i,j) where $i \in g_0$ and $j \in g_1$, we only need to sort the g_0 rows (with columns in g_{01}). For the efficiency of the algorithm, subjects are sorted by dosage values (0, 1, 2) before input into the xBCov algorithm.

The computation time of xBCov is only 1/4 of the original BCov (both implemented in C and called from R) for sample size n = 1000, and the fraction further lowers to 1/10 for sample size n = 4940. The run time of $n = 2k$ is averaged across 10 runs. The run times are summarized in the following table:

Run Time (s)	Average of m Runs		BCov/xBCov
	BCov	xBCov	
n=1k	0.68361	0.15963	4.3
n=2k	4.3445	0.6402	6.8
n=5k	51.8925	4.8999	10.6
2k/1k	6.4	4.0	—
5k/1k	75.9	30.7	—

Compute $BCor_n$

Since $\Delta^{XX}_{ij,n} = \frac{1}{n}\sum_{k=1}^{n}\delta^{X}_{ij,k} = \Delta^{X}_{ij,n}$,

$$\begin{aligned} BCov^2_n(X) &= \frac{1}{n^2}\sum_{i,j=1}^{n}\left[\Delta^X_{ij,n} - \left(\Delta^X_{ij,n}\right)^2\right]^2 \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n}\left[\Delta^X_{ij,n}\left(1-\Delta^X_{ij,n}\right)\right]^2 \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n} D^2_{ij}, \end{aligned} \tag{6.15}$$

where $D_{ij} = \Delta^X_{ij,n}\left(1-\Delta^X_{ij,n}\right)$. According to the table for $\Delta^X_{ij,n}$, we have

D_{ij}		X_j		
		0	1	2
X_i	0	D_0	D_2	0
	1	0	D_1	0
	2	0	D_0	D_2

where $D_0 = a_0(1 - a_0)$, $D_1 = a_1(1 - a_1)$, and $D_2 = a_2(1 - a_2)$. Therefore,

$$BCov_n^2(X) = \left[(n_0^2 + n_1 n_2)D_0^2 + n_1^2 D_1^2 + (n_2^2 + n_0 n_1)D_2^2\right]/n^2. \tag{6.16}$$

Since $n\Delta_{ij,n}^Y = r_{ij}$,

$$\begin{aligned} BCov_n^2(Y) &= \frac{1}{n^2}\sum_{i.j=1}^{n}\left[\Delta_{ij,n}^Y\left(1 - \Delta_{ij,n}^Y\right)\right]^2 \\ &= \frac{1}{n^2}\sum_{i.j=1}^{n}\left[\frac{r_{ij}}{n}\left(1 - \frac{r_{ij}}{n}\right)\right]^2 \\ &= \frac{1}{n^6}\sum_{i.j=1}^{n}[r_{ij}(1 - r_{ij})]^2. \end{aligned} \tag{6.17}$$

Then, the empirical Ball correlation is calculated as follows:

$$BCor_n^2(X,Y) = \frac{BCov_n^2(X,Y)}{\sqrt{BCov_n^2(X)BCov_n^2(Y)}}. \tag{6.18}$$

APPENDIX 2: METHOD 1 - FGWAS RESULT DETAILS

The details of the FGWAS results run on surface radial distance of both the left and right hippocampus from the ADNI cohort are reported as follows.

GSIS

Of the top 2000 SNPs selected by the GSIS step for the right hippocampus, only one SNP rs17305227 (chr19:58,986,294) is among the SNPs involved in the 73 significant SNP-region pairs identified by our method for the right hippocampus. This SNP ranked 884 in the top 2000 SNPs from the GSIS step of FGWAS, with global p-value equal to 0.0016; in our method, the SNP-region pair p-value of this SNP is 6.96×10^{-11}. For the left hippocampus, the top 2000 SNPs selected by the GSIS step include no SNP involved in the 47 significant SNP-region pair detected on the left hippocampus by our method.

Among the top 2000 SNPs from GSIS of FGWAS for each side of the hippocampus, none of their global test is significant after adjusting for the total number of SNPs using wild bootstrap based on 500 bootstrap samples for each side of the hippocampus. The maximum global test statistic reported from GSIS is 0.067 (with raw p-value 3.56×10^{-7}) for the right side and 0.078 (with raw p-value 1.15×10^{-6}) for the left side. The range of global test statistics (gstat) and their bootstrap distributions for each side of the hippocampus are listed below:

	Left	Right
Global Test Statistics (gstat)	2.08×10^{-4} to 7.81×10^{-2}	2.58×10^{-4} to 6.73×10^{-2}
gstat from Wild Bootstrap	1.45 to 1.52	1.37 to 1.46

The global test statistic is the averaged local test statistics across all voxels. Therefore, the global signal could be diluted by regions with weak or no signal. On the other hand, the global test answers the question of whether a marker is associated with the entire surface of the hippocampus, which is of less interest to use than a local test, which answers the question of whether a given marker is associated with a region of the hippocampus surface.

Local Tests

Local test is also performed at each voxel in FGWAS. No significant local signal was found among the screened SNPs after adjusting for the number of SNPs and number of voxels through wild bootstrap based on 500 bootstrap samples for each side of the hippocampus. The adjusted p-values equal to 1 for all top SNPs across all 15,000 voxels on both left and right sides of the hippocampus surface. The range of local test statistics (lstat) and their bootstrap distributions are listed below for the left and right hippocampus, respectively:

	Left	Right
Local Test Statistics (lstat)	0 to 31	0 to 27
lstat from Wild Bootstrap	33 to 58	34 to 53

Cluster-Size Analysis

The p-value of cluster-size is based on the size of the largest connected region with raw local p-value below 0.005, rather than the strength of the association at each voxel. That is to say, strong association on a small region may be ignored in cluster-size analysis.

Of the cluster-size analysis for the screened SNPs, 382 significant SNPs were found for the left hippocampus after adjusting for the number of SNPs through wild bootstrap, with the smallest p-value 0.004 for 290 SNPs; while no significant SNPs was found for the right hippocampus. With 500 bootstrap samples, the bootstrap-based p-value of 0.004 means that there are 2 bootstrap samples with $A_{bstp}^{(g)} \geq A^{(g)}$. For the right hippocampus, $A^{(g)}$ ranges from 0 to 3, and $A_{bstp}^{(g)}$ ranges from 3 to 6. For the left hippocampus surface, $A_{bstp}^{(g)}$ ranges from 3 to 7.

APPENDIX 3: METHOD 2 - EVALUATION OF ESTIMATION

For $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 = (0.5, -0.2, 0.3, 0, 0, 0, 0, 0, 0, -0.7)^T$ and $\boldsymbol{\gamma} = 4\boldsymbol{\gamma}_0$, simulate data under the following 6 models:

1. $x_i = \boldsymbol{z}_i\boldsymbol{\gamma} + \epsilon_i$

2. $x_i = |\boldsymbol{z}_i\boldsymbol{\gamma}| + \epsilon_i$

3. $x_i = \sin(\boldsymbol{z}_i\boldsymbol{\gamma}) + \epsilon_i$

4. $x_i = \exp(\boldsymbol{z}_i\boldsymbol{\gamma}) + \epsilon_i$

5. $x_i = \log|\boldsymbol{z}_i\boldsymbol{\gamma}| + \epsilon_i$

6. $x_i = \frac{1}{1+|\boldsymbol{z}_i\boldsymbol{\gamma}|} + \epsilon_i$

where $\epsilon_i \sim N(0, 0.2)$, $z_{ij} \sim N(0, 1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, sample size $n = 1000$, and $p = 10$. 100 repeated datasets are generate for each of the above 12 settings.

Apply the dCor-optimization coefficient estimation algorithm to estimate $\hat{\boldsymbol{\gamma}}$ by maximizing $dCor(Z\boldsymbol{\gamma}, X)$ for the above 12 sets of simulated data. The boxplots of $\hat{\boldsymbol{\gamma}}$ estimated across the 100 replicates for each setting is shown in Figure 6.3. We can see that $\hat{\boldsymbol{\gamma}}$ coincides with $\boldsymbol{\gamma}_0$ for most of the settings, except for model 3 with $\boldsymbol{\gamma} = 4\boldsymbol{\gamma}_0$. The number of iterations are mostly below 100 and run time approximately half a minute, except for model 4 with $\boldsymbol{\gamma} = 4\boldsymbol{\gamma}_0$ that has the number of iterations reaches the maximum for most of the replicates and run time around 7 minutes.

