HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2016 April 22.

Published in final edited form as:

J Am Stat Assoc. 2015 ; 110(509): 289–302. doi:10.1080/01621459.2014.892008.

SPReM: Sparse Projection Regression Model For High-dimensional Linear Regression *

Qiang Sun [Ph.d student],
Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7420.

Hongtu Zhu [Professor of Biostatistics],
Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7420.

Yufeng Liu [Professor of Statistics],
Department of Statistics and Operation Research, University of North Carolina at Chapel Hill, CB 3260, Chapel Hill, NC 27599.

Joseph G. Ibrahim [Alumni Distinguished Professor of Biostatistics], and
Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7420.

for the Alzheimer';s Disease Neuroimaging Initiative

Qiang Sun: qsun@bios.unc.edu; Hongtu Zhu: hzhu@bios.unc.edu; Yufeng Liu: yiu@email.unc.edu; Joseph G. Ibrahim: ibrahim@bios.unc.edu

Abstract

The aim of this paper is to develop a sparse projection regression modeling (SPReM) framework to perform multivariate regression modeling with a large number of responses and a multivariate covariate of interest. We propose two novel heritability ratios to simultaneously perform dimension reduction, response selection, estimation, and testing, while explicitly accounting for correlations among multivariate responses. Our SPReM is devised to specifically address the low statistical power issue of many standard statistical approaches, such as the Hotelling's T^2 test statistic or a mass univariate analysis, for high-dimensional data. We formulate the estimation problem of SPREM as a novel sparse unit rank projection (SURP) problem and propose a fast optimization algorithm for SURP. Furthermore, we extend SURP to the sparse multi-rank projection (SMURP) by adopting a sequential SURP approximation. Theoretically, we have systematically investigated the convergence properties of SURP and the convergence rate of SURP estimates. Our simulation results and real data analysis have shown that SPReM out-performs other state-of-the-art methods.

*Address for correspondence and reprints: Hongtu Zhu, Ph.D., hzhu@bios.unc.edu; Phone No: 919-966-7272.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp−content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. We thank the Editor, the Associate Editor, and two anonymous referees for valuable suggestions, which greatly helped to improve our presentation.

Keywords

heritability ratio; imaging genetics; multivariate regression; projection regression; sparse; wild bootstrap

1 Introduction

Multivariate regression modeling with a multivariate response $\mathbf{y} \in \mathbb{R}^q$ and a multivariate covariate $\mathbf{x} \in \mathbb{R}^p$ is a standard statistical tool in modern high-dimensional inference, with wide applications in various large-scale applications, such as genome-wide association studies (GWAS) and neuroimaging studies. For instance, in GWAS, our primary problem of interest is to identify genetic variants (**x**) that cause phenotypic variation (**y**). Specifically, in imaging genetics, multivariate imaging measures (**y**), such as volumes of regions of interest (ROIs), are phenotypic variables, whereas covariates (**x**) include single nucleotide polymorphisms (SNPs), age, and gender, among others. The joint analysis of imaging and genetic data may ultimately lead to discoveries of genes for neuropsychiatric and neurological disorders such as autism and schizophrenia (Scharinger et al., 2010; Paus, 2010; Peper et al., 2007; Chiang et al., 2011a,b). Moreover, in many neuroimaging studies, there is a great interest in the use of imaging measures (**x**), such as functional imaging data and cortical and subcortical structures, to predict multiple clinical and/or behavioral variables (**y**) (Knickmeyer et al., 2008; Lenroot and Giedd, 2006). This motivates us to systematically investigate a multivariate linear model with a multivariate response **y** and a multivariate covariate **x**.

Throughout this paper, we consider n independent observations $(\mathbf{y}_i, \mathbf{x}_i)$ and a Multivariate Linear Model (MLM) given by

$$\boldsymbol{Y} = \mathbf{XB} + \mathbf{E}, \text{ or} \mathbf{y}_i = \mathbf{B}^T \mathbf{x}_i + \mathbf{e}_i, \quad (1)$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, and $\mathbf{B} = (\beta_{jl})$ is a $p \times q$ coefficient matrix with $\text{rank}(\mathbf{B}) = r^* \leq \min(p, q)$. Moreover, the error term $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^T$ has $E(\mathbf{e}_i) = 0$ and $\text{Cov}(\mathbf{e}_i) = \Sigma_R$ for all i, where Σ_R is a $q \times q$ matrix. Many hypothesis testing problems of interest, such as comparison across groups, can often be formulated as

$$H_0: \mathbf{CB} = \mathbf{B}_0 \text{ v.s. } H_1: \mathbf{CB} \neq \mathbf{B}_0, \quad (2)$$

where **C** is an $r \times p$ matrix and $\mathbf{B}_0$ is an $r \times q$ matrix. Without loss of generality, we center the covariates, standardize the responses, and assume $\text{rank}(\mathbf{C}) = r$.

We focus on a specific setting that q is relatively large, but p is relatively small. Such a setting is general enough to cover two-sample (or multi-sample) hypothesis testing for high-dimensional data (Chen and Qin, 2010; Lopes et al., 2011). There are at least three major challenges including (i) a large number of regression parameters, (ii) a large covariance matrix, and (iii) correlations among multivariate responses. When the number of responses and the number of covariates are even moderately high, fitting the conventional MLM usually requires estimating a $p \times q$ matrix of regression coefficients, whose number pq can

be much larger than n. Although accounting for complicated correlations among multiple responses is important for improving the overall prediction accuracy of multivariate analysis (Breiman and Friedman, 1997; Cook et al., 2010), it requires estimating $q(q+1)/2$ unknown parameters in an unstructured covariance matrix.

There is a great interest in the development of efficient methods for handling MLMs with large q. Four popular traditional methods include the mass univariate analysis, the Hotelling's T^2 test, partial least squares regression, and dimension reduction methods. As pointed by Klei et al. (2008) and many others, testing each response variable individually in the mass univariate analysis requires a substantial penalty of controlling for multiplicity. The Hotelling's T^2 test is not well-defined, when $q > n$. Even when $q \leq n$, the power of the Hotelling's T^2 can be very low if q is nearly as large as n. Partial least squares regression (PLSR) aims to find a linear regression model by projecting **y** and **x** to a smaller latent space (Chun and Keles, 2010; Krishnan et al., 2011), but it focuses on prediction and classification. Although dimension reduction techniques, such as principal component analysis (PCA), are considered to reduce the dimensions of both the response and covariates (Formisano et al., 2008; Kherif et al., 2002; Rowe and Hoffmann, 2006; Teipel et al., 2007), most of the methods ignore the variation of covariates and their associations with responses. Thus, such methods can be sub-optimal for our problem.

Some recent developments primarily include regularization methods and envelope models (Peng et al., 2010; Tibshirani, 1996; Breiman and Friedman, 1997; Cook et al., 2010, 2013; Lin et al., 2012). Cook, Li and Chiaromonte (2010) developed a powerful envelope modeling framework for MLMs. Such envelope methods use dimension reduction techniques to remove the immaterial information, while achieving efficient estimation of the regression coefficients by accounting for correlations among the response variables. However, the existing envelope methods are limited to the $n > \max(p, q)$ scenario. Recently, much attention has been given to regularization methods for enforcing sparsity in **B** (Peng et al., 2010; Tibshirani, 1996). These regularization methods, however, do not provide a standard inference tool (e.g., standard deviation) on the regression coefficient matrix **B**. Lin et al. (2012) developed a projection regression model (PRM) and its associated estimation procedure to assess the relationship between a multivariate phenotype and a set of covariates without providing any theoretical justification.

This paper presents a new general framework, called sparse projection regression model (SPReM), for simultaneously performing dimension reduction, response selection, estimation, and testing in a general high dimensional MLM setting. We introduce two novel heritability ratios, which extend the idea of principal components of heritability from familial studies (Klei et al., 2008; Ott and Rabinowitz, 1999), for MLM and overcome over-fitting and noise accumulation in high dimensional data by enforcing the sparsity constraint. We develop a fast algorithm for both **s**parse **u**nit **r**ank **p**rojection (SURP) and **s**parse **mu**lti-**r**ank **p**rojection (SMURP). Furthermore, a test procedure based on the wild-bootstrap method is proposed, which leads to a single p–value for the test of an association between all response variables and covariates of interest, such as genetic markers. Simulations show that our method can control the overall Type I error well, while achieving high statistical power.

Section 2 of this paper introduces the SPReM framework. We introduce a novel deflation procedure to extract the most informative directions for testing hypotheses of interest. Simulation studies and an imaging genetic example are used to examine the finite sample performance of SPReM in Section 3. We present concluding remarks in Section 4.

2 Sparse Projection Regression Model

2.1 Model Setup and Heritability Ratios

We introduce SPReM as follows. The key idea of our SPReM is to appropriately project $\mathbf{y}_i$ in a high-dimensional space onto a low-dimensional space, while accounting for the correlation structure Σ_R among the response variables and the hypothesis test in (2). Let $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$ be a $q \times k$ nonrandom and unknown direction matrix, where $\mathbf{w}_j$ are $q \times 1$ vectors. A projection regression model (PRM) is given by

$$\mathbf{W}^T\mathbf{y}_i=(\mathbf{BW})^T\mathbf{x}_i+\mathbf{W}^T\mathbf{e}_i=\boldsymbol{\beta}_{\mathbf{w}}^T\mathbf{X}_i+\varepsilon_i, \quad (3)$$

where $\beta_{\mathbf{w}}$ is a $p \times k$ regression coefficient matrix and the random vector ε_i has $E(\varepsilon_i) = \mathbf{0}$ and $\mathrm{Cov}(\varepsilon_i) = \mathbf{W}^T\Sigma_R\mathbf{W}$. When $k = 1$, PRM reduces to the pseduo-trait model considered in (Amos et al., 1990; Amos and Laing, 1993; Klei et al., 2008; Ott and Rabinowitz, 1999). If $k << \min(n, q)$ and $\mathbf{W}$ were known, then one could use likelihood (or estimating equation) based methods to efficiently estimate $\beta_{\mathbf{w}}$, and (2) would reduce approximately to

$$H_{0W}:\mathbf{C}\boldsymbol{\beta}_{\mathbf{w}}=\mathbf{b}_0 \text{ v.s. } H_{1W}:\mathbf{C}\boldsymbol{\beta}_{\mathbf{w}} \neq \mathbf{b}_0, \quad (4)$$

where $\mathbf{C}\beta_{\mathbf{w}} = \mathbf{CBW}$ and $\mathbf{b}_0 = \mathbf{B}_0\mathbf{W}$. In this case, the number of null hypotheses in (4) is much smaller than that of (2). It is also expected that different $\mathbf{W}$'s strongly influence the statistical power of testing the hypotheses in (2).

A fundamental question arises "how do we determine an 'optimal' $\mathbf{W}$ to achieve good statistical power of testing (2)?" To determine $\mathbf{W}$, we develop a novel deflation approach to sequentially determine each column of $\mathbf{W}$ at a time starting from $\mathbf{w}_1$ to $\mathbf{w}_k$. We focus on how to determine $\mathbf{w}_1$ below and then discuss how to extend it to the scenario with $k > 1$.

To determine an optimal $\mathbf{w}_1$, we consider two principles. The first principle is to maximize the mean value of the square of the signal-to-noise ratio, called the heritability ratio, for model (3). For each i, the signal-to-noise ratio in model (3) is defined as the ratio of mean to standard deviation of a signal or measurement $\mathbf{w}^T\mathbf{y}_i$, denoted by $\mathrm{SNR}_i = \mathbf{w}^T\mathbf{B}^T\mathbf{x}_i/(\mathbf{w}^T\Sigma_R\mathbf{w})^{0.5}$. Thus, the heritability ratio (HR) is given by

$$\mathrm{HR}(\mathbf{w})=n^{-1}\sum_{i=1}^{n}\mathrm{SNR}_i^2=\frac{\mathbf{w}^T\mathbf{B}^T S_X\mathbf{Bw}}{\mathbf{w}^T\Sigma_R\mathbf{w}}, \quad (5)$$

where $S_X=n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$. The HR has several important interpretations. If the $\mathbf{x}_i$ are independently and identically distributed (i.i.d) with $E(\mathbf{x}_i) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{x}_i) = \Sigma_X$, then as $n \to \infty$, we have

$$\mathrm{HR}(\mathbf{w}) \to^{p} \frac{\mathbf{w}^T\mathbf{B}^T\Sigma_X\mathbf{B}\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}} = \frac{\mathrm{Var}(\mathbf{w}^T\mathbf{B}^T\mathbf{x}_i)}{\mathrm{Var}(\varepsilon_i)},$$

where $\to^p$ denotes convergence in probability. Thus, HR(**w**) is close to the ratio of the variance of signal $\mathbf{w}^T\mathbf{B}^T\mathbf{x}_i$ to that of noise ε_i. Moreover, HR(**w**) is close to the heritability ratio considered in (Amos et al., 1990; Amos and Laing, 1993; Klei et al., 2008; Ott and Rabinowitz, 1999) for familial studies, but we define HR from a totally different perspective. With such new perspective, one can easily define HR for more general designs, such as cross-sectional or longitudinal design. One might directly maximize HR(**w**) to calculate an 'optimal' $\mathbf{w}_1$, but such a $\mathbf{w}_1$ can be sub-optimal for testing the hypotheses in (2) as discussed below.

