LEARNING METHODS IN REPRODUCING KERNEL HILBERT SPACE BASED ON HIGH-DIMENSIONAL FEATURES

Hojin Yang

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.

Chapel Hill
2016

Approved by:
Joseph G. Ibrahim
Hongtu Zhu
John H. Gilmore
Yun Li
Eric Bair
ABSTRACT

Hojin Yang: Learning Methods in Reproducing Kernel Hilbert Space Based on High-dimensional Features
(Under the direction of Joseph G. Ibrahim and Hongtu Zhu)

The first topic focuses on the dimension reduction method via the regularization. We propose the selection for principle components via LASSO. This method assumes that some unknown latent variables are related to the response under the highly correlate covariates structure. $L_1$ regularization plays a key role in adaptively finding a few liner combinations in contrast to the persistent idea that is to employ a few leading principal components. The consistency of regression coefficients and selected model are asymptotically proved and numerical performances are shown to support our suggestion. The proposed method is applied to analyze microarray data and cancer data.

Second and third topics focus on the approaches of the independent screening and the dimension reduction with the machine learning approach using positive definite kernels. A Key ingredient matter of these papers is to use reproducing kernel Hilbert space (RKHS) theory. Specifically, we proposed Multiple Projection Model (MPM) and Single Index Latent Factor Model (SILFM) to build an accurate prediction model for clinical outcomes based on a massive number of features. MPM and SILFM can be summarized as three-stage estimation, screening, dimension reduction, and nonlinear fitting. Screening and dimension reduction are unique approaches of two novel methods. The convergence property of the proposed screening method and the risk bound for SILFM are systematically investigated. The results from several simulation scenarios are shown to support it. The proposed method is applied to analyze brain image data and its clinical behavior response.
ACKNOWLEDGMENTS

First and foremost, I would like to thank my principal academic advisor Professor Joseph G. Ibrahim. I am grateful to him for his unwavering support for my academic activities. Apart from that, I am privileged to have had the opportunity to be acquainted with such a wonderful person. I would like to thank Professor Hongtu Zhu for his insightful comments about my research, for his painstaking efforts in improving the quality of my writing and for serving as my advisor. I thank Professor John H. Gilmore for his encouragement and advice and also for serving in my orals committee. I thank Professor Yun Li for serving in my orals committee. I thank Professor Eric Bair for his consistent advice and serving in my orals committee.

I am thankful to the entire faculty and the staff members of the department, for their efforts at resolving any problems I faced, and for providing a wonderful academic environment. I would like to thank Statisticians Allison Deal and Dominic Moore, my supervisors at Lineberger Comprehensive Cancer Center for their wonderful guidance and constant support.

It has also been a very pleasant experience as a student at University of North Carolina and for that I must thank to my colleagues Eunjee Lee, Jingwen Zhang, Yue Wang, Chao Huang, Dehan Kong and Mihye Ahn at BIAS lab and many friends Heejoon Jo, Yunro Chung, Wonhee Yang, Kyungsoo Kim, Wansuk Choi, Byeongyeob Choi, Jungin Kim and Hyowon An at the department. Finally, I lack enough words to express my gratitude towards my family - my parents, Dongjin Yang, and Nayeong Kim - whose encouragement, affection and sacrifices have been the biggest driving forces in this endeavor.
# TABLE OF CONTENTS

LIST OF TABLES ................................................................. viii

LIST OF FIGURES ............................................................... x

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW ............... 1

1.1 Introduction ........................................................................ 1

1.2 Principal Component Analysis ........................................... 2
  1.2.1 Introduction for Principal Components ......................... 2
  1.2.2 Unsupervised vs Supervised Learning ......................... 3
  1.2.3 Limiting Spectral and Spike Model ......................... 5
  1.2.4 Sparse Principal Components ......... 6

1.3 Variable Selection ........................................................... 7
  1.3.1 Classical Regularization .................................... 7
  1.3.2 Modern Regularization .................................... 8
  1.3.3 Sure Independent Screening ........................ 13
  1.3.4 Two Stage Model ...................................... 15

1.4 Dimension Reduction ....................................................... 15
  1.4.1 Parametric Dimension Reduction ........................ 15
  1.4.2 Nonparametric Dimension Reduction ................ 18
  1.4.3 Semiparametric Dimension Reduction ................ 19

1.5 Kernel Methods in Machine Learning ............................... 20
  1.5.1 Positive Definite Kernels and Reproducing Kernel Hilbert Space 20
1.5.2 Reviewing Methodologies in Machine Learning on RKHS ........................................... 22
1.5.3 Statistics in Reproducing Kernel Hilbert Space ......................................................... 23

CHAPTER 2: PRINCIPAL COMPONENTS SELECTION VIA LASSO ............................................ 26
  2.1 Introduction ................................................................. 26
  2.2 Description .................................................................. 28
    2.2.1 Regression Method .................................................. 28
    2.2.2 Motivation .............................................................. 31
    2.2.3 Choice Principle Components by LASSO in Regression ......................... 34
    2.2.4 Choice Principle Components by LASSO in Survival Analysis ................. 36
    2.2.5 Tuning Parameter Selection ......................................... 39
    2.2.6 Relation with the Eigenarray Space ..................................... 42
  2.3 Simulation Studies ........................................................ 42
    2.3.1 Simulation 1: Single PC Cases ...................................... 42
    2.3.2 Simulation 2: Multiple PC Cases .................................... 47
  2.4 Real Data Analysis ......................................................... 49
  2.5 Asymptotic Analysis ....................................................... 54
  2.6 Conclusion ................................................................. 59

CHAPTER 3: MPM: MULTIPLE PROJECTION MODEL ......................................................... 60
  3.1 Introduction ................................................................. 60
  3.2 Description ................................................................. 62
    3.2.1 Model Setup .......................................................... 62
    3.2.2 Sure independent Screening via Positive Definite Kernel ......................... 63
    3.2.3 Local Dimension Reduction ......................................... 66
    3.2.4 Kernel Machine Learning ........................................... 67
3.3 Simulation Studies ......................................................... 69
  3.3.1 Simulation 1: Prediction in Regression Problem .................. 69
  3.3.2 Simulation 2: Prediction in Classification Problem ............... 74
  3.3.3 Simulation 3: Prediction in Functional Curve Problem .......... 75

3.4 Real Data Analysis ..................................................... 76

3.5 Conclusion ................................................................. 79

CHAPTER 4: SILFM: SINGLE INDEX LATENT FACTOR
MODEL BASED ON HIGH-DIMENSIONAL FEATURES ................. 83
  4.1 Introduction ............................................................. 83
  4.2 SILFM: Single Index Latent Factor Model ......................... 87
    4.2.1 Model Setup ...................................................... 87
    4.2.2 Estimation Procedure .......................................... 88

  4.3 Simulation Studies ..................................................... 92
    4.3.1 Simulation 1: Continuous Response (I) ....................... 93
    4.3.2 Simulation 2: Continuous Response (II) ..................... 94
    4.3.3 Simulation 3: Binary Response on Random
      Design ................................................................. 97

  4.4 Real Data Analysis .................................................... 99

  4.5 Asymptotic Analysis ................................................. 103

  4.6 Conclusion ............................................................. 108

CHAPTER 5: DISCUSSION .................................................... 109

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2 ................. 110

APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 4 ................. 124

BIBLIOGRAPHY .............................................................. 136
## LIST OF TABLES

2.1 Model consistency in example 1. .......................................................... 45
2.2 Goodness of fit in example 1. ............................................................ 45
2.3 Model consistency in example 2. ........................................................ 48
2.4 Goodness of fit in example 2. ............................................................ 48
2.5 Description of data sets and results ($n$: number of data points, $p$: number of data dimensions, $E$: number of total events, $\hat{\gamma}_{\text{min}}$: minimum penalty in Figure 2.2, $\hat{\gamma}^*$: criterion on 2.23 and PCs: number of selected PCs). .................. 54

3.1 Model consistency in regression. ......................................................... 71
3.2 Sum of prediction error in regression. ............................................... 71
3.3 Model consistency in classification. .................................................. 72
3.4 Average of test error in classification. .............................................. 72
3.5 Model consistency in functional curve. .............................................. 73
3.6 Average of prediction error in functional curve. ................................ 73
3.7 Sum of prediction error in hippocampal surfaces data. ...................... 78

4.1 Performance of SIS methods in simulation 1: HSIC-SIS: screening with HSIC, DC-SIS: screening with correlation distance, SIS: screening with Pearson correlation, KD-SIS: screening with Kendall correlation, SP-SIS: screening with Spearman correlation, $P_{tp}$: true positive rate, $P_A$: screening accuracy, $P_{tn}$: true negative rate, and $P_{fp}$: the false positive rate. .................................................. 95

4.2 Average of prediction errors in simulation 1: Each column indicates the number of $|\hat{M}|$ used for the prediction model. Each row presents a learning method. ............................... 95

4.3 Performance of SIS methods in simulation 3: HSIC-SIS: Screening with HSIC, KS-SIS: Screening with Kolmogrov-Sminov statistic, KD-SIS: Screening with Kendall statistic T-SIS: Screening with t statistic SIS: Screening with Pearson correlation. .................................................. 100
4.4 Average of test error in simulation 3: SILFM: Projection by SILFM, KPCA: Projection by kernel PCs, PCA: Projection by PCs SPCA: Projection by sparse PCs. 100

4.5 Sum of prediction errors in hippocampal surfaces data. 101

4.6 Prediction for behavior score in hippocampal surfaces data: \(\hat{Y}_{\text{number}}\) is the prediction value, where the subscript number is the reduced feature dimension. Sums of prediction errors for \(\hat{Y}_{1000}, \hat{Y}_{700}\) and \(\hat{Y}_{500}\) were 117.3, 118.07 and 118.01 for 180 test observations. Correlation thresholding values were 0.3, 0.5 and 0.7. Estimated smoothing values for them were 0.02, 0.03 and 0.04, respectively. 101
LIST OF FIGURES

2.1 LASSO path for microarray datasets: (A) lung cancer data, (B) breast cancer data, (C) acute myeloid leukemia data and (D) DLBCL data. ................................................................. 50

2.2 Cross validation error for microarray datasets: (A) lung cancer data, (B) breast cancer data, (C) acute myeloid leukemia data and (D) DLBCL data. ................................................................. 51

2.3 Survival probability for microarray datasets: (A) lung cancer data, (B) breast cancer data, (C) acute myeloid leukemia data and (D) DLBCL data. ................................................................. 53

3.1 Images of correlation matrix: (A) left vs right, (B) reorder for both, (C) thresholding at 0.5 (D) thresholding at 0.7. ................................................................. 81

3.2 Smoothing plots: (1) 30000, (2) 1000, (3) MMP at 1000. ................................................................. 81

3.3 Contour plots: (A) MMP, (B) KREG, (C) KPCA, (D) KSPCA. ................................................................. 82

3.4 Smoothing plots: (A) MMP, (B) KREG, (C) KPCA, (D) KSPCA. ................................................................. 82

4.1 Results from simulated data sets: the top row includes a selected covariate image \( x_i \), an informative covariate image \( \tilde{x}_i \), a selected key feature \( z_i \) versus a selected feature of \( \tilde{x}_i \); the second row includes the scatter plot \( y_i \) and the selected feature of \( z_i \) and the average prediction error plots for SILFM1 (red) and SILFM2 (blue) and other competing methods (KRR (orange), SVM (skyblue), LASSO (green) SCAD (darkgreen), PLS (pink), SPLS (hotpink), PCA (black), SPCA (gray), and SSDR (purple)) in two simulation scenarios in subsection 3. ................................................................. 85

4.2 Path diagram of SILFM estimation procedure. ................................................................. 88

4.3 Averages of true positive rate for five different screening methods in simulation 2: panels A \( (\sigma_y = 0.2, \sigma_x = 0.2) \), B \( (\sigma_y = 2, \sigma_x = 0.2) \), C \( (\sigma_y = 0.2, \sigma_x = 1) \), and D \((\sigma_y = 2, \sigma_x = 1) \) report the results based on the different levels of \( \sigma_y \) and \( \sigma_x \). ................................................................. 96
4.4 Averages of prediction errors for 11 different predictive methods in simulation 2: all panels A ($\sigma_y = 0.2, \sigma_x = 0.2$), B ($\sigma_y = 2, \sigma_x = 0.2$), C ($\sigma_y = 0.2, \sigma_x = 1$), and D ($\sigma_y = 2, \sigma_x = 1$) report the results based on the different levels of $\sigma_y$ and $\sigma_x$. .......................... 97

4.5 Correlation matrix for the top 1,000 features selected from the left and right hippocampi. (A) left vs right, (B) reorder for both, (C) thresholding at 0.5 (D) thresholding at 0.7. .......................... 102
CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Accompanying the development of sciences and technologies, recent scientific data has characteristics of increasing in both complexity and size. One tendency of such complexity is the massive amount of available covariates called the ultrahigh dimensionality, which makes it hard the relationship between a response variables and the collection of the covariates. There are also challenges of noise accumulation, computational expediency, statistical accuracy and algorithmic stability due to such complexity. These challenges may need new statistical modeling techniques.

Aims of this paper is to propose several dimension reduction methods via variable selection approaches and to improve the prediction performance by utilizing them in analyzing the ultrahigh dimensional data sets in numerous fields of science and engineering. The main subjective of this thesis is to concentrate on some specific aspects of dimension reduction, variable selection, and machine learning techniques.

To suggest acceptable methodology, we assume more realistic scenarios within the ultrahigh dimensionality circumstance. The first scenario is that the dimensionality of the covariate $p$ also goes to infinity as $n$ goes to infinity. Since traditionally, the dimensionality of the covariate frequently has been assumed as fixed constant, regarding $p$ as sequence is more acceptable in ultrahigh dimensional data sets. The second scenario is that the circumstance of the covariates is a highly correlated circumstance. Since conventionally, it is assumed that the structure of covariates is independent of each other, assuming the covariance matrix not the identity matrix but the low rank matrix is appropriate and general in the ultrahigh dimensionality. The third scenario is that it is believed that there is an arbitrary relationship between the covariate and the
response. While traditionally, the conditional mean in regression was described as the linear function, we assume that the conditional mean in the modeling is described as an arbitrary relationship where it may include both the linear and the non-linear relationship.

We will devise the methodology based on two stage model to address the ultrahigh dimensionality. Our two stage model is divided by the dimension reduction stage and the estimating stage for the conditional mean. To reduce the problem of the large number of covariates, mainly two different approaches are available in statistical literature. Variable selection is the first approach where it is believed that only a few covariates, among all the available covariates, truly are relevant to the response, others are irrelevant and have no real explanatory effect. Various regularization and independent screening methods have been developed for a recent couple of decades to deal with ultra high dimensional data and these are clearly the representative methods in this area. It is the goal of variable selection to identify a few significant covariates. Dimension reduction is the second approach where it is believed that only a few linear combinations of the many covariates relates to the response. In contrast to the variable selection approach, all covariates could have explanatory effect, but the effect is only represented in a few linear combinations. To identify a few linear combinations from all covariates is the goal of dimension reduction. To estimate an unknown function within an arbitrary relationship, we employ the positive definite kernel based on theory of reproducing kernel Hilbert space. We will incorporate specific aspects of dimension reduction, variable selection, and machine learning techniques in favor of our circumstance in this paper.

1.2 Principal Component Analysis

1.2.1 Introduction for Principal Components

Principal component analysis (PCA) is one of the oldest and best popular methods for reducing dimensionality in multivariate problems. Let $X_1, X_2, \ldots, X_n$ be the vectors in $\mathbb{R}^p$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ and $\theta_1, \theta_2, \ldots, \theta_p$ be the eigenvalues and corresponding eigenvectors of $\Sigma = \text{Var}(X)$. Then, the population version of principal components are defined to be linearly the
transformed variables, \( \{\theta'_1 X, \theta'_2 X, \ldots, \theta'_p X\} \). \( \theta_j \) is called the \( j \)-th principal component direction. The sample version of principal components are \( \{\hat{\theta}'_1 X, \hat{\theta}'_2 X, \ldots, \hat{\theta}'_p X\} \) where \( \hat{\theta}_1 \geq \hat{\theta}_2 \geq \cdots \geq \hat{\theta}_p \) and \( \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p \) are the eigenvalues and eigenvectors of the ordinal sample covariance matrix \( \hat{\Sigma} \).

Adcock (1878) wrote the first page of history on principal components, finding the the most probable straight line or hyperplane determined by \( p \) dimensional coordinates of \( n \) points. Pearson (1901) and Hotelling (1933) attempted to explain and analyze the complex statistical variables into principal components. Theoretically, Anderson (1963) derived the consistency of the principal components and the asymptotic distribution for fixed \( p \).

There have been several reasons for why such reduction has been practical in statistical literature for a long time. In terms of traditional point of view, mitigating the effect of collinearity in regression problem and recovering signal structure in denoising process may be included as reasons. Since principal component analysis may be the best linear approximation capturing the maximum variability in covariates, it has been widely used in real data world.

### 1.2.2 Unsupervised vs Supervised Learning

Principle components have also been utilized to reduce dimensionality in the regression problem. Let \( Y \) be the response and suppose that our goal is to reduce the dimensionality of \( X \) before fitting a regression with the response \( Y \) for handling collinearity, providing low dimensional structure or constructing predictable and interpretable model.

One persistent idea for the reduction accompanying principal components is that leading principal components can be employed in place of the covariates \( X \). Many prominent statisticians such as Fisher (1925), Kendall (1965), Hocking (1976) and Mosteller and Tukey (1977) believed that by replacing the covariates by a few leading principal components, the loss information for the response \( y \) could be minimized since the first few leading principal components essentially include the most part of information for the covariates. Fisher (1925) pointed out that reduced statistics to be employed in the regression must not be chosen with considering the response. He
particularly placed an emphasis on not doing the practice where reducible variables are selected by figuring out the relationship between the individual covariate and the response. Their belief may be in the spirit of the idea of the sufficiency. Consider the concept of the sufficiency

\[ X|\{T(X), Y\} \sim X|T(X) \]

where \( T \) is statistics. In other words, the reduction should be done, not depending on the response \( Y \) but containing the same information for the covariates \( X \).

Cox (1968) however suggested entirely opposed thought in his article. He believed that there is no logical reason why the response should not be closely related to the least important principal component. He mentioned that if the covariates \( X \) and the response \( Y \) have a joint distribution or there is an omitted variable \( Z \) which can be an external variable or an internal variable obtained by decomposing the linear combination derived from principal components, employing a few leading principal components is not appropriate. Hotelling (1957) attempted to explain this issue in factor model frame and Hawkins and Fatti (1984) explored the data representation through using minor principal components. Their belief may be in the spirit of the idea of the latent variable. Consider the regression model with the latent variable

\[ Y = f(Z) + \epsilon_y, \quad X = g(Z) + \epsilon_x \]

where \( f \) and \( g \) are arbitrary functions. Consequently, the final goal of the dimension reduction can be the problem to estimate an omitted variable \( Z \). The dimension reduction methodologies involving with the response \( Y \) is an appropriate approach since the response \( Y \) contains the information for \( Z \). Partial least square estimator introduced by Helland (1988) and supervised principal components proposed by Bair et al. (2006) are representative methodologies utilizing the information for the response \( Y \) to reduce the dimensionality of the covariates.

There has been a no decisive rule for reducing its dimensionality within both approaches. There are not also other different approaches excepting these two approaches. Selecting appropriate
principal components in regression model may be a simple and tricky problem. Jolliffe (2005) commented that the choice of principal components for the regression problem remains an open question.

1.2.3 Limiting Spectral and Spike Model

Principal components obtained by the reduction of dimensionality are vectors with \( p \) variables in \( n \) observation. Contemporary data sets have much larger \( p \) than \( n \), there arises the concern for the accuracy of principal components since Johnstone and Lu (2009) proved inconsistency of principal components as \( p \) and \( n \) goes to infinity while Anderson (1963) showed the consistency of them for fixed \( p \). In other words, Johnstone’s condition that \( p/n \rightarrow \gamma \in [0, \infty) \) is more flexible than Anderson’s condition \( p/n \rightarrow 0 \) in high dimensional data sets. Similar assertion happens to sample eigenvalues. Suppose \( X_1, X_2, \ldots, X_n \) are observations from Gaussian model where mean zero and the covariance matrix is the identity matrix. Bai (1999) showed in his review paper that when the population covariance is the identity, the smallest and the largest eigenvalues of sample covariance matrix converge almost surely to \((1-\sqrt{\gamma})^2\) and \((1+\sqrt{\gamma})^2\), respectively. Under the circumstance \( p/n \rightarrow \gamma \), if \( 0 < \gamma < 1 \), then eigenvalues converge almost surely to near 1. There may be no great change but if \( \gamma \geq 1 \), there may be a great change between the sample covariance matrix and the population covariance matrix in high dimensional data sets. Johnstone (2001b) derived the asymptotic distribution for the largest eigenvalue under this circumstance.

Assume that \( X_1, X_2, \ldots, X_n \) are observations from Gaussian model with mean zero and the covariance

\[
\Sigma = diag(\lambda_1, \lambda_2, \ldots, \lambda_M, 1, \ldots, 1)
\]

where \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M > 1 \) and this covariance model is called spiked covariance. The literature on the asymptotic behavior of sample eigenvalues for the spiked covariance has been
recently introduced. Baik and Silverstein (2006) proved $\lambda_j \to (1 + \sqrt{\gamma})^2$ when $\lambda_j \leq 1 + \sqrt{\gamma}$ and Bai et al. (2004) derived the asymptotic distribution of sample eigenvalue $\hat{\lambda}_j$ for the same circumstance. Paul (2007) showed the asymptotic distribution of sample eigenvalues for $\lambda_j > 1 + \sqrt{\gamma}$.

There are two importance in this subsection. The first message is that the result from the asymptotic spectral behavior, we may have an insight why principal components may not be a good tool in a high dimensional setting. The significant condition $p/n \to \gamma \in [0, \infty)$ showed by Johnstone and Lu (2009) is closely related to the asymptotic spectral behavior for the sample covariance. The second message is that phase transition phenomenon may be happened in a high dimensional data sets. This means when the ratio $\gamma$ is larger than zero, the eigenvalues and sample covariance matrix are no longer reliable even if $\Sigma = I$. Hence, spiked covariance model is a realistic structure in a high dimensional setting by magnifying the scale of eigenvalues or being comparable to $\gamma$.

1.2.4 Sparse Principal Components

Donoho (1993) and Johnstone (2003) introduced weak-$l^q$ space $wl^q(C)$ space where $C$ and $q$ are positive constants. Suppose that the coordinates of a vector $\theta \in \mathbb{R}^p$ are $|v|_1, |v|_2, \ldots, |v|_p$ where $|v|_k$ denotes the $k$-th largest element in absolute value. Then, $wl^q(C)$ space is defined as

$$\theta \in wl^q(C) \Leftrightarrow |v|_k \leq Ck^{-1/q}, \quad k = 1, 2, \ldots, p.$$ 

$wl^q(C)$ space is more general space than $L^q(C)$ space since

$$\theta \in \mathbb{R}^p \cap L^q(C) \Rightarrow \sum_{i=1}^p |v_i|^q \leq C^q \Rightarrow \theta \in wl^q(C).$$ 

Johnstone and Lu (2009) assumed that the unknown population principal components $\theta$ satisfy
for some positive constants $q$ and $C$,

$$\theta(n) \in w l^q(C) \quad \text{as } p \to \infty \text{ and } n \to \infty$$

where this condition is called a uniform sparsity condition. Then, they proposed an estimator for $k$-th principal component

$$\hat{\theta}_k^I = \sum_{j \in I} \hat{\theta}_j^k e_j$$

where $\hat{\theta}_j^k = \hat{\theta}_j^k I\{|\hat{\theta}_j^k| \geq \delta\}$ is given by hard thresholding, $I$ is the set of indices $1 \leq j \leq p$ corresponding to the largest $M$ variances ($\hat{\sigma}_j^2 = \widehat{\text{Var}}(v_{ij})$) and $\hat{\theta}_j^k$ is $k$-th component of $j$-th principal component. This thresholded estimator $\hat{\theta}_k^I$ is called sparse principal component and Johnstone and Lu (2009) showed its consistency under the uniform sparsity condition.

There are several different types of sparse principal components for the effort to obtain interpretable and predictable principal components. Zou et al. (2006) and Witten and Tibshirani (2008) suggested variant sparse principal components via $L_1$ regularization for $\theta$. Those sparse principal components have an similar objective function and constraints as the following

$$(\hat{\alpha}, \hat{\theta}) = \min_{\alpha, \theta} \sum_{i=1}^n (X_i - \alpha \theta' X_i)^2 + \mu \sum_{j=1}^p |\theta_j|$$

subject to $\alpha' \alpha = 1$.

From the consequence of this subsection, we reviewed that approaches of principal components in a high dimensional setting have been developed the work relevant to reducing each component in a principal component through threshold or regularization. Details about regularization method will be keeping in the next subsection.
1.3 Variable Selection

1.3.1 Classical Regularization

Suppose that we have \((X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\) from the population \((X, y)\) where it is assumed that the conditional expectation of \(Y\) given \(X\) is a linear function, \(\beta'X\) with \(\beta = (\beta_1, \beta_2, \ldots, \beta_p)'\). Goals of variable selection is to identify all significant variables whose coefficients are not zero and to estimate effective coefficients. It is one of important topics on not only parametric modeling but also nonparametric modeling. Akaike (1998) introduced AIC that minimizes Kullback-Leibler (KL) distance between the selected model and the true model for achieving aims. He showed that the asymptotic expansion of the estimated KL distance is given by

\[-l_n(\hat{\beta}) + \lambda \sum_{j=1}^{p} I(\hat{\beta}_j \neq 0)\]

where \(l_n\) is the log-likelihood function and \(\lambda = 1\). Schwarz et al. (1978) devised BIC for \(\lambda = \log n\). It is known that BIC with the larger observations provides a better result for estimating the dimension of true model than AIC. For the normal log-likelihood case, \(l_n(\hat{\beta})\) becomes \(RSS_d/\sigma^2\) where \(RSS_d\) is the residual sum of squares with the best \(d\) subset. Mallows (1973) proposed \(C_p = RSS_d/s^2 + 2d - n\) statistics where \(s^2\) is the variance estimator of the full model. Cross-validation proposed by Allen (1974) and generalized cross-validation introduced by Wahba and Wold (1975) also play a similar role in variable selection. All statistics mentioned here could be explained as the approach of penalized log-likelihood framework with \(L_0\) norm. Li (1987) and Shao (1997) showed that those statistics are asymptotically equivalent for varying \(\lambda\) in the linear model.

1.3.2 Modern Regularization

Traditionally, \(L_0\) regularization introduced in the previous subsection has been employed for improving the prediction performance and choosing the significant covariates. However, there
may be two difficulties. One difficulty mentioned by Fan and Li (2001) is that the existence of
the most extreme value among coefficients makes the lack of stability and stochastic errors
inherited in there procedure are ignored. Another difficulty is that computation is infeasible in
a high-dimensional setting. Other regularization techniques have been developed for the last
fifteen years to cope with difficulties in high dimensionality. Such regularization techniques can
be generalized as

$$l_n(\beta) - \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

where $p_\lambda(\cdot)$ is a penalty function index by regularization parameter $\lambda \geq 0$.

$L_q$ regularization can be one special case of this general frame by $p_\lambda(\cdot) = \lambda \| \cdot \|_q$. Frank
and Friedman (1993) proposed the bridge regression for $0 < q < 2$, Hoerl and Kennard (1970)
introduced the ridge regression for $q = 2$, Tibshirani (1996) suggested LASSO regression for
$q = 1$ and Zou and Hastie (2005) proposed the elastic regression for the convex combination of
$q = 1$ and $q = 2$ when the response $Y$ is observation from Gaussian distribution. It may be the
natural question which one is better.

Fan and Li (2001) mentioned several characteristics of the appropriate penalty function in his
article.

1. **Unbiasedness**: The resulting estimator is nearly unbiased when the true unknown parameter
   is large to avoid unnecessary modeling bias.

2. **Sparsity**: The resulting estimator is a thresholding rule, which automatically sets small
   estimated coefficients to zero to reduce model complexity.

3. **Continuity**: The resulting estimator is continuous in data to avoid instability in model
   prediction.

As a good penalty function, he pointed out that it should result in an estimator with three
properties. How these properties should be mathematically related to the penalty function is
shown by taking the first order derivative of the univariate penalized least square problem

$$\hat{\beta} = \min_\beta \frac{1}{2} (z - \beta)^2 + p_\lambda(|\beta|)$$

where the solution is given $\text{sgn}(\beta)\{||\beta| + p'_\lambda(|\beta|)\} - z$ by Donoho et al. (1995).

1. **Unbiasedness**: if $p'_\lambda(|\beta|) = 0$ for large $\beta$

2. **Sparsity**: if $\min_\beta \{\beta + p'_\lambda(|\beta|)\} > |z|$

3. **Continuity**: $\min_\beta \{\beta + p'_\lambda(|\beta|)\} < |z|$

$L_q$ penalty with $q > 1$ does not hold the sparsity property, $L_1$ does not hold the unbiasedness property and the $L_q$ penalty with $0 \leq q < 1$ does not hold the continuity condition. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty function and its first order derivative is given by

$$p'_\lambda(|\beta|) = \lambda \{I(|\beta| \leq \lambda) + \frac{(a\lambda - \beta)}{(a - 1)\lambda} I(\beta > \lambda)\}$$

where $a > 2$. SCAD satisfies all three conditions simultaneously and several papers showed that SCAD has nice theoretical results associated with the sampling property and performance as variable selection methodology.

The sampling properties of regularization method in variable selection have been considerably studied and many results in the literature have been developed within four types of endeavors: consistency, persistency, selection consistency and oracle property mentioned by Fan and Li (2006) and Zhao and Yu (2006). Consistency means the accuracy of estimated model parameter and it appears in many statistical contexts where it investigates the limiting behavior of estimated model parameter and identify the asymptotic distribution of estimated model parameter. The persistency means the accuracy of the expected loss of the estimated model and it frequently appears in machine learning problem where it investigates the limiting behavior of the risk. Selection consistency means consistency of the selected model. Oracle property means that
for any model selection procedure, when true parameter is decomposed as sparse subset and non-sparse subset, estimator corresponding to sparse subset goes to zero with the probability one and estimator corresponding to non-sparse subset attains an certain information bound mimicking the information of true parameter.

Knight and Fu (2000) provided two important results for the consistency of LASSO. Assuming that \( \lambda / \sqrt{n} \to O_p(1) \) and that \( \lambda / n \to o_p(1) \) and \( \lambda / \sqrt{n} \to \infty \), respectively, results are summarized as

\[
\hat{\beta}^L \overset{p}{\to} \beta, \quad \frac{n}{\lambda} (\hat{\beta}^L - \beta) \overset{d}{\to} V
\]

where \( V \) is some random variable. For LASSO, it is difficulty to be compatible that these two results are simultaneously satisfied. Either the rate of consistency or consistency in variable selection may be sacrificed.

Fan and Li (2001) also provided two significant results for the consistency and the oracle property of SCAD. Assuming that \( \lambda = o(\min_{1 \leq j \leq s} |\beta_j|) \) and \( \lambda \sqrt{n} \to \infty \) results are given as

\[
\hat{\beta}^S \overset{p}{\to} \beta, \quad \sqrt{n}(\hat{\beta}^S - \beta) \overset{d}{\to} N(0, I(\beta)^{-1})
\]

where \( j \in M_s = \{1 \leq j \leq p \mid \beta_j \neq 0\} \) and \( s = |M_s| \). For SCAD, both the rate of consistency and consistency in variable selection are achievable.

Zou (2006) proposed the adaptive LASSO by using an adaptively weighted \( L_1 \) penalty term, \( \lambda \sum w_j |\beta_j| \) to address inconsistency issue of LASSO where the weight is suggested by \( |\hat{\beta}|^{-\gamma} \) and \( \gamma > 0 \). Assuming that \( \lambda / \sqrt{n} \to o_p(1) \) and \( \lambda n^{(\gamma - 1)/2} \to \infty \), he showed that

\[
\hat{\beta}^{aL} \overset{p}{\to} \beta, \quad \sqrt{n}(\hat{\beta}^{aL} - \beta) \overset{d}{\to} N(0, I(\beta)^{-1})
\]

By taking weight as least square estimator, since it is root-n consistency estimator, achieving both the rate of consistency and consistency in variable selection are possible for the adaptive
LASSO.

Those rates of the regularization parameter for LASSO and SCAD are not only way to show the model selection consistency. The condition on the design matrix can be also involved with the aim of variable selection. Zhao and Yu (2006) studied the model selection consistency for LASSO, and introduced the conception called the sign consistency. It is shown that

$$\mathbb{P}(\text{sgn}(\hat{\beta}_L) = \text{sgn}(\beta)) \xrightarrow{p} 1, \quad \text{as } n \to \infty$$

if the design matrix satisfies \(|X'McX'M(X'MX)^{-1} \text{sgn}(\beta_M)|_\infty < 1\). The condition mentioned above is called as irrepresentable condition. Irrepresentable plays a pivotal role in restricting the covariance between each covariate contained sparse subset and each covariate contained non-sparse subset scaled by the variance of covariates contained sparse subset. It is similar to restricted isometries condition devised by Candes and Tao (2005) in linear optimization problem.

The model selection consistency by LASSO is defined to be

$$\mathbb{P}(\hat{\mathcal{M}}_L = \mathcal{M}) \xrightarrow{p} 1, \quad \text{as } n \to \infty$$

where \(\hat{\mathcal{M}}_L = \{1 \leq j \leq p | \hat{\beta}_j^L \neq 0\}\). As Zhao and Yu (2006) pointed out, the consistency of parameter estimation does not necessarily select the correct model. However, reverse is true. Since the sign consistency is equivalent to the model selection consistency, the sign consistency is stronger than the consistency.

Sparse representation by \(L_1\) regularization appears as variant methodologies in statistics, machine learning and operation research. As one of variant methods, Candes and Tao (2007) proposed Dantzig selector as the solution of the following linear optimization problem

$$\min_{\beta} \|eta\|_1 \quad \text{subject to } \|n^{-1}X'(Y - X\beta)\|_\infty \leq \lambda.$$
selector uses the maximum covariance between each covariate and the residual error as the penalty. Since goodness of fit function is involved with the constraint, Danzig selector provides the estimator that minimizing its $L_1$ norm near least square estimator. Under uniform uncertainty principle (UUP) that any submatrices with $n \times s$ of the design matrix, $X$ are uniformly close to orthonormal matrices in a sparsity scenario, Dantzig estimator attains the risk of the oracle estimator up to a logarithmic factor $\log p$,

$$\|\hat{\beta}^D - \beta\|^2 \leq C \sqrt{2 \log p} / n (\sigma^2 + \sum_{j \in M} \max (\beta_j^2, \sigma^2))^{1/2}$$

where $C$ is constant and $\lambda$ is chosen as $\sqrt{2 \log p} / n$. Subsequent study for the relation between Dantzig selector and LASSO under the nonparametric regression model was done by Bunea et al. (2007) and Bickel et al. (2009). Simplifying nonparametric function as the linear function, They showed that the order of the oracle inequality for LASSO is the same to the order of that for Dantzig selector and that Dantzig selector is asymptotically equivalent to LASSO.

Among regularization methods aforementioned, LASSO has received much attention and the huge number of literature devoted to study properties of LASSO has been studied and still ongoing.

### 1.3.3 Sure Independent Screening

In spite of the remarkable development of the regularization methods, there are several concerns about applying it to ultra high dimensional problem. First concern is that the computational cost for the large number of covariates is very expensive such as implementing optimization tools. Second is that as the dimensionality increases, the risk of the ideal estimator is also increased since its convergence rate is captured as logarithmic factor $\log p$, which was treated as constant term in previous subsection. Lastly, the notion of the irrepresentable condition (Zhao and Yu 2006) or the uniform uncertainty principle (Candes and Tao 2007) may not be held. Consequently, it means that there is no guarantee the consistency, the model selection
consistency, the persistency and the oracle property in ultra high-dimensional setting.

Fan and Lv (2008) introduced a novel concept called sure screening property where all important variables survival after applying a variable screening procedure with the probability tending to one. If sure screening property holds, variable screening could be a desirable variable selection method. He proposed a simple sure screening method employing marginal regression or equivalently a correlation learning. Specifically, for any given \( s \), take the selected submodel to be

\[
\widehat{M}_s^S = \{ 1 \leq j \leq p \mid |\rho_j| \text{ is among the first } s \text{ largest of all} \}
\]

where \( \rho_j \) is the marginal correlation of \( j \)-th covariate and the response \( Y \).

It is believed that according to the marginal correlation with the response, such a correlation learning ranks significant variables and filters out variables with the weaker marginal correlation. They called this correlation learning method Sure Independence Screening (SIS) for the reason that individual covariate or feature is used independently and its usefulness is determined by how it is related to the response.

Let \( \mathcal{M}_s = \{ 1 \leq j \leq p \mid \beta_j \neq 0 \} \) be the true sparse model and non-sparsity size \( s = |\mathcal{M}_s| \) under the circumstance, \( p \gg n \) and Gaussian model. Fan and Lv (2008) showed the theoretical finding of SIS, assuming the following conditions for \( \alpha \in (0, 1 - 2\kappa) \)

\[
\min_{j \in \mathcal{M}} |\beta_j| \geq cn^{-k}, \quad \min_{j \in \mathcal{M}} |Cov(\beta_j^{-1}Y, X_j)| \geq c,
\]

\[
\log p = O(n^\alpha) \quad \text{and} \quad \lambda_{\text{max}}(\Sigma) = O(n^\tau)
\]

where \( \tau, \kappa \) and \( c \geq 0 \) and \( \lambda_{\text{max}} \) is the largest eigenvalue. If these conditions are held, there exists \( \gamma \in (2\kappa + \tau, 1) \) such that if \( 2\kappa + \tau < 1 \) and \( s = O_p(n^\gamma) \) then

\[
P(\mathcal{M}_s \subset \widehat{M}_s^S) = 1 - O_p(p e^{-Cn^{1-2\kappa}/\log n}).
\]
where $C$ is constant. It is shown that SIS reduces exponentially high dimensionality to a large scale $s$, estimated model $\hat{M}_s$ includes all significant variables, tending the probability 1. It is chosen as $s = n - 1$ or $s = [n/\log n]$ in conservative circumstance or $s > n$ in containing significant variables with large probability.

There are various applications to generalized linear models, classification problems under different loss functions or even nonparametric learning for keeping significant features to construct a successful model and to obtain better prediction performance for a recent few years.

Fan and Song (2010) proposed to use the ranking of marginal coefficients in generalized linear model and showed its sure screening property. Sequentially Fan et al. (2011) extended those to nonparametric circumstance. Zhao and Li (2012) employed the ranking of those in Cox proportional hazard regression model and proved its sure property. To address the restriction on model assumption, Li et al. (2012a) suggested to use the ranking of robust rank statistics for keeping features in a parametric or nonparametric relationship with the response and showed its sure property and Zhu et al. (2011) extended it as semiparametric approach. Also, Li et al. (2012b) introduced the screening via distance correlation learning (Szekely et al. 2007) to achieve the same purposes. Similar development has been done in classification problem. Mai and Zou (2013) proposed marginal Kolmogorov-Smirnov statistic and Cui et al. (2014) introduced to employ marginal empirical cumulative statistics.

1.3.4 Two Stage Model

Developing various regularizations and independent sure screening methods in variable selection, there is an attempt to incorporate these two different approaches as a unified methodology. Fan and Lu (2008) proposed two stage model where the first stage is to obtain significant features in ultrahigh dimensional modeling issue by using variable screening with the sure screening property and the second is to employ any kind of regularizations to achieve the ideal risk or the best prediction. The procedure such as SIS-SCAD, SIS-LASSO and SIS-Danzig selector showed to deal with the aforementioned several challenges better than other methods in his article. It has
been clearly one of practical methodologies on variable selection to address noise accumulation, computational expediency, statistical accuracy and algorithmic stability happened in the ultrahigh dimensionality.

