8-2016

Model-Free Variable Screening, Sparse Regression Analysis and Other Applications with Optimal Transformations

Qiming Huang
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Qiming Huang

Entitled
Model-Free Variable Screening, Sparse Regression Analysis and Other Applications with Optimal Transformations

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Michael Yu Zhu
Chair

Chuanhai Liu

Hyonho Chun

Hao Zhang

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Michael Yu Zhu

Approved by: Jun Xie 5/4/2016
Head of the Departmental Graduate Program Date

MODEL-FREE VARIABLE SCREENING, SPARSE REGRESSION ANALYSIS AND OTHER APPLICATIONS WITH OPTIMAL TRANSFORMATIONS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Qiming Huang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2016

Purdue University

West Lafayette, Indiana

To my family.

ACKNOWLEDGMENTS

My first and foremost thanks go to my advisor, Professor Michael Yu Zhu. His far-reaching vision, valuable guidance, and inspirational encouragement have helped me explore and develop ideas and overcome incredible challenges throughout my PhD. Michael has spent a tremendous amount of time and energy on guiding me through interesting research questions and helping me develop as a researcher. This dissertation would not be possible without his guidance and patience. Thank you for being a fantastic advisor and friend.

I deeply appreciate insightful comments and encouragements from Professor Hyonho Chun, Professor Chuanhai Liu and Professor Hao Zhang who serves as members of my thesis committee.

I'd like to thank Professor Anirban DasGupta and Professor Chuanhai Liu for their extraordinary courses. I'd like to thank Professor Todd Kelley, Professor Louis Tay, Professor Brenda Capobianco and Dr. Chell Nyquist for their guidances and collaborations on psychometrics and SLED project with four-year financial supports. My thanks go to Dr. Sergey Kirshner and Professor Olga Vitek for their supervisions and supports at the early stage of my PhD. Thanks also go to all members of my research group: Longjie Cheng, Zhaonan Sun, Han Wu, Pan Chao, Bing Yu, Rongrong Zhang, for their critical discussions and helps on various research topics.

I'd like to thank all my friends at Purdue. A big thank you goes to Jeff Li for being a great mentor, roommate and friend; You are like a brother to me. I am very grateful for the generous helps from Jin Xia, Youran Fan, Cheng Liu, Han Wu and Bowen Zhou. It's quite an unforgetable memory preparing for qualifying exams with Yang Zhao and Xiaoguang Wang. I'm fortunate to have Xiaosu Tong as my intern partner and thank you for the wonderful and fruitful summer we had together. I had lots of fun fishing with Wei Sun. I learned a lot from various short chats with Zach Haas, Whitney Huang, Qi Wang, Yixuan

Qiu and Wei Sun. I want to thank Zhuo Chen, Shuqian Zhang, Weiwei Zhang and Bingrou Zhou for all the wonderful meals and their hospitality.

Last, but not least, I would like to thank my beloved parents, my brother Jeff, my sister-in-law Kami, and my dearest wife Emily, for their amazing support, tolerance, understanding, and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ACE	Alternating Conditional Expectation
BMC	B-spline-based Maximum Correlation, using the largest eigenvalue
CV	Cross Validation
DC-SIS	Distance Correlation-based Sure Independence Screening
IQR	Inter-Quantile Range
LLA	Local Linear Approximation
MBMC	Multivariate version of B-spline-based Maximum correlation, using the largest eigenvalue
MC-SIS	Maximum Correlation-based Sure Independence Screening
MMS	Mimimal Model Size
MSE	Mean Squared Error
NIS	Nonparametric Independence Screening
RKHS	Reproducing Kernel Hilbert Space
RSD	Robust Standard Deviation
SICA	Smooth Integration of Counting and Absolute deviation
SIS	Sure Indepedence Screening
SPAM	SParse Additive Model
SPOT	SParse Optimal Transformation
SPOT-LASSO	SParse Optimal Transformation with L_1 penalty
SPOT-SICA	SParse Optimal Transformation with SICA penalty
T-BMC	B-spline-based Maximum Correlation, using Trace
T-MBMC	Multivariate version of B-spline-based Maximum Correlation, using Trace

ABSTRACT

Huang, Qiming PhD, Purdue University, August 2016. Model-Free Variable Screening, Sparse Regression Analysis and Other Applications with Optimal Transformations . Major Professor: Michael Yu Zhu.

Variable screening and variable selection methods play important roles in modeling high dimensional data. Variable screening is the process of filtering out irrelevant variables, with the aim to reduce the dimensionality from ultrahigh to high while retaining all important variables. Variable selection is the process of selecting a subset of relevant variables for use in model construction. The main theme of this thesis is to develop variable screening and variable selection methods for high dimensional data analysis. In particular, we will present two relevant methods for variable screening and selection under a unified framework based on optimal transformations.

In the first part of the thesis, we develop a maximum correlation-based sure independence screening (MC-SIS) procedure to screen features in an ultrahigh-dimensional setting. We show that MC-SIS possesses the sure screen property without imposing model or distributional assumptions on the response and predictor variables. MC-SIS is a model-free method in contrast with some other existing model-based sure independence screening methods in the literature. In the second part of the thesis, we develop a novel method called SParse Optimal Transformations (SPOT) to simultaneously select important variables and explore relationships between the response and predictor variables in high dimensional nonparametric regression analysis. Not only are the optimal transformations identified by SPOT interpretable, they can also be used for response prediction. We further show that SPOT achieves consistency in both variable selection and parameter estimation.

Besides variable screening and selection, we also consider other applications with optimal transformations. In the third part of the thesis, we propose several dependence measures, for both univariate and multivariate random variables, based on maximum correlation and B-spline approximation. B-spline based Maximum Correlation (BMC) and Trace BMC (T-BMC) are introduced to measure dependence between two univariate random variables. As extensions to BMC and T-BMC, Multivariate BMC (MBMC) and Trace Multivariate BMC (T-MBMC) are proposed to measure dependence between multivariate random variables. We give convergence rates for both BMC and T-BMC.

Numerical simulations and real data applications are used to demonstrate the performances of proposed methods. The results show that the proposed methods outperform other existing ones and can serve as effective tools in practice.

1. INTRODUCTION

One common goal for data analysis is to discover the underlying dependence structure between the response Y and predictor vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, which can be fully captured by the conditional distribution $P(Y|\mathbf{X})$. Different regression models have been proposed to characterize the dependence structure, from a limited sample of Y and $\mathbf{X}$. Regression models differ in several aspects, such as model flexibility, interpretability, computational efficiency and prediction accuracy.

Model flexibility and interpretability have been recognized to play key roles in practical data analysis. A general nonparametric regression model,

$$Y = f(\mathbf{X}, \epsilon), \tag{1.1}$$

or a simplified version $Y = f(\mathbf{X}) + \epsilon$ where ϵ is a random error, is the most flexible model in regression setting. It assumes no structure constraints on the function f, and can accommodate any possible interactions among those predictor variables. However, this approach suffers severely from the curse of dimensionality, and would generally result in poor estimation efficiency. Moreover, the generation process of the response is described much like a 'black-box' mechanism by the single joint multivariate function f which consists of all covariates, making the model hard to be interpretable. A linear model

$$Y = \sum_{j=0}^{p} \beta_j X_j + \epsilon, \tag{1.2}$$

on the other extreme, is highly interpretable due to its assumed linear additive structure. Moreover, the additive structure provides a convenient assessment of the individual contribution from each predictor variable. However, a reliance on the rigid parametric form limits its ability to model nonlinear effects of the predictor variables.

Different approaches have been proposed to remedy the disadvantages of general nonparametric regression models and linear models, which can achieve a higher degree of model flexibility than linear models, and obtain better interpretability and computation efficiency than nonparametric regression models. One approach is to transform response Y such as Box-Cox transformations, which lead to

$$T(Y) = \sum_{j=0}^{p} \beta_j X_j + \epsilon. \tag{1.3}$$

Box and Cox (1964) proposed a family of power transformations on the response for $T(Y)$, which aims to make the assumptions of linearity, normality and homogeneous variance in linear models more appropriate after transformation. Additive models in Stone (1985),

$$Y = \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{1.4}$$

which are different from Box-Cox transformations, allow transformations on each predictor variable. Additive models assume that each additive component is a univariate smooth function of a single predictor variable, thus providing nonparametric extensions of linear models and can offer a higher degree of flexibility. And the additive combination of univariate functions is more interpretable and easier to fit than general nonparametric models. Despite the popularity of Box-Cox transformations and additive models, their effectiveness are still vulnerable to model mis-specifications, and they could be ineffective for simple cases like $Y = \log(X_1 + X_2^2 + \epsilon)$. In addition, another drawback of Box-Cox transformations is that the parametric form of transformation on the response can be restrictive in some applications.

To further improve the model flexibility and interpretability from Box-Cox transformations and additive models, transformation models are proposed, where general nonparametric transformations are applied to both Y and $\mathbf{X}$. Transformation models are formulated as

$$h(Y) = \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{1.5}$$

where h and $f_j, j = 1, \ldots, p$, are arbitrary measurable functions of corresponding random variables. Under certain conditions, it is shown that transformations h and $f_j, j = 1, \ldots, p$, are identifiable and different estimation procedures have been proposed (Linton et al., 2008;

Chiappori et al., 2015). With the strengths provided by nonlinear transformations and additive structure, transformation models achieve a good balance in model flexibility and interpretability.

For data analysis, it is an ideal case that the underlying dependence structure between Y and $\mathbf{X}$ is known so that a precise model can be specified and corresponding model parameters can be accurately estimated. However, such prior knowledge is seldom given in practice. To explore their relationship, it is a common practice to apply different model structures to approximate the true structure. The choice of a specific model involves different considerations over various factors such as model flexibility, interpretability, computational efficiency, prediction accuracy, etc. To combine the advantages of both nonlinear transformation and the additive structure as in (1.5), we consider optimal transformation defined in Breiman and Friedman (1985) and propose several methods in the areas of variable screening, variable selection, dependence measure, sufficient dimension reduction, etc.

1.1 Optimal Transformation

1.1.1 Formal Definition

Breiman and Friedman (1985) proposed to apply general nonparametric transformations to both Y and $\mathbf{X}$ and considered optimal transformations by solving a minimization problem.

$$\begin{aligned} \min_{h \in L^2(P_Y), f_j \in L^2(P_{X_j})} \quad & \mathrm{E}\Big[\{h(Y) - \sum_{j=1}^{p} f_j(X_j)\}^2\Big], \\ \text{s.t.} \quad & \mathrm{E}[h(Y)] = \mathrm{E}[f_j(X_j)] = 0; \\ & \mathrm{E}[h^2(Y)] = 1, \mathrm{E}[f_j^2(X_j)] < \infty. \end{aligned} \tag{1.6}$$

Here, P_Y and P_{X_j} denote the marginal distributions of Y and X_j, respectively, and $L^2(P)$ denotes the class of square integrable functions under the measure *P*. We denote the solution to (1.6) as h^* and $f_j^*(j = 1, \ldots, p)$, which are referred to as the optimal transformations for Y and $\mathbf{X}$, respectively. Problem (1.6) tries to find transformations that produce the best-fitting additive model. Knowledge of such transformations can aid in the interpretation and understanding the relationship between the response and predictors. From the aspect of applying transformation, both Box-Cox transformations and additive models can be considered as special cases of optimal transformations.

A set of sufficient conditions is given in Breiman and Friedman (1985, Section 5.2) for the existence of optimal transformations. Note that under some restrictive conditions, the optimal transformations from (1.6) are equivalent to the transformations in regression model (1.5). However, the equivalence property does not hold in general. The necessary conditions which ensure the equivalence property is still an open research question. Despite this theoretical gap, the optimal transformation approach is still a useful statistical tool in exploring the relationship between the response and predictor variables. In addition, it provides a general framework under which several methods can be proposed.

1.1.2 Applications of Optimal Transformations

Based on optimal transformation, we propose several methods to deal with different statistical problems in next few chapters, including variable screening, sparse nonparametric regression, dependence measure and sufficient dimension reduction. Here, we briefly introduce these methods and show their connections with optimal transformations.

Variable Screening

Variable screening is the process of filtering out irrelevant variables, with the aim to reduce the dimensionality from ultrahigh to high while retaining all important variables prior to model building. In Chapter 2, we propose a screening procedure based on a dependence measure maximum correlation (Rényi, 1959), which is defined by

$$\rho^*(Y, X) = \sup_{\theta,\phi}\{\rho\left(\theta(Y), \phi(X)\right) : 0 < \mathrm{E}\{\theta^2(Y)\} < \infty, 0 < \mathrm{E}\{\phi^2(X)\} < \infty\}, \quad (1.7)$$

where ρ is the Pearson correlation, and θ and ϕ are Borel-measurable functions of univariate random variables Y and X.

Breiman and Friedman (1985) derived the relationship between the optimal transformations from (1.6) and maximum correlation. For bivariate cases where $p = 1$, the optimal transformations are equivalent to the transformations that yield maximum correlation. Since maximum correlation is a measure that can sensitively capture dependence between the response and the predictor variable in univariate cases, we build a screening procedure which ranks the predictor variables according to their marginal maximum correlations with the response. Maximum correlation is not directly computable because the maximization in (1.7) is taken over infinite-dimensional spaces. Therefore, we approximate the optimal transformations in order to numerically evaluate maximum correlation. The resulting procedures are essentially proposed based on optimal transformations for univariate cases with $p = 1$.

Sparse Nonparametric Regression

Optimal transformations only enjoy good statistical and computational behaviors when the number of variables p is not large to the sample size n, their usefulness is limited in the high dimensional setting. In Chapter 3, we extend optimal transformations to deal with high dimensional problems by proposing a sparse version of optimal transformations, which penalizes the sum of L_2 norm of each function component f_j in (1.6). The resulting optimal transformations encourage parsimonious solutions and perform model selection and parameter estimation simultaneously. To make the optimal transformation interpretable and suitable for regression analysis, we further consider monotone transformation on the response Y.

Dependence Measures

Due to the fact that maximum correlation between random variables X and Y is zero if and only if X and Y are independent, maximum correlation can be applied in testing the hypothesis "random variables X and Y are independent". Beside the maximum correlation r_1 and the optimal transformations θ_1, ϕ_1 defined by

$$r_1 = \max_{\theta_1, \phi_1 \in L_2(P)} \rho\left(\theta_1(Y), \phi_1(X)\right), \tag{1.8}$$

one can also define subsequent maximum correlations and optimal transformations. For functions $\{\theta_i, \phi_i; i = 1, 2, \ldots\}$ with bounded positive second moments, let

$$\begin{aligned} r_i = \max_{\theta_i, \phi_i \in L_2(P)} & \rho\left(\theta_i(Y), \phi_i(X)\right), \\ & \langle \theta_i(Y), \theta_j(Y) \rangle_{L_2(P_Y)} = 0, \\ & \langle \phi_i(X), \phi_j(X) \rangle_{L_2(P_X)} = 0, \end{aligned} \tag{1.9}$$

for all $j = 1, \ldots, i-1$. Here, $\langle \cdot, \cdot \rangle$ is the inner product defined in corresponding L_2 spaces.

Under independence of random variables X and Y, all the values of r_i's are zero. Based on this property, we propose several independence measures. Since all correlations r_i are not directly computable, we again approximate optimal transformations in order to numerically evaluate maximum correlation. Under the framework of optimal transformations, we develop dependence measures by approximating optimal transformations using B-spline basis functions. Given a sample, the optimal transformations are obtained by solving an equivalent eigen problem. Additionally, eigenvalues from the eigen problem correspond to the values of r_i's. In Chapter 4, we apply the leading eigenvalue, as well as the sum of all eigenvalues for measuring dependence.

Sufficient Dimension Reduction

The goal of a traditional linear sufficient dimension reduction procedure is to find a few linear combinations $\beta_1^\top \mathbf{X}, \ldots, \beta_d^\top \mathbf{X}$ that can fully represent $\mathbf{X}$, without loss of information on Y. It is required that those linear combinations satisfy the constraints,

$$Y \perp\!\!\!\perp \mathbf{X} | \{\beta_1^\top \mathbf{X}, \ldots, \beta_d^\top \mathbf{X}\}.$$

That is, Y is conditionally independent of $\mathbf{X}$ given $\{\beta_1^\top \mathbf{X}, \ldots, \beta_d^\top \mathbf{X}\}$. Equivalently, the dependence structure of Y on $\mathbf{X}$ is expressed by the regression model

$$Y = f(\beta_1^\top \mathbf{X}, \ldots, \beta_d^\top \mathbf{X}, \epsilon).$$

Li (1991) proposed Sliced Inverse Regression (SIR) that can recover the space spanned by $\beta_1, \ldots, \beta_d$ under some mild conditions. SIR is connected to a maximization problem as follows. Define

$$R^2(b) = \max_T \rho(T(Y), b^\top \mathbf{X}) \tag{1.10}$$

where ρ is the Pearson correlation, T is any squared integrable function, and b is a vector of length p. We look for the direction b_1 which maximizes $R^2(b)$, and continue to find subsequent directions $b_2, \ldots, b_d$, satisfying the following conditions.

$$\begin{aligned} &\mathrm{Cov}(b_i^\top \mathbf{X}, b_j^\top \mathbf{X}) = 0, \text{ for } i \neq j \\ &R^2(b_i) = \max_b R^2(b) \end{aligned} \tag{1.11}$$

It is shown in Chen and Li (1998) that the resulting directions $b_1, \ldots, b_d$ are equivalent to the directions obtained by SIR. Therefore, solving the maximization problem above can be viewed as a procedure to recover the space spanned by $\{\beta_1, \ldots, \beta_d\}$.

One possible way to improve SIR is to generalize dimension reduction from linear to nonlinear cases, where we consider additive terms of transformed $\mathbf{X}$ instead of its linear combinations. We apply the optimal transformations and extract the transformations of $\mathbf{X}$ successively, similar to the procedure described in (1.8) and (1.9) of extracting the sequence of maximum correlations. This direction of research is briefly discussed at the end of Chapter 4.

For comparison purposes, we review some existing methods on variable screening and variable selection in high dimension data analysis.

1.2 Review on Variable Screening Methods

In a seminar paper, Fan and Lv (2008) proposed Sure Independence Screening (SIS) for screening variables in linear models. More screening procedures are developed after SIS for other specific models, including screening methods for generalized linear models (Fan and Song, 2010), multi-index models (Zhu et al., 2011) and additive models (Fan et al., 2011), varying coefficient models (Fan et al., 2014), etc. Another kind of screening procedures is developed without imposing any specific model assumption, for example, the distance correlation-based sure independence screening Li et al. (2012b). In this section, we review three typical screening methods.

1.2.1 Sure Independence Screening (SIS)

Consider a linear regression model

$$Y = \sum_{j=0}^{p} \beta_j X_j + \epsilon \tag{1.12}$$

where ϵ is a random error. Fan and Lv (2008) suggested ranking all predictors according to their marginal Pearson correlations with the response and select the top predictors with relatively larger Pearson correlation values with a given sample. Let $w_j = \rho(Y, X_j)$ where ρ denotes the Pearson correlation, and $\widehat{w_j}$ be its sample estimates from n observations. SIS retains the following set of predictors.

$$\widehat{\mathcal{M}_\gamma} = \{1 \le j \le p : |\widehat{w_j}| \text{ is among the first } [\gamma n] \text{ largest of all}\}$$

where γ is a pre-defined constant with $\gamma \in (0, 1)$, and $[\gamma n]$ denote the integer part of γn. For linear model (1.12), the true set of important predictors is defined as

$$\mathcal{M}_\star = \{1 \le j \le p : \beta_j \ne 0\}.$$

Under some regularity conditions, Fan and Lv (2008) showed that SIS possesses the sure screening property in the ultrahigh dimensional setting, that is,

$$\Pr(\mathcal{M}_\star \subseteq \widehat{\mathcal{M}_\gamma}) \to 1, \text{ as } n \to \infty.$$

1.2.2 Nonparameteric Independence Screening (NIS)

To screening features in ultrahigh dimensional additive model

$$Y = \sum_{j=0}^{p} m_j(X_j) + \epsilon \tag{1.13}$$

where $\mathrm{E}\{m_j(X_j)\} = 0$. Fan et al. (2011) proposed to rank all predictors according to $\mathrm{E}\{f_j^2(X_j)\}$ where $f_j(X_j) = \mathrm{E}(Y|X_j)$ is the projection of Y on X_j. Given data $\{Y_i\}_{i=1}^n$

and $\{X_{ij}\}_{i=1}^n$, the function $f_j(X_j)$ can be estimated through any basis expansion methods such as B-splines. Denote its sample estimate as $\widehat{f_{nj}}$, NIS retains the following set of predictors.

$$\widehat{\mathcal{M}_\nu} = \{1 \le j \le p : ||\widehat{f_{nj}}||_n^2 \ge \nu_n\}$$

where $||\widehat{f_{nj}}||_n^2 = n^{-1}\sum_{i=1}^n \widehat{f_{nj}}(X_{ij})$ and ν_n is a pre-specified value. For additive model (1.13), the true set of important predictors is defined as

$$\mathcal{M}_\star = \{1 \le j \le p : \mathrm{E}m_j^2(X_j) > 0\}.$$

Under some regularity conditions, Fan et al. (2011) showed that NIS possesses the sure screening property for additive models.

1.2.3 Distance Correlation-based Sure Independence Screening (DC-SIS)

Both SIS and NIS are proposed for targeted classes of specified models and may become ineffective when the model is mis-specified. To overcome this difficulty, Li et al. (2012b) proposed a model-free screening procedure, DC-SIS, to screen features in the ultrahigh dimensional setting, without imposing any specific model assumptions. DC-SIS uses a dependence measure called distance correlation introduced in Szekely et al. (2007) to rank the predictor variables. The distance correlation between two random vector $\mathbf{u} \in R^{d_u}$ and $\mathbf{v} \in R^{d_v}$ is defined by

$$\mathrm{dcorr}(\mathbf{u}, \mathbf{v}) = \frac{\mathrm{dcov}(\mathbf{u}, \mathbf{v})}{\sqrt{\mathrm{dcov}(\mathbf{u}, \mathbf{u})\mathrm{dcov}(\mathbf{v}, \mathbf{v})}}$$

where $\mathrm{dcov}(\cdot, \cdot)$ is called distant covariance and defined as follows.

$$\mathrm{dcov}^2(\mathbf{u}, \mathbf{v}) = \int_{R^{d_u+d_v}} \|\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s}) - \phi_{\mathbf{u}}(\mathbf{t})\phi_{\mathbf{v}}(\mathbf{s})\|^2 \, w(\mathbf{t}, \mathbf{s}) \, d\mathbf{t} \, d\mathbf{s}$$

where $\phi_{\mathbf{u}}(\mathbf{t})$ and $\phi_{\mathbf{v}}(\mathbf{s})$ are the respective characteristic functions of the random vectors $\mathbf{u}$ and $\mathbf{v}$, $\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s})$ is the joint characteristic function of $\mathbf{u}$ and $\mathbf{v}$, and

$$w(\mathbf{u}, \mathbf{v}) = \{c_{d_u} c_{d_v} ||\mathbf{t}||_{d_u}^{1+d_u} ||\mathbf{s}||_{d_v}^{1+d_v}\}^{-1}$$

with $c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$ and Γ being the Gamma function.

Distant correlation is a generalization of the Pearson correlation and can be used to capture nonlinear relationships between any two random vectors. Denote the sample estimates of distant correlation between Y and X_j by $\widehat{\mathrm{dcorr}}(Y, X_j)$, DC-SIS ranks the predictors according to $\widehat{\mathrm{dcorr}}^2(Y, X_j)$ and retains the set of predictors

$$\widehat{\mathcal{M}} = \{1 \leq j \leq p : \widehat{\mathrm{dcorr}}^2(Y, X_j) \geq cn^{-\kappa}\}.$$

Define the true set of important predictors by

$$\mathcal{M}_\star = \{1 \leq j \leq p : F(Y|\mathbf{X}) \text{ functionally depends on } X_j\},$$

Li et al. (2012b) proved that DC-SIS has the sure screening property under some regularity conditions, without imposing any specific model assumptions.

1.3 Review of Variable Selection Method in Regression

Classical variable selection procedures, which differ from variable screening, perform model selection and parameter estimation simultaneously. The majority of these procedures select variables by minimizing a penalized objective function with the following form.

$$\text{Loss function} + \text{Penalization} \tag{1.14}$$

The most popular choices of loss functions are least squares, negative log-likelihood, and their variants. The penalization part penalizes model complexity and encourages sparsity in the final model. Early methods of variable selection include best subset selection or stepwise (forward/backward) selection with a criterion like Akaike information criterion (AIC) (Akeike, 1973), Bayesian information criterion (BIC) (Schwarz et al., 1978), Mallow's C_p (Mallows, 1973), etc. These methods are computational expensive and quickly becomes infeasible as dimensionality grows. Furthermore, the subset selection approaches suffer from instability and their theoretical properties are difficult to examine (Breiman, 1996). In high dimensional data analysis, regularization methods have been proposed to overcome these difficulties. We review some popular methodologies on variable selection in both linear models and additive models.

1.3.1 The Lasso and Its Variants

For linear models (1.2), a standard way of performing variable selection is to penalized least square with a proper choice of the penalty function. One example is the bridge estimator (Frank and Friedman, 1993) which uses the ℓ_q-norm ($q > 0$) of the slope coefficients. When $0 < q \leq 1$, some slope estimate can be exactly zero with proper choices of tuning parameters.

Among all bridge estimators with different choices of q, the most popular estimator is the one with $q = 1$, known as the least shrinkage and selection operator (Lasso) proposed in Tibshirani (1996). The Lasso estimates of the coefficients are the solution to the following optimization problem.

$$\min_{\beta_1,\ldots,\beta_p} \mathrm{E}\left[\left(Y - \sum_{j=0}^{p} \beta_j X_j\right)^2\right] \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq L; \tag{1.15}$$

which is also equivalent to the standard form as in (1.14),

$$\min_{\beta_1,\ldots,\beta_p} \quad \mathrm{E}\left[\left(Y - \sum_{j=0}^{p} \beta_j X_j\right)^2\right] + \lambda \sum_{j=1}^{p} |\beta_j|; \tag{1.16}$$

where L and λ are tuning parameters.

Least Angle Regression (LARS) algorithm (Efron et al., 2004) gives the entire solution path of the Lasso estimate. In addition, Lasso estimates can also be computed efficiently via coordinate descent algorithms (Fu, 1998; Friedman et al., 2007). It is shown that Lasso can consistently select the true model under the Irrepresentable Condition (Zhao and Yu, 2006).

Other variants of Lasso includes the grouped lasso (Yuan and Lin, 2006), the elastic net (Zou and Hastie, 2005), the fussed lasso (Tibshirani et al., 2005), the adaptive lasso (Zou, 2006), etc. Beside the ℓ_1 penalty, other penalty functions are investigated in the literature, examples include the SCAD (Fan and Li, 2001) and MCP (Zhang, 2010).

1.3.2 Variable Selection in Additive Models

There are several approaches to generalize variable selection from linear to non-linear models, in particular, the additive models (1.4). One typical example is the Sparse Additive Model (SPAM) proposed in Ravikumar et al. (2007). They consider a modification of standard additive model optimization problem as follows.

$$\begin{aligned} \min_{g_j \in \mathcal{H}_{\mathcal{X}_j}} \quad & \mathrm{E}\left[\left(Y - \sum_{j=1}^{p} \beta_j g_j(X_j)\right)^2\right] \\ \text{s.t.} \quad & \sum_{j=1}^{p} |\beta_j| \le L, \mathrm{E}[g_j^2(X_j)] = 1; \end{aligned} \tag{1.17}$$

where L is a pre-defined constant.

Denote $\beta = (\beta_1, \dots, \beta_p)^\top$. Then, the constraint that β lies in the ℓ_1 ball $\{\beta : ||\beta||_1 \le 1\}$ encourages sparsity of the estimated β, just as for the Lasso (Tibshirani, 1996).

Let $f_j = \beta_j g_j$, we can re-express the minimization problem (1.17) in the following equivalent Lagrangian form:

$$\frac{1}{2}\mathrm{E}\left[\left(Y - \sum_{j=1}^{p} \beta_j f_j(X_j)\right)^2\right] + \lambda \sum_{j=1}^{p} \sqrt{\mathrm{E}[f_j^2(X_j)]} \tag{1.18}$$

where λ is the regularization parameter.

Ravikumar et al. (2007) developed a backfitting algorithm, named SPAM, to estimate the functions f_j $(j = 1, \dots, p)$ for a given sample. They further showed that SPAM can consistently select all important functional components under some regularity conditions.

Other approaches of variable selection in additive models include Meier et al. (2009), Huang et al. (2010) and Balakrishnan et al. (2012), where different penalty functions are used to produce sparse estimates of the functional components.

1.4 Outline

In this thesis, we study and propose several new methodologies for variable screening, sparse nonparametric regression, dependence measures and dimension reduction, under the unified framework with optimal transformations. In Chapter 2, we develop a maximum correlation-based sure independence screening (MC-SIS) procedure to screen features in an ultrahigh-dimensional setting. In Chapter 3, we develop a novel method called SParse Optimal Transformations (SPOT) to simultaneously select important variables and explore relationships between the response and predictor variables in high dimensional nonparametric regression analysis. In Chapter 4, we propose several dependence measures based on maximum correlation and B-spline approximation, and discuss the application of optimal transformations in nonlinear sufficient dimension reduction. Chapter 5 summaries the results of this thesis.

2. MODEL-FREE SURE SCREENING VIA MAXIMUM CORRELATION

2.1 Introduction

With the rapid development of modern technology, various types of high-dimensional data are collected in a variety of areas such as next-generation sequencing and biomedical imaging data in bioinformatics, high-frequency time series data in quantitative finance, and spatial-temporal data in environmental studies. In those types of high-dimensional data, the number of variables *p* can be much larger than the sample size *n*, which is referred to as the 'large p small n' scenario. To deal with this scenario, a commonly adopted approach is to impose the sparsity assumption that the number of important variables is small relative to p. Based on the sparsity assumption, a variety of regularization procedures have been proposed for high-dimensional regression analysis such as the lasso (Tibshirani, 1996), the smoothly clipped absolute deviation method (Fan and Li, 2001), and the elastic net (Zou and Hastie, 2005). All these methods work when p is moderate. However, when applied to analyze ultrahigh-dimensional data where dimensionality grows exponentially with sample size (e.g., $p = \exp(n^{\alpha})$ with $\alpha > 0$), their performances will deteriorate in terms of computational expediency, statistical accuracy and algorithmic stability (Fan et al., 2009). To address the challenges of ultrahigh dimensionality, a number of marginal screening procedures have been proposed under different model assumptions. They all share the same goal that is to reduce dimensionality from ultrahigh to high while retaining all truly important variables. When a screening procedure achieves this goal, it is said to have the sure screening property in the literature.

Fan and Lv (2008) proposed to use the Pearson correlation for feature screening and showed that the resulting procedure possesses the sure screening property under the linear model assumption. They refer to the procedure as the Sure Independence Screening

(SIS) procedure. Fan and Song (2010) extended SIS from linear models to generalized linear models by using maximum marginal likelihood values. Fan et al. (2011) developed a Nonparametric Independence Screening (NIS) procedure and proved that NIS has the sure screening property under the additive model. Li et al. (2012b) proposed to use distance correlation to rank the predictor variables, and showed that the resulting procedure, denoted as DC-SIS, has the sure screening property without imposing any specific model assumptions. Compared with the other screening procedures discussed previously, DC-SIS is thus model-free. Distance correlation was introduced in Szekely et al. (2007), which uses joint and marginal characteristic functions to measure the dependence between two random variables. We briefly review the SIS, NIS, DC-SIS procedures here.