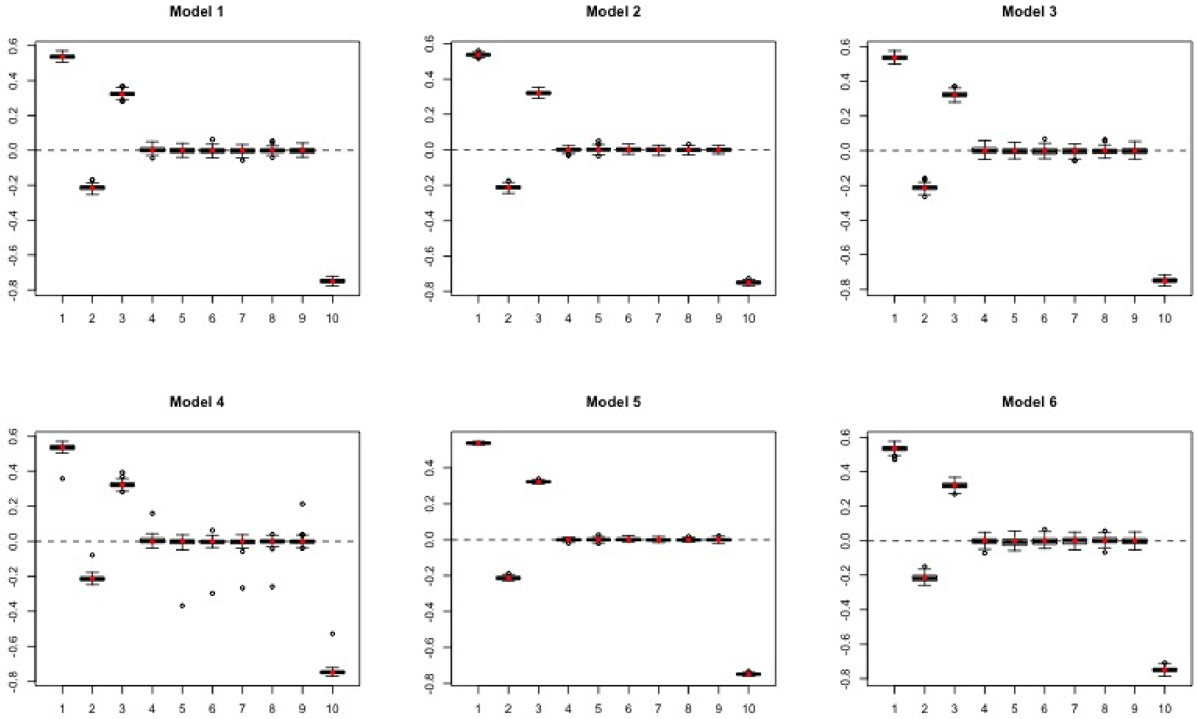


Figure 6.1: Boxplots of $\hat{\boldsymbol{\gamma}}$ vs $\boldsymbol{\gamma}$ under different settings. The red dot stands for values of true $\boldsymbol{\gamma}$.

APPENDIX 4: METHOD 2 - VOXELWISE TYPE I ERROR

f	Method	p									
		1	2	3	4	5	6	7	8	9	10
f_1	FPLS-DC	0.46	0	0	0	2.6×10^{-6}	0	0	0	2.0×10^{-6}	4.0×10^{-5}
	FPLS	0.46	0	3.7×10^{-4}	0	0	0	0	0	1.1×10^{-3}	6.0×10^{-4}
	wdCor	0.46	0	0	0	0	1.3×10^{-6}	0	0	1.9×10^{-4}	3.0×10^{-4}
	PC	0.014	0	0	0	0	0	0	0	0	0
f_2	FPLS-DC	0.43	0	0	0	0	0	4.0×10^{-6}	0	1.0×10^{-5}	4.7×10^{-6}
	FPLS	0.43	0	2.5×10^{-5}	0	0	0	0	0	4.5×10^{-5}	3.1×10^{-5}
	wdCor	0.43	0	0	0	0	1.9×10^{-5}	4.7×10^{-5}	9.0×10^{-5}	2.8×10^{-4}	3.6×10^{-4}
	PC	0.51	0	0	0	0	0	0	0	0	0

Table 6.1: Raw Voxelwise Type I Error Rates tested under the non-causal SNP. The 0's in the table means no rejection is found across all 15,000 voxels in the 100 replicates, and thus has an estimated rejection rate of less than 6.67×10^{-7}.

f	Method	p									
		1	2	3	4	5	6	7	8	9	10
f_1	FPLS-DC	0.43	0	0	0	0	2.9×10^{-5}	1.1×10^{-4}	2.3×10^{-4}	2.8×10^{-4}	2.7×10^{-4}
	FPLS	0.42	6.2×10^{-4}	0.076	0.083	0.064	0.043	0.043	0.044	0.14	0.16
	wdCor	0.42	0	0	0	1.1×10^{-5}	0	1.7×10^{-5}	1.6×10^{-5}	5.0×10^{-4}	6.4×10^{-4}
	PC	0.61	0	0	0	0	0	0	2.3×10^{-4}	1.6×10^{-5}	1.1×10^{-5}
f_2	FPLS-DC	0.40	0	0	0	1.3×10^{-5}	0	0	1.8×10^{-5}	4.6×10^{-5}	3.5×10^{-5}
	FPLS	0.40	0	6.4×10^{-3}	1.9×10^{-3}	4.1×10^{-3}	0.012	0.027	0.031	0.14	0.12
	wdCor	0.40	1.4×10^{-6}	0	0	0	0	6.5×10^{-4}	7.9×10^{-4}	6.9×10^{-4}	1.3×10^{-3}
	PC	0.64	0	0	0	0	0	0	0	0	0

Table 6.2: Raw Voxelwise Type I Error Rates in the non-affected region for the causal SNP. The 0's in the table means no rejection is found across all 14,683 non-affected voxels in the 100 replicates, and thus has an estimated rejection rate of less than 6.81×10^{-7}.

APPENDIX 5: METHOD 2 - NON-THRESHOLDED GLOBAL TEST

We examine the Type I Error and power of the non-thresholded global test using the same selected causal SNP and non-causal SNP. The p-value of the global test was calculated using the dcov.gamma() function in the R package 'kpcalg'. The rejection rates are calculated from 100 replicates, and the results are shown in Table 6.3. The first 10 rows of the table shows the rejection rates based on the dCov test performed directly on the projected image along $\hat{b}(s)$ and the SNP, while the last row shows the rejection rates based on the dCov test performed on the projected image long the thresholded $\hat{b}(s)$ and the SNP as proposed in section 4.2.3. From the table, we can see that the thresholded test gives good control of the Type I Error, while does not suffer too much in power.

	Type I Error		Power	
$f(x)$	x	$(x-0.4)^2$	x	$(x-0.4)^2$
$p=1$	0.02	0.00	0.04	0.04
$p=2$	0.03	0.03	0.06	0.04
$p=3$	0.09	0.10	0.71	0.18
$p=4$	0.41	0.44	0.98	0.51
$p=5$	0.42	0.48	0.98	0.51
$p=6$	0.42	0.48	0.98	0.50
$p=7$	0.42	0.49	0.98	0.49
$p=8$	0.41	0.49	0.98	0.52
$p=9$	0.40	0.47	0.98	0.53
$p=10$	0.42	0.48	0.98	0.50
$p=10^*$	0.00	0.00	0.84	0.11

Table 6.3: Type I Error and Power for $p=1,\ldots,10$ and $f(x)=x$ or $(x-0.4)^2$. $p=10^*$ denotes the thresholded global test with $p=10$, where the image is projected along $b(s)I\{|T(s)|>\Phi^{-1}(1-3.33\times10^{-6})\}$ to test its dependency with the SNP through dCov test using gamma approximation, where $T(s)$ is the local test statistic at s, Φ^{-1} is the inverse cdf of standard normal distribution, and $3.33\times10^{-6}=\frac{0.05}{15000}$ is the significance level adjusting for the number of 15000 voxels in the simulated image.

APPENDIX 6: METHOD 3 - THEORETICAL DETAILS

Theoretical Properties for One Study

Treat $Y(\cdot)$ as random and $Y_0(\cdot)$ given and fixed, then the expected sum of squared errors (SSE) at $\boldsymbol{s}$ is

$$\begin{aligned}
&E\Big[\big\|Y_0(\boldsymbol{s}) - \hat{Y}_0(\boldsymbol{s})\big\|_F^2\Big] = E\Big[\big\|Y_0(\boldsymbol{s}) - X_0\hat{B}(\boldsymbol{s})\big\|_F^2\Big] \\
=&E\Big[\sum_{i=1}^{n_0}\sum_{j=1}^{J}\big\{y_{0ij}(\boldsymbol{s}) - \boldsymbol{x}_{0i}^T\hat{\boldsymbol{\beta}}_j(\boldsymbol{s})\big\}^2\Big] \\
=&\sum_{i=1}^{n_0}\sum_{j=1}^{J}\Big\{Var\big[\boldsymbol{x}_{0i}^T\hat{\boldsymbol{\beta}}_j(\boldsymbol{s})\big] + \big[y_{0ij}(\boldsymbol{s}) - \boldsymbol{x}_{0i}^T E\big\{\hat{\boldsymbol{\beta}}_j(\boldsymbol{s})\big\}\big]^2\Big\} \\
=&tr\Big(R_0\sum_{j=1}^{J}Var\big[\hat{\boldsymbol{\beta}}_j(\boldsymbol{s})\big]\Big) + \big\|Y_0(\boldsymbol{s}) - X_0E\big[\hat{B}(\boldsymbol{s})\big]\big\|_F^2 \qquad (6.19)
\end{aligned}$$

with

$$E\big[\hat{B}(\boldsymbol{s})\big] = R^{-1}X^TE\big[\tilde{Y}(\boldsymbol{s})\big] = \sum_{v=1}^{V}c_h(\boldsymbol{s},\boldsymbol{s}_v)B(\boldsymbol{s}_v) \qquad (6.20)$$

$$Var\big[\hat{\boldsymbol{\beta}}_j(\boldsymbol{s})\big] = Var\big[R^{-1}X^T\tilde{\boldsymbol{y}}_j(\boldsymbol{s})\big] = R^{-1}X^TVar\big[\tilde{\boldsymbol{y}}_j(\boldsymbol{s})\big]XR^{-1}, \qquad (6.21)$$