The second principle is to explicitly account for the hypotheses in (2) under model (1) and the reduced ones in (4) under model (3). We define four spaces associated with the null and alternative hypotheses of (2) and (4) as follows:

$$S_{H_0}=\{\mathbf{B}:\mathbf{CB}=\mathbf{B}_0\}, S_{H_W}=\{\mathbf{B}:\mathbf{CBW}=\mathbf{B}_0\mathbf{W}\},\ S_{H_1}=\{\mathbf{B}:\mathbf{CB}\neq\mathbf{B}_0\}, S_{H_{1W}}=\{\mathbf{B}:\mathbf{CBW}\neq\mathbf{B}_0\mathbf{W}\}.$$

It can be shown that they satisfy the following relationship:

$$S_{H_0}\subset S_{H_W}\ \text{and}\ S_{H_{1W}}\subset S_{H_1}\ \text{for any}\ \mathbf{W}\neq\mathbf{0}.$$

Due to potential information loss during dimension reduction, both $S_{HW} - S_{H0}$ and $S_{H1} - S_{H1W}$ may not be the empty set, but we need to choose **W** such that $S_{H1} - S_{H1W} \approx$ ø. The next question is how to achieve this.

We consider a data transformation procedure. Let $\mathbf{C}_1$ be a $(p-r) \times p$ matrix such that

$$\mathrm{rank}[\mathbf{C}^T\mathbf{C}_1^T]=p\ \text{and}\ \mathbf{C}\mathbf{C}_1^T=\mathbf{0}. \quad (6)$$

Let $\mathbf{D}=[\mathbf{C}^T\mathbf{C}_1^T]^T$ be a $p \times p$ matrix and $\tilde{\mathbf{x}}_i=(\tilde{\mathbf{x}}_{i1}^T,\tilde{\mathbf{x}}_{i2}^T)=\mathbf{D}^{-T}\mathbf{x}_i$ be a $p \times 1$ vector, where $\tilde{\mathbf{x}_{i1}}$ and $\tilde{\mathbf{x}_{i2}}$ are, respectively, the $r \times 1$ and $(p - r) \times 1$ subvectors of $\tilde{\mathbf{x}_i}$. We define $\tilde{\mathbf{B}}=[\tilde{\mathbf{B}}_1^T\tilde{\mathbf{B}}_2^T]^T=\mathbf{DB}$, or $\mathbf{B} = \mathbf{D}^{-1}\tilde{\mathbf{B}}$, where $\tilde{\mathbf{B}_1}$ and $\tilde{\mathbf{B}_2}$ are, respectively, the first r rows and the last $p - r$ rows of $\tilde{\mathbf{B}}$. Therefore, model (3) can be rewritten as

$$\mathbf{W}^T\mathbf{y}_i=(\mathbf{D}^{-1}\tilde{\mathbf{B}}\mathbf{W})^T\mathbf{x}_i+\mathbf{W}^T\mathbf{e}_i=\mathbf{W}^T(\tilde{\mathbf{B}}_1-\mathbf{B}_0)^T\tilde{\mathbf{x}}_{i1}+\mathbf{W}^T\mathbf{B}_0^T\tilde{\mathbf{x}}_{i1}+\mathbf{W}^T\tilde{\mathbf{B}}_2^T\tilde{\mathbf{x}}_{i2}+\mathbf{W}^T\mathbf{e}_i. \quad (7)$$

In (7), due to (6), we only need to consider the transformed covariate vector $\tilde{\mathbf{x}_{i1}}$, which contains useful information associated with $\tilde{\mathbf{B}_1} - \mathbf{B}_0 = \mathbf{CB} - \mathbf{B}_0$.

We define a generalized heritability ratio based on model (7). Specifically, for each i, we define a new signal-to-noise ratio as the ratio of mean to standard deviation of signal $\mathbf{w}^T(\tilde{\mathbf{B}_1}$

$-\mathbf{B}_0)^T\mathbf{x}_{i1}^{\sim}+\mathbf{w}^T\mathbf{e}_i$, denoted by $\mathrm{SNR}_{i,\mathrm{C}}=\mathbf{w}^T(\mathbf{B}_1^{\sim}-\mathbf{B}_0)^T\mathbf{x}_{i1}^{\sim}/(\mathbf{w}^T\Sigma_R\mathbf{w})^{0.5}$. The generalized heritability ratio is then defined as

$$\mathrm{GHR}(\mathbf{w};\mathbf{C})=n^{-1}\sum_{i=1}^{n}\mathrm{SNR}_{i,\mathrm{C}}^2=\frac{\mathbf{w}^T(\tilde{\mathbf{B}}_1-\mathbf{B}_0)^T S_{\tilde{X}_1}(\tilde{\mathbf{B}}_1-\mathbf{B}_0)\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}, \quad (8)$$

where $S_{\tilde{X}_1}=n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{x}}_{i1}\tilde{\mathbf{x}}_{i1}^T$. If the $\mathbf{x}_i$s are random, then we have

$$\mathrm{GHR}(\mathbf{w};\mathbf{C})\to^p\frac{\mathbf{w}^T(\tilde{\mathbf{B}}_1-\mathbf{B}_0)^T\mathrm{Cov}(\tilde{\mathbf{x}}_{i1})(\tilde{\mathbf{B}}_1-\mathbf{B}_0)\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}=\frac{\mathbf{w}^T\Sigma_C\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}, \quad (9)$$

where $\Sigma_C=(\mathbf{B}_1^{\sim}-\mathbf{B}_0)^T(\mathbf{D}^{-T}\Sigma_X\mathbf{D}^{-1})_{(r,r)}(\mathbf{B}_1^{\sim}-\mathbf{B}_0)$, and $(\mathbf{D}^{-T}\Sigma_X\mathbf{D}^{-1})_{(r,r)}$ is the upper $r\times r$ submatrix of $\mathbf{D}^{-T}\Sigma_X\mathbf{D}^{-1}$. Particularly, if $\mathbf{C}=[\mathbf{I}_r\ \mathbf{0}]$, then Σ_C reduces to $\mathbf{w}^T(\mathbf{B}_1^{\sim}-\mathbf{B}_0)^T(\Sigma_X)_{(1,1)}(\mathbf{B}_1^{\sim}-\mathbf{B}_0)\mathbf{w}$, in which $(\Sigma_X)_{(1,1)}$ is the upper $r\times r$ submatrix of Σ_X. Thus, GHR(**w; C**) can be interpreted as the ratio of the variance of $\mathbf{w}^T(\mathbf{B}_1^{\sim}-\mathbf{B}_0)^T\mathbf{x}_{i1}^{\sim}$ relative to that of $\mathbf{w}^T\mathbf{e}_i$. We propose to calculate an optimal $\mathbf{w}_*$ as follows:

$$\mathbf{w}_*=\underset{\mathbf{w}}{\mathrm{argmax}}\,\mathrm{GHR}(\mathbf{w};\mathbf{C}). \quad (10)$$

We expect that such an optimal $\mathbf{w}_*$ can substantially reduce the size of both $S_{H_1}-S_{H_1W}$ and $S_{H_W}-S_{H_0}$ and thus the use of such an optimal $\mathbf{w}_*$ can enhance the power of testing the hypotheses in (2). Without loss of generality, we assume $\mathbf{B}_0=\mathbf{0}$ from now on.

We consider a simple example to illustrate the appealing properties of GHR(**w; C**).

Example We consider model (1) with $p=q=5$ and want to test the nonzero effect of the first covariate on all five responses. In this case, $r=1$, $\mathbf{C}=(1,0,0,0,0)$, $\mathbf{B}_0=(0,0,0,0,0)$, and $\mathbf{D}=I_5$, which is a 5×5 identity matrix. Without loss of generality, it is assumed that $(\Sigma_X)_{(1,1)}=1$.

We consider three different cases of Σ_R and **B**. In the first case, we set $\Sigma_R=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_5^2)$ and the first column of **B** to be (1, 0, 0, 0, 0). It follows from (8) that

$$\mathrm{GHR}(\mathbf{w};\mathbf{C})=\frac{w_1^2}{\sigma_1^2w_1^2+\sigma_2^2w_2^2+\cdots+\sigma_5^2w_5^2}\ \text{and}\ \mathbf{w}_*^T=(c_0,0,0,0,0),$$

where c_0 is any nonzero scalar. Therefore, $\mathbf{w}_*$ picks out the first response, which is the sole one that is associated with the first covariate.

In the second case, we set $\Sigma_R=\mathrm{diag}(\sigma_1^2,\cdots,\sigma_5^2)$ with $\sigma_1^2\geq\cdots\geq\sigma_5^2$ and the first row of **B** to be (1, 1, 0, 0, 0). It follows from (8) that

$$\mathrm{GHR}(\mathbf{w};\mathbf{C})=\frac{(w_1+w_2)^2}{\sigma_1^2w_1^2+\sigma_2^2w_2^2+\cdots+\sigma_5^2w_5^2}\ \text{and}\ \mathbf{w}_*^T=(\frac{\sigma_2^2}{\sigma_1^2}c_0,c_0,0,0,0),$$

where c_0 is any nonzero scalar. Therefore, $\mathbf{w}_*$ picks out both the first and second response with larger weight on the second component. This is desirable since β_{11} and β_{21} are equal in terms of strength of effect and the noise level for the second response is smaller than that of the first one.

In the third case, we set the first row of $\mathbf{B}$ to be (1, 1, 0, 0, 0) and the first and second columns of Σ_R are set as $\sigma^2(1, \rho, 0, 0, 0)$ and $\sigma^2(\rho, 1, 0, 0, 0)$, respectively. It follows from (8) that

$$\mathrm{GHR}(\mathbf{w};\mathbf{C})=\frac{(w_1+w_2)^2}{\sigma^2 w_1^2+2\sigma_2^2\rho w_1 w_2+\sigma_2^2 w_2^2+Q(w_3,w_4,w_5)}\text{and}\mathbf{w}_*^T=(c_0,c_0,0,0,0),$$

where $Q(w_3,w_4,w_5)$ is a non-negative quadratic form of (w_3,w_4,w_5). Thus, the optimal $\mathbf{w}_*$ chooses the first two responses with equal weight, since they are correlated with each other with same variance and $\beta_{11} = \beta_{21} = 1$.

For high dimensional data, it is difficult to accurately estimate $\mathbf{w}_*$, since the sample covariance matrix estimator $\hat{\Sigma_R}$ can be either ill-conditioned or not invertible for large $q > n$. One possible solution is to focus only on a small number of important features for testing. However, a naive search for the best subset is NP-hard. We develop a penalized procedure to address these two problems, while obtaining a relatively accurate estimate of $\mathbf{w}$. Let $\tilde{\Sigma_R}$ and $\hat{\Sigma_C}$ be, respectively, estimators of Σ_R and Σ_C. Here we use $\tilde{\Sigma_R}$ to denote the covariance estimator other than sample covariance matrix $\hat{\Sigma_R}$. To obtain $\hat{\Sigma_C}$, we need to plug $\hat{\mathbf{B}}$, an estimator of $\mathbf{B}$, into Σ_C. Without loss of generality, we consider the ordinary least squares estimate of $\mathbf{B}$. By imposing a sparse structure on $\mathbf{w}_1$, we recast the optimization problem as

$$\max\{\frac{\mathbf{w}^T\hat{\Sigma}_C\mathbf{w}}{\mathbf{w}^T\tilde{\Sigma}_R\mathbf{w}}\}\text{s.t.}\|\mathbf{w}\|_1\le t, \quad (11)$$

where $\|\cdot\|_1$ is the L_1 norm and $t > 0$.

2.2 Sparse Unit Rank Projection

When $r = 1$, we call the problem in (10) as the unit rank projection problem and its corresponding sparse version in (11) as the sparse unit rank projection (SURP) problem. Actually, many statistical problems, such as two-sample test and marginal effect test problems, can be formulated as the unit rank projection problem (Lopes et al., 2011). We consider two cases including $\ell = (\mathbf{CB})^T = \mathbf{0}$ and $\ell = (\mathbf{CB})^T \neq \mathbf{0}$. When $\ell = (\mathbf{CB})^T = \mathbf{0}$, the solution set of (8) is trivial, since any $\mathbf{w} \neq \mathbf{0}$ is a solution of (8). As discussed later, this property is extremely important for controlling the type I error rate.