From the review above, it is clear that the standard approach to developing a valid screening procedure consists of two steps. First, a proper dependence measure between the response and predictor variables needs to be defined and further used to rank-order all the predictor variables; and second, the sure screening property needs to be established for the screening procedure based on the dependence measure. The screening methods discussed previously differ from each other in these two steps. For example, SIS uses the Pearson correlation as the dependence measure and possesses the sure screening property under linear models, whereas NIS uses the goodness of fit measure of the nonparametric regression between the response and predictor variable as the dependence measure and possesses the sure screening property under additive models.

For the purpose of screening in an ultrahigh dimensional setting, we argue that an effective screening procedure should employ a sensitive dependence measure and satisfy the sure screening requirement without model specifications. The goal of screening is not to precisely select the true predictors, instead, it is to reduce the number of predictor variables from ultrahigh to high while retaining the true predictor variables. Therefore, false positives or selections can be tolerated to a large degree, and sensitive dependence measures are more preferred than insensitive measures. In ultrahigh dimensional data, there usually does not exist information about the relationship between the response and predictor variables, and it is extremely difficult to explore the possible relationship due to the presence of a large number of predictors. Therefore, model assumptions should be avoided as much as possible in ultrahigh dimensional screening, and we should prefer screening procedures that possess the sure screening property without model specifications. In other words, model-free sure screening procedures are more preferable. Among the existing screening procedures discussed previously, only DC-SIS is model-free because it does not require any restrictive model assumption. However, the distance correlation measure used by DC-SIS may not be sensitive especially when the sample size is small, because empirical characteristic functions are employed to estimate distance correlations.

A more sensitive dependence measure between the response and a predictor variable is the maximum correlation, which was originally proposed by Gebelein (1941) and studied by Rényi (1959) as a general dependence measure between two random variables. Rényi (1959) listed seven fundamental properties that a reasonable dependence measure must have, and maximum correlation is one of a few measures that can satisfy this requirement. The definition and estimation of maximum correlation involve maximizations over functions (see Section 2.2.1), and thus it is fairly sensitive even when the sample size is small. Recently, there have been some revived interests in using maximum correlation as a proper dependence measure in high-dimensional data analysis (Bickel and Xu, 2009; Hall and Miller, 2011; Reshef et al., 2011; Speed, 2011).

We propose to use maximum correlation as a dependence measure for ultrahigh dimensional screening, and prove that the resulting procedure has the sure screening property without imposing model specifications (see Theorem 2.2.2 in Section 2.2.4). We adopt the B-spline functions-based estimation method from Burman (1991) to estimate maximum correlation. We refer to our proposed procedure as the Maximum Correlation-based Sure Independence Screening procedure, or in short, the MC-SIS procedure. Numerical results show that MC-SIS is competitive to other existing model-based screening procedures, and is more sensitive and robust than DC-SIS when the sample size is small or the distributions of the predictor variables have heavy tails.

The rest of this chapter is organized as follows. In Section 2.2, we introduce maximum correlation and the B-spline functions-based method for estimating maximum correlation, propose the MC-SIS procedure, and establish the sure screening property for MC-SIS. In Section 2.3, we develop a three-step procedure for selecting tuning parameters for MC-SIS in practice. Section 2.4 presents results from simulation studies and a real life screening application. Section 2.5 provides additional remarks on the screening methods and future research directions. The proofs of the theorems are given in Section 2.6.

2.2 Independence Screening via Maximum Correlation

In this section, we formally introduce the proposed screening procedure MC-SIS, which uses maximum correlation as the dependence measure. We first introduce its connection to optimal transformation in Section 2.2.1, and then propose to use B-spline function to approximate optimal transformation in Section 2.2.2, which leads to a proper approximated evaluation of maximum correlation. Based on the approximation, we propose MC-SIS in Section 2.2.3. Sure screening property of MC-SIS is established in Section 2.2.4.

2.2.1 Maximum correlation and optimal transformation

Recall that Y is the response variable and $\mathbf{X} = (X_1, \ldots, X_p)$ the vector of predictor variables. We assume the supports of Y and X_j $(j = 1, \ldots, p)$ are compact, and they are further assumed to be [0,1] without loss of generality. For any given j, consider a pair of random variables (X_j, Y). The maximum correlation coefficient between X_j and Y, denoted as ρ_j^*, is defined as follows.

$$\rho_j^*(X_j, Y) = \sup_{\theta,\phi}\{\rho\left(\theta(Y), \phi(X_j)\right) : 0 < \mathrm{E}\{\theta^2(Y)\} < \infty, 0 < \mathrm{E}\{\phi^2(X_j)\} < \infty\}, \quad (2.1)$$

where ρ is the Pearson correlation, and θ and ϕ are Borel-measurable functions of Y and X_j. We further denote θ_j^* and ϕ_j^* as the optimal transformations that attain the maximum correlation.

Maximum correlation coefficient enjoys the following properties given in Rényi (1959):

(a) $0 \leq \rho_j^*(X_j, Y) \leq 1$;

(b) $\rho_j^*(X_j, Y) = 0$ if and only if X_j and Y are independent;

(c) $\rho_j^*(X_j, Y) = 1$ if there exist Borel-measurable functions θ^* and ϕ^* such that $\theta^*(Y) = \phi^*(X_j)$;

(d) if X_j and Y are jointly Gaussian, then $\rho_j^*(X_j, Y) = |\rho(X_j, Y)|$.

Some other properties of maximum correlation coefficient are discussed in Szekely and Mori (1985), Dembo et al. (2001), Bryc and Dembo (2005), and Yu (2008). Due to Property (d), it is clear that maximum correlation is a natural extension of the Pearson correlation. Note that the Pearson correlation does not possess Properties (b) and (c). For Property (c), there are cases that the Pearson correlation coefficient can be as low as zero when Y is functionally determined by X_j. For example, if $Y = X_1^2$ where $X_1 \sim \mathcal{N}(0, 1)$, the Pearson correlation between Y and X_1 is zero, whereas the maximum correlation is one. Therefore, maximum correlation is a more proper measure of the dependence between two random variables than the Pearson correlation.

Rényi (1959) established the existence of maximum correlation under certain sufficient conditions, and a different set of sufficient conditions are given in Breiman and Friedman (1985). Breiman and Friedman (1985) also showed that optimal transformations θ_j^* and ϕ_j^* can be obtained via the following minimization problem.

$$\begin{aligned} \min_{\theta_j, \phi_j \in L_2(P)} \quad & e_j^2 = \mathrm{E}[\{\theta_j(Y) - \phi_j(X_j)\}^2], \\ \text{subject to} \quad & \mathrm{E}\{\theta_j(Y)\} = \mathrm{E}\{\phi_j(X_j)\} = 0; \\ & \mathrm{E}\{\theta_j^2(Y)\} = 1. \end{aligned} \tag{2.2}$$

Here, *P* denotes the joint distribution of $(X_j$,$Y)$ and $L_2(P)$ is the class of square integrable functions under the measure *P*. Let e_j^{*2} be the minimum of e_j^2. Breiman and Friedman (1985) derived two critical connections between e_j^{*2}, squared maximum correlation ρ_j^{*2}, and optimal transformation ϕ_j^*, which we state as *Fact 0* below.

Fact 0.

$$e_j^{*2} = 1 - \rho_j^{*2}; \tag{2.3a}$$

$$\mathrm{E}(\phi_j^{*2}) = \rho_j^{*2}. \tag{2.3b}$$

Fact 0 suggests that the minimization problem (2.2) is equivalent to the optimization problem (2.1). Furthermore, the squared maximum correlation coefficient is equal to the expectation of the squared optimal transformation ϕ_j^*.

Various algorithms have been proposed in the literature to compute maximum correlation, including Alternating Conditional Expectations (ACE) in (Breiman and Friedman, 1985), B-spline approximation in Burman (1991), and polynomial approximation in Bickel and Xu (2009) and Hall and Miller (2011). Equation (2.3b) indicates that maximum correlation coefficient ρ_j^* can be calculated through the optimal transformation ϕ_j^*. In this chapter, we apply Burman's approach to first estimate ϕ_j^*, and then estimate ρ_j^*, which will be further used in screening.

2.2.2 B-spline estimation of optimal transformations

Let $\mathcal{S}_n$ be the space of polynomial splines of degree $\ell \geq 1$ and $\{B_{jm}, m = 1, \ldots, d_n\}$ denote a normalized B-spline basis with $||B_{jm}||_{\sup} \leq 1$, where $||\cdot||_{\sup}$ is the sup-norm. We have $\theta_{nj}(Y) = \boldsymbol{\alpha}_j^\top \mathbf{B}_j(Y)$, $\phi_{nj}(X_j) = \boldsymbol{\beta}_j^\top \mathbf{B}_j(X_j)$ for any $\theta_{nj}(Y), \phi_{nj}(X_j) \in \mathcal{S}_n$, where $\mathbf{B}_j(\cdot) = (B_{j1}(\cdot), \ldots, B_{jd_n}(\cdot))^\top$ denotes the vector of d_n basis functions. Additionally, we let k be the number of knots where $k = d_n - \ell$. One example of the B-spline basis functions is depicted in Figure 2.1.

The population version of B-spline approximation to the minimization problem (2.2) can be written as follows.

$$\begin{aligned} \min_{\theta_{nj}, \phi_{nj} \in \mathcal{S}_n} \quad & \mathrm{E}[\{\theta_{nj}(Y) - \phi_{nj}(X_j)\}^2], \\ \text{subject to} \quad & \mathrm{E}\{\theta_{nj}(Y)\} = \mathrm{E}\{\phi_{nj}(X_j)\} = 0; \\ & \mathrm{E}\{\theta_{nj}^2(Y)\} = 1. \end{aligned} \tag{2.4}$$

Burman (1991) applied a technique to remove the first constraint $\mathrm{E}\{\theta_{nj}(Y)\} = \mathrm{E}\{\phi_{nj}(X_j)\} = 0$ in the optimization problem above as follows. First, let $\mathbf{z}_1, \ldots, \mathbf{z}_{d_n - 1}$ ($\mathbf{z}_i = (z_{i1}, \ldots, z_{id_n})^\top$ for $i = 1, \ldots, d_n - 1$) be d_n-dimensional vectors which are orthogonal to each other, orthogonal to the vector of 1's and $\mathbf{z}_i^T \mathbf{z}_i = 1$ for $i = 1, \ldots, d_n - 1$. Second, obtain matrix $\mathbf{D}_j$ with the (s, m)-entry $\mathbf{D}_{j,sm} = z_{sm}/(kb_{jm})$ where $b_{jm} = \mathrm{E}\{B_{jm}(X_j)\}$, for $s = 1, \ldots, d_n - 1$

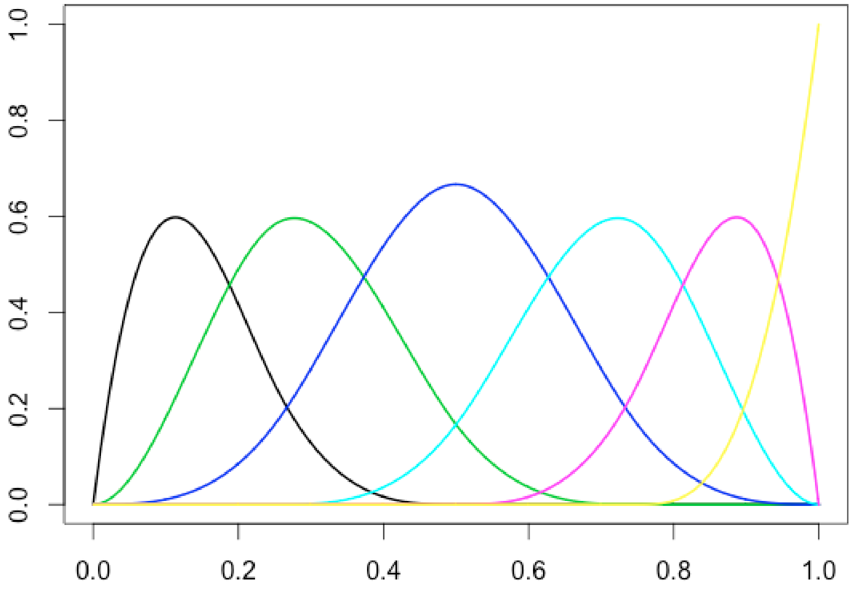

Figure 2.1. A example of cubic B-spline basis functions

and $m = 1, \ldots, d_n$. Third, let $\phi_{nj}(X_j) = \boldsymbol{\eta}_j^\top \boldsymbol{\psi}_j(X_j)$ where $\boldsymbol{\psi}_j(X_j) = \mathbf{D}_j \mathbf{B}_j(X_j)$. With this construction, it is easy to verify that $\mathrm{E}\{\phi_{nj}(X_j)\} = 0$, and the minimization of $\mathrm{E}[\{\theta_{nj}(Y) - \phi_{nj}(X_j)\}^2]$ subject to $\mathrm{E}\{\theta_{nj}^2(Y)\} = 1$ ensures that $\mathrm{E}\{\theta_{nj}(Y)\} = 0$. Burman (1991) showed the equivalence between the optimization problem (2.4) and the one stated below.

$$\begin{aligned} \min_{\theta_{nj}, \phi_{nj} \in \mathcal{S}_n} \quad & \mathrm{E}[\{\theta_{nj}(Y) - \phi_{nj}(X_j)\}^2], \\ \text{subject to} \quad & \mathrm{E}\{\theta_{nj}^2(Y)\} = 1. \end{aligned} \tag{2.5}$$

For fixed $\theta_{nj}(Y)$ (i.e., fixed $\boldsymbol{\alpha}_j$), the minimizer of (2.5) with respect to $\boldsymbol{\eta}_j$ and $\phi_{nj}(X_j)$ are

$$\begin{aligned} \boldsymbol{\eta}_j &= [\mathrm{E}\{\boldsymbol{\psi}_j(X_j)\boldsymbol{\psi}_j^\top(X_j)\}]^{-1}\mathrm{E}\{\boldsymbol{\psi}_j(X_j)\mathbf{B}_j^\top(Y)\}\boldsymbol{\alpha}_j, \\ \phi_{nj}(X_j) &= \boldsymbol{\psi}_j^\top(X_j)[\mathrm{E}\{\boldsymbol{\psi}_j(X_j)\boldsymbol{\psi}_j^\top(X_j)\}]^{-1}\mathrm{E}\{\boldsymbol{\psi}_j(X_j)\mathbf{B}_j^\top(Y)\}\boldsymbol{\alpha}_j. \end{aligned} \tag{2.6}$$

By plugging $\phi_{nj}(X_j)$ back in (2.5), we obtain the following maximization problem,

$$\begin{aligned}&\max_{\boldsymbol{\alpha}_j\in\mathbb{R}^{d_n}} \quad \boldsymbol{\alpha}_j^\top \mathrm{E}\{\mathbf{B}_j(Y)\boldsymbol{\psi}_j^\top(X_j)\}[\mathrm{E}\{\boldsymbol{\psi}_j(X_j)\boldsymbol{\psi}_j^\top(X_j)\}]^{-1}\mathrm{E}\{\boldsymbol{\psi}_j(X_j)\mathbf{B}_j^\top(Y)\}\boldsymbol{\alpha}_j,\\ &\text{subject to} \quad \boldsymbol{\alpha}_j^\top \mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(Y)\}\boldsymbol{\alpha}_j = 1.\end{aligned} \tag{2.7}$$

Following the notation in Burman (1991), we denote

$$\mathbf{A}_{j00} = \mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(Y)\}, \qquad \mathbf{A}_{jXX} = \mathrm{E}\{\boldsymbol{\psi}_j(X_j)\boldsymbol{\psi}_j^\top(X_j)\},$$
$$\mathbf{A}_{jX0} = \mathrm{E}\{\boldsymbol{\psi}_j(X_j)\mathbf{B}_j^\top(Y)\}, \quad \text{and} \quad \mathbf{A}_{j0X} = \mathbf{A}_{jX0}^\top.$$

It is clear that (2.7) is a generalized eigenvalue problem, which can be solved by the largest eigenvalue and its corresponding eigenvector of $\mathbf{A}_{j00}^{-1/2}\mathbf{A}_{j0X}\mathbf{A}_{jXX}^{-1}\mathbf{A}_{jX0}\mathbf{A}_{j00}^{-1/2}$. We denote the largest eigenvalue by λ_{j1}^*, which is equal to $||\mathbf{A}_{j00}^{-1/2}\mathbf{A}_{j0X}\mathbf{A}_{jXX}^{-1}\mathbf{A}_{jX0}\mathbf{A}_{j00}^{-1/2}||$, where $||\cdot||$ is the operator norm, and further denote the corresponding eigenvector by $\boldsymbol{\alpha}_j^*$. Let $\phi_{nj}^*(X_j) = \boldsymbol{\psi}_j^\top(X_j)[\mathrm{E}\{\boldsymbol{\psi}_j(X_j)\boldsymbol{\psi}_j^\top(X_j)\}]^{-1}\mathrm{E}\{\boldsymbol{\psi}_j(X_j)\mathbf{B}_j^\top(Y)\}\boldsymbol{\alpha}_j^*$. ϕ_{nj}^* can be considered the spline approximation to the optimal transformation ϕ_j^* defined previously. Note that the target function in (2.7) is $\mathrm{E}(\phi_{nj}^{*2})$, and we also have $\mathrm{E}(\phi_{nj}^{*2}) = \lambda_{j1}^*$.

Given the data $\{Y_u\}_{u=1}^n$ and $\{X_{uj}\}_{u=1}^n$, we estimate $\mathbf{A}_{j00}$, $\mathbf{A}_{jXX}$, $\mathbf{A}_{jX0}$, and $\mathbf{A}_{j0X}$ as follows.

$$\widehat{\mathbf{A}_{j00}} = n^{-1}\sum_{u=1}^n \mathbf{B}_j(Y_u)\mathbf{B}_j^\top(Y_u), \qquad \widehat{\mathbf{A}_{jXX}} = n^{-1}\sum_{u=1}^n \widehat{\boldsymbol{\psi}}_j(X_{uj})\widehat{\boldsymbol{\psi}}_j^\top(X_{uj}),$$
$$\widehat{\mathbf{A}_{jX0}} = n^{-1}\sum_{u=1}^n \widehat{\boldsymbol{\psi}}_j(X_{uj})\mathbf{B}_j^\top(Y_u), \quad \text{and} \quad \widehat{\mathbf{A}_{j0X}} = \widehat{\mathbf{A}_{jX0}}^\top,$$

where $\widehat{\boldsymbol{\psi}}_j(X_{uj}) = \widehat{\mathbf{D}_j}\mathbf{B}_j(X_{uj})$, the (s,m)-entry of $\widehat{\mathbf{D}_j}$ is $\widehat{\mathbf{D}}_{j,sm} = z_{sm}/(k\widehat{b_{jm}})$, and $\widehat{b_{jm}} = n^{-1}\sum_{u=1}^n B_{jm}(X_{uj})$, for $s = 1,\ldots,d_n - 1$ and $m = 1,\ldots,d_n$. Then, λ_{j1}^* is estimated by

$$\widehat{\lambda_{j1}^*} = ||\widehat{\mathbf{A}_{j00}}^{-1/2}\widehat{\mathbf{A}_{j0X}}\widehat{\mathbf{A}_{jXX}}^{-1}\widehat{\mathbf{A}_{j0X}}^\top\widehat{\mathbf{A}_{j00}}^{-1/2}||,$$

and $\boldsymbol{\alpha}_j^*$ is estimated by the eigenvector of $\widehat{\mathbf{A}_{j00}}^{-1/2}\widehat{\mathbf{A}_{j0X}}\widehat{\mathbf{A}_{jXX}}^{-1}\widehat{\mathbf{A}_{j0X}}^\top\widehat{\mathbf{A}_{j00}}^{-1/2}$ corresponding to $\widehat{\lambda_{j1}^*}$, which we denote as $\widehat{\boldsymbol{\alpha}_j^*}$. Therefore, the optimal transformation of Y is estimated by $\widehat{\theta_{nj}^*} = \widehat{\boldsymbol{\alpha}_j^*}^\top B_j(Y)$. Furthermore, based on (2.6), the optimal transformation of X_j can be obtained by $\widehat{\phi_{nj}^*} = \widehat{\boldsymbol{\eta}_j^*}^\top \boldsymbol{\psi}_j(X_j)$ with $\widehat{\boldsymbol{\eta}_j^*} = \widehat{\mathbf{A}_{jXX}}^{-1}\widehat{\mathbf{A}_{jX0}}\widehat{\boldsymbol{\alpha}_j^*}$.

Based on the two relationships (i) $\mathrm{E}(\phi_j^{*2}) = (\rho_j^*)^2$ and (ii) $\mathrm{E}(\phi_{nj}^{*2}) = \lambda_{j1}^*$, and the fact that ϕ_{nj}^* is the optimal spline approximation to ϕ_j^*, we propose to screen important variables using the magnitudes of $\widehat{\lambda_{j1}^*}$ for $1 \le j \le p$.

2.2.3 MC-SIS procedure

Let ν_n be a pre-specified threshold, and $\widehat{\mathcal{D}_{\nu_n}}$ the collection of selected important variables. Then, our proposed screening procedure can be defined as

$$\widehat{\mathcal{D}_{\nu_n}} = \{1 \leq j \leq p\colon \widehat{\lambda_{j1}^*} \geq \nu_n\}. \tag{2.8}$$

Empirically, the threshold value ν_n is often set so that $|\widehat{\mathcal{D}_{\nu_n}}| = n$ or $[n/\ln n]$ as in Fan and Lv (2008) and Fan et al. (2011), where $|\widehat{\mathcal{D}_{\nu_n}}|$ is the cardinality of $\widehat{\mathcal{D}_{\nu_n}}$ and $[a]$ denotes the integer part of a. Since $\widehat{\lambda_{j1}^*}$ is the estimate of λ_{j1}^*, which is an approximation to the squared maximum correlation coefficient ρ_j^{*2}, we refer to the procedure as the MC-SIS procedure.

2.2.4 Sure Screening Property

We establish the sure screening property of the MC-SIS procedure in this section. The sure screening property is a property under the asymptotic regime that the sample size n goes to infinity and the number of predictor variables (denoted as p_n) may grow with n. The regime with a fixed number of predictor variables, which is $p_n = p$ for all $n > 0$, can be considered a special case. We first introduce some notations.

For any given n, following Li et al. (2012b), we use $F_n(Y|\mathbf{X})$ to denote the conditional distribution of Y given $\mathbf{X}$. Note that the subscript n in F_n is used to indicate that the conditional distribution of Y given $\mathbf{X}$ can depend on n because both Y and $\mathbf{X}$ depend on p_n and p_n may grow with n. Define $\mathcal{A}_n = \{j : F_n(y|\mathbf{X})$ functionally depends on $X_j\}$ and $\mathcal{E}_n = \{j : \rho_j^*(Y, X_j) > 0\}$. Let $\mathcal{A}_n^c = \{j : F_n(y|\mathbf{X})$ does not functionally depend on $X_j\}$ and $\mathcal{E}_n^c = \{j : \rho_j^*(Y, X_j) = 0\}$. Note that both $\mathcal{A}_n$ and $\mathcal{E}_n$ can change with n as n goes to infinity.

The predictor variables in $\mathcal{A}_n$ are the true predictors that jointly affect the response variable Y. The predictor variables in $\mathcal{E}_n$ are those that have positive maximum correlations with Y. In some cases, $\mathcal{A}_n$ is a subset of $\mathcal{E}_n$, whereas in some other cases, $\mathcal{A}_n$ is not a subset of $\mathcal{E}_n$. It is known that a predictor variable can be a true predictor variable, but it is marginally independent of Y; When this happens, like other existing marginal screening procedures in the literature, our proposed MC-SIS procedure will fail to retain the true predictor variable. Define $\mathcal{D}_n = \mathcal{A}_n \cap \mathcal{E}_n$, and $\mathcal{D}_n^c = \mathcal{A}_n^c \cup \mathcal{E}_n^c$. We refer to the predictor variables in $\mathcal{D}_n$ as the *active predictor variables*, and those in $\mathcal{D}_n^c$ the *inactive predictor variables*.

The goal of the MC-SIS procedure is to retain the active predictor variables. Recall that $\widehat{\mathcal{D}_{\nu_n}}$ is the collection of predictor variables selected by MC-SIS. The probability that $\widehat{\mathcal{D}_{\nu_n}}$ contains $\mathcal{D}_n$, which is $\Pr(\mathcal{D}_n \subseteq \widehat{\mathcal{D}_{\nu_n}})$, is not expected to be one when based on a finite sample. Instead, we aim to identify reasonable sufficient conditions under which the probability $\Pr(\mathcal{D}_n \subseteq \widehat{\mathcal{D}_{\nu_n}})$ converges to one as n goes to infinity. This property is referred to as the sure screening property in the literature.

We first consider the special case in which the active set $\mathcal{D}_n$ is fixed. Under this special case, there exists a positive constant c, such that $\min_{j \in \mathcal{D}_n} \rho_j^*(Y, X_j) > c > 0$ for any n, indicating that the marginal maximum correlation coefficients between the response and active predictor variables are always bounded away from zero by the constant c. For this special case, we can show that as the sample size n goes to infinity and some additional conditions hold, the probability that MC-SIS can retain $\mathcal{D}_n$ converges to one; In other words, MC-SIS possesses the sure screening property.

We next consider the general case in which the active set $\mathcal{D}_n$ can change and diverge as n increases. For each n, define $c_n = \min_{j \in \mathcal{D}_n} \rho_j^*(Y, X_j)$. Clearly, c_n is the smallest maximum correlation coefficient between the response and the active predictor variables for given n. Under the assumption that there exists a constant $c > 0$ such that asymptotically c_n is bounded away from zero by the constant c, that is, $\liminf_{n\to\infty} c_n > c$, the sure screening property of MC-SIS can be established. Although this assumption is broader than the special case discussed previously, it is still too restrictive.

When the sample size increases, we should allow the possibility that c_n may decrease to zero. The rate at which c_n decreases to zero plays a critical role in determining whether MC-SIS possesses the sure screening property. If the rate is too fast, the correlation between the response and some active predictor variables becomes too weak, and MC-SIS may fail to retain those active predictor variables, and thus MC-SIS fails to possess the sure screening property. On the other hand, d_n, which is the number of B-spline basis functions used in MC-SIS, critically affects the performance of MC-SIS. The success of MC-SIS hinges on the interplay of d_n and c_n as n goes to infinity. In this article, we impose a mild condition on this interplay between c_n and d_n, which controls the relative rates of c_n and d_n as n goes to infinity. This condition is listed as Condition 5 or (C5) below. Under (C5) and other regularity conditions, we show that MC-SIS indeed possesses the sure screening property (see Theorem 2.2.2). Note that the two special cases discussed above automatically satisfy (C5); Therefore, Theorem 2.2.2 implies that MC-SIS is a sure screening procedure for these two special cases.

Before stating the theorems regarding the theoretical properties of MC-SIS, we first list the conditions below.

(C1) If the transformations θ_j and ϕ_j with zero means and finite variances satisfy

$$\theta_j(Y) + \phi_j(X_j) = 0 \text{ a.s., then each of them is zero a.s.}$$

(C2) The conditional expectation operators $\mathrm{E}\{\phi_j(X_j) \mid Y\} : H_2(X_j) \to H_2(Y)$ and $\mathrm{E}\{\theta_j(Y) \mid X_j\} : H_2(Y) \to H_2(X_j)$ are all compact operators. $H_2(Y)$ and $H_2(X_j)$ are Hilbert spaces of all measurable functions with zero mean, finite variance and usual inner product.

(C3) The optimal transformations $\{\theta_j^*, \phi_j^*\}_{j=1}^p$ belong to a class of functions $\mathcal{F}$, whose *r*th derivative $f^{(r)}$ exists and is Lipschitz of order α_1, that is, $\mathcal{F} = \{f : |f^{(r)}(s) - f^{(r)}(t)| \leq K|s-t|^{\alpha_1}$ for all $s, t\}$ for some positive constant *K*, where *r* is a nonnegative integer and $\alpha_1 \in (0, 1]$ such that $d = r + \alpha_1 > 0.5$.

(C4) The joint density of Y and X_j $(j = 1, \ldots, p)$ is bounded and the marginal densities of Y and X_j are bounded away from zero.

(C5) The number of B-spline basis functions d_n satisfies that $d_n \leq \min_{j \in \mathcal{D}_n}(\rho_j^{*2})/(2c_1 n^{-2\kappa})$, for some constant $c_1 > 0$ and constant κ where $0 \leq \kappa < d/(2d+1)$.

(C6) There exist positive constant C_1 and constant $\xi \in (0,1)$ such that $d_n^{-d-1} \leq c_1(1-\xi)n^{-2\kappa}/C_1$.

Conditions (C1) and (C2) are adopted from (Breiman and Friedman, 1985), which ensure that the optimal transformations exist. Conditions (C3) and (C4) are from Burman (1991), but modified for our two-variable scenario. Condition (C5) above is similar to Condition 3 in Fan and Lv (2008), Condition C in Fan et al. (2011), and Condition (C2) in Li et al. (2012b), which all require that the dependence between the response and active predictor variables cannot be too weak. As discussed earlier in this section, this condition is necessary since a marginal screening procedure will fail when the marginal dependence between the response and an active predictor variable is too weak.

The following lemma shows that the maximum correlations achieved by B-spline-based transformations are at the same level as the original maximum correlations.

Lemma 2.2.1 *Under conditions (C3) – (C6), we have* $\min_{j \in \mathcal{D}_n} \lambda_{j1}^* \geq c_1 \xi d_n n^{-2\kappa}$.

Based on condition (C1) – (C6), we establish the following sure screening property for MC-SIS.

Theorem 2.2.2 *(a) Under conditions (C1) – (C4), for any* $c_2 > 0$, *there exist positive constants* c_3 *and* c_4 *such that*

$$\Pr(\max_{1 \leq j \leq p} |\widehat{\lambda_{j1}^*} - \lambda_{j1}^*| \geq c_2 d_n n^{-2\kappa}) \leq \mathcal{O}\left(p\zeta(d_n, n)\right). \tag{2.9}$$

where $\zeta(d_n, n) = d_n^2 \exp(-c_3 n^{1-4\kappa} d_n^{-4}) + d_n \exp(-c_4 n d_n^{-7})$.

(b) Additionally, if conditions (C5) and (C6) hold, by taking $\nu_n = c_5 d_n n^{-\kappa}$ with $c_5 \leq c_1\xi/2$, we have that

$$\Pr(\mathcal{D}_n \subseteq \widehat{\mathcal{D}_{\nu_n}}) \geq 1 - \mathcal{O}\left(s\zeta(d_n, n)\right), \tag{2.10}$$

where s is the cardinality of $\mathcal{D}_n$.

Note that Theorem 2.2.2 is stated in terms of a fixed number of predictor variables p. In fact, the same theorem holds for a divergent number of predictor variables, which is denoted as p_n. As long as $p_n\zeta(d_n, n)$ goes to zero asymptotically, MC-SIS can possess the sure screening property. We remark that the number of basis functions d_n affects the final performance of MC-SIS. To obtain the sure screening property, an upper bound of d_n is $o(n^{1/7})$. Since d_n is determined by the choices of the degree of B-spline basis functions and the number of knots, different combinations of degree and the number of knots can lead to different screening results. Additionally, knots placement can further affect the behavior of B-spline functions, and in practice, knots are usually equally spaced or placed at sample quantiles. In next section, we will propose a data-driven three-step procedure for determining d_n for MC-SIS in practice. The optimal choice of d_n and knots placement are beyond the scope of this thesis and can be an interesting topic for future research.