where $\tilde{\boldsymbol{y}}_j(\boldsymbol{s})$ is the jth column of $\tilde{Y}(\boldsymbol{s})$ and

$$\begin{aligned}
&Var\big[\tilde{\boldsymbol{y}}_j(\boldsymbol{s})\big] = Var\Big[\sum_{v=1}^{V}c_h(\boldsymbol{s},\boldsymbol{s}_v)\ \boldsymbol{y}_j(\boldsymbol{s}_v)\Big] \\
=&\sum_{v_1=1}^{V}\sum_{v_2=1}^{V}Cov\Big[c_h(\boldsymbol{s},\boldsymbol{s}_{v_1})\ \boldsymbol{y}_j(\boldsymbol{s}_{v_1}),\ c_h(\boldsymbol{s},\boldsymbol{s}_{v_2})\ \boldsymbol{y}_j(\boldsymbol{s}_{v_2})\Big] \\
=&\sum_{v_1=1}^{V}\sum_{v_2=1}^{V}a_h(\boldsymbol{s},\boldsymbol{s}_{v_1},\boldsymbol{s}_{v_2})\ Cov\Big[\boldsymbol{y}_j(\boldsymbol{s}_{v_1}),\boldsymbol{y}_j(\boldsymbol{s}_{v_2})\Big], \qquad (6.22)
\end{aligned}$$

with the second-order smoothing coefficient $a_h(\boldsymbol{s},\boldsymbol{s}_{v_1},\boldsymbol{s}_{v_2}) = c_h(\boldsymbol{s},\boldsymbol{s}_{v_1})\cdot c_h(\boldsymbol{s},\boldsymbol{s}_{v_2})$.

Theoretical Properties for Multiple Studies

Since

$$E\big[\hat{B}_e(\boldsymbol{s})\big] = E\big[\hat{B}_m(\boldsymbol{s})\big] = \sum_{v=1}^{V} c_h(\boldsymbol{s}, \boldsymbol{s}_v) B(\boldsymbol{s}_v) \qquad \text{and} \tag{6.23}$$

$$Var\big[\hat{\boldsymbol{\beta}}_{e,j}(\boldsymbol{s})\big] = \sum_{k=1}^{K} w_k^2 \; Var\big[\hat{\boldsymbol{\beta}}_{k,j}(\boldsymbol{s})\big], \tag{6.24}$$

according to (6.19), (5.16) is equivalent to

$$\sum_{\boldsymbol{s}\in\mathcal{S}} tr\Big(R_0 \sum_{j=1}^{J}\sum_{k=1}^{K} w_k^2 \; Var\big[\hat{\boldsymbol{\beta}}_{k,j}(\boldsymbol{s})\big]\Big) \;\le\; \sum_{\boldsymbol{s}\in\mathcal{S}} tr\Big(R_0 \sum_{j=1}^{J} Var\big[\hat{\boldsymbol{\beta}}_{m,j}(\boldsymbol{s})\big]\Big), \tag{6.25}$$

where $\hat{\boldsymbol{\beta}}_{e,j}(\boldsymbol{s})$, $\hat{\boldsymbol{\beta}}_{k,j}(\boldsymbol{s})$, and $\hat{\boldsymbol{\beta}}_{m,j}(\boldsymbol{s})$ are the jth columns of $\hat{B}_e(\boldsymbol{s})$, $\hat{B}_k(\boldsymbol{s})$, and $\hat{B}_m(\boldsymbol{s})$, respectively.

Let $\Sigma_r(\boldsymbol{s},\boldsymbol{t}) = \Sigma_\eta(\boldsymbol{s},\boldsymbol{t}) + \Sigma_\epsilon(\boldsymbol{s},\boldsymbol{t})$ and $\boldsymbol{y}_{k,j}(\boldsymbol{s})$ denote the jth column of $Y_k(\boldsymbol{s})$, then

$$Cov\big[\boldsymbol{y}_{k,j}(\boldsymbol{s}), \boldsymbol{y}_{k,j}(\boldsymbol{t})\big] = Z_k G_j(\boldsymbol{s},\boldsymbol{t}) Z_k^T + \big(\Sigma_r(\boldsymbol{s},\boldsymbol{t})\big)_{jj} I_{n_k}. \tag{6.26}$$

Utilizing (6.21) and (6.22), in the left hand side of (6.25)

$$\begin{aligned}
& tr\Big(R_0 Var\big[\hat{\boldsymbol{\beta}}_{k,j}(\boldsymbol{s})\big]\Big) \\
=& tr\Big(R_0 R_k^{-1} X_k^T \sum_{v_1,v_2} a_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) \big[Z_k G_j(\boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) Z_k^T + \big(\Sigma_r(\boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2})\big)_{jj} I_{n_k}\big] X_k R_k^{-1}\big]\Big) \\
=& tr\Big(Z_0^T Z_0 \sum_{v_1,v_2} a_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) G_j(\boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2})\Big) + \sum_{v_1,v_2} a_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) \big(\Sigma_r(\boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2})\big)_{jj} tr\Big(R_k^{-1} R_0\Big)
\end{aligned} \tag{6.27}$$

Let $\tilde{G}.(\boldsymbol{s}) = \sum_{v_1,v_2} a_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) \sum_{j=1}^{J} G_j(\boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2})$, which is also a $q \times q$ diagonal matrix, with diagonal elements $\tilde{\sigma}_{\cdot l}^2(\boldsymbol{s}) = \sum_{v_1,v_2} a_h(\boldsymbol{s}, \boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2}) \sum_{j=1}^{J} \sigma_{jl}^2(\boldsymbol{s}_{v_1}, \boldsymbol{s}_{v_2})$ $(l = 1, \ldots, q)$, and $\tilde{\sigma}_r^2(\boldsymbol{s}) =$

$\sum_{v_1,v_2} a_h(\boldsymbol{s},\boldsymbol{s}_{v_1},\boldsymbol{s}_{v_2}) tr\big(\Sigma_r(\boldsymbol{s}_{v_1},\boldsymbol{s}_{v_2})\big)$; denote $G_{\mathcal{S}} = \sum_{\boldsymbol{s}\in\mathcal{S}} \tilde{G}.(\boldsymbol{s})$ and $\sigma^2_{r,\mathcal{S}} = \sum_{\boldsymbol{s}\in\mathcal{S}} \tilde{\sigma}^2_r(\boldsymbol{s})$, then

$$\begin{aligned}
&\sum_{\boldsymbol{s}\in\mathcal{S}} tr\Big(R_0 \sum_{j=1}^{J}\sum_{k=1}^{K} w_k^2 \ Var\big[\hat{\boldsymbol{\beta}}_{j,k}(\boldsymbol{s})\big]\Big) \\
=&\sum_{\boldsymbol{s}\in\mathcal{S}} \Big(\sum_{k=1}^{K} w_k^2\Big) tr\Big(\tilde{G}.(\boldsymbol{s}) Z_0^T Z_0\Big) + \tilde{\sigma}^2_r(\boldsymbol{s}) \sum_{k=1}^{K} w_k^2 tr\Big(R_k^{-1} R_0\Big) \\
=&\Big(\sum_{k=1}^{K} w_k^2\Big) tr\Big(G_{\mathcal{S}} Z_0^T Z_0\Big) + \sigma^2_{r,\mathcal{S}} \sum_{k=1}^{K} w_k^2 tr\Big(R_k^{-1} R_0\Big).
\end{aligned} \tag{6.28}$$

Let $\boldsymbol{y}_{m,j}(\boldsymbol{s})$ denote the jth column of $Y_m(\boldsymbol{s})$, and $n_m = \sum_{k=1}^{K} n_k$. $Cov\big[\boldsymbol{y}_{m,j}(\boldsymbol{s}), \boldsymbol{y}_{m,j}(\boldsymbol{t})\big]$ is an $n_m \times n_m$ block diagonal matrix, with the kth block being $Cov\big[\boldsymbol{y}_{k,j}(\boldsymbol{s}), \boldsymbol{y}_{k,j}(\boldsymbol{t})\big]$, and

$$X_m^T Cov\big[\boldsymbol{y}_{m,j}(\boldsymbol{s}), \boldsymbol{y}_{m,j}(\boldsymbol{t})\big] X_m = \sum_{k=1}^{K} X_k^T \Big[Z_k G_j(\boldsymbol{s},\boldsymbol{t}) Z_k^T + \big(\Sigma_r(\boldsymbol{s},\boldsymbol{t})\big)_{jj} I_{n_k}\Big] X_k. \tag{6.29}$$

Therefore, the right hand side of (6.25)

$$\begin{aligned}
&\sum_{\boldsymbol{s}\in\mathcal{S}} tr\Big(R_0 \sum_{j=1}^{J} Var\big[\hat{\boldsymbol{\beta}}_{m,j}(\boldsymbol{s})\big]\Big) \qquad (6.30) \\
=&\sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{j=1}^{J} \sum_{v_1,v_2} a_h(\boldsymbol{s},\boldsymbol{s}_{v_1},\boldsymbol{s}_{v_2}) tr\Big(R_0 R_m^{-1} \sum_{k=1}^{K} X_k^T \Big[Z_k G_j(\boldsymbol{s}_{v_1},\boldsymbol{s}_{v_2}) Z_k^T + \big(\Sigma_r(\boldsymbol{s}_{v_1},\boldsymbol{s}_{v_2})\big)_{jj} I_{n_k}\Big] X_k R_m^{-1}\Big) \\
=&\sum_{\boldsymbol{s}\in\mathcal{S}} tr\Big(R_m^{-1} \sum_{k=1}^{K} X_k^T Z_k \tilde{G}.(\boldsymbol{s}) Z_k^T X_k R_m^{-1} R_0\Big) + \tilde{\sigma}^2_r(\boldsymbol{s}) tr\Big(R_m^{-1} R_0\Big) \\
=&tr\Big(R_m^{-1} \sum_{k=1}^{K} X_k^T Z_k G_{\mathcal{S}} Z_k^T X_k R_m^{-1} R_0\Big) + \sigma^2_{r,\mathcal{S}} tr\Big(R_m^{-1} R_0\Big).
\end{aligned}$$