1.4 Dimension Reduction

1.4.1 Parametric Dimension Reduction

Suppose that the linear regression model is given by

\[ Y = \alpha \beta' X + \epsilon \]

where \( \alpha \in \mathbb{R}^m \) and \( \epsilon \sim N(0, \sigma_Y^2) \). Let \( \beta \) be a matrix with \( p \times m \) and \( S(\beta) \) be the subspace spanned by basis for column space of \( \beta \). Li (1991) defined a dimension reduction space for the regression of \( Y \) on \( X \) to be any subspace \( S(\beta) \) such that

\[ Y \perp X | \beta' X \]

where \( \perp \) denotes independence and \( m \leq p \). The conditional independence of the response \( Y \) and the covariates \( X \) given on \( \beta' X \) implies the following fact

\[ P(Y|X) = P(Y|\beta' X) \]

and thus, \( \beta' X \) contains sufficient information for the regression. The smallest dimension reduction space is defined to be the central space (Cook 2009) denoted by \( S_{Y|X} \). If it is assumed that there exists such an identifiable \( \beta \), the main issue of dimension reduction problems is to estimate \( S_{Y|X} \) by finding the smallest number of basis \( m \) for \( S(\beta) \). Cook and Li (2002) similarly defined a mean dimension reduction space for the regression \( Y \) on \( X \) to be any subspace \( S(\beta) \) such that

\[ Y \perp \mathbb{E}(Y|X)|\beta' X. \]
From the conditional independence, it also implies

$$\mathbb{E}(Y|X) = \mathbb{E}(Y|\beta'X).$$

The *central mean space* denoted by $S_{\mathbb{E}(Y|X)}$ is defined as the smallest mean dimension reduction space to be estimated. It is known that $S_{\mathbb{E}(Y|X)}$ is a subspace of $S_{Y|X}$. Due to the weaker assumption and the replaceable reason, it is often that estimating $S_{\mathbb{E}(Y|X)}$ is useful when the conditional mean is the main concern which happens in multivariate regression.

Inverse regression methods are utilized for estimating $S_{Y|X}$ where the conception of inverse regression is reversing the relationship between the response and the covariates. Inverse regression methods need two necessary conditions given by

$$\mathbb{E}(X|\beta'X) = P_\beta X, \quad Cov(X|\beta'X) = I - P_\beta.$$

where $P_\beta$ is a projection operator onto $S(\beta)$. These required conditions are called the linearity condition that the expectation of the covariates conditional on $\beta'X$ needs be laid on the column space of $\beta$ and the constant variance condition that the covariance of that conditional on $\beta'X$ needs to be laid on the orthogonal complement space of that. Hall and Li (1993), Box and Cox (1964) and Cook and Nachtsheim (1994) proposed treatments by re-weighting or transformation when these conditions are not satisfied. Suppose these two conditions are held. Then, $\beta$ is estimated as the solution of the following optimization problem

$$\max_{\beta} Cov(\mathbb{E}(X|Y)) \quad \text{subject to} \quad \beta'\beta = I_m$$

where the solution $\hat{\beta}'$ is given as eigenvectors for the objection function. The different approaches of estimating the covariance of the conditional mean of $X$ on $Y$ is the main difference among inverse regression methods. Li (1991) suggested slice inverse regression method by partitioning the range of $Y$ into slices, by obtaining averages over each slice and by estimating the sample
covariance of them. Right after, Cook and Weisberg (1991) introduced slices average variance estimation (SAVE) by using different objective function

$$\mathbb{E}(I - Cov(X|Y))^2.$$ 

Li et al. (2005) introduced contour regression to find $S_{Y|X}$ and Cook and Li (2002) proposed the principal hessian direction (PHD) to capture $S_{E(Y|X)}$. Cook and Ni (2005) theoretically showed that the inverse regression estimator is asymptotically efficient estimator and has an asymptotic chi-squared distribution. If the conditional normality of the covariates given on the response is assumed in contrast to two traditional moment assumptions mentioned above, the different approach may be needed to estimate $S_{Y|X}$. Cook et al. (2008) suggested the inverse regression by using principal fitted components under such a circumstance.

Cook et al. (2010) introduced *envelope model* as the effort to reduce $Y$ in multivariate linear regression. Suppose the following multivariate linear regression model

$$Y = \alpha + \beta X + \epsilon$$

where $Y \in \mathbb{R}^r$ and $\epsilon \sim N_r(0, \Sigma_y)$. If there exists a orthogonal matrix $M = (\Gamma_1, \Gamma_0)$ such that

$$S(\beta) \subset S(\Gamma_1) \quad \text{and} \quad \Gamma'_1 Y \perp \Gamma'_0 Y | X$$

$\Gamma'_0 Y$ does not depend on $X$ and $\Gamma'_1 Y$ does not lose all the information on $X$. Cook et al. (2010) defined $M$-envelope of $\beta$ denoted by $E_M(\beta)$ to be the smallest subspaces of $M$ possessing two conditions above. The notion of envelope conceptually seems to be similar to the central space. He specified the likelihood function as the function of $E_{\Sigma_y}(\beta)$ and derived maximum likelihood estimator. Cook et al. (2013) used uses $E_{\Sigma_x}(\beta)$ and partial least square method to estimate $\beta = P_{E_{\Sigma_x}(\beta)}\hat{\beta}$ in the same model.
1.4.2 Nonparametric Dimension Reduction

Suppose that the single index model introduced by Ichimura (1993) is given by

\[ Y = f(\beta'X) + \epsilon \]

where \( \beta \in \mathbb{R}^{p \times m} \), \( \epsilon \) has a certain distribution such that \( \mathbb{E}(\epsilon|X) = 0 \) and \( f(\cdot) \) is an unknown smooth function. Aim of nonparametric dimension reduction is to estimate \( S(\beta) \) under the presence of an unknown function. \( \beta \) is estimated as the solution of the following optimization problem

\[
\min_{\beta} (\mathbb{E}(Y - f(\beta'X))^2).
\]

Nonparametric regression technique is required technique since the objective function is involved with an unknown distribution or regression function. Xia et al. (2002) introduced the minimum average variance estimation (MAVE) method which is a representative method among nonparametric dimension reduction tools. To approximate an unknown function and address unknown distribution, utilizes the first order of Taylor expansion for an unspecified regression function at a local point

\[
f(\beta'X) \approx f(\beta'X_0) + \nabla f(\beta'X_0)'\beta'(X - X_0)
\]

Then, assuming \( \beta'\beta = I_m \) for the identifiable issue, the objective function is given by

\[
\min_{\alpha,\gamma,\beta} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \alpha_j - \gamma_j'\beta'(X_i - X_j)\}^2 K_h(\beta'(X_i - X_j))
\]

where \( \alpha_j = f(\beta'X_j) \), \( \gamma_j = \nabla f(\beta'X_j)'\beta'(X_i - X_j) \) and \( K_h(\cdot) \) is an density kernel function depending the smoothing parameter \( h \). The solution \( \hat{\beta}^M \) is given as eigenvector for \( \frac{1}{n} \sum_{j=1}^{n} \gamma_j\gamma_j' \). MAVE method employs a local polynomial regression method to estimate \( S_{Y|X} \). Xia (2007)
proposed the density based MAVE (dMAVE) to estimate $S_{Y|X}$.

Yin et al. (2008) utilized a different object function which is associated with Kullback-Leibler distance to recover the central mean space. Nonparametric dimension reduction does not need two moment restrictions but there is a limitation on it. Since all covariates are associated with the local smoothing procedure, all covariates should be continuous variable. Appearing discrete covariates, it may not be an appropriate dimension reduction method.

1.4.3 Semiparametric Dimension Reduction

Semiparametric dimension reduction problem is relatively new notion and introduced by Ma and Zhu (2013a). The complete family of influence functions introduced by Bickel et al. (1993) and Tsiatis (2007) are used for estimate $S_{Y|X}$. The advantages of semiparametric reduction does not require moment conditions in parametric approach or continuous condition in nonparametric approach. It is more flexible dimension reduction method.

The fundamental idea of semiparametric dimension reduction is similar to the parameter estimation in the presence of nuisance parameter. Influence function provides the advantage that allows to avoid the problem estimating nuisance parameters. The likelihood of one observation is decomposed as

$$l(X, Y : \beta) = f(X)g(Y, \beta'X)$$

where $f(\cdot)$ is the marginal density function of covariates and $g(\cdot)$ is the conditional density function of $Y|\beta'X$. Aim of semiparametric dimension reduction is to estimate $\beta$ in the conditional density function and marginal density function involved with nuisance part. One complete family of influence function is defined as

$$\{h(Y, X) - \mathbb{E}(h|\beta'X, Y) \mid \mathbb{E}(h|X) = \mathbb{E}(h|\beta'X), \forall h \in \mathcal{F}\}$$

where $\mathcal{F}$ is an arbitrary functional space. By replacing each likelihood function by the influence
function with complete family, estimation equation is given by

$$\sum_{i=1}^{n} \left\{ g_i(Y, \beta'X) - \mathbb{E}(g_i|\beta'X) \right\} \left\{ f_i(X) - \mathbb{E}(f_i|\beta'X) \right\}$$

where $g_i$ and $f_i$ are any functions in $\mathcal{F}$. This frame has one strong advantage since it is equipped with double robustness. For misspecification of parametric assumption, conditional density function allows for holding estimation equation while for misspecification of nonparametric assumption, marginal density function allows for holding estimation equation.

1.5 Kernel Methods in Machine Learning

1.5.1 Positive Definite Kernels and Reproducing Kernel Hilbert Space

Learning method using positive definite kernels have become popular in machine learning for a couple decades. Due to its strong mathematical background, many statistician and mathematic-ian have a great interest in these methods. Conventionally, they have developed theory and algorithms of machine learning and statistics in the linear frame. However, since it is devised for the linear case, there may be some limitation on real data analysis problem. It is necessary for nonlinear case to capture the dependent relationship leading successful prediction.

Positive definite kernel is characterized if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

for every $x_1, x_2, \ldots, x_n \in \mathcal{X}$ and for every $c_1, c_2, \ldots, c_n \in \mathbb{R}$ where we denote $\mathcal{X}$ as the input space and $x_i$ as the covariate from $i$-th observation in this subsection. By utilizing positive definite kernel, it can have the best performance on both linear and nonlinear case. To construct RKHS, mathematically, we need a required feature map from the input space into the some inner
product space called feature space, specifically

\[ \phi : x \in \mathcal{X} \rightarrow \phi(x) \in \mathcal{F}. \]

If it is given, we can obtain Hilbert space whose dense set is isomorphic to the feature space with the inner product that is the form of the limit for the inner product between two Cauchy sequences from the completion procedure. Detail proof is shown by Berlinet and Thomas-Agnan (2004). From such a completion, we have a Hilbert space \( \mathcal{H} \) with the inner product given by

\[ (\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}) = (\mathcal{F}^{\ast}, \lim_{n \rightarrow \infty} \langle \cdot, \cdot \rangle_{\mathcal{F}^{\ast}}) \]

where \( \mathcal{F}^{\ast} \) whose all the elements are Cauchy sequence is a dense set of \( \mathcal{H} \). If \( \forall f \in \mathcal{F} \) is represented as constant sequence by \( f = (f, f, f, \ldots) \) in \( \mathcal{F}^{\ast} \), Feature space \( \mathcal{F} \) and dense set \( \mathcal{F}^{\ast} \) have isomorphic relation. Then, by defining evaluation functional on the space for all input points,

\[ e_{x} : \mathcal{H} \rightarrow k(x, \cdot) \in \mathcal{H}, \text{ such that } e_{x}(f) = f(x) \]

where it is called reproducing property, there is a reproducing kernel Hilbert Space (RKHS) distinguished from functional \( L_{2} \) space.

\[ (\mathcal{H}_{k}, \langle \cdot, \cdot \rangle_{\mathcal{H}}) \]

By Aronszajn (1950)’s theorem, for every positive definite kernel on input space, there exists a unique RKHS and vice versa. Also, since \( \langle f, k(x_{i}, \cdot) \rangle_{\mathcal{H}} = 0 \) implies \( f = 0 \) from the reproducing property, the set \( C = \{ k(x_{1}, \cdot), k(x_{2}, \cdot), \ldots, k(x_{n}, \cdot) \} \) can be complete orthogonal
system (CONS). Thus, we have

$$\mathcal{F}^* = \{ f \mid f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \forall \alpha_i \in \mathbb{R} \}.$$ 

As a result, in real world, utilizing one positive definite kernel implies that the input space which is normally Euclidean space is transformed into the corresponding RKHS. Positive definite kernel is sometimes called kernel and its sample version is called Gram matrix. Substituting infinite dimensional inner product on RKHS by the kernel value is called kernel trick in this area. Such a mathematical background provides the justification on employing a kernel or gram matrix for figuring out the complex nonlinear relationship and employing a kernel trick for allowing the computational price to be cheaper.

### 1.5.2 Reviewing Methodologies in Machine Learning on RKHS

As mentioned it on previous subsection, kernel methodology is analysis by transforming data into a high dimensional feature space given by RKHS. Schölkopf and Smola (2001) and Hofmann et al. (2008) organized various learning methods on the RKHS frame. Schölkopf et al. (1998) applied the principal component on RKHS, maximizing the variance of the feature map where it is called Kernel principal component capturing nonlinear feature. Scholkopft and Mullert (1999) and Melzer et al. (2001) introduced a kernel Fisher discriminant analysis, maximizing the ratio that is the between variance on RKHS divided by the within variance on RKHS. It provides more flexible nonlinear discriminant function on RKHS rather than the straight line on Euclidean. Vapnik (1998) showed that the support vector machine with kernel provides better classifier for non-separable problem and that its solution is given by solving a quadratic optimization for its dual form on RKHS. Shawe-Taylor and Cristianini (2004) proposed canonical correlation analysis (CCA) on RKHS and Fukumizu et al. (2007a) showed its consisten -cy.

Other variant methods such as kernel logistic regression by Zhu and Hastie (2001), kernel
partial least square by Rosipal and Trejo (2002), kernel independent component by Bach and Jordan (2003), support vector regression by Basak et al. (2007) and kernel supervised principal component by Barshan et al. (2011) are suggested for recent fifteen years. All these approached have the common significant characteristic in finding the final optimal solution under the given loss function.

The solution of all minimization or maximization is achieved at some function in RKHS. Since we have CONS for the dense set, The optimal functional $f^*$ is represented by CONS. This principle is called the representation theorem. By the representation theorem, the optimization in an high or infinite dimensional space can be reduced to the optimization in a subspace of sample size dimension. Hence, various classical linear methods of data analysis can be extended with the kernel or gram matrix which leads the linear solution for each optimization problem on RKHS.

1.5.3 Statistics in Reproducing Kernel Hilbert Space

By Riesz’s representation theorem, the existence and uniqueness of the feature map are guaranteed for RKHS. It is possible to extend the notion of feature map to the embedding of a probability measure where details are explained by Sriperumbudur et al. (2008). If we assume that there exists another bounded linear functional such that it is an injective mapping from $\mathbb{P}$ to RKHS, specifically

$$T : \mathbb{P} \in \mathcal{P} \rightarrow T(\mathbb{P}) = \int f d\mathbb{P} = \mathbb{E}(f)$$

by applying Riesz’s theorem again, there exists an unique representer $\mu_\mathbb{P} \in \mathcal{H}$ such that $\langle f, \mu_\mathbb{P} \rangle_{\mathcal{H}} = \mathbb{E}(f)$. Riesz’s representer $\mu_\mathbb{P}$ particularly for the bounded expectation operator is called mean element in RKHS. As the mean statistics on Euclidean space includes the central information, $\mu_\mathbb{P}$ on RKHS contains the central information on RKHS. If the positive define kernel has the nice property such as characteristic and universality explained by Sriperumbudur et al. (2011), $\mu_\mathbb{P}$ uniquely determines a probability measure. Empirical average of the feature map, $\frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot)$
is estimator for $\mu_P$. Berlinet and Thomas-Agnan (2004) showed

$$\|\hat{\mu}_P - \mu_P\|_2 = O_p(1/\sqrt{n}), \quad \sqrt{n}(\hat{\mu}_P - \mu_P) \xrightarrow{w} G(\cdot)$$

where $G(\cdot)$ is Gaussian process with the covariance function given by

$$\text{Cov}(G(t), G(s)) = \int (G - \mu_P, k(t, \cdot)) \otimes (G - \mu_P, k(s, \cdot)) d\mathbb{P}.$$ 

Gretton et al. (2012) extended the mean element for two samples case and employed two mean elements for the homogeneity test and showed the asymptotic distribution for test statistics where it is the difference of two mean elements.

As the variance or covariance also contains information for the random variable on Euclidean space, there similarly exists cross variance-covariance operator which is defined to be tensor product form of two separated RKHSs. Cross variance-covariance operator is defined as

$$\Sigma_{yx} : \mathcal{H}_x \rightarrow \mathcal{H}_y \quad \text{or} \quad \Sigma_{yx} = \mathbb{E}(\phi(y) \otimes \phi(x)) - \mu_y \otimes \mu_x$$

where it has the property that $\langle g, \Sigma_{yx} f \rangle_{\mathcal{H}_y} = \text{Cov}(g, f)$. Gretton et al. (2005) suggested empirical Hilbert Schmidt norm of the operator $\frac{1}{n} \sum_{i=1} k(y_i, \cdot) \otimes k(x_i, \cdot) \|_{HS}$ as estimators and showed that

$$\|\hat{\Sigma}_{yx} - \Sigma_{yx}\|_{HS} = O_p(1/\sqrt{n})$$

by using U-statistics theory. Gretton et al. (2005) also employed the cross-covariance operator for the independent test. Fukumizu et al. (2007b) extended the cross-covariance operator to conditional cross-covariance operator and proposed its empirical estimator and derived the consistency of it. Fukumizu et al. (2004) used the cross-covariance and the conditional cross-covariance for the purpose of the dimension reduction problem and Fukumizu et al. (2009) showed theoretical results of them in the dimension reduction issue.
CHAPTER 2: PRINCIPAL COMPONENTS SELECTION VIA LASSO

2.1 Introduction

In multiple regression problems, one of the major difficulties with the ordinary least squares (OLS) estimators is the problem of multicollinearity. Principal components (PC) regression is one possible method to overcome this problem. As its name implies, PC regression performs principal components analysis (PCA) on the set of predictor variables and uses the scores for a subset of the PC's as predictors (in place of the original predictor variables).

If all the PC’s are included in the PC regression model, the PC regression estimates will be equivalent to the OLS estimates (and hence will suffer from the same problems resulting from multicollinearity). Thus, one typically selects a subset of the PC’s as predictors. Selecting a subset of the PC’s will result in biased regression estimates. However, selecting an appropriate set of PC’s can substantially reduce the variance in the model, resulting in better predictive accuracy. See Hastie et al. (2009) for details and additional information.

PC regression requires users to select a subset of the PC’s to use as predictors. Several methods for selecting these PC’s have been proposed. One strategy is to delete PCs with eigenvalues less than a prespecified cutoff value Marquardt (1970). However, in some examples, PC’s with low variance are nevertheless strongly associated with the outcome, and excluding these PC’s will reduce the predictive accuracy of the model Kung and Sharif (1980), Jolliffe (1982). On the other hand, simply selecting the PC’s that are most strongly associated with the outcome variable can also produce unsatisfactory results, since the power to detect an association between low-variance PC’s and the outcome is low Mason and Gunst (1985). Such an approach is also prone to overfitting. Other methods seek to identify a subset of PC’s that directly maximize the
predictive accuracy of the model Lott (1972), Soofi (1988), Mertens et al. (1995). For example, Hastie et al. (2009) recommend choosing the \( M \) PC’s with the largest variance and selecting \( M \) using cross validation. However, the problem of identifying the optimal subset of PC’s remains an open question Jolliffe (2005).

As the motivation in this paper, we consider the choice of PCs in regression analysis by using LASSO method. Using LASSO to choose PCs seems to be very simple idea. The reason for doing this simple idea is that we intend to add the model consistency issue in variable selection problem to traditional open problem choosing PCs in regression as the new criterion. For instance, suppose we observe the covariates from Gaussian distribution with mean zero and the covariance of the block diagonal structure of two blocks. All the covariates may be involved with either one of two blocks and they are correlated within each block. Clearly, there would be two distinct PCs corresponded to each block structure at least and the rest of PCs would be corresponded to the eigenvalue of the multiplicity root. If the conditional mean of the response variable is defined as linear function of PC only from one block, then the component choice problem would be equivalent problem to the variable selection in regression model where the goal of variable selection is to find the significant covariates in the model. Can the method depending on the size of variance be justifiable when the portion of variance for the significant block is small portion ? Can the method relying on MSE or PRESS for estimates be justifiable in this situation? Perhaps, these two major methods choosing PCs may not be appropriate. This simple example implicates that there may be the particular situation where the issue for the model selection consistency in variable selection would be needed rather than traditional issues for the size of variance or the minimum MSE when the selection of PCs is main interest.

By using LASSO, nice theoretical properties and effective computation algorithm derived from \( L_1 \) regularization are available. In practice, the method to choose the PCs by using LASSO can be applied to other statistical methodology involved with the issue for the choice of the number of PCs such as factor analysis and independent component analysis as the general methodology. As special cases, it is possible to apply it to supervised principal components
(SPC) analysis proposed by Bair et al. (2006) with issue for selecting the proper number of SPCs and preconditioning method proposed by Paul et al. (2008).

In this chapter, we suggest the method for selecting the appropriate number of PCs by LASSO. We will investigate the fact that the method suggested can correctly pick up PCs when predictors have a certain covariance or correlation structure and the response is associated with that structure in regression analysis in terms of the theoretical view and simulation studies. this chapter is organized as follows. Section 2 gives the details of the suggested method and the motivation, as well as brief summary of PC method in regression and in Cox proportional hazard model. Section 3 discusses asymptotic properties of our method. Section 4 gives a simulation study and we will apply our method to 4 microarray datasets in Section 5. We concludes with the discussion of limitations and future work in Section 6. The Appendix contains details of some proofs for this chapter.

2.2 Description

2.2.1 Regression Method

Let $X$ be an $n \times p$ matrix on $n$ observations and $p$ predictors and $y$ be the outcome measureme-nt of the length $n$. The outcome is a quantitative variable. We assume that the columns of $X$ are centered as mean 0 in the regression model setting so that the intercept term is omitted in this paper. In order to remind PC regression and mention the issue for the choice PCs, we will begin with the brief summary of PC regression method. Almost notations are similar to Jolliffe (2005)’s notation (Chapter 8).

Consider the standard regression model, that is,

$$ y = X\beta + \epsilon \quad (2.1) $$

where $(i, j)$ element of $X$ is the value of the $j$-th predictor variable for the $i$-th observation and $\epsilon_i$’s are independent normal random variables with mean 0 and variance $\sigma^2$. 
The PC score for each observation are given by

\[ Z = XV, \]  

(2.2)

where \((i, k)\) element of \(Z\) is the score of the \(k\)-th PC for the \(i\)-th observation, and \(V\) is a \(p \times p\) matrix whose \(k\)-th column is the \(k\)-th eigenvector of \(X'X\). Since \(V\) is orthogonal, \(X\beta\) can be transformed as \(XVV'\beta = Z\gamma\), where \(\gamma = V'\beta\). Equation (1) can be written as

\[ y = Z\gamma + \epsilon \]  

(2.3)

which has simply replaced the predictor variables by their PCs in regression model. PC regression can be defined as the use of the model (3) or of the reduced model

\[ y = Z_m\gamma_m + \epsilon \]  

(2.4)

where \(\gamma_m\) is a vector of \(m\) elements that are subset of elements of \(\gamma\), \(Z_m\) is an \(n \times m\) matrix whose columns are the corresponding subset of columns of \(Z\). Using least square to estimate \(\gamma\) in (3) and then finding an estimate for \(\beta\) from the transformation

\[ \hat{\beta} = Z\hat{\gamma} \]  

(2.5)

is equivalent to finding \(\hat{\beta}\) by applying least square directly to (1). Since the vector \(\hat{\gamma}\) is

\[ \hat{\gamma} = (Z'Z)^{-1}Z'y = L^{-2}Z'y \]  

(2.6)

where \(L\) is the diagonal matrix whose \(k\)-th diagonal element is \(l_k^{1/2}\) and \(l_k\) is \(k\)th largest eigenvalue of \(X'X\), \(\hat{\beta}\) is given by

\[ \hat{\beta} = \sum_{k=1}^{p} l_k^{-1}v_kv_k'y \]  

(2.7)
where \( v_k \) is the \( k \)th columns of \( V \). Thus, the variance-covariance matrix of \( \hat{\beta} \) is

\[
V(\hat{\beta}) = \sigma^2 \sum_{k=1}^{p} l_k^{-1} v_k v_k'.
\]  

(2.8)

If none of the first \( m \) eigenvalues \( l_k \) is very small and the other eigenvalues are very small value, we can know that eigenvalues corresponding to small values caused to the variance of the large value from equation (8). To avoid from inflating variance, the reduced model in (4) can be one of alternative method and another biased \( \tilde{\beta} \) in the reduced model with \( \hat{\gamma}_m \) can be given by

\[
\tilde{\beta} = \sum_{k=1}^{m} l_k^{-1} v_k v_k' X' y.
\]  

(2.9)

Instead of, deleting terms from (8) corresponding to small eigenvalues, it is also possible to delete terms where the corresponding elements of \( \gamma \) are not significantly different from zero. This point is our motivation in this paper and will discuss detail in the next subsection with simple specific examples. Before discuss the problem for choosing PCs, we will introduce SVD notation in PCs regression and will use SVD notation as main notation since it is convenient and general in the case where the number of predictors, \( p \) is more than the number of observations, \( n \).

Recall that the SVD writes \( X \) in the form

\[
X = ULV'.
\]

Then, \( X\beta \) can be rewritten \( ULV' = U\delta \), where \( \delta = LV'\beta \), so that \( \beta = VL^{-1}\delta \). The least square estimator for \( \delta \) is given by

\[
\hat{\delta} = (U'U)^{-1}U'y = U'y,
\]

leading to \( \hat{\delta} = LV'\hat{\beta} \). The relationship between \( \gamma \), defined earlier, and \( \delta \) is shown as

\[
\gamma = V'{\beta} = V'(VL^{-1}\delta) = (V'V)L^{-1}\delta = L^{-1}\delta.
\]  

(2.10)
So that setting a subset of elements of $\delta$ equal to zero is equivalent to setting the same subset of element of $\gamma$ equal to zero. This result means that the SVD can provide an alternative computational approach for PC regression equations, which provides the efficient algorithm because of orthogonality of $U$.

### 2.2.2 Motivation

In this subsection, we will introduce the motivation for our method with simple examples and extend it to general case. Here, as the first example, we will see that the response, $y$ under the equation (1) can be characterized by the structure of the covariance matrix for predictors, $X$ by linking the principal component derived from the covariance matrix with some pattern with the response. PC regression could be more proper than ordinal regression in such a situation. Consider the following design matrix, parameter and its variance structure under (3).

$$X = \begin{bmatrix} 2 & 1 \\ 2 & -1 \\ -2 & 1 \\ -2 & -1 \end{bmatrix}, \quad \beta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad X'X = \begin{bmatrix} 16 & 0 \\ 0 & 4 \end{bmatrix}, \quad Cov(X) = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}.$$ 

We can easily know that the first and second eigenvalues for the $Cov(X)$ are $l_1 = 4$ and $l_2 = 1$, respectively and corresponding eigenvectors are given by

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$
By using SVD for $X$, we obtain orthonormal basis, $u_1$ and $u_2$ of $\mathbb{R}^4$

$$
\begin{bmatrix}
1/2 \\
1/2 \\
-1/2 \\
-1/2
\end{bmatrix}, \quad
\begin{bmatrix}
1/2 \\
-1/2 \\
1/2 \\
-1/2
\end{bmatrix}
$$

where singular values for $X$ given are 4 and 2.

By (10), in SVD notation, PC regression model can be rewritten by

$$
y = \begin{bmatrix}
1/2 & 1/2 \\
1/2 & -1/2 \\
-1/2 & 1/2 \\
-1/2 & -1/2
\end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \epsilon 
$$

where the transformed coefficient, $\delta = (0, 2)'$ is resulted from multiplying $\gamma$ by the diagonal matrix whose diagonal entries are singular values of $X$.

From (11), we can confirm that the response is associated with $u_2$. Since $u_2$ is the function of the second principal component, $v_2$, we can see that $y$ under (1) is determined by the second principal component. This example is very trivial but there are a few points that in spite of (1) model, the response $y$ can be affected by the particular covariance structure of $X$. Also, we can see the interesting aspect that some coefficients in PC regression are exactly zeros. We pay attention on this aspect while we naturally think of the use for $L_1$ penalty.

As the second example, we will see the possibility that LASSO selects a significant PC for the transformed coefficient. Donoho et al. (1995) derived the solution for $L_2$ loss function with $L_1$ constraint as shrinkage estimator and it is known by

$$
\hat{\delta}_1^L = sgn(\hat{\delta}_{1se}^L)(|\hat{\delta}_{1se}^L| - \gamma)_+ \quad \hat{\delta}_2^L = sgn(\hat{\delta}_{2se}^L)(|\hat{\delta}_{2se}^L| - \gamma)_+ 
$$

(2.12)
where $\hat{\delta}^{lse}_i = u'_i y$, $u_i = \frac{1}{\sqrt{\pi}} X v_i$ and $\gamma$ is the positive penalty parameter. Since $y$ is characterized by the second PC where $y = u_2 \delta_2 + \epsilon$, we can know that

$$\hat{\delta}^{lse}_1 = u'_1 (u_2 + \epsilon) = u'_1 \epsilon \equiv Z_1 \quad \text{and} \quad \hat{\delta}^{lse}_2 = u'_2 (u_2 + \epsilon) = 1 + u'_1 \epsilon \equiv 1 + Z_2$$

where $Z_1$ and $Z_2$ have $N(0, 1)$ distribution independently, from the fact $\|u_1\|^2 = 1$. Hence, if we can find $\gamma$ such that $|Z_1| \leq \gamma \leq |Z_2 + 1|$, then selecting significant component would be possible and PC regression via LASSO can provide the model selection consistency. In practice, it is difficult for finding such $\gamma$ since both $Z_1$ and $Z_2$ are random variables and if we further assume the general covariance structure with the large $p$ case, pursuing selection consistency can be much more complicated issues. However, in the second example, we can expect that there is the possibility that we are able to find the correct component and quantify that possibility as

$$\mathbb{P}(|Z_1| \leq \gamma \leq |Z_2 + 1|).$$

To move the issue from two simple examples to realistic issue, we consider multiple component model (factor model). Assuming that we have $n$ observation $x_i$ with $p$ predictors, viewed as a $p$ dimensional column vectors, this model is given by

$$x'_i = \mu + \sum_{j=1}^{m} \sqrt{\lambda_j} u_j \theta_j + \sigma_x \epsilon, \quad i = 1, \ldots, n \quad (2.13)$$

Here $\mu$ is the mean function, which is assumed known, and hence is taken to be zero. Each $u_j$ is random effect and each orthonormal basis $\theta_j$ is unknown. We will use $\hat{\theta}_j$ obtained by the sample covariance as estimates for $\theta_j$. We will develop theoretical properties under the factor model frame and discuss details about this in the appendix.
2.2.3 Choice Principle Components by LASSO in Regression

Here, as the main method for choosing PCs, we apply the LASSO suggested by Tibshirani (1996) to PC regression. We regard $u_1, \ldots, u_m$ as predictors and fit the LASSO regression model with the outcome $y$. LASSO estimator is given by minimizing

$$
\hat{\delta}^L_\lambda = \min_{\delta} \frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} u_{ij}\delta_j)^2 \quad s.t. \quad \sum_{i=1}^{m} |\delta_i| \leq t \tag{2.14}
$$

where $m = \text{rank}(X)$.

Equivalently, the LASSO is the solution for minimizing the Lagrange objective function,

$$
f(\delta) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} u_{ij}\delta_j)^2 + \gamma \sum_{i=1}^{m} |\delta_i|,
$$

where $\gamma$ is the penalty parameter ($\gamma \geq 0$) and $\gamma$ has a one-to-one correspondence relation between $\gamma$ and $t$.

If we assume $\gamma = 0$, $\hat{\delta}$ would be simply least square estimator (LSE) and we denote $\hat{\delta}^{lse}$. Since $u_1, \ldots, u_m$ are mutually orthonormal vectors, our LSE is given by

$$
\hat{\delta}^{lse} = U y.
$$

For each predictor, the LASSO solution suggested by Donoho et al. (1995) is

$$
\hat{\delta}_j^{lasso} = S(\hat{\delta}_j^{lse}, \gamma) = \text{sgn}(\hat{\delta}_j^{lse})(|\hat{\delta}_j^{lse}| - \gamma)_+
$$

$$
= \begin{cases} 
\hat{\delta}_j^{lse} - \gamma & \text{if } \hat{\delta}_j^{lse} > 0 \text{ and } \gamma < |\hat{\delta}_j^{lse}| \\
\hat{\delta}_j^{lse} - \gamma & \text{if } \hat{\delta}_j^{lse} < 0 \text{ and } \gamma < |\hat{\delta}_j^{lse}| \\
0 & \text{if } \gamma \geq |\hat{\delta}_j^{lse}|.
\end{cases}
$$

where $S$ is soft-threshold operator for a given $\gamma$ and this solution is known as the soft-threshold
By defining the quantity, \( \tilde{y}^{(j)}_i = \sum_{k \neq j}^m u_{ik} \tilde{\delta}_k(\gamma) \) and the partial residual for the \( j \)-th predictor as \( y_i - \tilde{y}^{(j)}_i \) at each \( j \)-th step, applying the coordinate optimization method suggested by Friedman et al. (2007) with the partial residual as a response variable is feasible. In this framework, the Lagrange objective function at \( j \)-th step is rewritten as

\[
f(\hat{\delta}) = \frac{1}{2} \sum_{i=1}^N \left( y_i - \sum_{k \neq j}^m u_{ik} \tilde{\delta}_k - u_{ij} \delta_j \right)^2 + \gamma \sum_{k \neq j}^m |\tilde{\delta}_i| + \gamma |\delta_i|.\]

For \( j = 1, \ldots, m, 1, 2, \ldots \) until convergence, we update at each \( j \)-step

\[
\hat{\delta}_j^L \leftarrow S \left( \sum_{i=1}^n u_{ij} (y_i - \tilde{y}^{(j)}_i), \gamma \right).
\]

Since our design is the orthogonal design, it would require computation of \( O(m) \) and this would provide the speed up algorithm.

Now, we fit the PC regression model with predictors \( u_1, \ldots, u_m \) and the response \( y \) as the following

\[
\hat{y}^{PC}_\gamma = U \hat{\delta}^L.
\]

This fitted value would be the prediction value by the PCs choice via LASSO. It would be the distinct value from the fitted value by the all PCs, due to \( \hat{y}^{lse} = U \hat{\delta}^{lse}_j \). Hence, we are able to select some principle components among all principle components by \( L_1 \) penalty. Also, we only select a few leading principle components, our estimators in PC regression could be shrinkage estimator.

According to (10), our estimators can be viewed as the ordinal regression model and the PC regression, specifically,

\[
\hat{y}^{PC}_\gamma = U \hat{\delta}^L = XL^{-1} \hat{\delta}^L
\]

35
\[ ZL^{-1}\hat{\delta}^{L} = Z \sum_{i=1}^{m} \frac{1}{l_i} \hat{\delta}^{L} = Z\hat{\gamma} \text{PCL} \]

\[ XW\hat{\delta}^{L} = X \sum_{i=1}^{m} \frac{v_i}{l_i} \hat{\delta}^{L} = X\hat{\beta} \text{PCL} \]

where each \( l_i \) is a singular value of \( X \) and denote superscript "PCL" as the expression that the related coefficients are produced based on PCs choice via LASSO.

### 2.2.4 Choice Principle Components by LASSO in Survival Analysis

We have data \((y_1, n_1, x_1'), \ldots, (y_n, n_n, x_n')\) where \( y_1 \) is observed time and \( n_i \) is censoring status. Let \( t_{(1)} < \ldots < t_{(q)} \) be the unique failure times and \( R_i \) is the collection of indices for observations, that is, all collection of \( j \) such that \( y_j \geq t_i \). Almost notations are similar to Fleming and Harrington (2011)'s notations. Cox proportional hazards model assumes a semi-parametric form for the hazard

\[ h_i(t) = h_0(t)e^{x_i\beta} \]

where \( h_0(t) \) is a baseline hazard and \( \beta \) is a length \( p \) vector. In Cox model, the inference for coefficients is achieved under the frame of the partial likelihood and the partial likelihood is defined as

\[ L(\beta) = \prod_{k=1}^{q} \frac{e^{x'_{(k)}\beta}}{\sum_{j \in R_k} e^{x'_{j}\beta}}. \]

Thus, the partial log likelihood is given by

\[ \log L(\beta) = \sum_{k=1}^{q} \left[ x'_{(k)}\beta - \log \left( \sum_{j \in R_k} e^{x'_{j}\beta} \right) \right]. \]

We can easily derive the following score statistic and information through simple algebra after the first and second derivation of the log partial likelihood with respect to \( \beta \). The score

36
statistics for $\beta$ is

$$U(\beta) = \sum_{k=1}^{q} \left[ x_{(k)} - \frac{\sum_{j \in R_k} x_j e^{x_j^T \beta}}{\sum_{j \in R_k} e^{x_j^T \beta}} \right]$$

and the information for $\beta$ is given by

$$I(\beta) = \sum_{k=1}^{q} \left[ \sum_{j \in R_k} x_j^2 \frac{e^{x_j^T \beta}}{\sum_{j \in R_k} e^{x_j^T \beta}} - \left( \frac{\sum_{j \in R_k} x_j e^{x_j^T \beta}}{\sum_{j \in R_k} e^{x_j^T \beta}} \right)^2 \right].$$

Next, consider the LASSO regularization in the Cox model as Tibshirani et al. (1997) proposed. If we regard $u_1, \ldots, u_m$ as predictors in the Cox model, we want to maximize

$$\hat{\delta}_{\text{lasso}} = \max_{\delta} \sum_{k=1}^{q} \left[ u'_{(k)} \delta - \log \left\{ \sum_{j \in R_k} e^{u_j^T \delta} \right\} \right]. \text{ s.t. } \sum_{i=1}^{m} |\delta_i| \leq t \quad (2.16)$$

Similarly, the LASSO is the solution for maximizing the Lagrange objective function,

$$f(\delta) = \sum_{k=1}^{q} \left[ u'_{(k)} \delta - \log \left\{ \sum_{j \in R_k} e^{u_j^T \delta} \right\} \right] - \gamma \sum_{i=1}^{m} |\delta_i|,$$

where $\gamma$ is the penalty parameter ($\gamma \geq 0$) and $\gamma$ has a one-to-one correspondence relationship between $\gamma$ and $t$.

If we assume $\gamma = 0$, the maximizer, $\hat{\delta}$ would be the partial maximum likelihood estimator (PMLE) and we denote $\hat{\delta}^{\text{pmle}}$. To obtain the maximizer of equation (16), we use the re-weighted least square algorithm strategy. By Taylor expansion, we approximate partial log likelihood at $\hat{\delta}$,
where \( \eta = U\delta \) and \( \hat{\eta} = U\hat{\delta} \). We add and subtract the term \( \frac{1}{2} \frac{\partial l(\hat{\eta})}{\partial \eta} \frac{\partial^2 l(\hat{\eta})}{\partial \eta^2} \) in the last equation. Then, we obtain

\[
\begin{align*}
    l(\delta) & \approx -\frac{1}{2}(\hat{\eta} - \eta)' \frac{\partial^2 l(\hat{\eta})}{\partial \eta^2} (\hat{\eta} - \eta) \\
    & + l(\hat{\delta}) - \frac{1}{2} \frac{\partial l(\hat{\eta})}{\partial \eta} \frac{\partial^2 l(\hat{\eta})}{\partial \eta^2}. 
\end{align*}
\]

(2.17)

The second term in right hand side in equation (17) is not depend on \( \delta \) in optimization problem and by multiplying it by minus one, the maximization problem would be equivalent to the minimization problem. Define the quantity \( y^* = \hat{\eta} - \frac{\partial \eta^2}{\partial^2 l(\hat{\eta})} \) and treat this as the new response outcome. Also, we approximate the off diagonal elements of \( \frac{\partial^2 l(\hat{\eta})}{\partial \eta^2} \) to 0 since the off diagonal entries are small in comparison to the diagonal elements, based on the argument of Hastie and Tibshirani (1990). Denoting the \( i \)-th diagonal element of \( \frac{\partial^2 l(\hat{\eta})}{\partial \eta^2} \) by \( w_i \), we can approximate the original objective function as the \( L_2 \) type objective function. As the result, it is given by

\[
\hat{\delta}_L^\lambda = \min_\delta \frac{1}{2} \sum_{i=1}^{N} w_i (y_i^* - \sum_{j=1}^{m} u_{ij}\delta_j)^2 + \gamma \sum_{i=1}^{m} |\delta_i|. 
\]

For \( j = 1, \ldots, m, 1, 2, \ldots \) until convergence, we update both coefficient and linear predictor at each \( j \)-th step

\[
\begin{align*}
    \hat{\delta}_j^L & \leftarrow S \left( \sum_{i=1}^{m} w_i u_{ij}(y_i^* - \hat{y}_i^{(j)}), \gamma \right) / \sum_{i=1}^{m} w_i u_{ij}^2 + \gamma, \\
    \hat{\eta} & \leftarrow U\hat{\delta}_L. 
\end{align*}
\]

Repeat updating \( \hat{\delta}_L \) and \( \hat{\eta} \) until convergence of \( \hat{\delta}_L \).