The sure screening property from Theorem 2.2.2 guarantees that MC-SIS retains the active set. The size of the selected set can be much larger than the size of the active set. Therefore, it is of interest to assess the size of the selected set. Following an approach in Fan et al. (2011), we establish such a result for MC-SIS and state it in the next theorem.

Theorem 2.2.3 *Under Conditions (C1) – (C6), we have that for any $\nu_n = c_5 d_n n^{-\kappa}$, there exist positive constants c_3 and c_4 such that*

$$\Pr\{|\widehat{\mathcal{D}_{\nu_n}}| \leq \mathcal{O}\left(n^{2\kappa}\lambda_{\max}(\boldsymbol{\Sigma})\right)\} \geq 1 - \mathcal{O}\left(p_n\zeta(d_n, n)\right), \tag{2.11}$$

where $|\widehat{\mathcal{D}_{\nu_n}}|$ is the cardinality of $\widehat{\mathcal{D}_{\nu_n}}$, $\lambda_{\max}(\boldsymbol{\Sigma})$ is the largest eigenvalue of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{\psi}\boldsymbol{\psi}^\top)$, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^\top, \ldots, \boldsymbol{\psi}_{p_n}^\top)^\top$, p_n is the divergent number of predictor variables, and $\zeta(d_n, n)$ is *defined in Theorem 2.2.2.*

From Theorem 2.2.3, we have that when $\lambda_{\max}(\boldsymbol{\Sigma}) = \mathcal{O}(n^{\tau})$, the cardinality of the selected set by MC-SIS will be of order $\mathcal{O}(n^{2\kappa+\tau})$. Thus, by applying MC-SIS, we can reduce dimensionality from the original exponential order to a polynomial order, while retaining the entire active set.

2.3 Tuning Parameter Selection

In the previous section, we show that in order to achieve the sure screening property of MC-SIS, we need to impose several conditions on the choice of d_n. Recall $d_n = k + \ell$, where k is the number of knots and ℓ is the degree of the B-spline basis functions. These conditions are of theoretical interest, but cannot be directly implemented in practice. It is well known that the performance of B-spline functions in nonparametric regression depends on the choices of k and ℓ as well as the placement of knots. This is also the case for the performance of MC-SIS under a given finite sample.

Several rules of thumb have been proposed to choose d_n for B-spline basis functions when used for the purpose of screening in the literature. For example, cubic splines with $d_n = \left[n^{1/5}\right] + 2$ were used in Fan et al. (2011), and cubic splines with $d_n = \left[2n^{1/5}\right]$ were proposed in Fan et al. (2014), and in both works, the knots were placed at the sample quantiles. These rules of thumb can also be applied to MC-SIS, however, we found their performances are not so satisfactory in some models we have experimented with. In this section, we propose a more effective approach for selecting ℓ and k (or d_n) of the B-spline basis functions for MC-SIS.

There are two major factors in the use of B-spline basis functions, which affect the performance of MC-SIS. The first factor is the complexity of the B-spline basis functions characterized by ℓ and k. The larger ℓ and k are, the more complex the B-spline basis functions. Using more complex basis functions can clearly lead to the overfitting problem for inactive predictor variables, many of which may be retained due to their falsely inflated empirical correlations with the response variable. On the other hand, using less complex basis functions with small ℓ and k can lead to the underfitting problem for active predictor variables, that is, the maximum correlations between the response variable and active predictor variables may be underestimated, and some active predictor variables may be ranked lower due to underestimated maximum correlations. Therefore, the proper selection of ℓ and k hinges on the balance between the overfitting and underfitting problems.

The other factor that affects the performance of MC-SIS is whether the same choices of ℓ and k are used for all predictor variables, which is referred to as the *unified scheme*, or different choices of ℓ and k are used for different predictor variables, which is referred to as the *separate scheme*. The unified scheme treats all predictor variables the same way and is relatively simple, but it may be appropriate for some variables while being inappropriate for other variables. It is difficult to find a unified scheme that simultaneously fits all predictor variables. On the other hand, the separate scheme allows individual variables to choose their most suitable basis functions, but it has two drawbacks. The first drawback is that it may exacerbate the overfitting problem for inactive predictor variables, and the second is that its computational demand is high.

Based on the discussion above, it is clear that for the purpose of screening, an ideal scheme for choosing basis functions for MC-SIS is to use the unified scheme with simple basis functions for inactive predictor variables and the separate scheme with complex basis functions for active predictor variables. This ideal scheme is not feasible in practice because we do not know which predictor variables are active and which are inactive ahead of time. In what follows next, we instead propose a data-driven three-step approach to approximate the ideal scheme. Because B-spline basis functions of degree higher than three are seldom used in practice, we only consider $\ell \in \{1, 2, 3\}$. Furthermore, we always place knots at sample quantiles.

In the first step, we use the unified scheme with B-spline basis functions of degree one. In other words, we fix $\ell = 1$. The number of knots k is then determined as follows. Consider a set of candidate values for k, for example, $K_1 \leq k \leq K_2$, where K_1 and K_2 are pre-specified integers. For each k, we first calculate the maximum correlations between the response variable and the predictor variables using k knots and $\ell = 1$, and then we fit a two-component Gaussian mixture distribution to the calculated maximum correlations and denote the resulting component means as $\mu_1(k)$ and $\mu_2(k)$, respectively. The Gaussian mixture distribution is used to cluster predictor variables into two groups with one group including large maximum correlations and the other including small ones. Let $d(k) = |\mu_1(k) - \mu_2(k)|$, which is a measure of separability of those two groups. The larger $d(k)$ is, the more separable the two groups are. We want to choose the value of k that can separate the two groups the most. A natural choice is $\tilde{k} = \min_{K_1 \leq k \leq K_2} d(k)$. Then, we apply MC-SIS with $\ell = 1$ and $k = \tilde{k}$ to all of the predictor variables, and retain B_1 predictor variables with the largest B_1 maximum correlations, where B_1 is a pre-specified number. The purpose of using the unified scheme with linear B-spline basis functions in this step is to avoid the overfitting problem and screen out a large number of inactive predictor variables.

In the second step, we employ the separate scheme. For each remaining predictor variable, an M-fold Cross-Validation (CV) procedure is used to select $\ell \in \{1, 2\}$ and k (where $K_1 \leq k \leq K_2$), where M is a pre-defined integer. The maximum correlation between the predictor variable and the response variable is then calculated using B-spline basis functions with the selected ℓ and k. Subsequently, we rank-order the predictors using their corresponding maximum correlations and retain the top B_2 predictor variables, where B_2 is a pre-specified number. The M-fold CV procedure uses the correlation between the response variable and the predictor variable as the score function. The purpose of using the separate scheme and B-spline basis functions of higher degree is to correct the under-fitting problem possibly suffered by the active predictor variables in the first step.

The third step is similar to the second step. The only difference is that the degree ℓ for B-spline basis functions is selected from $\{1, 2, 3\}$ instead of $\{1, 2\}$. In other words, for individual remaining predictor variables, B-spline basis functions of degree up to three may be used to calculate their maximum correlations. The purpose of using cubic spline basis functions is to provide sufficient capacity to calculate the maximum correlations of active predictor variables.

The maximum correlations for all pairs of (Y, X_j) are calculated based on their selected tuning parameters. The predictor variables are then sorted, and the top B_3 are retained as the final output of MC-SIS, where B_3 is a pre-specified number.

Note that the three-step procedure proposed above requires three pre-specified numbers, B_1, B_2 and B_3. The choices of B_1, B_2 and B_3 can vary from one problem to another and depend on a number of factors, including the sample size n, the number of predictor variables p, the signal strengths of the active variables, the noise level, etc. How to optimally determine B_1, B_2 and B_3 is beyond the scope of this thesis. Here, we instead provide some general guidelines for the user in practice. Suppose the user has a conservative lower bound, denoted as q_1, for the number of predictor variables that are independent of the response, and a conservative upper bound, denoted as q_2, for the number of active variables. For example, suppose there are 500 predictor variables in an application problem. Applying the sparsity principle, the user believes that a half of the predictors are independent of the response variable. Then, q_1 can be assumed to be 200. Furthermore, the user believes that the number of true variables is less than 20. Then, q_2 can be set as 20.

The goal of the first step in the three-step procedure is to screen out inactive predictor variables which are independent of the response, and B_1 is the number of predictor variables that can enter the second step. A proper choice of B_1 is $B_1 = p - q_1$. In the previous example, B_1 then becomes 300, and is conservative in that the first step eliminates 200 out of all 250 predictors that are independent of the response. Similarly, B_2 is the number of predictor variables that can enter the third step so that the maximum correlations of active predictor variables can be accurately evaluated. In order to not leave out any active variables from the third step, a proper choice of B_2 is $B_2 = q_2$. Again in the previous example, B_2 is set to be 20. Because B_3 is the number of predictor variables that are retained in the output set $\widehat{\mathcal{D}_{\nu_n}}$ defined in Section 2.2.3, the choice of B_3 is equivalent to the choice of ν_n. Therefore, B_3 can be chosen in the same way as ν_n as discussed in Section 2.2.3.

2.4 Numerical Results

We illustrate the MC-SIS procedure by studying its performance under different model settings and distributional assumptions of the predictor variables. For all examples, we compare MC-SIS with SIS, NIS, and DC-SIS. As mentioned at the end of Section 2.2.1, the ACE algorithm in Breiman and Friedman (1985) can also be used to calculate the maximum correlation coefficient. Therefore, the ACE algorithm can also be used to perform maximum correlation-based screening, and we refer to the resulting procedure as the ACE-based MC-SIS procedure. We also include the ACE-based MC-SIS procedure in our simulation study. To avoid confusion, we refer to our proposed procedure as the B-spline-based MC-SIS procedure in this section. For each simulation example, we set $p = 1000$ and choose $n \in \{200, 300, 400\}$.

Following Fan and Lv (2008) and Fan et al. (2011), we measure the effectiveness of MC-SIS using average minimum model size (MMS) and robust estimate of its standard deviation (RSD). MMS is defined as the minimum number of selected variables, i.e., the size of the selected set, that is required to include the entire active set. The average MMS is the average of MMS over 100 replicated simulation runs. RSD is defined as IQR/1.34, where IQR is the interquartile range of MMS. When constructing B-spline basis functions, we choose the degree and the number of knots according to the procedure proposed in Section 2.3, and set $K_1 = 3$, $K_2 = 6$, $B_1 = 200$, $B_2 = 50$ and $M = 10$.

Example 2.4.1 *(1.a): $Y = \boldsymbol{\beta}^{*\top}\mathbf{X} + \varepsilon$, with the first s components of $\boldsymbol{\beta}^*$ taking values ± 1 alternatively and the remaining being 0, where $s = 3, 6$ or 12; X_k are independent and identically distributed as $\mathcal{N}(0,1)$ for $1 \leq k \leq 950$; $X_k = \sum_{j=1}^{s} X_j (-1)^{j+1}/5 + (1 - s\varepsilon_k/25)^{1/2}$ where ε_k are independent and identically distributed as $\mathcal{N}(0,1)$ for $k = 951, \ldots, 1000$; and $\varepsilon \sim \mathcal{N}(0,3)$. Here, $\mathcal{D}_n = \{1, \ldots, s\}$.*

(1.b): $Y = X_1 + X_2 + X_3 + \varepsilon$, where X_k are independent and identically distributed as $\mathcal{N}(0,1)$ for $k = 1,$ and $3 \leq k \leq 1000$; $X_2 = X_1^3/3 + \tilde{\varepsilon}$, and $\tilde{\varepsilon} \sim \mathcal{N}(0,1)$; and

$\varepsilon \sim \mathcal{N}(0, 3)$. *Here,* $\mathcal{D}_n = \{1, 2, 3\}$.

The first example is from Fan et al. (2011) and the simulation results are presented in Table 2.1. Under model (1.a), SIS demonstrates the best performance across all cases, which is expected since SIS is specifically developed for linear models. Under the models (1.a) with $s = 3$ or 6, when $n = 200$, MC-SIS underperforms all other methods. However, when sample size increases to 300 or 400, MC-SIS becomes comparable to others. For the case with $s = 12$, MC-SIS underperforms other methods for all choices of n. The cause for the relatively poor performance of MC-SIS is due to the weak signal. With $s = 12$, it requires more samples for MC-SIS to estimate maximum correlation coefficient, without taking advantages of linearity assumptions.

In model (1.b), SIS fails because there exists a nonlinear relationship between X_1 and X_2. NIS demonstrates the best performance as NIS is designed for dealing with nonparametric additive models. The ACE-based MC-SIS procedure demonstrates the second best performance. The B-spline-based MC-SIS procedure performs better than DC-SIS.

Example 2.4.2 *(2.a):* $Y = X_1X_2 + X_3X_4 + \varepsilon$*;* $\mathcal{D}_n = \{1, 2, 3, 4\}$*; (2.b):* $Y = X_1^2 + X_2^3 + X_3^2X_4 + \varepsilon$*;* $\mathcal{D}_n = \{1, 2, 3, 4\}$*; (2.c):* $Y = X_1 \sin(X_2) + X_2 \sin(X_1) + \varepsilon$*;* $\mathcal{D}_n = \{1, 2\}$*; (2.d):* $Y = X_1 \exp(X_2) + \varepsilon$*;* $\mathcal{D}_n = \{1, 2\}$*; (2.e):* $Y = X_1 \ln(|c_0 + X_2|) + \varepsilon$*;* $\mathcal{D}_n = \{1, 2\}$*; (2.f):* $Y = X_1/(c_0 + X_2) + \varepsilon$*;* $\mathcal{D}_n = \{1, 2\}$*. Here* $X_1, \ldots, X_{1000}$ *and* ϵ *are generated independently from* $\mathcal{N}(0, 1)$*, and* $c_0 = 10^{-4}$.

The eight models considered in this example are non-additive, and the simulation results are presented in Table 2.2. Due to the presence of non-additive structures, we notice that SIS and NIS fail in all models, and increasing sample size does not help improve the performances of SIS and NIS for most models. Both MC-SIS and DC-SIS work well in this example, but MC-SIS outperforms DC-SIS for almost all the models in terms of MMS. Even when the sample size is as small as 200, MC-SIS can effectively retain the active set

Table 2.1.
Average MMS and RSD (in parentheses) for Example 2.4.1

Model	n	SIS	NIS	DC-SIS	MC-SIS	MC-SIS
					(ACE)	(B-spline)
1.a	200	*5.8(3.0)*	6.4(3.0)	6.8(3.2)	11.9(7.7)	36.6(20.7)
(s = 3)	300	*4.6(0.9)*	4.9(1.5)	5.1(1.5)	5.9(3.0)	15.0(6.7)
	400	*3.3(0.0)*	3.4(0.0)	3.6(0.8)	3.6(0.8)	6.8(3.7)
1.a	200	*57.4(2.4)*	68.7(9.7)	60.2(3.7)	140.5(60.8)	175.0(50.2)
(s = 6)	300	*56.0(0.0)*	58.2(0.2)	57.1(0.0)	67.4(5.2)	94.7(27.8)
	400	*55.8(0.0)*	55.9(0.0)	55.9(0.0)	56.8(0.8)	68.0(9.0)
1.a	200	*119.4(42.9)*	250.6(133.2)	195.2(55.8)	484.6(181.9)	500.4(197.4)
(s = 12)	300	*73.4(7.5)*	120.6(35.3)	80.3(10.6)	211.2(108.4)	248.9(103.9)
	400	*64.5(0.8)*	82.21(6.7)	69.7(1.5)	118.2(90.8)	178.2(41.2)
1.b	200	443.6(455.2)	*26.5(6.7)*	136.1(113.4)	56.8(32.8)	115.7(84.7)
	300	394.5(379.7)	*7.3(0.0)*	59.9(48.5)	21.9(5.4)	51.9(27.4)
	400	410.0(361.2)	*3.2(0.0)*	41.1(36.8)	5.6(0.8)	20.0(4.7)

Table 2.2.
Average MMS and RSD (in parentheses) for Example 2.4.2

Model	n	SIS	NIS	DC-SIS	MC-SIS	MC-SIS
					(ACE)	(B-spline)
2.a	200	709.3(239.0)	651.5(285.5)	440.6(231.2)	*248.7(242.5)*	324.3(228.2)
	300	724.1(194.6)	631.2(251.7)	350.5(186.0)	*117.8(88.3)*	197.8(152.6)
	400	795.3(194.8)	636.5(256.3)	280.0(148.9)	*59.3(26.1)*	118.2(92.2)
2.b	200	617.5(308.2)	300.5(298.7)	186.5(132.5)	*104.2(103.0)*	176.5(135.1)
	300	608.5(305.0)	277.8(250.0)	163.6(150.2)	*78.4(44.6)*	125.1(71.6)
	400	597.4(291.6)	262.0(228.9)	114.7(103.7)	*54.9(13.9)*	63.8(32.1)
2.c	200	574.5(352.2)	511.7(389.0)	113.6(80.2)	*18.1(2.24)*	30.9(15.1)
	300	616.4(342.2)	521.8(321.6)	51.0(30.0)	*8.4(0.8)*	9.6(3.2)
	400	622.4(306.3)	547.8(337.9)	21.4(14.0)	13.0(0.0)	*4.8(2.2)*
2.d	200	536.5(285.1)	181.8(168.5)	*2.0(0.0)*	2.3(0.8)	9.7(3.2)
	300	268.6(307.1)	172.8(190.9)	*2.0(0.0)*	*2.0(0.0)*	6.4(3.0)
	400	272.1(331.0)	176.3(178.7)	*2.0(0.0)*	*2.0(0.0)*	4.7(2.2)
2.e	200	580.2(152.8)	512.2(405.6)	191.0(152.8)	55.1(20.3)	*26.6(14.2)*
	300	588.7(299.4)	641.0(295.3)	107.1(70.3)	40.7(1.5)	*11.5(4.5)*
	400	602.1(258.4)	568.0(311.9)	66.2(44.6)	19.8(0.0)	*7.6(3.7)*
2.f	200	928.8(59.3)	654.5(417.9)	140.5(123.5)	*30.0(9.9)*	40.8(11.9)
	300	936.7(37.7)	768.8(292.0)	61.6(46.6)	23.4(2.2)	*17.5(6.0)*
	400	942.0(39.9)	821.7(175.2)	60.9(22.8)	17.8(0.8)	*12.6(3.7)*

under models (2.c), (2.e) and (2.f). This example demonstrates the advantages of MC-SIS and DC-SIS over SIS and NIS for non-additive models as well as the effectiveness of MC-SIS over DC-SIS.

Example 2.4.3 *The models considered in this example are modifications of the models considered in Example 2.4.2. First, the error term ϵ in each original model is removed; and second, the predictor variables $X_1, \ldots, X_p$ are drawn independently from $Cauchy(0, 1)$ instead of $\mathcal{N}(0, 1)$. The resulting models are denoted as (3.a)-(3.f), correspondingly. Simulation results based on these models are presented in Table 2.3.*

Intuitively, the absence of the error terms in the models is expected to help the screening methods, but the use of heavy-tailed distributions for the predictor variables is expected to hinder the methods. The exact performance of a screening method in this example depends on the trade-off between those two changes. Comparing Table 2.3 with Table 2.2, we can see that the performances of SIS and NIS have improved, though they are still far from being satisfactory. The performance of DC-SIS has improved in models (3.a) and (3.c), but has much deteriorated in the other models, which indicates that DC-SIS is susceptible to heavy-tailed distributions. In the presence of heavy tails, Condition (C1) in Li et al. (2012b) is violated, and DC-SIS may not have the sure screening property. The performances of ACE-based and B-spline-based MC-SIS are better over DC-SIS in most models, which indicates the robustness of MC-SIS towards heavy-tailed distributions.

Example 2.4.4 *In this example, we consider a real data set from Segal et al. (2003), which contains the expression levels of 6319 genes and the expression levels of a G protein-coupled receptor (Ro1) in 30 mice. The same data set has been analyzed in Hall and Miller (2009) and in Li et al. (2012b) using DC-SIS. The goal is to identify the most influential genes for Ro1.*

We apply SIS, NIS, DC-SIS, ACE-based MC-SIS and B-spline-based MC-SIS to select the top two most important genes, separately. For B-spline-based MC-SIS, as the number of observations is small, we set $K_1 = 1$, $K_2 = 4$, $B_1 = 100$, $B_2 = 30$ and $M = 3$ for the tuning parameter selection procedure in Section 2.3. B-spline-based MC-SIS ranks *Msa.2437.0* and *Msa.26751.0* as the top two genes. We note that gene *Msa.2437.0* is

Table 2.3.
Average MMS and RSD (in parentheses) for Example 2.4.3

Model	n	SIS	NIS	DC-SIS	MC-SIS	MC-SIS
					(ACE)	(B-spline)
3.a	200	338.8(284.3)	296.6(175.4)	90.3(54.3)	124.1(39.2)	*78.7(26.5)*
	300	310.2(241.6)	310.8(253.7)	64.6(32.5)	72.2(14.7)	*44.6(9.1)*
	400	273.3(242.4)	303.1(260.6)	48.3(29.9)	*41.5(7.1)*	*34.5(6.0)*
3.b	200	617.5(305.2)	617.5(256.7)	478.9(286.6)	117.8(36.6)	*79.6(56.0)*
	300	665.8(348.3)	689.2(256.2)	511.2(258.8)	72.0(8.6)	*42.1(6.2)*
	400	619.8(297.0)	696.8(250.0)	507.8(265.1)	32.7(6.7)	*32.2(6.9)*
3.c	200	136.5(80.2)	106.6(70.7)	23.7(12.7)	*11.9(5.2)*	22.8(6.9)
	300	116.1(82.1)	90.1(56.2)	13.4(6.3)	*8.7(4.5)*	17.3(6.2)
	400	90.4(36.0)	67.9(39.2)	9.9(4.7)	*7.3(3.2)*	13.7(5.2)
3.d	200	409.5(367.0)	434.8(409.0)	412.3(401.1)	*15.4(3.7)*	19.3(6.0)
	300	485.1(320.0)	486.7(411.0)	493.8(397.0)	*7.8(2.4)*	14.1(3.7)
	400	460.8(342.0)	493.4(360.1)	480.7(407.3)	12.5(0.0)	*11.5(3.7)*
3.e	200	252.2(193.8)	250.2(228.5)	124.0(99.1)	55.8(11.4)	*39.6(8.2)*
	300	332.9(332.7)	340.0(289.0)	188.7(120.9)	42.9(4.5)	*36.1(7.5)*
	400	314.3(315.5)	334.6(308.6)	121.1(98.0)	37.8(4.1)	*22.8(6.0)*
3.f	200	779.8(172.0)	737.0(244.2)	507.7(249.6)	37.5(6.9)	*27.4(6.0)*
	300	808.4(149.8)	855.7(120.9)	498.6(336.0)	28.7(4.5)	*20.7(5.2)*
	400	806.7(149.1)	837.6(143.5)	432.6(281.9)	34.3(3.7)	*17.3(3.9)*

Table 2.4.
Top ranked (Rank 1 and Rank 2) genes for Example 2.4.4

	SIS	NIS	DC-SIS	MC-SIS (ACE)	MC-SIS (B-spline)
Rank 1 gene	Msa.2877.0	Msa.2877.0	Msa.2134.0	Msa.8081.0	Msa.2437.0
Rank 2 gene	Msa.964.0	Msa.1160.0	Msa.2877.0	Msa.2437.0	Msa.26751.0

ranked in the second place by ACE-based MC-SIS and in the 15th place by NIS. Gene *Msa.26751.0* is ranked in the 22nd place by ACE-based MC-SIS and in the 41st place by SIS. Additionally, we note that almost all of the procedures considered here, including B-spline-based MC-SIS, consistently ranked *Msa.741.0*, *Msa.2134.0* and *Msa.2877.0* among the top genes. The top-ranked two genes by individual procedures are reported in Table 2.4.

To further compare the performances of the screening procedures, we fit regression models for the response, which is the expression level of Ro1, using the top two genes selected by the procedures. Three different models are considered, which are the linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, the additive model $Y = \ell_1(X_1) + \ell_2(X_2) + \varepsilon$, and the optimal transformation model $\theta^*(Y) = \phi_1^*(X_1) + \phi_2^*(X_2) + \varepsilon$, where θ^*, ϕ_1^* and ϕ_2^* are the optimal transformations in Breiman and Friedman (1985). For each procedure, all three models are fitted using the top ranked one gene as well as using the top ranked two genes, and the resulting adjusted R^2 values are reported in Table 2.5.

Under the linear model, as expected, SIS achieves the largest adjusted R^2 values, whereas the adjusted R^2 values of MC-SIS are rather poor. The major cause of the difference between SIS and MC-SIS is that the former is specifically developed for screening under the linear model, whereas the latter is for screening under the optimal transformation model. Under the additive model, when the top one gene is used, NIS achieves the largest adjusted R^2 value; and when the top two genes are used, DC-SIS achieves the largest adjusted R^2 value. Under the optimal transformation model, both ACE-based and

Table 2.5.
Adjusted R^2 (in percentage) of fitting 3 different models for Example 2.4.4

	SIS		NIS		DC-SIS		MC-SIS (ACE)		MC-SIS (B-spline)	
Model	top 1	top 2	top 1	top 2	top 1	top 2	top 1	top 2	top 1	top 2
Linear	74.5	82.3	74.5	75.8	58.4	77.6	13.8	16.9	12.7	40.5
Additive	80.0	84.2	80.0	84.5	65.7	96.8	58.9	68.7	68.5	68.8
Transformation	84.5	88.1	84.5	88.0	90.0	94.7	94.1	96.9	94.1	96.2

B-spline-based MC-SIS achieve the largest adjusted R^2 values with the top one gene as well as top two genes. When plotting the expression levels of Ro1 against the expression levels of various selected genes, different patterns including linear and nonlinear patterns emerge for different screening methods. In practice, we believe that the top ranked genes by different methods are all worth further investigation.

2.5 Discussions

2.5.1 On Tuning Parameter Selection

The performances and results of B-spline-based MC-SIS depend on the choice of degree and the number of knots for B-spline basis functions. In this chapter, we have developed a data-driven three-step procedure to construct B-spline basis functions for MC-SIS in practice. The proposed procedure demonstrates satisfactory performance in simulation study as well as real data application. We plan to investigate and characterize the theoretical property of the procedure in future research.

Most existing marginal screening procedures under nonparametric model assumptions, including MC-SIS, make use of independent measures, whose estimation typically involves nonparametric model fitting and tuning parameter selection. Nonparametric methods are known to be sensitive to tuning parameter selection. Therefore, this can also become a drawback for those screening procedures. On the other hand, there are various independence measures that are based on cumulative distribution functions, and the estimation of those measures does not involve nonparametric fitting and tuning parameter selection. Two examples include Hoeffding's test in Hoeffding (1948) and Heller-Heller-Gorfine(HHG) tests in Heller et al. (2012). It will be of interest to explore further on the application of the parameter-free measures for screening and the potential of using these methods for variable selection after screening. As an example of dependence measure without tuning parameters, we briefly review the recently proposed HHG tests.

HHG Tests and associated screening procedure

The main idea of HHG tests is described as follows. Let $d(\cdot,\cdot)$ be a pre-specified distance measure, and $(x_1, y_1), \dots, (x_n, y_n)$ be a given sample of (X, Y). For any pair of observations (x_i, y_i) and (x_j, y_j) $(j \neq i)$, the remaining $n-2$ observations are classified to four categories based on their coordinate-wise distances to (x_i, y_i) as follows. Let

$$\begin{aligned}
\mathscr{A}_{11}(i,j) &= \{(x_k, y_k) : d(x_i, x_k) \le d(x_i, x_j) \text{ and } d(y_i, y_k) \le d(y_i, y_j)\},\\
\mathscr{A}_{12}(i,j) &= \{(x_k, y_k) : d(x_i, x_k) \le d(x_i, x_j) \text{ and } d(y_i, y_k) > d(y_i, y_j)\},\\
\mathscr{A}_{21}(i,j) &= \{(x_k, y_k) : d(x_i, x_k) > d(x_i, x_j) \text{ and } d(y_i, y_k) \le d(y_i, y_j)\},\\
\mathscr{A}_{22}(i,j) &= \{(x_k, y_k) : d(x_i, x_k) > d(x_i, x_j) \text{ and } d(y_i, y_k) > d(y_i, y_j)\}.
\end{aligned}$$

The frequences of $\mathscr{A}_{11}(i,j), \mathscr{A}_{12}(i,j), \mathscr{A}_{21}(i,j), \mathscr{A}_{22}(i,j)$ are denoted as $A_{11}(i,j)$, $A_{12}(i,j)$, $A_{21}(i,j)$, $A_{22}(i,j)$, which form a 2×2 contingency table as follows:

$A_{11}(i,j)$	$A_{12}(i,j)$
$A_{21}(i,j)$	$A_{22}(i,j)$

Denote $A_{1\cdot}(i,j) = A_{11}(i,j) + A_{12}(i,j)$, $A_{\cdot,1}(i,j) = A_{11}(i,j) + A_{21}(i,j)$, $A_{2\cdot}(i,j) = A_{21}(i,j) + A_{22}(i,j)$ and $A_{\cdot 2}(i,j) = A_{12}(i,j) + A_{22}(i,j)$. If the two random variables X and Y are independent, we have that $E[A_{kl}(i,j)] = E[A_{k\cdot}(i,j)]E[A_{\cdot l}(i,j)]/(n-2)$ for $k, l = 1, 2$. Therefore, Pearson's χ^2 test can be used to test the independence between X and Y, and the test statistic based on the 2×2 contingency table above is denoted by $S(i,j)$. Heller et al. (2012) proposed to combine $S(i,j)$ of all possible pairs (x_i, y_i) and (x_j, y_j), and use the sum as a test statistic, which is $T = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} S(i,j)$. The sampling distribution of T is difficult to obtain, so Heller et al. (2012) further proposed to use permutation distributions to calculate p-values.

One can develop a screening procedure based on HHG tests similar to other existing screening procedures. In particular, one can use the p-values of the observed test statistics to rank and screen variables. The HHG tests demonstrate powerful performances in hypothesis testing problems (Heller et al., 2012). However, when applied in variable screening, the HHG-based screening procedure may have two major drawbacks. First, the procedure is extremely computationally intensive because the p-values are calculated from permutation distributions, which is time-consuming to obtain. Second, the purposes of screening and testing are different. The former aims to reduce the dimensionality from ultrahigh to high while retaining the active variables, therefore, screening can tolerate false positives to a certain degree in order to gain in speed. Most of existing screening procedures do not employ formal testing. After screening, formal testing methods can be further used to single out the active variables.

2.5.2 On Marginal Screening Procedure

Similar to other existing screening procedures, MC-SIS may fail to retain predictor variables that are functionally related to, but marginally independent of, the response variable. Under the linear regression model, Fan and Lv (2008) proposed an iterative procedure to recover such predictor variables. Similarly, we have developed an iterative version of MC-SIS with the hope to recover active predictor variables missed by MC-SIS.

Iterative MC-SIS

To overcome the drawback that MC-SIS fails to identify important predictors that are marginally independent with the response, we adopt an iterative approach originally proposed in Zhu et al. (2011). This approach relies on iteratively applying MC-SIS, which is given as follows.

1. Apply MC-SIS to data $\{\mathbf{Y}, \mathbf{X}\}$ where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and $\mathbf{X}$ is an $n \times p$ data matrix $(X_{uj})_{1\le u\le n, 1\le j\le p}$. Suppose p_1 predictors are selected. Denote the selected set by $\widehat{\mathcal{D}_1}$, its corresponding $n \times p_1$ data matrix by $\mathbf{X}_{\widehat{\mathcal{D}_1}}$, and the remaining $n \times (p - p_1)$ data matrix as $\mathbf{X}^c_{\widehat{\mathcal{D}_1}}$.