When $\ell = (\mathbf{CB})^T \neq \mathbf{0}$, (8) reduces to the following optimization problem:

$$\mathbf{w}_*=\underset{\mathbf{w}^T\Sigma_R\mathbf{w}=1}{\operatorname{argmax}}\ \mathbf{w}^T\Sigma_C\mathbf{w}=\underset{\mathbf{w}^T\Sigma_R\mathbf{w}\le 1}{\operatorname{argmax}}\ \mathbf{w}^T\Sigma_C\mathbf{w}=\underset{\mathbf{w}^T\Sigma_R\mathbf{w}\le 1}{\operatorname{argmax}}\ \mathbf{w}^T\ell, \quad (12)$$

where ℓ is the sole eigenvector of Σ_C, since Σ_C is a unit-rank matrix. To impose an L_1 sparsity on $\mathbf{w}$, we propose to solve the penalized version of (12) given by

$$\mathbf{w}_\lambda = \underset{\mathbf{w}^T\Sigma_R\mathbf{w}\le 1}{\operatorname{argmax}}\ \mathbf{w}^T\ell - \lambda\|\mathbf{w}\|_1. \quad (13)$$

Although (13) can be solved by using some standard convex programming methods, such methods are too slow for most large-scale applications, such as imaging genetics. We therefore reformulate our problem below. Without special saying, we focus on $\ell = (\mathbf{CB})^T \neq \mathbf{0}$.

By omitting a scaling factor $\|\Sigma_R^{-1/2}\ell\|_2$, which will not affect the generalized heritability ratio, we note that (12) is equivalent to the following

$$\mathbf{w}_0 = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} - \mathbf{w}^T\ell. \quad (14)$$

We consider a penalized version of (14) as

$$\mathbf{w}_{0,\lambda} = \underset{\mathbf{w}}{\operatorname{argmin}} f(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} - \mathbf{w}^T\ell + \lambda\|\mathbf{w}\|_1. \quad (15)$$

A nice property of (15) is that it does not explicitly involve the inequality constraint, which leads to a fast computation. We define (14) as the oracle, since $\mathbf{w}_\lambda$ converges to $\mathbf{w}_0$ as $\lambda \to 0$. It can be shown that

$$\mathbf{w}_0 = \sum\nolimits_R^{-1}\ell. \quad (16)$$

We obtain an equivalence between (15) and (13) as follows.

Theorem 2.1 *Problem (15) is equivalent to problem (13) and* $\mathbf{w}_\lambda \propto \mathbf{w}_{0,\lambda}$.

We discuss some connections between our SURP problem and the optimization problem considered in Fan et al. (2012) for performing classification in high dimensional space.

However, rather than recasting the problem as in (12) and then (15), they formulate it as

$$\mathbf{w}_c = \underset{\|\mathbf{w}\|_1 \le c, \mathbf{w}^T\ell = 1}{\operatorname{argmin}}\ \mathbf{w}^T\Sigma_R\mathbf{w},$$

which can further be reformulated as

$$\mathbf{w}_\lambda = \underset{\mathbf{w}^T\ell = 1}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} + \lambda\|\mathbf{w}\|_1. \quad (17)$$

Since (17) involves a linear equality constraint, they replace it by a quadratic penalty as

$$\mathbf{w}_{\lambda,\gamma}=\operatorname{argmin}\frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w}+\lambda\|\mathbf{w}\|_1+\frac{1}{2}\gamma(\mathbf{w}^T\ell-1)^2. \quad (18)$$

This new formulation requires the simultaneously tuning of λ and γ, which can be computationally intensive. However, in Fan et al. (2012), they stated that the solution to (18) is not sensitive to γ, since solution is always in the direction of $\Sigma_R^{-1}\ell$ when λ = 0, as validated by simulations. Their formulation (17) is close to the formulation (15). This result sheds some light on why $\mathbf{w}_{\lambda,\gamma}$ is not sensitive to γ. Finally, we can show that the solution path to (15) has a piecewise linear property.

Proposition 2.2 *Let $\ell \in \mathbb{R}^q$ be a constant vector and Σ_R be positive definite. Then, $\mathbf{w}_{0,\lambda}$ is a continuous piecewise linear function in λ.*

We derive a coordinate descent algorithm to solve (15). Without loss of generality, suppose that $\mathbf{w}=(\tilde{w}_1,\tilde{\mathbf{w}}_2^T)^T=(\tilde{w}_1,\cdots,\tilde{w}_q)^T$, w_j for all $j \geq 2$ are given, and we need to optimize (15) with respect to w_1. In this case, the objective function (15) becomes

$$f_1(\tilde{w}_1,\tilde{\mathbf{w}}_2)=\frac{1}{2}(\tilde{w}_1,\tilde{\mathbf{w}}_2^T)\begin{pmatrix}\sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix}\tilde{w}_1\\ \tilde{\mathbf{w}}_2\end{pmatrix}-(\tilde{\ell}_1\tilde{w}_1+\tilde{\ell}_2^T\tilde{\mathbf{w}}_2)+\lambda|\tilde{w}_1|+\lambda\|\tilde{\mathbf{w}}_2\|_1,$$

where $\ell=(\tilde{\ell}_1,\tilde{\ell}_2^T)$ and σ11, Σ12, and Σ22 are subcomponents of Σ_R. Then, by taking the sub-gradient with respect to w_1, we have

$$f_1'(\tilde{w}_1,\tilde{\mathbf{w}}_2)=\tilde{w}_1\sigma_{11}+\Sigma_{12}\tilde{\mathbf{w}}_2+\lambda\Gamma_1-\tilde{\ell}_1$$

where $\Gamma_1 = \operatorname{sign}(w_1)$ for $w_1 \neq 0$ and is between −1 and 1 if $w_1 = 0$. Let $S_\lambda(t) = \operatorname{sign}(t)(|t|-\lambda)^+$ be the soft-thresholding operator. By setting $f_1'(\tilde{w}_1,\tilde{\mathbf{w}}_2)=0$, we have $w_1 = S_\lambda(\tilde{\ell}_1 - \Sigma_{12}\tilde{\mathbf{w}_2})/\sigma_{11}$. Based on this result, we can obtain a coordinate descent algorithm as follows.

Algorithm 1

a. *Initialize* **w** *at a starting point* $\mathbf{w}^{(0)}$ *and set* $m = 0$.

b. *Repeat:*

- *(b.1) Increase m by* 1*:* $m \leftarrow m + 1$
- *(b.2) for* $j \in 1, \ldots, p$*, if* $\tilde{w}_j^{(m-1)}=0$*, then set* $\tilde{w}_j^{(m)}=0$*; otherwise:* $\tilde{w}_j^{(m)}=\operatorname{argmin} f(\tilde{w}_1^{(m)},\cdots,\tilde{w}_{j-1}^{(m)},\tilde{w}_j,\tilde{w}_{j+1}^{(m-1)},\cdots,\tilde{w}_q^{(m-1)})$

c. *Until numerical convergence: we require* $|f(\mathbf{w}^{(m)})-f(\mathbf{w}^{(m-1)})|$ *to be sufficiently small.*

2.3 Extension to Multi-rank Cases

In this subsection, we extend the sparse unit rank projection procedure to handle multiple rank test problems when $r > 1$. We propose the k–th projection direction as the solution to the following problem:

$$\operatorname{argmax}\frac{\mathbf{w}_k^T\Sigma_C\mathbf{w}_k}{\mathbf{w}_k^T\Sigma_R\mathbf{w}_k}\text{s.t.}\mathbf{w}_k^T\Sigma_R\mathbf{w}_j=0,\forall_j<k. \quad (19)$$

It can be shown that (19) is equivalent to

$$\operatorname{argmax}\mathbf{w}_k^T\Sigma_C\mathbf{w}_k\text{s.t.}\mathbf{w}_k^T\Sigma_R\mathbf{w}_k\le 1,\mathbf{w}_k^T\Sigma_R\mathbf{w}_j=0,\forall_j<k. \quad (20)$$

Following the reasoning in Witten and Tibshirani (2011), we recast (20) into an equivalent problem.

Proposition 2.3 *Problem* (20) *is equivalent to the following problem:*

$$\operatorname*{argmax}_{\mathbf{w}}\frac{\mathbf{w}^T\mathbf{B}^T\mathbf{C}^T\Sigma_{11}^{1/2}P_\perp^{k-1}\Sigma_{11}^{1/2}\mathbf{C}\mathbf{B}\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}, \quad (21)$$

where $P_\perp^{k-1}$ *is the projection matrix onto the orthogonal space spanned by* $\{\Sigma_{11}^{1/2}\mathbf{C}\mathbf{B}\mathbf{w}_j, 1\le j\le k-1\}$, *in which* $\Sigma_{11} = (D^{-T}\Sigma_X D^{-1})_{(r,r)}$.

Based on Proposition 2.3, we consider several strategies of imposing the sparsity structure on $\mathbf{w}_k$. A simple strategy is to consider the following problem given by

$$\operatorname*{argmax}_{\mathbf{w}_k}\mathbf{w}_k^T\Sigma_C^k\mathbf{w}_k-\lambda\|\mathbf{w}_k\|_1\text{s.t.}\mathbf{w}_k^T\Sigma_R\mathbf{w}_k\le 1, \quad (22)$$

where $\Sigma_C^k=\mathbf{B}^T\mathbf{C}^T\Sigma_{11}^{1/2}P_\perp^{k-1}\Sigma_{11}^{1/2}\mathbf{C}\mathbf{B}$. When the rank of $\mathbf{C}$ is greater than 1, the problem in (22) is no longer convex, since it involves maximizing an objective function that is not concave. A potential solution is to use the minorization-maximization (MM) algorithm (Lange et al., 2000). Specifically, for any fixed $\mathbf{w}^{(m)}$, we take a Taylor series expansion of $\mathbf{w}_k^T\boldsymbol{\Sigma}_C^k\mathbf{w}_k$ at $\mathbf{w}^{(m)}$ and get

$$\mathbf{w}_k^T\Sigma_C^k\mathbf{w}_k-\lambda\|\mathbf{w}_k\|_1\ge 2\mathbf{w}_k^T\Sigma_C^k\mathbf{w}_k^{(m)}-\mathbf{w}_k^{(m)T}\Sigma_C^k\mathbf{w}_k^{(m)}-\lambda\|\mathbf{w}_k\|_1. \quad (23)$$

Thus, the right hand side of (23) minorizes the objective function (22) at $\mathbf{w}_k^{(m)}$ and is a convex function, which can be solved by using some convex optimization methods. However, based on our extensive experience, the MM algorithm is too slow for most large-scale problems, such as imaging genetics.

To further improve computational efficiency, we consider a surrogate of (22). Recall the discussion in the second principle, we are only interested in extracting informative directions

for testing hypotheses of interest. We consider a spectral decomposition of $(D^{-T}\Sigma_X D^{-1})_{(r\times r)}$ as $(D^{-T}\Sigma_X D^{-1})_{(r\times r)}=\sum_{j=1}^{r}\gamma_j\ell_j\ell_j^T$, where (γ_j, ℓ_j) are eigenvalue-eigenvector pairs with $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_r$. Then, instead of solving (22), we propose to solve r SURP problems as

$$\mathbf{w}_\lambda^k=\operatorname{argmin}\frac{1}{2}\mathbf{w}_k^T\Sigma_R\mathbf{w}_k-\sqrt{\gamma_k}\ell_k^T\mathbf{C}\mathbf{B}\mathbf{w}_k+\lambda_k\|\mathbf{w}_k\|_1 \text{for} 1\leq k\leq r. \quad (24)$$

Solving (24) leads to r sparse projection directions. In (24), since we sequentially extract the direction vector according to the input signal Σ_C, it may produce a less informative direction vector compared with those from (22). However, such formulation leads to a fast computational algorithm and our simulation results demonstrate its reasonable performance. Thus, (24) is preferred in practice.