The two sides of (6.25) are (6.28) and (6.30), respectively, both of which are consist of a term related to $G_{\mathcal{S}}$ and a term multiplied by $\sigma^2_{r,\mathcal{S}}$:

	$G_{\mathcal{S}}$ term	$\sigma^2_{r,\mathcal{S}}$ term
Ensemble	$A_e = \big(\sum_{k=1}^{K} w_k^2\big) tr\Big(G_{\mathcal{S}} Z_0^T Z_0\Big)$	$B_e = \sum_{k=1}^{K} w_k^2 \ tr\Big(R_k^{-1} R_0\Big)$
Merge	$A_m = tr\Big(R^{-1} \sum_{k=1}^{K} X_k^T Z_k G_{\mathcal{S}} Z_k^T X_k R^{-1} R_0\Big)$	$B_m = tr(R^{-1} R_0)$

Therefore, the comparison of the prediction accuracy of the two learners from ensembling and merging in (5.16) can be reduced to the relationship between $G_{\mathcal{S}}$ and $\sigma^2_{r,\mathcal{S}}$ through terms of the design matrix, i.e. (6.25) is equivalent to

$$
\begin{aligned}
& & A_e + \sigma^2_{r,\mathcal{S}} \cdot B_e \le A_m + \sigma^2_{r,\mathcal{S}} \cdot B_m & \\
& i.e. & A_m - A_e \ge \sigma^2_{r,\mathcal{S}}(B_e - B_m) & \qquad (6.31)
\end{aligned}
$$

Proof of Theorem 1

Denote $\sigma^2_{\cdot l,\mathcal{S}} = \sum_{\boldsymbol{s}\in\mathcal{S}} \tilde{\sigma}^2_{\cdot l}(\boldsymbol{s})$ for $l \in \{1,\dots,q\}$, then $G_{\mathcal{S}} = diag(\sigma^2_{\cdot 1,\mathcal{S}},\dots,\sigma^2_{\cdot q,\mathcal{S}})$. We can further reduce the relationship between $G_{\mathcal{S}}$ and $\sigma^2_{r,\mathcal{S}}$ from (6.13) to a simpler form. $\forall l \in \{1,\dots,q\}$, let $\boldsymbol{u}_l$ be a $q \times 1$ vector with all elements equal to 0 except for the l-th element which equals to 1, s.t. $Z_k\boldsymbol{u}_l$ gives the lth column of Z_k. Therefore, $\boldsymbol{u}_l\boldsymbol{u}_l^T$ is a $q \times q$ matrix with all elements equal to 0 except for the lth diagonal element which equals to 1, and $\sum_{l=1}^q \boldsymbol{u}_l\boldsymbol{u}_l^T = I_q$. Therefore, $G_{\mathcal{S}} = \sum_{l=1}^q \sigma^2_{\cdot l,\mathcal{S}}\ \boldsymbol{u}_l\boldsymbol{u}_l^T$,

$$
\begin{aligned}
A_e =& \Big(\sum_{k=1}^K w_k^2\Big) tr\Big(G_{\mathcal{S}} Z_0^T Z_0\Big) \\
=& \Big(\sum_{k=1}^K w_k^2\Big) tr\Big(\sum_{l=1}^q \sigma^2_{\cdot l,\mathcal{S}}\ \boldsymbol{u}_l\boldsymbol{u}_l^T Z_0^T Z_0\Big) \\
=& \Big(\sum_{k=1}^K w_k^2\Big) \sum_{l=1}^q \sigma^2_{\cdot l,\mathcal{S}}\ \boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l, \qquad (6.32)
\end{aligned}
$$

and

$$
\begin{aligned}
A_m =& tr\Big(R_m^{-1} \sum_{k=1}^K X_k^T Z_k G_{\mathcal{S}} Z_k^T X_k R_m^{-1} R_0\Big) \\
=& tr\Big(R_m^{-1} \sum_{k=1}^K X_k^T Z_k \Big[\sum_{l=1}^q \sigma^2_{\cdot l,\mathcal{S}}\ \boldsymbol{u}_l\boldsymbol{u}_l^T\Big] Z_k^T X_k R_m^{-1} R_0\Big) \\
=& \sum_{l=1}^q \sigma^2_{\cdot l,\mathcal{S}}\ tr\Big(R_m^{-1} \sum_{k=1}^K X_k^T Z_k \boldsymbol{u}_l\boldsymbol{u}_l^T Z_k^T X_k R_m^{-1} R_0\Big). \qquad (6.33)
\end{aligned}
$$

Denote $\sigma^2_{\mathcal{S}} = \frac{1}{q} tr\Big(G_{\mathcal{S}}\Big) = \frac{1}{q}\sum_{l=1}^{q} \sigma^2_{\cdot l,\mathcal{S}}$, for $l \in \{1, \dots, q\}$, $A_{el} = \Big(\sum_{k=1}^{K} w_k^2\Big) \cdot \boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l$ and $A_{ml} = tr\Big(R_m^{-1} \sum_{k=1}^{K} X_k^T Z_k \boldsymbol{u}_l \boldsymbol{u}_l^T Z_k^T X_k R_m^{-1} R_0\Big)$, then the left hand side of (6.31)

$$A_m - A_e = \sum_{l=1}^{q} \sigma^2_{\cdot l,\mathcal{S}} \left(A_{ml} - A_{el}\right). \tag{6.34}$$

Therefore,

$$q\sigma^2_{\mathcal{S}} \min_l \{A_{ml} - A_{el}\} \le A_m - A_e \le q\sigma^2_{\mathcal{S}} \max_l \{A_{ml} - A_{el}\}. \tag{6.35}$$

Therefore, a sufficient condition for (6.31) $\iff$ (6.25) $\iff$ (5.16) is

$$q\sigma^2_{\mathcal{S}} \min_l \{A_{ml} - A_{el}\} \ge \sigma^2_{r,\mathcal{S}}(B_e - B_m). \tag{6.36}$$

Similarly, a sufficient condition for

$$E\left[\sum_{\boldsymbol{s}\in\mathcal{S}} \left\| Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s})\right\|_F^2\right] \quad \ge \quad E\left[\sum_{\boldsymbol{s}\in\mathcal{S}} \left\| Y_0(\boldsymbol{s}) - \hat{Y}_{0,m}(\boldsymbol{s})\right\|_F^2\right], \tag{6.37}$$

i.e.

$$A_m - A_e \le \sigma^2_{r,\mathcal{S}}(B_e - B_m), \tag{6.38}$$

is therefore

$$q\sigma^2_{\mathcal{S}} \max_l \{A_{ml} - A_{el}\} \le \sigma^2_{r,\mathcal{S}}(B_e - B_m). \tag{6.39}$$

Write $\sigma^2_{\mathcal{S}}$ on the left of the inequality, we have the conclusions in Theorem 1.

Proof of Theorem 2

If the variance of the random effect $\boldsymbol{\gamma}_{kjl}(\boldsymbol{s})$ remains the same across $l = 1, \dots, q$, where $\boldsymbol{\gamma}_{kjl}(\boldsymbol{s})$ is the lth element in the $q \times 1$ vector $\boldsymbol{\gamma}_{kj}(\boldsymbol{s})$, then $\sigma^2_{\cdot l,\mathcal{S}} = \sigma^2_{\mathcal{S}}$ for $l \in \{1, \dots, q\}$. Therefore,

$G_{\mathcal{S}} = \sigma_{\mathcal{S}}^2 \cdot I_q$, and

$$A_m = \sigma_{\mathcal{S}}^2 \cdot tr\Big(R^{-1} \sum_{k=1}^{K} X_k^T Z_k Z_k^T X_k R^{-1} R_0\Big) \tag{6.40}$$

$$A_e = \sigma_{\mathcal{S}}^2 \cdot \Bigg(\sum_{k=1}^{K} w_k^2\Bigg) tr\big(Z_0^T Z_0\big). \tag{6.41}$$

Let $A_{m0} = tr\Big(R^{-1} \sum_{k=1}^{K} X_k^T Z_k Z_k^T X_k R^{-1} R_0\Big)$ and $A_{e0} = \big(\sum_{k=1}^{K} w_k^2\big) tr\big(Z_0^T Z_0\big)$, then

$$A_m - A_e = \sigma_{\mathcal{S}}^2 \cdot (A_{m0} - A_{e0}). \tag{6.42}$$

Therefore, given (5.16) $\iff$ (6.25) $\iff$ (6.31), we have the conclusion in Theorem 2.