Similarly, we can estimate the linear prediction through predictors \( u_1, \ldots, u_m \) and coefficients \( \hat{\delta}_L \). We define this linear predictions as PCs in the Cox model. With the similar argument on
equation (10), we can reconsider the hazard function in the Cox model as the following,

\[ h_i(t) = h_0(t) \exp\{u_i \hat{\delta}_i \} \]
\[ = h_0(t) \exp\left\{ \sum_{i=1}^{m} \frac{v_i}{l_i} \hat{\delta}_i \right\} = h_0(t) \exp\{x_i' \hat{\beta}_{PCL} \} \]
\[ = h_0(t) \exp\left\{ \sum_{i=1}^{m} \frac{1}{l_i} \hat{\delta}_i \right\} = h_0(t) \exp\{z_i' \hat{\gamma}_{PCL} \}. \]

where each \( l_i \) is a singular value of \( X \) and denote superscript "PCL" as the expression that the related coefficients are produced based on PCs choice via LASSO.

### 2.2.5 Tuning Parameter Selection

We have tuning parameter \( \gamma \) in a path of solutions for LASSO. We will try to use cross validation in the regression model and in the Cox proportional hazard model to acquire optimal \( \gamma \). \( k \)-fold cross validation which randomly splits the data in \( k \) pieces, uses \( k-1 \) pieces to build the model, and tests the rest \( k \)th piece to validate the model will be used for the reason that it has low variance. See the Hastie et al. (2001)’s argument (Chapter 7) about cross validation.

Denote the \( j \)-th part of the data removed as \(-j\) and \( K \) as the set of indices for folds \( \{1, \ldots, k\} \). Then cross validation estimate for prediction error in the regression is given by

\[ CV(\gamma) = \frac{1}{k} \sum_{j \in K} \sum_{i \in j} (y_i - x_i' \hat{\beta}_{-j, PCL})^2. \] (2.18)

We choose 10 as the value of \( K \) which is typical choices. For each fold \( j \in K \), the optimal value of \( \gamma \) is required to compute the complete cross validation.

The same partition procedure and 5 folds cross validation method are applied to the Cox model. However, the more sophisticated partition strategy is required in the survival setting than in the regression setting. Each partition indexed by \( j \) should have the similar number of events to the number of events in other partition. In addition, it would be better to use partial log likelihood ratio statistic for accessing the prediction error than squared loss function. Thus, we suggest the
cross validation estimate for prediction in the Cox model which is given by

$$CV(\gamma) = \frac{2}{k} \sum_{j \in K} \sum_{i \in j} \left( l_{-j}((\hat{\beta}_{j,PCL}(\gamma)) - l((\hat{\beta}_{j,PCL}(\gamma))) \right)$$ (2.19)

where $l_{-j}$ is partial log likelihood contributed by the data that $j$th partition was removed and $l$ is partial log likelihood contributed by the full data. Equation (19) is very similar to deviance residuals in survival analysis. This aspect allows for considering that martingale residuals might be used for selecting the optimal value $\gamma$, accessing prediction error.

As counting process notation, let $Y_i(t)$ be predictable 0,1 valued process and $N_i(t)$ be counting process at time $t$. Then, the multiplicative intensity model is given by

$$\lambda\{t|x_i(t)\}dt = e^{x_i(t)\alpha}d\Lambda_0(t)$$

where $\Lambda_0(t) = \int_0^t \lambda_0(t)dt$ is an unspecified baseline cumulative hazard function.

Martingale residual in the multiplicative intensity model is defined as

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)e^{x_i(t)\alpha}d\Lambda_0(s).$$ (2.20)

According to our method, the unspecified baseline hazard $\Lambda_0(t)$ can be estimated by

$$\hat{\Lambda}_0(t) = \int_0^t \left[ \sum_{i=1}^n Y_i(s)e^{x_i(t)\alpha} \right]^{-1} \sum_{i=1}^n dN_i(s)$$ (2.21)

If $\hat{M}_i(\infty) \equiv \hat{M}_i$, then the estimator for the martingale residual is given by

$$\hat{M}_i(t) = N_i(t) - \int_0^\infty Y_i(s)e^{x_i(t)\alpha}d\hat{\Lambda}_0(s).$$
We suggest following another measure for cross validation

\[
CV(\gamma) = \frac{1}{k} \sum_{j \in K} \sum_{i \in j} (N_i - \int_0^t Y_i(s) \exp\{u_i'\hat{\delta}_L\} d\hat{\Lambda}_0^{-j}(s))^2
\]  

(2.22)

where the problem about choosing \( \gamma \) would require the same method to previous argument. This is quite similar to martingale residuals and can be used for selecting the tuning parameters \( \gamma \) in the Cox proportional hazard model.

We have been the attention to the criterion that minimizes the cross validation error with the notion that the value of \( \hat{\gamma} \) allows for determining how many PCs are chosen. However, this criterion does not guarantee the model selection consistency. We may need just a few PCs in real data analysis since the case that many PCs are selected is not informative in a certain data analysis. For example, especially, in micro array data analysis, suppose that we utilize PC regression method for the dimension reduction. In this situation, selecting many numbers of PCs may not be meaningful since not only all PCs are spanned by the column space of all the covariates, but also the final purpose is to find out significant features (some meaningful predictors). If we just require a few PCs to satisfy this practical needed, we empirically suggest to use

\[
\hat{\gamma}^* = \hat{\gamma}_{min} + SE(CV(\gamma))
\]

(2.23)

where \( SE \) is standard deviation.

Equation (23) provides a few PCs since the penalty value is increased. This method seems to be ad-hoc method but it can be valid in the issue for the PCs choice than the usual regression analysis. We will discuss later it in section 4. If we use PC regression in usual data analysis, our issue can be done with the decision for the PCs choice. However, as mentioned above, if we are interested in finding significant feature in micro analysis, one more procedure is remained. We will use the important score to find significant features and briefly introduce it on the next subsection.
2.2.6 Relation with the Eigenarray Space

It is known that the principle components of the columns of $X$ are referred to as the eigengene space, and the principle components of row of $X$ are referred to as the eigenarray space in microarray data analysis, as Witten and Tibshirani (2008) pointed out argument. We suggest the LASSO regression in the eigengene space rather then in eigenarray space. LPC suggested by Witten and Tibshirani (2008) where it is LASSO regression in the eigenarray space is a good counterfactual. If the outcome is quantitative, LASSO in the eigengene space has the advantage since it is intuitive and reasonable to interpret final model.

2.3 Simulation Studies

2.3.1 Simulation 1: Single PC Cases

In this subsection we compare performances for PCs choices by LASSO based on $\hat{\gamma}^*$ and $\hat{\gamma}_{min}$ with the true PC in given models where for all given models, we assume that the single PC is true case. We also compare model assessments for these criterions with those for the conventional PCs choice method. We evaluate each performance by Monte Carlo simulation method. Our simulation studies were conducted using R code.

To obtain the covariate characterized by the pattern of its correlation structure where we are interested in its correlation matrix rather than its covariance matrix, we first generate orthonormal basis $\theta_j \in \mathbb{R}^p$ for its correlation matrix as the following.

$$\theta_{ij} = \begin{cases} 
0.158 & \text{if } i \leq 40, \quad j = 1 \\
0.183 & \text{if } 41 \leq i \leq 70, \quad j = 2 \\
0.224 & \text{if } 71 \leq i \leq 90, \quad j = 3 \\
0.316 & \text{if } 91 \leq i \leq 100, \quad j = 4 \\
0 & \text{if } 101 \leq i \leq p, \quad 1 \leq j \leq 4
\end{cases}$$
These orthonormal basis $\theta_1, \ldots, \theta_4$ are column vectors in $\mathbb{R}^p$ and have two distinct points. One is that nonzero coordinate components in one basis do not overlap with nonzero coordinate components in other basis. Another is that they are sparse in that most coordinate components of one basis are zeros. The pattern of such this 4 orthonormal basis characterizes the particular correlation matrix structure which has the block diagonal structure of 4 groups. By adding a few other orthonormal basis, we can also characterize the correlation structure within one block structure. Actually, we tried to generate additional orthonormal basis. However, we present only the result for 4 representative basis case here since almost results based on additional basis similar to the result we report in this paper.

With $\theta_1, \ldots, \theta_4$ and corresponding eigenvalues $l = (l_1, l_2, l_3, l_4) = (40, 30, 20, 10)$ we generate covariate $X$ as

$$X = \sum_{i=1}^{4} \sqrt{l_i} u_i \theta'_i + E$$

where $u_i$ and $\epsilon_{x,i}$ are both $n$ dimensional random vector from $N(0, I_n)$, $E$ is $n$ by $p$ matrix whose $i$-th column vector is $\epsilon_{x,i}$ and $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (41, 31, 21, 11)$.

Thus, we have covariates whose correlation matrix structure has the block diagonal structure of 4 groups and all the correlations between groups are zero whereas all the correlations within groups are ones.

Next, we generate responses from the following four models.

(1.1) : $y_i = \frac{\sum_{j=1}^{40} x_{ij}}{40} + \epsilon_{y,i}$

(1.2) : $y_i = \frac{\sum_{j=41}^{70} x_{ij}}{30} + \epsilon_{y,i}$

(1.3) : $y_i = \frac{\sum_{j=71}^{90} x_{ij}}{20} + \epsilon_{y,i}$

(1.4) : $y_i = \frac{\sum_{j=91}^{100} x_{ij}}{10} + \epsilon_{y,i}$

where $\epsilon_{y,i} \sim N(0, I_n)$.

Each outcome $y_i$ in each given model is proportional to the sum of each group of covariates.
These models provide the views that each response in each model is characterized by the sample mean of the particular group as the coefficients and Gaussian noise in the view of regression or is characterized by the particular principal component as the signal and noise in the view of PC regression. For example, the response in model (1.2) is generated by the sample mean of the second group as regression coefficients or generated by the second principal component as the signal.

We vary the sample size $n$ to be 80, 150 and 300 for the two cases where $p$ are 150 and 2000, respectively. We denote that $\hat{u}_i$ is the estimator for $u_i$ obtained by SVD, the set, $A$ is indices set of true principal component (or orthogonal basis) and the set, $\hat{A}$ is indices set of selected principal component by LASSO. We repeat each experiment 300 times and evaluate the performance for the selection through the following three criterions.

- $P_{\hat{u}_i \in A}$: the proportion that index $i$ for the estimated orthonormal basis belong to the indices set for true orthonormal basis in a given model in the 300 replications.
- $P_{A \subset \hat{A}}$: the proportion that all indices for selected orthonormal basis contain the true orthonormal basis indices for the given model in the 300 replications.
- $B$: the average cardinality for overestimated indices, that is $|\hat{A} - A|$ in the 300 replications.

To make a clearness, $\hat{u}_i \in A$ is simply the event that $\hat{\theta}_i$ belongs to $A$ since selecting $\hat{u}_i$ corresponded to $\hat{\theta}_i$ estimated by LASSO implies selecting $\hat{\theta}_i$. Thus, we can know that the $P_{\hat{u}_i \in A}$ is used to marginally measure the model consistency for each estimated basis whereas $P_{A \subset \hat{A}}$ is used to jointly measure the model consistency. However, if the size of $\hat{A}$ is getting increased, $P_{A \subset \hat{A}}$ would be also increased. This require another measure. In this sense, $B$ is used to determine whether or not estimated indices set are overestimated or not.

We used LASSO on $\hat{U}$ with following penalties.

- $\hat{\gamma}^*$: The estimated penalty parameter is added one standard deviation of CV error to the value that minimizes CV error.
Table 2.1: Model consistency in example 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>$n$</th>
<th>$p$</th>
<th>$\gamma^*$</th>
<th>$\gamma_{\text{min}}$</th>
<th>$\gamma_{\text{min}}$</th>
<th>$\gamma_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>80</td>
<td>150</td>
<td>0.93</td>
<td>0.65</td>
<td>2.13</td>
<td>1.97</td>
</tr>
<tr>
<td>(1)</td>
<td>150</td>
<td>150</td>
<td>0.93</td>
<td>0.65</td>
<td>2.13</td>
<td>1.97</td>
</tr>
<tr>
<td>(1)</td>
<td>300</td>
<td>150</td>
<td>0.93</td>
<td>0.65</td>
<td>2.13</td>
<td>1.97</td>
</tr>
</tbody>
</table>

Table 2.2: Goodness of fit in example 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>$n$</th>
<th>$p$</th>
<th>$R^2$</th>
<th>$R^2$</th>
<th>$P^2$</th>
<th>$MSE$</th>
<th>$P^2$</th>
<th>$MSE$</th>
<th>$\gamma_{\text{PC}}$</th>
<th>$\gamma_{\text{PC}}$</th>
<th>$\gamma_{\text{SSE}}$</th>
<th>$\gamma_{\text{SSE}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>80</td>
<td>150</td>
<td>0.93</td>
<td>0.65</td>
<td>2.13</td>
<td>0.98</td>
<td>0.47</td>
<td>1.23</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>(1)</td>
<td>150</td>
<td>150</td>
<td>0.93</td>
<td>0.65</td>
<td>2.13</td>
<td>0.98</td>
<td>0.47</td>
<td>1.23</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>(1)</td>
<td>300</td>
<td>150</td>
<td>0.93</td>
<td>0.65</td>
<td>2.13</td>
<td>0.98</td>
<td>0.47</td>
<td>1.23</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>
• $\hat{\gamma}_{\min}$: The estimated penalty parameter is the value that minimizes CV error.

• $\gamma_{\mid A}$: The penalty value is able to select true orthonormal basis.

Table 2.1 depicts results of simulations for $P_{u_{i} \in A}$, $P_{A \subset \hat{A}}$ and $B$. The marginal selection performances ($P_{u_{i} \in A}$) of PC selections with both $\hat{\gamma}^{*}$ and $\hat{\gamma}_{\min}$ show nice results when comparing them with the true selection with $\gamma_{\mid A}$ in models (1.1), (1.2), (1.3) and (1.4). For all model cases, the marginal selection performances for both $\hat{\gamma}^{*}$ and $\hat{\gamma}_{\min}$ converges to 1, as $n$ is increasing on $p=150$ and $p=2000$, respectively. The joint selection performances ($P_{A \subset \hat{A}}$) of PC selections with both values shows the same result since here we are considering single PC case. PC selection by LASSO seems to allow for the correct choice as $n$ or $p$ increase. However, The selection with $\hat{\gamma}_{\min}$ shows relatively large numbers in $B$. Since this implies overestimation in model selection, we conclude that the selection with $\hat{\gamma}^{*}$ presents not only nice results for the marginal and joint selection performances but also the stable selection result that overestimation does not occur.

Table 2.2 indicates results for the residual sum of square, the prediction mean squared error, and mean squared error relevant to true coefficients or principal components in this simulation. These three measures for the selection with both $\hat{\gamma}^{*}$ and $\hat{\gamma}_{\min}$ are compared with the true selection with $\gamma_{\mid A}$ and conventional PC selections with the use of 50% and 75% portion of the total variance. Even if selections with both cases have greater values for the residual sum of square, they have smaller values for the prediction mean squared error and mean squared error (MSE) than conventional selections. The selection with $\hat{\gamma}^{*}$ shows that it has nice performances in terms of model selection criterions but that it has similar or a little inferior performances in terms of the prediction mean squared error and MSE to the conventional selection with the use of 50% of total variance as $n$ is large case. However, as $n$ is smaller case, the selection by $\hat{\gamma}^{*}$ provides better values than the conventional selection. We can reconfirm that the conventional selections with 75% or 100% (LSE) that they are not applicable to our selection procedure show the smallest values for the residual sum of square but these no longer improve the accuracy of the prediction and MSE.
2.3.2 Simulation 2: Multiple PC Cases

In this subsection, we are interested in cases that response is generated by multiple orthonormal basis. We generate new responses from the following four models.

\[(1.5) \quad y_i = \frac{\sum_{j=1}^{70} x_{ij}}{70} + \epsilon_{y,i}\]

\[(1.6) \quad y_i = \frac{\sum_{j=41}^{90} x_{ij}}{50} + \epsilon_{y,i}\]

\[(1.7) \quad y_i = \frac{\sum_{j=71}^{100} x_{ij}}{30} + \epsilon_{y,i}\]

\[(1.8) \quad y_i = \frac{\sum_{j=41}^{70} x_{ij} + \sum_{i=91}^{100} x_{ij}}{60} + \epsilon_{y,i}\]

where \(\epsilon_{y,i} \sim \mathcal{N}(0, I_n)\).

Similar to the previous subsection, each outcome \(y_i\) in each given model is proportional to the sum of several groups of covariates. These models give the views that each response in each model is generated by the sample mean of the pooled groups as the coefficients and noise in the regression view or generated by particular principal components as the signal and noise in the PC regression view. For example, the response in model (1.5) is generated by the sample mean of the pooled groups between the first group and second group as regression coefficients or generated by the first and second principal component as the signal. Similarly, the response in model (1.6) is generated by the sample mean of the pooled groups between the second group and the third group as regression coefficients or generated by the second and third principal component as the signal.

Table 2.3 depicts results of simulations for measures for the model consistency. The marginal selection performances \((P_{u_i \in A})\) of PC selections with both \(\hat{\gamma}^*\) and \(\hat{\gamma}_{\text{min}}\) show nice results when comparing them with the true selection with \(\gamma_{|A|}\) in models (1.5) and (1.6). For (1.5) and (1.6) case, the marginal selection performances for both \(\hat{\gamma}^*\) and \(\hat{\gamma}_{\text{min}}\) converges to 1, as \(n\) is increasing when \(p=150\) and \(p=2000\), respectively. However, in the case of the selection with \(\hat{\gamma}^*\) in the
Table 2.3: Model consistency in example 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>p</th>
<th>$\gamma^*$</th>
<th>$\gamma_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\gamma_{\text{C}}$</td>
<td>$\gamma_{\text{P}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\text{RSE}$</td>
<td>$\text{PSE}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\gamma_{\text{C}}$</td>
<td>$\gamma_{\text{P}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\text{RSE}$</td>
<td>$\text{PSE}$</td>
</tr>
</tbody>
</table>

Table 2.4: Goodness of fit in example 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>p</th>
<th>$\gamma^*$</th>
<th>$\gamma_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\gamma_{\text{C}}$</td>
<td>$\gamma_{\text{P}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\text{RSE}$</td>
<td>$\text{PSE}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\gamma_{\text{C}}$</td>
<td>$\gamma_{\text{P}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\text{RSE}$</td>
<td>$\text{PSE}$</td>
</tr>
</tbody>
</table>

48
model (1.7), the marginal selection performances for both the third and fourth PCs present the nice results when \( n > p \) but the marginal selection performance for the fourth PCs shows the inferior result at \( n = 80 \) when \( n < p \). It is getting better while \( n \) is increasing as 150 and 300 when \( n < p \). This pattern does not happen in the selection with \( \hat{\gamma}_{\text{min}} \) in the model (1.7).

Also, in the case of the selection with \( \hat{\gamma}^* \) in the model (1.8), Table 2.3 shows that the marginal selection performance for the second PC is nice but those performances for the fourth PCs are poor at \( n = 80 \) and \( n = 150 \). Similar results are observed when \( n < p \). The joint selection performances \( (P_{A \subset \hat{A}}) \) of PC selections with both cases shows similar nice performances in models (1.5), (1.6) and (1.7). However, they do not show good performances in the selection with \( \hat{\gamma}^* \) in model (1.8). We expect that this pattern happened since the fourth orthonormal basis is relatively weaker than other other orthonormal basis and let the ratio of the signal to noise be a future research. Compared both cases with the cardinality of indices overestimated, \( B \) for the true selection, it is difficult to accept the selection with \( \hat{\gamma}_{\text{min}} \). Based on this result, we can confirm that the criterion minimizing MSE does not guarantee the nice model selection and the selection with \( \hat{\gamma}^* \) would give a better model selection but it may depend on the magnitude of signal.

Table 2.4 indicates results for measures for model assessments in this simulation. We just points out one distinction since overall results are similar to the results in Table 2.2 case. In the selection with the selection with \( \hat{\gamma}^* \) in (1.7) and (1.8), the increasing extents between MSE when \( p > n \) and MSE when \( p < n \) are relatively larger than the extents between MSE when \( p > n \) and MSE when \( p < n \) in the selection with the use of 50% of portion of the total variance. Similar phenomenon happens for prediction means squared error case.

### 2.4 Real Data Analysis

In order to apply our method to real data analysis, we assume that the approach for selecting groups of genes that may together determine the outcome, survival time, would be better rather than the approach for selecting several genes in a biological sense.

In a biological sense, this is similar to the hypothesis that a gene that truly is associated with
Figure 2.1: LASSO path for microarray datasets: (A) lung cancer data, (B) breast cancer data, (C) acute myeloid leukemia data and (D) DLBCL data.
survival time will be involved in a biological pathway related to the clusters that involved many genes. By identifying the clusters of relevant genes based on $L_1$ penalty to remove irrelevant sources of other clusters and by fitting Cox proportional hazard model on the eigengen space of all clusters of genes, we would essentially find the genes that have been highly correlated with the clusters of relevant gene via other method, for example, the important score. As a result, only some particular clusters would be selected in Cox model and genes with the expression pattern not similar to those clusters would be excluded.

![Figure 2.2](image)

Figure 2.2: Cross validation error for microarray datasets: (A) lung cancer data, (B) breast cancer data, (C) acute myeloid leukemia data and (D) DLBCL data.
In this sense, we focus on finding groups of genes. To illustrate our proposed method, we apply our method to several datasets for predicting survival based on microarray data. First, we examine a lung cancer dataset Bair and Tibshirani (2004). There were 7,129 genes and 86 patients. Second, we consider a breast cancer dataset van’t Veer et al. (2002). There were 4,751 genes and 78 patients. Third, we consider a dataset of acute myeloid leukemia patients Bullinger et al. (2004). It consisted of 6,283 genes and 116 patients. Finally, we consider the DLBCL dataset Rosenwald et al. (2002). There were 7,399 genes, 240 patients.

Figure 2.1 depicts paths for major 10 PCs for 4 datasets, denoted by A for Lung Cancer data, B for Breast Cancer Data, C for Acute Myeloid Leukemia Data and D for DLBCL Data. Path for A in Figure 2.1 shows that the 2nd PC, the 7th PC and the 5th PC dominates other PCs with varying penalty values. This is one of the important finding. In many data analysis related to PC issue, the 1st PC normally dominates other PCs and it becomes major part of the results. Also, in terms of view of traditional PC choice, the lower order of PCs tends to be selected with the higher probability than higher order of PCs since many PC choice methods bases to use PC’s portion of total variance. However, our method based on $L_1$ penalty allows for identifying the 2nd PC as the cluster of relevant genes for survival time. Path for B in Figure 2.1 shows that the 1st PC dominates almost other PCs. We may identify the 1st PC as the cluster of relevant genes for the survival time of the breast cancer. Similarly, path for C in Figure 2.1 presents that the 1st PC and the 3rd PC are determined as the main relevant clusters for the survival time of the Acute Myeloid Leukemia. Lastly, path for D in Figure 2.1 shows that the 3rd PC and the 5th PC dominates other and are identified by relevant clusters for the survival time, which is also interesting finding that the relatively higher order PCs compared to the 1st or 2nd PC are selected in the analysis related PC issues.

To estimate optimal penalty values for each dataset, we partitioned each dataset into training set and test set and used CV error based on the partial likelihood function for each case.

Figure 2.2 depicts the relationship between CV error and penalty values and Table 2.5 shows
optimal penalty values ($\hat{\gamma}_{\text{min}}$) for each dataset. PCs column in Table 2.5 presents the number of the selected PCs for the optimal penalty with the minimum criterion for each data. For A and B, since the number of the selected PCs are informative, these datasets do not need the penalty correction procedure ($\hat{\gamma}^*$). Thus, we can select the 2nd PC (coefficient: $-0.0213$) for the survival time in dataset A and similarly, choose the 1st PC (coefficient: -0.1104), 2nd PC (coefficient: -0.0058) and 5th PC (0.2599) for the survival time in dataset B. However, for C and D, the number of the selected PCs is not informative. We expect that overestimation
occurs. To obtain the smaller number of PCs than current selected PCs, we used the selection with $\hat{\gamma}^*$ and confirmed the number of selected PCs in PCs column in Table 2.5 for these two dataset. Hence, we could select the 1st PC (coefficient:-0.0815), 3rd PC (coefficient:0.0621) and 12th PC (coefficient:-0.0621) for the survival time in data D and similarly, choose the 3rd PC (coefficient:0.0618) and 5th PC (coefficient:-0.0730).

Figure 2.3 presents the estimated baseline survival probability with estimated coefficients in the Cox proportional hazard model for each dataset. Based on these we can estimate the survival probability for each datasets.

Table 2.5: Description of data sets and results ($n$: number of data points, $p$: number of data dimensions, $E$: number of total events, $\hat{\gamma}_{\text{min}}$: minimum penalty in Figure 2.2, $\hat{\gamma}^*$: criterion on 2.23 and PCs: number of selected PCs).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$p$</th>
<th>$n$</th>
<th>$E$</th>
<th>$\hat{\gamma}_{\text{min}}$</th>
<th>$\hat{\gamma}^*$</th>
<th>Selected PCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7129</td>
<td>86</td>
<td>24</td>
<td>17.6</td>
<td>17.6</td>
<td>PC2</td>
</tr>
<tr>
<td>B</td>
<td>4751</td>
<td>78</td>
<td>34</td>
<td>14.8</td>
<td>14.8</td>
<td>PC1, PC2, PC5</td>
</tr>
<tr>
<td>C</td>
<td>6283</td>
<td>116</td>
<td>67</td>
<td>9.7</td>
<td>19.8</td>
<td>PC1, PC3, PC12</td>
</tr>
<tr>
<td>D</td>
<td>7399</td>
<td>240</td>
<td>138</td>
<td>6.3</td>
<td>29.2</td>
<td>PC3, PC5</td>
</tr>
</tbody>
</table>

2.5 Asymptotic Analysis

Asymptotic properties of regularization method in variable selection have been considerably studied and many results in the literature have been developed within four types of efforts: consistency, persistency, selection consistency and oracle property mentioned by Fan and Li (2006) and Zhao and Yu (2006). Consistency means the accuracy of estimated model parameter and it appears in many statistical contexts where it investigates the limiting behavior of estimated model parameter and identify the asymptotic distribution of estimated model parameter. The persistency means the accuracy of the expected loss of the estimated model and it frequently appears in machine learning problem where it investigates the limiting behavior of the risk. Selection consistency means consistency of the selected model. Oracle property means that
for any model selection procedure, when true parameter is decomposed as sparse subset and non-sparse subset, estimator corresponded to sparse subset goes to zero with the probability one and estimator corresponded to non-sparse subset attains an certain information bound mimicking the information of true parameter.

In this section, we will investigate several theoretical properties of our method. We have three purposes, the first one that how our estimator is close to true parameter value as the consistency, the second one that how selected components based on our method are close to true components as the selection consistency and the last one that our selection procedure possesses the oracle property.

**Theorem 2.5.1.** Suppose that (A.1)–(A.2) hold under the given model in the appendix, for each $k \in K$, the estimator $\frac{1}{n} \langle \hat{u}_k, y \rangle$ is a consistent estimator for $\delta_k$.

Theorem 2.5.1 provides the basic consistency concept for $\delta$ in the assumption. Even if this result is very similar to the result of consistency for the ordinal least square estimator, there are a few significant differences in proving it. One difference is the status of the covariate $X$. Traditionally, the covariate $X$ is regarded as the fixed matrix in regression analysis since the conditional expectation of $y$ given $X$ was main interest. However, we regard $X$ as the random matrix rather than the fixed matrix in this asymptotic section. Hence, Theorem 2.5.1 contains such a difference. Also, it is known that in contrast to Anderson (1963)’s the consistency of principal component for fixed $p$, Johnstone and Lu (2009) proved inconsistency of principal components as $p$ and $n$ goes to infinity. Specifically, Johnstone’s condition that $\frac{p}{n} \rightarrow \gamma \in [0, \infty)$ is more general than Anderson’s condition $\frac{p}{n} \rightarrow \gamma \in 0$. However, if it is assumed that the covariance has the spiked model structure in the appendix, we do not lose the consistency of principal components as $p$ and $n$ goes to infinity. Theorem 2.5.1 is derived under this circumstance. In addition, Theorem 2.5.1 helps to show the oracle property as Zou (2006) proved that the oracle property of the adaptive LASSO estimator was derived by using the weighted $L_1$ penalty function where the consistency estimator was used as the weight.
Theorem 2.5.2. Suppose that (A.1)–(A.2) hold under the given model in the appendix. If $\gamma_n = o_p(n)$, LASSO estimator $\hat{\delta}^L$ is consistency estimator for $\delta$.

Theorem 2.5.2 characterizes the consistency of our estimator and addresses the first aim of this section. It indicates that our estimator, $\hat{\delta}^L$, converges to $\delta$ in probability when $\gamma_n = o_p(n)$. Its consistency depends on the rate of $\gamma_n$ and this shows the similar conclusion to the result derived by Knight and Fu (2000) the result. Even if it provides the constrained result for consistency, based on this fact, the simple notion that trying to select principal components by $L_1$ penalty is able to be one of justifiable notions in the issue for the PCs choice. Next, we are naturally curious for the behavior of $\hat{\delta}^L$ in the situation that $\gamma_n$ moves more slowly. Theorem 2.5.3 gives this answer for this question.

Theorem 2.5.3. Suppose that (A.1)–(A.2) hold under the given model in the appendix.

If $\gamma_n / \sqrt{n} \xrightarrow{a.s.} \gamma_0 \geq 0$,

$$\sqrt{n}(\hat{\delta}^L - \delta) \overset{d}{\to} W = \min_w f(w)$$

where

$$f(w) = -2w'C + w'Tw + \gamma_0 \sum_{j=1}^k [w_j \text{sgn}(\delta_j)I(\delta_j \neq 0) + |\delta_j|I(\delta_j = 0)].$$

Theorem 2.5.3 shows the fact that $\hat{\delta}^L$ converges to $\delta$ in distribution when $\gamma_n = o_p(\sqrt{n})$. This result may indicates that our estimator is somewhat inferior to the estimator in Theorem 3.1. Especially, if the dimension of $\delta$ is large, the bias of our estimator for $\gamma_n > 1$ may also be unacceptably large. However, this problem may be weaker in the issue for the choice of PCs than usual regression problem since the dimension of $\delta$ is usually equal or less than the number of data points and it is assumed that the spiked model has finite ranks. For this reason, in practice, we can take some advantage when using the empirical chosen penalty value, where it is estimated as the value added by one standard deviation of the CV error on the usual penalty value on the minimum criterion for CV error.

Theorem 2.5.3 also provides the fundamental concept with Theorem 2.5.1 to show the oracle property. With the slightly different argument for Theorem 2.5.3 and the similar technique in Zou...
(2006)’s article, we can show the oracle property for the adaptive type of our estimator rather than our original estimator. As far as, we have been interested in the first aim in this section that how our estimator is close to true parameter value as consistency. Next, we will investigate the second issue that how selected components based on our method are close to true components.

**Theorem 2.5.4.** Suppose that (A.1)–(A.2) hold under the given model in the appendix. Suppose that the number of nonzero coefficients is $K$ among $m_s$ components. For fixed $m$, if $\gamma_n/n \xrightarrow{a.s.} 0$ and $\gamma_n/n^{1+c} \xrightarrow{a.s.} \infty$ with $0 \leq c < 1$, we have

$$P(\hat{A}_n = A) = 1 - \exp\{-n^c\}$$

where $\hat{A}_n$ is indices for nonzero component coefficients by $\delta$ and $A$ is indices for true nonzero component coefficients in the given model.

Theorem 2.5.4 gives that the probability of our estimator choosing the true components converges to 1 at the exponential rate. We already had the conditions, $\gamma_n = o_p(n)$ for the consistency in Theorem 2.5.2 and $\gamma_n = o_p(\sqrt{n})$ for the asymptotic normality in Theorem 2.5.3. Thus, theorem 2.5.4 allows for the selection consistency, the consistency and the asymptotic normality, simultaneously. Here, we took another advantage of the choice for PCs different from the regression model selection issue. Zhao and Yu (2006) introduced irrepresentable condition and used to show sign consistency in his paper. In that paper, irrepresentable condition is important sufficient condition for showing it. Since our main purpose is to select correct PCs and we use the orthogonal design transformed rather than use original principal scores on PCs, irrepresentable condition was able to be automatically held as the trivial case. Hence, we could obtain this result without irrepresentable condition.

**Theorem 2.5.5.** Suppose that (A.1)–(A.2) hold under the given model in the appendix.

If $\gamma_n/\sqrt{n} \xrightarrow{a.s.} 0$ and $a_n \gamma_n/\sqrt{n} \xrightarrow{a.s.} \infty$ where $a_n$ is a sequence such that $a_n(\hat{\delta} - \delta) = O_p(1)$ then

$$\lim_{n \to \infty} P(\hat{A}_n = A) = 1$$
\[ \sqrt{n}(\hat{\delta}_A - \delta_A) \xrightarrow{d} N(0, \sigma^2_y I_{|A|}) \]

where \( A \) is indices for true nonzero component coefficients in the given model.

Theorem 2.5.5 indicates that the oracle property for the adaptive LASSO type of our estimator is held where such a estimator is derived from \( L_1 \) weighted penalty function as Zou (2006)’s article introduced. This result may be the answer of the last aim and shows the improvement aspect of our selection procedure when the consistency estimator is used as the weight in the weighted \( L_1 \) penalty function. With Theorem 2.5.4 or Theorem 3.5, we could approach on the main issue for selecting components in PC regression in the model selection point of view rather than previous point of view depending on the portion of variance.

As one extension case, we try to apply our theoretical results to supervised principal components analysis Bair et al. (2006). The only difference between principal component regression and supervised principal component (SPC) method is the design matrix. In the case of PC regression, candidate principal components are derived from the covariance matrix of the whole covariate matrix, \( X \) while in the case of SPC, candidate principal components are derived from the covariance matrix of the reduced covariate matrix, \( X_\theta \) where \( \theta \) is some thresholding value able to discriminate whether or not covariates are related to the response \( y \). Bair also mentioned that there is no decisive method for the problem how many supervised principal components should be chosen in his paper. In this sense, we try to apply our theoretical results to the issue for the choice for supervised principal components. Here, we only investigate the asymptotic property of the screening process related to \( X_\theta \). It is possible to extend our results in SPC issue if we can set up the asymptotic property of the screening process since the screening process is the only difference between PCs and SPCs.

**Theorem 2.5.6.** Suppose that (A.1)–(A.3) hold under the given model in the appendix.

Let \( \hat{M} \) denote the set of indices selected by screening scheme, based on score statistic. For any \( b > 2 \),

\[ \mathbb{P}(\mathcal{M}^* \subset \hat{M}) \geq 1 - n^{-c} \]

58
where $c > 1$ and $\mathcal{M}^*$ denotes the set of indices for true model.

Theorem 2.5.6 states that we can restrict our analysis to the set $\hat{\mathcal{M}}$ while using the given screening procedure and the probability of that the given screening procedure identifies the true covariates related to the response converges to 1 at the given rate. With this result, we can extend our results in SPC issue.

2.6 Conclusion

This paper presented the method for selecting PCs in PCs regression and Cox proportional hazard model based on $L_1$ penalty. Our method can be one approach to dimensionality reduction in even large $p$ data. It is a simple idea but practical methodology in real situation due to the computational efficiency. Also, we demonstrated asymptotic properties such as the consistency and the model selection consistency for our selecting method under the conventional latent variable assumption. Although it may not be realistic in practice, it is a reasonable starting point. We showed that the result of simulation study supports the result of the theoretical justification in terms of the model consistency. As real data analysis, we have explored its application to four gene expression studies and been able to obtain meaningful clusters of genes.

We already mentioned that this study can be general framework to investigate supervised principal component analysis in the previous section. As future research topics, it would be possible to extend our method to supervised principal components analysis with systematic methods. Also, we pointed out that the magnitude of signal to the noise affects the selection result. This implies that we may require other novelty choice method for principal components to overcome this limitation. Lastly, it would be worth to study for the selection method in complicated orthonormal basis structures or nonlinear structure.
CHAPTER 3: MPM: MULTIPLE PROJECTION MODEL

3.1 Introduction

In recent years, due to the development of sciences and technologies, recent scientific data has characteristics of increasing in both complexity and size. One tendency of such complexity is the massive amount of available covariates called the ultrahigh dimensionality, which makes it hard to figure out the relationship between a response variables and the collection of the covariates. There are also subchallenges of noise accumulation, computational expediency, statistical accuracy and algorithmic stability due to such complexity. These challenges may need new statistical modeling techniques.

Two different approaches are mainly available in statistical literature for handling with the issue for the high dimensionality. One is the variable selection where there are only a few covariates are truly related to the response. Regularization methods such as the LASSO (Tibshirani 1996), the SCAD (Fan and Li 2001), the adaptive LASSO (Zou 2006) and the Danzig selector (Candes and Tao 2007) have received much attention and the huge number of literature has been devoted to study those methods for the last fifteen years. In spite of the remarkable development of the regularization methods, there are several concerns about applying it to ultra high dimensional problem. First concern is that the computational cost for the large number of covariates is very expensive such as implementing optimization tools. Second is that as the dimensionality increases, the risk of the ideal estimator is also increased since its convergence rate is captured as logarithmic factor $\log p$, which was treated as constant term in most literatures. Lastly, the notion of the irrepresentable condition (Zhao and Yu 2006) or the uniform uncertainty principle (Candes and Tao 2007) may not be held which are essential conditions to achieve
four significant criteria: the consistency, the model selection consistency, the persistency and the oracle property. Consequently, it means that there is no regularization method to guarantee to hold nice properties aforementioned in ultra high-dimensional setting. Fan and Lv (2008) introduced a novel concept called sure screening property where all important variables survival after applying a variable screening procedure with the probability tending to one. If sure screening property is held, variable screening could be a desirable variable selection method. He proposed a simple sure screening method employing marginal correlation. Subsequent studies are done by Fan and Song (2010), Zhao and Li (2012), Fan et al. (2011), Li et al. (2012a), Mai and Zou (2013). As the strong model assumption, the rankings of marginal coefficients of generalized linear and Cox proportional hazard model were suggested by Fan and Song (2010) and Zhao and Li (2012). Fan et al. (2011) extended those to nonparametric circumstance. As the weak model assumption, the rankings of marginal rank statistics and Kolmogorov-Sminov statistic were also proposed by Li et al. (2012a) and Mai and Zou (2013). However, since most of screening methods need a certain model assumption, sure independent property may not be held if the assumption breaks down. We may need an alternative screening method able to detect more general types of dependent relationship between the response and the covariate without the specific model assumption. This aspect encourages us to devise the novel approach as one motivation.