2. Regress $\mathbf{X}^c_{\widehat{\mathcal{D}_1}}$ on $\mathbf{X}_{\widehat{\mathcal{D}_1}}$ to obtain the residuals as

$$\mathbf{X_r} = \{\mathbf{I_n} - \mathbf{X}_{\widehat{\mathcal{D}_1}}(\mathbf{X}^T_{\widehat{\mathcal{D}_1}}\mathbf{X}_{\widehat{\mathcal{D}_1}})^{-1}\mathbf{X}^T_{\widehat{\mathcal{D}_1}}\}\mathbf{X}^c_{\widehat{\mathcal{D}_1}}.$$

Apply MC-SIS to data $\{\mathbf{Y}, \mathbf{X_r}\}$. Suppose p_2 predictors are selected, and denote this selected set by $\widehat{\mathcal{D}_2}$. Update $\widehat{\mathcal{D}_1}$ with $\widehat{\mathcal{D}_1} \bigcup \widehat{\mathcal{D}_2}$.

3. Repeat step 2 until the total number of selected predictors reaches N.

Here, N is a pre-defined value for the size of the selected set. And the number of the selected predictors in each iteration can be either pre-specified or determined by the number of the predictors with marginal maximum correlations exceeding a user-specified threshold value.

In step 2, we compute the residuals of the remaining variables against the selected variables, which are the projection of the remaining variables onto the orthogonal complement space of the variables selected in the previous steps. This step serves two purposes. First, it can make a previously undetectable active variable, due to its marginal independence with the response, detectable; and second, it can decrease the correlation between irrelevant variables and the response, and thus make the selection of the remaining active variables easier.

2.6 Technical Proofs

2.6.1 Notation

n : sample size

p : dimension size

ℓ : degree of polynomial spline

k : number of knots

d_n : dimension of B-spline basis

$\mathcal{D}_n$: active set

$\mathcal{D}_n^c$: inactive set

θ_j : transformation of response Y for pair (X_j, Y), $j = 1, \ldots, p$

ϕ_j : transformation of X_j for pair (X_j, Y)

ρ_j : the Pearson correlation of pair (X_j, Y)

e_j^2 : squared error by regressing ϕ_j on θ_j

θ_j^* : optimal transformation of response Y for pair (X_j, Y)

ϕ_j^* : transformation of X_j for pair (X_j, Y)

ρ_j^* : maximum correlation of pair (X_j, Y)

e_j^{*2} : squared error by regressing ϕ_j^* on θ_j^*

θ_{nj}^* : spline approximation to optimal transformation θ_j^*

ϕ_{nj}^* : spline approximation to optimal transformation ϕ_j^*

s : cardinality of active set $\mathcal{D}_n$

$||\cdot||$: operator norm

$||\cdot||_{\text{sup}}$: sup norm

2.6.2 Bernstein's Inequality and Four Facts

Lemma 2.6.1 *(Bernstein's inequality, Lemma 2.2.9, (Van der Vaart and Wellner, 1996)) For independent random variables $Y_1, \ldots, Y_n$ with bounded ranges $[-M, M]$ and 0 means,*

$$\Pr(|Y_1 + \ldots + Y_n| > x) \leq 2\exp[-x^2/\{2(v + Mx/3)\}]$$

for $v \geq \text{var}(Y_1 + \ldots + Y_n)$.

Under conditions (C3) and (C4), the following four facts hold when $\ell \geq d$.

Fact 1. (Burman (1991)) There exists a positive constant C_1 such that

$$\text{E}\{(\phi_j^* - \phi_{nj}^*)^2\} \leq C_1 k^{-d} \tag{2.12}$$

Fact 2. (Stone et al. (1985); Huang et al. (2010)) There exists a positive constant C_2 such that

$$\text{E}\{B_{jm}^2(\cdot)\} \leq C_2 d_n^{-1} \tag{2.13}$$

Fact 3. (Burman (1991); Zhou et al. (1998)) There exist positive constants c_{11}, c_{12} such that

$$\begin{gathered} c_{11} d_n^{-1} \leq \lambda_{\min}\left(\text{E}\{\mathbf{B}_j(\cdot)\mathbf{B}_j^\top(\cdot)\}\right) \leq \lambda_{\max}\left(\text{E}\{\mathbf{B}_j(\cdot)\mathbf{B}_j^\top(\cdot)\}\right) \leq c_{12} d_n^{-1} \\ c_{11} k^{-1} \leq \lambda_{\min}\left(\text{E}\{\boldsymbol{\psi}_j(X_j)\boldsymbol{\psi}_j^\top(X_j)\}\right) \leq \lambda_{\max}\left(\text{E}\{\boldsymbol{\psi}_j(X_j)\boldsymbol{\psi}_j^\top(X_j)\}\right) \leq c_{12} k^{-1} \end{gathered} \tag{2.14}$$

Fact 4. (Burman (1991); Faouzi et al. (1999)) There exists a positive constant C_3 such that

$$C_3 k^{-1} \leq b_{jm} \leq 1, \qquad 0 \leq \widehat{b_{jm}} \leq 1 \tag{2.15}$$

Remark 2.6.1 *The choice of knots plays a role in establishing the sure screening property. When the knots of the B-splines are placed at the sample quantiles, $\widehat{b_{jm}}$ is positive. When knots are uniform placed, $\widehat{b_{jm}}$ can be zero with a small probability. According to Burman (1991, section 6a), when the marginal density $f_{X_j}(x) > \gamma_0 > 0$ by Condition (C4) for each X_j, we have $\Pr(\widehat{b_{jm}} = 0\, \textit{for some}\, m = 1, \ldots, d_n) \leq k \exp(-\gamma_0 n/k)$. The results in Burman (1991) are based on equally spaced knots, and our proof for MC-SIS use the same choice of knots, as the probability of $\widehat{b_{jm}}$ being zero is a small probability, we just acknowledge $\widehat{b_{jm}} > 0$ in the proof. In fact, sure screening property still hold when the event $\widehat{b_{jm}} = 0$ is included.*

Remark 2.6.2 *With ℓ fixed, k and d_n are of the same order, we replace k with d_n in the following proof for convenience.*

2.6.3 Proof of Lemma 2.2.1

Proof By Cauchy-Schwarz inequality, we have

$$\mathrm{E}(\phi_j^{*2}) \leq 2\mathrm{E}\{(\phi_j^* - \phi_{nj}^*)^2\} + 2\mathrm{E}(\phi_{nj}^{*2})$$

Therefore,

$$\mathrm{E}(\phi_{nj}^{*2}) \geq \frac{1}{2}\mathrm{E}(\phi_j^{*2}) - \mathrm{E}\{(\phi_j^* - \phi_{nj}^*)^2\}$$

Lemma 2.2.1 follows from condition (C5) together with $\mathrm{E}(\phi_{nj}^{*2}) = \lambda_{j1}^*$. ■

2.6.4 Proof of Eight Basic Results

We list and prove eight results (R1) – (R8) which together form the major parts in proving sure screening property of MC-SIS. For the rest of this chapter, we use P_n to denote the sample average.

R1. With c_{11} in *Fact 3*, we have that,

$$||\mathbf{A}_{j00}^{-1/2}|| \leq c_{11}^{-1/2} d_n^{1/2} \tag{2.16}$$

Proof $||\mathbf{A}_{j00}^{-1/2}|| = \lambda_{\min}^{-1/2}(\mathbf{A}_{j00})$, result follows by *Fact 3*. ■

R2. There exist positive constant c_{13} such that

$$||\mathbf{A}_{j0X}|| \leq c_{13} d_n^{-1/2} \tag{2.17}$$

Proof Let $\mathbf{u} = (u_1, \ldots, u_{d_n})^\top \in R^{d_n}$ with $\sum_{m=1}^{d_n} u_m^2 = 1$.

$$\begin{aligned}
\mathbf{u}^\top \mathrm{E}\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(Y)\}\mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}\mathbf{u} &= \sum_{i=1}^{d_n}\left[\int \{\sum_{m=1}^{d_n} u_m B_{jm}(X_j)\} B_{ji}(Y) dF\right]^2 \\
&\leq \int \{\sum_{m=1}^{d_n} u_m B_{jm}(X_j)\}^2 dF \times \sum_{i=1}^{d_n}\{\int B_{ji}^2(Y) dF\} \\
&\leq \lambda_{\max}[\mathrm{E}\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}] \times d_n \max_i \mathrm{E}\{B_{ji}^2(Y)\}
\end{aligned}$$

Then, $||\mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| \leq (c_{12}C_2/d_n)^{1/2}$ by *Fact 2* and *Fact 3*.

It can be easily shown that, for $\mathbf{u} \in R^{d_n - 1}$ with $\sum_{i=1}^{d_n-1} u_i^2 = 1$,

$$\mathbf{u}^\top \mathbf{D}_j \mathbf{D}_j^\top \mathbf{u} = \sum_{m=1}^{d_n} \frac{1}{k^2 b_{jm}^2}\left(\sum_{i=1}^{d_n-1} u_i z_{im}\right)^2 \leq C_3^{-2} \sum_{m=1}^{d_n}\left(\sum_{i=1}^{d_n-1} u_i z_{im}\right)^2 \leq C_3^{-2}$$

which indicates $||\mathbf{D}_j^\top|| \leq C_3^{-1}$.

Then, $||\mathbf{A}_{j0X}|| \leq ||\mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ ||\mathbf{D}_j^\top|| \leq c_{13} d_n^{-1/2}$ with $c_{13} = (c_{12}C_2)^{1/2} C_3^{-1}$.

■

R3. For any given constant c_4, there exists a positive constant c_8 such that

$$\Pr\{||\widehat{\mathbf{A}_{j00}}^{-1/2}|| \geq \left((c_8+1)c_{11}^{-1} d_n\right)^{1/2}\} \leq 2d_n^2 \exp(-c_4 n d_n^{-3}) \tag{2.18}$$

Proof Since $||\widehat{\mathbf{A}_{j00}}^{-1/2}|| = \sqrt{||[P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(Y)\}]^{-1}||}$. R3 can be obtained via equation (26) in Fan et al. (2011), which is $\Pr\{||[P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(Y)\}]^{-1}|| \geq (c_8+1)c_{11}^{-1}d_n\} \leq 2d_n^2\exp(-c_4nd_n^{-3})$. ■

R4. There exist some positive constants c_6, c_7 such that,

$$\Pr\{||\widehat{\mathbf{A}_{j0X}}|| \geq c_6d_n^{-1/2}\} \leq 4d_n^2\exp(-c_7nd_n^{-2}) \tag{2.19}$$

Proof As $||\widehat{\mathbf{A}_{j0X}}|| = ||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}\widehat{\mathbf{D}_j}^\top|| \leq ||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ ||\widehat{\mathbf{D}_j}^\top||$, we firstly deal with $||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||$.

For any square matrix $\mathbf{A}$ and $\mathbf{B}$, $||\mathbf{A}+\mathbf{B}|| \leq ||\mathbf{A}|| + ||\mathbf{B}||$. We have

$$||\mathbf{A}|| - ||\mathbf{B}|| \leq ||\mathbf{A}-\mathbf{B}|| \qquad \text{and} \qquad ||\mathbf{B}|| - ||\mathbf{A}|| \leq ||\mathbf{B}-\mathbf{A}||$$

Then,

$$|\ ||\mathbf{A}|| - ||\mathbf{B}||\ | \leq ||\mathbf{A}-\mathbf{B}||$$

Let $\mathbf{V}_j = P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\} - \mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}$. It follows that,

$$|\ ||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| - ||\mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ | \leq ||\mathbf{V}_j||$$

It is easy to verify that,

$$|\ ||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| - ||\mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ | \leq d_n||\mathbf{V}_j||_{\mathrm{sup}}$$

Since $||B_{jm}(\cdot)||_{\mathrm{sup}} \leq 1$ and using *Fact 2*, we have

$$\mathrm{var}(B_{jm_1}(Y)B_{jm_2}(X_j)) \leq \mathrm{E}\{B_{jm_1}^2(Y)B_{jm_2}^2(X_j)\} \leq \mathrm{E}\{B_{jm_1}^2(Y)\} \leq C_2d_n^{-1}$$

By Bernstein's inequality, for any $\delta > 0$,

$$\Pr\{|(P_n - \mathrm{E})\{B_{jm_1}(Y)B_{jm_2}(X_j)\}| \geq \delta/n\} \leq 2\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\} \tag{2.20}$$

Therefore,

$$\Pr\{|\ ||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| - ||\mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ | \geq d_n\delta/n\} \leq 2d_n^2\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\}$$

Recalling R2, we have,

$$\Pr\{||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| \geq d_n\delta/n + (c_{12}C_2/d_n)^{1/2}\} \leq 2d_n^2 \exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\}$$

By taking $\delta = c_8(c_{12}C_2)^{1/2}nd_n^{-3/2}$, we obtain that for some positive constant c_4,

$$\Pr\{||(P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\})|| \geq (c_8+1)(c_{12}C_2/d_n)^{1/2}\} \leq 2d_n^2\exp(-c_4nd_n^{-2}) \quad (2.21)$$

Next we deal with $||\widehat{\mathbf{D}_j}^\top||$. Using Bernstein's inequality, we obtain that,

$$\Pr\{|\widehat{b_{jm}} - b_{jm}| \geq \delta/n\} \leq 2\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\} \quad (2.22)$$

Since $b_{jm} \geq C_3k^{-1}$, by taking $\delta = C_3w_1nd_n^{-1}$ for $w_1 \in (0,1)$, we have that there exists some positive constant c_5 such that

$$\Pr\{\widehat{b_{jm}} \leq C_3(1-w_1)d_n^{-1}\} \leq 2\exp(-c_5nd_n^{-1}) \quad (2.23)$$

For $\mathbf{u} = (u_1,\ldots,u_{d_n-1})^\top \in R^{d_n-1}$ with $\sum_{i=1}^{d_n-1}u_i^2 = 1$,

$$\mathbf{u}^\top\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top\mathbf{u} = \sum_{m=1}^{d_n}\frac{1}{k^2\widehat{b_{jm}}^2}\left(\sum_{i=1}^{d_n-1}u_iz_{im}\right)^2 \leq \max_m\frac{1}{k^2\widehat{b_{jm}}^2} \quad (2.24)$$

Combing (2.22), (2.23) and (2.24), we have that

$$\begin{aligned}\Pr\{||\widehat{\mathbf{D}_j}^\top|| \geq C_3^{-1}(1-w_1)^{-1}\} &\leq \Pr\{\max_m\frac{1}{k\widehat{b_{jm}}} \geq C_3^{-1}(1-w_1)^{-1}\}\\ &\leq \Pr\{\min_m\widehat{b_{jm}} \leq C_3(1-w_1)k^{-1}\} \\ &\leq 2d_n\exp(-c_5nd_n^{-1})\end{aligned} \quad (2.25)$$

Combining (2.21), (2.25), and $||\widehat{\mathbf{A}_{j0X}}|| \leq ||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ ||\widehat{\mathbf{D}_j}^\top||$, we have

$$\begin{aligned}&\Pr\{||\widehat{\mathbf{A}_{j0X}}|| \geq (c_8+1)(c_{12}C_2)^{1/2}d_n^{-1/2}C_3^{-1}(1-w_1)^{-1}\}\\ \leq &\Pr\{||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| \geq (c_8+1)(c_{12}C_2)^{1/2}d_n^{-1/2}\} + \Pr\{||\widehat{\mathbf{D}_j}^\top|| \geq C_3^{-1}(1-w_1)^{-1}\}\\ \leq &2d_n^2\exp(-c_4nd_n^{-2}) + 2d_n\exp(-c_5nd_n^{-1})\end{aligned} \quad (2.26)$$

Result in R4 follows by choosing c_6, c_7 accordingly. ■

R5. There exist some positive constants c_9, c_{10} such that, for any $\delta > 0$,

$$\begin{aligned}&\Pr\{||\widehat{\mathbf{A}_{j0X}} - \mathbf{A}_{j0X}|| \geq c_9 d_n^2\delta^2/n^2 + c_{10}d_n\delta/n\}\\ &\qquad\qquad \leq 8d_n^2\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\} + 4d_n\exp(-c_5nd_n^{-1})\end{aligned} \tag{2.27}$$

Proof It is easy to derive

$$\begin{aligned}&||\widehat{\mathbf{A}_{j0X}} - \mathbf{A}_{j0X}|| = ||P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}\widehat{\mathbf{D}_j}^\top - \mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}\mathbf{D}_j{}^\top||\\ &\leq ||(P_n-\mathrm{E})\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ ||\widehat{\mathbf{D}_j}^\top - \mathbf{D}_j^\top|| + ||\mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ ||\widehat{\mathbf{D}_j}^\top - \mathbf{D}_j^\top||\\ &\qquad + ||(P_n-\mathrm{E})\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||\ ||\mathbf{D}_j{}^\top||\end{aligned} \tag{2.28}$$

It is proved in R2 that $||\mathrm{E}\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| \leq (c_{12}C_2/d_n)^{1/2}$ and that $||\mathbf{D}_j{}^\top|| \leq C_3^{-1}$. Combining (2.20) and the fact that

$$||(P_n-\mathrm{E})\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| \leq d_n||(P_n-\mathrm{E})\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}||_{\sup},$$

we have that,

$$\Pr\{||(P_n-\mathrm{E})\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(X_j)\}|| \geq d_n\delta/n\} \leq 2d_n^2\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\}. \tag{2.29}$$

For $\mathbf{u} \in R^{d_n-1}$ with $\sum_{i=1}^{d_n-1} u_i^2 = 1$,

$$\begin{aligned}\mathbf{u}^\top(\widehat{\mathbf{D}_j} - \mathbf{D}_j)(\widehat{\mathbf{D}_j} - \mathbf{D}_j)^\top\mathbf{u} &= \sum_{m=1}^{d_n}\left(\frac{1}{k\widehat{b_{jm}}} - \frac{1}{kb_{jm}}\right)^2\left(\sum_{i=1}^{d_n-1} u_iz_{im}\right)^2\\ &\leq C_3^{-2}\max_m\frac{(\widehat{b_{jm}} - b_{jm})^2}{\widehat{b_{jm}}^2}\end{aligned} \tag{2.30}$$

From (2.22), (2.23) and (2.30), we have that,

$$\begin{aligned}&\Pr\{||\widehat{\mathbf{D}_j}^\top - \mathbf{D}_j^\top|| \geq C_3^{-2}(1-w_1)^{-1}d_n\delta/n\}\\ &\leq \Pr\{C_3^{-1}\max_m\frac{|\widehat{b_{jm}} - b_{jm}|}{\widehat{b_{jm}}} \geq C_3^{-1}\frac{\delta/n}{C_3(1-w_1)d_n^{-1}}\}\\ &\leq \Pr\{\max_m|\widehat{b_{jm}} - b_{jm}| \geq \delta/n\} + \Pr\{\min_m\widehat{b_{jm}} \leq C_3(1-w_1)d_n^{-1}\}\\ &\leq 2d_n\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\} + 2d_n\exp(-c_5nd_n^{-1})\end{aligned} \tag{2.31}$$

Therefore, together with (2.28), (2.29), (2.31) and union bound of probability, we have

$$\Pr\{||\widehat{\mathbf{A}_{j0X}} - \mathbf{A}_{j0X}|| \geq \frac{d_n^2\delta^2/n^2}{C_3^2(1-w_1)} + \frac{(c_{12}C_2)^{1/2}d_n^{1/2}\delta/n}{C_3^2(1-w_1)} + C_3^{-1}d_n\delta/n\}$$

$$\leq 4d_n^2\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\} + 4d_n\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\} + 4d_n\exp(-c_5nd_n^{-1})$$

Result in R5 can be obtained by adjusting the values of c_9 and c_{10}. ■

R6. For given c_4 and c_5, there exist positive constants c_{15} and c_{16} such that,

$$\begin{aligned}&\Pr\{||\widehat{\mathbf{A}_{jXX}}^{-1}|| \geq c_{16}d_n\}\\ &\quad\leq 2d_n^2\exp(-c_4nd_n^{-3}) + 2d_n^3\exp(-c_{15}nd_n^{-7}) + 2d_n^3\exp(-c_5nd_n^{-1})\end{aligned} \tag{2.32}$$

Proof Follow the proof in Lemma 5 of Fan et al. (2011), we have that,

$$|\lambda_{\min}(\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top) - \lambda_{\min}(\mathbf{D}_j\mathbf{D}_j^\top)| \leq d_n||\mathbf{V}_j||_{\sup},\ \text{where } \mathbf{V}_j = \widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top - \mathbf{D}_j\mathbf{D}_j^\top$$

The (s,m)-entry of $\mathbf{V}_j$ is

$$\begin{aligned}(\mathbf{V}_j)^{(s,m)} &= |\sum_{i=1}^{d_n}\frac{z_{si}z_{mi}}{k^2}\left(\frac{1}{\widehat{b_{ji}}^2} - \frac{1}{b_{ji}^2}\right)| = |\sum_{i=1}^{d_n}\frac{z_{si}z_{mi}}{k^2b_{ji}^2}\left(\frac{b_{ji}^2 - \widehat{b_{ji}}^2}{\widehat{b_{ji}}^2}\right)|\\ &\leq C_3^{-2}d_n\max_i|\frac{b_{ji}^2 - \widehat{b_{ji}}^2}{\widehat{b_{ji}}^2}| \leq 2C_3^{-2}d_n\max_i|\frac{b_{ji} - \widehat{b_{ji}}}{\widehat{b_{ji}}^2}|\end{aligned}$$

It is clear that $||\mathbf{V}_j||_{\sup} \leq 2C_3^{-2}d_n\max_i|(b_{ji} - \widehat{b_{ji}})/\widehat{b_{ji}}^2|$. Together with (2.22) and (2.23) , we have

$$\begin{aligned}&\Pr\{|\lambda_{\min}(\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top) - \lambda_{\min}(\mathbf{D}_j\mathbf{D}_j^\top)| \geq 2C_3^{-4}(1-w_1)^{-2}d_n^4\delta/n\}\\ &\leq \Pr\{2C_3^{-2}d_n^2\max_i|\frac{b_{ji} - \widehat{b_{ji}}}{\widehat{b_{ji}}^2}| \geq 2C_3^{-2}d_n^2\delta/n \times C_3^{-2}(1-w_1)^{-2}d_n^2\}\\ &\leq \Pr\{\max_m|\widehat{b_{jm}} - b_{jm}| \geq \delta/n\} + \Pr\{\min_m\widehat{b_{jm}} \leq C_3(1-w_1)d_n^{-1}\}\\ &\leq 2d_n\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\} + 2d_n\exp(-c_5nd_n^{-1})\end{aligned}$$

which indicates that there exists a positive constant c_{14},

$$\begin{aligned}&\Pr\{|\lambda_{\min}(\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top) - \lambda_{\min}(\mathbf{D}_j\mathbf{D}_j^\top)| \geq c_{14}d_n^4\delta/n\}\\ &\qquad\leq 2d_n\exp\{-\frac{\delta^2}{2(C_2nd_n^{-1}+2\delta/3)}\} + 2d_n\exp(-c_5nd_n^{-1})\end{aligned} \tag{2.33}$$

Due to the facts that

$$c_{11}k^{-1} \le \lambda_{\min}(\mathbf{D}_j \mathrm{E}\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}\mathbf{D}_j^\top) \le$$
$$\lambda_{\max}(\mathrm{E}\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\})\lambda_{\min}(\mathbf{D}_j\mathbf{D}_j^\top) \le c_{12}k^{-1}\lambda_{\min}(\mathbf{D}_j\mathbf{D}_j^\top)$$

and that

$$c_{11}k^{-1}\lambda_{\max}(\mathbf{D}_j\mathbf{D}_j^\top) \le \lambda_{\min}(\mathrm{E}\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\})\lambda_{\max}(\mathbf{D}_j\mathbf{D}_j^\top) \le$$
$$\lambda_{\max}(\mathbf{D}_j \mathrm{E}\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}\mathbf{D}_j^\top) \le c_{12}k^{-1}$$

we have

$$\frac{c_{11}}{c_{12}} \le \lambda_{\min}(\mathbf{D}_j\mathbf{D}_j^\top) \le \lambda_{\max}(\mathbf{D}_j\mathbf{D}_j^\top) \le \frac{c_{12}}{c_{11}}$$

By taking $\delta = w_2/c_{14}nd_n^{-4} \times c_{11}/c_{12}$ in (2.33) for any $w_2 \in (0,1)$, there exists a positive constant c_{15} such that,

$$\Pr\{|\lambda_{\min}(\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top) - \lambda_{\min}(\mathbf{D}_j\mathbf{D}_j^\top)| \ge w_2\lambda_{\min}(\mathbf{D}_j\mathbf{D}_j^\top)\}$$
$$\le 2d_n\exp(-c_{15}nd_n^{-7}) + 2d_n\exp(-c_5nd_n^{-1})$$

By following a similar argument in proving inequality (26) in NIS Fan et al. (2011), we have,

$$\Pr\{\lambda_{\min}^{-1}(\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top) \ge (c_8+1)c_{12}/c_{11}\} \le 2d_n\exp(-c_{15}nd_n^{-7}) + 2d_n\exp(-c_5nd_n^{-1}) \quad (2.34)$$

Similarly, it is easy to obtain

$$\Pr\{\lambda_{\min}^{-1}(P_n\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}) \ge (c_8+1)c_{11}^{-1}d_n\} \le 2d_n^2\exp(-c_4nd_n^{-3}) \quad (2.35)$$

Due to the fact that $\lambda_{\max}(\mathbf{H}^{-1}) = \lambda_{\min}^{-1}(\mathbf{H})$, we have

$$||\widehat{\mathbf{A}_{jXX}}^{-1}|| = \lambda_{\min}^{-1}(\widehat{\mathbf{A}_{jXX}}) \le \lambda_{\min}^{-1}(P_n\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\})\ \lambda_{\min}^{-1}(\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top)$$

Together with (2.34) and (2.35), we can obtain that

$$\Pr\{||\widehat{\mathbf{A}_{jXX}}^{-1}|| \ge (c_8+1)^2c_{12}c_{11}^{-2}d_n\}$$
$$\le \Pr\{\lambda_{\min}^{-1}(P_n\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\})\ \lambda_{\min}^{-1}(\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top) \ge (c_8+1)^2c_{12}c_{11}^{-2}d_n\}$$
$$\le \Pr\{\lambda_{\min}^{-1}(P_n\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}) \ge (c_8+1)c_{12}/c_{11}\} + \Pr\{\lambda_{\min}^{-1}(\widehat{\mathbf{D}_j}\widehat{\mathbf{D}_j}^\top) \ge (c_8+1)c_{11}^{-1}d_n\}$$
$$\le 2d_n^2\exp(-c_4nd_n^{-3}) + 2d_n\exp(-c_{15}nd_n^{-7}) + 2d_n\exp(-c_5nd_n^{-1})$$

Therefore, R6 follows by choosing $c_{16} = (c_8+1)^2c_{12}c_{11}^{-2}$. ■

R7. For any $\delta > 0$, given positive constant c_4, there exists a positive constant c_{17} such that,

$$\Pr\{||\widehat{\mathbf{A}_{j00}}^{-1/2} - \mathbf{A}_{j00}^{-1/2}|| \geq c_{17} d_n^{5/2}\delta/n\} \leq 2d_n^2 \exp(-c_4 n d_n^{-3}) + 2d_n^2 \exp\{-\frac{\delta^2}{2(C_2 n d_n^{-1} + 2\delta/3)}\} \tag{2.36}$$

Proof Using perturbation theory from Katō (1995), it is proved in Burman (1991, *Lemma 6.3*) that for some $c_{18} > 0$,

$$||\widehat{\mathbf{A}_{j00}}^{-1/2} - \mathbf{A}_{j00}^{-1/2}|| \leq c_{18}\tilde{\gamma}^{-3/2}||\widehat{\mathbf{A}_{j00}} - \mathbf{A}_{j00}|| \tag{2.37}$$

where $\tilde{\gamma}$ is the minimum of the smallest eigenvalues of $\widehat{\mathbf{A}_{j00}}$ and $\mathbf{A}_{j00}$. $\tilde{\gamma}$ is positive by definition. Therefore,

$$\tilde{\gamma}^{-1} = \max\{\lambda_{\min}^{-1}(\widehat{\mathbf{A}_{j00}}), \lambda_{\min}^{-1}(\mathbf{A}_{j00})\} = \max\{||\widehat{\mathbf{A}_{j00}}^{-1}||, ||\mathbf{A}_{j00}^{-1}||\}$$

From *Fact 3* and R3, we have,

$$c_{12}^{-1} d_n \leq ||\mathbf{A}_{j00}^{-1}|| \leq c_{11}^{-1} d_n \tag{2.38a}$$

$$\Pr\{||[P_n\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(Y)\}]^{-1}|| \geq (c_8+1)c_{11}^{-1} d_n\} \leq 2d_n^2 \exp(-c_4 n d_n^{-3}) \tag{2.38b}$$

Combining (2.38a) and (2.38b) yields

$$\Pr\{\tilde{\gamma}^{-1} \geq \max\big((c_8+1)c_{11}^{-1} d_n, c_{11}^{-1} d_n\big)\} \leq 2d_n^2 \exp(-c_4 n d_n^{-3})$$

which is,

$$\Pr\{\tilde{\gamma}^{-1} \geq (c_8+1)c_{11}^{-1} d_n\} \leq 2d_n^2 \exp(-c_4 n d_n^{-3}) \tag{2.39}$$

Additionally, as proved in equation (33) in Fan et al. (2011), we have large deviation bound for $||(P_n - \mathrm{E})\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(Y)\}||$,

$$\Pr\{||(P_n - \mathrm{E})\{\mathbf{B}_j(Y)\mathbf{B}_j^\top(Y)\}|| \geq d_n\delta/n\} \leq 2d_n^2 \exp\{-\frac{\delta^2}{2(C_2 n d_n^{-1} + 2\delta/3)}\} \tag{2.40}$$

By (2.37), (2.39), (2.40) and under the union bound of probability, we have that,

$$\begin{aligned}
&\Pr\{||\widehat{\mathbf{A}_{j00}}^{-1/2} - \mathbf{A}_{j00}^{-1/2}|| \geq c_{18}(c_8+1)^{3/2}c_{11}^{-3/2}d_n^{5/2}\delta/n\} \\
&\leq \Pr\{c_{18}\tilde{\gamma}^{-3/2}||\widehat{\mathbf{A}_{j00}} - \mathbf{A}_{j00}|| \geq c_{18}(c_8+1)^{3/2}c_{11}^{-3/2}d_n^{3/2}\ d_n\delta/n\} \\
&\leq \Pr\{\tilde{\gamma}^{-1} \geq (c_8+1)c_{11}^{-1}d_n\} + \Pr\{||\widehat{\mathbf{A}_{j00}} - \mathbf{A}_{j00}|| \geq d_n\delta/n\} \\
&\leq 2d_n^2\exp(-c_4 n d_n^{-3}) + 2d_n^2\exp\{-\frac{\delta^2}{2(C_2 n d_n^{-1} + 2\delta/3)}\}
\end{aligned} \tag{2.41}$$

Therefore, R7 follows by choosing $c_{17} = c_{18}(c_8+1)^{3/2}c_{11}^{-3/2}$. ■

R8. For any $\delta > 0$, given positive constant c_4, there exists a positive constant c_{19} such that,

$$\begin{aligned}
\Pr\{||\widehat{\mathbf{A}_{jXX}}^{-1} - \mathbf{A}_{jXX}^{-1}|| \geq c_{19}(d_n^5\delta^3/n^3 + d_n^3\delta/n)\} \leq 8d_n^2\exp\{-\frac{\delta^2}{2(C_2 n d_n^{-1} + 2\delta/3)}\}+ \\
4d_n^2\exp(-c_4 n d_n^{-3}) + 2d_n\exp(-c_{15} n d_n^{-7}) + 6d_n\exp(-c_5 n d_n^{-1})
\end{aligned} \tag{2.42}$$