Derivation of Optimal Ensembling Weights

With the ensembled learner $\hat{B}_e(\boldsymbol{s}) = \sum_{k=1}^{K} w_k \hat{B}_k(\boldsymbol{s})$, it is of interest to derive the ensembling weights w_k $(k = 1, \dots, K)$ that achieve the minimum expected prediction error

$$E\Bigg[\sum_{\boldsymbol{s}\in\mathcal{S}} \Big\|Y_0(\boldsymbol{s}) - \hat{Y}_{0,e}(\boldsymbol{s})\Big\|_F^2\Bigg].$$

According to (6.19), the expected prediction error above is equivalent to

$$\sum_{\boldsymbol{s}\in\mathcal{S}} tr\Big(R_0 \sum_{j=1}^{J} Var\big[\hat{\boldsymbol{\beta}}_{e,j}(\boldsymbol{s})\big]\Big) + \sum_{\boldsymbol{s}\in\mathcal{S}} \Big\|Y_0(\boldsymbol{s}) - X_0 \sum_{v=1}^{V} c_h(\boldsymbol{s}, \boldsymbol{s}_v) B(\boldsymbol{s}_v)\Big\|_F^2,$$

where the second term is constant given $\big(X_0, Y_0(\cdot)\big)$ and h. Therefore, we aim to minimize the first term above w.r.t. w_k's. According to (6.28), our objective of minimization

$$\sum_{\boldsymbol{s}\in\mathcal{S}} tr\Big(R_0 \sum_{j=1}^{J} Var\big[\hat{\boldsymbol{\beta}}_{e,j}(\boldsymbol{s})\big]\Big) = \Big(\sum_{k=1}^{K} w_k^2\Big) tr\big(G_{\mathcal{S}} Z_0^T Z_0\big) + \sigma_{r,\mathcal{S}}^2 \sum_{k=1}^{K} w_k^2 tr\big(R_k^{-1} R_0\big).$$

Denote

$$f(w_1, \dots, w_K) = \Big(\sum_{k=1}^{K} w_k^2 \Big) tr\Big(G_{\mathcal{S}} Z_0^T Z_0 \Big) + \sigma_{r,\mathcal{S}}^2 \sum_{k=1}^{K} w_k^2 tr\Big(R_k^{-1} R_0 \Big),$$
$$g(w_1, \dots, w_K) = \sum_{k=1}^{K} w_k.$$

Since $\sum_{k=1}^{K} w_k = 1$, we can write this optimization problem as

$$\text{minimize} \quad f(w_1, \dots, w_K)$$
$$\text{subject to} \quad g(w_1, \dots, w_K) = 1.$$

Using Lagrange multipliers, we get the system of equations

$$\frac{\partial}{\partial w_k} f(w_1, \dots, w_K) = \lambda \frac{\partial}{\partial w_k} g(w_1, \dots, w_K) \qquad (\text{for } k = 1, \dots, K)$$
$$g(w_1, \dots, w_K) = 1.$$

Since

$$\frac{\partial}{\partial w_k} f(w_1, \dots, w_K) = 2 w_k \Big[tr(G_{\mathcal{S}} Z_0^T Z_0) + \sigma_{r,\mathcal{S}}^2 tr(R_k^{-1} R_0) \Big],$$
$$\frac{\partial}{\partial w_k} g(w_1, \dots, w_K) = 1,$$

the system of equations leads to

$$w_k = 0.5 \lambda \Big[tr(G_{\mathcal{S}} Z_0^T Z_0) + \sigma_{r,\mathcal{S}}^2 tr(R_k^{-1} R_0) \Big]^{-1}$$

for $k = 1, \dots, K$, with $\sum_{k=1}^{K} w_k = 1$. Therefore, the optimal ensembling weights

$$w_{k,opt} = \frac{a_k}{\sum_{k=1}^{K} a_k}, \tag{6.43}$$

with

$$a_k = \left[tr(G_{\mathcal{S}} Z_0^T Z_0) + \sigma^2_{r,\mathcal{S}} tr(R_k^{-1} R_0)\right]^{-1} \tag{6.44}$$

is the w_k-related term $\sum_{\boldsymbol{s}\in\mathcal{S}} tr\Big(R_0 \sum_{j=1}^J Var\big[\hat{\boldsymbol{\beta}}_{k,j}(\boldsymbol{s})\big]\Big)$ in $\sum_{\boldsymbol{s}\in\mathcal{S}} \big\|Y_0(\boldsymbol{s}) - X_0 \hat{B}_k(s)\big\|_F^2$.

Proof of b's > 0

1. Prove $b_{1l} = tr\Big(R_m^{-1} \sum_{k=1}^K X_k^T Z_k \boldsymbol{u}_l \boldsymbol{u}_l^T Z_k^T X_k R_m^{-1} R_0\Big) - \Big(\sum_{k=1}^K w_k^2\Big) \cdot \boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l > 0$ under equal ensemble weights $w_k = \frac{1}{K}$ for $k = 1, \ldots, K$:

For $k = 1, \ldots, K$, suppose the columns of X_k are linearly independent. Let $R_k = V_k D_k V_k^{-1}$ be an eigendecomposition where $D_k = diag(d_1^{(k)}, \ldots, d_p^{(k)})$ ($d_l^{(k)} > 0$ for $l = 0, 1, \ldots, p$) and V_k is an orthonormal matrix of eigenvectors, and $R_m = VDV^{-1}$ be an eigendecomposition where $D = diag(d_1, \ldots, d_p)$ ($d_l > 0$ for $l = 1, \ldots, p$) and V is an orthonormal matrix of eigenvectors. Assume $V_k = V$ for $k = 1, \ldots, K$, i.e. the relationship between the p covariates remain the same across studies, then $D = \sum_{k=1}^K D_k$, i.e. $d_l = \sum_{k=1}^K d_l^{(k)}$ for $l = 1, \ldots, p$, because $R_m = \sum_{k=1}^K R_k$.

Since $Z_k = X_k U$,

$$\begin{aligned}
& tr\Big(R_m^{-1} \sum_{k=1}^K X_k^T Z_k \boldsymbol{u}_l \boldsymbol{u}_l^T Z_k^T X_k R_m^{-1} R_0\Big) \\
= & \sum_{k=1}^K \boldsymbol{u}_l^T Z_k^T X_k R_m^{-1} R_0 R_m^{-1} X_k^T Z_k \boldsymbol{u}_l \\
= & \sum_{k=1}^K \boldsymbol{u}_l^T U^T V D_k D^{-1} D_0 D^{-1} D_k V^{-1} U \boldsymbol{u}_l \\
= & \boldsymbol{u}_l^T U^T V D_0 \ diag(\ldots, \frac{\sum_{k=1}^K \big[d_{l'}^{(k)}\big]^2}{d_{l'}^2}, \ldots) \ V^{-1} U \boldsymbol{u}_l
\end{aligned}$$

and

$$\boldsymbol{u}_l^T Z_0^T Z_0 \boldsymbol{u}_l = \boldsymbol{u}_l^T U^T V D_0 V^{-1} U \boldsymbol{u}_l.$$

Since $\sum_{k=1}^K w_k^2$ reaches its minimum value $\frac{1}{K}$ at equal ensemble weights $w_k = \frac{1}{K}$ for $k = 1, \ldots, K$, we reduces to comparing $K \sum_{k=1}^K \Big[d_{l'}^{(k)}\Big]^2$ and $d_{l'}^2 = \Big[\sum_{k=1}^K d_{l'}^{(k)}\Big]^2$ for $l = 1, \ldots, p$. Suppose

$d_{l'}^{(k)} \neq d_{l'}^{(k')}$ for some $k, k' = 1, \ldots, K$ and $l' = 1, \ldots, p$. then,

$$K \sum_{k=1}^{K} \left[d_{l'}^{(k)} \right]^2 - \left[\sum_{k=1}^{K} d_{l'}^{(k)} \right]^2 = \sum_{k=2} K \sum_{k'=1}^{k-1} \left[d_{l'}^{(k)} - d_{l'}^{(k')} \right]^2 > 0,$$

i.e.

$$\frac{\sum_{k=1}^{K} \left[d_{l'}^{(k)} \right]^2}{d_{l'}^2} > \frac{1}{K}.$$

Therefore,

$$\boldsymbol{u}_l^T U^T V D_0 \; diag(\ldots, \frac{\sum_{k=1}^{K} \left[d_{l'}^{(k)} \right]^2}{d_{l'}^2}, \ldots) \; V^{-1} U \boldsymbol{u}_l > \frac{1}{K} \boldsymbol{u}_l^T U^T V D_0 V^{-1} U \boldsymbol{u}_l,$$

i.e. $b_{1l} > 0$ under equal ensemble weights $w_k = \frac{1}{K}$ for $k = 1, \ldots, K$.

$b_2 = tr\Big(R^{-1} \sum_{k=1}^{K} X_k^T Z_k Z_k^T X_k R^{-1} R_0 \Big) - \big(\sum_{k=1}^{K} w_k^2 \big) \, tr\big(Z_0^T Z_0 \big) > 0$ can be proved similarly.