2.4 Test Procedure

We consider three statistics for testing H_{0W} against H_{1W} in (4). Based on model (3), we calculate the ordinary least squares estimate of $\beta_{\mathbf{w}}$, given by

$$\hat{\boldsymbol{\beta}}_{\mathbf{w}}=\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{y}_i^T\mathbf{W}.$$

Subsequently, we calculate a $k \times k$ matrix, denoted by T_n, as follows:

$$T_n=(\mathbf{C}\hat{\boldsymbol{\beta}}_{\mathbf{w}}-\mathbf{b}_0)^T\Sigma_{\tilde{\Omega}}^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_{\mathbf{w}}-\mathbf{b}_0), \quad (25)$$

where $\Sigma_{\tilde{\Omega}}$ is a consistent estimate of the covariance matrix of $\mathbf{C}\hat{\boldsymbol{\beta}}_{\mathbf{w}}-\mathbf{b}_0$. Specifically, let $\tilde{\beta}_{\mathbf{w}}$ be the restricted least squares (RLS) estimate of β under H_0, which is given by

$$\tilde{\boldsymbol{\beta}}_{\mathbf{w}}=\hat{\boldsymbol{\beta}}_{\mathbf{w}}-(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_{\mathbf{w}}-\mathbf{b}_0).$$

Then, we can set $\Sigma_{\tilde{\Omega}}=\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\sum_{i=1}^{N}a_i^2\mathbf{x}_i\tilde{\boldsymbol{\varepsilon}}_i^T\tilde{\boldsymbol{\varepsilon}}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$, where $a_i=1/\{1-\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\}$ and $\tilde{\boldsymbol{\varepsilon}}_i=\mathbf{W}^T\mathbf{y}_i-\tilde{\boldsymbol{\beta}}_{\mathbf{w}}^T\mathbf{x}_i$. When $k > 1$, we use the determinant, trace and eigenvalues of T_n as test statistics, which are given by

$$W_n=\det(T_n), \mathrm{Tr}_n=\operatorname{trace}(T_n), \text{and} \quad \mathrm{Roy}_n=\max(\operatorname{eig}(T_n)), \quad (26)$$

where det, trace, and eig, respectively, denote the determinant, trace and eigenvalues of a symmetric matrix. When $k = 1$, all three statistics in (26) reduce to the Wald-type (or Hotelling's T^2) test statistic. For simplicity, we focus on Tr_n throughout the paper.

We propose a wild bootstrap method to improve the finite sample performance of the test statistic Tr_n. First, we fit model (1) under the null hypothesis (2) to calculate the estimated regression coefficient matrix, denoted by $\hat{\mathbf{B}}_0$, with corresponding residuals $\hat{\mathbf{e}}_i=\mathbf{y}_i-\hat{\mathbf{B}}_0^T\mathbf{x}_i$ for $i = 1, \ldots, n$. Then we generate G bootstrap samples $\mathbf{z}_i^{(g)}=(\hat{\mathbf{B}}_0)^T\mathbf{x}_i+\eta_i^{(g)}\hat{\mathbf{e}}_i$ for $i = 1, \ldots, n$, where $\eta_i^{(g)}$ are independently and identically distributed as a distribution F, which is chosen

to be ±1 with equal probability. For each generated wild-bootstrap sample, we repeat the estimation procedure for estimating the optimal weights and the calculation of the test statistic $\mathrm{Tr}_n^{(g)}$. Subsequently, the *p*-value of Tr_n is computed as $\frac{1}{G}\sum_{g=1}^{G} 1(\mathrm{Tr}_n^{(g)} \geq \mathrm{Tr}_n)$, where **1**(·) is an indicator function.

2.5 Tuning Parameter Selection

We consider several methods to select the tuning parameter λ. The first one is cross validation (CV), which is primarily a way of measuring the predictive performance of a statistical model. However, the CV technique can be computationally expensive for large-scale problems. The second one is the information criterion, which has been widely to measure the relative goodness of fit of a statistical model. However, neither of these two methods are applicable for SURP, since our primary interest is to find informative directions for appropriately testing the null and alternative hypotheses of (2). If the null hypothesis is true, it is expected that $\widehat{\mathbf{CB}}$ only contains noisy components and the estimated direction vectors should be random. In this case, the test statistics Tr_n, W_n, and Roy_n should not be sensitive to the value of λ. This motivates us to use the rejection rate to select the tuning parameter as follows:

$$\hat{\lambda} = \underset{0 \leq \lambda \leq \lambda_{\max}}{\operatorname{argmax}} \{(\text{Rejection Rate})_\lambda\}, \quad (27)$$

where $\lambda_{\max}$ is the largest λ to make **w** nonzero.

3 Asymptotic Theory

We investigate several theoretical properties of SURP and its associated estimator. By substituting $\tilde{\Sigma}_R$ and $\widehat{\ell} = \widehat{\mathbf{CB}}$ into (15), we can calculate an estimate of $\mathbf{w}_0$ as

$$\hat{\mathbf{w}}_\lambda = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T \tilde{\Sigma}_R \mathbf{w} - \mathbf{w}^T \hat{\ell} + \lambda \|\mathbf{w}\|_1. \quad (28)$$

The following question arises naturally:

how close is $\hat{\mathbf{w}}_\lambda$ to $\mathbf{w}_0$?

We address this question in Theorems 3.1 and 3.2.

We consider the scenario that there are a few nonzero components in $\mathbf{w}_0$, that is, a few response variables are associated with the covariates of interest. Such a scenario is common in many large-scale problems. We make a note here that the sparsity of $\mathbf{w}_0 = \Sigma_R^{-1}\ell$ does not require neither Σ_R^{-1} nor ℓ to be sparse, and hence are more quite flexible. Let $S_0 = \{j : w_{0,j} \neq 0\}$ be the active set of $\mathbf{w}_0 = (w_{0,1}, \ldots, w_{0,q})^T$ and s_0 is the number of elements in S_0. We use the banded covariance estimator of Σ_R (Bickel and Levina, 2008) such that $\|\tilde{\Sigma}_R - \Sigma_R\|_2 = O_p((\frac{\log q}{n})^{\frac{\alpha}{2(\alpha+1)}})$ for some well behaved covariance class $\mathscr{U}(\varepsilon_0, \alpha, C_1)$, which is defined as

$$\mathscr{U}(\varepsilon_0,\alpha,C_1)=\{\Sigma=(\sigma_{jj'}):\max_j\sum_{j'}\{|\sigma_{j'j}|:|j'-j|>k\}\le C_1k^{-\alpha}\text{for all}k>0\text{and}0<\varepsilon_0\le\lambda_{\min}(\Sigma)\le\lambda_{\max}(\Sigma)\le1/\varepsilon_0\}.$$

We have the following results.

Theorem 3.1 *Assume that* $\Sigma_R \in \mathscr{U}(\varepsilon_0, \alpha, C_1)$ *and*

$$\lambda=\max\{(k_nt_1^0+C_1k_n^{-\alpha})\|\mathbf{w}_0\|_2,t_2^0\}\asymp(\frac{\log(q\vee n)}{n})^{\frac{\alpha}{2(\alpha+1)}}\|\mathbf{w}_0\|_2, \quad (29)$$

where $k_n\asymp(\frac{\log(q\vee n)}{n})^{-\frac{1}{2(\alpha+1)}}, t_1^0:=\sqrt{2(\eta_1+1)\frac{1}{\gamma(\varepsilon_0,\delta)}}\sqrt{\frac{\log(q\vee n)}{n}}$, *and* $t_2^0:=\frac{C_0}{\varepsilon_0}\sqrt{2(\eta_2+1)}\sqrt{\frac{\log(q\vee n)}{n}}$, *in which* $\gamma(\varepsilon_0, \delta)$ *and* $\delta = \delta(\varepsilon_0)$ *only depends on* $(\varepsilon_0$. *Then, with probability at least* $1-(q\vee n)^{-\eta_1}-(q\vee n)^{-\eta_2}$, we have

$$\|\hat{\mathbf{w}}_\lambda-\mathbf{w}_0\|_2\le C\lambda\sqrt{s_0}, \quad (30)$$

where C is a constant not depending on q and n. Furthermore, for $\|\ell\|_2 > \delta_0$, *we have*

$$\|\frac{\hat{\mathbf{w}}_\lambda}{\|\hat{\mathbf{w}}_\lambda\|_2}-\frac{\mathbf{w}_0}{\|\mathbf{w}_0\|_2}\|_2\le\frac{2C\lambda\sqrt{s_0}}{\|\mathbf{w}_0\|_2}. \quad (31)$$

Theorem 3.1 gives an oracle inequality and the L_2 convergence rate of $\hat{\mathbf{w}}_\lambda$ in the sparse case, which indicates direction consistency and is important to ensure the good performance of test statistics. This result has several important implications. If $\sqrt{s_0}(\frac{\log q}{n})^{\frac{\alpha}{2(\alpha+1)}}=o(1)$, then $\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_2$ converges to zero in probability. Therefore, our SURP should perform well for the extremely sparse cases with $s_0 << n$. This is extremely important in practice, since the extremely sparse cases are common for many large-scale problems. Although we consider the banded covariance estimator of Σ_R in Theorem 3.1 (Bickel and Levina, 2008), the convergence rate of $\hat{\mathbf{w}}_\lambda$ can be established for other estimators of Σ_R and ℓ as follows.

Theorem 3.2 *Suppose that we have* $\|\tilde{\Sigma_R} - \Sigma_R\|_2 = O_p(a_n) = o_p(1)$ *and* $\|\hat{\ell} - \ell\|_\infty = O_p(b_n) = o_p(1)$, *then*

$$\|\hat{\mathbf{w}}_\lambda-\mathbf{w}_0\|_2=O_p((a_n\vee b_n)\sqrt{s_0}). \quad (32)$$

Furthermore, for $\|\ell\|_2 > \delta_0$, *we have*

$$\|\frac{\hat{\mathbf{w}}_\lambda}{\|\hat{\mathbf{w}}_\lambda\|_2}-\frac{\mathbf{w}_0}{\|\mathbf{w}_0\|_2}\|_2=O_p(\frac{(a_n\vee b_n)\sqrt{s_0}}{\|\mathbf{w}_0\|_2}). \quad (33)$$

Theorem 3.2 gives the L_2 convergence rate of $\hat{\mathbf{w}}_\lambda$ for any possible estimators of Σ_R and ℓ. A direct implication is that we can consider other estimators of Σ_R in order to achieve better estimation of Σ_R under different assumptions of Σ_R. For instance, if Σ_R has an approximate factor structure with sparsity, then we may consider the principal orthogonal complement thresholding (POET) method in Fan et al. (2013) to estimate Σ_R. Moreover, if we can achieve good estimation of ℓ for large p, then we can extend model (1) to the scenario with large p. We will systematically investigate these generalizations in our future work.

Remark The SPReM estimator $\hat{\mathbf{w}}_\lambda$ is closely connected with those estimators in Witten and Tibshirani (2011) and Fan et al. (2012) in the framework of penalized linear discriminant analysis. However, little is known about the theoretical properties of such estimators. To the best of our knowledge, Theorems 3.1 and 3.2 are the first results on the convergence rate of such estimators under the restricted eigen-vectors of problem (11).

Remark The SPReM estimator $\hat{\mathbf{w}}_\lambda$ does not have the oracle property due to the asymptotic bias introduced by the L_1 penalty. See detailed discussions in (Fan and Li, 2001; Zou, 2006). However, our estimation procedure may be modified to achieve the oracle property by using some non-concave penalties or adaptive weights. We will investigate this issue in more depth in our future work.

4 Numerical Examples

4.1 Simulation 1: Two Sample Test in High Dimensions

In this subsection, we consider high-dimensional two-sample test problems and compare SPReM with the High-dimensional Two-Sample test (HTS) method in Chen and Qin (2010) and the Random Projection (RP) method proposed by Lopes et al. (2011). Both HTS and RP are the state-of-the-art methods for detecting a shift between the means of two high-dimensional normal distributions. It has been shown in Lopes et al. (2011) that the random projection method outperforms several competing methods when q/n converges to a constant or ∞.

We simulated two sets of samples $\{\mathbf{y}_1, \ldots, \mathbf{y}_{n_1}\}$ and $\{\mathbf{y}_{n_1+1}, \ldots, \mathbf{y}_n\}$ from $N(\beta_1, \Sigma_R)$ and $N(\beta_2, \Sigma_R)$, respectively, where β_1 and β_2 are $q \times 1$ mean vectors and $\Sigma_R = \sigma^2(\rho_{jj'})$, in which $(\rho_{jj'})$ is a $q \times q$ correlation matrix. We set $n = 2n_1 = 100$ and the dimension of the multivariate response q is 50, 100, 200, 400, and 800, respectively. We are interested in testing the null hypothesis $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$. This two-sample test problem can be formulated as a special case of model (1) with $n = n_1 + n_2$. Moreover, we have $\mathbf{B}^T = [\beta_1, \beta_2]$ and $\mathbf{C} = (1, -1)$. Without loss of generality, we set $\beta_1 = \beta_2 = \mathbf{0}$ to assess type I error rate and then introduce a shift in the first ten components of β_2 to be 1 to assess power. We set σ^2 to be 1 and 3 and consider three different correlation matrices as follows.

- Case 1 is an independent covariance matrix with $(\rho_{jj'}) = \text{diag}(1, \ldots, 1)$.
- Case 2 is a weak correlation matrix with $\rho_{jj'} = \mathbf{1}(j' = j) + 0.3 + \mathbf{1}(j' \neq j)$.
- Case 3 is a strong correlation covariance matrix with $\rho_{jj'} = 0.8^{|j'-j|}$.