The other is the dimension reduction technique where it has also remarkably improved for the recent years under the assumption that there are only a few linear combinations of the many covariates related to the response. Principal component analysis (Jolliffe 2005) is a classical method that provides the best linear approximation capturing the maximum variability in data and is clearly one of the most popular dimension reduction techniques. In spite of its reputation, there are some limitations for it to be the perfect tool as a dimension reduction technique. Principal component analysis (PCA) may not be free from noise accumulations especially in high dimensional issues since each principal components (PCs) is a linear combination of the collection of the covariates with $p$ dimension. Also, it may not be possible to control important
or interesting features. Many PCA variants have been developed to address these limitations. For example, Zou et al. (2006) proposed the sparse PCA and it is practical as the interpretable dimension reduction tool.

Bair et al. (2006) suggested the supervised PCs derived from the selected covariates whose dependent relationship with the response is strong. However, since these PCs concentrated on the first few leading principal components, as Cox (1968) pointed out, there is no logical reason why the response should not be closely related to the least important principal component. He mentioned that if the covariates $X$ and the response $Y$ have a joint distribution or there is an omitted variable $Z$ which can be an external variable or an internal variable obtained by decomposing the linear combination derived from principal components, employing a few leading principal components is not appropriate.

The aim of this chapter is to develop a supervised dimension reduction in a reproducing kernel Hilbert space (RKHS), called multiple projection model (MPM) to address the ultrahigh dimensionality. MPM is summarized as three stages model with two novel approaches. In the first stage, feature selection will be done by our new screening method on RKHS able to detect an arbitrary functional relationship between the marginal covariate and the response. In this second stage, dimension reduction will be done by our projection able to capture local information. In the third stage, we will transform the condensed information on Euclidean space into the information on RKHS by employing positive definite kernel and predict the outcome by utilizing an appropriate kernel machine for the purpose. Our MPM is devised to specifically improve the prediction accuracy for statistical learning problems accompanying the ultrahigh dimensionality and the complex local correlation structure in the regression, the classification or the functional linear model framework. This paper is outlined as follows. In Section 2, we discuss the general MPM framework and its details. In Section 3, simulation study and real data analysis are conducted to evaluate the improvement of our MPM over other commonly used supervised learning methods. Concluding remarks and discussions are given in the Last Section.
3.2 Description

3.2.1 Model Setup

We introduce multiple projection model as the following. The key idea of MPM is to appropriately project predictor $x_i$ in a high-dimensional space onto a few low-dimensional spaces, while accounting for the relationship with the response $y_i$ and the local structure among the covariates. Let $P_1, P_2, \ldots, P_d$ be unknown projection operators onto low dimensional spaces and $C(Z)$ be the space of bounded continuous function defined on $Z$. Multiple projection model (MPM) is given by

$$y_i = f(P_1(x_i), P_2(x_i), \ldots, P_d(x_i)) + \sigma_y \epsilon_i$$

(3.1)

where random variable $\epsilon_i$ has distribution $N(0, 1)$, $\sigma_y$ is a fixed constant and link function $f \in C(\mathcal{X})$. A primary question of this paper is how to find optimal projection operators to achieve better prediction accuracy. To construct $P_1, \ldots, P_d$, we develop two novel approaches: independent screening and local dimension reduction. Independent screening is described as the global approach to find the common direction that all projection operators should possess while local dimension reduction is described as the local approach to characterize particular structure among the covariates where individual projection operator should independently possess. Without specific saying, we focus on the case that for all the projection operator, $P_d(x_i) \in \mathbb{R}^1$.

3.2.2 Sure independent Screening via Positive Definite Kernel

We develop of the first stage of MPM to find a global property of candidate projection operators by explicitly accounting for any functional relationship between covariate $x_i$ and response $y_i$ under the ultrahigh dimensionality. We borrow the notion for the independent screening technique from variable selection techniques to find a global property. We need a certain measurement being able to measure the dependent relationship in model (3.1). There are various independent screening methods available but it may not sufficiently guarantee to detect an arbitrary dependent
relationship in MPM model frame since they were devised under specific model assumptions. MPM requires finding an arbitrary functional relationship not only linear but also nonlinear between marginal covariate and response. Two principles, One that the novel measurement should be similar to the measurement of existing screening methods and another that it is able to detect both linear and nonlinear relationship are considered. To cope with these principles, the novel independent screening measurement for marginal covariate denoted by \( \beta_j \) is defined as

\[
\beta_j = \sup_{f \in \mathcal{C}(X_j), g \in \mathcal{C}(Y)} (\mathbb{E}_P[f(x)g(y)] - \mathbb{E}_Q[f(x)g(y)])
\]

(3.2)

where \( \mathbb{P} \) is a joint probability measure as \( \mathbb{P}_{X,Y} \) and \( \mathbb{Q} \) is a pair of marginal probability measures as \( \mathbb{P}_X \mathbb{P}_Y \). This parametrization works because of two reasons. Firstly, it turns out that if the positive definite kernel corresponding to the given RKHS is an universal or characteristic kernel, Riesz’s representer in RKHS for the expectation of the product between any bounded continuous functions defined on each sample space, \( f(\cdot)g(\cdot) \) uniquely determines those probability measures \( \mathbb{P} \) and \( \mathbb{Q} \). Secondly, since our interested functional class is relatively unrestricted functional class compared to functional class in the existing screening frameworks, focusing on the unrestricted functional class completely provides model free approach. Due to this injective property on RKHS and unrestricted class, the novel measurement, \( \beta_j \) has an important interpretation that if higher \( \beta_j \) means higher functional dependent relationship and vice versa. In fact, \( \beta_j \) is close to the maximum mean discrepancy (Gretton et al. 2012), Hilbert-Schmidt norm (Gretton et al. 2005) or Energy distance (Sejdinovic et al. 2013) but we define \( \beta_j \) from a different perspective and extend it to address ultrahigh dimensionality as the model free approach.

To estimate (3.2), it requires some theory in RKHS. We briefly introduce fundamental importa

-nt notions (Gretton et al. 2012, Sriperumbudur et al. 2011). Without the loss generality, we skip subscript \( j \) for convenience for a while. Suppose that there is any continuous evaluation functional for each of two Hilbert spaces \( \mathcal{H}_f \) and \( \mathcal{H}_g \). By the existence of bounded evaluation functional for each space, we have a dual RKHS \( \mathcal{H}_f \otimes \mathcal{H}_g \) with the feature map \( \phi(\cdot)\psi(\cdot) : \)
\((x, y) \rightarrow \phi(x)\psi(y) \in \mathcal{F} \otimes \mathcal{G}\) such that \(\langle \phi(x)\psi_j(y), \phi(x')\psi(y') \rangle = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_f} \langle \psi(y), \psi(y') \rangle_{\mathcal{H}_g} = K(x, x')L(y, y')\) where \(K(\cdot, \cdot)\) and \(L(\cdot, \cdot)\) are bounded and positive definite kernels in \(\mathcal{H}_f\) and \(\mathcal{H}_g\), respectively. We extend the notion of feature map to the embedding of a probability measure. If we assume that there exists another bounded linear functional such that it is an injective mapping from \(\mathbb{P}\) as the following,

\[
E_Pfg = \int f(x)g(y)d\mathbb{P} \tag{3.3}
\]

then by the Riesz representation theorem, there exist \(\mu_\mathbb{P}\) such that \(E_Pfg = \langle fg, \mu_\mathbb{P} \rangle\). Also by reproducing property, we can formulate as the following

\[
\mu_\mathbb{P}((\cdot, s), (\cdot, t)) = \langle \mu_\mathbb{P}, \phi(s)\psi(t) \rangle_{\mathcal{H}_f \otimes \mathcal{H}_g} = \langle \mu_\mathbb{P}, K(\cdot, s)L(\cdot, t) \rangle_{\mathcal{H}_f \otimes \mathcal{H}_g} = \int K(x, s)L(y, t)d\mathbb{P}.
\]

If we restrict our interest to the functional space whose norm is bounded by unit norm, our embedding of probability is given by

\[
\beta^2 = \sup_{\|f\|_{\mathcal{H}_f} \leq 1, \|g\|_{\mathcal{H}_g} \leq 1} (\mathbb{E}_P[f(x)g(y)] - \mathbb{E}_Q[f(x)g(y)])^2
\]

\[
= \sup_{\|f\|_{\mathcal{H}_f} \leq 1, \|g\|_{\mathcal{H}_g} \leq 1} \langle \mu_\mathbb{P} - \mu_\mathbb{Q}, ff \rangle_{\mathcal{H}_f \otimes \mathcal{H}_g}^2
\]

\[
= \|\mu_\mathbb{P} - \mu_\mathbb{Q}\|_{\mathcal{H}_f \otimes \mathcal{H}_g}^2. \tag{3.4}
\]

Given an RKHS, we can obtain an empirical estimate for (3.4) as the sum of U-statistics averages

\[
\hat{\beta}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i, x_j)L(y_i, y_j) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(x_i, x_j) \sum_{i=1}^{n} \sum_{j=1}^{n} L(y_i, y_j)
\]

\[
- \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} K(x_i, x_k)L(y_k, y_j). \tag{3.5}
\]
It can be shown that (3.5) is equivalent to the square of Hilbert-Schmidt norm

\[ \hat{\beta}^2 = \frac{1}{n^2} tr(KHLH) \]

where \( H \) is a centering matrix \( I - \frac{1}{n} 11' \).

**Proposition 3.2.1.** Suppose that \( \mathcal{F} \) and \( \mathcal{G} \) be a unit open ball for each RKHS, respectively and that \( K(\cdot, \cdot) \) and \( L(\cdot, \cdot) \) is a a universal kernel. Then, \( \beta = 0 \) if and only if \( \mathbb{P} = \mathbb{Q} \).

The proof of Proposition 3.2.1 is followed from Gretton et al. (2012). Based on Proposition 3.2.1, we take the sub-model containing covariates whose measurement is higher than cut-off value

\[ \hat{M}_{r_n} = \{1 \leq j \leq p : |\hat{\beta}_j| \geq \gamma_n\}. \]

where \( \gamma_n \) is a predefined cut-off value. Equivalently, we can employ the ranking of marginal measurement for the predefined sub-model size \( \gamma_n \)

\[ \hat{M}_{r_n} = \{1 \leq j \leq p : |\hat{\beta}_j| \text{ is among the first } \gamma_n \text{ largest of all}\}. \quad (3.6) \]

In section 3, we will show that the sure independence screening property holds where it is defined to be the event that the true set \( \mathcal{M} \) belongs to the estimated indices set \( \hat{M}_{r_n} \) with high probability for an appropriate selection of \( \gamma_n \) under certain conditions.

### 3.2.3 Local Dimension Reduction

When the sparsity condition is held, by setting the size of submodel (3.6) as \( \gamma_n < n \), projection operator \( P(x_i) = x_i \) as the identity operator and the total number of projection as \( d = \gamma_n \), the stage of the local dimension reduction can be skipped and jumped into the next stage to estimate the link function \( f(\cdot) \) in model (3.1). However, there may need a further dimensional reduction for the reason that the dimensionality for significant features is still large \( n < \gamma_n \ll p \) in a practical problem. As Fan and Lv (2008) pointed out, it can chosen as \( s > n \) to contain
significant variables with large probability in conservative circumstance. Extracting PCs from the sample covariance matrix of the covariates contained in the submodel (3.6) built on our feature selection will be primarily done to construct projection operators. As aforementioned in Section 3.1, the noise accumulation of PCs followed from the size of submodel \((n < \gamma_n)\) may happen and employing the first few leading PCs may ignore significant covariance information corresponding to insignificant eigenvalues of the sample covariance when extracting usual PCs. To address these issues, we will use the covariance thresholding introduced by Bickel and Levina (2008) and the spectral clustering method. Let \(s_h\) be thresholding value and \(T_{s_h}\) be thresholding operator. Thereby, we have \(T_{s_h}(\hat{\Sigma}_n) = \{ \hat{\sigma}_{ij} I(|\hat{\sigma}_{ij}| \geq s_h) \}\) and it is expected that the noise resulted from the high dimensionality is removed. Let \(\hat{\pi}\) be spectral clustering function such that \(j \in \hat{M}_{r_n} \mapsto k \in \hat{C}_\pi = \{1, 2, \ldots, C\}\) for \(T_{s_h}(\hat{\Sigma}_n)\). Then, the global property on the submodel is decomposed into the local property on the sub-submodel

\[
\hat{M}_{r_n} = \bigcup_{k \in \hat{C}_{s_h}} \hat{L}^s_{r_n,k} \quad \text{where} \quad \hat{L}^s_{r_n,k} = \{1 \leq j \leq p : j \in \hat{M}_{r_n} \cap \hat{\pi}^{-1}(k)\}.
\]

According to the result of the spectral clustering algorithm, it is believed that the first sub-submodel, \(\hat{L}^s_{r_n,1}\) includes the covariates corresponding to the largest eigenvalue of \(T_{s_h}(\hat{\Sigma}_n)\) and the \(k\)-th sub-submodel, \(\hat{L}^s_{r_n,k}\) is vice versa. If PCs are derived from the \(k\)-th submatrix on \(\hat{L}^s_{r_n,k}\) or SVD for \(X_{j \in \hat{L}^s_{r_n,k}} = \sum_{m=1}^{\lambda} v_{m,r,s_h,k} u_{m,r,s_h,k}^t\), we can characterize the first PC, \(v_{1,r,s_h,k}\) as \(k\)-th local characteristic at the level of \(s_h\), removing accumulated noise and capturing the local covariance information corresponding to the least important eigenvalues.

As a result, we construct projection operator based on the first leading PC and its projection is given by \(P_{s_h,k}(x_i) = P_{s_h,k}(x_{i,j \in \hat{L}}) = \langle v_{h,k}, x_{i,j \in \hat{L}} \rangle\) for all \(k\). This single projection approach can be easily extended to multiple projections by defining thresholding sequence \(\{s_h\}_{h=1}^{H}\), and \(T_{s_1}(\Sigma_{\gamma_n}), \ldots, T_{s_H}(\Sigma_{\gamma_n})\). If we apply the same procedure for each thresholded sample matrices, then we have \(\hat{M}_{r_n} = \bigcup_{k \in \hat{C}_1} \hat{L}^{s_1}_{r_n,k} = \cdots = \bigcup_{k \in \hat{C}_H} \hat{L}^{s_H}_{r_n,k}\). Finally, a series of \(P_{s_1,1}, \ldots P_{s_1,n_1}, \ldots, P_{s_h,1}, \ldots P_{s_h,n_h}\) are obtained where \(n_1 = |\hat{C}_1|, \ldots, n_h = |\hat{C}_h|\) and \(d = \sum_{h=1}^{H} |\hat{C}_h|\).
3.2.4 Kernel Machine Learning

Let denote $Z$ as the input space and $z_i = (P_1(x_i), P_2(x_i), \ldots, P_d(x_i))$ as $i$-th observation. By utilizing positive definite kernel, we can have the best performance on both linear and nonlinear case. To construct RKHS, mathematically, we need a required feature map from the input space into the some inner product space called feature space, specifically

$$\phi : z \in Z \rightarrow \phi(z) \in \mathcal{F}.$$ 

If it is given, we can obtain Hilbert space whose dense set is isomorphic to the feature space with the inner product that is the form of the limit for the inner product between two Cauchy sequences from the completion procedure. Detail proof is shown by Berlinet and Thomas-Agnan (2004). From such a completion, we have a Hilbert space $\mathcal{H}$ with the inner product given by

$$(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}) = (\mathcal{F}^*, \lim_{n \to \infty} \langle \cdot, \cdot \rangle_{\mathcal{F}^*})$$

where $\mathcal{F}^*$ whose all the elements are Cauchy sequence is a dense set of $\mathcal{H}$. If $\forall f \in \mathcal{F}$ is represented as constant sequence by $f = (f, f, f, \ldots)$ in $\mathcal{F}^*$, Feature space $\mathcal{F}$ and dense set $\mathcal{F}^*$ have isomorphic relation. Then, by defining evaluation functional on the space for all input points called reproducing property,

$$e_z : \mathcal{H} \rightarrow k(z, \cdot) \in \mathcal{H}, \quad \text{such that } e_z(f) = f(z)$$

there is a RKHS, $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ distinguished from functional $L_2$ space.

By Aronszajn (1950)'s theorem, for every positive definite kernel on input space, there exists a unique RKHS and vice versa. Also, since $\langle f, k(z_i, \cdot) \rangle_{\mathcal{H}} = 0$ implies $f = 0$ from the reproducing property, the set $C = \{k(z_1, \cdot), k(z_2, \cdot), \ldots, k(z_n, \cdot)\}$ can be complete orthogonal system (CONS). Thus, we have

$$\mathcal{F}^* = \{ f | f = \sum_{i=1}^{n} \alpha_i k(z_i, \cdot), \forall \alpha_i \in \mathbb{R} \}.$$ 

As a result, $\alpha$ is the only parameter to be estimated for the link function $f(\cdot)$ and $\hat{\alpha}$ is obtained
by the solution of minimization or maximization achieved at some function in RKHS as the linear approach since we have CONS for the dense set. Positive definite kernel is sometimes called kernel and its sample version is called Gram matrix. Substituting infinite dimensional inner product on RKHS by the kernel value is called kernel trick in this area. Such a mathematical background provides the justification on employing a kernel or Gram matrix for figuring out the complex nonlinear relationship and employing a kernel trick for allowing the computational price to be cheaper.

3.3 Simulation Studies

3.3.1 Simulation 1: Prediction in Regression Problem

In the first scenario, we generate the regression response from one model involving four projection predictors, $U_1, U_2, U_3, U_4$. Four scaled B spine basis function with the order 1 having knots $(50, 100, 150, 200)$ is used for the basis to be $\theta_1, \theta_2, \theta_3, \theta_4$. The dimension of the covariate $x_i$ is set to be $p = 2500$. The first 200 coordinates are set up to be nonzero significant feature coordinates. For each observation, $z_{i1}, z_{i2}, z_{i3}, z_{i4}$ have distribution $N(1, 1)$. Coefficient functions for basis are given by $f_1(z_{i1}) = tan^{-1}(z_{i1})cos(tan^{-1}(z_{i1})), f_2(z_{i1}) = tan^{-1}(z_{i2})sin(tan^{-1}(z_{i2})), f_3(z_{i3}) = z_{i3}, f_4(z_{i4}) = 3z_{i4}^3 + 2z_{i4}^2 - 2z_{i4} + 1$. Based on these 4 domains, we generate covariate $x_i$ as

$$x_i = \sum_{m=1}^{4} \theta_m f_m(z_{im}) + \sigma_x E_i$$

where $E_i$ is $2500 \times 1$ a random vector from $N(0, I_{2500})$ and $\sigma_x = 0.1$. The regression response is generated as the following

$$y_i = -2 * sin(z_{i1}) + 3 * cos(z_{i2}) + 4 * cos(z_{i3}) + z_{4i} + \sigma_y \epsilon_i$$
where $\epsilon_i$ is from $N(0, 1)$ and $\sigma_y = 0.2$. We fix the sample size $n$ to be 60 for training set and to be 40 for test set. Each experiment is repeated for 100 times. We evaluate the performance for the sure screening property and the prediction accuracy in MMP with the kernel machine.

- $P_{M \subset \hat{M}}$: the proportion for how many true nonzero significant covariates can be included in the estimated indices (submodel) set by each independent screening with the size of $|\hat{M}|$ in a given model in the 100 replications.

- $P_{M \subset \hat{M}}$: the proportion for how many of the estimated indices (submodel) set by each independent screening with the size of $|\hat{M}|$ can be true nonzero significant covariates in the 100 replications.

If $|\hat{M}|$ is a large number, it may include almost of true nonzero features. $P_{M \subset \hat{M}}$ measures this phenomenon. However, it can not guarantee all estimated entries in submodel are true nonzero significant features for that case. It may include many of zero features. $P_{M \subset \hat{M}}$ measures this phenomenon.

Table 3.1 compares the independent screening with the order 2 of polynomial kernel and Gaussian kernel over existing screening methods based on Spearman, Kendall and Pearson correlations. The number of top features as 100, 200 and 300 is used for the number of selected covariates and for each case, estimated $P_{M \subset \hat{M}}$ and $P_{M \subset \hat{M}}$ are provided with portions close to one demonstrating a good estimate. All the results are averaged over 100 simulation runs. As we expect, the best result are obtained when the kernel screening is used even in the small sample size $n = 60$.

Table 3.2 reports the sum of prediction error for MPM model fitted by the different kernel machines denoted in each row compared to the usual model fitted by the kernel machine and LASSO model, based on the number of $|\hat{M}|$ denoted in each column as test sample size is $n = 40$. Seven projections are selected for MPM for predefined thresholding sequence 0, 0.1, 0.2, 0.3 on the correlation matrix where those are from PCs derived by four submatrices included in estimated covariance matrix at the level 0, the PCs derived by the first and second submatrix.
Table 3.1: Model consistency in regression.

<table>
<thead>
<tr>
<th>Method</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>70%</th>
<th>95%</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Oracle</strong></td>
<td>10.95</td>
<td>50.75</td>
<td>100.5</td>
<td>150.25</td>
<td>190.05</td>
<td>1.000</td>
<td>1.000</td>
<td>200</td>
</tr>
<tr>
<td>$\hat{\beta}_P$</td>
<td>7.69</td>
<td>33.80</td>
<td>61.25</td>
<td>94.45</td>
<td>165.45</td>
<td>0.500</td>
<td>1.000</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{\beta}$</td>
<td>90.32</td>
<td>115.00</td>
<td>150.97</td>
<td>176.05</td>
<td>295.05</td>
<td>0.496</td>
<td>0.992</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{\delta}$</td>
<td>47.49</td>
<td>151.25</td>
<td>406.05</td>
<td>1075.37</td>
<td>2198.41</td>
<td>0.258</td>
<td>0.517</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{\tau}$</td>
<td>204.10</td>
<td>486.07</td>
<td>1099.52</td>
<td>1781.55</td>
<td>2347.66</td>
<td>0.077</td>
<td>0.155</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{\rho}$</td>
<td>136.31</td>
<td>645.22</td>
<td>1222.65</td>
<td>1868.02</td>
<td>2345.44</td>
<td>0.039</td>
<td>0.079</td>
<td>100</td>
</tr>
<tr>
<td>$\hat{\beta}_P$</td>
<td>10.95</td>
<td>50.90</td>
<td>101.12</td>
<td>155.65</td>
<td>202.40</td>
<td>0.968</td>
<td>0.968</td>
<td>200</td>
</tr>
<tr>
<td>$\hat{\beta}$</td>
<td>14.44</td>
<td>72.01</td>
<td>133.75</td>
<td>225.68</td>
<td>1427.66</td>
<td>0.833</td>
<td>0.833</td>
<td>200</td>
</tr>
<tr>
<td>$\hat{\delta}$</td>
<td>33.67</td>
<td>121.80</td>
<td>430.57</td>
<td>1348.52</td>
<td>2268.74</td>
<td>0.452</td>
<td>0.452</td>
<td>200</td>
</tr>
<tr>
<td>$\hat{\tau}$</td>
<td>173.29</td>
<td>483.73</td>
<td>1146.70</td>
<td>1799.86</td>
<td>2358.64</td>
<td>0.137</td>
<td>0.137</td>
<td>200</td>
</tr>
<tr>
<td>$\hat{\rho}$</td>
<td>129.18</td>
<td>136.31</td>
<td>612.32</td>
<td>1857.13</td>
<td>2350.76</td>
<td>0.081</td>
<td>0.081</td>
<td>200</td>
</tr>
<tr>
<td>$\hat{\beta}_P$</td>
<td>42.19</td>
<td>89.01</td>
<td>141.00</td>
<td>272.11</td>
<td>1650.01</td>
<td>0.997</td>
<td>0.794</td>
<td>300</td>
</tr>
<tr>
<td>$\hat{\beta}$</td>
<td>20.74</td>
<td>82.03</td>
<td>157.25</td>
<td>751.23</td>
<td>2017.26</td>
<td>0.966</td>
<td>0.644</td>
<td>300</td>
</tr>
<tr>
<td>$\hat{\delta}$</td>
<td>41.19</td>
<td>88.06</td>
<td>140.15</td>
<td>268.90</td>
<td>1578.12</td>
<td>0.801</td>
<td>0.658</td>
<td>300</td>
</tr>
<tr>
<td>$\hat{\tau}$</td>
<td>16.09</td>
<td>76.20</td>
<td>151.05</td>
<td>769.92</td>
<td>2132.85</td>
<td>0.794</td>
<td>0.664</td>
<td>300</td>
</tr>
<tr>
<td>$\hat{\rho}$</td>
<td>16.09</td>
<td>76.25</td>
<td>151.10</td>
<td>754.96</td>
<td>2123.88</td>
<td>0.714</td>
<td>0.664</td>
<td>300</td>
</tr>
</tbody>
</table>

The first 200 hundred of covariates are true nonzero features. The first 5 columns denote quantiles for indices for true nonzero features. $\hat{\beta}_P$: Screening with polynomial kernel, $\hat{\beta}$: Screening with Gaussian kernel, $\hat{\delta}$: Spearman correlation, $\hat{\tau}$: Kendall correlation and $\hat{\rho}$: Pearson correlation (SIS).

Table 3.2: Sum of prediction error in regression.

<table>
<thead>
<tr>
<th>Method</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MPM + KREG</strong></td>
<td>37.20</td>
<td>16.74</td>
<td>17.86</td>
</tr>
<tr>
<td><strong>MPM + KPCA</strong></td>
<td>21.86</td>
<td>15.51</td>
<td>17.33</td>
</tr>
<tr>
<td><strong>MPM + KSPCA</strong></td>
<td>21.89</td>
<td>15.41</td>
<td>17.31</td>
</tr>
<tr>
<td><strong>SIS + KPC</strong></td>
<td>90.46</td>
<td>90.46</td>
<td>43.69</td>
</tr>
<tr>
<td><strong>SIS + KSPC</strong></td>
<td>89.25</td>
<td>89.24</td>
<td>90.46</td>
</tr>
<tr>
<td><strong>SIS + LASSO</strong></td>
<td>75.21</td>
<td>74.21</td>
<td>72.86</td>
</tr>
</tbody>
</table>

Each column presents the size of $|\hat{M}|$ used for the prediction model. Each row indicates that MPM submodel fitted by the kernel regression machine, MPM submodel fitted by the kernel principal component machine, MPM submodel fitted by the kernel supervised principal component, SIS submodel fitted by the kernel principal component machine, SIS submodel fitted by the kernel supervised principal component machine, and SIS submodel fitted by the LASSO method for submodel identified by SIS.
Table 3.3: Model consistency in classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>5%</th>
<th>25%</th>
<th>50%</th>
<th>70%</th>
<th>95%</th>
<th>(P_{\hat{\mathcal{M}} \subseteq \mathcal{M}})</th>
<th>(P_{\mathcal{M} \subseteq \hat{\mathcal{M}}})</th>
<th>(\hat{\mathcal{M}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>10.95</td>
<td>50.75</td>
<td>100.5</td>
<td>150.25</td>
<td>190.05</td>
<td>1.000</td>
<td>1.000</td>
<td>200</td>
</tr>
<tr>
<td>(\hat{\beta})</td>
<td>68.66</td>
<td>104.25</td>
<td>136.55</td>
<td>164.25</td>
<td>190.53</td>
<td>0.499</td>
<td>0.999</td>
<td>100</td>
</tr>
<tr>
<td>(\hat{\kappa})</td>
<td>71.57</td>
<td>105.05</td>
<td>138.47</td>
<td>166.00</td>
<td>191.47</td>
<td>0.497</td>
<td>0.995</td>
<td>100</td>
</tr>
<tr>
<td>(\hat{\tau})</td>
<td>93.22</td>
<td>119.50</td>
<td>145.50</td>
<td>170.80</td>
<td>194.51</td>
<td>0.497</td>
<td>0.995</td>
<td>100</td>
</tr>
<tr>
<td>(\hat{t})</td>
<td>91.65</td>
<td>118.75</td>
<td>145.27</td>
<td>171.16</td>
<td>194.01</td>
<td>0.499</td>
<td>0.998</td>
<td>100</td>
</tr>
<tr>
<td>(\hat{\rho})</td>
<td>70.77</td>
<td>97.4</td>
<td>125.60</td>
<td>151.32</td>
<td>176.00</td>
<td>0.495</td>
<td>0.990</td>
<td>100</td>
</tr>
<tr>
<td>(\hat{\beta})</td>
<td>12.24</td>
<td>56.27</td>
<td>109.70</td>
<td>159.75</td>
<td>190.53</td>
<td>0.951</td>
<td>0.951</td>
<td>200</td>
</tr>
<tr>
<td>(\hat{\kappa})</td>
<td>20.41</td>
<td>64.93</td>
<td>121.45</td>
<td>171.40</td>
<td>194.51</td>
<td>0.892</td>
<td>0.892</td>
<td>200</td>
</tr>
<tr>
<td>(\hat{\tau})</td>
<td>41.11</td>
<td>88.06</td>
<td>140.15</td>
<td>268.90</td>
<td>175.12</td>
<td>0.801</td>
<td>0.801</td>
<td>200</td>
</tr>
<tr>
<td>(\hat{t})</td>
<td>42.19</td>
<td>89.01</td>
<td>141.00</td>
<td>272.11</td>
<td>165.01</td>
<td>0.794</td>
<td>0.794</td>
<td>200</td>
</tr>
<tr>
<td>(\hat{\rho})</td>
<td>36.24</td>
<td>99.60</td>
<td>179.07</td>
<td>1023.12</td>
<td>2160.41</td>
<td>0.794</td>
<td>0.794</td>
<td>300</td>
</tr>
</tbody>
</table>

The first 5 columns denote quantiles for indices for true nonzero features. \(\hat{\beta}\): Screening with Gaussian kernel, \(\hat{\kappa}\): Kolmogrov-Sminov statistic, \(\hat{\tau}\): Kendall statistic, \(\hat{t}\): t statistic, and \(\hat{\rho}\): Pearson correlation (SIS).

Table 3.4: Average of test error in classification.

| Method  | \(|\hat{\mathcal{M}}|\) | KNN | LDA | REG | SVM | KLDA | KREG | KSVM |
|---------|----------------|-----|-----|-----|-----|------|------|------|
| Oracle  | 200 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| MPM     | 100 | 0.151 | 0.116 | 0.112 | 0.112 | 0.106 | 0.340 | 0.103 | 0.116 |
| MPM     | 200 | 0.113 | 0.106 | 0.105 | 0.105 | 0.101 | 0.218 | 0.101 | 0.103 |
| MPM     | 300 | 0.142 | 0.178 | 0.217 | 0.126 | 0.133 | 0.282 | 0.136 | 0.126 |
| KPC     | 100 | 0.278 | 0.430 | 0.500 | 0.496 | 0.288 | 0.202 | 0.287 | 0.371 |
| KPC     | 200 | 0.370 | 0.487 | 0.500 | 0.500 | 0.408 | 0.353 | 0.408 | 0.486 |
| KPC     | 300 | 0.347 | 0.500 | 0.498 | 0.500 | 0.387 | 0.342 | 0.386 | 0.500 |
| PCA     | 100 | 0.230 | 0.202 | 0.275 | 0.196 | 0.210 | 0.238 | 0.210 | 0.190 |
| PCA     | 200 | 0.320 | 0.310 | 0.346 | 0.298 | 0.291 | 0.295 | 0.293 | 0.290 |
| PCA     | 300 | 0.341 | 0.337 | 0.388 | 0.333 | 0.311 | 0.322 | 0.310 | 0.300 |
| SPC     | 100 | 0.332 | 0.282 | 0.347 | 0.278 | 0.287 | 0.288 | 0.286 | 0.285 |
| SPC     | 200 | 0.255 | 0.223 | 0.315 | 0.240 | 0.208 | 0.202 | 0.210 | 0.193 |
| SPC     | 300 | 0.332 | 0.362 | 0.397 | 0.368 | 0.308 | 0.321 | 0.312 | 0.298 |

Method column presents MPM: Multiple projection, KPCA: projection by the kernel principal component, PCA: projection by the principal component, SPCA: projection by the sparse principal component, and \(|\hat{\mathcal{M}}|\) column reports the number of features used for the prediction model.
Table 3.5: Model consistency in functional curve.

| Method | 5%  | 25%  | 50%  | 70%  | 95%  | $P_{\tilde{M} \subseteq M}$ | $P_{\tilde{M} \subseteq \hat{M}}$ | $|\mathcal{M}|$ |
|--------|-----|-----|-----|-----|-----|-----------------|----------------|--------|
| Oracle | 10.95 | 50.75 | 100.5 | 150.25 | 190.05 | 1.000 | 1.000 | 200 |
| $\beta$ | 6.05 | 25.85 | 50.6 | 75.35 | 95.15 | 0.500 | 1.000 | 100 |
| $\beta_P$ | 6.30 | 26.10 | 50.85 | 75.60 | 95.40 | 0.500 | 1.000 | 100 |
| $\delta$ | 6.15 | 25.95 | 50.70 | 75.45 | 95.25 | 0.500 | 1.000 | 100 |
| $\tau$ | 26.76 | 100.73 | 196.15 | 291.77 | 372.72 | 0.255 | 0.510 | 100 |
| $\beta$ | 10.95 | 50.75 | 100.50 | 156.1 | 211.01 | 0.966 | 0.966 | 200 |
| $\beta_P$ | 10.95 | 50.75 | 100.50 | 150.25 | 194.05 | 0.986 | 0.986 | 200 |
| $\delta$ | 10.95 | 50.75 | 100.50 | 151.48 | 233.00 | 0.946 | 0.946 | 200 |
| $\tau$ | 22.44 | 101.3 | 201.12 | 303.67 | 377.22 | 0.496 | 0.496 | 200 |
| $\beta$ | 15.95 | 75.75 | 150.50 | 231.95 | 315.35 | 0.990 | 0.660 | 300 |
| $\beta_P$ | 15.95 | 75.75 | 150.50 | 228.50 | 297.55 | 0.995 | 0.663 | 300 |
| $\delta$ | 15.95 | 75.75 | 150.50 | 245.85 | 341.65 | 0.986 | 0.658 | 300 |
| $\tau$ | 21.23 | 101.09 | 203.95 | 302.09 | 379.41 | 0.739 | 0.493 | 300 |

The first 200 hundred of covariates are true nonzero features. The first 5 columns present the denoted quantile for true nonzero features. $\hat{\beta}$: global direction with Gaussian kernel, $\hat{\beta}_P$: global direction with polynomial kernel, $\hat{\delta}$: Spearman correlation, and $\hat{\tau}$: Kendall correlation.

Table 3.6: Average of prediction error in functional curve.

| Method | $|\mathcal{M}|$ | FREG | KREG | KPC | KSPC |
|--------|----------------|------|------|-----|------|
| MPM    | 100            | 0.091| 0.058| 0.070| 0.060|
| MPM    | 200            | 0.098| 0.055| 0.062| 0.055|
| MPM    | 300            | 0.071| 0.082| 0.095| 0.084|
| Method | $|\mathcal{M}|$ | FREG | Lasso | Danzig$_1$ | Danzig$_2$ |
|--------|----------------|------|-------|-------------|-------------|
| FPCA   | 500            | 0.938| 0.210 | 0.372       | 0.203       |

FREG: functional linear regression model, KREG: kernel regression, KPC: kernel principal component, KSPC: kernel supervised principal component, FPCA: functional principal component Lasso: Lasso on the constraint for basis function, Danzig$_1$: Danzig selector on the constraint for the first derivative of basis function, and Danzig$_2$: Danzig selector on the constraint for the second derivative of basis function.
included in estimated covariance matrix at the level 0.1, PCs obtained by the third and the fourth submatrices contained in the covariance matrix at the level 0.3 and the rest 4 PCs derived by each submatrix at the level 0.3. Overall, MPM fitted by the kernel machine greatly improves the prediction accuracy than usual model fitted by the kernel machine and the LASSO model. In the reality, the number of the true nonzero feature is unknown so that there may be either overestimated or underestimated in the estimation of the independent screening. Our MPM also shows the robustness for both two cases in the prediction accuracy.

3.3.2 Simulation 2: Prediction in Classification Problem

In the second scenario, we generate the same number of the classification response \{-1, 1\} from two different multivariate normal distributions. \(z_{i1} | y_i = 1, z_{i2} | y_i = 1, z_{i3} | y_i = 1, z_{i4} | y_i = 1\) are distributed from \(N(1/2, 1/2)\) while \(z_{i1} | y_i = -1, z_{i2} | y_i = -1, z_{i3} | y_i = -1, z_{i4} | y_i = -1\) are distributed from \(N(-1/2, 1/2)\). The projections, basis and covariate \(x_{i} | y_i\) are generated in the same fashion to the previous simulation.

Table 3.3 compares the global direction with Gaussian kernel over existing methods Kolmogorov-Smirnov, Kendall, \(t\) and Pearson filters that are popular filter in the classification. Similarly, \(|\hat{M}|\) was chosen as the "top" number of 100, 200 and 300 for each filter. According to \(P_{M|\hat{M}}\) and \(P_{M_=\hat{M}}\), we can observe an underestimation for \(|\hat{M}| = 100\) and an overestimation for \(|\hat{M}| = 300\). However, for all cases, as expected, the best results are obtained as the global direction case on the small sample size \(n = 60\) in the classification problem. Seven projections are selected as MPM with the predefined thresholding sequence \(0, 0.1, 0.2, 0.3\) on the correlation matrix and those projections are from the PCs derived by the same block diagonal structure mentioned in Simulation 1.

Table 3.4 reports numerical summaries for the average of test error comparing MPM to existing projections by the kernel principal component, the principal component and the sparse principal component without the kernel machine denoted by the first four columns and with the kernel machine denoted by the last four columns. Test sample \(n = 40\) with the equal number for
each class was used to evaluate the prediction accuracy. Again, MPM with the kernel machine and without the kernel machine are superior to projections by KPCA, PCA and SPCA with the kernel machine and without kernel machine. For the underestimated case $|\hat{M}| = 100$, PCA gives a slightly better test error estimate but MPM are substantially superior to almost competing methods. For the overestimated case $|\hat{M}| = 300$, SPCA gives a slightly better test error estimate but MPM also outperforms in almost competing methods. Overall, MPM fitted by the kernel machine with the linear discriminant on $L_2$ penalty, support vector machine and regression greatly improve the prediction accuracy over all competing methods.

3.3.3 Simulation 3: Prediction in Functional Curve Problem

In the third scenario, we generate the functional scalar response from one model involving four projection predictors, $U_1, U_2, U_3$. After generating seven B spline basis function with the order 4 having knots $(100, 200, 300)$, we used the first 5 B spline functions are used for the basis to be $\theta_1, \theta_2, \theta_3, \theta_4$ and $\theta_5$. The observed point of the curve $X_i(t)$ is set to be 400 points equally spaced on $1 \leq t \leq 400$. We extend our MMP to multiple single index model as the following,

$$g(y_i) = f\left(\int_{R_1} X_i(t)\beta(t)dt, \int_{R_2} X_i(t)\beta(t)dt, \int_{R_3} X_i(t)\beta(t)dt \right) + \epsilon_i, x(t).$$

The observed values of teh predictors $X_i(t)$ were generated by

$$X_i(t) = \sum_{m=1}^{5} \sqrt{\lambda_i} z_{im}\theta_m(t) + \epsilon_i, x(t).$$

where $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) = (10, 8, 6, 4, 2)$, $\epsilon_i(t)$ is distributed from $N(0, 0.3)$ for each $t$ and $z_{i1}, \ldots, z_{i5}$ is distributed from $N(0, 1)$.

The corresponding index function was generated by

$$\beta(t) = \sum_{m=1}^{5} \theta_m(t)\eta_m$$
where $\eta_1, \ldots, \eta_5$ have distribution $N(0, 1)$. Let $R_1 = [1, 200]$, $R_2 = [1, 45]$ and $R_3 = [46, 200]$ be restricted regions and $P_1 = \int_{R_1} X_i(t)\beta(t)dt$, $P_2 = \int_{R_2} X_i(t)\beta(t)dt$ and $P_3 = \int_{R_3} X_i(t)\beta(t)dt$ be corresponding projections. Then, the scalar response came from the nonlinear model

$$y_i = \frac{2P_1\pi}{50} \cos \frac{P_1\pi}{50} - \frac{2P_2\pi}{100} \sin \frac{P_2\pi}{100} + \sin \frac{P_3\pi}{400} + \sigma_y \epsilon_i$$

where $\epsilon_i$ is from $N(0, 1)$ and $\sigma_y = 0.01$. We also fix the sample size $n$ to be 60 for training set and to be 40 for test set. Before conducting this simulation scenario, we estimated the curve $X_i(t)$ by the smoothing spline method with the assumption that B spline basis function is known.