2. Prove $b_0 = \sum_{k=1}^{K} w_k^2 \; tr\big(R_k^{-1} R_0 \big) - tr(R^{-1} R_0) > 0$:

Using the same eigendecomposition above,

$$\begin{aligned} b_0 &= \sum_{k=1}^{K} w_k^2 \; tr\big(R_k^{-1} R_0 \big) - tr(R^{-1} R_0) \\ &= \sum_{k=1}^{K} w_k^2 \; tr(D_k^{-1} D_0) - tr(D^{-1} D_0) \\ &= tr\Big(diag\big(\ldots, \left[\sum_{k=1}^{K} \frac{w_k^2}{d_l^{(k)}} \right] - \frac{1}{d_l}, \ldots \big) D_0 \Big) \\ &= \sum_{l=1}^{p} \left\{ \left[\sum_{k=1}^{K} \frac{w_k^2}{d_l^{(k)}} \right] - \frac{1}{d_l} \right\} d_l^{(0)}. \end{aligned}$$

According to Jensen's inequality,

$$\sum_{k=1}^{K} \frac{w_k^2}{d_l^{(k)}} \geq \left[\sum_{k=1}^{K} w_k^2 d_l^{(k)} \right]^{-1}.$$

Since $0 < w_k < 1$ for $k = 1, \ldots, K$, $w_k^2 d_l^{(k)} < d_l^{(k)}$,

$$\sum_{k=1}^{K} \frac{w_k^2}{d_l^{(k)}} > \left[\sum_{k=1}^{K} d_l^{(k)}\right]^{-1} = d_l^{-1}.$$

Therefore, $b_0 > 0$.

APPENDIX 7: METHOD 3 - MORE ON TRANSITION POINT

Under equal variance (i.e. homogeneous) and equal weights ($w_k = frac14$), using the same correlation structure as in the main simulation setting, we examined the values of tau_2 under different sample sizes. Let n_k vary from 100 to 10,000 by 100 for $k = 1, \dots, 4$. Consider at 2 scenarios of n_0: 1) $n_0 = n_1 + n_2 + n_3 + n_4$, and 2) $n_0 = 500$. For each scenario, the change of $A_m, A_e, B_m, B_e, d_0, d_1$, and τ_2 are plotted below. We can see that as n increases, τ_2 decreases in both scenario.

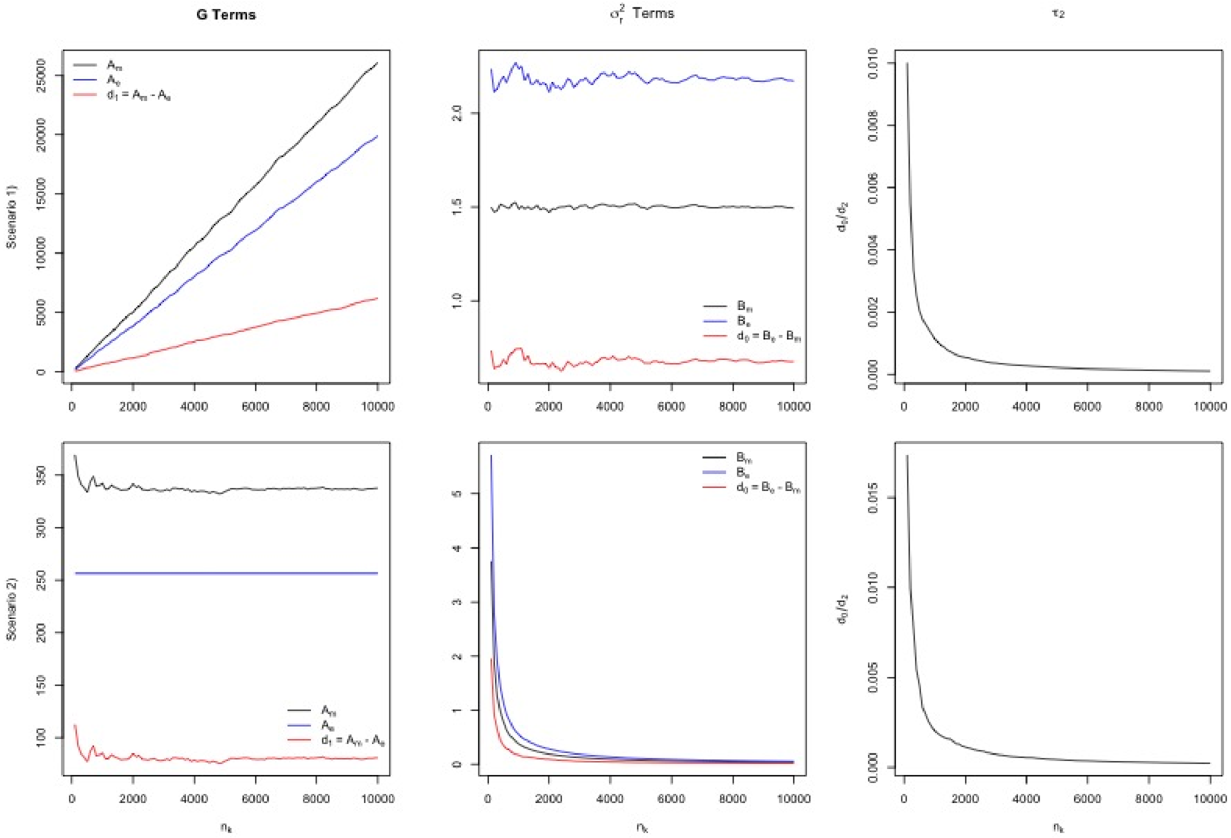


Figure 6.2: Values of terms in τ_2 for varied sample size.

REFERENCES

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30,** 97–101.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62(5),** 1198–211.

Asimit, J. L., Day-Williams, A. G., Morris, A. P., and Zeggini, E. (2012). Ariel and amelia: testing for an accumulation of rare variants using next-generation sequencing data. *Human Heredity* **73(2),** 84–94.

Basser, P. J., Mattiello, J., and Lebihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance, Series B* **103,** 247–254.

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67,**.

Bellamy, S. L., Li, Y., Lin, X., and Ryan, L. M. (2005). Quantifying pql bias in estimating cluster-level covariate effects in generalized linear mixed models for group-randomized trials. *Statistica Sinica* **15(4),** 1015–1032.

Bishop, C. M. (2006). Pattern recognition and machine learning. *Journal of Electronic Imaging* **16,** 049901.

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics* **32(2),** 485–498.

Bravata, D. M. and Olkin, I. (2001). Simple pooling versus combining in meta-analysis. *Evaluation & the health professions* **24,** 218–230.

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, ..., and Neale, B. M. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47(3),** 291–295.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., and Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209.

Cai, C., Chen, R., and Xie, M.-g. (2020). Individualized inference through fusion learning. *WIREs Computational Statistics* **12,** e1498.

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., ..., and Dale, A. M. (2018). The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience* **32,** 43–54.

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., and Raschellà, G. (2017). Zinc-finger proteins in health and disease. *Cell Death Discovery* **3,** 17071.

Chasioti, D., Yan, J., Nho, K., and Saykin, A. J. (2019). Progress in polygenic composite scores in alzheimer′s and other complex diseases. *Trends in Genetics* **35(5),** 371–382.

Cheng, J. Q., Liu, R. Y., and ge Xie, M. (2017). Fusion learning. *Wiley StatsRef: Statistics Reference Online* pages 1–8.

Chiu, C.-y., Jung, J., Chen, W., Weeks, D. E., Ren, H., Boehnke, M., Amos, C. I., Liu, A., Mills, J. L., Ting Lee, M.-l., Xiong, M., and Fan, R. (2017). Meta-analysis of quantitative pleiotropic traits for next-generation sequencing with multivariate functional linear models. *European Journal of Human Genetics* **25,** 350–359.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10,** 101–129.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal* **Complex Systems,** 1695.

Dahmani, L., Courcot, B., Near, J., Patel, R., Amaral, R. S. C., Chakravarty, M. M., and Bohbot, V. D. (2020). Fimbria-fornix volume is associated with spatial memory and olfactory identification in humans. *Frontiers in Systems Neuroscience* **13,** 87.

Deb, N. and Sen, B. (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *arXiv* page 1909.08733v1.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics* **40,** 322–352.

Derkach, A., Lawless, J. F., and Sun, L. (2013). Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology* **37(1),** 110–21.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7(3),** 177–88.

DerSimonian, R. and Laird, N. (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials* **45,** 139–145.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *In International workshop on multiple classifier systems* **1875,** 1–15.

Dima, D. and Breen, G. (2015). Polygenic risk scores in imaging genetics: Usefulness and applications. *Journal of Psychopharmacology* **29(8),** 867–71.

Du, L., Liu, K., Zhang, T., Yao, X., Yan, J., Risacher, S. L., Han, J., Guo, L., Saykin, A. J., Shen, L., and for the Alzheimer's Disease Neuroimaging Initiative (2018). A novel scca approach via truncated ℓ_1-norm and truncated group lasso for brain imaging genetics. *Bioinformatics* pages 278–285.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* **9(3),** e1003348.

Dutta, D., Scott, L., Boehnke, M., and Lee, S. (2018). Multi-skat: General framework to test for rare-variant association with multiple phenotypes. *Genetic Epidemiology* **43(1),** 4–23.

Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its applications. monographs on statistics and applied probability 66. *Chapman & Hall, London* .

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B* **70,** 849–911.

Fan, R., Wang, Y., Boehnke, M., Chen, W., Li, Y., Ren, H., Lobach, I., and Xiong, M. (2015). Gene Level Meta-Analysis of Quantitative Traits by Functional Linear Models. *Genetics* **200,** 1089–1104.

Feng, S., Liu, D., Zhan, X., Wing, M. K., and Abecasis, G. R. (2014). Raremetal: fast and powerful meta-analysis for rare variants. *Bioinformatics (Oxford, England)* **30(19),** 2828–2829.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., and Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161,** 149–170.

Ganjgahi, H., Winkler, A. M., Glahn, D. C., Blangero, J., Donohue, B., Kochunov, P., and Nichols, T. E. (2018). Fast and powerful genome wide association of dense genetic data with high dimensional imaging phenotypes. *Nature Communications* **9(1),** 3254.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., and Nichols, T. E. (2012). Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *NeuroImage* **63(2),** 858–873.