Simulation results are summarized in Tables 1 and 2. As expected, both HTS and RP perform worse as q gets larger, whereas our SPReM works very well even for relatively large q. This is consistent with our theoretical results in Theorems 3.1 and 3.2. Moreover, HTS and RP cannot control the type I error rate well in all scenarios, whereas our SPReM based on the wild bootstrap method works reasonably well. According to the best of our knowledge, none of the existing methods for the two sample test in high dimensions work well in this sparse setting. For cases (ii) and (iii), $\Sigma_R^{-1}(\beta_1 - \beta_2)$ is not sparse, but SPReM performs reasonably well under the correlated scenarios. This may indicate the potential of extending SPReM and its associated theory to non-sparse cases. As expected, increasing σ^2 decreases statistical power in rejecting the null hypothesis. Since both SPReM and RP significantly outperform HTS, we increased q to 2,000 and presented some additional comparisons between SPReM and RP based on 100 simulated data sets in Figure 1.

4.2 Simulation 2: Multiple Rank Cases

In this subsection, we evaluate the finite sample performance of SMURP. The simulation studies were designed to establish the association between a relatively high-dimensional imaging phenotype with a genetic marker (e.g., SNP or haplotype), which is common in imaging genetics studies, while adjusting for age and other environmental factors. We set the sample size $n = 100$ and the dimension of the multivariate phenotype q to be 50, 100, 200, 400 and 800, respectively, and then simulated the multivariate phenotype according to model (1). The random errors were simulated from a multivariate normal distribution with mean 0 and covariance matrix with diagonal elements 1. For the off-diagonal elements in the covariance matrix, which characterize the correlations among the multivariate phenotypes, we categorized each component of the multivariate phenotype into three categories: high correlation, medium correlation and very low correlation with the corresponding number of components (1, 1, $q - 2$) in each category, and then we set the three degrees of correlation among the different components of the multivariate phenotype according to Table 3. The final covariance matrix is set to be $\Sigma_R = \sigma^2(\rho_{jj'})$, where $(\rho_{jj'})$ is the correlation matrix. We considered $\sigma^2 = 1$ and 3.

For the covariates, we included two SNPs with an additive effect and 3 additional continuous covariates. We varied the minor allele frequency (MAF) of the first SNP, whereas we fixed the MAF of the second SNP to be 0.5. For the first SNP, we considered 6 scenarios assuming the MAFs are 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5, respectively. We simulated the three additional continuous covariates from a multivariate normal distribution with mean 0, standard deviation 1, and equal correlation 0.3. We first set $\mathbf{B} = 0$ to assess type I error rate. To assess power, we set the first response to be the only components of the multivariate phenotype associated with the first SNP and the second response to be the component related to the second SNP effect. Specifically, we set the coefficients of the two SNPs to be 1 for the selected responses and all other regression coefficients to be 0. We are interested in testing the joint effects of the two SNPs on phenotypic variance.

We applied SPReM to 100 simulated data sets. Note that to the best of our knowledge, no other methods can be used to test the multi-rank test problem and thus we only focus on SPReM here. Table 4 presents the estimated rejection rates corresponding to different

MAFs, q, and σ^2. Our SPReM works very well even for relatively large q under both $\sigma^2 = 1$ and 3. Specifically, the wild bootstrap method can control the type I error rate well in all scenarios. For the power, SPReM performs reasonably well under the small MAFs and $q = 800$. It may indicate that our method can perform well for much larger q if the sample size gets larger. As expected, increasing σ^2 decreases statistical power in rejecting the null hypothesis.

4.3 Alzheimer's Disease Neuroimaging Initiative (ADNI) Data Analysis

The development of SPReM is motivated by the joint analysis of imaging, genetic, and clinical variables in the ADNI study. "Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year publicprivate partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment 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 to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org."

The Huamn 610-Quad BeadChip (Illumina, Inc. San Diego, CA) was used to genotype 818 subjects in the ADNI-1 database, which resulted in a set of 620,901 SNPs and copy number variation (CNV) markers. Since the Apolipoprotein E (ApoE) SNPs, rs429358 and rs7412, are not on the Human 610-Quad Bead-Chip, they were genotyped separately and added to the data set manually. For simplicity, we only considered the 10, 479 SNPs collected on the chromosome 19, which houses the famous ApoE gene commonly suspected of having association with Alzheimer's disease. A complete GWAS of ADNI will be reported elsewhere. The SNP data were preprocessed by standard quality control steps including dropping any SNP that has more than 5% missing data, imputing the missing values in each SNP with its mode, dropping SNPs with minor allele frequency < 0.05, and screening out SNPs violating the Hardy-Weinberg equilibrium. Finally, we obtained 8, 983 SNPs on chromosome 19, including the ApoE allele as the last SNP in our dataset.

Our problem of interest is to perform a genome-wide search for establishing the association between the 10, 479 SNPs collected on the chromosome 19 and the brain volume of 93 regions of interest (ROIs). We fitted model (1) with all 93 ROIs as responses and a covariate vector including an intercept, a specific SNP, age, gender, whole brain volume, and the top 5 principal components to account for population stratification. To reduce population stratification effects, we only used 761 Caucasians from all 818 subjects. Subjects with missing values were removed, which leads to 747 subjects. We set $\lambda = \lambda_{\max}$ in our SPReM for computational efficiency. To test the SNP effect on all 93 ROIs, we calculated the test statistic and its *p*–value for each SNP. We further performed a standard massive univariate analysis. Specifically, we fitted a linear model with the same set of covariates and calculated a *p*–value for every pair of ROIs and SNPs.

We developed a computationally efficient strategy to approximate the *p*–value of each SNP with different MAFs. In the real data analysis, we considered a pool of SNPs consisting of 6 MAF groups including MAF∈ (0.05, 0.075], MAF∈ (0.075, 0.15], MAF∈ (0.15, 0.25], MAF∈ (0.25, 0.35], MAF∈ (0.35, 0.45], and MAF∈ (0.45, 0.50]. Each MAF group contains 40 SNPs. For each SNP, we generated 10,000 wild bootstrap samples under the null hypothesis to obtain 10,000 bootstrapped test statistics. Then, based on 40 × 10, 000 bootstrapped samples for each MAF group, we use the Satterthwaite method to approximate the null distribution of the test statistic by a Gamma distribution with parameters (a_T, b_T). Specifically, we set $a_T = \varepsilon^2/\nu$ and $b_T = \nu/\varepsilon$ by matching the mean (ε) and the variance (ν) of the test statistics and those of the Gamma distribution. The histograms and the fitted gamma distributions along with the QQ-plots are, respectively, presented in Figures 2–3. Figures 2 and 3 reveal that our gamma approximations work reasonably well for a wide range of MAFs when $\lambda = \lambda_{\max}$. Since we only use Gamma(a_T, b_T) to approximate the *p*–value of large test statistic, we only need a good approximation at the tail of the Gamma distribution. See Figure 3 for details. For each SNP, we matched its MAF with the closest MAF group in the pool and then calculated the *p*–value of the test statistic based on the approximated gamma distribution. We present the manhattan plot in Figure 4 and the top 10 SNPs with their *p*–values for SPReM and the mass univariate analysis in Table 5 for $\lambda = \lambda_{\max}$.

We have several important findings. The ApoE allele was identified as the top one significant covariate with $-\log_{10}(p) \sim 15$ and 9 respectively, indicating a strong association between the ApoE allele and imaging phenotype, a biomarker of Alzheimer's disease diagnosis. This finding agrees with the previous result in Vounou et al. (2012). We also found some interesting results regarding rs207650 on the TOMM40 gene, which is one of the top 10 significant SNPs with $-\log_{10}(p) \sim 5$ and 4 respectively. The TOMM40 gene is located in close proximity to the ApoE gene and has also been linked to AD in some recent studies (Vounou et al., 2012). We are also able to detect some additional SNPs, such as rs11667587 on the NOVA2 gene, among others, on the chromosome 19, which are not identified in existing genome-wide association studies. The new findings may shed more light on further Alzheimer's research. The *p*–values for those top 10 SNPs calculated from SPReM are much smaller than those calculated from the mass univariate analysis. In other words, to achieve comparable *p*–values, the mass univariate analysis requires many more samples. This strongly demonstrates the effectiveness of our proposed method.

5 Discussion

In this paper, we have developed a general SPReM framework based on the two heritability ratios. Our SPReM methodology has a wide range of applications, including sparse linear discriminant analysis, two sample tests, and general hypothesis tests in MLMs, among many others. We have systematically investigated the L_2 convergence rate of $\hat{\mathbf{w}}_\lambda$ in the ultrahigh dimensional framework. We further extend the SURP problem to the SMURP and offered a sequential SURP approximation algorithm. We carried out simulation studies and examined a real data set to demonstrate the excellent performance of our SPReM framework compared to other state-of-the-art methods.

6 Assumptions and Proofs

Throughout the paper, the following assumptions are needed to facilitate the technical details, although they may not be the weakest conditions.

Assumption A1. $\mathbf{C}(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T \asymp 1$, that is, there exists constant c_0 and C_0 such that $c_0 \le \mathbf{C}(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T \le C_0$.

Assumption A2. $0 \le \varepsilon_0 \le \lambda_{\min}(\Sigma_R) \le \lambda_{\max}(\Sigma_R) \le 1/\varepsilon_0$.

Assumption A3. The covariance estimator $\tilde{\Sigma}_R$ satisfies: $\|\tilde{\Sigma}_R - \Sigma_R\|_2 = O_p(a_n) \le o_p(1)$.

Remark : Assumption A1 is a very weak and standard assumption for regression models. Assumption A2 has been widely used in the literature. Assumption A3 requires a relatively accurate covariance estimator in terms of spectral norm convergence. We may use some good penalized estimators of Σ_R under different assumptions of Σ_R (Bickel and Levina, 2008; Cai et al., 2010; Lam and Fan, 2009; Rothman et al., 2009; Fan et al., 2013).

Proof of Theorem 2.1 The Karush-Kuhn-Tucker (KKT) conditions for problem (13) are given by:

$$\ell - \lambda\Gamma - \gamma\Sigma_R\mathbf{w}=0, \gamma \ge 0, \gamma(\frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} - \frac{1}{2})=0, \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} \le \frac{1}{2},$$

where Γ is a $q \times 1$ vector and equals the subgradient of $\|\mathbf{w}\|_1$ with respect to $\mathbf{w}$. We consider two scenarios. First, suppose that $|\ell_j| > \lambda$ for some j. We must have $\gamma\Sigma_R\mathbf{w} \ne 0$, which leads to $\gamma > 0$ and $\mathbf{w}^T\Sigma_R\mathbf{w} = 1$. Thus, the KKT conditions reduce to

$$\ell - \lambda\Gamma - \gamma\Sigma_R\mathbf{w}=0, \gamma \ge 0, \mathbf{w}^T\Sigma_R\mathbf{w}=1.$$

If we write $\tilde{\mathbf{w}} = \gamma\mathbf{w}$, this is equivalent to solving problem (15) with $\tilde{\mathbf{w}}$ and then take normalization. Second, if $|\ell_j| \le \lambda$ for any j, then $\mathbf{w} = 0$ and $\gamma = 0$, which is the solution of (15) as well. This finishes the proof.

Proof of Proposition 2.2 It follows from Theorem 2 of Rosset and Zhu (2007).

Proof of Proposition 2.3 The proof is similar to that of Proposition 1 of Witten and Tibshirani (2011). Letting $\tilde{\mathbf{w}}_k = \Sigma_R^{1/2}\mathbf{w}_k$, then problem (20) can be rewritten as

$$\operatorname{argmax}\tilde{\mathbf{w}}_k^T \Sigma_R^{-1/2}\mathbf{B}^T\mathbf{C}^T\Sigma_{11}^{1/2}\mathbf{C}\mathbf{B}\Sigma_{\mathbf{R}}^{-\mathbf{1/2}}\hat{\mathbf{w}}_{\mathbf{k}}\text{s.t.}\|\hat{\mathbf{w}}_{\mathbf{k}}\|^{\mathbf{2}} \leq \mathbf{1},$$

which is equivalent to

$$\operatorname{argmax}\tilde{\mathbf{w}}_k \mathbf{A} P_{\perp}^{k-1}\mathbf{u}_k \text{s.t}\|\tilde{\mathbf{w}}_k\|^2 \leq 1, \|\mathbf{u}_k\|^2 \leq 1, \quad (34)$$

where $\mathbf{A}=\mathbf{B}^T\mathbf{C}^T\Sigma_{11}^{1/2}$. Thus, $\tilde{\mathbf{w}_k}$ and $\mathbf{u}_k$ that solve problem (34) are the k-th left and right singular vectors of $\mathbf{A}$ (Witten and Tibshirani, 2011). Therefore, we have $P_{\perp}^{k-1}=\mathbf{I}-\sum_{j=1}^{k-1}\mathbf{u}_j\mathbf{u}_j^T$ and $\mathbf{u}_k$ is the k-th eigenvector of $\mathbf{A}^T\mathbf{A}$, or equivalently the k-th right singular vector of $\mathbf{A}$. For problem (34), $\tilde{\mathbf{w}_k}$ is the k-th left singular vector of $\mathbf{A}$. Therefore, the solution of (21) is the k-th discriminant vector of (20).