Table 3.5 also shows the result that is consistent with results presented in previous in regression and classification cases. Spearman rank based correlation gave a better results in this simulation but our global direction with Gaussian kernel outperforms over all others.

Table 3.6 reports the average of prediction error for MPM model fitted by functional linear model and by the kernel machines involved on each column compared to the projection from the functional principal component fitted by functional linear model and by LASSO and Danzig selector. Three projections are selected for MPM for predefined thresholding sequence 0, 0.1, 0.2 on the covariance function where those are from the PC derived by one entire block at 0, the PC derived by one block at 0.1 and the PC derived by the other block at 0.1. For the last row in Table 3.6, the projection from the functional principal component was used for fitting functional linear model. LASSO constraint was given to $\eta$, $Danzig_1$ constraint was given to the first derivative for $\theta$. Similarly, $Danzig_2$ constraint was given to the second derivative for $\theta$. Overall, MPM fitted by the kernel machine greatly improve the prediction accuracy compared to other methods.

### 3.4 Real Data Analysis

Hippocampal surfaces data can be represented as a $30000 \times 7$ matrix at the baseline for each subject, where the 7 numbers in row indicate the 7 features of a vertex on the hippocampus.
To be specific, the first column is the radial distance, the second to the fourth columns are the multivariate tensor-based morphometry features and the fifth column is the determinant of Jacobian matrix. These information are taken from a specific location on the brain surfaces. The first $15000 \times 7$ matrix presents data obtained from different locations on left hippocampus and the second $15000 \times 7$ matrix presents data obtained from the right hippocampus. In this data analysis subsection, we focus on the variable radial distance, which is the first column of data. Besides the hippocampal surface data, there are several clinical and genetic covariates. We include the diagnostic covariates, gender and age. Data analysis also includes the APOE genetic covariates since relevant studies have shown that the APOE4 genotypes have significant effect on the subject. Our goal is to predict subject’s behavior score at the time when it is five year after the baseline, based on our method. We have 406 individuals and among them, hippocampal data for 226 individuals and for 180 individuals used for training data and for test data, respectively.

Figure 3.1 depicts images of the correlation matrix for 1000 radial distances that are selected based on our method. Image for A shows the raw image of the correlation where the first 500 coordinates on each vertical and horizontal line obtained from the left part of hippocampal data and the rest 500 coordinates on each vertical and horizontal line obtained from the right part of data. Then, we permuted all selected radial distances and properly reordered them for conducting the local dimension reduction. Image for B shows reordered correlation image. Images for C and D present reordered correlations with thresholding at the level of 0.5 and 0.7, respectively. According to Figure 1, we can figure out that selected radial distance has several distinct subgroups of the different scale while varying thresholding level.

Shown in Figure 3.3 are the contour of cross validation values taken over the level of the correlation thresholding versus the level of the smoothing parameter of the Gaussian kernel for each method. Caption A shows the results of the case employing usual regression based on MMP and the others B, C and D are associated with the result of the case employing kernel regression, kernel principal component analysis and kernel supervised principal component analysis based on MMP, respectively. On the average, the optimal thresholding level for the correlation was
observed to be closer 0.7 except for A case. The discussion for the level of the smoothing parameter will be done late after we point out some issue.

Figure 3.4 presents the relative behavior of the relationship between the cross validation value and the level of smoothing parameter when the reduction dimension was varied in each method. The first row in Figure 3 are associated with its relationship when the dimension of data is 30000 on each method. Since the degree of the noise accumulation is proportion to the size of the data dimension, they show that there is the difficulty of selecting the optimal level of smoothing parameter of the kernel. In this sense, we may require dimension reduction method even when employing kernel machine. As we suggested in section 2, we reduced the dimension of our data based on our method and shown on the second row in Figure 3 are the relationship when the reduced dimension is 1000. They seem to be improved compared to the case where the dimension is 30000 but it is still needed to be further dimension reduction. As further reduction method, by employing MMP, the third row in Figure 3 shows that the signal recovery is possible. As the result, if we apply our MMP method especially to kernel machine, we can obtain the kernel matrix in an efficient way in computation sense, can search on smoothing parameter in the small number of candidate values in a given range and can avoid from the problem related to the noise accumulation, These are benefits from employing MMP method.

Figure 3.2 depicts the relationship between the cross validation value and the level of smoothing parameter for employing MMP at the case where the dimension is 1000 on each method. We finally estimated 1 as the smoothing parameter level for using only MMP case and 0.02 and 0.045 were estimated as the smoothing levels for using MMP and kernel machines in each method, respectively.

Table 3.7 reports the sum of prediction error for MPM model fitted by the different kernel machine denoted in the row compared to the usual model fitted by the kernel machine and LASSO model, based on the selected dimension number denoted in the column. Overall, MPM fitted by the kernel machine greatly improve the prediction accuracy over the different level of the screening with Gaussian kernel than usual model fitted by the kernel machine and the LASSO
Table 3.7: Sum of prediction error in hippocampal surfaces data.

<table>
<thead>
<tr>
<th>Method</th>
<th>1000</th>
<th>700</th>
<th>500</th>
<th>30000</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPM + REG</td>
<td>383.19</td>
<td>384.29</td>
<td>383.39</td>
<td></td>
</tr>
<tr>
<td>MPM + KREG</td>
<td>152.40</td>
<td>152.53</td>
<td>152.28</td>
<td></td>
</tr>
<tr>
<td>MPM + KPCA</td>
<td>152.44</td>
<td>159.19</td>
<td>160.07</td>
<td></td>
</tr>
<tr>
<td>MPM + KSPCA</td>
<td>151.92</td>
<td>151.56</td>
<td>152.36</td>
<td></td>
</tr>
<tr>
<td>KPC</td>
<td>175.09</td>
<td>165.71</td>
<td>160.59</td>
<td>237.36</td>
</tr>
<tr>
<td>KSPC</td>
<td>175.79</td>
<td>165.61</td>
<td>160.86</td>
<td>415.67</td>
</tr>
<tr>
<td>LASSO</td>
<td>381.70</td>
<td>382.00</td>
<td>382.04</td>
<td>237.36</td>
</tr>
</tbody>
</table>

Each column is the size of $|\hat{M}|$. Each row presents that MPM submodel fitted by the kernel regression machine, MPM submodel fitted by the kernel principal component machine, MPM submodel fitted by the kernel supervised principal component, KSIS submodel fitted by the kernel principal component machine, KSIS submodel fitted by the kernel supervised principal component machine, and KSIS submodel fitted by the LASSO method.

Based on Table 3.7, we finally decide to employ MMP method as the dimension reduction method among others for hippocampal surfaces data. While we was treating diagnostic status, gender, age and APOE genetic information as fixed covariates for the purpose of the study, we estimated least square coefficients for these variables on the response subtracted fitted values on kernel method (our method) from the behavior score. The estimated coefficients are given by

$\beta_{\text{age}} = -0.26434$, $\beta_{D} = 13.01$, $\beta_{A_{p}=3} = 0.23$, $\beta_{A_{p}=4} = 1.26$, $\beta_{A_{p}=3} = 14.39$, $\beta_{A_{p}=4} = 14.92$, and $\beta_{\text{male}} = -0.68$ where we use abbreviation D as Alzheimer’s disease status and AP as APOE genetic information.

3.5 Conclusion

In this paper, we suggest multiple projection model (MPM) as a general framework with applications including the regression and the classification. By incorporating the global and local properties, MPM enables to estimate candidate projections and to predict target response by an appropriate kernel machine. It is shown that it greatly improves the prediction accuracy in nonlinear assumption.

There may be some contributions of this paper. Firstly, we propose general type of independent screening statistic which can detect both linear and nonlinear relationship. Since it does not
require strong assumptions, we expect that it would be an useful technique in variable selection problem. Secondly, it is possible to capture the complex local information among covariates by employing the local dimension reduction.

We believe that it would be the novel conception to detect the correlation structure. Also, we asymptotically study for the sure screening property and the uniform convergence of thresholded estimator under our frame in the ultrahigh dimensionality. Extensive simulations show our method outperforms the other state-of-the art method. As the future research, we expect to extend our notion to complicated functional data analysis issues.
Figure 3.1: Images of correlation matrix: (A) left vs right, (B) reorder for both, (C) thresholding at 0.5 (D) thresholding at 0.7.

Figure 3.2: Smoothing plots: (1) 30000, (2) 1000, (3) MMP at 1000.
Figure 3.3: Contour plots: (A) MMP, (B) KREG, (C) KPCA, (D) KSPCA.

Figure 3.4: Smoothing plots: (A) MMP, (B) KREG, (C) KPCA, (D) KSPCA.
CHAPTER 4: SILFM: SINGLE INDEX LATENT FACTOR MODEL BASED ON HIGH-DIMENSIONAL FEATURES

4.1 Introduction

We consider a high-dimensional prediction problem based on a set of \( n \) independent observations \( \{(x_i, y_i) : i = 1, \ldots, n\} \), where \( x_i \) is a \( p_x \times 1 \) vector of all candidate features and \( y_i \) is an outcome variable, such as diagnostic status. Without loss of generality, we consider a nonparametric prediction model given by

\[
y_i = f(x_i) + \sigma_y \epsilon_{iy} = f(x_{i1}, \ldots, x_{ip_x}) + \sigma_y \epsilon_{iy},
\]

where \( f(\cdots) \) is a generic link function and \( \epsilon_{iy} \sim N(0,1) \). In the classical setting with \( n \gg p_x \), various parametric and nonparametric regression models have been developed to find a linear/nonlinear combination of predictors which can efficiently characterize \( y_i \) (Hastie et al. 2009, Zhang and Singer 2010, Clarke et al. 2009). Although there is a large literature on the development of supervised learning methods for prediction problems (Hastie et al. 2009, Friedman 1991, Zhang and Singer 2010, Clarke et al. 2009), most of them suffer from the curse of dimensionality due to diverging spectra and noise accumulation in the high dimensional feature space with \( p_x \gg n \) (Fan and Lv 2008, Bickel and Levina 2004). For instance, in medical imaging studies, it is interesting to study the predictive value of image signals at millions of locations (or voxels) \( (p_x \sim 10^6) \) for clinical outcomes. High variance and overfitting have been major concerns in this setting. Therefore, it is imperative to use dimension reduction and/or regularization methods, such as projection, screening methods, or the Lasso, to extract and select ‘low-dimensional’ and ‘informative’ features, while avoiding overfitting (Liu et al. 2011, Zou
and Hastie 2005, Bair et al. 2006, Fan and Fan 2008, Chun and Keles 2010, Krishnan et al. 2011). Although many marginal variable screening techniques, such as the Sure Independence Screening (SIS) procedure, are shown to be able to filter out many uninformative variables in many scenarios (Fan and Lv 2008, Li et al. 2012a, Mai and Zou 2013, Fan and Song 2010). These ‘informative’ features selected from such screening methods can be highly correlated and non-sparse.

Throughout the paper, we use \( \tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip_x}) \) to denote a \( p_x \times 1 \) vector of relatively low-dimensional informative features for predicting \( y_i \). In this case, model (4.1) reduces to

\[
y_i = f(\tilde{x}_i) + \sigma_y \epsilon_{iy} = f(\tilde{x}_{i1}, \ldots, \tilde{x}_{ip_x}) + \sigma_y \epsilon_{iy}.
\]  

(4.2)

In many applications, such as genetics or neuroimaging, \( x_i \) and/or \( \tilde{x}_i \) can be highly correlated, and moreover, the number of important features can be *non-sparse*, that is, \( p_x \gg \tilde{p}_x \gg n \). Such a high correlation structure and non-sparsity are notoriously difficult for existing dimension reduction and regularization methods (Fan and Lv 2010, Zou 2006, Zou et al. 2006, Buhlmann et al. 2012, Fan and Fan 2008, Tibshirani 1996). For instance, almost all regularization methods for high-dimensional regression strongly depend on some assumptions on the correlation structure of \( x_i \) and the sparsity (Buhlmann et al. 2012, Zhao and Yu 2006, Candes and Tao 2007). Moreover, individual features can be weakly correlated with the response, whereas their joint effect can be strong. Therefore, it is imperative to aggregate these correlated and informative features into \( p_z \) key features with \( p_z \ll n \).

Let \( z_i = (z_{i1}, \ldots, z_{ip_z})^T \) be a \( p_z \times 1 \) vector of such key features. Finally, model (4.1) may be approximated by

\[
y_i = f(z_i) + \sigma_y \epsilon_{iy} = f(z_{i1}, \ldots, z_{ip_z}) + \sigma_y \epsilon_{iy}.
\]  

(4.3)

When \( z_i = \Gamma_z x_i \), in which \( \Gamma_z \) is a \( p_z \times p_x \) matrix, model (4.3) reduces to the well-known semi-parametric index model in the dimensional reduction literature (Cook et al. 2010, Li et al. 2005, Cook et al. 2010, Li et al. 2005, 84
Cook and Ni 2005, Li 1991, Xia et al. 2002, Xia 2007, Sheng and Yin 2013, Zhang and Yin 2014). Most existing dimension reduction methods focus on the scenario when $p_x$ is smaller than $n$. See Ma and Zhu (2013b) for a comprehensive review on dimension reduction. Little has been done on the scenario when $p_x$ is much larger than $n$ and/or $x_i$ is highly correlated due to many statistical and computational challenges (Li 2007, Yin and Hilafu 2015, Ma and Zhu 2013b, Yu et al. 2013). For instance, many sufficient variable selection methods require the calculation of a large sample covariance matrix of $x_i$ and its inverse, which can be non-trivial (Yin and Hilafu 2015, Chen et al. 2010, Yu et al. 2013).

Figure 4.1: Results from simulated data sets: the top row includes a selected covariate image $x_i$, an informative covariate image $\tilde{x}_i$, a selected key feature $z_i$ versus a selected feature of $\tilde{x}_i$; the second row includes the scatter plot $y_i$ and the selected feature of $z_i$ and the average prediction error plots for SILFM1 (red) and SILFM2 (blue) and other competing methods (KRR (orange), SVM (skyblue), LASSO (green) SCAD (darkgreen), PLS (pink), SPLS (hotpink), PCA (black), SPCA (gray), and SSDR (purple)) in two simulation scenarios in subsection 3.
The aim of this paper is to develop a single index latent factor model (SILFM) framework using (4.3) to predict \( y_i \) using \( x_i \). SILFM can be regarded as an extension/integration of the well-known single index model, the high-dimensional linear model (HLM), and the latent factor model in the literature. Our results in the simulations and real data analysis show that such integration is very powerful for handling high-dimensional correlated features. As an illustration, we consider simulated data with \( n = 100 \) from model (4.3) (See simulation Section. for this scenario). The first row of Figure 1 presents the observed covariate image \( x \) with \( p_x = 3000 \) (panel A), the informative covariate image \( \tilde{x} \) with \( \tilde{p}_x = 1000 \) (panel B), one component of the key feature \( z \) with \( p_z = 3 \) versus one feature of \( \tilde{x} \) (panel C), and one component of \( z \) versus the outcome variable \( y \) (panel D). We have compared SILFM with eleven state-of-the-art methods, such as sequential sufficient dimension reduction (Yin and Hilafu 2015) and sparse partial least squares (SPLS) (Chung et al. 2012). Compared to all these competing methods, SILFM shows very promising results in terms of prediction accuracy for some interesting scenarios (panels E and F). For more details, please see simulation Section.

Compared with the existing literature, we make at least four major contributions in this paper:

- Model (4.3) differs from most single index models considered in the literature, in which \( z_i = \Gamma_z x_i \). Specifically, we introduce a latent factor model to characterize the potential relationship between \( z_i \) and \( x_i \). Such a latent factor model can be useful and powerful for handling weak and correlated individual signals, but strong joint effects.

- Moreover, model (4.3) differs from those models considered in many contemporary works on variable selection, where the signals are mostly rare but strong. For instance, to deal with the “curse-of-dimensionality”, it is common to assume an additive structure with \( f(x_i) = \sum_{j=1}^{p_x} f_j(x_{ij}) \) and a sparse signal \( \#\{j : f_j(\cdot) \neq \text{constant}\} \ll n \).

- A comprehensive three-stage estimation procedure is developed to adaptively and sequentially improve prediction accuracy. Our estimation procedure includes screening, aggregating, and nonlinear fitting. Each step is computationally efficient even for the high-dimensional
scenario with $p_x \gg n$.

- We investigate several theoretical properties of SILFM, such as the sure independence screening property and risk bounds.

This chapter is organized as follows. In Section 2, we introduce the general SILFM framework. In Sections 3, simulation studies are conducted to evaluate the small-sample performance of SILFM. In Section 4, we apply SILFM to the analysis of hippocampus data obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset. In Section 5, we systematically investigate the theoretical properties of SILFM. Concluding remarks are given in Section 6.

### 4.2 SILFM: Single Index Latent Factor Model

#### 4.2.1 Model Setup

The measurement models of SILFM are specified by

\[
y_i = f(z_i) + \sigma_y \epsilon_{iy}, \quad (4.4)
\]

\[
\tilde{x}_i = F_R(x_i) = G(z_i) + \epsilon_{ix}, \quad (4.5)
\]

where $\epsilon_{ix}$ is a $\tilde{p}_x \times 1$ vector of measurement errors with zero mean, $F_R(\cdot) : R^{p_x} \rightarrow R^{\tilde{p}_x}$ is a dimension reduction function of $x_i$, and $G(\cdot) : R^{p_z} \rightarrow R^{\tilde{p}_x}$ is a smooth function of $z_i$. SILFM includes many well-known models as special cases. For instance, if $z_i = x_i$ and $f(x_i) = \sum_{j=1}^{p_x} f_j(x_{ij})$, then SILFM reduces to additive models. Furthermore, if $f(x_i) = \sum_{j=1}^{p_x} x_{ij} \beta_j$, then SILFM reduces to a high-dimensional linear model. Moreover, when $\tilde{x}_i = x_i = z_i + \epsilon_{ix}$, SILFM reduces to a measurement error model.

Model (4.5) is a generalized version of standard latent factor models when $G(z_i) = \Lambda_G z_i$, in which $\Lambda_G$ is a $\tilde{p}_x \times p_z$ matrix. Model (4.5) includes many well-known models as special cases.
Specifically, if $G(z_i) = \Lambda_G z_i$ and $\Lambda_G$ is full column rank, then $z_i$ can be rewritten as

$$z_i = \Gamma_G \tilde{x}_i - \Gamma_G \epsilon_{ix} = (\Lambda_T^T \Lambda_G)^{-1} \Lambda_T^T F_R(x_i) - (\Lambda_T^T \Lambda_G)^{-1} \Lambda_T^T G \epsilon_{ix}, \quad (4.6)$$

where $\Gamma_G = (\Lambda_T^T \Lambda_G)^{-1} \Lambda_T^T$. Furthermore, if $\epsilon_{ix} = 0$ and $F_R(x_i) = \Lambda_R x_i$, in which $\Lambda_R$ is a $p_x \times p_x$ matrix, then $z_i$ can be written as $(\Lambda_T^T \Lambda_G)^{-1} \Lambda_T^T \Lambda_R x_i$ and model (4.4) reduces to the well-known single-index model. When $\tilde{x}_i = x_i$ and $G(\cdot)$ is a nonlinear function, model (4.5) reduces to a standard model for the nonlinear dimension reduction.

A unique feature of SILFM is that (4.5) integrates both a selection process and dimension reduction into a single formulation. Specifically, $F_R(\cdot)$ and $G(\cdot)$ can be regarded as a feature selection map and a dimension reduction map, respectively. It may allow us to efficiently deal with weak and correlated individual signals, that may have strong joint effects on $y_i$. By using $F_R(\cdot)$, we may be able to eliminate many individual signals unrelated to prediction. The use of $G(\cdot)$ allows us to aggregate many weak and correlated individual signals into a few strong and independent signals.

### 4.2.2 Estimation Procedure

We develop a three-stage estimation procedure in order to sequentially estimate $F_R(\cdot)$, $G(\cdot)$, and $f(\cdot)$, while achieving better prediction accuracy. Our estimation procedure SILFM is a three-stage process consisting of screening, aggregating, and nonlinear fitting as follows:

$$\begin{align*}
(y_i, x_i) & \implies \tilde{x}_i & \implies z_i & \implies y_i = \hat{f}(z_i).
\end{align*} \quad (4.7)
$$

See Figure 4.2 for an overview of our procedure, whose three stages are given as follows.

- **Stage (I).** Use a Sure Independence Screening (SIS) procedure based on a Hilbert-Schmidt Independence Criterion (HSIC) to select a set of important features $\tilde{x}_i$.

- **Stage (II).** Extract the key features $z_i$ from the selected important features.
Figure 4.2: Path diagram of SILFM estimation procedure.
• Stage (III). Use a kernel ridge regression to build a prediction method based on the extracted key features.

Stage (I) is a fully nonparametric robust screening method based on HSIC. The key steps of Stage (I) include three steps as follows.

• Step (I.1). Use HSIC and its associated \( p \)-value to measure the relationship of each feature individually to the response.

• Step (I.2). Rank marginal HSIC values or their \( p \)-values according to their size (or their degree of dependence to the response).

• Step (I.3). Filter out all noisy features whose size is smaller than a given threshold.

The HSIC statistic is a two-variable independence test in Reproducing Kernel Hilbert Spaces (RKHS) (Gretton et al. 2005). As shown in Sejdinovic et al. (2013), the HSIC statistic is consistent when a characteristic kernel is used and is equivalent to the distance covariance (DC) test of multivariate independence when the distance-induced kernel in HSIC is chosen (Székely et al. 2007). Moreover, the HSIC test can be more sensitive than DC when other kernels are used, and the HSIC test can be readily extended to many metric spaces. It should be noted that the use of HSIC is not critical in Stage (I) and any other independence test, such as the fused Kolmogorov filter developed in Mai and Zou (2013), can be used here.

We review the key ideas of HSIC for testing the independence between two random variables. Let \( Z \sim \mathbb{P}_Z \) and \( Y \sim \mathbb{P}_Y \) be, respectively, random variables on \( \mathcal{Z} \) and \( \mathcal{Y} \), which are two nonempty topological spaces. Let \( \mathbb{P}_{Z,Y} \) be the joint probability measure of \((Z,Y)\). Let \( K_Z \) and \( K_Y \) be kernels on \( \mathcal{Z} \) and \( \mathcal{Y} \) with respective RKHSs \( \mathcal{H}_{K_Z} \) and \( \mathcal{H}_{K_Y} \). Then, it is well known that \( K_{Z\times Y}((z,y),(z',y')) = K_Z(z,z')K_Y(y,y') \) is a kernel on the product space \( \mathcal{Z} \times \mathcal{Y} \) with RKHS \( \mathcal{H}_{K_{Z\times Y}} \) that is isomorphic to the tensor product \( \mathcal{H}_{K_Z} \otimes \mathcal{H}_{K_Y} \). The HSIC of \( Z \) and \( Y \) is defined as

\[
\text{HSIC}(Z,Y)^2 = \int \int K_{Z\times Y}d([\mathbb{P}_{Z,Y} - \mathbb{P}_Z\mathbb{P}_Y] \times [\mathbb{P}_{Z,Y} - \mathbb{P}_Z\mathbb{P}_Y]). \tag{4.8}
\]
A fundamental result is that if $K_Z$ and $K_Y$ are universal kernels, then $\text{HSIC}(Z,Y) = 0$ if and only if $P_{Z,Y} = P_Z P_Y$.

We construct an empirical estimate of HSIC. Let $H_n$ be a centering matrix $I_n - n^{-1}1_n1_n^T$, where $I_n$ is an $n \times n$ identity matrix and $1_n = (1, \cdots, 1)^T$ is an $n \times 1$ vector with all elements 1. Let $K_{Z,n}$ be an $n \times n$ matrix with the $(i,j)$th element $K_Z(z_i, z_j)$, and let $K_{Y,n}$ be an $n \times n$ matrix with the $(i,j)$th element $K_Y(y_i, y_j)$. Given an independently and identically distributed sample $\{(z_i, y_i)\}_{i=1}^n$, we can construct an empirical estimate of HSIC as the sum of U-statistics given by

$$\hat{\text{HSIC}}(Z,Y)^2 = n^{-2}\text{tr}(K_{Z,n}H_nK_{Y,n}H_n).$$

The estimated $n\hat{\text{HSIC}}(Z,Y)^2$ has some nice statistical properties, which form the theoretical foundation of the HSIC screening procedure. Statistically, as $n \to \infty$, $n\hat{\text{HSIC}}(Z,Y)^2$ converges to the weighted sum of $\chi^2(1)$ random variables in distribution (Gretton et al. 2005, Sejdinovic et al. 2013, Székely et al. 2007). Since different features may have different patterns, such as scale, we use a computationally fast approach based on a spectral method to approximate the $p-$value of $\hat{\text{HSIC}}$ for each feature. Specifically, for the $j-$th component of $x_i$, we calculate its HSIC and $p-$value. However, for computational simplicity, it is more convenient to directly use the value of the estimated HSIC to filter out 'noisy' features. In this case, for a given threshold $\gamma_n$, we can form the set of important features according to

$$\hat{M}_{\gamma_n} = \{1 \leq j \leq p_x : |n\hat{\text{HSIC}}(X_j,Y)^2| \geq \gamma_n\},$$

where $X_j$ and $Y$ are, respectively, the random variables for the $j-$th component of $x$ and $y$. Theoretically, we will show that our variable screening procedure enjoys the sure independence screening property under some mild conditions. Compared to test marginal screening methods, Stage (I) aims to use a relatively small $\gamma_n$ in order to increase the chance of keeping all important and/or week signals.
Stage (II) is not only a dimension-reduction method, but it is also an information aggregation method. Consider the true active set $\mathcal{M} = \{1, \cdots, \tilde{p}_x\}$ for the variables in $\tilde{x}_i$. Stage (II) includes three steps as follows:

- **Step (II.1).** Calculate the (kernel) correlation matrix of the selected features, denoted by $R_{\tilde{x}} = (r_{jk})_{1 \leq j, k \leq \tilde{p}_x}$.
- **Step (II.2).** Use the covariance thresholding method introduced by Bickel and Levina (2008) and the spectral clustering method to partition $\mathcal{M}$ into $p_{z,s}$ multiple disjoint clusters $\mathcal{M} = \bigcup_{k=1}^{p_{z,s}} \mathcal{M}_{k,s}$ with $\mathcal{M}_{k,s} \cap \mathcal{M}_{k',s} = \emptyset$ for $k \neq k'$ and $s = 1, \cdots, S$, where $\mathcal{M}_{k,s}$ is a subset of $\mathcal{M}$ and $p_{z,s}$ is an integer, which may vary across $s$. For each $s$, let $\tilde{r}_s$ be a given thresholding value and $T_{\tilde{r}_s}$ be thresholding operator such that $T_{\tilde{r}_s}(R_{\tilde{x}}) = (r_{jk}I(|r_{jk}| \geq \tilde{r}_s))$. Let $\Pi$ be a spectral clustering function that maps each $j \in \tilde{\mathcal{M}}_{\gamma_n}$ into a unique cluster $\mathcal{M}_{k,s}$ based on $T_{\tilde{r}_s}(R_{\tilde{x}})$. That is, $\Pi(\cdot, \cdot)$ is defined as $\Pi(j, T_{\tilde{r}_s}(R_{\tilde{x}})) \in \mathcal{M}_{k,s}$ for each $j \in \tilde{\mathcal{M}}_{\gamma_n}$.
- **Step (II.3).** For each $\mathcal{M}_{k,s}$, we calculate the sample (kernel) covariance matrix of these features with their indices in $\mathcal{M}_{k,s}$, denoted as $S_{X,k,s}$, and the eigenvalue-eigenvector pairs of $S_{X,k,s}$. Finally, we extract the key features $z_i$ based on the scores from the eigenvectors corresponding to the $r_{k,s}$ algebraically largest eigenvalues of $S_{X,k,s}$.

Stage (II) can be regarded as a novel generalization of the supervised PCA method (Bair et al. 2006), since it conducts standard PCA on marginally selected features with their indices in each cluster. A key difference is that in Step (II.2), we choose a series of $0 \leq \tilde{r}_1 < \cdots < \tilde{r}_S < 1$ so that we can threshold the correlation matrix at different levels. It is expected that the larger $\tilde{r}_s$ is, the larger $p_{z,s}$ is. Equivalently, for large $\tilde{r}_s$, we only use group features that are highly correlated with each other. This point is similar to $\tau$-separation (Buhlmann et al. 2012) to seek the finest group features where the number of the group feature is chosen by $\tau$ to deal with the features with strong dependency. Compared to the hierarchical bottom-up agglomerative clustering algorithm for estimating $\tau$-separation, we can take the advantage of the computational expediency when
using the spectral clustering algorithm. As varying a series of thresholds, we extract information from the selected features at different degrees of correlation, which allow us to select the most informative projected features that have the largest prediction power in Stage (III).

4.3 Simulation Studies

In this section, we conducted three simulation studies in order to examine the small sample performance of SILFM and we compare SILFM with other competing methods. In order to compare different methods, we examined two types of performance measures for dimension reduction and prediction accuracy. For each scenario, 100 simulated data sets were generated, while each simulated data set consists of a training set with $n = 100$ and a test set with $n = 100$. First, for dimension reduction, we consider the true positive rate defined as $P_{tp} = \frac{\left|\hat{M}_{\gamma_n} \cap M\right|}{\left|M\right|}$, the screening accuracy defined as $P_A = \frac{\left|\hat{M}_{\gamma_n} \cap M\right|}{\left|\hat{M}_{\gamma_n}\right|}$, and the true negative rate defined as $P_{tn} = \frac{\left|M^c - \hat{M}^c_{\gamma_n}\right|}{\left|M^c\right|}$ where $M^c$ and $\hat{M}^c_{\gamma_n}$ are, respectively, the compliment of $M$ and $\hat{M}_{\gamma_n}$. Second, for prediction accuracy, we computed the empirical squared prediction error of the test data set as $n^{-1} \sum_{i=1}^{n} (y^\text{test}_i - \hat{f}(x^\text{test}_i))^2$, where $\hat{f}(\cdot)$ is the prediction model built from the training set and $(x^\text{test}_i, y^\text{test}_i)$s’ are observations in the test set.

We simulated data from different $f(\cdot)$s in SILFM with both fixed and random designs for $X$ and $\epsilon \sim N_n(0, I)$ with $n = 100$, and $p_x = 3000$. We generated $X$ from a multivariate normal distribution $N_{p_x}(0, \Sigma)$. Among the $x_i$, we set the number of informative features as $\tilde{p}_x = s = 600$ and created them from two clusters with each having $m_k = |M_k| = 300$ features for $k = 1, 2$, respectively. Specifically, the correlation structure among all informative features $\Sigma_s$ consists of two block-diagonal structures $\Sigma_s = \text{diag}(\Sigma_1, \Sigma_2)$ where we used a highly correlated structure $\rho_w = 0.9$ as the within cluster correlation and $\rho_b = 0.6$ as the between cluster correlation. Also, we set the number of insignificant features as $s^c = |M^c| = 2400$ and the correlation structure among all of them ($\Sigma_s^c$) consists of the identity matrix, $I_{s^c}$. Therefore, the correlation structure among all features $\Sigma$ consists of $\Sigma = \text{diag}(\Sigma_s, \Sigma_s^c)$. 

93
4.3.1 Simulation 1: Continuous Response (I)

We directly simulated \( x_i \) from \( \mathcal{N}(0, \Sigma) \) with \( \Sigma \) being a block diagonal matrix mentioned above. For \( F_R(\cdot) \), we consider \( \mathcal{M} = \{1 \leq j \leq 600\} \) as the active set for the variables in \( \tilde{x}_i \) where \( F_R(x_i) = \tilde{x}_i \) and \( |\mathcal{M}| = \hat{p}_x = 600 \). We consider a nonlinear model as follows:

\[
y_i = \sqrt{z_{i1}^2 + z_{i2}^2 + z_{i3}^2} + \log(\sqrt{z_{i1}^2 + z_{i2}^2 + z_{i3}^2}) + \sigma_y \epsilon_{iy},
\]

where \( \sigma_y = 0.2 \) and \( z_i = (z_{i1}, z_{i2}, z_{i3})^T \) is a \( 3 \times 1 \) vector of key features specified by \( z_{i1} = 1^T_{600} \tilde{x}_i / 600 \), \( z_{i2} = (1^T_{300}, 0^T_{300}) \tilde{x}_i / 300 \), and \( z_{i3} = (0^T_{300}, 1^T_{300}) \tilde{x}_i / 300 \), in which \( 1_k \) is a \( k \times 1 \) vector of ones and \( 0_k \) is a \( k \times 1 \) vector of zeros. Major important features are in the first block of the highly correlated variables, while the number of such features is distinguished as non-sparse \( (p_x = 3000 \gg \hat{p}_x = 600 \gg n = 100) \).

In Stage (I) of SILFM, we compare the HSIC-SIS procedure with other SIS procedures based on distance correlation (DC-SIS), Pearson correlation (SIS), Spearman correlation (SP-SIS), and Kendall correlation (KD-SIS). We set the number of selected features \( \hat{\mathcal{M}} = \hat{p}_x \) to be 450, 650, and 850, respectively.

In Stage (II) of SILFM, we set the number of thresholdings \( S = 3 \), \( (\tilde{r}_1, \tilde{r}_2, \tilde{r}_3) = (0.0, 0.4, 0.7) \), estimated the total number of the cluster group for each \( \tilde{r}_k \) by using a spectral clustering method, and used the first principal component score from each cluster for all \( k \) to estimate \( \hat{z}_i \). It should be noted that the number of clusters may be different or the same for a fixed \( \tilde{r}_k \) in each simulation run, depending on the spectral algorithm. However, inducing different numbers of clusters between thresholdings is more important than the dependency on the clustering algorithm since these differences lead to distinct features in the aggregation procedure. For this reason, we do not need to use a large number of thresholdings.

In Stage (III) of SILFM, we consider three different kernel learning methods including KRR, KPCA, and KSPCA. For simplicity, we only report the results of KRR, since the results of KPCA and KSPCA are almost the same as those of KRR for this simulation setting.
For an extensive and fair comparison, we consider an integration of the HSIC-SIS procedures and eleven learning methods including KRR, SILFM with KRR (SILFM1), support vector regression (SVR), SILFM with SVR (SILFM2), LASSO, SCAD, partial least squares (PLS), sparse partial least squares (SPLS), principal components (PC), sparse principal components (SPC) and sequential sufficient dimension reduction (SSDR). We used the R packages kernlab (Karatzoglou et al. 2004), e1071 (Meyer and Wien 2014), glmnet (Friedman et al. 2010), ncvreg (Breheny and Huang 2011), pls (Mevik and Wehrens 2007), spls (Chung et al. 2012), elasticnet (Zou et al. 2006), and the SSDR code (Yin and Hilafu 2015). For all kernel related methods, the Gaussian RBF kernel \( \exp(-\|x_1 - x_2\|^2/\sigma) \) was used, in which we set \( 1/\sigma \) as \( 1/p \) and set \( p \) to be the number of selected features \( \hat{p}_x \). Then, we set the regularization parameter \( \lambda \) at \( \lambda = 0.001(n^{-1/4}) \). For other prediction methods, we used an optimized tuning parameter that minimizes a cross-validation error. Moreover, we fixed the dimension of key features at 4 for some methods that need the choice of dimension, such as the principal components and partial least squares methods.

We have the following simulation results. Table 4.1 presents five selected quantiles of indices for the true important features, empirical true positive, accuracy, true negative and false positive measures for each estimated size as top features. It is observed that HSIC outperforms other existing correlation measures. Table 4.2 reports the average of prediction error for each method calculated from the 100 test data sets. Our methods including SILFM1 and SILFM2 significantly outperform all other competing methods when the number of selected important features is close to \( \hat{p}_x = 600 \). Figure 4.1 shows the boxplots of the estimated prediction errors as \( \hat{p}_x = 450, 650, \) and \( 850 \). Figure 4.1 reveals that the SILFMs show smaller prediction errors and variabilities compared to the others.

### 4.3.2 Simulation 2: Continuous Response (II)

We directly simulated \( x_i \) from \( N(0, \Sigma) \) with \( \Sigma \) being a block diagonal matrix mentioned above. For \( F_R(\cdot) \), we consider \( M = \{1 \leq j \leq 600\} \) as the active set for the variables in \( \tilde{x}_i \) and
### Table 4.1: Performance of SIS methods in simulation 1

| Method     | 5%   | 25%  | 50%  | 70%  | 95%  | $P_{tp}$ | $P_A$ | $P_{tn}$ | $P_{fp}$ | $|\mathcal{M}|$ |
|------------|------|------|------|------|------|----------|-------|----------|----------|----------|
| Oracle     | 30.95| 150.75| 300.50| 450.25| 570.05| 1.00     | 1.00  | 1.00     | 0.00     | 450      |
| HSIC-SIS   | 24.25| 117.75| 234.25| 384.10| 557.89| 0.750    | 1.000 | 1.000    | 0.250    | 450      |
| DC-SIS     | 24.45| 117.57| 234.50| 385.97| 565.53| 0.744    | 0.993 | 0.998    | 0.255    | 450      |
| SIS        | 246.39| 679.62| 1261.05| 1896.17| 2734.66| 0.231    | 0.308 | 0.870    | 0.768    | 450      |
| KD-SIS     | 364.36| 765.57| 1413.00| 2228.52| 2839.12| 0.148    | 0.198 | 0.849    | 0.851    | 450      |
| SP-SIS     | 151.73| 747.00| 1483.30| 2219.42| 2837.05| 0.154    | 0.205 | 0.851    | 0.845    | 450      |
| HSIC-SIS   | 33.45| 163.25| 325.50| 487.75| 1389.92| 1.000    | 0.923 | 0.979    | 0.000    | 650      |
| DC-SIS     | 33.45| 163.25| 325.80| 489.85| 1500.51| 0.994    | 0.917 | 0.977    | 0.006    | 650      |
| SIS        | 214.48| 624.57| 1207.15| 1891.55| 2767.22| 0.349    | 0.322 | 0.816    | 0.651    | 650      |
| KD-SIS     | 295.58| 686.90| 1387.80| 2203.95| 2828.43| 0.231    | 0.213 | 0.787    | 0.768    | 650      |
| SP-SIS     | 140.58| 729.75| 1486.25| 2229.77| 2840.43| 0.224    | 0.207 | 0.785    | 0.775    | 650      |


### Table 4.2: Average of prediction errors in simulation 1

<table>
<thead>
<tr>
<th>Method</th>
<th>450</th>
<th>650</th>
<th>850</th>
<th>3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>SILFM-1</td>
<td>0.151</td>
<td>0.775</td>
<td>0.752</td>
<td>2.363</td>
</tr>
<tr>
<td>SILFM-2</td>
<td>0.157</td>
<td>0.834</td>
<td>0.822</td>
<td>2.362</td>
</tr>
<tr>
<td>KRR</td>
<td>2.249</td>
<td>2.255</td>
<td>2.259</td>
<td>2.363</td>
</tr>
<tr>
<td>SVR</td>
<td>2.252</td>
<td>2.258</td>
<td>2.261</td>
<td>2.362</td>
</tr>
<tr>
<td>LASSO</td>
<td>2.316</td>
<td>2.815</td>
<td>3.092</td>
<td>2.354</td>
</tr>
<tr>
<td>SCAD</td>
<td>2.431</td>
<td>3.027</td>
<td>2.967</td>
<td>2.677</td>
</tr>
<tr>
<td>PLS</td>
<td>2.874</td>
<td>2.799</td>
<td>2.636</td>
<td>2.456</td>
</tr>
<tr>
<td>SPLS</td>
<td>2.490</td>
<td>2.758</td>
<td>2.634</td>
<td>2.565</td>
</tr>
<tr>
<td>PCA</td>
<td>2.340</td>
<td>2.451</td>
<td>2.457</td>
<td>2.482</td>
</tr>
<tr>
<td>SPCA</td>
<td>2.373</td>
<td>2.376</td>
<td>2.377</td>
<td>2.861</td>
</tr>
<tr>
<td>SSDR</td>
<td>2.368</td>
<td>2.369</td>
<td>2.474</td>
<td>2.483</td>
</tr>
</tbody>
</table>

Table 4.2: Average of prediction errors in simulation 1: Each column indicates the number of $|\hat{\mathcal{M}}|$ used for the prediction model. Each row presents a learning method.
set $F_R(\mathbf{x}_i) = \tilde{\mathbf{x}}_i$ and $|\mathcal{M}| = \tilde{p}_x = 600$. We consider a nonlinear model as follows:

$$y_i = 4 \sin \left(3z_{i1} \frac{\pi}{10}\right) + 10z_{i2}^2 + 2.2z_{i3} + \sigma_y \epsilon_{iy},$$

where $\epsilon_{iy} \sim N(0, 1)$.