Proof It's obvious that

$$||\widehat{\mathbf{A}_{jXX}}^{-1} - \mathbf{A}_{jXX}^{-1}|| \leq ||\mathbf{A}_{jXX}^{-1}||\ ||\mathbf{A}_{jXX} - \widehat{\mathbf{A}_{jXX}}||\ ||\widehat{\mathbf{A}_{jXX}}^{-1}|| \tag{2.43}$$

and that

$$\begin{aligned}
&||\widehat{\mathbf{A}_{jXX}} - \mathbf{A}_{jXX}|| = ||\widehat{\mathbf{D}_j}P_n\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}\widehat{\mathbf{D}_j}^\top - \mathbf{D}_j\mathrm{E}\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}\mathbf{D}_j^\top|| \\
&\leq ||\widehat{\mathbf{D}_j} - \mathbf{D}_j||\ ||(P_n - \mathrm{E})\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}||\ ||\widehat{\mathbf{D}_j}^\top - \mathbf{D}_j^\top|| + 2||P_n\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}||\times \\
&\quad ||\widehat{\mathbf{D}_j}^\top - \mathbf{D}_j^\top|| + ||\mathbf{D}_j^\top||\ ||(P_n - \mathrm{E})\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}||\ ||\mathbf{D}_j||
\end{aligned} \tag{2.44}$$

From the similar reasoning in proving (2.21) and (2.29), it is easy to obtain that

$$\Pr\{||P_n\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}|| \geq (c_8+1)c_{13}d_n^{-1}\} \leq 2d_n^2\exp(-c_4 n d_n^{-3}) \tag{2.45}$$

$$\Pr\left(||(P_n - \mathrm{E})\{\mathbf{B}_j(X_j)\mathbf{B}_j^\top(X_j)\}|| \geq d_n\delta/n\right) \leq 2d_n^2\exp\{-\frac{\delta^2}{2(C_2 n d_n^{-1} + 2\delta/3)}\} \tag{2.46}$$

With c_{19} chosen properly, results in R8 follows by combining *Fact 3*, (2.31), (2.32), (2.43), (2.44), (2.45), (2.46) and the fact $||\mathbf{D}_j^\top|| < C_3^{-1}$. ■

2.6.5 Proof of Theorem 2.2.2

Proof of Theorem 2.2.2. Recall that

$$\lambda^*_{j1} = ||\mathbf{A}_{j00}^{-1/2}\mathbf{A}_{j0X}\mathbf{A}_{jXX}^{-1}\mathbf{A}_{jX0}\mathbf{A}_{j00}^{-1/2}||$$

and that

$$\widehat{\lambda^*_{j1}} = ||\widehat{\mathbf{A}_{j00}}^{-1/2}\widehat{\mathbf{A}_{j0X}}\widehat{\mathbf{A}_{jXX}}^{-1}\widehat{\mathbf{A}_{j0X}}^\top\widehat{\mathbf{A}_{j00}}^{-1/2}||$$

Let $\mathbf{a} = \mathbf{A}_{j00}^{-1/2}$, $\mathbf{b} = \mathbf{A}_{j0X}$, $\mathbf{H} = \mathbf{A}_{jXX}^{-1}$, $\mathbf{a_n} = \widehat{\mathbf{A}_{j00}}^{-1/2}$, $\mathbf{b_n} = \widehat{\mathbf{A}_{j0X}}$, $\mathbf{H_n} = \widehat{\mathbf{A}_{jXX}}^{-1}$,

$$\begin{aligned}
&\widehat{\lambda^*_{j1}} - \lambda^*_{j1} = ||\mathbf{a_n}^\top\mathbf{b_n}^\top\mathbf{H_n}\mathbf{b_n}\mathbf{a_n}|| - ||\mathbf{a}^\top\mathbf{b}^\top\mathbf{H}\mathbf{b}\mathbf{a}|| \\
&\leq ||(\mathbf{a}_n - \mathbf{a})^\top\mathbf{b}_n^\top\mathbf{H}_n\mathbf{b}_n(\mathbf{a}_n - \mathbf{a})|| + 2||(\mathbf{a}_n - \mathbf{a})^\top\mathbf{b}_n^\top\mathbf{H}_n\mathbf{b}_n\mathbf{a}|| + ||\mathbf{a}^\top(\mathbf{b}_n^\top\mathbf{H}_n\mathbf{b}_n - \mathbf{b}^\top\mathbf{H}\mathbf{b})\mathbf{a}|| \\
&\quad \triangleq S_1 + S_2 + S_3
\end{aligned} \tag{2.47}$$

We denote the terms in r.h.s as S_1, S_2 and S_3 respectively. Furthermore, we let the r.h.s of inequalities (2.19),(2.27),(2.32),(2.36),(2.42) as Q_4, Q_5, Q_6, Q_7, Q_8.

Note that

$$S_1 \leq ||\mathbf{a}_n - \mathbf{a}||^2\ ||\mathbf{b}_n||^2\ ||\mathbf{H}_n|| \tag{2.48}$$

By (2.19),(2.32),(2.36), we have that there exists a positive constant c_{20} such that,

$$\Pr\{S_1 \geq c_{20}d_n^5\delta^2/n^2\} \leq Q_4 + Q_6 + Q_7 \tag{2.49}$$

As to S_2,

$$S_2 \leq ||\mathbf{a}_n - \mathbf{a}||\ ||\mathbf{b}_n||^2\ ||\mathbf{H}_n||\ ||\mathbf{a}|| \tag{2.50}$$

By (2.16),(2.19),(2.32),(2.36), we have that there exists a positive constant c_{21} such that,

$$\Pr\{S_2 \geq c_{21}d_n^3\delta/n\} \leq Q_4 + Q_6 + Q_7 \tag{2.51}$$

As to S_3,

$$\begin{aligned} S_3 &\leq ||\mathbf{a}||^2 \; ||\mathbf{b}_n^\top \mathbf{H}_n B_n - \mathbf{b}^\top \mathbf{H}\mathbf{b}|| \\ &\leq ||\mathbf{a}||^2 (||(\mathbf{b}_n - \mathbf{b})^\top \mathbf{H}_n (\mathbf{b}_n - \mathbf{b})|| + 2||(\mathbf{b}_n - \mathbf{b})^\top \mathbf{H}_n \mathbf{b}|| + ||\mathbf{b}^\top (\mathbf{H}_n - \mathbf{H})\mathbf{b}||) \\ &\quad \triangleq ||\mathbf{a}||^2 (S_{31} + 2S_{32} + S_{33}) \end{aligned} \tag{2.52}$$

Note that

$$S_{31} \leq ||\mathbf{b}_n - \mathbf{b}||^2 \; ||\mathbf{H}_n|| \tag{2.53}$$

By (2.27),(2.32), we have that there exists a positive constant c_{22} such that,

$$\Pr\{S_{31} \geq c_{22} d_n^5 (\delta^2/n^2 + \delta/n)^2\} \leq Q_5 + Q_6 \tag{2.54}$$

As to S_{32},

$$S_{32} \leq ||\mathbf{b}_n - \mathbf{b}|| \; ||\mathbf{H}_n|| \; ||\mathbf{b}|| \tag{2.55}$$

By (2.17),(2.27),(2.32),(2.36), we have that there exists a positive constant c_{23} such that,

$$\Pr\{S_{32} \geq c_{23} d_n^{5/2} (\delta^2/n^2 + \delta/n)\} \leq Q_5 + Q_6 \tag{2.56}$$

As to S_{33},

$$S_{33} \leq ||\mathbf{b}||^2 \; ||\mathbf{H}_n - \mathbf{H}|| \tag{2.57}$$

By (2.17),(2.42), we have that there exists a positive constant c_{24} such that,

$$\Pr\{S_{33} \geq c_{24} (d_n^4 \delta^3/n^3 + d_n^2 \delta/n)\} \leq Q_8 \tag{2.58}$$

Combining (2.16),(2.52),(2.53),(2.55),(2.57), we have

$$\begin{aligned} \Pr\{S_3 \geq c_{22} d_n^6 (\delta^2/n^2 + \delta/n)^2 + c_{23} d_n^{7/2} (\delta^2/n^2 + \delta/n) + c_{24} (d_n^5 \delta^3/n^3 + d_n^3 \delta/n)\} \\ \leq 2Q_5 + 2Q_6 + Q_8 \end{aligned} \tag{2.59}$$

Define $\varsigma(d_n, \delta) = c_{20} d_n^5 \delta^2/n^2 + c_{21} d_n^3 \delta/n + c_{22} d_n^6 (\delta^2/n^2 + \delta/n)^2 + c_{23} d_n^{7/2} (\delta^2/n^2 + \delta/n) + c_{24} (d_n^5 \delta^3/n^3 + d_n^3 \delta/n)$. Then, from (2.47),(2.49),(2.51),(2.59), we have that due to symmetry,

$$\Pr\{|\widehat{\lambda_{j1}^*} - \lambda_{j1}^*| \geq \varsigma(d_n, \delta)\} \leq 4Q_4 + 4Q_5 + 8Q_6 + 4Q_7 + 2Q_8 \tag{2.60}$$

By properly choosing the value of δ (i.e., taking $\delta = c_2(c_{22}+c_{23})^{-1}d_n^{-5/2}n^{1-2\kappa}$), we can make $\varsigma(d_n,\delta) = c_2 d_n n^{-2\kappa}$, for any $c_2 > 0$. Then, we have

$$\Pr(|\widehat{\lambda_{j1}^*} - \lambda_{j1}^*| \geq c_2 d_n n^{-2\kappa}) \leq \mathcal{O}\left(d_n^2 \exp(-c_3 n^{1-4\kappa} d_n^{-4}) + d_n \exp(-c_4 n d_n^{-7})\right) \quad (2.61)$$

The first part of Theorem 2.2.2 follows via the union bound of probability.

To prove the second part, we define an event

$$\mathcal{B}_n \equiv \{\max_{j\in\mathcal{D}_n} |\widehat{\lambda_{j1}^*} - \lambda_{j1}^*| \leq c_1\xi d_n n^{-2\kappa}/2\}$$

By Lemma 2.2.1, we have

$$\widehat{\lambda_{j1}^*} \geq c_1 \xi d_n n^{-2\kappa}/2, \forall j \in \mathcal{D}_n \quad (2.62)$$

Thus, by choosing $\nu_n = c_5 d_n n^{-2\kappa}$ with $c_5 \leq c_1\xi/2$. We have that $\mathcal{D}_n \subseteq \widehat{\mathcal{D}_{\nu_n}}$. Therefore,

$$\Pr(\mathcal{B}_n^c) \leq \mathcal{O}\left(s\{d_n^2 \exp(-c_3 n^{1-4\kappa} d_n^{-4}) + d_n \exp(-c_4 n d_n^{-7})\}\right)$$

Then, the probability bound for the second part of Theorem 2.2.2 is attained.

2.6.6 Proof Sketch of Theorem 2.2.3

Proof of Theorem 2.2.3. From subsection 2.2.2, we have that $\lambda_{j1}^* = \mathrm{E}(\phi_{nj}^{*2})$ and $\widehat{\lambda_{j1}^*} = P_n(\phi_{nj}^{*2})$.

From equation (2.5), after obtaining θ_{nj}^* where $\mathrm{var}(\theta_{nj}^*) = 1$, ϕ_{nj}^* can be obtained via the following optimization problem.

$$\underset{\phi_{nj}\in\mathcal{S}_n}{\arg\min} \quad \mathrm{E}[\{\theta_{nj}^*(Y) - \phi_{nj}(X_j)\}^2], \text{ where } \phi_{nj}(X_j) = \boldsymbol{\eta}_j^\top \boldsymbol{\psi}_j(X_j).$$

Therefore, $\phi_{nj}^* = \boldsymbol{\psi}_j^\top \mathrm{E}\left(\boldsymbol{\psi}_j \boldsymbol{\psi}_j^\top\right)^{-1} \mathrm{E}\boldsymbol{\psi}_j \theta_{nj}^*$.

We notice that the only difference between our proof and the proof of Theorem 2.2.3 in Fan et al. (2011) is the role of Y. As MC-SIS essentially uses a transformation of Y, we can not deal directly with Y. However, from the formulation above, θ_{nj}^* here plays the same role as Y in Fan et al. (2011). With this connection, our proof follows immediately by replacing Y in the proof of Theorem 2.2.3 in Fan et al. (2011) with θ_{nj}^*.

3. SPARSE OPTIMAL TRANSFORMATION

3.1 Introduction

Regression analysis is arguably one of the most commonly used tools for data analysis in practice. Suppose Y is the response variable of interest and $\mathbf{X} = (X_1, \ldots, X_p)$ is the vector of p predictor variables. Based on a finite sample of Y and $\mathbf{X}$, regression analysis is commonly used to discover how and to which degree the predictor variables X_j's affect Y. In its generality, regressing Y against $\mathbf{X}$ is to infer the dependence structure of Y on $\mathbf{X}$. However, most existing regression methods are usually focused on certain characteristics of Y such as the mean (i.e. $\mathrm{E}(Y|\mathbf{X})$), median (i.e. 50th percentile of $Y|\mathbf{X}$), or other quantiles of $Y|\mathbf{X}$. These methods are useful when the chosen characteristics are of primary interests, but may fail to capture the full dependence structure of Y on $\mathbf{X}$. A number of attempts were made in the literature to directly estimate the conditional distribution $P(Y|\mathbf{X})$ (Rosenblatt, 1969; Fan et al., 1996; Sugiyama et al., 2010). The resulting approaches, unfortunately, suffer severely from the curse of dimensionality and are thus not practical (Efromovich, 2007).

Another approach to exploring the dependence of Y on $\mathbf{X}$ is to first apply transformations to Y and $\mathbf{X}$ and then perform regression analysis to the transformed variables. Intuitively, different transformations can lead to the discovery of different aspects of the dependence structure of Y on $\mathbf{X}$. The well-known Box-Cox transformation and additive model can be considered two such approaches. Box and Cox (1964) proposed to apply power transformation to the response variable Y in regression analysis, which aims to make the assumptions of linearity, normality, and homogeneity more appropriate. Different from the Box-Cox transformation, the additive model assumes that Y depends on the transformations of individual predictor variables in an additive fashion, and fitting the additive model is to identify those transformations (Hastie and Tibshirani, 1990). Despite the popularity of the Box-Cox transformation and additive model, their effectiveness can be compromised due to their susceptibility to model mis-specification. For example, both will fail in simple cases like $Y = \log(X_1 + X_2^2 + \epsilon)$.

Breiman and Friedman (1985) proposed to apply general nonparametric transformations to both Y and $\mathbf{X}$, and further to identify the *optimal transformations* that achieve the maximum correlation between them. The optimal transformations can be equivalently stated as the solution to the following least squares problem.

$$\begin{aligned} \min_{h\in L^2(P_Y), f_j \in L^2(P_{X_j})} \quad & \mathrm{E}\Big[\{h(Y) - \sum_{j=1}^{p} f_j(X_j)\}^2\Big], \\ \text{s.t.} \quad & \mathrm{E}[h(Y)] = \mathrm{E}[f_j(X_j)] = 0; \\ & \mathrm{E}[h^2(Y)] = 1, \mathrm{E}[f_j^2(X_j)] < \infty. \end{aligned} \tag{3.1}$$

Here, P_Y and P_{X_j} denote the marginal distributions of Y and X_j, respectively, and $L^2(P)$ denotes the class of square integrable functions under the measure *P*. We denote the solution to (3.1) as h^* and $f_j^*(j = 1, \ldots, p)$, which are referred to as the optimal transformations for Y and $\mathbf{X}$, respectively. A set of sufficient conditions are given in Breiman and Friedman (1985, Section 5.2) for the existence of optimal transformations.

Breiman and Friedman further developed the Alternating Conditional Expectation (ACE) algorithm to compute the optimal transformations. Although in general, the optimal transformations are not expected to fully capture the dependence structure of Y on $\mathbf{X}$, they represent in a certain sense the most important features of the dependence structure. Notice that the transformed predictors are additive for the transformed response. This additive structure is important because it ensures the interpretability of the captured dependence, that is, it shows how the predictors jointly affect the transformed response. In order to uncover the remaining dependence, intuitively, the idea of transformation can be iteratively applied. In this thesis, however, we will focus on the optimal transformations only.

The optimal transformations are subject to two limitations. Firstly, without any shape constraint, the transformation on the response $h^*(Y)$ may not be easily interpretable. In many real life applications such as modeling utility functions in economics, $h^*(Y)$ may not be meaningful if the order of the observations cannot be preserved after transformation. The problem becomes worse if the primary interest after transformation is to predict Y instead of $h^*(Y)$. Secondly, despite the additive structure, the estimation of optimal transformations can suffer from the curse of dimensionality when the number of predictor variables p is large. Even when the optimal transformations can be effectively estimated, the retention of a large number of spurious predictors can compromise their interpretability and prediction capacity.

To overcome those two limitations of the optimal transformations, in this chapter, we first propose to impose the monotonicity constraint on the transformation of Y. This constraint ensures that the transformed response variable is interpretable and invertible, and subsequently the prediction of Y can be performed. Second, in order to eliminate the spurious predictor variables, we regularize the estimation procedure of the optimal transformations by using a special type of penalty called the Smooth Integration of Counting and Absolute deviation (SICA) penalty (Lv and Fan, 2009). We refer to the resulting optimal transformations as the SParse Optimal Transformations or SPOT in short.

Existing methods that are closely related to SPOT include those developed for sparse additive models. Lin and Zhang (2006) proposed the COSSO procedure, which assumes that each component function belongs to a Reproducing Kernel Hilbert space (RKHS). COSSO uses the sum of the RKHS norms of the component functions as a penalty for simultaneous variable selection and model fitting. Ravikumar et al. (2007) introduced an approach called SPAM that penalizes the sum of L_2 norms of the component functions and is effectively a functional version of the group lasso Yuan and Lin (2006). Meier et al. (2009), Huang et al. (2010) and Balakrishnan et al. (2012) also developed different methods for sparse high-dimensional additive models by using different types of penalty functions.

Our proposed approach SPOT is distinct from the existing methods in two main aspects. Firstly, SPOT considers transformations on both Y and $\mathbf{X}$ with former being subject to the monotonicity constraint. The monotone transformation can be crucial for the cases where the usual additive model for Y does not hold. For such cases, the existing methods may fail to identify the dependence of Y on $\mathbf{X}$, whereas SPOT can still be successful. The monotone transformation clearly includes the identity function as a special case, therefore, SPOT is expected to work well when the additive model for Y indeed holds. Secondly, the SICA penalty used in SPOT enjoys many advantages over other types of penalty existing in the literature. The family of SICA functions proposed by Lv and Fan (2009) forms a smooth homotopy between the L_0 and L_1 types of penalty, and include the L_0 and L_1 penalty as limiting cases. SICA can avoid the drawbacks of the L_0 and L_1 penalties while combining their strengths and lead to more stable estimates of model parameters and less stringent conditions under which variable selection consistency can be established (See Section 3.3.2 for more details).

Due to the use of monotone transformation on Y and the SICA penalty, SPOT produces sparse optimal transformations that are interpretable and can be further used for prediction. We extended the ACE algorithm to compute the estimates of the sparse optimal transformations. Furthermore, we establish the consistency results for SPOT under various regularity conditions and assumptions. Our simulation study and real data application provide convincing evidence of SPOT's effectiveness in performing variable selection, exploring complex dependence structures, and performing prediction for the response. We believe SPOT can become an effective tool for high dimensional exploratory regression analysis in practice.

The rest of the chapter is organized as follows. Section 3.2 introduces basic notations used in this chapter. In Section 3.3, we formally define the sparse optimal transformation problem, propose the SICA penalty and the monotone transformation, and further develop the algorithm for estimating the sparse optimal transformations. The theoretical results on the estimation and selection consistency of sparse optimal transformation are given in Section 3.4. Experimental results based on simulation study and real data applications are reported in Section 3.5. Section 3.6 provides some discussions of the proposed methods. The proofs of the theorems and more simulations results are included in Section 3.7.

3.2 Notations and Assumptions

Let $h(Y)$ and $f_j(X_j)$ denote the transformations of Y and X_j, $j = 1, \ldots, p$. We assume the supports of Y and X_j's are compact, and they are further assumed to be [0,1] without loss of generality. Throughout this chapter, $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ is assumed to be an i.i.d. sample of $\mathbf{X}$ and Y.

For each $j = 1, \ldots, p$, let $\mathcal{H}_{\mathcal{X}_j}$ denote the Hilbert space of measurable functions $f_j(X_j)$ with $\mathrm{E}[f_j(X_j)] = 0$ and the inner product $\langle f_j, f_j' \rangle = \mathrm{E}(f_j f_j')$, where f_j' is an arbitrary function in $\mathcal{H}_{\mathcal{X}_j}$. Note that the expectations are taken over the probability distribution of X_j and $\mathrm{E}[f_j^2(X_j)] < \infty$. Let $\mathcal{H}_{\mathcal{X}}^{+} = \mathcal{H}_{\mathcal{X}_1} \oplus \mathcal{H}_{\mathcal{X}_2} \oplus \cdots \oplus \mathcal{H}_{\mathcal{X}_p}$ be the Hilbert space of functions of $\mathbf{X}$ that take an additive form: $\mathbf{f}(\mathbf{X}) = \sum_{j=1}^{p} f_j(X_j)$, with $f_j \in \mathcal{H}_{\mathcal{X}_j}$. Let $L^2[0,1]$ be the Hilbert space of square integrable functions under the Lebesgue measure and $\{\psi_{jk} : k = 1, 2, \ldots\}$ denote a uniformly bounded, orthonormal basis of $L^2[0,1]$. To impose smoothness on each f_j, we only consider $f_j \in \mathcal{T}_j$, where $\mathcal{T}_j$ is the Sobolev ball of order two, that is, $\mathcal{T}_j = \{f_j \in \mathcal{H}_{\mathcal{X}_j} : f_j = \sum_{k=1}^{\infty} \beta_{jk}\psi_{jk}, \sum_{k=1}^{\infty} \beta_{jk}^2 k^4 \leq C^2\}$ for some $0 < C < \infty$. To impose smoothness on h, we require that h should be r times continuously differentiable and its r-th derivative be Hölder continuous: $|h^{(r)}(y_1) - h^{(r)}(y_2)| \leq c|y_1 - y_2|^v$ for all y_1 and y_2, for some $0 < v \leq 1$ and $c > 0$. We use $\mathcal{M}$ to denote the set of functions satisfying this condition.

3.3 Sparse Optimal Transformations

3.3.1 Sparse Optimal Transformation Problem

Different from SPAM, we consider an additional transformation on the response Y, which aims to model more complex structures from data. As discussed in the Section 3.1, in order to make the transformation of Y interpretable and suitable for prediction, the transformation h needs to be a monotone function. Without loss of generality, we require h

to be monotone increasing in this thesis. Then, the sparse optimal transformation (SPOT) problem can be defined as follows.

$$\begin{aligned} \min_{h\in\mathcal{M},\mathbf{f}:f_j\in\mathcal{T}_j} \quad & \mathcal{L}(h,\mathbf{f})+\lambda\Omega\left(\mathbf{f}\right), \\ \text{s.t.} \quad & \mathrm{E}[h^2]=1, \quad h'\geq 0; \end{aligned} \tag{3.2}$$

where

$$\mathcal{L}(h,\mathbf{f})=\frac{1}{2}\mathrm{E}\left[\left(h(Y)-\sum_{j=1}^{p}f_j(X_j)\right)^2\right],$$

$$\Omega(\mathbf{f})=\sum_{j=1}^{p}\rho\left(\sqrt{\mathrm{E}[f_j^2(X_j)]}\right)$$

Here, $\mathcal{M},\mathcal{T}_j$ are the function spaces defined previously in Section 3.2, λ is the regularization parameter and ρ is a pre-specified penalty function.

3.3.2 SICA Penalty

As discussed in the Introduction, we choose to use the SICA penalty as ρ, which is denoted as $\rho := \rho_a(t)$ where

$$\rho_a(t)=\left(\frac{t}{a+t}\right)I(t\neq 0)+\left(\frac{a}{a+t}\right)t,\ t\in[0,\infty), \tag{3.3}$$

and

$$\rho_0(t)=\lim_{a\to 0+}\rho_a(t)=I(t\neq 0);$$

$$\rho_\infty(t)=\lim_{a\to\infty}\rho_a(t)=t.$$

A visulization for the SICA penalty for a few a values are depicted in Figure 3.1.

It is clear that $\rho_0(\cdot)$ and $\rho_\infty(\cdot)$ correspond to the L_0 and L_1 penalty functions, respectively. As a changes from zero to infinity, $\rho_a(\cdot)$ forms a smooth homotopy between the L_0 and L_1 penalty functions. Therefore, the SICA penalty with $0<a<\infty$ represents a compromise between the L_0 and L_1 penalty functions, while the L_0 and L_1 penalty functions can be considered the limiting cases.

Regularized regression methods using the L_0 penalty demonstrate different performances in parameter estimation, variable selection and computing than those using the

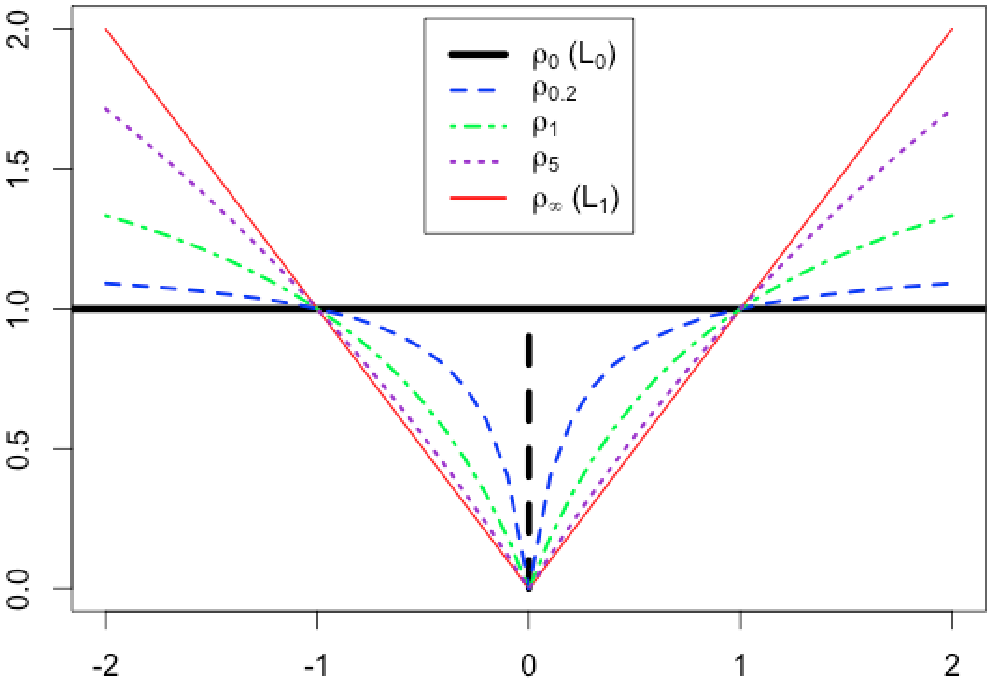

Figure 3.1. Plot of SICA penalty functions for a few a values.

L_1 penalty. The L_0 penalty is directly imposed on the number of variables, and thus is the original measure of model complexity. The L_0 penalty does not cause bias in estimation and can lead to consistency in variable selection under fairly general conditions (e.g. BIC of Schwarz (1978)). It however suffers from the instability problem (Breiman, 1996) and can become quickly infeasible in computing when the number of variables increases. On the other hand, as a convex relaxation of the L_0 penalty, the L_1 penalty enjoys the advantages of stability and simplicity in computing (Tibshirani, 1996), but it can lead to noticeably large bias in estimation (Fan and Li, 2001) and achieve variable selection consistency only under stringent conditions such as the irrepresentable condition for the lasso (Zhao and Yu, 2006).

From (3.3), it can be seen that the SICA penalty in some sense can be considered a combination of the L_0 and L_1 penalty with the weights being dependent on t, and the tuning parameter a determines where the SICA penalty stands between the L_0 and L_1 penalty. Lv and Fan (2009) proposed a unified framework for regularizing least squares-based methods using the SICA penalty and investigated the properties of the resulting estimator under the linear model. It turns out that the SICA penalty possesses a number of advantages. Firstly, not like the L_0 penalty, the SICA penalty is continuous in t, therefore, stable and efficient algorithms can be developed to solve the SICA-regularized least squares problem. Secondly, the condition under which the SICA penalty can lead to variable selection consistency is much less restrictive than the irrepresentable condition under the L_1 or lasso penalty. The fundamental reason for the second advantage is given as follows. When the tuning parameter a approaches to zero, the SICA penalty approaches the L_0 penalty, and helps the regularized method explore a broader solution or model space. (We note that a cannot get too close to zero in practice; otherwise, the computation will start to become unstable.) In summary, the SICA penalty manages to combine the strengths of the L_0 and L_1 penalty while avoiding their limitations. We believe that the SICA penalty is not simply a variant of the popularly used L_1 penalty, and it is in fact a significant improvement and should be widely adopted in practice. Other good properties related the SICA penalty can be found in Nikolova (2000), Lin and Lv (2013) and Lv and Liu (2014).

When the tuning parameter a is sufficiently large, the behavior of the SICA penalty is very similar to the L_1 penalty, and in such a case, we propose to directly use the L_1 penalty. Therefore, we include both the SICA penalty and the L_1 penalty when we implement SPOT in a computing package. When the SICA penalty is used, we refer to our procedure as SPOT-SICA, and when the L_1 penalty is used, we refer to our procedure as SPOT-LASSO. We remark that SPOT-LASSO can be considered a special case of SPOT-SICA with $a = \infty$, and SPAM a special case of SPOT-LASSO with $h(Y) = Y$.

3.3.3 Monotone Transformation on Response

Let $\mathcal{S}_{q\ell_n}$ be the space of polynomial splines of degree $q \geq 1$ with equally-spaced knots. Let $\{B_m, m = 1, \ldots, \ell_n\}$ denote a normalized B-spline basis with $||B_m||_{\sup} \leq 1$, where

$||\cdot||_{\text{sup}}$ is the sup-norm. Then, $\widetilde{h}(Y) = \sum_{m=1}^{\ell_n} \alpha_m B_m(Y)$ for any $\widetilde{h}(Y) \in \mathcal{S}_{q\ell_n}$. One example of the B-spline basis functions is depicted in Figure 2.1.

It is shown in De Boor (2001) that for any $h \in \mathcal{M}$ defined in Section 3.2, there exists a function $\tilde{h} \in S_{q\ell_n}$ such that $||\tilde{h} - h||_{\text{sup}} = O(\ell_n^{-(r+v)})$, with $q \geq r + v$. The constraint that the transformation h is monotone increasing in the SPOT problem (3.2) can be readily accommodated in B-spline approximation. According to Schumaker (1981), a sufficient condition for a polynomial spline $\tilde{h}(Y)$ to be monotone increasing is that its coefficients satisfy the linear constraints $\alpha_m \geq \alpha_{m-1}$ for $m = 2, \ldots, \ell_n$. When using the centered B-spline basis, the linear constraints become $\alpha_1 \geq 0, \alpha_m \geq \alpha_{m-1}$ for $m = 2, \ldots, \ell_n$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{\ell_n})^\top$. The linear constraints can be further written as $D^\top \boldsymbol{\alpha} \geq 0$, where D is the $\ell_n \times \ell_n$ matrix defined as

$$D = I_{\ell_n - 1} - \begin{bmatrix} \mathbf{0}_{\ell_n-1} & I_{\ell_n-1} \\ 0 & \mathbf{0}_{\ell_n-1}^\top \end{bmatrix}.$$

Here, I_k is the $k \times k$ identity matrix, and $\mathbf{0}_{\ell_n - 1}$ is the $\ell_n - 1$ dimensional vector of 0's. Denote $\mathbf{B}$ as the $n \times \ell_n$ matrix where $\mathbf{B}(i,k) = B_k(Y_i)$. Then, in terms of the sample, we have $\tilde{h}(\mathbf{Y}) = \mathbf{B}\boldsymbol{\alpha}$ where $\tilde{h}(\mathbf{Y}) = (\tilde{h}(Y_1), \ldots, \tilde{h}(Y_n))^\top$.