Proof of Theorem 3.1 In this theorem, we specifically use the banded covariance estimator $\tilde{\Sigma_R} = B_{k_n}(\tilde{\Sigma_R})$, where $B_k(\Sigma) = [\sigma_{jj'}I(|j'-j| \leq k)]$ and $\hat{\Sigma_R}$ is the sample covariance matrix of $\mathbf{y}_i - \hat{\mathbf{B}^T}\mathbf{x}_i$.

First, we define $\mathscr{J} = \{\|\tilde{\Sigma_R} - B_{k_n}(\Sigma_R)\|_\infty \leq t_1\} \cap \{\|\hat{\boldsymbol{\ell}} - \boldsymbol{\ell}\|_\infty \leq t_2\}$, where t_1 and t_2 are specified as in Lemma 6.2. Then, it follows from Lemma 6.2 that $P(\mathscr{J}) \geq 1-3(q \vee n)^{-\eta_1} - 2(q \vee n)^{-\eta_2}$.

On the set $\mathscr{J}$, by taking $\lambda=\max\{k_n t_1+C_1 k_n^{-\alpha}, t_2\}$ and using Lemma 6.1, we have

$$\begin{aligned}
&\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda \\
&\quad - \mathbf{w}_0)+\lambda\|\hat{\mathbf{w}}_\lambda\|_1 \leq (\mathbf{w}_0^T(\Sigma_R \\
&\quad - \tilde{\Sigma}_R)+\varepsilon^T)(\hat{\mathbf{w}}_\lambda \\
&\quad - \mathbf{w}_0)+\lambda\|\mathbf{w}_0\|_1 \leq \|\tilde{\Sigma}_R - \Sigma_R\|_2\|\mathbf{w}_0\|_2\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_1 \\
&+\|\varepsilon\|_\infty\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_1 \\
&+\lambda\|\mathbf{w}_0\|_1 \leq (k_n t_1 \\
&\quad +C_1 k_n^{-\alpha})\|\mathbf{w}_0\|_2\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_1 \\
&+t_2\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_1 \\
&+\lambda\|\mathbf{w}_0\|_1 \leq \lambda\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_1 \\
&+\lambda\|\mathbf{w}_0\|_1.
\end{aligned}$$

Let $\mathbf{w}_0$,$S_0 = [w_{0,j}I(j \in S_0)]$, where $w_{0,j}$ is the j–th component of $\mathbf{w}_0$. The above equation can be rewritten as

$$(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T(\tilde{\Sigma}_R - \Sigma_R + \Sigma_R)(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda\|\hat{\mathbf{w}}_{\lambda,S_0}\|_1 + \lambda\|\hat{\mathbf{w}}_{\lambda,S_0^c}\|_1 \leq \lambda\|\hat{\mathbf{w}}_{\lambda,S_0} - \mathbf{w}_{0,S_0}\|_1 + \lambda\|\mathbf{w}_{0,S_0}\|_1 + \lambda\|\hat{\mathbf{w}}_{\lambda,S_0^c}\|_1,$$

which yields

$$\{\lambda_{\min} - O(1)(\frac{\log(q)}{n})^{\frac{\alpha}{2(\alpha+1)}}\}\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_2^2 \leq 2\lambda\sqrt{s_0}\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_2.$$

Finally, we obtain the following inequality

$$\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_2 \leq \frac{2\lambda\sqrt{s_0}}{\lambda_{min} - O(1)(\frac{\log(q)}{n})^{\frac{\alpha}{2(\alpha+1)}}} \leq C\lambda\sqrt{s_0},$$

which finishes the proof.

Proof of Theorem 3.2 It follows from Lemma (6.1) that

$$\begin{aligned}
&\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda \\
&\quad - \mathbf{w}_0) + \lambda\|\hat{\mathbf{w}}_\lambda\|_1 \leq (\mathbf{w}_0^T(\Sigma_R \\
&\quad - \tilde{\Sigma}_R) + (\hat{\ell} - \hat{\ell}))(\hat{\mathbf{w}}_\lambda \\
&\quad - \mathbf{w}_0) + \lambda\|\mathbf{w}_0\|_1 \leq (\|\tilde{\Sigma}_R - \Sigma_R\|_2\|\mathbf{w}_0\|_2 \\
&\quad + \|\hat{\ell} - \hat{\ell}\|_\infty)\|\hat{\mathbf{w}}_{\lambda,S} - \mathbf{w}_{0,S}\|_1 \\
&\quad + \lambda\|\mathbf{w}_{0,S}\|_1
\end{aligned}$$

Note that $\|\hat{\mathbf{w}}_\lambda\|_1 \leq \|\mathbf{w}_{0,S_0}\|_1 - \|\mathbf{w}_{0,S_0} - \hat{\mathbf{w}}_{\lambda,S_0}\|_1 + \|\hat{\mathbf{w}}_{\lambda,S_0^c}\|_1$. Then, by taking

$$\lambda = \|\tilde{\Sigma}_R - \Sigma_R\|_2\|\mathbf{w}_0\|_2 + \|\hat{\ell} - \hat{\ell}\|_\infty \asymp O_p(a_n\|\mathbf{w}_0\|_2 \vee b_n),$$

we have

$$\begin{aligned}\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda \\ - \mathbf{w}_0) \le (\|\tilde{\Sigma}_R - \Sigma_R\|_2\|\mathbf{w}_0\|_2 \\ +\|\hat{\ell} - \hat{\ell}\|_\infty)(\|\hat{\mathbf{w}}_{\lambda,S_0} - \mathbf{w}_{0,S_0}\|_1 + \|\hat{\mathbf{w}}_{\lambda,S_0^c}\|_1) - \lambda(\|\mathbf{w}_{0,S_0}\|_1 \\ - \|\mathbf{w}_{0,S} - \hat{\mathbf{w}}_{\lambda,S_0}\|_1 \\ +\|\hat{\mathbf{w}}_{\lambda,S_0^c}|) + \lambda|\mathbf{w}_{0,S_0}\|_1 \\ = O_p(a_n \vee b_n)\|\hat{\mathbf{w}}_{\lambda,S_0} - \mathbf{w}_{0,S_0}\|_1 \le O_p(a_n \vee b_n)\sqrt{s_0}\|\hat{\mathbf{w}}_{\lambda,S} - \mathbf{w}_{0,S_0}\|_2.\end{aligned}$$

By using Weyl's inequality, we have

$$(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) \ge (\lambda_{\min}(\Sigma_R) - \|\tilde{\Sigma}_R - \Sigma_R\|_2)\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_2^2$$

where $\|\tilde{\Sigma_R} - \Sigma_R\|_2 = O_p(a_n) = o_p(1)$. Finally, we have

$$\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_2^2 \le \frac{O_p(a_n \vee b_n)\sqrt{s_0}\|\hat{\mathbf{w}}_{\lambda,S} - \mathbf{w}_{0,S}\|_2}{\lambda_{\min}(\Sigma) - O_p(a_n)}, \quad (35)$$

which finishes the proof.

Lemma 6.1 *We have the following basic inequality*

$$\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda\|\hat{\mathbf{w}}_\lambda\|_1 \le \{\mathbf{w}_0^T(\Sigma_R - \tilde{\Sigma}_R) + (\hat{\ell} - \ell)^T\}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda\|\mathbf{w}_0\|_1. \quad (36)$$

Proof We rewrite the optimization problem (28) as

$$\hat{\mathbf{w}}_\lambda = \operatorname{argmin}\frac{1}{2}(\mathbf{w} - \tilde{\Sigma}_R^{-1}\hat{\ell})^T\tilde{\Sigma}_R(\mathbf{w} - \tilde{\Sigma}_R^{-1}\hat{\ell}) + \lambda\|\mathbf{w}\|_1.$$

Thus, we have

$$\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \tilde{\Sigma}_R^{-1}\hat{\ell})^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda - \tilde{\Sigma}_R^{-1}\hat{\ell}) + \lambda\|\hat{\mathbf{w}}_\lambda\|_1 \le \frac{1}{2}(\mathbf{w}_0 - \tilde{\Sigma}_R^{-1}\hat{\ell})^T\tilde{\Sigma}_R(\mathbf{w}_0 - \tilde{\Sigma}_R^{-1}\hat{\ell}) + \lambda\|\mathbf{w}_0\|_1,$$

which yields

$$\frac{1}{2}\|\hat{\mathbf{w}}_\lambda - \mathbf{w}_0\|_{\tilde{\Sigma}_R}^2 + \lambda\|\hat{\mathbf{w}}_\lambda\|_1 \le (\hat{\ell} - \tilde{\Sigma}_R\mathbf{w}_0)^T(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) = \{\mathbf{w}_0^T(\Sigma_R - \tilde{\Sigma}_R) + (\hat{\ell} - \ell)^T\}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda\|\mathbf{w}_0\|_1,$$

in which we have used $\hat{\boldsymbol{\ell}} = \Sigma_R\mathbf{w}_0 + \hat{\ell} - \ell$ in the last equality.

Lemma 6.2 *For all* $t_1 \ge t_1^0$ *and* $t_2 \ge t_2^0$, *we have*

$$P(\mathscr{J}) \geq 1 - 3(q \vee n)^{-\eta_1} - 2(q \vee n)^{-\eta_2}. \quad (37)$$

Proof First, it follows from Lemma A.3 of Bickel and Levina (2008) that

$$P(\|\tilde{\Sigma}_R - B_{k_n}(\Sigma_R)\|_\infty \geq t_1) \leq 2(k+1)q\exp\{-n(t_1^0)^2\gamma(\varepsilon_0,\delta)\} \leq 2(k_n+1)(q \vee n)\exp\{-2n(\eta_1+1)\frac{1}{\gamma(\varepsilon_0,\delta)}\frac{\log(q \vee n)}{n}\gamma(\varepsilon_0,\delta))\} \leq 3((q \vee n)k_n)\exp\{-(\eta_1+1)\log((q \vee n)k_n))\} \leq 3((q \vee n)k_n)^{-(\eta_1+1)+1} \leq 3(q \vee n)^{-\eta_1},$$

where $t_1^0 = \sqrt{2(\eta_1+1)\frac{1}{\gamma(\varepsilon_0,\delta)}}\sqrt{\frac{\log(q \vee n)}{n}}$.

Second, we know that $\frac{\sqrt{n}(\hat{\ell}_j - \ell_j)}{\sigma_j C_X}$ is *Sub*(1)-distributed, where $C_X = \mathbf{C}(\frac{1}{n}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$. Then by the union sum inequality, we have

$$P(\max_j|\frac{\sqrt{n}(\hat{\ell}_j - \ell_j)}{C_0/\varepsilon_0}| \geq t_2) \leq P(\max_j|\frac{\sqrt{n}(\hat{\ell}_j - \ell_j)}{\sigma_j C_X}| \geq t_2) \leq 2(q \vee n)\exp\{-\frac{(t_2^0)^2}{2}\}. \quad (38)$$

By taking $(t_2^0)^2 = 2\eta_2\log(q \vee n)$, we can rewrite the above inequality as

$$P(\|\hat{\ell} - \ell\|_\infty \geq \frac{C_0}{\varepsilon_0}\sqrt{\frac{(2\eta_2+2)\log(q \vee n)}{n}}) \leq 2(q \vee n)^{-\eta_2}$$

Finally, we get

$$P(\mathscr{J}) \geq 1 - P(\|\tilde{\Sigma}_R - B_k(\Sigma_R)\|_\infty \geq t_1^0) - P(\|\hat{\ell} - \ell\|_\infty \geq \frac{C_0}{\varepsilon_0}\sqrt{\frac{(2\eta_2+2)\log(q \vee n)}{n}}) \geq 1 - 3(q \vee n)^{-\eta_1} - 2(q \vee n)^{-\eta_2},$$

which finishes the proof.

Acknowledgments

The research of Drs. Zhu and Ibrahim was supported by NIH grants RR025747-01, GM70335, CA74015, P01CA142538-01, MH086633, EB005149-01 and AG033387. The research of Dr. Liu was supported by NSF Grant DMS-07-47575 and NIH Grant NIH/NCI R01 CA- 149569. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech,

Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

References

Amos CI, Elston RC, Bonney GE, Keats BJB, Berenson GS. A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. Am. J. Hum. Genet. 1990; 47:247–254. [PubMed: 2378349]

Amos CI, Laing AE. A comparison of univariate and multivariate tests for genetic linkage. Genetic Epidemiology. 1993; 84:303–310.

Bickel PJ, Levina E. Regularized estimation of large covariance matrices. The Annals of Statistics. 2008; 36:199–227.