Moreover, $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})^T$ is a $3 \times 1$ vector of key features specified as

$$\mathbf{z}_i = \Gamma G(\tilde{\mathbf{x}}_i - \epsilon_{ix}) = \begin{pmatrix} B_{i1} & \cdots & B_{i200} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{i201} & \cdots & B_{i400} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & B_{i401} & \cdots & B_{i600} \end{pmatrix} (\tilde{\mathbf{x}}_i - \epsilon_{ix}) (4.9)$$

where $\epsilon_{ix} \sim N(0_{600}, \sigma_x^2 I_{600})$, in which $I_k$ denotes a $k \times k$ identity matrix, and the $B_{i,j}$ were sampled from $\mathcal{B} = \{2/(600), 4/(600), 6/(600), \ldots, 2\} = \{B_1, B_2, B_3, \ldots, B_{600}\}$ with the same probability without replacement.

Figure 4.3: Averages of true positive rate for five different screening methods in simulation 2: panels A ($\sigma_y = 0.2$, $\sigma_x = 0.2$), B ($\sigma_y = 2$, $\sigma_x = 0.2$), C ($\sigma_y = 0.2$, $\sigma_x = 1$), and D ($\sigma_y = 2$, $\sigma_x = 1$) report the results based on the different levels of $\sigma_y$ and $\sigma_x$.

This simulation setting is also similar to the first one such that the number of important features is distinguished as *non-sparse* and their effect on $y_i$ is nonlinear. Similar to the first simulation, we compared SILFM with the same set of competing methods by varying the values
Figure 4.4: Averages of prediction errors for 11 different predictive methods in simulation 2: all panels A ($\sigma_y = 0.2$, $\sigma_x = 0.2$), B ($\sigma_y = 2$, $\sigma_x = 0.2$), C ($\sigma_y = 0.2$, $\sigma_x = 1$), and D ($\sigma_y = 2$, $\sigma_x = 1$) report the results based on the different levels of $\sigma_y$ and $\sigma_x$.

of $\sigma_x$ and $\sigma_y$. Specifically, we consider four different levels of $(\sigma_x, \sigma_y)$ including $(0.2, 0.2)$, $(0.2, 2)$, $(1, 0.2)$, and $(1, 2)$. Figures 4.3 and 4.4 present the simulation results based on 100 independent simulation runs. Figure 4.3 reveals that HSIC-SIS and DC-SIS have similar performance and both outperform other SIS methods as $(\sigma_x, \sigma_y)$ varies. The HSIC-SIS achieves better true positive rates when all components of $\Gamma_C$ are not the same everywhere and this advantage is more applicable to the model with a highly correlated covariance structure. As expected, the standard SIS method is much worse than all other SIS methods across all configurations and has difficulty in finding the true significant features. Figure 4.4 shows that the SILFM methods have the best overall prediction performance compared to all competing methods for each estimated number of selected features across different levels of $(\sigma_x, \sigma_y)$ when the number of selected informative features is greater than the true $\tilde{p}_x$. Moreover, SILFM2 outperforms SILFM1 across different levels of $(\sigma_x, \sigma_y)$. As expected, increasing $(\sigma_x, \sigma_y)$ can increase prediction error.

4.3.3 Simulation 3: Binary Response on Random Design

We generated the same number of binary responses $\{0, 1\}$ and then independently generated a $4 \times 1$ key feature vector $z_i$ from two different multivariate normal distributions, $z_i \sim N_4(\mu_1, 0.5\mathbf{I}_2)$
for $y_i = 1$ and $z_i \sim N_4(\mu_0, 0.5I_2)$ for $y_i = 0$ where the two mean vectors are given by $\mu_1 = 1_4/2$ and $\mu_0 = -1_4/2$ in the third scenario.

We generated $\tilde{x}_i = \Lambda_G h(z_i) + \sigma_x \epsilon_{ix}$ based on (4.5) where $\Lambda_G$ is a $200 \times 4$ orthogonal matrix, that is, $\Lambda_G = 1_{50} \oplus 1_{50} \oplus 1_{50} \oplus 1_{50}$, and $h(z_i) = (h_1(z_{i1}), h_2(z_{i2}), h_3(z_{i3}), h_4(z_{i4}))^T$ is a $4 \times 1$ vector. Specifically, we used four functions $h_1(z_{i1}) = \tan^{-1}(z_{i1}) \cos(\tan^{-1}(z_{i1}))$, $h_2(z_{i2}) = \tan^{-1}(z_{i2}) \sin(\tan^{-1}(z_{i2}))$, $h_3(z_{i3}) = z_{i3}$, and $h_4(z_{i4}) = 3z_{i4}^3 + 2z_{i4}^2 - 2z_{i4} + 1$. Note that these 200 components of $\tilde{x}_i$ are set to be informative. In contrast, the vector of all insignificant features $\tilde{x}_i^c$ is set to be a $2300 \times 1$ vector and independently sampled from $N(0, \sigma_x^2 I_{2300})$ in which we fixed $\sigma_x = 0.1$. Thus, we consider $\mathcal{M} = \{1 \leq j \leq 200\}$, $p_x = 2500$ and $\tilde{p}_x = 200$ in this scenario.

Using Bayes’s rule, the probability of $Y$ given $Z$ is a nonlinear function of $Z$. We set $n = 60$ for the training set and 40 for the test set. We simulated 100 datasets and then evaluated the variable screening and prediction accuracy of SILFM.

In Stage (I) of SILFM, we compared the HSIC-SIS procedure with the Gaussian kernel to four other SIS procedures based on the Kolmogrov-Sminov (KS-SIS), Kendall (KD-SIS), $t$ ($t$-SIS), and Pearson correlation (SIS), which are commonly used for the classification problem. In Stage (II), we set the number of selected features $\tilde{p}_x$ as being 100, 200, and 300, respectively, and also consider three additional dimension reduction methods including the kernel principal component analysis (KPCA), the principal component analysis (PCA), and the sparse principal component analysis (SPCA) at each $\tilde{p}_x$. We set the number of thresholdings $S = 3$ and $(\tilde{r}_1, \tilde{r}_2, \tilde{r}_3) = (0.0, 0.1, 0.2)$, estimated the total number of the cluster group for each $\tilde{r}_k$, and obtained the number of the clustering group as 1, 2, and 4 for $\tilde{r}_1$, $\tilde{r}_2$, and $\tilde{r}_3$, respectively. Then, the first principal component score from each group was used so that the total dimension of key features is equal to 7.

In Stage (III) of SILFM, we consider eight different classification methods including kernel nearest neighbor (KNN), linear discriminant analysis (LDA), regression (REG), support vector machine (SVM) with linear and Gaussian kernel, kernel linear discriminant (KLD) analysis with
and without an $L_2$ penalty term, and KRR.

Tables 4.3 and 4.4 present the simulation results. Table 4.3 includes results from an underestimation scenario when $|\hat{M}| = 100$ and other results from an overestimation scenario when $|\hat{M}| = 300$. The HSIC-SIS procedures outperform all other SIS procedures. Table 4.4 reports the classification error of each method. Our dimension reduction method in Stage (II) dramatically improves the classification performance of almost all classification methods except KLDA. Overall, Stage (II) outperforms all competing dimension reduction methods.

### 4.4 Real Data Analysis

To illustrate the usefulness of SILFM, we consider a data set from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) study. The goal of this data analysis is to integrate imaging, genetic, and clinical variables at the baseline together in order to predict behavior scores (MMSE) five years from baseline. Such a prediction model not only predicts the cognitive trajectory, but also potentially provides new approaches for early diagnosis of AD. This earlier identification would allow for more efficient selection of samples for clinical trials and possibilities for earlier disease treatment. The sample in our investigation includes $n = 406$ subjects: 223 healthy controls (HC) (107 females and 116 males) and 183 individuals with AD (87 females and 96 males). We randomly split the whole sample into 226 individuals for the training data and 180 individuals used for the test data. The features in $x_i$ include diagnostic status, gender, age, APOE4, and 30,000 radial distance measures of the left and right hippocampus surfaces. Thus, we have 30,005 possible features in our data analysis. The hippocampus, as a part of the limbic system, plays an important role in the consolidation of information from short-term memory to long-term memory and spatial navigation.

We applied SILFM as follows. In Stage (I), we used HSIC-SIS to select the top 1,000, 700, and 500 features. In Stage (II), we also set $S = 3$, $(\tilde{r}_1, \tilde{r}_2, \tilde{r}_3) = (0.0, 0.5, 0.7)$, and $p_z = 10$. In Stage (III), we used KRR as our learning method. As a comparison, we directly applied seven learning methods including SVR, Lasso, SCAD, PLS, PCA, SSDR and RBM to the important
### Table 4.3: Performance of SIS methods in simulation 3: HSIC-SIS: Screening with HSIC, KS-SIS: Screening with Kolmogrov-Sminov statistic, KD-SIS: Screening with Kendall statistic T-SIS: Screening with t statistic SIS: Screening with Pearson correlation.

| Method | 5%  | 25%  | 50%  | 70%  | 95%  | $P_p$ | $P_A$ | $P_{tn}$ | $P_{fp}$ | $|\mathcal{M}|$ |
|--------|-----|------|------|------|------|-------|-------|----------|----------|----------|
| Oracle | 10.95 | 50.75 | 100.5 | 150.25 | 190.05 | 1.000 | 1.000 | 1.000 | 1.000 | 200 |
| HSIC-SIS | 68.66 | 104.25 | 136.55 | 164.25 | 190.53 | 0.499 | 0.999 | 0.999 | 0.500 | 100 |
| KS-SIS | 71.57 | 105.05 | 138.47 | 166.00 | 191.47 | 0.497 | 0.995 | 0.999 | 0.502 | 100 |
| KD-SIS | 93.22 | 119.50 | 145.50 | 170.80 | 194.51 | 0.497 | 0.995 | 0.999 | 0.502 | 100 |
| T-SIS | 91.65 | 118.75 | 145.27 | 171.16 | 194.01 | 0.499 | 0.998 | 0.999 | 0.501 | 100 |
| SIS | 70.77 | 97.40 | 125.60 | 151.32 | 176.00 | 0.495 | 0.990 | 0.999 | 0.504 | 100 |
| HSIC-SIS | 12.24 | 56.27 | 109.70 | 159.75 | 504.26 | 0.951 | 0.951 | 0.995 | 0.048 | 200 |
| KS-SIS | 20.41 | 64.93 | 121.45 | 171.40 | 1003.45 | 0.892 | 0.892 | 0.990 | 0.107 | 200 |
| KD-SIS | 41.11 | 89.61 | 141.85 | 297.11 | 1565.43 | 0.793 | 0.793 | 0.982 | 0.206 | 200 |
| T-SIS | 41.19 | 88.06 | 140.15 | 268.90 | 1578.12 | 0.801 | 0.801 | 0.982 | 0.198 | 200 |
| SIS | 42.19 | 89.01 | 141.00 | 272.11 | 1650.01 | 0.794 | 0.794 | 0.982 | 0.206 | 200 |
| HSIC-SIS | 16.09 | 76.20 | 151.05 | 769.92 | 2132.85 | 0.997 | 0.664 | 0.956 | 0.002 | 300 |
| KS-SIS | 20.74 | 82.03 | 157.25 | 751.23 | 2017.26 | 0.966 | 0.644 | 0.953 | 0.033 | 300 |
| KD-SIS | 35.94 | 99.05 | 183.00 | 1012.03 | 2138.66 | 0.882 | 0.588 | 0.946 | 0.117 | 300 |
| T-SIS | 35.24 | 98.15 | 177.82 | 1009.42 | 2151.77 | 0.887 | 0.591 | 0.946 | 0.112 | 300 |
| SIS | 36.24 | 99.60 | 179.07 | 1023.12 | 2160.41 | 0.880 | 0.586 | 0.946 | 0.119 | 300 |

### Table 4.4: Average of test error in simulation 3: SILFM: Projection by SILFM, KPCA: Projection by kernel PCs, PCA: Projection by PCs SPCA: Projection by sparse PCs.

| Method | $|\mathcal{M}|$ | KNN | LDA | REG | SVM | KLD$_2$ | KLD | KRR | SVM$_k$ |
|--------|---------------|-----|-----|-----|-----|--------|------|-----|---------|
| SILFM  | 100           | 0.151 | 0.116 | 0.170 | 0.112 | 0.106 | 0.340 | 0.103 | 0.116 |
| SILFM  | 200           | 0.113 | 0.106 | 0.170 | 0.105 | 0.101 | 0.218 | 0.101 | 0.103 |
| SILFM  | 300           | 0.142 | 0.178 | 0.217 | 0.126 | 0.133 | 0.282 | 0.136 | 0.126 |
| KPCA   | 100           | 0.278 | 0.430 | 0.500 | 0.496 | 0.288 | 0.202 | 0.287 | 0.371 |
| KPCA   | 200           | 0.370 | 0.487 | 0.500 | 0.500 | 0.408 | 0.353 | 0.408 | 0.486 |
| KPCA   | 300           | 0.347 | 0.500 | 0.498 | 0.500 | 0.387 | 0.342 | 0.386 | 0.500 |
| PCA    | 100           | 0.230 | 0.202 | 0.275 | 0.196 | 0.210 | 0.238 | 0.210 | 0.190 |
| PCA    | 200           | 0.320 | 0.310 | 0.346 | 0.298 | 0.291 | 0.295 | 0.293 | 0.290 |
| PCA    | 300           | 0.341 | 0.337 | 0.388 | 0.333 | 0.311 | 0.322 | 0.310 | 0.300 |
| SPCA   | 100           | 0.332 | 0.282 | 0.347 | 0.278 | 0.287 | 0.288 | 0.286 | 0.285 |
| SPCA   | 200           | 0.255 | 0.223 | 0.315 | 0.240 | 0.208 | 0.202 | 0.210 | 0.193 |
| SPCA   | 300           | 0.332 | 0.362 | 0.397 | 0.368 | 0.308 | 0.321 | 0.312 | 0.298 |

Table 4.3: Performance of SIS methods in simulation 3: HSIC-SIS: Screening with HSIC, KS-SIS: Screening with Kolmogrov-Sminov statistic, KD-SIS: Screening with Kendall statistic T-SIS: Screening with t statistic SIS: Screening with Pearson correlation.

Table 4.4: Average of test error in simulation 3: SILFM: Projection by SILFM, KPCA: Projection by kernel PCs, PCA: Projection by PCs SPCA: Projection by sparse PCs.
### Table 4.5: Sum of prediction errors in hippocampal surfaces data.

<table>
<thead>
<tr>
<th>Method</th>
<th>1000</th>
<th>700</th>
<th>500</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>SILFM</td>
<td>84.11</td>
<td>83.47</td>
<td>82.84</td>
<td>108.98</td>
</tr>
<tr>
<td>SVR</td>
<td>127.07</td>
<td>127.07</td>
<td>126.80</td>
<td>111.64</td>
</tr>
<tr>
<td>LASSO</td>
<td>97.71</td>
<td>97.71</td>
<td>97.75</td>
<td>107.92</td>
</tr>
<tr>
<td>SCAD</td>
<td>113.58</td>
<td>134.98</td>
<td>155.34</td>
<td>107.92</td>
</tr>
<tr>
<td>PLS</td>
<td>88.33</td>
<td>85.59</td>
<td>84.10</td>
<td>113.83</td>
</tr>
<tr>
<td>PCA</td>
<td>103.80</td>
<td>103.86</td>
<td>103.36</td>
<td>118.57</td>
</tr>
<tr>
<td>SSDR</td>
<td>131.27</td>
<td>131.26</td>
<td>131.26</td>
<td>121.13</td>
</tr>
<tr>
<td>RBM</td>
<td>91.72</td>
<td>91.33</td>
<td>89.12</td>
<td>111.64</td>
</tr>
</tbody>
</table>

Table 4.6: Prediction for behavior score in hippocampal surfaces data: $\hat{Y}_{\text{number}}$ is the prediction value, where the subscript number is the reduced feature dimension. Sums of prediction errors for $\hat{Y}_{1000}$, $\hat{Y}_{700}$ and $\hat{Y}_{500}$ were 117.3, 118.07 and 118.01 for 180 test observations. Correlation thresholding values were 0.3, 0.5 and 0.7. Estimated smoothing values for them were 0.02, 0.03 and 0.04, respectively.

<table>
<thead>
<tr>
<th>ID</th>
<th>$Y_{\text{true}}$</th>
<th>$Y_{1000}$</th>
<th>$Y_{700}$</th>
<th>$Y_{500}$</th>
<th>Age</th>
<th>Status</th>
<th>Gender</th>
<th>AP1</th>
<th>AP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>270</td>
<td>27.33</td>
<td>27.934</td>
<td>27.746</td>
<td>27.559</td>
<td>85.2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>268</td>
<td>13.33</td>
<td>14.895</td>
<td>14.733</td>
<td>15.238</td>
<td>82.8</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>265</td>
<td>25</td>
<td>24.378</td>
<td>24.306</td>
<td>24.033</td>
<td>66</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>309</td>
<td>17</td>
<td>17.754</td>
<td>17.979</td>
<td>18.341</td>
<td>65.1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>318</td>
<td>30.67</td>
<td>32.029</td>
<td>32.237</td>
<td>31.742</td>
<td>88.2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>283</td>
<td>20</td>
<td>18.348</td>
<td>18.631</td>
<td>19.043</td>
<td>80.1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>300</td>
<td>23</td>
<td>22.772</td>
<td>23.408</td>
<td>23.849</td>
<td>78</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>304</td>
<td>24</td>
<td>22.873</td>
<td>23.128</td>
<td>22.894</td>
<td>80.1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>307</td>
<td>17.67</td>
<td>16.163</td>
<td>16.126</td>
<td>16.267</td>
<td>76.2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>312</td>
<td>17.33</td>
<td>17.484</td>
<td>17.532</td>
<td>17.42</td>
<td>72.3</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>280</td>
<td>20.67</td>
<td>22.537</td>
<td>20.688</td>
<td>20.092</td>
<td>72</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>302</td>
<td>17.67</td>
<td>17.947</td>
<td>17.425</td>
<td>17.506</td>
<td>80.1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>345</td>
<td>10.33</td>
<td>10.132</td>
<td>11.517</td>
<td>12.233</td>
<td>80.7</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>343</td>
<td>3.67</td>
<td>4.035</td>
<td>3.502</td>
<td>3.977</td>
<td>73.8</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>337</td>
<td>7.33</td>
<td>6.809</td>
<td>7.012</td>
<td>7.758</td>
<td>72.6</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>361</td>
<td>9.67</td>
<td>10.466</td>
<td>10.794</td>
<td>10.992</td>
<td>85.8</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>344</td>
<td>13</td>
<td>14.677</td>
<td>14.689</td>
<td>14.61</td>
<td>70.8</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>364</td>
<td>11</td>
<td>11.557</td>
<td>11.718</td>
<td>11.742</td>
<td>70.8</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>401</td>
<td>0.67</td>
<td>1.12</td>
<td>1.175</td>
<td>1.082</td>
<td>74.1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>380</td>
<td>7</td>
<td>6.866</td>
<td>5.476</td>
<td>5.961</td>
<td>72.9</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>353</td>
<td>8.67</td>
<td>8.123</td>
<td>7.431</td>
<td>7</td>
<td>83.7</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>342</td>
<td>7</td>
<td>6.099</td>
<td>5.075</td>
<td>5.224</td>
<td>76.8</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>349</td>
<td>5</td>
<td>4.22</td>
<td>4.007</td>
<td>3.946</td>
<td>76.8</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>335</td>
<td>3.33</td>
<td>4.001</td>
<td>3.287</td>
<td>2.823</td>
<td>73.5</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
features selected from Stage (I). We used the Gaussian RBF kernel \( \exp(-\|x_1 - x_2\|^2/\sigma) \) with \( \sigma = p \), the regularization parameter \( \lambda \) at \( \lambda = 0.001(n^{-1/4}) \). For other prediction methods, we used an optimized tuning parameter that minimizes a cross-validation error and for the choice of dimension, we used \( p_z \) for PLS, PCA, and RBM (Restricted Boltzmann Machine) methods. Table 4.5 presents the sum of prediction errors for all methods. The SILFM methods based on the kernel machine methods dramatically improve the prediction accuracy. The last column in Table 4.5 showed the prediction results when we used only demographic covariates such as diagnostic status, gender, age and APOE4. Compared the first three columns to the last column, we see that there is the numerically additional contribution of using the imaging data for this analysis. Table 4.6 presents the true behavioral score, clinical covariates, and prediction value of 24 randomly selected subjects. For instance, we consider two subjects with ID270 and ID268, which have the same clinical covariates, but different behavioral scores. Our prediction model is able to recover most differences between their behavioral scores, indicating that imaging data do add important predictive value to the prediction problem considered here. Figures 4.5 (A)
and (B) show the correlation matrices of the 1000 top radial distances selected from HSIC-SIS. For a better illustration, we also thresholded the correlation matrix at the level of 0.5 and 0.7, respectively (Figure 4.5 (C) and (D)). It is clear that most selected features are highly correlated with each other.

4.5 Asymptotic Analysis

We investigate several theoretical properties of SILFM including sure independence screening property, $\sqrt{n}$ consistency, and risk bound for SILFM. For simplicity, it is assumed that $\{(x_i, y_i) : i = 1, \ldots, n\}$ are independent and identically distributed. The following conditions are used to facilitate the technical details. Although they may not be the weakest conditions, they do help to simplify the proof. We have stated the following theorems, whose detailed proofs can be found in a supplementary document.

First, we will show that the features selected by Stages (I) enjoy the sure screening property under some general conditions. Let $K(\cdot, \cdot)$ be a reproducing kernel on each marginal feature space and $L(\cdot, \cdot)$ is a reproducing kernel defined on the output space $\mathcal{Y}$. We define $\beta_j$ and $\hat{\beta}_j$ as $\text{HSIC}(X_j, Y)$ and $\hat{\text{HSIC}}(X_j, Y)$, respectively.

- **Condition 1.** Kernels $K(\cdot, \cdot)$ and $L(\cdot, \cdot)$ are measurable, bounded, continuous, and positive definite kernels.

- **Condition 2.** The $K(\cdot, \cdot) \otimes L(\cdot, \cdot)$ is dense in $L^2(\mathbb{P})$ for all probability distributions $\mathbb{P}$ on $(\mathcal{X}_j, \mathcal{Y})$.

- **Condition 3.** $\min_{j \in \mathcal{M}} |\beta_j| \geq c_0 n^{-\kappa}$ for some $0 < \kappa < 1/2$ and $c_0 > 0$.

Condition 1 is a condition to define a standard RKHS on both the feature and output spaces. Condition 2 is to ensure the existence of a characteristic kernel in $L^2(\mathbb{P})$, which allows us to define a cross covariance operator on $\mathcal{H}_x \otimes \mathcal{H}_y$. Condition 3 is analogous to Condition 3 in Fan and Lv (2008), where $\kappa$ controls the rate of probability error in recovering the true sparse model.
Theorem 4.5.1. Let $\mathcal{F}_j = \{f_j \in \mathcal{H}_{xj} \mid \|f_j\|_{\mathcal{H}_{xj}} \leq 1\}$ and $\mathcal{G} = \{g \in \mathcal{H}_y \mid \|g\|_{\mathcal{H}_y} \leq 1\}$ be functional classes with the unit ball in an RKHS on each marginal and outcome domain. Under Conditions 1-2, for any positive constant $c_1$, there exists a constant $c_2$ such that

$$\mathbb{P}(\max_{1 \leq j \leq p_x} \max_{f_j \in \mathcal{F}_j, g \in \mathcal{G}} |\hat{\beta}_j - \beta_j| \geq c_1 n^{-\kappa}) \leq p_x \exp(-2c_2 n^{1-2\kappa}).$$

(4.10)

If Condition 3 also holds, then taking $r_n = c_3 n^{-\kappa}$ with $c_3 \leq c_0/2$ leads to

$$\mathbb{P}(\mathcal{M} \subset \mathcal{M}_{r_n}) \overset{p}{\to} 1,$$

(4.11)

where $\overset{p}{\to}$ denotes convergence in probability.

Theorem 4.5.1 establishes the sure screening property of our screening method. Compared with the existing literature (Fan and Lv 2008, Fan and Song 2010), our results in Theorem 4.5.1 focus on a non-sparse circumstance ($p_x \gg \tilde{p}_x \gg n$) and thus the maximum dimension holds as $\log p_x/n^{1-2\kappa} \to 0$. Therefore, it is unnecessary to assume the standard tail probability condition for each feature.

Second, we establish the consistency of Stage (II) under a latent factor model given by

$$\tilde{x}_i = \mathbf{B}_{\tilde{f}_i} + \sigma_x e_i$$

(4.12)

where $\mathbf{B} = (b_{1}^{(1)}, \ldots, b_{M}^{(1)}, \ldots, b_{1}^{(G)}, \ldots, b_{M}^{(G)})$ is a $\tilde{p}_x \times GM$ matrix, $\tilde{f}_i = (f_{i1}^{(1)}, \ldots, f_{iM}^{(G)})^T$ is an $GM \times 1$ vector of latent factors and $f_{i}^{(r)} = (f_{i1}^{(r)}, \ldots, f_{iM}^{(r)})^T$ is an $M \times 1$ vector of latent factors corresponding to $r$th group feature. Under model (4.12), we assume the independence between $\tilde{f}_i$ and $e_i$ such that $\text{Cov}(\tilde{f}_i) = \mathbf{I}_{GM}$ and that the columns of $\mathbf{B}$ are orthogonal. Thus, $\text{Cov}(\tilde{x}_i)$ is given by $\Sigma = \mathbf{B}\mathbf{B}^T + \sigma_x^2 \mathbf{I}_{\tilde{p}_x}$. For simplicity, the components of $\tilde{f}_i$ and $e_i$ are assumed to be independently and identically distributed as $N(0,1)$.

Let $\Pi(\cdot) : \mathcal{M} \to \{1, \ldots, G\}$ be a clustering map such that $\mathcal{M} = \bigcup_{r=1}^{[G]} \mathcal{M}_r$ with $\mathcal{M}_r \cap \mathcal{M}_{r'} = \emptyset \ (r \neq r')$ and $\Sigma^{(r)}$ be a covariance matrix based on $\mathcal{M}_r$(the rth cluster). Let $\hat{\Pi}(\cdot) : \hat{\mathcal{M}} \to \hat{\mathcal{M}}_r.$
\{1, \ldots, G\} be a spectral clustering map such that \( \hat{\mathcal{M}} = \bigcup_{r=1}^{G} \hat{\mathcal{M}}_r \) with \( \hat{\mathcal{M}}_r \cap \hat{\mathcal{M}}_{r'} = \emptyset \ (r \neq r') \) \( S(r) \) be a covariance matrix based on \( \hat{\mathcal{M}}_r \). We denote \( \hat{\theta}_k^{(r)} \) as the \( k \)th eigenvector of \( S(r) \). Note that \( \hat{\theta}_k^{(r)} = b_k^{(r)}/\|b_k^{(r)}\|_2 \) as the \( k \)th eigenvector corresponding to the \( k \)th eigenvalue denoted by \( \lambda_k^{(r)} = \|b_k^{(r)}\|^2 \) for \( \Sigma^{(r)} \) for \( k = 1, \ldots, M \) and \( r = 1, \ldots, G \) where it follows from Fan et al. (2015). The following conditions are needed:

- **Condition 4.** \( P(\hat{\Pi}^{(\hat{\mathcal{M}})} \neq \Pi(\mathcal{M})) \to^p 0 \) for any feature \( j = 1, \ldots, \tilde{p}_x \).
- **Condition 5.** \( \frac{\lambda_k^{(r)}}{\lambda_1^{(r)}} \to \rho_k \) for all \( k \) and \( r \) with \( 1 = \rho_1 > \rho_2 > \ldots > \rho_M > 0 \) as \( n \to \infty \).
- **Condition 6.** for any \( r \), \( \min(\tilde{p}_x, \lambda_1^{(r)}) \to \infty \) and \( \frac{\tilde{p}_x(\lambda_k^{(r)}+1)}{n(\lambda_1^{(r)})^2} \to 0 \) as \( n \to \infty \).

Condition 4 is required for establishing consistency in the Stage (II) setting. It is stated that the clustering consistency should hold for all true group features and it also implies that the consistency of the selected features for all important features holds. Condition 5 is an asymptotic identifiability condition such that significant eigenvalues are separated at the ratio of the first leading eigenvalue. Condition 6 follows from Paul (2005), which is required for controlling the rate of convergence of the high dimensional principal components to zero in the sense of \( \| \cdot \|_2 \).

From Condition 6, Stage (II) includes even high dimensional principal components as a special case for the single cluster.

**Theorem 4.5.2.** Under the Condition 4, suppose that \( \tilde{x}_1, \ldots, \tilde{x}_n \) follow factor model (4.12) and all the eigenvalues \( \lambda_k^{(r)} \) of \( \Sigma \) satisfy Conditions 5-6. Let \( \hat{\theta}_k^{(r)} \) be the \( k \)th principal component of \( S(r) \) and \( L(\hat{\theta}_k^{(r)}) = \|\hat{\theta}_k^{(r)} - \theta_k\|_2 \) be a loss function. The convergence rate of the risk function for \( \hat{\theta}_k^{(r)} \) under the given loss function with respect to the probability measure \( P \) on \( \tilde{X} \), \( \mathcal{R}_{L,P}(\hat{\theta}_k^{(r)}) = E\{L(\hat{\theta}_k^{(r)})\} \) is given by

\[
\mathcal{R}_{L,P}(\hat{\theta}_k^{(r)}) = \frac{|\mathcal{M}_r| - M(1 + \lambda_k^{(r)})}{n(\lambda_1^{(r)})^2} + \frac{\sum_{m \neq k} \sigma_x^2(\lambda_m^{(r)} + \lambda_k^{(r)}) + (\sigma_x^4 + \lambda_m^{(r)} \lambda_k^{(r)})}{n(\lambda_1^{(r)} - \lambda_k^{(r)})^2} \tag{4.13}
\]

as \( \sqrt{\log |\mathcal{M}_r|} \to 0 \).
where for some \(0 < \kappa < 1\).

Theorem 4.5.2 establishes that the linear operator derived as the principal component in Stages (II) can be consistent under some conditions. Note that the property of \(\hat{\theta}_k^{(r)}\) would be close to the property of \(\theta_k\) where \(\theta_k\) is the \(k\)th eigenvector corresponding to the \(k\)th eigenvalue for \(\Sigma\) as \(|\mathcal{M}_r| \to \tilde{p}_x\). The result of Theorem 4.13 jointly with Condition 6 shows that we have the risk bound even for the single cluster case.

Third, we investigate the consistency of the estimator of Stages (III) and the prediction performance of SILFM. We need to introduce some notation as follows. For a given measurable space \((\mathcal{X}, \mathcal{A}, P_X)\), let \(L_0(\mathcal{X})\) be the set of all real measurable functions on \(X \in \mathbb{R}^d\) and \(L_\infty(\mathcal{X})\) be the set of all real measurable and bounded functions, i.e., \(L_\infty(\mathcal{X}) = \{f \in L_0(\mathcal{X}) : \|f\|_\infty < \infty\}\), where \(\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|\). Let us define \(f_\theta(\cdot) = f(\theta^T \cdot)\) as \(f_\theta : \mathcal{X} \to \mathbb{R}\) for a fixed \(\theta\) and \(\mathcal{F}(\theta)\) be a functional class with \(\mathcal{F}(\theta) = \{f_\theta \in L_\infty(\mathcal{X}) : \|f(\theta^T \cdot)\|_\infty < \infty\}\). According to the representer theorem, our estimator in Stage (III) is obtained by finding a minimizer \(\hat{f}_{n, \lambda, \hat{\theta}}\) of the regularized empirical risk under the squared loss function \(L\)

\[
\mathcal{R}_{L,n,\lambda,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) = \lambda \|\hat{f}_{n,\lambda,\hat{\theta}}\|^2 + \mathcal{R}_{L,n,\lambda,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) = \inf_{f \in \mathcal{F}(\hat{\theta})} \lambda \|f\|^2 + \mathcal{R}_{L,n,\lambda,\hat{\theta}}(f)
\]

where \(\mathcal{R}_{L,n,\lambda,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) = n^{-1} \sum_{i=1}^n L(y_i, \hat{f}_{n,\lambda,\hat{\theta}}(x_i)) = n^{-1} \sum_{i=1}^n (y_i - \hat{f}_{n,\lambda,\hat{\theta}}(\hat{z}_i))^2\) and \(\hat{z}_i = \hat{\theta}^T x_i\). To delineate the property of \(\hat{f}_{n,\lambda,\hat{\theta}}\), we consider its limit \(f_{P,\lambda,\theta}\) as the solution of a risk function given by

\[
\mathcal{R}_{L,P,\lambda,\theta}(f_{P,\lambda,\theta}) = \lambda \|f_{P,\lambda,\theta}\|^2 + \mathcal{R}_{L,P,\lambda,\theta}(f_{P,\lambda,\theta}) = \inf_{f \in \mathcal{F}(\theta)} \lambda \|f\|^2 + \mathcal{R}_{L,P,\lambda,\theta}(f),
\]

where \(\mathcal{R}_{L,P,\theta}(f_{P,\lambda,\theta}) = \mathbb{E}\{L(y_i, f_{n,\lambda,\theta}(x_i))\} = \mathbb{E}\{y_i - f_{n,\lambda,\theta}(z_i)\}^2\), in which \(z_i = \theta^T x_i\). For classes \(\mathcal{F}(\theta)\) and \(\mathcal{F}(\hat{\theta})\), we define minimizers \(f_{P,\theta}^*\) and \(f_{P,\hat{\theta}}^0\) as follows:

\[
\mathcal{R}_{L,P,\theta}^* = \mathcal{R}_{L,P,\theta}(f_{P,\theta}^*) = \inf_{f \in \mathcal{F}(\theta)} \mathcal{R}_{L,P,\theta}(f) \quad \text{and} \quad \mathcal{R}_{L,P,\hat{\theta}}^* = \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\hat{\theta}}^0) = \inf_{f \in \mathcal{F}(\theta)} \mathcal{R}_{L,P,\hat{\theta}}(f).
\]
The following conditions are needed:

- **Condition 7.** For any \( f \in \mathcal{L}_\infty(P_X) \), and \( \theta_1 \) and \( \theta_2 \in \mathbb{R}^d \) with \( \max\{\|\theta_1\|, \|\theta_2\|\} < \infty \), there exists a constant \( |L|_2 \) such that \( |\mathcal{R}_{L,P,\theta_1}(f_{\theta_1}) - \mathcal{R}_{L,P,\theta_2}(f_{\theta_2})| \leq |L|_2\|\theta_1 - \theta_2\|_2 \).

- **Condition 8.** For any \( f \in \mathcal{L}_\infty(P_X) \), and \( \theta_1 \) and \( \theta_2 \in \mathbb{R}^d \) with \( \max\{\|\theta_1\|, \|\theta_2\|\} < \infty \), there exists a constant \( |L|_3 \) such that \( |f_{\theta_1}(x) - f_{\theta_2}(x)| \leq |L|_3\|\theta_1 - \theta_2\|_2 \) for any \( x \).

- **Condition 9.** There exists a constant \( M > 0 \) such that \( \mathcal{R}_{L,P}(0) = E\{L(y_i, 0)\} \leq M \).

- **Condition 10.** There is \( c > 0 \) such that \( \lambda\|f_{P,\lambda,\theta}\|^2 + \mathcal{R}_{L,P,\theta}(f_{P,\lambda,\theta}) - \mathcal{R}_{L,P,\theta}(f_{P,\theta}^0) \leq c\lambda^\beta \).

Conditions 7-8 are Lipschitz conditions on the risk function \( \mathcal{R}_{L,P,\theta}(f_{\theta}) \) and \( f_{\theta}(\cdot) \). Condition 7 assumes that \( \mathcal{R}_{L,P,\cdot}(f_{\cdot}) : \Theta \rightarrow \mathbb{R} \) as a function of \( \theta \) is Lipschitz. Similarly, Conditions 8 assumes that \( f_{\cdot}(x) : \Theta \rightarrow \mathbb{R} \) is a Lipschitz function of \( \theta \) for each fixed \( x \). A sufficient condition of Condition 8 is that the reproducing kernel \( K((\cdot)^T x_i, (\cdot)^T x_j) : \Theta \rightarrow \mathbb{R} \) is a Lipschitz function of \( \theta \) for any \( x_i \) and \( x_j \). Conditions 9 and 10 assume that the second moment of the outcomes is bounded by a constant \( M \) and that the difference between the regularized risk and the minimum risk over \( \mathcal{F}(\theta) \) is bounded by \( c\lambda^\beta \).

**Theorem 4.5.3.** Let \( L(\cdot, \cdot) \) be a squared loss function and \( P \) be a distribution on \( \mathcal{X} \times \mathcal{Y} \). Let \( \mathcal{F} \subset L_\infty(X) \) be a non-empty and compact set. Moreover, let \( H \) be RKHS of a continuous kernel \( K \) on \( \mathcal{X} \) such that \( L_\infty(X) \subset H \) and \( K \) is \( m \) times continuously differentiable on \( \mathbb{R}^d \). Suppose Conditions 4-10 hold. Then for a fixed \( \tau > 0 \), the convergence rate of \( \hat{f}_{n,\lambda,\theta} \) is given by

\[
\lambda\|\hat{f}_{n,\lambda,\theta}\|^2 + \mathcal{R}_{L,P,\theta}(f_{n,\lambda,\theta}) - \mathcal{R}_{L,P,\theta}(f_{P,\theta}^*) \leq O_p(n^{-\frac{2m\beta}{(2m+d)(2\beta+1)}}) + O_p(\sqrt{\log |\mathcal{M}_r|}n^{-\kappa})
\]

as \( \|f_{P,\theta}^0 - f_{P,\theta}^*\| = o_p(1) \) and \( \lambda \rightarrow 0 \) where \( 0 < \kappa < 1 \).

Theorem 4.5.3 gives an integrated insight of various SILFM estimators on the prediction performance of SILFM. First, if \( d = O(p_x) \), then the SILFM estimator is at the rate of \( O_p(n^{-\frac{2m\beta}{(2m+d)(2\beta+1)}}) \). However, if \( d \) is relatively small, we can obtain a faster rate for the SILFM estimator, but we
have to pay the price for using the dimension reduction methods in Stages (I) and (II), which is \( O_p(\sqrt{\log |\mathcal{M}_r|n^{-\kappa}}) \). As \( \sqrt{\log |\mathcal{M}_r|n^{-\kappa}} \to 0 \), SILFM can truly achieve a better prediction risk, which has been justified by the simulations and real data analysis. This is one justification for using sequential estimators of SILFM under the challenging situation of ultrahigh dimensionality, highly correlated predictors and a complex functional relationship with the response.

### 4.6 Conclusion

We have developed a SILFM framework to build an accurate prediction model for clinical outcomes based on a massive number of features. SILFM as a three-stage estimation procedure integrating an independent screening method, a latent factor model, and kernel ridge regression. Theoretically, we have established several theoretical properties of SILFM, such as risk bound and selection consistency. Our simulation results and real data analysis show that SILFM outperforms many state-of-the-art methods in terms of prediction accuracy.
CHAPTER 5: DISCUSSION

There may be some contributions of this paper. Firstly, we propose general type of independent screening statistic which can detect both linear and nonlinear relationship. Since it does not require strong assumptions, we expect that it would be an useful technique in variable selection problem. Secondly, it is possible to capture the complex local information among covariates by employing the local dimension reduction. We believe that it would be the novel conception to detect the correlation structure. Also, we asymptotically study for the sure screening property and the risk bound under the ultrahigh dimensional circumstance. Extensive simulations show our method outperforms the other state-of-the art method.