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L., Auerbach, E. J., Behrens, T. E., ..., and Van Essen, D. C. (2016). The human connectome project′s neuroimaging approach. *Nature Neuroscience* **19(9),** 1175–1187.

Gong, W., Cheng, F., Rolls, E. T., Lo, C. Z., Huang, C. C., Tsai, S. J., Yang, A. C., Lin, C. P., and Feng, J. (2019). A powerful and efficient multivariate approach for voxel-level connectome-wide association studies. *Neuroimage* **188,** 628–641.

Gong, W., Wan, L., Lu, W., Ma, L., Cheng, F., Cheng, W., Grünewald, S., and Feng, J. (2018). Statistical testing and power analysis for brain-wide association study. *Medical Image Analysis* **47,** 15–30.

Gretton, A., Fukumizu, K., Teo, C. H., L. Song, B. S., and Smola, A. J. (2008). A kernel statistical test of independence. *Advances in Neural Information Processing Systems 20: 21st Annual Conference on Neural Information Processing Systems 2007* pages 585–592.

Grieve, S. M., Williams, L. M., Paul, R. H., Clark, C. R., and Gordon, E. (2007). Cognitive aging, executive function, and fractional anisotropy: A diffusion tensor MR imaging study. *American Journal of Neuroradiology* **28,** 226.

Guan, Z., Parmigiani, G., and Patil, P. (2019). Merging versus ensembling in multi-study machine learning: Theoretical insight from random effects. *arXiv* page 1905.07382.

Guillaume, B., Wang, C., Poh, J., Shen, M. J., Ong, M. L., Tan, P. F., Karnani, N., Meaney, M., and Qiu, A. (2018). Improving mass-univariate analysis of neuroimaging data by modelling important unknown covariates: Application to epigenome-wide association studies. *NeuroImage* **173,** 57–71.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., ..., and Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48(3),** 245–252.

Han, B. and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American Journal of Human Genetics* **88(5),** 586–598.

Han, F. and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity* **70(1),** 42–54.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1,** 297–310.

Heller, R., Heller, Y., and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100(2),** 503–510.

Hoeffding, W. (1948). A non-parametric test of independence. *Annals of Mathematical Statistics* **19(4),** 546–557.

Höffding, W. (1940). Maszstabinvariante korrelationstheorie. *Schr. Math. Inst. U. Inst. Angew. Math. Univ. Berlin* **5,** 181–233.

Hoffmann, T. J., Marini, N. J., and Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PloS One* **5(11),** e13584.

Hu, W., Zhang, A., Cai, B., Calhoun, V., and Wang, Y. P. (2019). Distance canonical correlation analysis with application to an imaging-genetic study. *Journal of Medical Imaging* **6(2),** 026501.

Huang, C. (2019). Advanced statistical learning methods for heterogeneous medical imaging data. *A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health* .

Huang, C., Thompson, P., Wang, Y., Yu, Y., Zhang, J., Kong, D., ..., and Alzheimer′s Disease Neuroimaging Initiative (2017). Fgwas: Functional genome wide association analysis. *NeuroImage* **159,** 107–121.

Huang, C., Zhu, H., and the Alzheimer′s Disease Neuroimaging Initiative (2021). Functional hybrid factor regression model for handling heterogeneity in imaging studies. *Preprint* .

Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., ..., and Alzheimer′s Disease Neuroimaging Initiative (2015). Fvgwas: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *NeuroImage* **118,** 613–627.

Hudson, R. R. and Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120(3),** 831–840.

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics* **7(2),** e1001289.

Ito, T. and Sugasawa, S. (2019). Improving the accuracy of confidence intervals and regions in multivariate random-effects meta-analysis. *arXiv:1906.08428v1 [stat.ME]* .

Jackson, D. and Riley, R. D. (2014). A refined method for multivariate meta-analysis and meta-regression. *Statistics in Medicine* **33,** 541–554.

Jahanshad, N., Kochunov, P. V., Sprooten, E., Mandl, R. C., Nichols, T. E., Almasy, L., Blangero, J., Brouwer, R. M., Curran, J. E., de Zubicaray, G. I., Duggirala, R., Fox, P. T., Hong, L. E., Landman, B. A., Martin, N. G., McMahon, K. L., Medland, S. E., Mitchell, ., and Glahn, D. C. (2013). Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: a pilot project of the ENIGMA-DTI working group. *Neuroimage* **81,** 455–469.

Jernigan, T. L., Brown, T. T., Hagler, Jr, D. J., Akshoomoff, N., Bartsch, H., Newman, E., ..., and Pediatric Imaging, Neurocognition and Genetics Study (2016). The pediatric imaging, neurocognition, and genetics (ping) data repository. *NeuroImage* **124(Pt B),** 1149–1154.

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *bioRxiv, doi: https://doi.org/10.1101/598110* .

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42,** 348–354.

Kang, H. M., Zaitlen, B. A., Wade, C. M., Kirby, A., Daly, D. H. M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178,** 1709–1723.

Kochunov, P., Jahanshad, N., Sprooten, E., Nichols, T. E., Mandl, R. C., Almasy, L., Booth, T., Brouwer, R. M., Curran, J. E., de Zubicaray, G. I., Dimitrova, R., Duggirala, R., Fox, P. T., Hong, L. E., Landman, B. A., Lemaitre, H., Lopez, L. M., Martin, N. G., ..., and Glahn, D. C. (2014). Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: Comparing meta and megaanalytical approaches for data pooling. *NeuroImage* **95,** 136–150.

Kong, D., An, B., Zhang, J., and Zhu, H. (2019). L2rm: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association* .

Kosch, R. and Jung, K. (2019). Conducting gene set tests in meta-analyses of transcriptome expression data. *Research Synthesis Methods* **10,** 99–112.

Kusic, D. M., Roberts, W. N., Jarvis, J. P., Zhang, P., Scheinfeldt, L. B., Rajula, K. D., Brenner, R., Dempsey, M. P., and Zajic, S. C. (2020). rs11670527 Upstream of ZNF264 Associated with Body Mass Index in the Coriell Personalized Medicine Collaborative. *Military Medicine* pages 649–655.

Lagani, V., Karozou, A. D., Gomez-Cabrero, D., Silberberg, G., and Tsamardinos, I. (2016). A comparative evaluation of data-merging and meta-analysis methods for reconstructing gene-gene interactions. *BMC Bioinformatics* **17,** S194.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics* **95(1),** 5–23.

Lee, S., Sun, W., Wright, F. A., and Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **104,** 303–316.

Lee, S., Teslovich, T. M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *American Journal of Human Genetics* **93(1),** 42–53.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics (Oxford, England)* **13(4),** 762–775.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3,** 1724–1735.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics* **83(3),** 311–321.

Lin, D. Y. and Tang, Z. Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics* **89(3),** 354–367.

Lin, D. Y., Tao, R., Kalsbeek, W., Zeng, D., Gonzalez, F., Fernández-Rhodes, L., Graff, M., Koch, G., North, K. E., and Heiss, G. (2014). Genetic association analysis under complex survey sampling: The hispanic community health study/study of latinos. *American Journal of Human Genetics* **95(6),** 675–688.

Lin, D. Y. and Zeng, D. (2010a). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology* **34(1),** 60–66.

Lin, D. Y. and Zeng, D. (2010b). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97(2),** 321–332.

Lin, D. Y., Zeng, D., and Tang, Z. Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences of the United States of America* **110(30),** 12247–12252.

Lin, J.-A., Zhu, H., Mihye, A., Sun, W., Ibrahim, J. G., and Alzheimer′s Neuroimaging Initiative (2014). Functional-mixed effects models for candidate genetic mapping in imaging genetic studies. *Genetic epidemiology* **38,** 680–691.

Lippert, C., Xiang, J., Horta, D., Widmer, C., Kadie, C., Heckerman, D., and Listgarten, J. (2014). Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics (Oxford, England)* **30(22),** 3206–3214.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63,** 1079–1088.

Liu, D. J. and Leal, S. M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genetics* **6(10),** e1001156.

Liu, L. Y.-F. (2018). Advanced statistical learning techniques for high-dimensional imaging data. *A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research* .

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., and Price, A. L. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47,** 284–290.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* **5(2),** e1000384.

Mirzaalian, H., Ning, L., Savadjiev, P., Pasternak, O., Bouix, S., Michailovich, O., Grant, G., Marx, C. E., Morey, R. A., Flashman, L. A., George, M. S., McAllister, T. W., Andaluz, N., Shutter, L., Coimbra, R., Zafonte, R. D., Coleman, M. J., Kubicki, M., Westin, C. F., Stein, M. B., Shenton, M. E., and Rathi, Y. (2016). Inter-site and inter-scanner diffusion mri data harmonization. *NeuroImage* **135,** 311–323.

Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutation Research* **615,** 28–56.

Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* **34(2),** 188–193.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., ..., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics* **7(3),** e1001322.

Pan, W. (2009). Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic Epidemiology* **33(6),** 497–507.

Pan, W., Wang, X., Zhang, H., Zhu, H., and Zhu, J. (2019). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association* **00(0),** 1–11.

Patil, P. and Parmigiani, G. (2018). Training replicable predictors in multiple studies. *Proceedings of the National Academy of Sciences* **115,** 2578–2583.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58,** 240–242.

Petersen, A. and Müller, H.-G. (2015). Functional data analysis for density functions by transformation to a hilbert space. *Annals of Statistics* **44,** 183–218.