Breiman L, Friedman J. Predicting multivariate responses in multiple linear regression. Journal of the Royal Statistical Society, Series B, Statistical Methodology. 1997; 59:3–54.

Cai T, Zhang C, Zhou H. Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics. 2010; 38:2118–2144.

Chen S, Qin Y. A two-sample test for high-dimensional data with applications to gene-set testing. The Annals of Statistics. 2010; 38:808–835.

Chiang MC, Barysheva M, Toga AW, Medland SE, Hansell NK, James MR, McMahon KL, de Zubicaray GI, Martin NG, Wright MJ, Thompson PM. BDNF gene effects on brain circuitry replicated in 455 twins. NeuroImage. 2011a; 55:448–454. [PubMed: 21195196]

Chiang MC, McMahon KL, de Zubicaray GI, Martin NG, Hickie I, Toga AW, Wright MJ, Thompson PM. Genetics of white matter development: A DTI study of 705 twins and their siblings aged 12 to 29. NeuroImage. 2011b; 54:2308–2317. [PubMed: 20950689]

Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society, Series B, Statistical Methodology. 2010; 72:3–25. [PubMed: 20107611]

Cook RD, Helland IS, Su Z. Envelopes and partial least squares regression. Journal of the Royal Statistical Society, Series B, Statistical Methodology. 2013 To appear.

Cook RD, Li B, Chiaromonte F. Envelope models for parsimonious and efficient multivariate linear regression. Statist. Sinica. 2010; 20:927–1010.

Fan J, Feng Y, Tong X. A road to classification in high dimensional space: the regularized optimal affine discriminant. Journal of the Royal Statistical Society, Series B, Statistical Methodology. 2012; 74:745–771. [PubMed: 23074363]

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Fan J, Liao Y, Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. Journal of Royal Statistical Society, Series B. 2013; 75:603–680.

Formisano E, Martino FD, Valente G. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magnetic Resonance Imaging. 2008; 26:921–934. [PubMed: 18508219]

Kherif F, Poline JB, Flandin G, Benali H, Simon O, Dehaene S, Worsley KJ. Multivariate model specification for fMRI data. Neuroimage. 2002; 16:1068–1083. [PubMed: 12202094]

Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principle components of heritability combine to increase power for association. Genetic Epidemiology. 2008; 32:9–19. [PubMed: 17922480]

Knickmeyer RC, Gouttard S, Kang C, Evans D, Wilber K, Smith JK, Hamer RM, Lin W, Gerig G, Gilmore JH. A structural MRI study of human brain development from birth to 2 years. J Neurosci. 2008; 28:12176–12182. [PubMed: 19020011]

Krishnan A, Williams LJ, McIntosh AR, Abdi H. Partial least squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage. 2011; 56:455–475. [PubMed: 20656037]

Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. The Annals of statistics. 2009; 37:4254–4278. [PubMed: 21132082]

Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions. Journal of Computational and Graphical Statistics. 2000; 9:1–20.

Lenroot RK, Giedd JN. Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. Neurosci Biobehav Rev. 2006; 30:718–729. [PubMed: 16887188]

Lin J, Zhu H, Knickmeyer R, Styner M, Gilmore J, Ibrahim J. Projection regression models for multivariate imaging phenotype. Genetic Epidemiology. 2012; 36:631–641. [PubMed: 22807230]

Lopes ME, Jacob LJ, Wainwright MJ. A more powerful two-sample test in high dimensions using random projection. 2011 arXiv preprint, arXiv:1108.2401.

Ott J, Rabinowitz D. A principle-components approach based on heritability for combining phenotype information. Hum Heredity. 1999; 49:106–111. [PubMed: 10077732]

Paus T. Population neuroscience: Why and how. Human Brain Mapping. 2010; 31:891–903. [PubMed: 20496380]

Peng J, Zhu J, Bergamaschi A, Han W, Noh D, Pollack JR, Wang P. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Annals of Applied Statistics. 2010; 4:53–77. [PubMed: 24489618]

Peper JS, Brouwer RM, Boomsma DI, Kahn RS, Pol HEH. Genetic inuences on human brain structure: A review of brain imaging studies in twins. Human Brain Mapping. 2007; 28:464–473. [PubMed: 17415783]

Rosset S, Zhu J. Piecewise linear regularized solution paths. The Annals of Statistics. 2007; 35:1012–1030.

Rothman AJ, Levina E, Zhu J. Generalized thresholding of large covariance matrices. Journal of the American Statistical Association. 2009; 104:177–186.

Rowe DB, Hoffmann RG. Multivariate statistical analysis in fMRI. IEEE Eng Med Biol Med. 2006; 25:60–64.

Scharinger C, Rabl U, Sitte HH, Pezawas L. Imaging genetics of mood disorders. NeuroImage. 2010; 53:810–821. [PubMed: 20156570]

Teipel SJ, Born C, Ewers M, Bokde ALW, Reiser MF, Moller HJ, Hampel H. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. NeuroImage. 2007; 38:13–24. [PubMed: 17827035]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, Statistical Methodology. 1996; 58:267–288.

Vounou M, Janousova E, Wolz R, Stein J, Thompson P, Rueckert D, Montana G. ADNI. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. Neuroimage. 2012; 60:700–716. [PubMed: 22209813]

Witten DM, Tibshirani R. Penalized classification using Fisher's linear discriminant. Journal of the Royal Statistical Society, Series B, Statistical methodology. 2011; 73:753–772. [PubMed: 22323898]

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

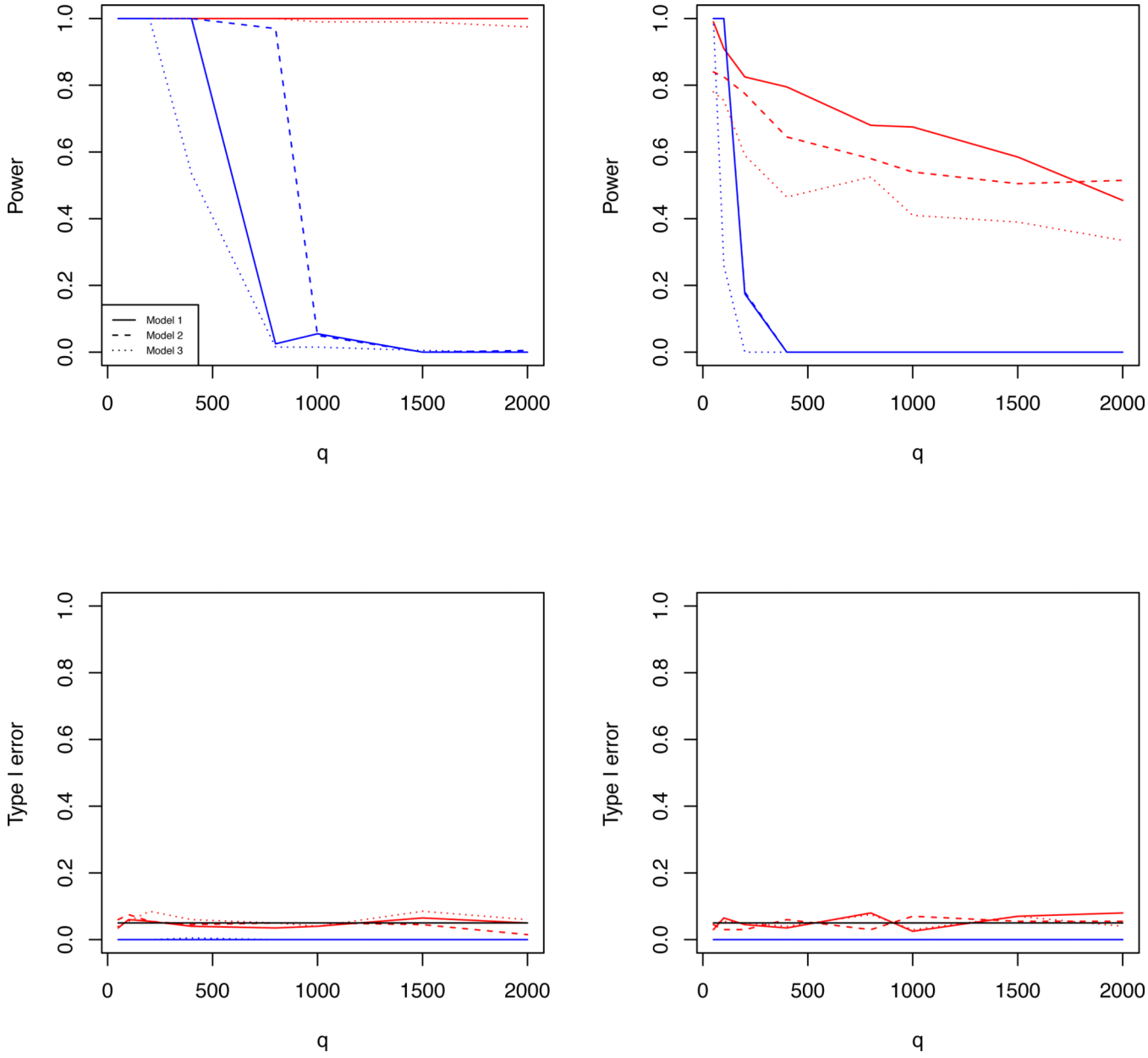

Fig. 1.

Simulation 1 results: the estimated rejection rates as functions of q for two different σ^2 values. The upper and lower rows are, respectively, for powers and for type I error rates, whereas the left and right columns correspond to $\sigma^2 = 1$ and $\sigma^2 = 3$, respectively. In all panels, the lines obtained from SPReM and RP are, respectively, presented in red and in blue, and the results for independence, weak, and strong correlation structures are, respectively, presented as thick, dashed, and dotted lines.

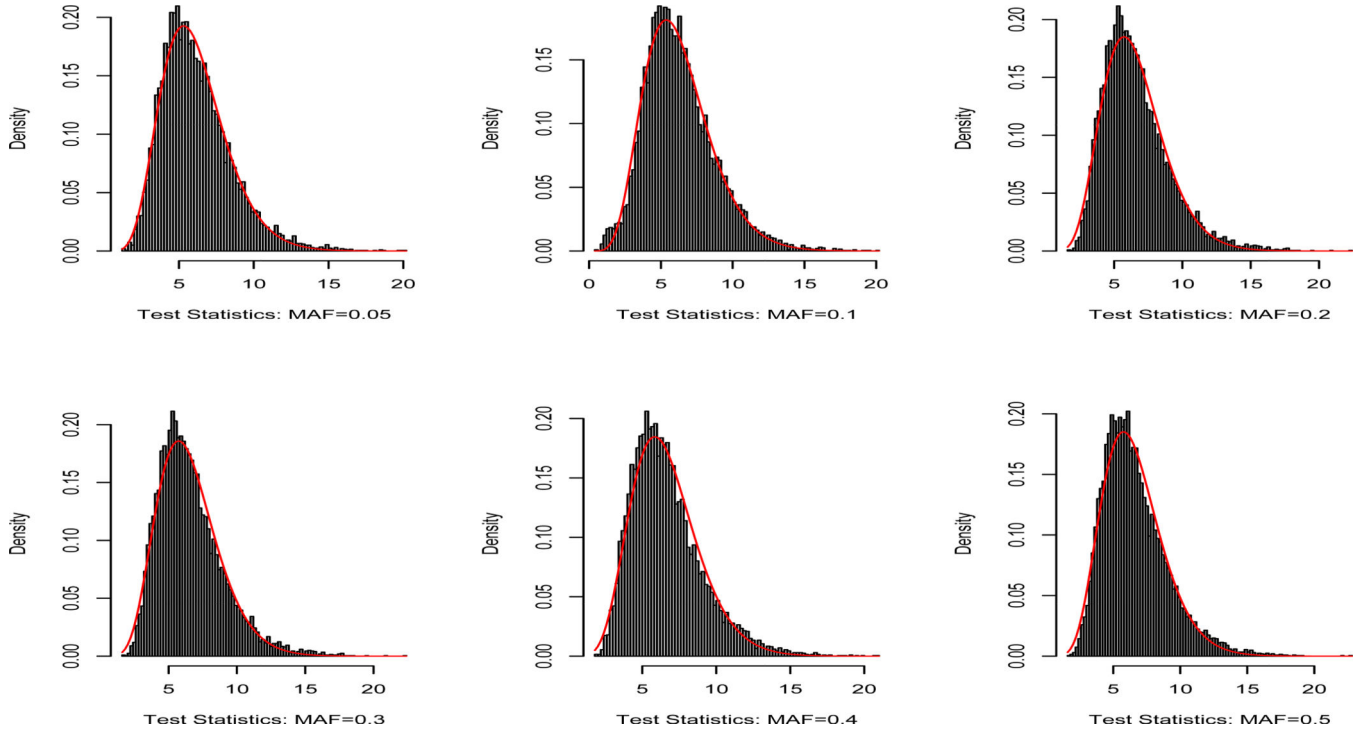

Fig. 2.