In terms of the high dimensional application, we have been interested in reducing the high dimensionality of the covariates to a few significant covariates related to the phenotype outcome on RKHS. Among many methods, we focused on sure independent screening (SIS) property where all important variables survive after applying a variable screening procedure with high probability and marginal correlation learning is well known method to do this. We proposed SIS procedure with universal positive definite kernel based on adopting Hilbert Schmidt independent criteria (HSIC) statistics on RKHS. Parameterizing the association between covariate and outcome as HSIC works because HSIC could detect not only linear relationship but also nonlinear relation. The novel SIS has successfully shown its performance in this dissertation.

As the future research, we positively expect the extension of various SIS methods with kernel on RKHS such as longitudinal and survival issues.
APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2

Suppose that the rows of $X$ are iid. Then we can formulate a population model as follows. Denoting the rows by $x_i' (i = 1, \ldots, n)$, We have the model

$$x_i' \sim N_p (\mu, \Sigma),$$

where $\Sigma (p \times p)$ is the covariance matrix. Without loss of generality, we assume that $\mu = 0$.

Suppose that the spectral decomposition of $\Sigma$ is given by $\Sigma = \sum_{i=1}^{p} \theta_i \theta_i'$, where $l_1 \geq \ldots \geq l_p \geq 0$ and $\theta_1, \ldots, \theta_p$ from an orthonormal basis of $\mathbb{R}^p$.

Suppose further that there exists an $M \geq 1$ such that

$$l_k = \lambda_k + \sigma_x^2, \quad k = 1, \ldots, M, \quad \text{and} \quad l_k = \sigma_x^2, \quad k = M + 1, \ldots, p,$$

where $\lambda_1 \geq \ldots \geq \lambda_M > 0$ and $\sigma_x > 0$.

We can equivalently express the predictors through the following factor analysis model. Note that under the Gaussian assumption the matrix $x_j'$ can be expressed as

$$x_j' = \sum_{i=1}^{M} \sqrt{\lambda_i} u_i \theta_i' + \sigma_x E$$

where $u_1, \ldots, u_M$ are i.i.d. $N_n (0, I)$ vectors, and $E$ is an $N \times p$ matrix, all entries are i.i.d. $N(0, 1)$ and is independent of $x_1, \ldots, x_M$.

Suppose the the following response model for $y$. Notice that we assume that $y$ has mean 0.

$$y = \sum_{i=1}^{K} \delta_i v_i + \sigma_y Z$$

Where $\sigma_y > 0$, $1 \leq K \leq M$, and $Z$ has $N_n (0, I)$ distribution and is independent of $X$.

Throughout the paper, the following assumptions are needed to facilitate the technical details.

A.1. The eigenvalues are such that $\frac{\lambda_k}{\lambda_1} \to \rho_k$ for $k = 1, \ldots, M$ with $1 = \rho_1 > \rho_2 > \ldots > \rho_M > 0$ and $\lambda_1 \to c > 0$ as $n \to \infty$. Moreover, $\sigma_x^2 \to \sigma^2 \in [0, \infty)$ as $n \to \infty$.

A.2. $p$ varies with $n$ in such a way that $\frac{\sigma_x^2 p}{n \lambda_1} \to 0$ as $n \to \infty$. 

111
A.3. For each \( j \in M = \{ j : \beta_j \neq 0 \} \) and for any \( b > 2 \), there exist an sequence \( \sqrt{b \log n} \) such that
\[
\frac{U_j(0)^2}{I_j(0)} \geq \sqrt{\frac{b \log n}{n}}.
\]

Note that A.1 is called an asymptotic identifiability condition for orthonormal basis \( \theta_1, \ldots, \theta_M \), that the the condition \( \lambda_1 \to c > 0 \), taken together with the first part of condition A.1, implies that all of the \( M \) eigenvalues \( \lambda_k \) converge to positive limits and that conditions A.1 and A.2 allow for the possibility that \( \frac{\lambda_1}{\sigma^2} \to \infty \) and \( \frac{p}{n} \) converges to a positive limit.

**Lemma A.0.1.** Suppose that the eigenvalues \( \lambda_1, \ldots, \lambda_M \) under given assumptions. If \( \log(\max(n, p)) = o(\min(n, p)) \), then for \( v = 1, \ldots, M \),
\[
\sup_{\theta} EL(\hat{\theta}_v, \theta_v) = \left[ \frac{p - M}{n h(\lambda_v)} + \frac{1}{n} \sum_{i \neq v} \frac{(\lambda_v + 1)(\lambda_i + 1)}{(\lambda_v - \lambda_i)^2} \right] (1 + o(1))
\]
where \( h(\lambda) = \frac{\lambda^2}{1+\lambda} \)

**Proof.** It follows from the results provided by Paul (2005).

**Lemma A.0.2.** For some \( T \in \mathbb{N} \), let \( A \) and \( B \) be two symmetric \( T \times T \) matrices. Set \( \lambda_0 = \infty \) and \( \lambda_{T+1} = -\infty \). We then have the following:

a. For all \( \gamma_1, \gamma_2 \in \{1, \ldots, T\} \) such that \( \lambda_{\gamma_1-1}(A) > \lambda_{\gamma_1}(A) = \lambda_{\gamma_1+1}(A) = \ldots = \lambda_{\gamma_2}(A) > \lambda_{\gamma_2+1}(A) \), we obtain
\[
\sum_{s=\gamma_1}^{\gamma_2} (\lambda_s(A + B) - \lambda_s(A)) = tr(P_{\gamma_1}(A)B) + R_1,
\]
where \( R_1 \) can be bounded by
\[
|R_1| \leq \min\{\gamma_2 - \gamma_1 + 1, T - \gamma_2 + \gamma_1 - 1\} \frac{6||B||^2}{\min_{\lambda \in EG(A), \lambda \neq \gamma_1} |\lambda - \lambda_{\gamma_1}|}.
\]
b. If $\lambda_\gamma(A)$ is a unique root of $A$ then,

$$p_\gamma(A + B) - p_\gamma(A) = -S_\gamma(A)BP_\gamma(A) + R_2,$$

where $\gamma(A) = \sum_{s \neq \gamma} \frac{1}{\lambda_{s} - \lambda_\gamma}P_{s}(A)$, and $R_2$ can be bounded by

$$\|R_2\|_2 \leq \frac{6\|B\|_2^2}{\min_{\lambda \in \text{EG}(A), \lambda \neq \lambda_\gamma} |\lambda - \lambda_\gamma|^2}.$$

**Proof.** It follows from the results provided by Kneip and Utikal (2001). $\square$

These results are useful to prove the results in Section 3. Lemma A.0.1 indicates that our sample version of principle components converges to the true principle components. The angle distance was used as the metric given by Paul (2005) paper. Lemma A.0.2 gives the bound for the perturbation of the covariance matrix in terms of orthonormal basis and its eigenvalue. Based on this result, we can obtain the bound for the difference between the eigenvalue of sample covariance matrix and the eigenvalue of true covariance matrix.

Suppose that the SVD of $X$ is given by

$$X = U \Lambda \Theta'$$

where $U$ is $N \times M$, is $M \times M$, and $\Theta$ is $p \times M$, with $M = \min(n, p)$. $n$ is the number of observations and $p$ is the number of covariates. Let $u_1, \ldots, u_M$ be the columns of $U$, $\theta_1, \ldots, \theta_M$ be the columns of $\Theta$, and $l_1, \ldots, l_M$ be the diagonal elements of $L$. Actually, all these quantities are obtained from the sample, $X$ through the SVD. Thus, we denote $\hat{u}_1, \ldots, \hat{u}_M, \hat{l}_1, \ldots, \hat{l}_M$ and $\hat{\theta}_1, \ldots, \hat{\theta}_M$ as estimators for each population quantities.

We denoted the singular value, $l_k$ ($k = 1, 2, \ldots, M$) as the PCA-based estimators of the $k$-th largest eigenvalue of $\Sigma$. The corresponding population quantity is $l_k = \lambda_k + \sigma_x^2$. If $\sigma_x^2$ is known, a natural estimator of $\lambda$ is $\hat{\lambda}_k = \max\{\hat{l}_k - \sigma_x^2, 0\}$. However, if $\sigma_x^2$ is unknown, we will estimate it, based on various strategies. One strategy is to use the median of the diagonal elements of
\( \frac{1}{n} \mathbf{X}' \mathbf{X}_1 \) as a estimate of \( \sigma_x^2 \) and then estimate of \( \tilde{\lambda}_k = \max \{ \hat{l}_k - \sigma_x^2, 0 \} \).

We will assume that either \( \sigma_x^2 \) is known or we already had a consistent estimator for \( \sigma_x^2 \) in this section.

Proof of Theorem 2.5.1. By the singular value decomposition of \( \mathbf{X} \),

\[
\hat{u}_k = \frac{1}{\sqrt{\hat{l}_k}} \mathbf{X} \hat{\theta}_k \\
= \frac{1}{\sqrt{\hat{l}_k}} \left[ \sum_{i=1}^{M} \sqrt{\lambda}_i \mathbf{u}_i \hat{\theta}_i' + \sigma_x \mathbf{E} \right] \hat{\theta}_k \\
= \frac{1}{\sqrt{\hat{l}_k}} \sum_{i=1}^{M} \sqrt{\lambda}_i \langle \hat{\theta}_i, \hat{\theta}_k \rangle \mathbf{u}_i + \frac{\sigma_x}{\sqrt{\hat{l}_k}} \begin{bmatrix} \epsilon'_1 \\ \vdots \\ \epsilon'_n \end{bmatrix} \hat{\theta}_k.
\]

Consider the quantity \( \langle \hat{u}_k, \mathbf{y} \rangle \),

\[
\langle \hat{u}_k, \mathbf{y} \rangle = \frac{1}{\sqrt{\hat{l}_k}} \sum_{i=1}^{M} \sum_{j=1}^{K} \sqrt{\lambda}_i \langle \hat{\theta}_i, \hat{\theta}_k \rangle \langle \mathbf{u}_i, \delta_j \mathbf{u}_j \rangle + \frac{\sigma_x}{\sqrt{\hat{l}_k}} \hat{\theta}_k' [\epsilon_1 \ldots \epsilon_n] [\sum_{j=1}^{K} \delta_j \mathbf{u}_j] \\
+ \frac{\sigma_y}{\sqrt{\hat{l}_k}} \sum_{i=1}^{M} \sqrt{\lambda}_i \langle \hat{\theta}_i, \hat{\theta}_k \rangle \langle \mathbf{u}_i, \mathbf{Z} \rangle + \frac{\sigma_x \sigma_y}{\sqrt{\hat{l}_k}} \hat{\theta}_k' [\epsilon_1 \ldots \epsilon_n] \mathbf{Z}.
\]

For \( \hat{\theta}_k \) by Lemma A.0.1 and for \( \hat{\theta}_k \) by the orthonormality, \( \| \hat{\theta}_k \|_2^2 = 1 \), we obtain that \( \hat{\theta}_k = \theta_k + O_p(\sqrt{\frac{p}{N}}) \) and \( \hat{\theta}_k' \epsilon_n \equiv \tilde{z}_k \) has \( N(0, 1) \) distribution. Thus,

\[
\langle \hat{u}_k, \mathbf{y} \rangle = \frac{\sqrt{\lambda}_k}{\sqrt{\hat{l}_k}} \sum_{j=1}^{K} \delta_j \langle \mathbf{u}_k, \mathbf{u}_j \rangle + \frac{\sigma_x}{\sqrt{\hat{l}_k}} [\tilde{z}_1 \ldots \tilde{z}_n] \left[ \sum_{j=1}^{K} \delta_j \mathbf{u}_j \right] \\
+ \frac{\sigma_y}{\sqrt{\hat{l}_k}} \langle \mathbf{u}_k, \mathbf{Z} \rangle + \frac{\sigma_x \sigma_y}{\sqrt{\hat{l}_k}} [\tilde{z}_1 \ldots \tilde{z}_n] \mathbf{Z}
\]
By Lemma 2, we have \( \hat{\lambda}_k = \lambda_k + O_p\left(\frac{p}{N}\right) \) and by given assumptions we have

\[
\frac{\lambda_k}{l_k} = \frac{1}{\lambda_k + \sigma_x^2} = \frac{1}{\lambda_k + \sigma_x^2} = \frac{1}{1 + O_p\left(\frac{p}{N}\right)} \xrightarrow{a.s.} 1
\]

\[
\frac{\sigma_x^2}{l_k} = \frac{1}{\sqrt{l_k}} + 1 \xrightarrow{a.s.} 0.
\]

Thus, we have

\[
\langle \hat{u}_k, y \rangle = \delta_k \langle u_k, u_k \rangle + \sum_{j=1}^{K-1} \delta_j \langle u_k, u_j \rangle + \frac{\sigma_y}{\sqrt{l_k}} \langle u_k, Z \rangle.
\]

By normalizing \( n \), we have again,

\[
\frac{\langle \hat{u}_k, y \rangle}{n} = \delta_k \frac{1}{n} \sum_{i=1}^{n} u_{ki} u_{ki} + \frac{\sigma_y}{\sqrt{l_k}} \frac{1}{n} \sum_{i=1}^{n} u_{ki} z_i + \frac{1}{n} \sum_{j=1}^{K-1} \delta_j \sum_{i=1}^{n} u_{ji} u_{ki}.
\]

Since \( \frac{1}{n} \sum_{i=1}^{n} u_{ki} u_{ki} \xrightarrow{p} E[u_{ki} u_{ki}] = 1 \), \( \frac{1}{n} \sum_{i=1}^{n} u_{ji} u_{ki} \xrightarrow{p} 0 \), and \( \frac{1}{n} \sum_{i=1}^{n} u_{ki} z_i \xrightarrow{p} 0 \) as \( n \to \infty \),

we obtain

\[
\frac{\langle \hat{u}_k, y \rangle}{n} \xrightarrow{p} \delta_k
\]

Thus, it is consistency estimator.

**Proof of Theorem 2.5.2.** Consider \( \langle \hat{u}_k, \hat{u}_k \rangle \) and \( \langle \hat{u}_k, \hat{u}_j \rangle \) under given models and assumptions.

\[
\langle \hat{u}_k, \hat{u}_k \rangle = \left[ \frac{1}{\sqrt{l_k}} \sum_{i=1}^{M} \sqrt{\lambda_i} \langle \theta_i, \hat{\theta}_k \rangle u_i' + \frac{\sigma_x}{\sqrt{l_k}} \hat{\theta}_k' \epsilon_1 \ldots \epsilon_n \right] 
\times \left[ \frac{1}{\sqrt{l_k}} \sum_{i=1}^{M} \sqrt{\lambda_i} \langle \theta_i, \hat{\theta}_k \rangle u_i + \frac{\sigma_x}{\sqrt{l_k}} \epsilon_1 \ldots \epsilon_n \hat{\theta}_k \right]
\]

By using orthonormality \( ||\hat{\theta}_k||^2 = 1 \), we obtain that \( \hat{\theta}_k' \epsilon_n \) has \( N(0, 1) \), denoting \( \hat{\theta}_k' \epsilon_n \) by \( \tilde{z}_k \). Also,
we already had the perturbation bound for eigenvector, thus

\[
\langle \hat{u}_k, \hat{u}_k \rangle = \left[ \frac{1}{\sqrt{l_k}} \sqrt{\lambda_i} u_k' + \frac{\sigma_x}{\sqrt{l_k}} [\hat{z}_1 \ldots \hat{z}_n] \right] \left[ \frac{1}{\sqrt{l_k}} \sqrt{\lambda_j} u_k + \frac{\sigma_x}{\sqrt{l_k}} \begin{bmatrix} \hat{z}_1' \\ \vdots \\ \hat{z}_n' \end{bmatrix} \right]
\]

\[
= \frac{\lambda_k}{l_k} \sum_{i=1}^n u_k^2 + 2 \frac{\sigma_x \sqrt{\lambda_k}}{l_k} [\hat{z}_1 \ldots \hat{z}_n] u_k + \frac{\sigma_x^2}{l_k} \sum_{i=1}^n \hat{z}_i^2
\]

The coefficient of the second term in right hand side is \( \frac{\sigma_x \sqrt{\lambda_k}}{l_k} = \frac{\sigma_x}{\sqrt{l_k}} \frac{\sqrt{\lambda_k}}{\sqrt{l_k}} \). Since \( \frac{\sqrt{\lambda_k}}{\sqrt{l_k}} \xrightarrow{a.s} 1 \) and \( \frac{\sigma_x}{\sqrt{l_k}} = \frac{\sigma_x}{\sqrt{\rho_k + \lambda_1}} \xrightarrow{a.s} 0 \), we get

\[
\frac{\langle \hat{u}_k, \hat{u}_k \rangle}{n} \xrightarrow{p} 1.
\]

Similarly, \( \hat{z}_j (\hat{\theta} \epsilon_n) \) has \( N(0, 1) \),

\[
\langle \hat{u}_k, \hat{u}_j \rangle = \frac{\sqrt{\lambda_k \lambda_j}}{\sqrt{l_k l_j}} \langle u_k, u_j \rangle + \frac{\sigma_x}{\sqrt{l_k l_j}} \left[ \sqrt{\lambda_j} \sum_{i=1}^n \hat{z}_i u_{ji} + \sqrt{\lambda_k} \sum_{i=1}^n \hat{z}_i u_{ki} \right]
\]

\[
+ \frac{\sigma_x^2}{\sqrt{l_k l_j}} \sum_{i=1}^n \hat{z}_i \hat{z}_i
\]

we get

\[
\frac{\langle \hat{u}_k, \hat{u}_j \rangle}{n} \xrightarrow{p} 0.
\]

The objective function is given by

\[
f(\phi) = \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^M u_{ij} \phi_j)^2 + \gamma \sum_{j=1}^M |\phi_j|
\]

which is minimized at \( \phi = \hat{\delta} \).
Rewrite the objective function in vector and matrix notation

\[
\frac{1}{n} \|y - \hat{U} \phi \|^2 = \frac{1}{n} y' y - \frac{2}{n} \phi' \hat{U}' y + \frac{1}{n} \phi' \hat{U}' \hat{U} \phi.
\]

Since

\[
\frac{1}{n} y' y \xrightarrow{p} \delta' \delta + \sigma_y^2,
\]

and

\[
\frac{1}{n} \hat{U}' y = \frac{1}{n} \begin{bmatrix} \hat{u}_1' \\
\vdots \\
\hat{u}_k'
\end{bmatrix} y \xrightarrow{p} \begin{bmatrix} \delta_1' \\
\vdots \\
\delta_k'
\end{bmatrix} = \delta,
\]

and

\[
\frac{1}{n} \hat{U}' \hat{U} \xrightarrow{p} I
\]

Thus, we can define convex function

\[
F(\phi) = \delta' \delta + \sigma_y^2 - 2\phi' \delta + \phi' \phi + \gamma_0 \|\phi\|_1
\]

\[
= (\phi - \delta)'(\phi - \delta) + \gamma_0 \|\phi\|_1.
\]

Thus,

\[
\frac{1}{n} \|y - \hat{U} \phi \|^2 + \frac{\gamma_n}{n} \|\phi\|_1 \xrightarrow{p} (\phi - \delta)'(\phi - \delta) + \gamma_0 \|\phi\|_1
\]

If \( \gamma_0 = 0 \) then minimizer \( \hat{\delta} \xrightarrow{p} \phi \). Therefore, we can say \( \hat{\delta}^{lasso} \) would be consistency estimator under some condition.

Proof of Theorem 2.5.3 Let \( \tilde{e} \) be \( y - \hat{U} \delta \). Then

\[
\tilde{e} = U \delta + \sigma_y Z - [\hat{u}_1, \ldots, \hat{u}_M] \delta
\]

\[
= (u_1 - \hat{u}_1) \delta_1 + \ldots + (u_k - \hat{u}_k) \delta_k + \sigma_y Z
\]
where $U$ is $n \times K$, that is being involved with the response $y$, $Z$ is $n \times 1$ and $\delta$ is $K \times 1$. Define the function,

$$f_n(w) = (y - \hat{U}(\delta + \frac{w}{\sqrt{n}}))' (y - \hat{U}(\delta + \frac{w}{\sqrt{n}})) - \tilde{e}'\tilde{e} + \gamma_n \sum_{j=1}^{K} [||\delta_j + \frac{w_j}{\sqrt{n}}|| - |\delta_j|]$$

Clearly, we can know the quantity $\sqrt{n}(\hat{\delta} - \delta)$ is minimizer of $f_n(w)$. We have

$$\langle \hat{u}_k, u_k \rangle = \frac{1}{\sqrt{l_k}} \sum_{i=1}^{M} \sqrt{\lambda_i} \langle \theta_i, \hat{\theta}_k \rangle \langle u_k, u_k \rangle + \frac{\sigma_x}{\sqrt{l_k}} \hat{\theta}_k [\epsilon_1, \ldots, \epsilon_n] u_k$$

$$= \frac{\sqrt{l_k}}{\lambda_k} \sum_{i=1}^{n} u_{ki}^2 + \frac{\sigma_x}{\sqrt{l_k}} [\tilde{z}_1, \ldots, \tilde{z}_n] u_k$$

$$= \frac{\sqrt{l_k}}{\lambda_k} \sum_{i=1}^{n} u_{ki}^2 + \frac{\sigma_x}{\sqrt{l_k}} \sum_{i=1}^{n} \tilde{z}_i u_{ki}^2$$

and

$$\langle \hat{u}_k, u_j \rangle = \frac{1}{\sqrt{l_k}} \sum_{i=1}^{M} \sqrt{\lambda_i} \langle \theta_i, \hat{\theta}_k \rangle \langle u_i, u_j \rangle + \frac{\sigma_x}{\sqrt{l_k}} [\epsilon_1, \ldots, \epsilon_n] u_j$$

$$= \frac{\sqrt{l_k}}{\lambda_k} \sum_{i=1}^{n} u_{ki} u_{ji} + \frac{\sigma_x}{\sqrt{l_k}} \sum_{i=1}^{n} \tilde{z}_i u_{ji}^2$$

Remind

$$\langle \hat{u}_k, \hat{u}_k \rangle = \lambda_k \frac{1}{l_k} \sum_{i=1}^{n} u_{ki}^2 + 2 \frac{\sigma_x}{l_k} \sum_{i=1}^{n} \tilde{z}_i u_{ki} + \sigma_x^2 \frac{1}{l_k} \sum_{i=1}^{n} \tilde{z}_i^2$$

$$\langle \hat{u}_k, \hat{u}_j \rangle = \frac{\sqrt{\lambda_k \lambda_j}}{\sqrt{l_k l_j}} \sum_{i=1}^{n} u_{ki} u_{ji} + \frac{\sigma_x}{\sqrt{l_k l_j}} \left[ \sqrt{\lambda_j} \sum_{i=1}^{n} \tilde{z}_i u_{ji} + \sqrt{\lambda_k} \sum_{i=1}^{n} \tilde{z}_i u_{ki} \right] + \frac{\sigma_x^2}{\sqrt{l_k l_j}} \sum_{i=1}^{n} \tilde{z}_i \tilde{z}_i$$
\( \langle \hat{u}_k, y \rangle = \frac{\sqrt{\lambda_k}}{\sqrt{l_k}} \sum_{i=1}^{K} \delta_j \langle u_k, u_j \rangle + \frac{\sigma_x}{\sqrt{l_k}} [\bar{z}_1 \ldots \bar{z}_n] \sum_{j=1}^{K} \delta_j u_j + \frac{\sigma_y}{\sqrt{l_k}} \sqrt{\lambda_k} \sum_{i=1}^{n} u_{ki} \bar{z}_i + \frac{\sigma_y \sigma_x}{\sqrt{l_k}} \sum_{i=1}^{n} \bar{z}_i \bar{z}_i \)

where \( z_i \) is from the structure of \( y \).

We only consider terms associated with \( w \) in \( f_n(w) \)

\[
f_n(w) = -\frac{2}{\sqrt{n}} w' \hat{U}' \hat{\epsilon} + \frac{1}{\sqrt{n}} w' \hat{U}' \hat{U} w + \gamma_n \sum_{j=1}^{K} |\delta_j| + \frac{w_j}{\sqrt{n}} - |\delta_j| \]

Consider \( \frac{1}{\sqrt{n}} v_i \hat{\epsilon} \)

\[
\frac{1}{\sqrt{n}} v_i \hat{\epsilon} = \frac{1}{\sqrt{n}} \langle \hat{u}_i, y \rangle - \frac{1}{\sqrt{n}} \langle \langle \hat{u}_1, \hat{u}_1 \rangle, \ldots, \langle \hat{u}_1, \hat{u}_k \rangle \rangle \delta
\]

\[
= -\frac{1}{\sqrt{n}} \left( \frac{\sqrt{\lambda_1}}{\sqrt{l_1}} \sum_{j=1}^{K} \delta_j \langle u_1, u_j \rangle + \frac{\sigma_x}{\sqrt{l_1}} \sum_{j=1}^{K} \delta_j \sum_{i=1}^{n} u_{ji} \bar{z}_i + \frac{\sigma_y}{\sqrt{l_1}} \sqrt{\lambda_1} \sum_{i=1}^{n} u_{ii} \bar{z}_i + \frac{\sigma_y \sigma_x}{\sqrt{l_1}} \sum_{i=1}^{n} \bar{z}_i \bar{z}_i \right)
- \frac{1}{\sqrt{n}} \left( \frac{\sqrt{\lambda_1 \lambda_2}}{\sqrt{l_2 l_1}} \sum_{i=1}^{n} u_{i1} u_{ii} \right)
+ \frac{\sigma_x}{\sqrt{l_1 l_1}} \sum_{i=1}^{n} \bar{z}_i \bar{z}_i \delta_1 - \ldots
- \frac{1}{\sqrt{n}} \left( \frac{\lambda_2}{l_2} \sum_{i=1}^{n} u_{i1} u_{ii} \right)
+ \frac{\sigma_x}{\sqrt{l_2 l_1}} \sum_{i=1}^{n} \bar{z}_i \bar{z}_i \delta_2 - \ldots
- \frac{1}{\sqrt{n}} \left( \frac{\sqrt{\lambda_1 \lambda_2}}{\sqrt{l_2 l_1}} \sum_{i=1}^{n} u_{ki} u_{ii} \right)
+ \frac{\sigma_x}{\sqrt{l_2 l_1}} \sum_{i=1}^{n} \bar{z}_i \bar{z}_i \delta_i - \ldots
+ \frac{\sigma_x^2}{\sqrt{l_2 l_1}} \sum_{i=1}^{n} \bar{z}_i \bar{z}_i \delta_k
\]

Thus, we obtain

\[
\frac{1}{\sqrt{n}} u_i' \hat{\epsilon} = \frac{\sqrt{\lambda_1}}{\sqrt{l_1}} \sigma_y \sum_{i=1}^{n} u_{ii} \bar{z}_i \frac{1}{\sqrt{n}} \frac{d}{Z} \sim N(0, 1)
\]
and
\[
\begin{bmatrix}
\frac{1}{\sqrt{n}} \hat{u}_i' \hat{e}' \\
\vdots \\
\frac{1}{\sqrt{n}} \hat{u}_k' \hat{e}'
\end{bmatrix} \xrightarrow{d} C \sim N_k(0, \sigma_y^2 I)
\]
since
\[
E\left[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{ki} z_i \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{li} z_i \right] = E\left[ \frac{\sigma_y^2}{\sqrt{n}} \left( \sum_{i=1}^{n} u_{ki} u_{li} z_i^2 + \sum_{i \neq j} u_{ki} u_{li} z_i z_j \right) \right]
\]
with the fact \( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_{ki} u_{li} \xrightarrow{p} 0 \). In addition,
\[
\gamma_n \sum_{j=1}^{K} |\delta_j + \frac{w_j}{\sqrt{n}}| - |\delta_j| \xrightarrow{a.s} \gamma_0 \sum_{j=1}^{K} [w_j sgn(\delta_j) I(\delta_j \neq 0) + |\delta_j| I(\delta_j = 0)]
\]
Thus,
\[
f_n(w) \xrightarrow{d} f(w)
\]
where \( f(w) = -2w'C + w'Tw + \gamma_0 \sum_{j=1}^{K} [w_j sgn(\delta_j) I(\delta_j \neq 0) + |\delta_j| I(\delta_j = 0)] \) and based on the property for the convex function defined on compact set \( \sqrt{n}(\hat{\delta} - \delta) \xrightarrow{d} \text{argmin}(f) \). This completes the proof of Theorem 2.5.3.

Proof of Theorem 2.5.4. Since \( \delta \) is the LASSO estimates as defined (7) by KKT (Karush-Kuhn-Tucker) conditions, we have the followings
\[
\frac{\partial ||y - \hat{U}\delta||^2}{\partial \delta_j} \bigg|_{\delta_j = \delta_j} = \gamma_n sgn(\hat{\delta}_j) \quad \text{for} \quad j \quad \text{s.t.} \quad \delta_j \neq 0
\]
\[
\frac{\partial ||y - \hat{U}\delta||^2}{\partial \delta_j} \bigg|_{\delta_j = \delta_j} \leq \gamma_n \quad \text{for} \quad j \quad \text{s.t.} \quad \delta_j = 0.
\]
We already had as the following in Theorem 2.5.3,
\[
f(w) = -2w'C + w'Tw + \gamma_0 \sum_{j=1}^{K} [w_j sgn(\delta_j) I(\delta_j \neq 0) + |\delta_j| I(\delta_j = 0)],
\]
where \( w^* = \sqrt{n}(\hat{\delta} - \delta) \) is minimizer of \( f(w) \). With the given condition \( \frac{n}{\sqrt{n}} \xrightarrow{a.s} 0 \), \( f(w) \) is
differentiable w.r.t. $w$ and  
\[
\frac{\partial F(w)}{\partial w} = 2\sqrt{n}(Tw - C)
\]
where 
\[
\begin{bmatrix}
\frac{1}{\sqrt{n}} \hat{u}'_{\epsilon} \\
\vdots \\
\frac{1}{\sqrt{n}} \hat{u}'_{K\epsilon}
\end{bmatrix} \xrightarrow{d} C \sim N_k(0, \sigma^2 I)
\]
and 
\[
\frac{1}{n} \hat{U}' \hat{U} \overset{p}{\to} I = T.
\]

Let $w(1), C(1)$ and $w(2), C(2)$ denote the first $q$ and last $K - q$ entries of $w$ and $C$ respectively. Then we have equivalent condition 
\[
\{ \text{sgn}(\hat{\delta}_j) = \text{sgn}(\delta_j), \forall j = 1, \ldots, q \} = \{ \text{sgn}(\delta(1))w(1) > -\sqrt{n}|\delta(1)| \}.
\]

Then by KKT conditions and uniqueness of LASSO solution, if there exist $w$ such that 
\[
T_{11}(w(1) - C(1)) = -\frac{\gamma_n}{2\sqrt{n}} \text{sgn}(\delta(1)) ,
\]
\[
-\frac{\gamma_n}{2\sqrt{n}} \mathbf{1} \leq T_{21}(w(1) - C(2)) \leq \frac{\gamma_n}{2\sqrt{n}} \mathbf{1}
\]
\[
|w(1)| < \delta(1)
\]
then \(\text{sgn}(\hat{\delta}(1)) = \text{sgn}(\delta(1))\) and \(\hat{\delta}(2) = w(2) = 0\). Substitute $w(1), w(2)$ and boundness for their absolute values, the existence of such $\gamma_n$ is given by 
\[
|(T_{11})^{-1}C(1)| < \sqrt{n}(|\delta(1)| + \frac{\gamma_n}{2\sqrt{n}} |(T_{11})^{-1} \text{sgn}(\delta(1))|),
\]
\[
|T_{21}(T_{11})^{-1}C(1) - C(2)| \leq \frac{\gamma_n}{2\sqrt{n}} (1 - |T_{21}(T_{11})^{-1} \text{sgn}(\delta(1))|).
\]
Let these two events be $E_1$ and $E_2$ respectively. We have

$$P(\hat{A}_n = A) = P(\{\text{sgn}(\hat{\delta}_j) = \text{sgn}(\delta); \forall j \in A\} \cap \{\hat{\delta}_j = 0; \forall j \in A^c\}) \geq P(E_1 \cap E_2).$$

By using complement event for $E_1$ and $E_2$, it is given

$$1 - P(E_1 \cap E_2) \leq P(E_1^c) + P(E_2^c) \leq \sum_{i=1}^q P(|z_i| \geq \sqrt{n}(|\delta_i| - \frac{\gamma_n}{2\sqrt{n}} b_i)) + \sum_{i=1}^{K-q} P(|\zeta_i| \geq \frac{\gamma_n}{2\sqrt{n}} \eta_i)$$

where $z = (z_1, \ldots, z_k)' = (T_{11})^{-1}C(1)$, $\zeta = (\zeta_1, \ldots, \zeta_{K-q})' = T_{21}(T_{11})^{-1}C(1) - C(2)$ and $b = (T_{11})^{-1} \text{sgn}(\delta_{(1)})$. We know their limiting distribution

$$(T_{11})^{-1}C(1) \xrightarrow{d} N(0, T_{11}^{-1}),$$

$$T_{21}(T_{11})^{-1}C(1) - C(2) \xrightarrow{d} N(0, T_{22})$$

where $T_{11}^{-1} = I_q$, $T_{21}T_{11}^{-1}T_{12} = 0$, and $T_{22} = I_{K-q}$.

For $t > 0$, the Gaussian distribution has its tail probability bounded by

$$1 - \Phi(t) < \frac{1}{t} \exp\{-\frac{1}{2}t^2\}$$

Since $\frac{\gamma_n}{n} \xrightarrow{a.s.} 0$ and $\frac{\gamma_n}{n^{1/2}} \xrightarrow{a.s.} \infty$, thus

$$\sum_{i=1}^q P(|z_i| \geq \sqrt{n}(|\delta_i| - \frac{\gamma_n}{2\sqrt{n}} b_i)) \leq (1 + o(1)) \sum_{i=1}^q (1 - \Phi(t))(1 + o(1)) \frac{1}{s} n^{3/2} |\delta_i|)$$

$$= exp\{-n^c\}$$
where $s$ is such that $E(z_i)^2 \leq s$ and $E(\zeta_i)^2 \leq s$ and

$$
\sum_{i=1}^{K-q} P(\rho_i \geq \frac{\gamma n}{2\sqrt{n}} \eta_i) = \sum_{i=1}^{K-q} (1 - \Phi(t)(\frac{1}{s} \frac{\gamma n}{2\sqrt{n}} \eta_i)) = \exp\{-n^c\}.
$$

This completes the proof of Theorem 2.5.4.

**Proof of Theorem 2.5.5.** Based on results from Theorem 2.5.1 and Theorem 2.5.3, it can be easily shown by using similar arguments provided by Zou (2006) the similar method.

**Proof of Theorem 2.5.6.** For $j \in M^c = \{j : \beta_j = 0\}$, the marginal score statistic $s_j = \frac{U_j(0)^2}{I_{jj}(0)}$ has the asymptotic $\chi^2(1)$ distribution. By Chi square oracle inequality Johnstone (2001a), we have

$$
P(\max_j \frac{U_j(0)^2}{I_{jj}(0)} \geq \sqrt{\frac{b \log n}{n}}) \leq \sum_{j \in M^c} P(\chi^2(n) \geq n \sqrt{\frac{b \log n}{n}}) \leq \sqrt{2p(b \log n)^{-1/2}}n^{-b/4}.
$$

If $p = O(n)$, then for any $b > 8$, by Borel-Cantelli,

$$
\max_j \frac{U_j(0)^2}{I_{jj}(0)} \leq \sqrt{\frac{b \log n}{n}} \quad \text{a.s. as } n \to \infty.
$$

On the even set

$$
A_n = \{\max_{j \in M^c} \left| \frac{U_j(0)^2}{I_{jj}(0)} - \frac{U_j(0)^2}{T_{jj}(0)} \right| \geq \sqrt{\frac{b \log n}{n}} \},
$$

by the above argument, we have,

$$
\frac{U_j(0)^2}{I_{jj}(0)} \leq \sqrt{\frac{b \log n}{n}}, \quad \text{for all } j \in M^c
$$
If we take $0 < c_1 < 1$, then $\mathcal{M}^* \subset \hat{\mathcal{M}} = \{ j : \frac{\hat{I}_{ij}(0)}{\hat{I}_{jj}(0)} \geq c_1 \sqrt{\frac{\log n}{n}} \}$. With Lemma 1 and 2 on consistency set for zero coefficients, we have

$$\mathbb{P}(\mathcal{M}^* \subset \hat{\mathcal{M}}) \geq 1 - n^{-c}$$

This completes the proof of Theorem 2.5.6.
Lemma B.0.3. (McDiarmid’s inequality) Consider independent random variables $X_1, \ldots, X_m$. Let $h(\cdot) : \mathcal{X}^m \to \mathbb{R}$ be a function such that for all $i = 1, \ldots, m$ and for all $x_1, \ldots, x_m, \tilde{x}_i \in \mathcal{X}$, $h(\cdot)$ satisfies

$$\sup_{x_j \in \mathcal{X}, \tilde{x}_i \in \mathcal{X}} |h(x_1, \ldots, x_m) - h(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_m)| \leq c_i.$$

Then, for all probability measures $\mathbb{P}$ and every $\epsilon > 0$,

$$\mathbb{P}(h(X_1, \ldots, X_m) - E\{h(X_1, \ldots, X_m)\} \geq \epsilon) \leq \exp\left( - \frac{2\epsilon^2}{\sum_{i=1}^{m} c_i^2} \right),$$

where $E$ denotes the expectation over the $m$ random variables $X_1, \ldots, X_m$.

Proof of Theorem 4.5.1. For the sake of space, we temporarily omit the subscript $j$ for the $j$th feature. Let $\mathcal{C}(\mathcal{X})$ be the space of bounded continuous functions on $\mathcal{X}$ and $\mathcal{F} = \{f \in \mathcal{C}(\mathcal{X}) \ | \ \|f\|_{\mathcal{H}_x} \leq 1\}$ and $\mathcal{G} = \{g \in \mathcal{C}(\mathcal{Y}) \ | \ \|g\|_{\mathcal{H}_y} \leq 1\}$ be functional classes with the unit ball in an RKHS on each marginal input and outcome domain.

The key idea of our proof is to use a concentration probability on $\hat{\beta}_j$ to prove Theorem 4.5.1. First, we use McDiarmid’s inequality to obtain an upper bound probability of $\max_{f \in \mathcal{F}, g \in \mathcal{G}} |\hat{\beta} - \beta|$. By using the reproducing property and tensor product, we have

$$\beta = \int f(x)g(y)d\mathbb{P} - \int f(x)g(y)d\mathbb{Q}$$
$$= \int \langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}_x} \langle g(\cdot), L(\cdot, y) \rangle_{\mathcal{H}_y} d\mathbb{P} - \int \langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}_x} \int \langle g(\cdot), L(\cdot, y) \rangle_{\mathcal{H}_y} d\mathbb{Q}$$
$$= \int \langle K(\cdot, x) \otimes L(\cdot, y), f(\cdot) \otimes g(\cdot) \rangle_{\mathcal{H}_x \otimes \mathcal{H}_y} d\mathbb{P} - \int \langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}_x} \int \langle g(\cdot), L(\cdot, y) \rangle_{\mathcal{H}_y} d\mathbb{Q}.$$
maximum discrepancy between \( \beta \) and \( \hat{\beta} \) subject to the given functional classes is given by

\[
\max_{f \in \mathcal{F}, g \in \mathcal{G}} |\hat{\beta} - \beta|
\]

\[
= \sup_{f, g \in C} \left| \{ \mathbb{P}_n f(x)g(y) - Q_n f(x)g(y) \} - \{ \mathbb{P} f(x)g(y) - Q f(x)g(y) \} \right|
\]

\[
\leq \sup_{f, g \in C} \left| \mathbb{P}_n f(x)g(y) - \mathbb{P} f(x)g(y) \right| + \sup_{f, g \in C} \left| Q_n f(x)g(y) - Q f(x)g(y) \right|. \tag{5.1}
\]

We employ Lemma B.0.3 to obtain an upper bound probability for controlling the maximum discrepancy between \( \beta \) and \( \hat{\beta} \). We define \( h(x_1, \ldots, x_n, y_1, \ldots, y_n) = (\mathbb{P}_n - \mathbb{P}) f(x)g(y) \) and

\[
h(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, y_{i-1}, \tilde{y}_i, y_{i+1}, \ldots, y_n) = (\tilde{\mathbb{P}}_{n(i)} - \mathbb{P}) f(x)g(y),
\]

where \( \tilde{\mathbb{P}}_{n(i)} \) is the empirical measure of \( \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (\tilde{x}_i, \tilde{y}_i), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\} \).