Pizer, S., Fritsch, D., Yushkevich, P., Johnson, V., and Chaney, E. (1999). Segmentation, registration, and measurement of shape variation via image object shape. *IEEE Transactions on Medical Imaging* **18,** 851–865.

Pluta, D., Ombao, H., Chen, C., Xue, G., Moyzis, R., and Yu, Z. (2018). Adaptive mantel test for association testing in imaging genetics data. *arXiv:1712.07270* .

Preda, C., Saporta, G., and Lévéder, C. (2007). Pls classification of functional data. *Computational Statistics* **22,** 223–235.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., and Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics* **86(6),** 832–838.

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

Sadigh-Eteghad, S., Talebi, M., and Farhoudi, M. (2012). Association of apolipoprotein e epsilon 4 allele with sporadic late onset alzheimer's disease. a meta-analysis. *Neurosciences Journal* **17,** 321–326.

Satterthwaite, T. D., Connolly, J. J., Ruparel, K., Calkins, M. E., Elliott, C. J. M. A., ..., and Gur, R. E. (2016). The philadelphia neurodevelopmental cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *NeuroImage* **124(Pt B),** 1115–1119.

Schölkopf, B., Smola, A. J., and Müller, K.-R. (1997). Kernel principal component analysis. *In: International Conference on Artificial Neural Networks. Springer* pages 583–588.

Schwarzer, G. (2007). meta: An r package for meta-analysis. *R News* **7(3),** 40–45.

Scuteri, A., Sanna, S., Chen, W. M., Uda, M., Albai, G., Strait, J., ..., and Abecasis, G. R. (2007). Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLoS Genetics* **3(7),** e115.

Shen, L. and Thompson, P. M. (2019). Brain imaging genomics: Integrated analysis and machine learning. *Proceedings of the IEEE* pages 1–38.

Shi, J., Thompson, P. M., Gutman, B., and Wang, Y. (2013). Surface fluid registration of conformal representation: Application to detect disease burden and genetic influence on hippocampus. *NeuroImage* **78,** 111–134.

Sidik, K. and Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54,** 367–384.

Somerville, L. H., Bookheimer, S. Y., Buckner, R. L., Burgess, G. C., Curtiss, S. W., Dapretto, M., Elam, J. S., Gaffrey, M. S., Harms, M. P., Hodge, C., Kandala, S., Kastman, E. K., Nichols, T. E., Schlaggar, B. L., Smith, S. M., Thomas, K. M., Yacoub, E., Van Essen, D. C., and Barch, D. M. (2018). The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5-21 year olds. *NeuroImage* **183,** 456–468.

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., ..., and Alzheimer's Disease Neuroimaging Initiative (2010). Voxelwise genome-wide association study (vgwas). *NeuroImage* **53(3),** 1160–1174.

Styner, M., Lieberman, J. A., Pantazis, D., and Gerig, G. (2004). Boundary and medial shape analysis of the hippocampus in schizophrenia. *Medical Image Analysis* **8,** 197–203. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12,** e1001779.

Sudlow, C. and et. al. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* **12(3),** e1001779.

Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology* **37(4),** 334–344.

Sun, R., Hui, S., Bader, G. D., and Kraft, X. L. P. (2019). Powerful gene set analysis in gwas with the generalized berk-jones statistic. *PLoS Genetics* **15(3),**.

Szekely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35,** 2769–2794.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of statistics* **35,** 2769–2794.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics* **35(6),** 2769–2794.

Taminau, J., Lazar, C., Meganck, S., and Nowé, A. (2014). Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN Bioinformatics* **2014,** 345106.

Tang, Z. Z., Bunn, P., Tao, R., Liu, Z., and Lin, D. Y. (2017). Premeta: a tool to facilitate meta-analysis of rare-variant associations. *BMC Genomics* **18(1),** 160.

Tang, Z. Z. and Lin, D. Y. (2013). Mass: meta-analysis of score statistics for sequencing studies. *Bioinformatics (Oxford, England)* **29(14),** 1803–1805.

Tang, Z. Z. and Lin, D. Y. (2014). Meta-analysis of sequencing studies with heterogeneous genetic associations. *Genetic Epidemiology* **38(5),** 389–401.

Thompson, P. and et. al. (2019). Enigma and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *PsyArXiv* pages 1–41.

Thompson, P. M., Hayashi, K. M., de Zubicaray, G. I., Janke, A. L., Rose, S. E., Semple, J., Hong, M. S., Herman, D. H., Gravano, D., Doddrell, D. M., and Toga, A. W. (2004). Mapping hippocampal and ventricular change in alzheimer disease. *NeuroImage* **22,** 1754–1766.

Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., ..., Alzheimer′s Disease Neuroimaging Initiative, EPIGEN Consortium, IMAGEN Consortium, and Saguenay Youth Study Group (2014). The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging and Behavior* **8(2),** 153–182.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* **61(3),** 611–622.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software* **36(3),** 1–48.

Villani, C. (2003). Topics in optimal transportation. graduate studies in mathematics 58. *American Mathematical Society, Providence, RI.* **MR1964483,**.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics* **45,** 1863–1894.

Wang, Q., Cagna, B., Chaminade, T., and Takerkart, S. (2020). Inter-subject pattern analysis: A straightforward and powerful scheme for group-level mvpa. *NeuroImage* **204,**.

Wang, Y., Ibrahim, J. G., and Zhu, H. (2000). Partial least squares for functional joint models with applications to the alzheimer′s disease neuroimaging initiative study. *Biometrics* **76,** 1109–1119.

Wang, Y., Liu, A., Mills, J. L., Boehnke, M., Wilson, A. F., Bailey-Wilson, J. E., Xiong, M., Wu, C. O., and Fan, R. (2015). Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genetic Epidemiology* **39,** 259–275.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., ..., and Alzheimer′s Disease Neuroimaging Initiative (2017). Recent publications from the alzheimer′s disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials. *Alzheimer′s & Dementia : the Journal of the Alzheimer's Association* **13(4),** e1–e85.

Wen, C., Yang, Y., Xiao, Q., Huang, M., Pan, W., and for the Alzheimer′s Disease Neuroimaging Initiative (2020). Genome-wide association studies of brain imaging data via weighted distance correlation. *Bioinformatics* pages 4942–4950.

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* **26(17),** 2190–2191.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare variant association testing for sequencing data using the sequence kernel association test (skat). *American Journal of Human Genetics* **89,** 82–93.

Xu, L., Tan, A. C., Winslow, R. L., and Geman, D. (2008). Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics* **9,** 125.

Zeng, D. and Lin, D. Y. (2015). On random-effects meta-analysis. *Biometrika* **102(2),** 281–294.

Zhang, J. (2018). Advanced methods for discovering genetic markers associated with high dimensional imaging data. *A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.* .

Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics* **35,** 1052–1079.

Zhang, Y., Bernau, C., Parmigiani, G., and Waldron, L. (2020). The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models. *Biostatistics* **2020,** 253–268.

Zhao, B., Ibrahim, J. G., Li, Y., Li, T., Wang, Y., Shan, Y., Zhu, Z., Zhou, F., Zhang, J., Huang, C., Liao, H., Yang, L., Thompson, P. M., and Zhu, H. (2019). Heritability of regional brain volumes in large-scale neuroimaging and genetic studies. *Cerebral Cortex* **29,** 2904–2914.

Zhao, B., Li, T., Smith, S. M., Xiong, D., Wang, X., Yang, Y., Luo, T., Zhu, Z., Shan, Y., Matoba, N., Sun, Q., Yang, Y., Hauberg, M. E., Bendl, J., Fullard, J. F., Roussos, P., Lin, W., Li, Y., Stein, J. L., and Zhu, H. (2020). Common variants contribute to intrinsic human brain functional networks. *bioRxiv* .

Zhao, B., Li, T., Yang, Y., Wang, X., Luo, T., Shan, Y., Zhu, Z., Xiong, D., Hauberg, M. E., Bendl, J., Fullard, J. F., Roussos, P., Li, Y., Stein, J. L., and Zhu, H. (2020). Common genetic variation influencing human white matter microstructure. *bioRxiv* .

Zhao, B., Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., Wang, X., Yang, L., Zhou, F., Zhu, Z., Zhu, H., Alzheimer′s Disease Neuroimaging Initiative, and Pediatric Imaging, Neurocognition and Genetics (2019). Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature Genetics* **51,** 1637–1644.

Zhao, B., Zhang, J., Ibrahim, J. G., Luo, T., Santelli, R. C., Li, Y., Li, T., Shan, Y., Zhu, Z., Zhou, F., Liao, H., Nichols, T. E., and Zhu, H. (2019). Large-scale gwas reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n = 17,706). *Molecular Psychiatry* .

Zhou, H., Zhang, Y., Ithapu, V., Johnson, S., Wahba, G., and Singh, V. (2017). When can multi-site datasets be pooled for regression? hypothesis tests, l2-consistency and neuroscience applications. *Proceedings of machine learning research* **70,** 4170–4179.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44(7),** 821–4.

Zhu, H., Li, R., and Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Annals of Statistics* **40(5),** 2634–2666.

Zhu, J., Pan, W., Zheng, W., and Wang, X. (2019). Ball: An r package for detecting distribution difference and association in metric spaces. *arXiv: 811.03750v3* .

Zhu, L., Xu, K., Li, R., and Zhong (2017). Projection correlation between two random vectors. *Biometrika* **104,** 829–843.