3.3.4 SPOT Algorithm

Recall that $\{\psi_{jk} : k = 1, 2, \ldots\}$ is an orthonormal basis and $f_j = \sum_{k=1}^{\infty} \beta_{jk}\psi_{jk}$. We use $\widetilde{f}_j = \sum_{k=1}^{d_n} \beta_{jk}\psi_{jk}$ to approximate f_j, where d_n is a truncation parameter. Thus, $\widetilde{f}_j$ is a smoothed approximation to f_j. It is well-known that for the second order Sobolev ball $\mathcal{T}_j$, we have $||f_j - \widetilde{f}_j||_2^2 = \mathcal{O}(1/d_n^4)$ for $f_j \in \mathcal{T}_j$. Let Ψ_j denote the $n \times d_n$ matrix where $\Psi_j(i,k) = \psi_{jk}(X_{ij})$ and $\boldsymbol{\beta}_j := (\beta_{j1}, \ldots, \beta_{jd_n})^\top$. We have $\widetilde{f}_j(X_j) = \Psi_j \boldsymbol{\beta}_j$ where $\widetilde{f}_j(X_j) = (\widetilde{f}_j(X_{1j}), \ldots, \widetilde{f}_j(X_{nj}))^\top$. Recall that $\tilde{h}(\mathbf{Y}) = \mathbf{B}\boldsymbol{\alpha}$ and D defined in Section

3.3.3. The sample version of the SPOT problem (3.2) with the SICA penalty can be written as follows.

$$\begin{aligned} \min_{\substack{\boldsymbol{\alpha}\in R^{\ell_n} \\ \boldsymbol{\beta}_j\in R^{d_n}}} \frac{1}{2n} & \left\| \mathbf{B}\boldsymbol{\alpha} - \sum_{j=1}^{p} \Psi_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda_n \sum_{j=1}^{p} \rho_a \left(\frac{\|\Psi_j \boldsymbol{\beta}_j\|_2}{\sqrt{n}} \right) \\ \text{s.t.} \qquad & \frac{1}{n} \boldsymbol{\alpha}^\top \mathbf{B}^\top \mathbf{B} \boldsymbol{\alpha} = 1; \quad D^\top \boldsymbol{\alpha} \geq 0. \end{aligned} \tag{3.4}$$

We develop a coordinate descent procedure to solve (3.4). The estimation procedure is summarized in Algorithm 1. To facilitate the calculation of the SICA penalty, we apply the local linear approximation (LLA) method proposed in Zou and Li (2008) to $\rho_a(t)$, which is $\rho_a(t) \approx \rho_a'(t_0)t + \rho_a(t_0) - \rho_a'(t_0)t_0$ and $\rho_a'(t_0) = a(a+1)/(a+t_0)^2$ in a neighborhood of t_0. We explain the two main components of Algorithm 1 below.

Suppose the current estimates of transformations are given as $\hat{h}^{(0)}, \hat{f}_1^{(0)}, \ldots, \hat{f}_p^{(0)}$, and we want to update f_j next. Denote $\hat{R}_j^{(0)} = \hat{h}^{(0)} - \sum_{k\neq j} \hat{f}_k^{(0)}$. Applying the LLA method, the objective function in (3.4) can be simplified to

$$\frac{1}{2n} \left\| \hat{R}_j^{(0)} - \Psi_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda_n w_j \frac{1}{\sqrt{n}} \|\Psi_j \boldsymbol{\beta}_j\|_2 , \tag{3.5}$$

where $w_j = a(a+1)/(a+t_j)^2$ and $t_j = \frac{1}{\sqrt{n}} ||\hat{f}_j^{(0)}||_2$. Notice that w_j only depends on the current estimate $\hat{f}_j^{(0)}$. Therefore, updating $\boldsymbol{\beta}_j$ in (3.5) is essentially equivalent to solving a weighted group lasso problem (Huang et al., 2012) with respect to one group, and the update of $\boldsymbol{\beta}_j$ admits an explicit expression as follows.

$$\hat{\boldsymbol{\beta}}_j = \left[1 - \frac{\lambda_n w_j \sqrt{n}}{||\Psi_j (\Psi_j^\top \Psi_j)^{-1} \Psi_j^\top \hat{R}_j^{(0)}||_2} \right]_+ (\Psi_j^\top \Psi_j)^{-1} \Psi_j^\top \hat{R}_j^{(0)}$$

where $[\cdot]_+$ denotes the positive part. Therefore, f_j can be updated as

$$\hat{f}_j = \Psi_j \hat{\boldsymbol{\beta}}_j = \left[1 - \frac{\lambda_n w_j \sqrt{n}}{||\hat{P}_j^{(0)}||_2} \right]_+ \hat{P}_j^{(0)} \tag{3.6}$$

where $\hat{P}_j^{(0)} = \Psi_j (\Psi_j^\top \Psi_j)^{-1} \Psi_j^\top \hat{R}_j^{(0)}$.

Note that the objective function for SPOT-LASSO is equivalent to (3.5) with $w_j = 1$. In such a case, we do not need to use the LLA method, and Algorithm 1 can be used directly to calculate SPOT-LASSO by specifying $w_j = 1$.

After updating f_j for $j = 1, \dots, p$, we fix $\hat{\mathbf{f}} = \hat{f}_1 + \dots + \hat{f}_p$ and further update h (i.e., update $\boldsymbol{\alpha}$). Problem (3.4) becomes

$$\begin{aligned} \min_{\boldsymbol{\alpha} \in R^{\ell n}} \quad & \frac{1}{2n} \left\| \mathbf{B}\boldsymbol{\alpha} - \hat{\mathbf{f}} \right\|_2^2 \\ \text{s.t.} \quad & \frac{1}{n} \boldsymbol{\alpha}^\top \mathbf{B}^\top \mathbf{B} \boldsymbol{\alpha} = 1; \\ & D^\top \boldsymbol{\alpha} \geq 0; \end{aligned} \tag{3.7}$$

which is equivalent to a typical quadratic programing problem. Standard numeric packages, such as the R package “quadprog”, can be used to solve problem (3.7), and we obtain $\widehat{\boldsymbol{\alpha}}$ as the estimate of $\boldsymbol{\alpha}$.

Algorithm 1 SPOT-SICA Coordinate Descent Algorithm

```
Input: Data $\{\mathbf{X}_i, Y_i\}_{i=1}^n$, tuning parameters $\lambda$, $a$
Initialize $\hat{h} = Y/||Y||_2$, $\hat{f}_j = 0$ for $j = 1, \dots, p$
Iterate (I) - (II) until convergence:
(I) Update $\hat{f}_j$, for each $j = 1, \dots, p$;
    Compute the residual: $\widehat{R}_j = \hat{h} - \sum_{k \neq j} \hat{f}_k$
    Calculate $\widehat{P}_j = \Psi_j (\Psi_j^\top \Psi_j)^{-1} \Psi_j^\top \widehat{R}_j$
    Compute weight $w_j$:
        $w_j = a(a+1)/(a + ||\hat{f}_j||_2/\sqrt{n})^2$ for finite $a$
        $w_j = 1$ for $a = \infty$ ($L_1$ penalty)
    Soft thresholding, $\hat{f}_j = \left[1 - \lambda w_j \sqrt{n}/||\widehat{P}_j||_2\right]_+ \widehat{P}_j$
    Centering, $\hat{f}_j = \hat{f}_j -$ mean $(\hat{f}_j)$
(II) Update $\hat{h}$;
    Solve $\widehat{\boldsymbol{\alpha}}$ in problem (3.7) by Quadratic Programming
    Obtain $\hat{h}$ = $\mathbf{B}\widehat{\boldsymbol{\alpha}}/||\mathbf{B}\widehat{\boldsymbol{\alpha}}||_2$
Output: Fitted functions $\hat{h}$ and $\hat{f}_j, j = 1, \dots, p$
```

3.4 Theoretical Properties

In this section, we discuss the theoretical properties of SPOT-SICA in variable selection and parameter estimation. In particular, we establish the consistency of SPOT-SICA under the transformation model and a given estimate of the response transformation. We assume that observations of $(Y, \mathbf{X})$ come from the following transformation model $h^*(Y) = \sum_{j=1}^{p} f_j^*(X_j) + \epsilon$, where h^* and f_j^* are the optimal transformations. Rewriting the transformation model in terms of an orthonormal basis $\{\psi_{jk}\}$, we have that

$$h^*(Y) = \sum_{j=1}^{p} \sum_{k=1}^{\infty} \beta_{jk}^* \psi_{jk}(X_j) + \epsilon. \tag{3.8}$$

The transformation model is a general model that encompasses a broad class of models in both statistics and econometrics (Linton et al., 2008; Jacho-Chávez et al., 2010; Chiappori et al., 2015). We use the transformation model to facilitate our theoretical discussion, and show that under the transformation model, SPOT-SICA can recover the true model with probability approaching one asymptotically. In the case that the true distribution of $\mathbf{X}$ and Y is more complex, the transformation model can also be used as an approximate model due to its flexibility, and our numerical results show that SPOT-SICA can still be used as an effective tool for variable selection.

Let S denote the set of true variables $S = \{j, f_j^* \neq 0\}$, and s_n the cardinality of S, and S^c its complement. We show that SPOT-SICA can correctly identify S and consistently estimate $\boldsymbol{\beta}_j^*$ in (3.8) for $j \in S$.

Recall that Ψ_j is the $n \times d_n$ matrix obtained from the sample, we use Ψ_S to denote the $n \times s_n d_n$ matrix formed by stacking the matrices $\Psi_j, j \in S$ one after another. We state the assumptions under which the main results hold.

Assumption 1 *We assume that the following assumptions hold.*

(A) Let $\tau_n = \min_{j \in S} ||\Psi_j \beta_j^*||/\sqrt{n}$. *It holds that* $n^{\alpha}\tau_n \to \infty$ *with* $\alpha \in (0, 1/2)$.

(B) It holds that $\rho'(\tau_n/2) = o(n^{-\alpha} d_n^{-1} \lambda_n^{-1} s_n^{-1/2})$ *and* $\sup_{t \geq \tau_n/2} \rho''(t) = o(\lambda_n^{-1})$.

(C) There exists a positive constant c_0 such that

$$c_0 \leq \min_{j \in S} \mathit{\Lambda}_{\min}\left(\frac{1}{n}\Psi_j^\top \Psi_j\right) \leq \mathit{\Lambda}_{\max}\left(\frac{1}{n}\Psi_S^\top \Psi_S\right) \leq c_0^{-1}$$

where $\mathit{\Lambda}_{\min}$ and $\mathit{\Lambda}_{\max}$ are the smallest and largest eigenvalues of a matrix, respectively.

(D) It holds that

$$\max_{j \in S^c} \left\| \Psi_j (\Psi_j^\top \Psi_j)^{-1} \Psi_j^\top \Psi_S \ (\Psi_S^\top \Psi_S)^{-1} \right\|_{\infty,2} \leq \frac{\sqrt{c_0}}{2\sqrt{n}} \frac{\rho'(0+)}{\rho'(\tau_n/2)} \tag{3.9}$$

where for a matrix A, $||A||_{\infty,2} = \sup_{||x||_\infty = 1} ||Ax||_2$ with x being a vector.

(E) The errors $\epsilon_i, i = 1, \ldots, n$, are independent and identically distributed as $N(0, \sigma^2)$.

This set of assumptions is adopted from Fan et al. (2015). Assumption (A) places a lower bound on the decaying rate of the signal strength of the true predictors $j \in S$. Assumption (B) can be satisfied by penalty functions with flat tails. Assumption (C) assumes that eigenvalues for the design matrix corresponding to true predictors are bounded from below and above. If Ψ_S is orthogonal, Assumption (C) is satisfied with $c_0 = 1$. Assumption (D) is similar to the Irrepresentable Condition for L_1 penalty that ensures selection consistency of Lasso (Zhao and Yu, 2006). When $a \to 0$ (e.g., $a = o(\tau_n)$), Condition (D) is automatically satisfied. More detailed discussions on these assumptions can be found in Fan et al. (2015, Appendix B).

It is worth noting that equation (3.9) reflects the restriction on the design matrix for SPOT-SICA to be consistent in variable selection. For fixed sample size n, the quantify $\rho'(0+)/\rho'(\tau_n/2)$ plays a critical role in Assumption (D). For the SICA penalty, we have $\rho'(0+)/\rho'(\tau_n/2) = (1 + \tau_n/(2a))^2$. The smaller a is, the less restrictive Assumption (D) becomes. As $a \to 0$, $\rho'(0+)/\rho'(\tau_n/2)$ approaches ∞. Therefore, for any given fixed design matrix, Assumption (D) will eventually be satisfied when a is sufficiently small, and SPOT-SICA will have a high probability of selecting the true model. On the other hand, as $a \to \infty$, ρ_a approaches the L_1 penalty ρ_∞, and $\rho'(0+)/\rho'(\tau_n/2)$ approaches 1; In other words, the right hand side of (3.9) becomes smaller and smaller, and Assumption (D) becomes more and more restrictive. When a is too large, Assumption (D) may fail to hold, and SPOT-SICA or its limiting version SPOT-LASSO may fail to select the true variables. In theory, it appears that a small a should always be preferred. Unfortunately, this is not true, because as we remarked previously, when a is too small, the SICA penalty in general will incur computational instability and produce inferior results.

Assumption 2 *There exits an L_∞-consistent estimate $\hat{h}^*$ of h^* and $||\hat{h}^* - h^*||_{L_\infty} = \sup_{y\in[0,1]} |\hat{h}^*(y) - h^*(y)| = \mathcal{O}_p(\upsilon_n)$ for some sequence $\upsilon_n = o(\lambda_n)$, where λ_n is the regularization parameter in (3.4).*

Assumption 2 assumes that there exists a good estimate of the transformation h^*. This assumption is valid since there are several procedures proposed for obtaining such an estimate in the literature (Linton et al., 2008; Chiappori et al., 2015). In particular, Chiappori et al. (2015) showed that under certain conditions, h^* can be estimated at the parametric rate for high dimensional data. For ease of presentation, we do not present the details along that direction but instead state it as an assumption.

Theorem 3.4.1 *Assume that $d_n + \log p = O(n\lambda_n^2)$, $\lambda_n n^\alpha d_n \sqrt{s_n} \to 0$, $\log(pd_n) = o(n^{1-2\alpha} s_n^{-1} d_n^{-2})$, and $s_n d_n^{-2} + \nu_n = o(\lambda_n)$. Then under Assumptions 1 and 2, with probability approaching one as n goes to infinity, there exits a local minimizer $\hat{\boldsymbol{\beta}}$ of (3.4) such that:*
(1) $\hat{\boldsymbol{\beta}}_{S^c} = 0$;
(2) $||\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*||_\infty \leq c_0^{1/2} n^{-\alpha} d_n^{-1/2}$;
where $||\cdot||_\infty$ stands for the infinity norm of a vector.

Theorem 3.4.1 establishes the weak oracle property for SPOT-SICA in that SPOT-SICA not only identifies the true model, but also estimates the true coefficients consistently. The sketch of the proof is given in Section 3.7.1 and the main idea of the proof follows Fan et al. (2015).

Remark 3.4.1 *Although Theorem 3.4.1 states a result of a local minimizer, it has been proved in Loh and Wainwright (2015) that any local minimizer will fall within statistical precision of the true parameter vector under appropriate conditions on the penalty function. Therefore, the results are extensible to all local minimizers with suitable constraints on the penalty.*

3.5 Numerical Results

In this section, we compare the performances of SPOT-SICA, SPOT-LASSO and SPAM in variable selection and prediction through both synthetic and real-life examples. For SPAM, we use its implementation in the R package “SAM”. Similar to the implementation of SPAM, we use B-spline bases for function approximation in SPOT-SICA and SPOT-LASSO.

3.5.1 Effectiveness on Synthetic Data

We test the methods using data sampled from two types of models, the additive model and the transformation model. In the first example, we consider an additive model where SPAM is expected to work well. In the second example, a typical transformation model is considered. For each training data set, we also generate a validation data set and a test data set. Validation datasets are used to choose the tuning parameters λ and a, and test datasets are used to measure the prediction accuracy of the estimated models in terms of mean squared error (MSE). The goal of using separate validation datasets and test datasets is to facilitate fair comparisons of different methods. We replicate each simulation 100 times, and report the averages and standard deviations (in parentheses) of precisions, recalls, sizes of the selected variables, F_1 scores, as well as MSEs of the estimated models on the test datasets. More simulation examples can be found in Section 3.7.2.

Example 3.5.1 *(Additive Model)*
$Y = \sum_{j=1}^{p} f_j(X_j)+\epsilon$ *where* $\epsilon \sim N(0, 8/9)$*; The functions are given by* $f_1(x) = -2\sin(2x)$,

$f_2(x) = x^2$, $f_3(x) = \frac{2\sin(x)}{2-\sin(x)}$, $f_4(x) = \exp(-x)$, $f_5(x) = x^3 + 1.5(x-1)^2$, $f_6(x) = x$, $f_7(x) = 3\sin(\exp(-0.5x))$, $f_8(x) = -5\Phi(x, 0.5, 0.8^2)$, and $f_j = 0$ for $j \geq 9$. Here, $\Phi(\cdot, \mu, \sigma^2)$ is the Gaussian cumulative distribution function with mean μ and standard deviation σ.

We generate covariates with a compound symmetry covariate structure as follows. Each covariate $X_j = (W_j + tU/3)/(1 + t/3), j = 1, \ldots, p$, where $W_1, \ldots, W_p$ and U are from $Unif(-2.5, 2.5)$. As t increases, the correlation between any two predictors will increase, which renders the variable selection problem more difficult in general. The sample size is $n = 200$, and we consider the dimension of covariates $p = 50$ and 200. Each component function $f_j (j = 1, \ldots, 8)$ are appropriately scaled as in Ravikumar et al. (2007) and Yin et al. (2012).

The simulation results are summarized in Table 3.1. We can see from Table 3.1 that SPOT-SICA always outperforms SPAM and SPOT-LASSO in terms of both F_1 score and prediction accuracy. The superior performance of SPOT-SICA is due to the use of SICA penalty for variable selection and estimation. Because of the advantages of SICA discussed in Section 3.3.2, SPOT-SICA can simultaneously screen out more spurious variables and produce less biased estimates of the function components, thus achieve better selection precision and lower prediction error. The performances of SPAM and SPOT-LASSO are mostly comparable in variable selection, because they both use the L_1 penalty. In terms of prediction, SPAM and SPOT-LASSO perform similarly in the cases when the predictors are sampled independently ($t = 0$). When data are sampled from more complex structures ($t = 3$ and $t = 6$), SPAM outperforms SPOT-LASSO, since SPOT-LASSO does not utilize the additive structure of the model.

Example 3.5.2 *(Transformation Model)*

$$Y = \log\left(4 + \sin(2\pi X_1) + |X_2| + X_3^2 + X_4^3 + X_5 + \epsilon\right)$$

Table 3.1.
Comparison of different methods on simulated data from Example 3.5.1.

p	t	Method	Precision	Recall	Size	F_1 score	MSE
50	0	SPAM	0.31 (0.07)	1.00 (0.00)	27.28 (5.72)	0.47 (0.08)	1.60 (0.21)
50	0	SPOT-LASSO	0.43 (0.13)	1.00 (0.00)	20.46 (6.04)	0.59 (0.12)	1.56 (0.22)
50	0	SPOT-SICA	0.83 (0.24)	1.00 (0.00)	11.23 (5.80)	0.88 (0.17)	1.37 (0.20)
50	3	SPAM	0.28 (0.06)	1.00 (0.00)	29.44 (5.63)	0.44 (0.07)	1.59 (0.22)
50	3	SPOT-LASSO	0.26 (0.08)	1.00 (0.00)	32.55 (7.70)	0.41 (0.09)	1.93 (0.36)
50	3	SPOT-SICA	0.84 (0.21)	1.00 (0.00)	10.83 (5.58)	0.89 (0.16)	1.40 (0.21)
50	6	SPAM	0.26 (0.05)	1.00 (0.00)	31.76 (5.68)	0.41 (0.06)	1.63 (0.22)
50	6	SPOT-LASSO	0.26 (0.10)	0.96 (0.09)	32.97 (9.81)	0.40 (0.10)	2.20 (0.33)
50	6	SPOT-SICA	0.75 (0.19)	0.95 (0.08)	11.14 (4.27)	0.82 (0.13)	1.58 (0.31)
200	0	SPAM	0.20 (0.05)	1.00 (0.00)	43.41 (11.13)	0.33 (0.07)	1.76 (0.24)
200	0	SPOT-LASSO	0.32 (0.11)	1.00 (0.00)	29.84 (16.17)	0.47 (0.13)	1.66 (0.30)
200	0	SPOT-SICA	0.85 (0.22)	1.00 (0.00)	11.18 (8.28)	0.90 (0.17)	1.36 (0.24)
200	3	SPAM	0.17 (0.04)	1.00 (0.00)	50.98 (12.96)	0.28 (0.06)	1.77 (0.28)
200	3	SPOT-LASSO	0.14 (0.06)	1.00 (0.00)	67.87 (31.14)	0.25 (0.10)	2.33 (0.47)
200	3	SPOT-SICA	0.89 (0.15)	1.00 (0.00)	9.44 (3.33)	0.94 (0.10)	1.39 (0.22)
200	6	SPAM	0.15 (0.03)	0.99 (0.03)	53.79 (12.25)	0.27 (0.05)	1.86 (0.28)
200	6	SPOT-LASSO	0.21 (0.18)	0.85 (0.16)	55.73 (36.24)	0.29 (0.16)	2.48 (0.37)
200	6	SPOT-SICA	0.70 (0.22)	0.89 (0.12)	11.84 (6.09)	0.75 (0.14)	1.75 (0.39)

where $\epsilon \sim N(0, 1/4)$. We sample covariates according to the same procedure in Example 3.5.1, except that we sample $W_1, \ldots, W_p$ and U now from $Unif(-1, 1)$. The change is to ensure that the term in the log-function is positive.

The simulation results are summarized in Table 3.2. We see that it is consistent in all cases that SPOT-SICA outperforms SPOT-LASSO, and SPOT-LASSO outperforms SPAM, in both variable selection precision and prediction accuracy. In addition, although the re-

Table 3.2.
Comparison of different methods on simulated data from Example 3.5.2.

p	t	Method	Precision	Recall	Size	F_1 score	MSE
50	0	SPAM	0.28 (0.11)	1.00 (0.00)	19.84 (5.89)	0.43 (0.11)	3.49 (0.46)
50	0	SPOT-LASSO	0.39 (0.17)	1.00 (0.02)	15.44 (7.04)	0.54 (0.16)	2.22 (0.45)
50	0	SPOT-SICA	0.83 (0.28)	1.00 (0.04)	8.00 (6.68)	0.87 (0.22)	2.08 (0.36)
50	3	SPAM	0.27 (0.08)	0.99 (0.05)	19.62 (5.36)	0.42 (0.09)	2.62 (0.37)
50	3	SPOT-LASSO	0.37 (0.16)	0.97 (0.08)	15.62 (6.24)	0.51 (0.15)	2.06 (0.29)
50	3	SPOT-SICA	0.78 (0.27)	0.95 (0.10)	7.67 (5.11)	0.82 (0.20)	1.95 (0.29)
50	6	SPAM	0.29 (0.11)	0.94 (0.10)	18.34 (6.37)	0.43 (0.11)	2.66 (0.37)
50	6	SPOT-LASSO	0.32 (0.16)	0.92 (0.13)	17.91 (8.26)	0.44 (0.14)	2.12 (0.29)
50	6	SPOT-SICA	0.70 (0.27)	0.86 (0.15)	8.09 (5.60)	0.72 (0.18)	2.00 (0.34)
200	0	SPAM	0.20 (0.09)	1.00 (0.00)	29.22 (10.92)	0.32 (0.11)	3.75 (1.23)
200	0	SPOT-LASSO	0.31 (0.19)	1.00 (0.02)	23.63 (18.69)	0.45 (0.20)	2.25 (0.83)
200	0	SPOT-SICA	0.79 (0.32)	1.00 (0.00)	10.68 (13.23)	0.84 (0.27)	2.06 (0.65)
200	3	SPAM	0.19 (0.08)	0.96 (0.09)	30.32 (12.71)	0.30 (0.11)	2.74 (0.43)
200	3	SPOT-LASSO	0.31 (0.22)	0.92 (0.13)	25.04 (21.33)	0.42 (0.20)	2.17 (0.36)
200	3	SPOT-SICA	0.79 (0.27)	0.94 (0.11)	8.12 (7.50)	0.82 (0.21)	1.93 (0.31)
200	6	SPAM	0.16 (0.06)	0.86 (0.15)	29.80 (11.37)	0.27 (0.08)	2.80 (0.45)
200	6	SPOT-LASSO	0.22 (0.13)	0.83 (0.17)	25.82 (15.54)	0.32 (0.13)	2.19 (0.37)
200	6	SPOT-SICA	0.62 (0.33)	0.81 (0.16)	11.19 (10.71)	0.63 (0.24)	2.04 (0.37)

sults on the average size of estimated supports are similar for SPAM and SPOT-LASSO, SPAM is much worse compared to SPOT-LASSO in terms of prediction accuracy. This observation supports the claim that the additional transformation on Y in SPOT is providing more flexibility in capturing the complex dependence structure. One sample of the estimated optimal transformations from SPOT-SICA is visualized in Figure 3.2, which match the true functions well. To further assess the variability of the transformation estimates, we

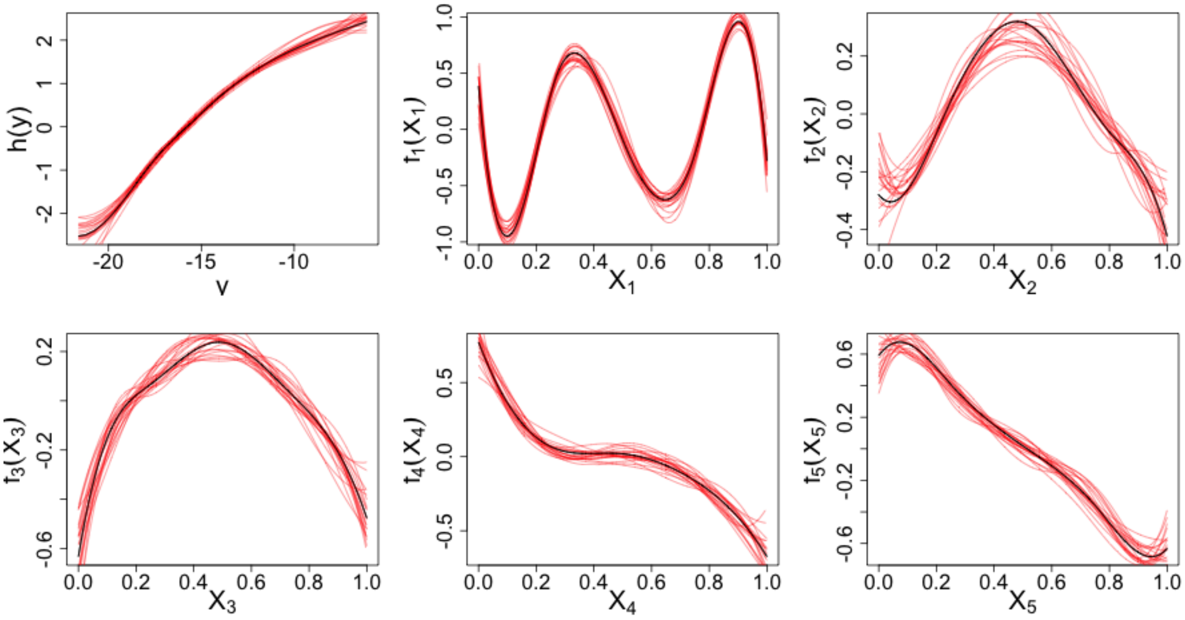

Figure 3.2. Transformations of Y and X_1 to X_5 obtained from SPOT-SICA ($a = 1$) in Example 3.5.2 ($p = 50, t = 0$). The black line is the estimated transformation from original data, red lines are estimated transformations from 20 bootstrapped samples.

run SPOT-SICA on bootstrapped samples and plot resulting transformations in Figure 3.2, as suggested in Breiman and Friedman (1985).

3.5.2 Role of Parameter a in Variable Selection

In this experiment, using the same model as in Example 3.5.2, we investigate the role played by the tuning parameter a in the SICA penalty in model selection accuracy. As discussed in Section 3.4, SPOT-SICA can achieve variable selection consistency under Assumption (D). The smaller a is, the less restrictive the assumption is. To demonstrate this effect, we choose two values of a, $a = 1$ and ∞, and compare their performances under different design matrices. We vary t from $\{1, 2, 3, 4, 5, 6\}$ to represent different levels of variable selection difficulty.

For any fixed a and a given sample, the performance of SICA depends on the regularization parameter λ. SPOT-SICA is declared to have a success if there exists a λ under which SPOT-SICA correctly select all true variables. For each value of t, we simulate 10 samples of $\mathbf{X}$ that leads to 10 design matrices. For each design matrix, we randomly sample the error term 100 times, and then apply SPOT-SICA and record their successes and failures. Consequently, we obtain 10 success rates, each over 100 random replicates. We plot these success rates at each value of t in Figure 3.3. From Figure 3.3, we see that SPOT-SICA ($a = 1$) outperforms SPOT-LASSO ($a = \infty$) by consistently selecting the correct model. As expected, when t increases, selecting the correct model becomes more difficult for both values of a. However, SPOT-SICA still has a higher chance to select the correct model even when SPOT-LASSO fails.

Next, we choose two fixed values of t, which are $t = 0$ and $t = 2$, but vary a from $\{0.05, 0.10, 0.50, 1.00, 2.00, 5.00\}$. For each fixed pair of t and a, we repeat the previous procedure and record the average success rate. Results are presented in Table 3.3. Results from the L_1 penalty ($a = \infty$) are also recorded in the last column. We see that as a becomes larger, the performance of the SICA penalty is approaching that of the L_1 penalty. When a gets closer to zero, the chance of selecting a true model will first increase and then decrease, this suggests that the computational difficulty increases as the SICA penalty approaches the L_0 penalty. The pattern exists for both $t = 0$ and $t = 2$. This phenomenon has also been pointed out in Lv and Fan (2009).