Consider a dual RKHS \( \mathcal{H}_x \otimes \mathcal{H}_y \) with the feature map \( \phi(\cdot)\psi(\cdot) : (x, y) \rightarrow \phi(x)\psi(y) \in \mathcal{F} \otimes \mathcal{G} \) such that

\[
\langle \phi(x)\psi(y), \phi(x')\psi(y') \rangle_{\mathcal{H}_x \otimes \mathcal{H}_y} = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_x} \langle \psi(y), \psi(y') \rangle_{\mathcal{H}_y} = K(x, x')L(y, y').
\]

Therefore, we have

\[
\sup_{\tilde{x}_i \in X, \tilde{y}_i \in Y} \left| h(x_1, \ldots, x_n, y_1, \ldots, y_n) - h(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_n, y_1, \ldots, y_{i-1}, \tilde{y}_i, y_{i+1}, \ldots, y_n) \right|
\]

\[
\leq \sup_{f \in \mathcal{F}, \tilde{x}_i \in X, g \in \mathcal{G}, \tilde{y}_i \in Y} \left| \mathbb{P}_n f(x)g(y) - \tilde{\mathbb{P}}_{n(i)} f(x)g(y) \right|
\]

\[
= \frac{1}{n} \sup_{\|f\|_{\mathcal{H}_x} \leq 1, \tilde{x}_i \in X, \|g\|_{\mathcal{H}_y} \leq 1, \tilde{y}_i \in Y} \left| \langle f, \phi(x_i) \rangle \langle g, \psi(y_i) \rangle - \langle f, \phi(\tilde{x}_i) \rangle \langle g, \psi(\tilde{y}_i) \rangle \right| \tag{5.2}
\]

\[
\leq \frac{1}{n} \langle \phi(x_i), \phi(x_i) \rangle^{1/2} \langle \psi(y_i), \psi(y_i) \rangle^{1/2} - \langle \phi(\tilde{x}_i), \phi(\tilde{x}_i) \rangle^{1/2} \langle \psi(\tilde{y}_i), \phi(\tilde{y}_i) \rangle^{1/2} \right| \leq \frac{2}{n} \tilde{K}^{1/2} \tilde{L}^{1/2},
\]
where $\tilde{K} = \|K\|_{\infty}$ and $\bar{L} = \|L\|_{\infty}$. In this case, we have $c_i = 2n^{-1}\tilde{K}^{1/2}\bar{L}^{1/2}$ and
\[
\sum_{i=1}^{n} c_i^2 = n\left(\frac{2}{n}\tilde{K}^{1/2}\bar{L}^{1/2}\right)^2 + n\left(\frac{2}{n}\tilde{K}^{1/2}\bar{L}^{1/2}\right)^2 = 4\tilde{K}\bar{L}\left(\frac{n+n}{nn}\right) = \frac{8\tilde{K}\bar{L}}{n}. \tag{5.3}
\]

By using McDiarmid’s inequality, (5.1) and (5.3), we have
\[
\mathbb{P}\left(\max_{f \in F, g \in G} |\hat{\beta} - \beta| \geq \mathbb{E}\left\{\sup_{f, g \in C} \left| (\mathbb{P}_n - \mathbb{P})(f(x)g(y)) \right| \right\} + \mathbb{E}\left\{\sup_{f, g \in C} \left| (Q_n - Q)(f(x)g(y)) \right| \right\} + \varepsilon \right) \\
\leq \mathbb{P}\left( \sup_{f, g \in C} \left| \mathbb{P}_n(f(x)g(y)) - \mathbb{P}(f(x)g(y)) \right| + \sup_{f, g \in C} \left| Q_n(f(x)g(y)) - Q(f(x)g(y)) \right| \geq \mathbb{E}\left\{\sup_{f, g \in C} \left| (\mathbb{P}_n - \mathbb{P})(f(x)g(y)) \right| \right\} + \mathbb{E}\left\{\sup_{f, g \in C} \left| (Q_n - Q)(f(x)g(y)) \right| \right\} + \varepsilon \right) \tag{5.4}
\leq \exp\left(-\frac{n\varepsilon^2}{4\tilde{K}\bar{L}}\right).
\]

Next, we will use the symmetrization theorem to show that an upper bound for the expectation term in (5.4) is given by
\[
\mathbb{E}\left\{\sup_{f, g \in C} \left| (\mathbb{P}_n - \mathbb{P})(f(x)g(y)) \right| \right\} + \mathbb{E}\left\{\sup_{f, g \in C} \left| (Q_n - Q)(f(x)g(y)) \right| \right\} \\
\leq 2\mathbb{E}\left\{\sup_{f, g \in C} \left| \mathbb{P}_n(\sigma f(x)g(y)) \right| + \sup_{f, g \in C} \left| Q_n(\sigma f(x)g(y)) \right| \right\},
\]

where $\sigma_1, \ldots, \sigma_n$ are Rademacher random variables. It follows from the reproducing property that
\[
\mathbb{E}\{\sup_{f, g \in C} \left| \mathbb{P}_n[\sigma f(x)g(y)] \right| \} = \mathbb{E}\{\sup_{||f||_{H_x} \leq 1} \left| \langle f, \mathbb{P}_n[\sigma\phi(x)\psi(y)] \rangle \right| \} \\
\leq \mathbb{E}\{n^{-1}\mathbb{P}_n\sigma^2\phi(x)^2\psi(y)^2 + \mathbb{P}_n\mathbb{P}_n\sigma\bar{\sigma}\phi(x)\psi(y)\phi(\bar{x})\psi(\bar{y}) \}^{1/2} \leq (\tilde{K}\bar{L}n^{-1})^{1/2}. \tag{5.5}
\]

By substituting the result of symmetrization (5.5) in (5.4) for all features and taking $\varepsilon = n^{-\kappa}$ for $\kappa \in (0, \frac{1}{2})$, we can obtain a concentration inequality as follows:
\[
\mathbb{P}\left(\max_{1 \leq j \leq p_x} \max_{f_j \in F_j, g \in G} |\hat{\beta}_j - \beta_j| \geq (4\sqrt{\tilde{K}\bar{L}n^{\kappa-1/2}} + 1)n^{-\kappa} \right) \leq p_x \exp\left(-\frac{n^{1-2\kappa}}{4\tilde{K}\bar{L}}\right). \tag{5.6}
\]
Thus, the first result in (4.10) follows by letting $c_1 = 4 \sqrt{K \log(n)} + 1$ and $c_2 = 1/8 K L$ where each subscript $j$ indicates each $j$th marginal feature.

Finally, in the event $A_n \equiv \{ \max_{j \in \mathcal{M}} \max_{f_j \in \mathcal{F}, g \in \mathcal{G}} |\hat{\beta}_j - \beta_j| \leq c_0 n^{-\kappa} / 2 \}$, we have $|\hat{\beta}_j| \geq c_0 n^{-\kappa} / 2$ for all $j \in \mathcal{M}$ by using Condition 3. By setting $\gamma_n = c_3 n^{-\kappa}$ with $c_3 \leq c_0 / 2$, we have

$$P(\mathcal{M} \subset \hat{\mathcal{M}}_{r_n}) \geq P(A_n^c) \geq 1 - p_x \exp(-2c_2 n^{1-2\kappa}),$$

which leads to (4.11). This completes the proof of Theorem 4.5.1.

We introduce two large deviation inequalities, which play a critical role in our proofs of Theorems 4.5.2 and 4.5.3. The first lemma is from Johnstone and Lu (2009) and the second lemma is from Paul (2005). We omit their proofs.

**Lemma B.0.4.** Let $x_{i1}$ and $x_{i2}$ for $i = 1, \ldots, n$ be two sequences of mutually independent $N(0, 1)$ random variables. Then for a large $n$, for any $\epsilon > 0$, and $0 < b \leq 1/2$, we have

$$P\left( \left| \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \right| \geq n(1 + \epsilon) \right) \leq \exp\left( -\frac{3n\epsilon^2}{16} \right)$$

and for any $b$ such that $0 < b \leq \sqrt{n}$

$$P\left( \left| \frac{1}{n} \sum_{i=1}^{n} x_{i1}x_{i2} \right| \geq \sqrt{\frac{b}{n}} \right) \leq 2 \exp\left( -\frac{3b}{2} \right)$$

**Lemma B.0.5.** Let $t_n = 6 (\frac{p}{n} \lor 1) \sqrt{\frac{\log(p/n)}{p/n}}$. Then, for any $c > 0$, there exists $n^* \geq 1$ such that for $n \geq n^*$,

$$B_n = \{ \left| \frac{1}{n} EE^T - I_p \right| \geq \sqrt{\frac{p}{n} \lor \frac{p}{n} + c t_n} \} \leq 2(p \lor n)^{-c^2}$$

where $E = (e_1, e_2, \ldots, e_n)$ is a $p \times n$ matrix, $I_p$ is the $p \times p$ identity matrix, and $e_i$ is a $p \times 1$ vector whose components are independent random variables from $N(0, 1)$.
Proof of Theorem 4.5.2. Without loss of generality, we focus on \( \hat{\theta}_1^{(r)} \) for a specific \( r \)th cluster. Specifically, we denote \( \theta_m^{(r)} = (\theta_{m1}^{(r)}, \ldots, \theta_{mM_r}^{(r)})^T \) as a \( |M_r| \times 1 \) vector, \( \tilde{e}_i = (e_{i1}, \ldots, e_{i|M_r|})^T \) as measurement errors corresponding to cluster \( M_r \), and \( f_m^{(r)} = (f_{1m}^{(r)}, \ldots, f_{M_m}^{(r)})^T \) is the \( m \)th latent factor vector for \( M_r \). We denote \( \| \cdot \|_2 \) as \( \| \cdot \| \) and \( \lambda_m^{(r)} \) as \( \lambda_m \) for notational brevity. Thus, the sample variance covariance matrix for the \( r \)th cluster, denoted by \( S^{(r)} \), is given by

\[
S^{(r)} = \frac{1}{n} \sum_{m=1}^{M} \lambda_m ||f_m^{(r)}||^2 \theta_m^{(r)} \theta_m^{(r)T} + \frac{1}{n} \sum_{m=1}^{M} \sum_{l \neq m} \sqrt{\lambda_m \lambda_l} \langle f_m^{(r)}, f_l^{(r)} \rangle \theta_m^{(r)} \theta_l^{(r)T} + \frac{2}{n} \sum_{m=1}^{M} \sqrt{\lambda_m \sigma_x} \tilde{E}_m \theta_m^{(r)} \theta_m^{(r)T} + \frac{\sigma^2}{n} \tilde{E} \tilde{E}^T. \tag{5.7}
\]

where \( \tilde{E} = (\tilde{e}_1, \ldots, \tilde{e}_n) \).

To compute the risk function of \( \hat{\theta}_1^{(r)} \), we consider a perturbation equation (Kneip and Utikal 2001) given by

\[
\| \hat{\theta}_1^{(r)} - \theta_1^{(r)} \| \leq -H_1(\Sigma^{(r)})S^{(r)}\theta_1^{(r)} + \max_{k \neq 1} \|S^{(r)} - \Sigma^{(r)}\|^2 / |\lambda_k - \lambda_1|^2 \tag{5.8}
\]

where

\[
H_1(\Sigma^{(r)}) = \sum_{k \neq 1} \frac{1}{\lambda_k - \lambda_1} \theta_k^{(r)} \theta_k^{(r)T} - \frac{1}{\lambda_1} (I - \sum_{m=1}^{M} \theta_m^{(r)} \theta_m^{(r)T}). \tag{5.9}
\]

It follows from (5.7) and (5.9) that \( H_1(\Sigma^{(r)})S^{(r)}\theta_1^{(r)} \) can be rewritten as

\[
\sum_{k \neq 1} \frac{\sigma_x}{\lambda_k - \lambda_1} \left\{ \sqrt{\frac{\lambda_k}{n}} \langle \tilde{E}_k^{(r)}, \theta_1^{(r)} \rangle + \sqrt{\frac{\lambda_m}{n}} \langle \theta_k^{(r)}, \tilde{E}_m^{(r)} \rangle \right\} \theta_k^{(r)}
+ \sum_{k \neq 1} \frac{1}{\lambda_k - \lambda_1} \left\{ \frac{\sigma^2}{n} \langle \tilde{E} \theta_1^{(r)}, \tilde{E} \theta_1^{(r)} \rangle + \frac{\sqrt{\lambda_k \lambda_1}}{n} \langle f_k^{(r)}, f_1^{(r)} \rangle \right\} \theta_k^{(r)}
- \frac{1}{n \lambda_1} (I - \sum_{m=1}^{M} \theta_m^{(r)} \theta_m^{(r)T}) \tilde{E} \tilde{E}^T \theta_1^{(r)} - \frac{1}{n \lambda_1} (I - \sum_{m=1}^{M} \theta_m^{(r)} \theta_m^{(r)T}) \tilde{E} \theta_1^{(r)}. \tag{5.10}
\]

We take the expectation of \( \|H_1(\Sigma_r)S^{(r)}\theta_1^{(r)}\|^2 \) as the major term in (5.8) and also simplify (5.10).
by using a conditional argument. Note that $\tilde{\mathbf{E}}f_m^{(r)} = (\sum_{i=1}^n e_{i1}f_{im}, \ldots, \sum_{i=1}^n e_{i|M_r|}f_{im})^T$ is a $|M_r| \times 1$ vector, and $\langle \theta_k^{(r)} , \tilde{\mathbf{E}}f_m^{(r)} \rangle$ is given by

$$
\langle \theta_k^{(r)} , \tilde{\mathbf{E}}f_m^{(r)} \rangle = \sum_{j=1}^{|M_r|} \sum_{i=1}^n e_{ij}f_{im}\theta_{kj}.
$$

(5.11)

Since $\langle \theta_k^{(r)} , \tilde{\mathbf{E}}f_m^{(r)} \rangle \sim N(0, \sum_{i=1}^n (f_{im})^2 \sum_{j=1}^{|M_r|} (\theta_{kj})^2)$ conditional on $f_m^{(r)}$, we have $\mathbb{E}\langle \theta_k^{(r)} , \tilde{\mathbf{E}}f_m^{(r)} \rangle = 0$ and $\mathbb{E}\langle \theta_k^{(r)} , \tilde{\mathbf{E}}f_m^{(r)} \rangle^2 = \sum_{i=1}^n (f_{im})^2 \sum_{j=1}^{|M_r|} (\theta_{kj})^2$ conditional on $f_m^{(r)}$. Taking unconditional expectations of the square of (5.11) leads to

$$
\mathbb{E}\langle \theta_k^{(r)} , \tilde{\mathbf{E}}f_m^{(r)} \rangle^2 = \sum_{j=1}^{|M_r|} (\theta_{kj})^2 \mathbb{E}[\chi_n^2] = n.
$$

(5.12)

It follows from (5.12) that the expectation of the square of the first term in (5.10) is given by

$$
\mathbb{E}\left[ \frac{\sqrt{\lambda_k}}{n} \langle \tilde{\mathbf{E}}f_k^{(r)} , \theta_1^{(r)} \rangle + \frac{\sqrt{\lambda_1}}{n} \langle \theta_k^{(r)} , \tilde{\mathbf{E}}f_1^{(r)} \rangle \right]^2 = \frac{1}{n} (\lambda_k + \lambda_1).
$$

(5.13)

Note that $\langle \tilde{\mathbf{E}}^T \theta_k^{(r)} , \tilde{\mathbf{E}}^T \theta_1^{(r)} \rangle = \sum_{i=1}^n \tilde{e}_i^{(k)} \tilde{e}_i^{(1)}$ where $\tilde{e}_i^{(k)} , \ldots, \tilde{e}_n^{(k)}$ are independently and identically distributed as $N(0, 1)$ and $\tilde{e}_1^{(1)} , \ldots, \tilde{e}_n^{(1)}$ are independently and identically distributed as $N(0, 1)$. Since $\tilde{e}_i^{(k)}$ and $\tilde{e}_i^{(1)}$ are independent, we have $\mathbb{E}\langle \tilde{\mathbf{E}}^T \theta_k^{(r)} , \tilde{\mathbf{E}}^T \theta_1^{(r)} \rangle = 0$ and $\mathbb{E}\langle \tilde{\mathbf{E}}^T \theta_k^{(r)} , \tilde{\mathbf{E}}^T \theta_1^{(r)} \rangle^2 = n$.

Also, we have $\langle f_k^{(r)} , f_1^{(r)} \rangle = \sum_{i=1}^n f_{ik}^{(r)} f_{i1}^{(r)}$ , $\mathbb{E}\langle f_k^{(r)} , f_1^{(r)} \rangle = 0$ and $\mathbb{E}\langle f_k^{(r)} , f_1^{(r)} \rangle^2 = n$. Since the cross product term between $\langle \tilde{\mathbf{E}}^T \theta_k^{(r)} , \tilde{\mathbf{E}}^T \theta_1^{(r)} \rangle$ and $\langle f_k , f_1 \rangle$ vanishes, we have

$$
\mathbb{E}\left[ \frac{\sigma^2}{n} \langle \tilde{\mathbf{E}}^T \theta_1^{(r)} , \tilde{\mathbf{E}}^T \theta_k^{(r)} \rangle + \frac{\sqrt{\lambda_k \lambda_1}}{n} \langle f_k , f_1 \rangle \right] = \left( \frac{\sigma^2}{n} + \frac{\lambda_k \lambda_1}{n} \right). \tag{5.14}
$$

For the last term (5.10), it follows that

$$
\mathbb{E}\left[ \frac{1}{n^2 \lambda_1^2} \theta_1^{(r)T} \tilde{\mathbf{E}} \left( I - \sum_{m=1}^M \theta_m^{(r)} \theta_m^{(r)T} \right) \tilde{\mathbf{E}}^T \theta_1^{(r)} \right] = \frac{(|M_r| - M)}{n \lambda_1^2}
$$

and

$$
\mathbb{E}\left[ \frac{1}{n^2 \lambda_1} \theta_1^{(r)T} \tilde{\mathbf{E}} \left( I - \sum_{m=1}^M \theta_m^{(r)} \theta_m^{(r)T} \right) \tilde{\mathbf{E}} \theta_1^{(r)} \right] = \frac{(|M_r| - M)}{n \lambda_1}. \tag{5.15}
$$
From (5.13), (5.14) and (5.15), we have

\[
\mathbb{E}\|H_1(\Sigma_r)S^{(r)}\theta_1^{(r)}\|^2 = \frac{(|\mathcal{M}_r| - M)(1 + \lambda_1)}{n\lambda_1^2} + \sum_{k \neq 1}^{M} \frac{\sigma_k^2(\lambda_k + \lambda_1) + (\sigma_k^2 + \lambda_k\lambda_1)}{n(\lambda_k - \lambda_1)^2}.
\] (5.16)

Finally, we show that \(\max_{k \neq 1} |\lambda_k - \lambda_1|^{-1}\|S^{(r)} - \Sigma^{(r)}\|\) is small in probability. We note

\[
\|S^{(r)} - \Sigma^{(r)}\| \leq \sum_{m=1}^{M} \lambda_m \frac{1}{n}\|f_m^{(r)}\|^2 - \mathbb{E}\|f_m^{(r)}\|^2 + \sum_{m=1}^{M} \sum_{l \neq m} \sqrt{\lambda_m\lambda_l} \frac{1}{n} \langle f_m^{(r)}, f_l^{(r)} \rangle
\]

\[+ 2 \sum_{m=1}^{M} \sqrt{\lambda_m - \sigma_x} \mathbb{E}|\tilde{f}_m^{(r)}| + \sigma_x^2 \frac{1}{n} \tilde{E}\tilde{E}^T - I|.
\] (5.17)

It follows from Lemmas B.0.4 and B.0.5 that

\[
\mathbb{P}\left( \max_{1 \leq m \leq M} \frac{1}{n}\|f_m^{(r)}\|^2 - \mathbb{E}\left( \frac{1}{n}\|f_m^{(r)}\|^2 \right) \geq \mathbb{E}\left( \frac{1}{n}\|f_m^{(r)}\|^2 \right) \epsilon \right) \leq M \exp\left( -\frac{3n\epsilon^2}{16} \right)
\]

\[
\mathbb{P}\left( \max_{1 \leq m \leq M} \max_{l \neq m} \frac{1}{n} \langle f_m^{(r)}, f_l^{(r)} \rangle \geq \sqrt{\mathbb{E}(\|f_m^{(r)}\|^2)\mathbb{E}(\|f_l^{(r)}\|^2)} \sqrt{\frac{\epsilon_2}{n}} \right) \leq M(M - 1) \exp\left( -\frac{3\epsilon_2^2}{2} \right)
\]

\[
\mathbb{P}\left( \max_{1 \leq m \leq M} \left\| \frac{1}{n} \tilde{E}f_m^{(r)} \right\| \geq \sqrt{\mathbb{E}(\|f_m^{(r)}\|^2)\sqrt{\frac{|\mathcal{M}_r|}{n}}(1 + \epsilon_1)} \right) \leq M \exp\left( -\frac{3n\epsilon_1^2}{16} \right)
\] (5.18)

and that

\[
\mathbb{P}\left( \left\| \frac{1}{n} \tilde{E}\tilde{E}^T - I \right\| \leq 2 \sqrt{\frac{|\mathcal{M}_r|}{n} + \frac{|\mathcal{M}_r|}{n} + \epsilon_3 t_n} \right) \leq 2(n \lor |\mathcal{M}_r|)^{-\epsilon_1^3}
\] (5.19)

where choose \(\epsilon_1 = \sqrt{\frac{\log |\mathcal{M}_r| + 1}{n^{2\epsilon_2} \log |\mathcal{M}_r|}}, \) a sufficiently small \(\epsilon_3\), and \(t_n = 6\left( \frac{|\mathcal{M}_r|}{n} \lor 1 \right) \sqrt{\frac{\log (|\mathcal{M}_r| \lor n)}{|\mathcal{M}_r| \lor n}}\).

Now, we obtain a new bound under the assumption and probability limit from (5.18) and (5.19):

\[
\mathbb{P}\left( \max_{1 \leq k \leq M} \frac{\|S^{(r)} - \Sigma^{(r)}\|}{|\lambda_k - \lambda_1|} \geq \epsilon(n, |\mathcal{M}_r|, \lambda) \right)
\]

\[\leq \mathbb{P}\left( \max_{1 \leq k \leq M} \sum_{m=1}^{M} \frac{\lambda_m - \lambda_m + 1}{\lambda_m - \lambda_{m+1}} \frac{1}{n}\|f_m^{(r)}\|^2 - \mathbb{E}(\|f_m^{(r)}\|^2) \geq c_1 \sum_{m=1}^{M} \frac{\lambda_m - \lambda_{m+1}}{n^{\kappa}} \sqrt{\frac{\log |\mathcal{M}_r|}{n^{\kappa}}} \right)\]
Hence, from (5.8), (5.10), (5.16) and (5.20), we obtain

\[
\begin{align*}
+ P \left( \max_{1 \leq k \leq M} \sum_{m=1}^{M} \sum_{l \neq m} \frac{\sqrt{\lambda_m \lambda_l}}{\lambda_m - \lambda_{m+1}} \left| \frac{1}{n} \left( f_m(f_l) f_{(r)}^l \right) \right| \right) & \geq c_2 \sum_{m=1}^{M} \sum_{l \neq m} \frac{\sqrt{\lambda_m \lambda_l}}{\lambda_m - \lambda_{m+1}} \sqrt{\log |M_r|} \\
+ P \left( \max_{1 \leq k \leq M} \sum_{m=1}^{M} \frac{\sqrt{\lambda_m}}{\lambda_m - \lambda_{m+1}} \frac{1}{n} \sigma_x \| \hat{E}_m \| \right) & \geq c_3 \frac{2}{n} \sum_{m=1}^{M} \frac{\sqrt{\lambda_m}}{\lambda_m - \lambda_{m+1}} \sqrt{\log |M_r|} \left( 1 + \sqrt{\log |M_r|} \right) \\
+ P \left( \max_{1 \leq k \leq M} \sum_{m=1}^{M} \frac{1}{\lambda_m - \lambda_{m+1}} \left| \frac{1}{n} \hat{E}^T - I \right| \right) & \geq \sum_{m=1}^{M} \frac{1}{\lambda_m - \lambda_{m+1}} \left( 2 \sqrt{\frac{|M_r|}{n}} + \frac{|M_r|}{n} + \epsilon_3 t_n \right) \\
& \leq 2M \exp \left( - \frac{3n^2 \log |M_r|}{16} \right) + 2M(M-1) \exp \left( - \frac{n^2 \log |M_r|}{2} \right) + 2(n \vee |M_r|)^{-\epsilon_3},
\end{align*}
\]

(5.20)

where \( \kappa_1 \) and \( \kappa_2 \in (0, 1/2) \), \( \kappa = (2\kappa_1 \wedge 1 - 2\kappa_2) \), \( c_1, c_2 \), and \( c_3 \) are constants from (5.18) and

\[
\begin{align*}
\epsilon(n, |M_r|, \lambda) = & \ c_1 \sum_{m=1}^{M} \frac{\lambda_m}{\lambda_m - \lambda_{m+1}} \sqrt{\frac{\log |M_r|}{n^\kappa}} + c_2 \sum_{m=1}^{M} \sum_{l \neq m} \frac{\sqrt{\lambda_m \lambda_l}}{\lambda_m - \lambda_{m+1}} \sqrt{\log |M_r|} \\
& + c_3 \frac{2}{n} \sum_{m=1}^{M} \frac{\sqrt{\lambda_m}}{\lambda_m - \lambda_{m+1}} \sqrt{\log |M_r|} \left( 1 + \sqrt{\log |M_r|} \right) \\
& + \sum_{m=1}^{M} \frac{1}{\lambda_m - \lambda_{m+1}} \left( 2 \sqrt{\frac{|M_r|}{n}} + \frac{|M_r|}{n} + \epsilon_3 t_n \right).
\end{align*}
\]

(5.21)

We note that \( \epsilon(n, |M_r|, \lambda) \) in (5.21) is sufficiently small under Conditions 5-6. It follows that

\[
\frac{\lambda_m/\lambda_1}{\lambda_m/\lambda_1 - \lambda_m+1/\lambda_1} \sqrt{\frac{\log |M_r|}{n^\kappa}} \rightarrow \rho \frac{\log |M_r|}{\rho \rho_m - \rho_{m+1}} + \rho_m = O_p \left( \sqrt{\frac{\log |M_r|}{n^\kappa}} \right).
\]

and that

\[
\sqrt{\frac{\lambda^2}{1 + \lambda_1 (\lambda_m - \lambda_{m+1})}} \sqrt{\frac{|M_r| (1 + \lambda_1)}{n^\kappa \lambda_1^2}} \left( 1 + O_p \left( \sqrt{\frac{\log |M_r|}{n^\kappa}} \right) \right) = o_p(1).
\]

Hence, from (5.8), (5.10), (5.16) and (5.20), we obtain

\[
\mathcal{R}_{L,P}(\hat{\theta}_{1}^{(r)}) = \mathbb{E} \left[ \| H_1(S_{r}) S^{(r)} \hat{\theta}_{1}^{(r)} \|^2 \right] + o_p(1)
\]

(5.22)

This completes the proof of Theorem 4.5.2.
Proof of Theorem 4.5.3. For a fixed $\epsilon > 0$, the compactness of $\mathcal{F}(\hat{\theta})$ shows that there exists an $\epsilon$-net $\mathcal{F}_\epsilon(\hat{\theta})$ of $\mathcal{F}(\hat{\theta})$ with $|\mathcal{F}_\epsilon(\hat{\theta})| = \mathcal{N}(\mathcal{F}(\hat{\theta}), \| \cdot \|_\infty, \epsilon) < \infty$. For $f \in \mathcal{F}(\hat{\theta})$, there exists a $g \in \mathcal{F}(\hat{\theta})$ with $\| f - g \|_\infty \leq \epsilon$. Thus, it follows from the existence of a constant $|L|_1$ because of the local Lipschitz continuity of the convex loss function (Steinwart and Christmann 2008) that

$$
|\mathcal{R}_{L,P,\hat{\theta}}(f) - \mathcal{R}_{L,n,\hat{\theta}}(f)|
\leq |\mathcal{R}_{L,P,\hat{\theta}}(f) - \mathcal{R}_{L,P,\hat{\theta}}(g)| + |\mathcal{R}_{L,P,\hat{\theta}}(g) - \mathcal{R}_{L,n,\hat{\theta}}(g)| + |\mathcal{R}_{L,n,\hat{\theta}}(g) - \mathcal{R}_{L,n,\hat{\theta}}(f)|
\leq 2\epsilon|L|_1 + |\mathcal{R}_{L,P,\hat{\theta}}(g) - \mathcal{R}_{L,n,\hat{\theta}}(g)|.
$$

By taking suprema on both sides, we obtain

$$
\sup_{f \in \mathcal{F}(\hat{\theta})} |\mathcal{R}_{L,n,\hat{\theta}}(f) - \mathcal{R}_{L,P,\hat{\theta}}(f)| \leq 2\epsilon|L|_1 + \sup_{g \in \mathcal{F}_\epsilon(\hat{\theta})} |\mathcal{R}_{L,P,\hat{\theta}}(g) - \mathcal{R}_{L,n,\hat{\theta}}(g)|. \quad (5.23)
$$

Furthermore, $\lambda \| \hat{f}_{n,\lambda,\hat{\theta}} \|^2 + \mathcal{R}_{L,n,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) \leq \lambda \| f_{P,\lambda,\hat{\theta}} \|^2 + \mathcal{R}_{L,n,\hat{\theta}}(f_{P,\lambda,\hat{\theta}})$ leads to an upper bound for the risk difference between a regularized empirical minimizer and a general solution. It is given by

$$
\lambda \| \hat{f}_{n,\lambda,\hat{\theta}} \|^2 + \mathcal{R}_{L,P,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\hat{\theta}}) - \left( \lambda \| f_{P,\lambda,\hat{\theta}} \|^2 + \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\hat{\theta}}) \right)
= \mathcal{R}_{L,P,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,n,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) + \lambda \| \hat{f}_{n,\lambda,\hat{\theta}} \|^2 + \mathcal{R}_{L,n,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\lambda,\hat{\theta}})
\leq \mathcal{R}_{L,P,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,n,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) + \mathcal{R}_{L,n,\hat{\theta}}(f_{P,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\lambda,\hat{\theta}})
\leq 2 \sup_{f \in \mathcal{F}(\hat{\theta})} |\mathcal{R}_{L,n,\hat{\theta}}(f) - \mathcal{R}_{L,P,\hat{\theta}}(f)|. \quad (5.24)
$$

Moreover, since it follows from Condition 9 that $\| f_{P,\lambda,\hat{\theta}} \| \leq \sqrt{\frac{\mathcal{R}_{L,P,\hat{\theta}}(0)}{\lambda}} \leq \frac{M}{\sqrt{\lambda}}$ holds for all distributions on $\mathcal{X} \times \mathcal{Y}$, for $f \leq \frac{M}{\sqrt{\lambda}} B_H$ and $B = M(|L|_1 \frac{1}{\sqrt{\lambda}} + 1)$, we have

$$
|L(x, y, f_{P,\lambda,\hat{\theta}})| \leq |L(x, y, f_{P,\lambda,\hat{\theta}}) - L(x, y, 0)| + L(x, y, 0) \leq B.
$$
where $B_H$ is a unit ball on $L(X)$. Thus, we conclude that $\sup_{f \in F} |L(x, y, f_{P, \hat{\theta}})| \leq B$.

Let $|\mathcal{F}_\epsilon(\hat{\theta})|$ be an $\epsilon$-net of $\frac{M}{\sqrt{\lambda}} B_H$ with cardinality

$$|\mathcal{F}_\epsilon(\hat{\theta})| = \mathcal{N}(\frac{M}{\sqrt{\lambda}} B_H, \| \cdot \|_\infty, \epsilon) = \mathcal{N}(B_H, \| \cdot \|_\infty, \frac{M}{\sqrt{\lambda}} \epsilon)$$

and $A_1(\lambda)$ is an approximation error function defined by $\lambda \|f_{P, \lambda, \hat{\theta}}\|^2 + \mathcal{R}_{L,P,\hat{\theta}}(f_{P, \lambda, \hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P, \hat{\theta}})$.

It follows from (5.23), (5.24) and Lemma B.0.3 that for $\tau' > 0$, we have

$$\mathbb{P}\left( \lambda \|\hat{f}_{n,\lambda,\hat{\theta}}\|^2 + \mathcal{R}_{L,P,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P, \hat{\theta}}) \geq A_1(\lambda) + B \sqrt{\frac{2\tau'}{n}} + 4\epsilon |L|_1 \right)$$

$$\leq \mathbb{P}\left( 2 \sup_{f \in \mathcal{F}(\hat{\theta})} |\mathcal{R}_{L,n,\hat{\theta}}(f) - \mathcal{R}_{L,n,\hat{\theta}}(f_{P, \hat{\theta}}) | \geq B \sqrt{\frac{2\tau'}{n}} + 4\epsilon |L|_1 \right)$$

$$\leq \mathbb{P}\left( \sup_{g \in \mathcal{F}_\epsilon(\hat{\theta})} |\mathcal{R}_{L,P,\hat{\theta}}(g) - \mathcal{R}_{L,n,\hat{\theta}}(g) | \geq B \sqrt{\frac{\tau'}{2n}} \right)$$

$$\leq 2\mathcal{N}(B_H, \| \cdot \|_\infty, \frac{M}{\sqrt{\lambda}} \epsilon) e^{-\tau}.$$ 

By choosing $\tau' = \tau + \log 2 \mathcal{N}(B_H, \| \cdot \|_\infty, \frac{M}{\sqrt{\lambda}} \epsilon)$, we have

$$\lambda \|\hat{f}_{n,\lambda,\hat{\theta}}\|^2 + \mathcal{R}_{L,P,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P, \hat{\theta}})$$

$$\leq A_1(\lambda) + 4\epsilon |L|_1 + (|L|_1 \frac{M}{\sqrt{\lambda}} + 1) \sqrt{2\tau + 2 \log (2\mathcal{N}(B_H, \| \cdot \|_\infty, \frac{M}{\sqrt{\lambda}} \epsilon))} \frac{2\tau + 2 \log (2\mathcal{N}(B_H, \| \cdot \|_\infty, \frac{M}{\sqrt{\lambda}} \epsilon))}{n}. \quad (5.25)$$

Since $K(\cdot, \cdot)$ is an $m$ times continuously differentiable kernel on $\mathbb{R}^d$, it follows from Theorem 6.26 (Steinwart and Christmann 2008) that we have the entropy $\log \mathcal{N}(B_H, \| \cdot \|_\infty, \epsilon) \leq \alpha \epsilon^{-2p}$, where $2p = d/m$ and $\alpha$ is a positive constant. For instance, the Gaussian kernel satisfies this entropy condition. Consequently, we use these kernels as candidate kernels for Stage (III). To derive the convergence rate, we choose

$$\epsilon = \left( \frac{p}{2} \right)^{1/(1+p)} \left( \frac{2\alpha}{n} \right)^{1/2(1+p)} \lambda^{-1/2}.$$
Equation (5.25) together with Lemma A.1.5 Steinwart and Christmann (2008) and \((p+1)(2/p)^{1/(1+p)} \leq 3\), establishes

\[
\lambda \| \hat{f}_{n,\lambda,\hat{\theta}} \|^2 + \mathcal{R}_{L,P,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\hat{\theta}}^0) \leq A_1(\lambda) + \frac{3}{\lambda^{1/2}} \left( 2 \left( \frac{2\alpha}{n} \right)^{1/(1+p)} + \left( \frac{2\tau}{n} \right)^{1/2} \right).
\]

From Condition 10, it follows that the asymptotically best choice for \(\lambda\) is shown by \(n^{-\frac{1}{(1+p)(2\beta+1)}}\), which leads the convergence rate

\[
\mathcal{R}_{L,P,\hat{\theta}}(\hat{f}_{n,\lambda,\hat{\theta}}) - \mathcal{R}_{L,P,\theta}(f_{P,\theta}^* - f_{P,\theta}^*) = O_p(n^{-\frac{1}{(1+p)(2\beta+1)}}).
\] (5.26)

By substituting \(2p = \frac{d}{m}\) into (5.26), we obtain the rate of \(O_p(n^{-\frac{2\beta}{(2m+d)(2\beta+1)}})\).

Next, it remains to show \(\mathcal{R}_{L,P,\hat{\theta}}(f_{P,\theta}^0) - \mathcal{R}_{L,P,\theta}(f_{P,\theta}^*)\) is asymptotically small in probability. If Conditions 9-10 hold, it follows that

\[
\mathcal{R}_{L,P,\hat{\theta}}(f_{P,\theta}^0) - \mathcal{R}_{L,P,\theta}(f_{P,\theta}^*) = \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\theta}^0) - \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\theta}^*) + \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\theta}^*) - \mathcal{R}_{L,P,\theta}(f_{P,\theta}^*)
\]

\[
\leq |L|_1 \| f_{P,\theta}^0 - f_{P,\theta}^* \| + |L|_2 \| \hat{\theta} - \theta \|
\]

\[
\leq |L|_1 \| f_{P,\theta}^0 - f_{P,\theta}^* \| + |L|_1 \| f_{P,\theta}^0 - f_{P,\theta}^* \| + |L|_1 \| f_{P,\theta}^* - f_{P,\theta}^* \| + |L|_2 \| \hat{\theta} - \theta \|
\]

\[
\leq |L|_1 \| f_{P,\theta}^0 - f_{P,\theta}^* \| + 2|L|_1 |L|_3 + |L|_2 \| \hat{\theta} - \theta \|
\]

Then, we consider an upper bound for \(\| \hat{\theta} - \theta \|\) from Theorem 4.5.2. Let define \(A_2 := C(\lambda_1^{(r)}) \times \left( \sqrt{\frac{\log |\mathcal{M}|}{n^\kappa}} + \frac{|\mathcal{M}||\lambda_1^{(r)}+1|}{n^\kappa} \right)\) where \(C(\lambda_1^{(r)})\) is an universal constant depending on \(\lambda_1^{(r)}\) for some \(0 < \kappa < 1\). Consequently, we have

\[
P \left( \mathcal{R}_{L,P,\hat{\theta}}(f_{P,\theta}^0) - \mathcal{R}_{L,P,\theta}(f_{P,\theta}^*) \geq (2|L|_1 |L|_3 + |L|_2) A_2 + |L|_1 \| f_{P,\theta}^0 - f_{P,\theta}^* \| \right)
\]

\[
\leq P \left( \| \hat{\theta} - \theta \| \geq A_2 \right) \to 0.
\]
Hence, we conclude that

$$\lambda \| \hat{f}_{n, \lambda, \hat{\theta}} \|^2 + R_{L,p,\theta}(\hat{f}_{n, \lambda, \hat{\theta}}) - R_{L,p,\theta}(f_{\hat{P},\theta}^*) \leq O_p(n^{- \frac{2m \beta}{(2m+d)(2\beta+1)}}) + O_p(\sqrt{\frac{\log |M_r|}{n^\kappa}}) + O_p(\frac{|M_r|(|\lambda_{1^*}^{(r)}| + 1)}{n(\lambda_{1^*}^{(r)})^2}) + O_p(||f_{\hat{P},\theta}^0 - f_{\hat{P},\theta}^*||)$$

This completes the proof of Theorem 4.5.3.
BIBLIOGRAPHY


— (2009), The Elements of Statistical Learning, New York, NY: Springer.


— (2003), Threshold selection in transform shrinkage, Springer.


Jolliffe, I. (2005), Principal component analysis, Wiley Online Library.


Kendall, M. G. (1965), course in multivariate analysis, Griffin, London.