Histograms and their gamma approximations based on the wild bootstrap samples under the null hypothesis for different MAFs for $\lambda = \lambda_{max}$.

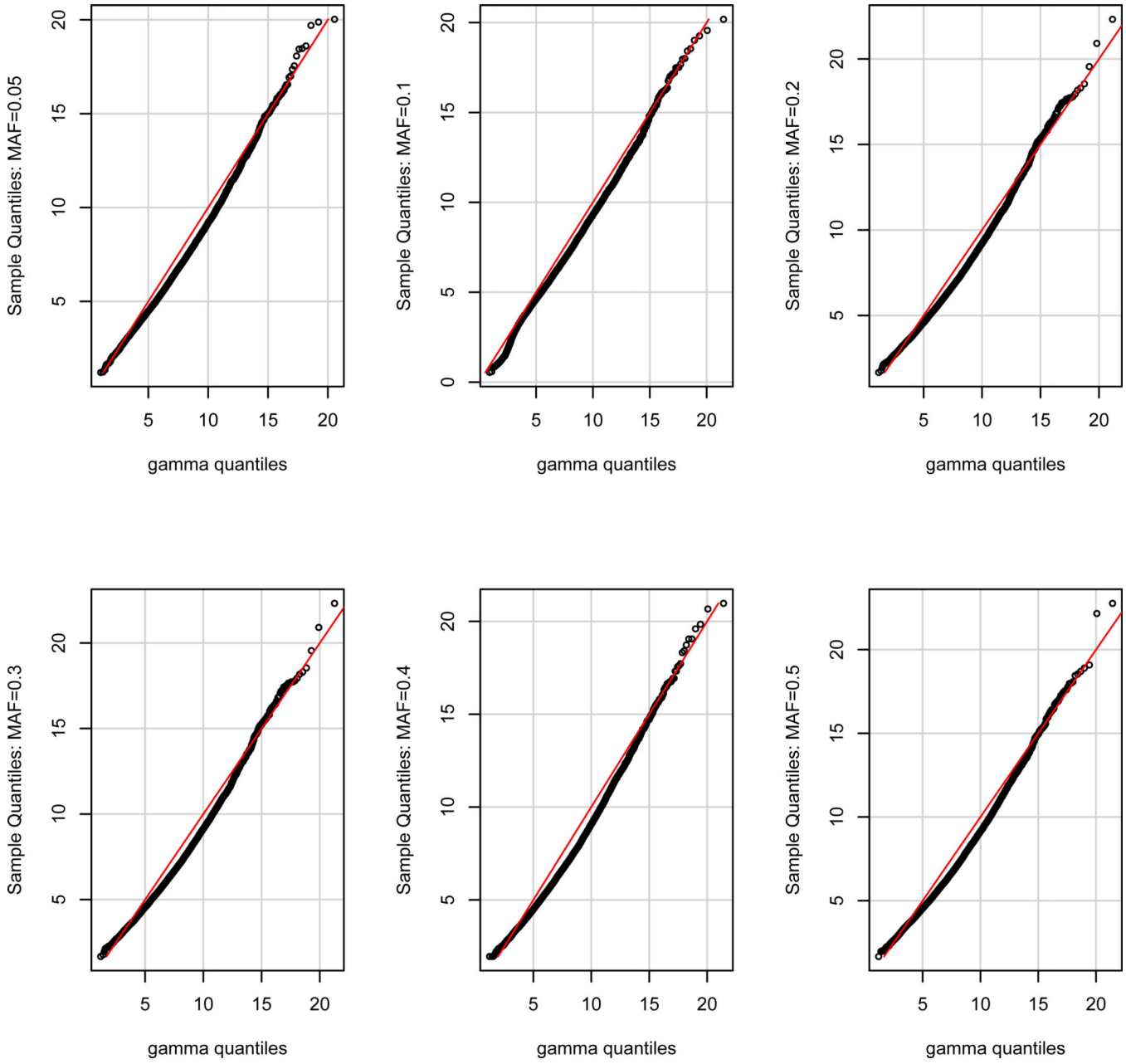

Fig. 3.

QQ-plot of the gamma approximations based on the wild bootstrap samples under the null hypothesis for different MAFs for $\lambda = \lambda_{max}$.

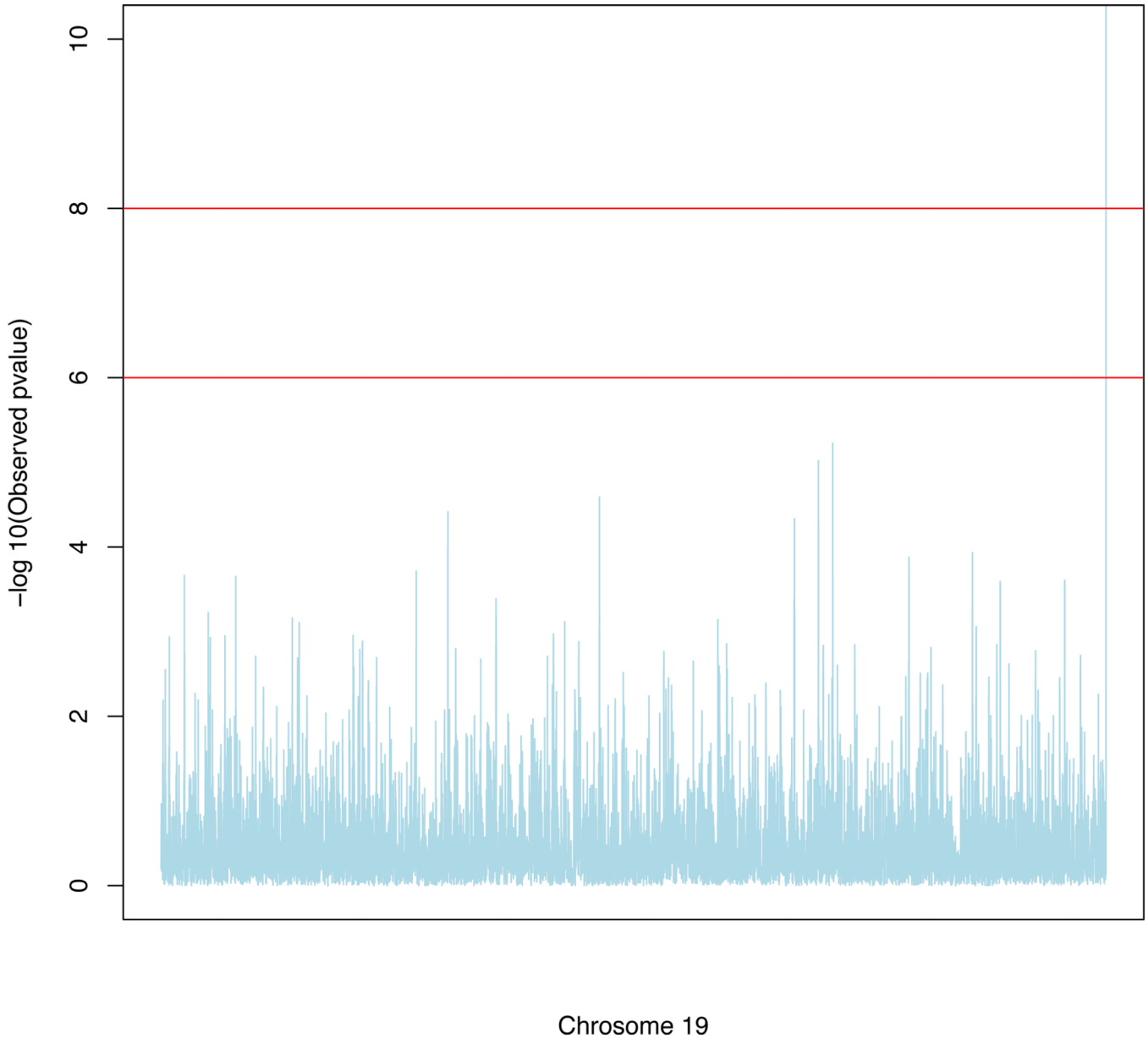

Fig. 4.

ADNI GWAS results: Manhattan plot of $-\log_{10}(p)$-values on chromosome 19 by SPReM for $\lambda = \lambda_{max}$.

Table 1

Simulation 1: power and type I error are reported for two sample test at 5 different *q*s at significance level $\alpha = 5\%$ when $\sigma^2 = 1$.

	Power					Type I error				
q	**50**	**100**	**200**	**400**	**800**	**50**	**100**	**200**	**400**	**800**
case 1										
SPReM	1.000	1.000	1.000	1.000	1.000	0.035	0.060	0.055	0.040	0.035
RP	1.000	1.000	1.000	1.000	0.025	0.000	0.000	0.000	0.000	0.000
HTS	0.965	0.320	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
case 2										
SPReM	1.000	1.000	1.000	1.000	1.000	0.060	0.075	0.055	0.045	0.050
RP	1.000	1.000	1.000	1.000	0.970	0.000	0.000	0.000	0.000	0.000
HTS	1.000	0.245	0.030	0.005	0.000	0.000	0.000	0.000	0.000	0.000
case 3										
SPReM	1.000	1.000	1.000	1.000	1.000	0.040	0.055	0.085	0.060	0.050
RP	1.000	1.000	1.000	0.535	0.015	0.000	0.000	0.000	0.005	0.000
HTS	1.000	0.140	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 2

Simulation 1: power and type I error are reported for two sample test at 5 different *q*s at significance level $\alpha = 5\%$ when $\sigma^2 = 3$.

	Power					Type I error				
q	**50**	**100**	**200**	**400**	**800**	**50**	**100**	**200**	**400**	**800**
case 1										
SPReM	0.990	0.910	0.825	0.795	0.680	0.030	0.065	0.045	0.035	0.080
RP	1.000	1.000	0.175	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HTS	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
case 2										
SPReM	0.840	0.825	0.775	0.645	0.580	0.045	0.030	0.030	0.060	0.030
RP	1.000	1.000	0.180	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HTS	0.105	0.015	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
case 3										
SPReM	0.780	0.755	0.590	0.465	0.525	0.050	0.055	0.050	0.040	0.075
RP	1.000	0.260	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HTS	0.095	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 3

Correlation matrix of responses used in the simulation

	High	Med	Low
High	0.9	0.6	0.3
Med	0.6	0.9	0.1
Low	0.3	0.1	0.1

Table 4

Simulation 2: the estimates of rejection rates were reported at 6 different MAFs, 5 different *q*s, and 2 different σ^2 values at significance level $\alpha = 5\%$. For each case, 100 simulated data sets were used.

	Power					Type I error				
					$\sigma^2 = 1$					
MAF\q	**50**	**100**	**200**	**400**	**800**	**50**	**100**	**200**	**400**	**800**
0.050	0.950	0.955	0.930	0.940	0.930	0.045	0.060	0.030	0.070	0.080
0.100	0.995	0.990	0.990	0.980	0.975	0.045	0.055	0.040	0.045	0.045
0.200	1.000	1.000	1.000	1.000	1.000	0.045	0.045	0.080	0.030	0.060
0.300	1.000	1.000	1.000	1.000	1.000	0.065	0.040	0.020	0.065	0.060
0.400	1.000	1.000	1.000	1.000	1.000	0.050	0.070	0.035	0.060	0.070
0.500	1.000	1.000	1.000	1.000	1.000	0.060	0.050	0.030	0.020	0.035
					$\sigma^2 = 3$					
0.050	0.915	0.875	0.765	0.795	0.735	0.050	0.040	0.030	0.050	0.065
0.100	0.970	0.960	0.940	0.875	0.865	0.040	0.055	0.070	0.080	0.050
0.200	0.995	0.985	0.975	0.975	0.970	0.015	0.050	0.060	0.010	0.065
0.300	1.000	1.000	0.990	0.970	0.955	0.045	0.055	0.055	0.080	0.040
0.400	0.995	1.000	1.000	0.990	0.985	0.055	0.035	0.045	0.050	0.070
0.500	0.995	1.000	1.000	0.985	0.980	0.085	0.055	0.055	0.065	0.030

Table 5

Comparison between SPReM and the massive univariate analysis (MUA) for ADNI data analysis: the top 10 SNPs and their $-\log_{10}(p)$ values for $\lambda = \lambda_{max}$.

SNP	apoe_allele	rs11667587	rs2075650	rs7248284	rs3745341
SPReM	5.04E-16	5.95E-06	9.58E-06	2.56E-05	3.83E-05
MUA	3.43E-11	4.42E-04	1.12E-04	8.75E-04	1.00E-03
SNP	rs4803646	rs8106200	rs2445830	rs8102864	rs740436
SPReM	4.65E-05	1.16E-04	1.32E-04	1.93E-04	2.17E-04
MUA	7.56E-04	3.70E-03	1.33E-02	9.34E-04	1.63E-03