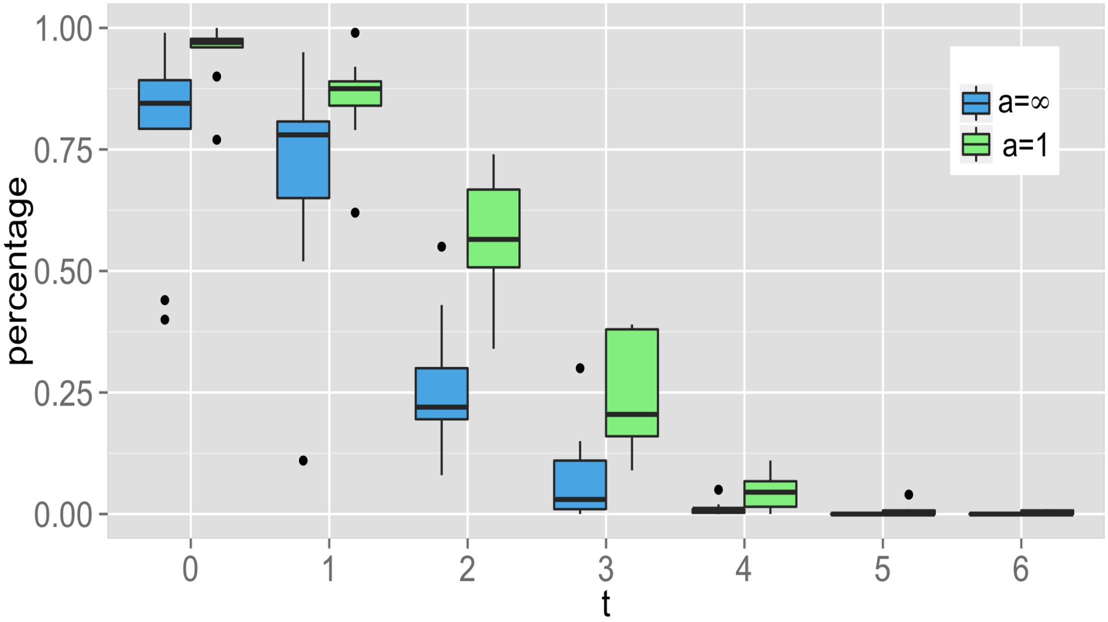

Figure 3.3. Impact of a on selection consistency of SPOT under different correlation structure controlled by t. Comparison between result from $a = 1$ and $a = \infty$, where $a = \infty$ corresponds to the L_1 penalty.

Table 3.3.

Average percentages of times that the true model can be selected by SPOT-SICA with different choices of a. The last column corresponds to the result from SPOT-LASSO.

a	0.05	0.10	0.50	1.00	2.00	5.00	∞
t=0	0.945	0.967	0.982	0.947	0.903	0.834	0.781
t=2	0.573	0.632	0.679	0.578	0.427	0.318	0.256

3.5.3 Real Data Application

We apply SPOT-SICA to two real datasets from the UCI Machine Learning Repository[1], which are the Boston Housing Data[2] and the Communities and Crime Data[3].

Boston Housing Data

The *Boston Housing Data* was collected to study the house values in the suburbs of Boston; The dataset contains $n = 506$ observations with $p = 10$ covariates, which are RM, AGE, DIS, TAX, PTRATIO, BLACK, LSTAT, CRIM, INDUS, NOX. To explore the variable selection property of SPOT-SICA, we follow the approach of Ravikumar et al. (2007) and add 20 noise variables in the analysis. The first ten noise variables are randomly drawn from $Unif(0, 1)$, and the other ten noise variables are a random permutation of the original ten covariates.

We adopt the commonly used “one-standard-error” rule with 10-fold cross-validation to select the tuning parameters λ and a, where we choose the most parsimonious model whose error is no more than one standard error above the error of the best model. We apply the SPOT-SICA to the 30-dimensional dataset with the selected tuning parameters. SPOT-SICA correctly zeros out both types of irrelevant variables, and it identifies five nonzero components out of the original ten covariates. The important variables are RM, DIS, TAX, PTRATIO, LSTAT. The estimated transformation functions are depicted in Figure 3.4. From Figure 3.4, we found that the monotone transformation of the response may be needed to yield a better-fitted model. Furthermore, aside from the commonly recognized important variables, which are RM, TAX, PTRATIO and LSTAT, SPOT-SICA suggests that DIS is also important, which exhibits a clear nonlinear effect on the response MEDV.

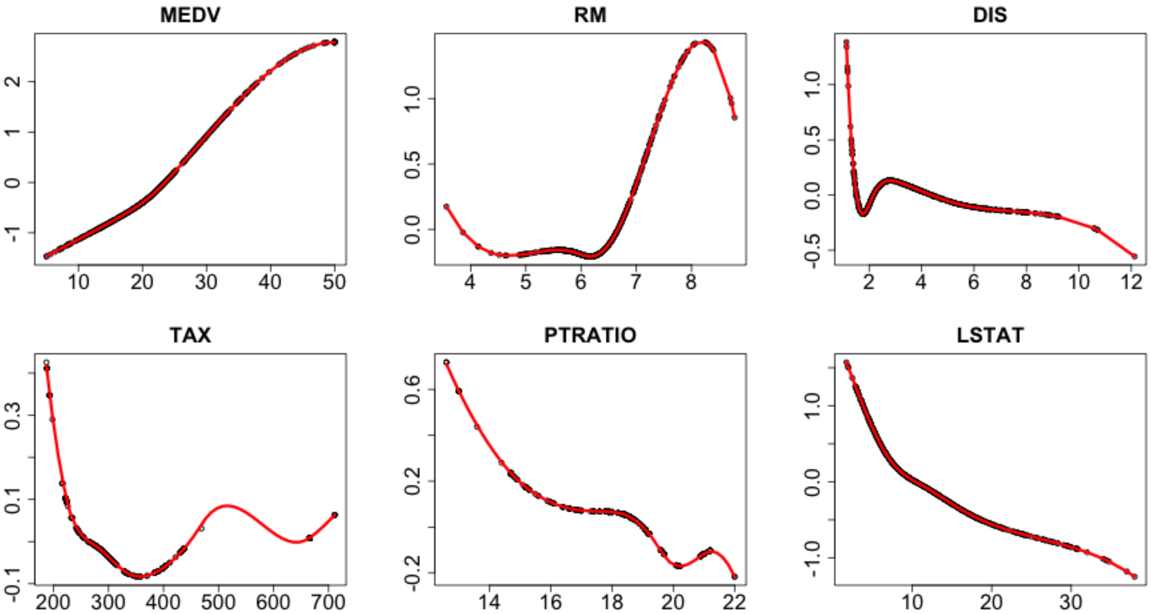

Figure 3.4. Estimated transformations of the response (MEDV) and selected predictors (RM, DIS, TAX, PTRATIO, LSTAT) by SPOT-SICA for the Boston Housing Data.

Communities and Crime Data

The *Communities and Crime Data* was first collected in Redmond and Baveja (2002) and it combines socio-economic data (the 1990 US Census), law enforcement data (the 1990 US LEMAS survey), and crime data (the 1995 FBI Uniform Crime Reporting) from several communities within the United States. The dataset consists of 1994 observations from 128 variables including ethnicity proportions, income, poverty rate, divorce rate etc., and was previously analyzed by Maldonado and Weber (2010); Song et al. (2011). We consider modeling the violent crime rate from other covariates in the dataset. By removing the covariates with missing values, we narrow down to 98 covariates.

[1] http://archive.ics.uci.edu/ml/

[2] https://archive.ics.uci.edu/ml/datasets/Housing

[3] http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

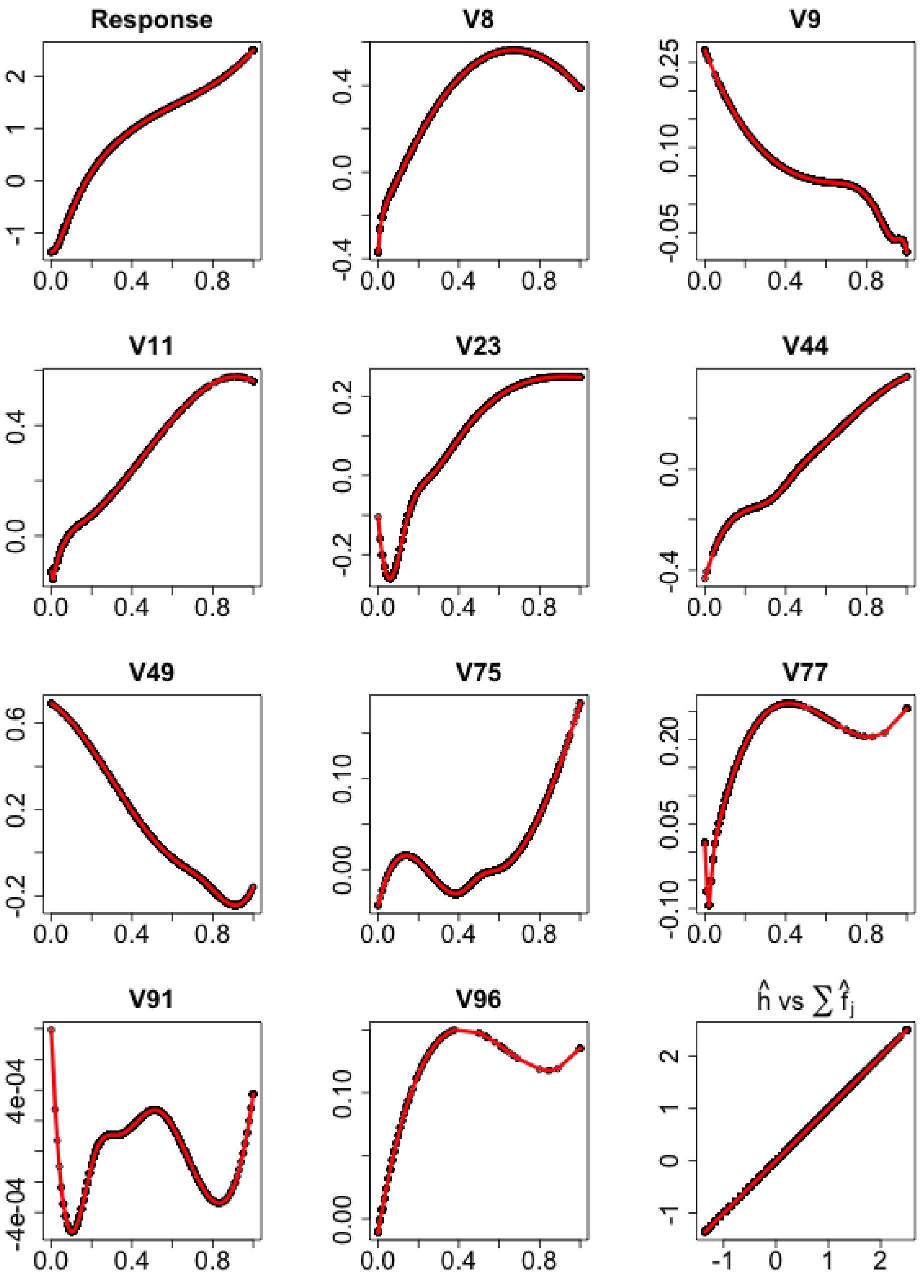

Figure 3.5. Estimated transformations of the response and 10 selected predictors by SPOT-SICA for the Communities and Crime Data. The labels above each graph corresponds to the orders of the covariates in the original data. The last graph is the plot of the estimated response transformation against the sum of all estimated transformations of selected variables.

We apply SPOT-SICA to the dataset, with tuning parameters selected by 10-fold cross-validation and the “one-standard-error” rule. As a result, SPOT-SICA selects 10 variables, which is fewer than 24 variables reported in Maldonado and Weber (2010). Moreover, the resulting estimates from SPOT-SCIA exhibit a higher prediction accuracy, with an average out-of-bag mean absolute error smaller than 0.093, which is better than the results from the proposed method in Maldonado and Weber (2010). Figure 3.5 shows the estimated transformations from applying SPOT-SICA. It is interesting to observe a clear nonlinear transformation of the response. Additionally, most selected variables exhibit nonlinear effects on the transformed response and a few others have nearly linear effects. Thus, our method effectively reduces the dimensionality of the data and is able to capture sensible linear and nonlinear relationships between the response and covariates.

3.6 Discussions

In this chapter, we develop a novel method called SPOT for exploring the dependence structure between the response Y and the predictor vector $\mathbf{X}$ in high dimensional data analysis. SPOT can consistently select important variables and automatically generate meaningful optimal transformations, under which the dependence structure can be best explored. SPOT demonstrates promising results on both simulated and real data in terms of selection consistency, estimation accuracy, prediction power, and interpretability.

One interesting direction to improve SPOT is to consider further transformations in addition to the optimal transformations, in order to capture the dependence of Y and $\mathbf{X}$ missed by optimal transformations. Another future direction is to investigate more relaxed conditions under which SPOT can possess selection and estimation consistency. We will pursue research in these two directions in the near future.

3.7 Technical Proofs and More Simulation Examples

We provide the proof sketch of Theorem 3.4.1 in Section 3.7.1. More simulation examples and results are included in Section 3.7.2.

3.7.1 Technical Proofs

Proof of Theorem 3.4.1

Proof Given $\hat{h}^*$, SPOT-SICA minimizes the following objective function

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\left\|\hat{h}^* - \sum_{j=1}^{p}\Psi_j\boldsymbol{\beta}_j\right\|_2^2 + \lambda_n\sum_{j=1}^{p}\rho\left(\frac{1}{\sqrt{n}}\|\Psi_j\beta_j\|_2\right). \tag{3.10}$$

Denote the difference between $f_j(X_{ij})$ and $\tilde{f}_j(X_{ij})$ as e_{ij} where

$$e_{ij} = f_j(X_{ij}) - \tilde{f}_j(X_{ij}) = \sum_{k=d_n+1}^{\infty}\beta^*_{jk}\psi_{jk}(X_{ij}) \tag{3.11}$$

With given $\hat{h}^*$, from the fact that $h^*(Y_i) = \sum_{j=1}^{p}\sum_{k=1}^{\infty}\beta^*_{jk}\psi_{jk}(X_{ij}) + \epsilon_i$, we have that

$$\hat{h}^*(Y_i) = \sum_{j=1}^{p}\sum_{k=1}^{\infty}\beta^*_{jk}\psi_{jk}(X_{ij}) + \epsilon^*_{i,1} + \epsilon^*_{i,2}, \tag{3.12}$$

where $\epsilon^*_{i,1} = \epsilon_i + \sum_{j=1}^{p}e_{ij}$ and $\epsilon^*_{i,2} = \hat{h}^*(Y_i) - h^*(Y_i)$.

Define

$$\mathcal{N} = \{\boldsymbol{\beta} \in R^{pd_n} : \boldsymbol{\beta}_{S^c} = 0, ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_\infty \leq \sqrt{c_0}d_n^{-1/2}n^{-\alpha}\} \tag{3.13}$$

Here are two lemmas which help the proof of Theorem 3.4.1.

Lemma 3.7.1 *Define the event $\mathcal{E}_1 = \{||\Psi_S^\top(\epsilon^*_{i,1}+\epsilon^*_{i,2})||_\infty \leq n\lambda_n/2\}$. Assume that $\lambda_n n^\alpha d_n\sqrt{s_n} \to 0$, then under Condition 1 and 2, and conditional on event $\mathcal{E}_1$, there exists a vector $\boldsymbol{\beta} \in \mathcal{N}$ such that $\boldsymbol{\beta}_S$ is a solution to the following nonlinear equations*

$$-\frac{1}{n}\Psi_S^\top\left(\hat{h}^* - \Psi_S\boldsymbol{\beta}_S\right) + \mathbf{v}_S(\boldsymbol{\beta}_S) = 0, \tag{3.14}$$

where $\mathbf{v}_S(\boldsymbol{\beta})$ is a vector obtained by stacking $\mathbf{v}_k(\boldsymbol{\beta}) = \rho'\left(\frac{1}{\sqrt{n}}\|\Psi_j\beta_j\|_2\right)\frac{1}{\sqrt{n}}\frac{\Psi_k^\top\Psi_k\beta_k}{\|\Psi_k\beta_k\|_2}$, $k \in S$ one underneath another.

Proof see Lemma 1 in Fan et al. (2015). ■

Lemma 3.7.2 *Define the event $\mathcal{E}_2 = \{||\Psi_{S^c}^\top(\epsilon_{i,1}^* + \epsilon_{i,2}^*)||_\infty \leq n\lambda_n/2\}$. Assume that $s_n d_n^{-2} + \nu_n = o(\lambda_n)$, $d_n + \log p = O(n\lambda_n^2)$, and $\lambda_n n^\alpha d_n \sqrt{s_n} \to 0$. Then, under Condition 1 and 2 and conditional on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\hat{\boldsymbol{\beta}} \in \mathcal{N}$.*

Proof The proof is similar to Lemma 2 in Fan et al. (2015). The only difference is that we have to prove

$$\max_{j \in S^c} ||\Psi_j(\Psi_j^\top \Psi_j)^{-1}\hat{\mathbf{v}}_j|| \leq n^{-1/2}\rho'(0+), \tag{3.15}$$

where

$$\hat{\mathbf{v}}_j = n^{-1}\Psi_j^\top(\hat{h}^* - \Psi_S\hat{\boldsymbol{\beta}}_S) = n^{-1}\Psi_j^\top\Psi_S(\boldsymbol{\beta}_S^* - \hat{\boldsymbol{\beta}}_S) + n^{-1}\Psi_j^\top(\epsilon_{i,1}^* + \epsilon_{i,2}^*) \tag{3.16}$$

By (3.14), we have

$$\boldsymbol{\beta}_S^* - \hat{\boldsymbol{\beta}}_S = (\Psi_S^\top\Psi_S)^{-1}(n\mathbf{v}_S - \Psi_S^\top(\epsilon_{i,1}^* + \epsilon_{i,2}^*)). \tag{3.17}$$

Plugging this into $\hat{\mathbf{v}}_j$, we obtain that

$$||\Psi_j(\Psi_j^\top\Psi_j)^{-1}\hat{\mathbf{v}}_j|| \leq I_{1,j} + I_{2,j} \tag{3.18}$$

where

$$\begin{aligned} I_{1,j} &= \left\|\Psi_j(\Psi_j^\top\Psi_j)^{-1}\Psi_j^\top\Psi_S(\Psi_S^\top\Psi_S)^{-1}\mathbf{v}_S\right\|_2 \\ I_{2,j} &= \frac{1}{n}\left\|\Psi_j(\Psi_j^\top\Psi_j)^{-1}\Psi_j^\top(\mathbf{I} - \Psi_S(\Psi_S^\top\Psi_S)^{-1}\Psi^\top)(\epsilon_{i,1}^* + \epsilon_{i,2}^*)\right\|_2 \end{aligned} \tag{3.19}$$

It is proved in Fan et al. (2015) that

$$I_{1,j} < \frac{1}{2\sqrt{n}}\rho'(0+), \tag{3.20}$$

$$\begin{aligned} I_{2,j} \leq &\underbrace{\frac{1}{n}\left\|\Psi_j(\Psi_j^\top\Psi_j)^{-1}\Psi_j^\top(\mathbf{I} - \Psi_S(\Psi_S^\top\Psi_S)^{-1}\Psi^\top)\epsilon_{i,1}^*\right\|_2}_{I_{2,1,j}} + \\ &\underbrace{\frac{1}{n}\left\|\Psi_j(\Psi_j^\top\Psi_j)^{-1}\Psi_j^\top(\mathbf{I} - \Psi_S(\Psi_S^\top\Psi_S)^{-1}\Psi^\top)\epsilon_{i,2}^*\right\|_2}_{I_{2,2,j}}, \end{aligned} \tag{3.21}$$

and

$$\max_{j\in S^c} I_{2,1,j} = o_p(n^{-1}(d_n^{1/2} + \sqrt{(\log p)})). \tag{3.22}$$

For all $j \in S^c$, we have that

$$I_{2,2,j} \leq n^{-1}||\hat{h}^*(Y) - h^*(Y)||_2 = O_p(n^{-1/2}\nu_n). \tag{3.23}$$

Combining (3.21), (3.22) and (3.23), we have

$$\begin{aligned}\max_{j\in S^c} I_{2,j} &= o_p(n^{-1}(d_n^{1/2} + \sqrt{\log p}) + n^{-1/2}\nu_n) \\ &= o_p(\lambda_n/\sqrt{n}) < \rho'(0+)/(2\sqrt{n})\end{aligned} \tag{3.24}$$

Therefore, from (3.18), (3.20) and (3.24), we can show

$$\max_{j\in S^c} ||\Psi_j(\Psi_j^\top\Psi_j)^{-1}\hat{\mathbf{v}}_j|| \leq n^{-1/2}\rho'(0+).$$

■

We now proof that $P\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \to 1$.

Note that

$$P\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) = 1 - P\left(||\Psi^\top(\epsilon_{i,1}^* + \epsilon_{i,2}^*)||_\infty \geq n\lambda_n/2\right), \tag{3.25}$$

and

$$||\Psi^\top(\epsilon_{i,1}^* + \epsilon_{i,2}^*)||_\infty \leq ||\Psi^\top\epsilon||_\infty + ||\Psi^\top\mathbf{e}||_\infty + ||\Psi^\top(\hat{h}^*(Y) - h^*(Y))||_\infty, \tag{3.26}$$

where $\mathbf{e}$ is the $n \times p$ matrix of the (i, j)-th element $\mathbf{e}(i, j) = e_{ij}$.

It is proved in Fan et al. (2015) that

$$P\left(||\Psi^\top\epsilon||_\infty > n\lambda_n/8\right) \to 0, \tag{3.27}$$

and

$$||\Psi^\top\mathbf{e}||_\infty \leq n\lambda_n/4. \tag{3.28}$$

In order to prove $P\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \to 1$, we only need to show that

$$P\left(||\Psi^\top(\hat{h}^*(Y) - h^*(Y))||_\infty > n\lambda_n/8\right) \to 0. \tag{3.29}$$

Equation (3.29) holds from the result that

$$P\left(||\Psi^\top(\hat{h}^*(Y) - h^*(Y))||_\infty > n\lambda_n/8\right) \leq P\left(||\hat{h}^*(Y) - h^*(Y)||_\infty > \lambda_n/8\right), \tag{3.30}$$

together with Condition 2 that $||\hat{h}^*(Y) - h^*(Y)||_\infty = O_p(\nu_n)$ where $\nu_n = o(\lambda_n)$.

Combining (3.27), (3.28) and (3.29), we have $P\left(\mathcal{E}_1 \cap \mathcal{E}_2\right) \to 1$

■

3.7.2 More Simulation Examples

We consider more simulations on transformation models in Example 3.7.1 and some general models in Example 3.7.2 to test the performances of the proposed methods on variable selection and prediction.

Example 3.7.1 *(More transformation models)*
Let $V = 4 + \sin(2\pi X_1) + |X_2| + X_3^2 + X_4^3 + X_5 + \epsilon$*, where* $\epsilon \sim N(0, 1/4)$. *We consider the following transformation models,*

(A.1) $Y = 20/V$*;*

(A.2) $Y = 10\sqrt{V}$*;*

(A.3) $Y = V^2/5$*;*

(A.4) $Y = \exp\{V/3\}$*;*

(A.5) $Y = 10\exp\{1/V\}$.

All predictors X_j are generated independently from $Unif(-1, 1)$. Sample size is $n = 200$. Results for $p = 50$ and $p = 200$ from all models in Example 3.7.1 are summarized in Table 3.4.

The results are consistent with the statement in Example 3.5.2. That is, SPOT-SICA consistently outperforms SPOT-LASSO, and SPOT-LASSO outperforms SPAM, in both variable selection precision and prediction accuracy.

Example 3.7.2 *(Some general models)*
We consider the following general models.

(B.1) $Y = \exp(X_1) + X_2^2\epsilon$*;*

(B.2) $Y = (1 + X_1)^{X_2} + 0.1\epsilon$*;*

(B.3) $Y = X_1^3 + X_2^2 X_3 + 0.1\epsilon$*;*

(B.4) $Y = X_1 + X_2 + (X_3 + X_4)^3 + 0.1\epsilon$*;*

where $\epsilon \sim N(0, 1)$.

All four models considered here do not belong to transformation models. In particular, Model (B.1) represents one case that heterogeneity exists in the model; Model (B.3) incorporates the interaction terms of X_2 and X_3, or it can be considered that X_2 and X_3 form a group in the model; Model (B.4) represents another group structure in the model, where the additive term $X_3 + X_4$ can be considered to be in one function. We test our methods on these models to see how they perform under more general model settings.

All predictors X_j are generated independently from $Unif(-1, 1)$. Sample size is $n = 200$. Results for $p = 19$ from all models in Example 3.7.2 are summarized in Table 3.5. It is expected that variable selection is more difficult in these models compared to additive models in Example 3.5.1 and transformation models in Examples 3.5.2 and 3.7.1. However, we see from the results that even when the assumption on the transformation model does not hold, our proposed method can still be applied as a fairly effective tool for variable selection.

Table 3.4.
Comparison of different methods on simulated data from Example 3.7.1.

Model	p	Method	Precision	Recall	Size	F_1 score	MSE
A.1	50	SPAM	0.28 (0.10)	1.00 (0.00)	20.11 (7.11)	0.43 (0.12)	1.08 (0.38)
A.1	50	SPOT-LASSO	0.39 (0.24)	1.00 (0.03)	17.95 (10.63)	0.53 (0.22)	0.71 (0.37)
A.1	50	SPOT-SICA	0.70 (0.30)	0.99 (0.03)	9.89 (7.67)	0.78 (0.24)	0.64 (0.23)
A.2	50	SPAM	0.28 (0.10)	1.00 (0.00)	19.88 (6.29)	0.43 (0.11)	3.61 (0.41)
A.2	50	SPOT-LASSO	0.41 (0.19)	1.00 (0.00)	14.58 (6.07)	0.56 (0.17)	2.21 (0.32)
A.2	50	SPOT-SICA	0.84 (0.26)	1.00 (0.00)	7.51 (5.95)	0.88 (0.20)	2.07 (0.34)
A.3	50	SPAM	0.28 (0.10)	1.00 (0.00)	19.85 (6.54)	0.43 (0.11)	2.62 (0.34)
A.3	50	SPOT-LASSO	0.40 (0.18)	1.00 (0.00)	14.70 (5.57)	0.55 (0.17)	1.59 (0.21)
A.3	50	SPOT-SICA	0.89 (0.20)	1.00 (0.00)	6.17 (2.63)	0.93 (0.14)	1.49 (0.22)
A.4	50	SPAM	0.28 (0.10)	1.00 (0.00)	19.81 (6.25)	0.43 (0.11)	2.49 (0.41)
A.4	50	SPOT-LASSO	0.40 (0.18)	1.00 (0.00)	14.84 (6.14)	0.55 (0.17)	1.52 (0.24)
A.4	50	SPOT-SICA	0.85 (0.24)	1.00 (0.00)	6.94 (4.16)	0.90 (0.18)	1.42 (0.23)
A.5	50	SPAM	0.28 (0.10)	1.00 (0.04)	20.36 (7.75)	0.43 (0.12)	0.56 (0.34)
A.5	50	SPOT-LASSO	0.39 (0.24)	1.00 (0.03)	18.34 (11.09)	0.52 (0.22)	0.38 (0.28)
A.5	50	SPOT-SICA	0.68 (0.30)	0.98 (0.07)	10.23 (7.83)	0.76 (0.24)	0.35 (0.23)
A.1	200	SPAM	0.19 (0.08)	0.99 (0.04)	29.67 (11.33)	0.32 (0.11)	1.26 (0.93)
A.1	200	SPOT-LASSO	0.30 (0.20)	1.00 (0.02)	28.14 (23.10)	0.43 (0.23)	0.82 (0.72)
A.1	200	SPOT-SICA	0.73 (0.29)	0.99 (0.06)	9.66 (9.95)	0.80 (0.23)	0.76 (0.70)
A.2	200	SPAM	0.18 (0.07)	1.00 (0.00)	30.10 (9.46)	0.31 (0.09)	3.84 (0.46)
A.2	200	SPOT-LASSO	0.30 (0.16)	1.00 (0.02)	24.61 (23.95)	0.44 (0.18)	2.33 (0.37)
A.2	200	SPOT-SICA	0.80 (0.29)	1.00 (0.00)	9.76 (12.51)	0.85 (0.24)	2.10 (0.30)
A.3	200	SPAM	0.18 (0.07)	1.00 (0.02)	30.77 (9.89)	0.30 (0.09)	2.76 (0.32)
A.3	200	SPOT-LASSO	0.29 (0.14)	1.00 (0.02)	21.79 (15.64)	0.44 (0.16)	1.66 (0.26)
A.3	200	SPOT-SICA	0.80 (0.30)	1.00 (0.02)	8.73 (7.29)	0.85 (0.24)	1.50 (0.21)
A.4	200	SPAM	0.18 (0.07)	0.99 (0.03)	30.96 (10.94)	0.30 (0.09)	2.62 (0.40)
A.4	200	SPOT-LASSO	0.28 (0.15)	1.00 (0.00)	23.40 (16.55)	0.42 (0.17)	1.59 (0.31)
A.4	200	SPOT-SICA	0.82 (0.27)	0.99 (0.04)	7.95 (6.71)	0.87 (0.21)	1.45 (0.29)
A.5	200	SPAM	0.20 (0.09)	0.98 (0.06)	29.55 (12.41)	0.32 (0.11)	0.98 (2.86)
A.5	200	SPOT-LASSO	0.31 (0.23)	0.99 (0.04)	29.56 (24.76)	0.43 (0.24)	0.74 (2.62)
A.5	200	SPOT-SICA	0.68 (0.28)	0.98 (0.07)	9.90 (9.02)	0.76 (0.22)	0.71 (2.62)

Table 3.5.
Comparison of different methods on simulated data from Example 3.7.2.

Model	Method	Precision	Recall	Size	F_1 score	MSE
B.1	SPAM	0.33 (0.21)	0.73 (0.25)	5.92 (3.25)	0.42 (0.20)	0.21 (0.04)
B.1	SPOT-LASSO	0.54 (0.28)	0.74 (0.25)	3.33 (1.78)	0.60 (0.25)	0.21 (0.04)
B.1	SPOT-SICA	0.58 (0.30)	0.76 (0.25)	3.28 (1.97)	0.63 (0.26)	0.21 (0.04)
B.2	SPAM	0.45 (0.36)	0.88 (0.26)	7.64 (6.42)	0.50 (0.29)	12.84 (37.45)
B.2	SPOT-LASSO	0.44 (0.33)	0.94 (0.23)	7.62 (6.35)	0.53 (0.32)	11.55 (34.64)
B.2	SPOT-SICA	0.58 (0.36)	0.94 (0.20)	5.81 (5.61)	0.65 (0.32)	11.29 (34.29)
B.3	SPAM	0.33 (0.15)	0.84 (0.17)	8.81 (3.52)	0.46 (0.15)	0.05 (0.01)
B.3	SPOT-LASSO	0.61 (0.32)	0.76 (0.15)	5.28 (3.44)	0.62 (0.21)	0.05 (0.01)
B.3	SPOT-SICA	0.79 (0.30)	0.74 (0.14)	3.63 (2.48)	0.72 (0.18)	0.04 (0.01)
B.4	SPAM	0.37 (0.13)	1.00 (0.00)	11.88 (3.57)	0.53 (0.13)	0.64 (0.11)
B.4	SPOT-LASSO	0.81 (0.28)	1.00 (0.00)	6.24 (4.03)	0.86 (0.21)	0.56 (0.12)
B.4	SPOT-SICA	0.93 (0.19)	1.00 (0.02)	4.81 (2.56)	0.95 (0.14)	0.53 (0.12)

4. MAXIMUM CORRELATION-BASED STATISTICAL DEPENDENCE MEASURES

4.1 Introduction

How to measure dependence between random variables is a classical problem in statistics and machine learning. Pearson correlation is one commonly used dependence measure, which is defined between two univariate random variables and is a powerful tool to capture linear dependence. Since the invention of Pearson correlation, many other measures have been developed to measure not only linear but also nonlinear dependence between both univariate variables and multivariate variables. Examples include Maximum Correlation (Lancaster, 1957), COnstraint COvariance (COCO) (Gretton et al., 2004), Kernel Canonical Correlation (KCCA) (Gretton et al., 2005b), Hilbert-Schmidt Information Criteria (HSIC) (Gretton et al., 2005a), Distance Correlation (dCor) (Szekely et al., 2007), Maximal Information Coefficient (MIC) (Reshef et al., 2011), Randomized Dependence Coefficient (RDC) (Lopez-Paz et al., 2013), and Copula Dependence Coefficient (CDC) (Jiang and Ding, 2014). Additionally, there are other dependence measures developed in the feature screening literature, which focus more on detecting associations under specific models; see Fan et al. (2011), Fan and Song (2010), Hall and Miller (2009), Li et al. (2012a), Shao and Zhang (2014), and others.

Of those dependence measures, maximum correlation is gaining resurgent interests. A number of algorithms have been proposed to approximate maximum correlation, including Alternating Conditional Expectation (ACE) (Breiman and Friedman, 1985), B-spline approximation (Burman, 1991), and polynomial approximations (Bickel and Xu, 2009; Hall and Miller, 2011). Additionally, KCCA can also be used to approximate maximum correlation when measuring dependence between univariate random variables, as long as a proper kernel is chosen. Recently, RDC is developed as an estimator of maximum correlation for multivariate random variables; and CDC, which is based on maximum correlation, is also proposed to measure dependence in multivariate cases.

In this chapter, we introduce dependence measures based on maximum correlation, in both univariate and multivariate cases. In univariate cases, we first introduce *B-spline based Maximum Correlation* (BMC), where we estimate maximum correlation by directly approximating optimal transformations using B-splines. The problem of estimating maximum correlation turns out to be a generalized eigenvalue problem, and maximum correlation can be approximated by the largest eigenvalue of the generalized eigenvalue problem. One variant (T-BMC) using all the eigenvalues is constructed to obtain a more robust measure of independence, which is also computationally faster. In multivariate cases, we propose MBMC and T-MBMC by making use of tensor product B-splines to approximate optimal transformations for multivariate random variables.

This chapter is organized as follows. Section 4.2 reviews the concepts of maximum correlation, optimal transformation and their connection introduced in Chapter 2. Based on the connection, dependence measures (BMC, T-BMC, MBMC, T-MBMC) using B-splines are defined and their properties are discussed in Section 4.3. Hypothesis testing procedures are proposed in Section 4.4. Numerical examples are given in Section 4.5 to validate the empirical performances of proposed measures.

4.2 Maximum Correlation Coefficient and Optimal Transformation

In Chapter 2, we have introduced the concepts of maximum correlation and optimal transformation. We briefly reviewed them as follows.

The maximum correlation coefficient between univariate random variables X and Y is defined as

$$\rho^*(X,Y) = \sup_{\theta,\phi}\{\rho(\theta(Y),\phi(X)) : 0 < \mathrm{E}|\theta(Y)|^2 < \infty, 0 < \mathrm{E}|\phi(X)|^2 < \infty\}$$

where $\rho(X,Y)$ is the Pearson correlation, θ and ϕ are Borel-measurable functions of Y and X. Furthermore, θ^* and ϕ^* are often denoted as optimal transformations that attain the maximum correlation. The existence of maximum correlation is guaranteed through conditions similar to Conditions (C1) and (C2) in Chapter 2. For simplicity, we assume that maximum correlation we consider always exist throughout this chapter. Breiman and Friedman (1985) showed that θ^* and ϕ^* in maximum correlation can be obtained via the optimal transformation problem defined in (2.2), which is restated as follows.

$$\begin{aligned} \min_{\theta,\phi\in L_2(P)} \quad & e^2 = \mathrm{E}[\{\theta(Y)-\phi(X)\}^2], \\ \text{subject to} \quad & \mathrm{E}\{\theta(Y)\} = \mathrm{E}\{\phi(X)\} = 0; \\ & \mathrm{E}\{\theta^2(Y)\} = 1. \end{aligned} \tag{4.1}$$

Let e^{*2} be the minimum of e^2. Breiman and Friedman (1985) showed that

$$e^{*2} = 1 - \rho^{*2}; \tag{4.2a}$$

$$\mathrm{E}(\phi^{*2}) = \rho^{*2}. \tag{4.2b}$$

Therefore, we can estimate maximum correlation coefficient by approximating either minimized regression error or the optimal transformations. Due to the flexibility and nice theoretical property of B-spline compared with other algorithms stated in the Introduction, we choose it as our main tool in estimating maximum correlation coefficient. Moreover, we propose several other efficient dependence measures based on the spline approximation.

4.3 Dependence Measure

In this section, we first summarize the procedure of B-spline approximation of maximum correlation coefficient as introduced in Chapter 2, and then propose a more robust version based on it. Extensions to multivariate cases are also developed.

4.3.1 Univariate Case: BMC and T-BMC

As defined in Section 2.2.2, $\mathcal{S}_n$ is the space of polynomial splines of degree $\ell \geq 1$ and $\mathbf{B}(\cdot) = (B_1(\cdot), \ldots, B_{d_n}(\cdot))^T$ denotes the vector of d_n normalized basis functions with $||B_m||_{sup} \leq 1$. We have $\theta_n(Y) = \alpha^T\mathbf{B}(Y)$, $\phi_n(X) = \beta^T\mathbf{B}(X)$ for any $\theta_n(Y), \phi_n(X) \in \mathcal{S}_n$.

BMC: B-spline-based Maximum Correlation

The population version of B-spline approximation to the minimization problem (4.1) can be written as follows.

$$\begin{aligned} \min_{\theta_n,\phi_n\in\mathcal{S}_n} \quad & \mathrm{E}[\{\theta_n(Y)-\phi_n(X)\}^2], \\ \text{subject to} \quad & \mathrm{E}\{\theta_n(Y)\}=\mathrm{E}\{\phi_n(X)\}=0; \\ & \mathrm{E}\{\theta_n^2(Y)\}=1. \end{aligned} \tag{4.3}$$

Given sample $\{x_i, y_i\}_{i=1}^n$, an empirical version of optimization problem (4.3) becomes

$$\begin{aligned} \min_{\alpha,\beta\in\mathcal{R}^{d_n}} \quad & \frac{1}{n}\sum_{i=1}^n \left[\alpha^T\mathbf{B}(y_i)-\beta^T\mathbf{B}(x_i)\right]^2, \\ \text{subject to} \quad & \frac{1}{n}\sum_{i=1}^n \left[\alpha^T\mathbf{B}(y_i)\right]=\frac{1}{n}\sum_{i=1}^n \left[\beta^T\mathbf{B}(x_i)\right]=0; \\ & \frac{1}{n}\sum_{i=1}^n \left[\beta^T\mathbf{B}(y_i)\right]^2=1. \end{aligned} \tag{4.4}$$

Algorithm 2 summarizes the procedure of solving optimization problem (4.4), more detailed derivations can be found in Chapter 2.

The output in Algorithm 2 is the estimate of $\mathrm{E}(\phi_n^{*2})$. We denote the population version of largest eigenvalue as $\lambda_1 := \mathrm{E}(\phi_n^{*2})$, and its sample estimate as $\widehat{\lambda_1}$. Then, the square root of the λ_1 is the B-spline approximation to maximum correlation coefficient, and the square root of $\widehat{\lambda_1}$ is the sample estimate of maximum correlation. We thus call λ_1 B-spline based Maximum Correlation (BMC). It has been shown in Chapter 2 that the screening procedure based on BMC enjoys some nice theoretical properties for screening variables in ultrahigh dimensional data analysis. We next discuss the theoretical properties of BMC as a dependence measure.

Theoretical Properties of BMC

Theoretical properties of BMC and its sample estimates rely on the following two conditions.

Algorithm 2 BMC: B-spline estimate of maximum correlation between univariate random variables

```
Require: data $x_u$,$y_u$, size n, spline degree $\ell$ , knots size k
Construct $\mathbf{B}(x_u)$ and $\mathbf{B}(y_u)$: B-splines for each observation $x_u$ and $y_u$ $(u = 1, \dots, n)$
    with degree $\ell$ and knots number k. $\mathbf{B}(x_u)$ and $\mathbf{B}(y_u)$ are vectors of length k + $\ell$.
Centering $\mathbf{B}(x_u)$ and $\mathbf{B}(y_u)$:
    $\mathbf{B}(x_u) \longleftarrow \mathbf{B}(x_u) - n^{-1}\sum_{u=1}^{n}\mathbf{B}(x_u)$,
    $\mathbf{B}(y_u) \longleftarrow \mathbf{B}(y_u) - n^{-1}\sum_{u=1}^{n}\mathbf{B}(y_u)$.
Calculate $\mathbf{A}_{yy} = n^{-1}\sum_{u=1}^{n}\mathbf{B}(y_u)\mathbf{B}^T(y_u)$,
    $\mathbf{A}_{yx} = n^{-1}\sum_{u=1}^{n}\mathbf{B}(y_u)\mathbf{B}^T(x_u)$,
    $\mathbf{A}_{xx} = n^{-1}\sum_{u=1}^{n}\mathbf{B}(x_u)\mathbf{B}^T(x_u)$.
Decompose $\mathbf{A}_{yy}$ by SVD: $\mathbf{A}_{yy} = \mathbf{R}^T\mathbf{D}\mathbf{R}$ due to symmetry of $\mathbf{A}_{yy}$, and $\mathbf{R}^T\mathbf{R} = \mathbf{I}$.
Return the largest eigenvalue of the objective matrix $\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{A}_{yx}\mathbf{A}_{xx}^{-1}\mathbf{A}_{yx}^T\mathbf{R}^T\mathbf{D}^{-\frac{1}{2}}$.
```

Condition 1. The optimal transformations $\{\theta^*, \phi^*\}$ belong to a class of functions $\mathcal{F}$, whose r-th derivative $f^{(r)}$ exists and is Lipschitz of order α_1, that is, $\mathcal{F} = \{f : |f^{(r)}(s) - f^{(r)}(t)| \leq K|s-t|^{\alpha_1}$ for all $s, t\}$ for some positive constant *K*, where *r* is a nonnegative integer and $\alpha_1 \in (0, 1]$ such that $w = r + \alpha_1 > 0.5$.

Condition 2. The joint density of Y and X is bounded and the marginal densities of Y, X are bounded away from zero on their support.

Theorem 4.3.1 *(Independence)* *When two random variables X and Y are independent,*

$$\lambda_1 = 0.$$

Under Condition 1 and Condition 2, when $\lambda_1 = 0$,

$$\rho^{*2} \leq \mathcal{O}\left(1/\sqrt{k}\right).$$

where k is the number of knots which increases with n.

There is no standard way to determine the optimal value of k, theoretically, it is set to be $\mathcal{O}\left(n^{\gamma}\right)$ with $0 < \gamma < 1$. We see that when $\lambda_1 = 0$, $\rho^{*2} \to 0$ as n increases, which makes BMC a good dependence measure under all circumstances.

Theorem 4.3.2 *(Distance between λ_1 and ρ^{*2}) Under Condition 1 and Condition 2, with the same constants c_1 and w in the proof of Theorem 4.3.1,*

$$|\lambda_1 - \rho^{*2}| \leq c_1 k^{-w} + 2\sqrt{c_1 k^{-w}}.$$

Lemma 4.3.3 *(Consistency of $\widehat{\lambda_1}$ to λ_1) Define $\zeta(d_n, n) = d_n^2 \exp(-c_3 n^{1-4\kappa} d_n^{-4}) + d_n \exp(-c_4 n d_n^{-7})$ for positive constants c_3, c_4 and $\kappa \in [0, w/(2w+1))$. Under Condition 1 and Condition 2, for any $c_2 > 0$,*

$$\Pr\left(|\widehat{\lambda_1} - \lambda_1| \geq c_2 d_n n^{-2\kappa}\right) \leq \mathcal{O}\left(\zeta(d_n, n)\right). \tag{4.5}$$

Here, κ is an important parameter which determines the optimal rate of d_n. According to Condition 1, $w > 1/2$. Therefore, the upper end limit of κ's range is at least $1/4$ and at most $1/2$. For the choice of d_n, on one hand, we want d_n as large as possible to fit the splines well; but on the other hand, d_n is required to be of $o(n^{1/7})$ in order to ensure estimation consistency in Lemma 4.3.3. Further from Lemma 4.3.3, we see that d_n should be no greater than $\min\{o(n^{1/7}), o(n^{2\kappa})\}$. Therefore, the exact value of κ is not that important as long as it is larger than $1/14$, and the optimal rate of d_n can be as large as $o(n^{1/7})$.

As $|\widehat{\lambda_1} - \rho^{*2}| \leq |\widehat{\lambda_1} - \lambda_1| + |\lambda_1 - \rho^{*2}|$, the following theorem is a direct result by combining Theorem 4.3.2 and Lemma 4.3.3.

Theorem 4.3.4 *(Consistency of $\widehat{\lambda_1}$ to ρ^{*2}) Under Condition 1 and Condition 2, with the same notations in Theorem 4.3.1 and Theorem 4.3.3,*

$$\Pr\left(|\widehat{\lambda_1} - \rho^{*2}| \geq c_2 d_n n^{-2\kappa} + c_1 k^{-w} + 2\sqrt{c_1 k^{-w}}\right) \leq \mathcal{O}\left(\zeta(d_n, n)\right). \tag{4.6}$$

T-BMC: Trace of BMC

Let $\mathbf{B}(Y)$ and $\mathbf{B}(X)$ be the random vectors from B-spline functions of random variables Y and X, from a specified choice of knots with size d_n. From canonical correlation analysis (Anderson, 1984; Johnson and Wichern, 2007), we know that square root of λ_1 is the largest canonical correlation between $\mathbf{B}(Y)$ and $\mathbf{B}(X)$. Let $\widehat{\lambda_i}$ be the i-th largest eigenvalue of $\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{A}_{yx}\mathbf{A}_{xx}^{-1}\mathbf{A}_{yx}^{T}\mathbf{R}^{T}\mathbf{D}^{-\frac{1}{2}}$, and λ_i be its counterpart in population version of B-spline space. In Fact, square root of λ_i is the i-th cannonical correlation between $\mathbf{B}(Y)$ and $\mathbf{B}(X)$. That is, for any given $1 \leq i \leq d_n$,

$$\lambda_i^{1/2} = \max_{\alpha_i,\beta_i \in \mathbf{R}^{d_n}} \rho\left(\alpha_i^T\mathbf{B}(Y), \beta_i^T\mathbf{B}(X)\right) \tag{4.7}$$

where $\rho\left(\alpha_i^T\mathbf{B}(Y), \alpha_j^T\mathbf{B}(Y)\right) = 0$ and $\rho\left(\beta_i^T\mathbf{B}(X), \beta_j^T\mathbf{B}(X)\right) = 0$ for all $j = 1, \ldots, i-1$.

The counterparts of $\lambda_i^{1/2}$ in the original L_2 space can be defined as follows. For functions $\{\theta_i, \phi_i; i = 1, 2, \ldots\}$ with bounded positive second moments, let

$$r_i = \max_{\theta_i,\phi_i \in L_2(P)} \rho\left(\theta_i(Y), \phi_i(X)\right),$$

where $\langle\theta_i(Y), \theta_j(Y)\rangle_{L_2(P_Y)} = 0$ and $\langle\phi_i(X), \phi_j(X)\rangle_{L_2(P_X)} = 0$ for all $j = 1, \ldots, i-1$. Here, $\langle\cdot,\cdot\rangle$ is the inner product defined in corresponding L_2 spaces.

From the definition above, it is clear that $\lambda_i^{1/2}$ is just a spline approximation to r_i defined in L_2 space. We notice that while maximum correlation ρ^* (or equivalently, r_1) captures the first layer of dependence, it excludes other information on the dependence which can be characterized by r_i $(i = 2, 3, \ldots)$. Thus, making use of the subsequent r_i $(i = 2, 3, \ldots)$ can provide more comprehensive understandings on the overall dependence level. In this sense, according to equation (4.7), $\lambda_i (i \neq 1)$ may contain extra information besides the largest eigenvalue λ_1 in quantifying the association between X and Y. As λ_1 and λ_i $(i = 2, \ldots, d_n)$ all preserve certain information on the internal dependence, measures which combine both λ_1 and subsequent λ_i $(i = 2, \ldots, d_n)$ are intuitively better than those which only make use of partial information (such as BMC).

In order to obtain a better measure of independence, we need to make use of the entire spectrum of the objective matrix $\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{A}_{yx}\mathbf{A}_{xx}^{-1}\mathbf{A}_{yx}^{T}\mathbf{R}^{T}\mathbf{D}^{-\frac{1}{2}}$, instead of using only the largest one in Algorithm 2. Various procedures can be proposed via making different use of all eigenvalues to achieve better measures of independence. One such example is to sum up all the eigenvalues, which is equivalent to the trace of that matrix. Similar to the development from COCO (Gretton et al., 2004) to HSIC (Gretton et al., 2005a), this extension makes use of trace, we therefore name it by T-BMC. We show later that asymptotically T-BMC, which sums up all eigenvalues, is indeed a robust measure than BMC for independent cases, and a more sensitive measure in terms of signal to noise ratio (SNR) for dependent cases.

The procedure to calculate T-BMC between univariate random variables is summarized in Algorithm 3. Another advantage of T-BMC over BMC is that T-BMC is faster than BMC in computation, since calculating trace is computationally much easier than obtaining the largest eigenvalue, especially for a large matrix. Similar to the notations of BMC, we denote the population version of T-BMC by $\eta := \sum_{i=1}^{d_n} \lambda_i$, and its sample estimate by $\widehat{\eta} := \sum_{i=1}^{d_n} \widehat{\lambda_i}$.

Algorithm 3 Calculate T-BMC

```
• Step 1 ∼ Step 5 in Algorithm 2.
• Return trace of $\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{A}_{yx}\mathbf{A}_{xx}^{-1}\mathbf{A}_{yx}^{T}\mathbf{R}^{T}\mathbf{D}^{-\frac{1}{2}}$.
```

Theoretical Properties of T-BMC

T-BMC has at least two main advantages over BMC, summarized as follows.

First, from Theorem 4.3.1, we have that for cases where random variables X and Y are independent, the corresponding largest eigenvalue will be zero. Since the matrix we construct is semi-positive definite, all its eigenvalues are non-negative. Therefore, all eigenvalues will be zero. In certain cases, the largest eigenvalue may be falsely enlarged, due to limited sample size or presence of outliers. Adding all eigenvalues (i.e., using the trace) is more stable than using any single eigenvalue (e.g., the largest eigenvalue).

Second, under dependence cases, the asymptotic behavior of eigenvalues for fixed $d_n := d$ can be similar to that of a random covariance matrice, which is characterized by the following theorem.

Lemma 4.3.5 *(Distribution of eigenvalues, Johnson and Wichern (2007))*
For a covariance matrix Σ of a p-dimensional random variable from a normal population, if its eigenvalues are distinct and positve so that $\lambda_1 > \lambda_2 > \ldots, \lambda_p > 0$, then approximately each estimate of λ_i behaves independently and identically from Gaussian distribution, and

$$\sqrt{n}(\widehat{\lambda_i} - \lambda_i) \sim \mathcal{N}_p\left(0, 2\lambda_i^2\right).$$

By analogy, under certain conditions, the estimate for each positive eigenvalue λ_i of the matrix $\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{A}_{yx}\mathbf{A}_{xx}^{-1}\mathbf{A}_{yx}^T\mathbf{R}^T\mathbf{D}^{-\frac{1}{2}}$ behaves approximately independent, following normal distribution with variance $2\lambda_i^2/n$. We can calculate the SNR for BMC by

$$\frac{\lambda_1}{\sqrt{2\lambda_1^2/n}} = \sqrt{\frac{n}{2}}.$$

The SNR for T-BMC for fixed $d_n = d$ is

$$\sqrt{\frac{n}{2}}\frac{\lambda_1 + \ldots + \lambda_d}{\sqrt{\lambda_1^2 + \ldots + \lambda_d^2}},$$

which is greater than that for BMC.

Therefore, by using the trace, T-BMC is more sensitive than BMC in detecting dependence, which is a desired property for confirming dependence.

Consistency property of T-BMC can also be established.

Theorem 4.3.6 *(Consistency of $\widehat{\eta}$ to η) With the same $\zeta(d_n, n)$ in Lemma 4.3.3. Under Condition 1 and Condition 2, for any $c_2 > 0$,*

$$\Pr\left(|\widehat{\eta} - \eta| \geq c_2 d_n^2 n^{-2\kappa}\right) \leq \mathcal{O}\left(d_n \zeta(d_n, n)\right). \tag{4.8}$$

One generalization of T-BMC is to use weighted sum of all eigenvalues, which can be written as $\sum_{i=1}^{d_n} w_i\lambda_i$, where $(w_1, \ldots, w_{d_n})$ are the weights which sum up to 1. T-BMC is obtained (up to a factor of d_n) by choosing equal weight for all eigenvalue (that is, $w_i = 1/d_n$ for $i = 1, \ldots, d_n$). We note that while using a weighted sum of all eigenvalues is more flexible than using trace as in T-BMC, tuning parameters $(w_1, \ldots, w_{d_n})$ becomes another issue, which makes this extension more complex. Moreover, this extension forfeits the computational advantage of T-BMC (unless it is T-BMC), and computation complexity will be no less than that of BMC.

4.3.2 Multivariate Case: MBMC and T-MBMC

In multivariate cases, given random vectors $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_q)$, one can measure the marginal dependence between every single X_i and Y_j. However, marginal dependence measure may fail to capture the dependence structure. Similar to equation (4.1), we have the following problem.

$$\begin{aligned} \min_{\theta,\phi \in L_2(P)} \quad & e^2 = \mathrm{E}[\{\theta(\mathbf{Y}) - \phi(\mathbf{X})\}^2], \\ \text{subject to} \quad & \mathrm{E}\{\theta(\mathbf{Y})\} = \mathrm{E}\{\phi(\mathbf{X})\} = 0; \\ & \mathrm{E}\{\theta^2(\mathbf{Y})\} = 1. \end{aligned} \tag{4.9}$$

we can approximate functions $\theta(\mathbf{Y})$ and $\phi(\mathbf{X})$ by tensor product B-splines. For example, $\theta(\mathbf{Y})$ can be approximated by $\theta_n(\mathbf{Y}) = \alpha^T\mathbf{B}(\mathbf{Y})$ where $\mathbf{B}(\mathbf{Y}) = \mathbf{B}(Y_1)\otimes\cdots\otimes\mathbf{B}(Y_q)$[1]. With tensor product B-splines, maximum correlation can be easily extended to measure dependence in multivariate cases.

Given n samples $\mathbf{x}_u = (x_{u,1}, x_{u,2}, \ldots, x_{u,p})$ and $\mathbf{y}_u = (y_{u,1}, y_{u,2}, \ldots, y_{u,q})$ where $u = 1, 2, \ldots, n$, we summarize the measures for mutlivariate cases as in Algorithm 4. Similar to the notations of BMC and T-BMC, we name the corresponding new measures as MBMC and T-MBMC.

By tensor product B-splines, the size of $\mathbf{B}(\mathbf{Y})$ is d_n^q if the size of each $\mathbf{B}(Y_j)$ is d_n, which will yield trivial solutions (i.e., MBMC $= 1$, independent of data) when $d_n^q > n$.

[1] Tensor product: $(a_1, a_2, \ldots, a_s)^T \otimes (b_1, b_2, \ldots, b_t)^T = (a_1b_1, a_1b_2, \ldots, a_1b_t, \ldots, a_sb_1, a_sb_2, \ldots, a_sb_t)^T$

```
**Algorithm 4** Calculate MBMC/ T-MBMC
• Construct $\mathbf{B}(\mathbf{x}_u)$ and $\mathbf{B}(\mathbf{y}_u)$ by tensor product B-splines,
    $\mathbf{B}(\mathbf{x}_u) = \mathbf{B}(x_{u,1}) \otimes \cdots \otimes \mathbf{B}(x_{u,p})$,
    $\mathbf{B}(\mathbf{y}_u) = \mathbf{B}(y_{u,1}) \otimes \cdots \otimes \mathbf{B}(y_{u,q})$.
• Step 2 $\sim$ Step 5 in Algorithm 2.
• **Return** the largest eigenvalue / trace of the matrix
        $\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{A}_{yx}\mathbf{A}_{xx}^{-1}\mathbf{A}_{yx}^{T}\mathbf{R}^{T}\mathbf{D}^{-\frac{1}{2}}$.
```

This problem can be alleviated when further assumptions are imposed. For example, if we consider additive structures on the transformations of $\theta(\mathbf{Y})$ and $\phi(\mathbf{X})$ in (4.9), where

$$\theta(\mathbf{Y}) = \theta_1(Y_1) + \ldots + \theta_q(Y_q),$$
$$\phi(\mathbf{X}) = \phi_1(X_1) + \ldots + \phi_p(X_p).$$

Then, splines $\mathbf{B}(\mathbf{X})$ and $\mathbf{B}(\mathbf{Y})$ in Algorithm 4 can be constructed by combining B-spline bases for each individual variable, where

$$\mathbf{B}(\mathbf{X}) = \big(\mathbf{B}^T(X_1), \cdots, \mathbf{B}^T(X_p)\big)^T,$$
$$\mathbf{B}(\mathbf{Y}) = \big(\mathbf{B}^T(Y_1), \cdots, \mathbf{B}^T(Y_q)\big)^T.$$

Then, under additive structures of $\theta(\mathbf{Y})$ and $\phi(\mathbf{X})$, the size of $\mathbf{B}(\mathbf{X})$ and $\mathbf{B}(\mathbf{Y})$ will be reduced to pd_n and qd_n, respectively. Given samples, the algorithms for obtaining MBMC and T-MBMC under additive structures are summarized in Algorithm 5.

```
Algorithm 5 Calculate MBMC/T-MBMC, additive cases
  • Construct $\mathbf{B}(\mathbf{x}_u)$ and $\mathbf{B}(\mathbf{y}_u)$ by tensor product B-splines,
      $\mathbf{B}(\mathbf{x}_u) = \big(\mathbf{B}^T(x_{u,1}), \cdots, \mathbf{B}^T(x_{u,p})\big)^T,$
      $\mathbf{B}(\mathbf{y}_u) = \big(\mathbf{B}^T(y_{u,1}), \cdots, \mathbf{B}^T(y_{u,q})\big)^T.$
  • Step 2 $\sim$ Step 5 in Algorithm 2.
  • Return the largest eigenvalue / trace of the matrix
        $\mathbf{D}^{-\frac{1}{2}}\mathbf{R}\mathbf{A}_{yx}\mathbf{A}_{xx}^{-1}\mathbf{A}_{yx}^T\mathbf{R}^T\mathbf{D}^{-\frac{1}{2}}.$
```

4.4 Hypothesis Testing

Consider the hypothesis “variables X and Y are independent”, which implies that $\lambda_1 = 0$ from Theorem 1, two testing procedures can be proposed to address this hypothesis testing problem.

Bartlett's approximation

In canonical correlation analysis, Bartlett's approximation (Mardia et al., 1979; Lopez-Paz et al., 2013) can be used for testing of independence between random variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_q)$. Under normality assumptions and for large sample size, if $\mathbf{X}$ and $\mathbf{Y}$ are independent,

$$-\left(n - \frac{p+q+3}{2}\right) \log \prod_{i=1}^{\min\{p,q\}} (1 - \widehat{\rho_i}^2) \sim \chi^2_{pq},$$

where $\widehat{\rho_i}$ is the i-th sample canonical correlation between $\mathbf{X}$ and $\mathbf{Y}$. Generally, theoretical distributions for $\{\widehat{\lambda_i}\}_{i=1}^{d_n}$ are difficult to obtain. From equation (4.7), we have that under certain assumptions, Barrlett's approximation holds for $\widehat{\lambda_i}$. That is, if variables X and Y are independent,

$$-\left(n - \frac{2d_n+3}{2}\right) \log \prod_{i=1}^{d_n} (1 - \widehat{\lambda_i}) \sim \chi^2_{d_n^2}$$

While Bartlett's approximation provides a computationally easy procedure to test the independence, it is common that the underlying assumptions are difficult to validate, or the sample size is limited. In those cases, Bartlett's approximation test will not be efficient. A more commonly used testing procedure which relaxes the distributional assumptions can be developed.

Permutation test

Here, we summarize the procedures of permutation test for BMC in Algorithm 6. Permutation test procedures for T-BMC, MBMC and T-MBMC can be developed similarly.

We reject the hypothesis that "variables X and Y are independent" if $\widehat{\lambda_1}$ caculated from the original data exceeds the $(1-\alpha)$-th quantile of $\{\widehat{\lambda_1^b}\}_{b=1}^{B}$, where α is usually set to be 0.05 or 0.1.

Algorithm 6 Permutation test of independence for BMC

```
Compute the BMC value for original observations $\{x_i, y_i\}_{i=1}^n$ by Algorithm 1, to obtain $\widehat{\lambda}_1$.
For $b = 1, \ldots, B$
permute the data by shuffling $\{y_i\}_{i=1}^n$ to obtain $b$-th permuted data $\{x_i, \widetilde{y_i}^b\}_{i=1}^n$, and compute BMC values for $\{x_i, \widetilde{y_i}^b\}_{i=1}^n$ by Algorithm 1, to obtain $\widehat{\lambda_1^b}$.
Return $\{\widehat{\lambda_1^b}\}_{b=1}^B$.
```

4.5 Numerical Results

In this section, we demonstrate the empirical performances of two proposed measures, BMC and T-BMC, on detecting dependence over different models. *Power* of a dependence measure is defined in Lopez-Paz et al. (2013) as *the ability to discern between dependent and independent samples that share equal marginal forms*. We use the same criteria and apply the same strategy here in evaluating our proposed measures.

4.5.1 Simulation Results for BMC/T-BMC

Pearson correlation, Kendall's τ coefficient, Distance Correlation, ACE, RDC and CDC are included for comparison. For BMC and T-BMC, we only reported the best results from a selective candidate of knots choices as an indication that, with proper parameter tuning (like cross-validation procedure in subsection **??**), our proposed methods can achieve better performance than other methods. Parameters for all other measures were set to the default values under the following considerations: Pearson correlation, Kendall's τ coefficient and Distance Correlation have no tuning parameters, ACE is fairly stable to its tuning parameters (Breiman and Friedman, 1985), CDC uses ACE for calculation, and the RDC authors stated that RDC is robust against the number of random features (i.e. its tuning parameters).

Example 4.5.1 *We consider ten different types of bivariate relationships as follows,*

1. $Y = X + L\epsilon/10$;
2. $Y = 4(X - 1/2)^2 + L\epsilon/10$;
3. $Y = 80(X - 1/3)^2 - 12(X - 1/3) + L\epsilon$;
4. $Y = \sin(16\pi X) + L\epsilon/10$;
5. $Y = \sin(4\pi X) + L\epsilon/5$;
6. $Y = X^{1/4} + L\epsilon/10$;
7. $Y = (2V - 1)\sqrt{1 - (2X - 1)^2} + L\epsilon/40$;
8. $Y = \mathcal{I}\{X > 1/2\} + L\epsilon/2$;
9. $Y = X\mathcal{I}\{U > 1/2\} + (1 - X)\mathcal{I}\{U \leq 1/2\} + L\epsilon/2$;

10. $Y = \mathcal{I}\{1/4 \leq X \leq 3/4\} + L\epsilon/2.$

Here $X, U \sim Unif[0,1]$, $\epsilon \sim \mathcal{N}[0,1]$, $V \sim Bernoulli(\frac{1}{2})$, and $L \in \{1, 2, \ldots, 30\}$. The true models without error terms ($L = 0$) are depicted inside small boxes in Figure 4.1.

For each model above, we first generated 500 datasets (*positive datasets*), each contains 320 data points (n = 320). Next, we re-generated input variable randomly, and combined it with the same response variable in *positive datasets*, to obtained another 500 datasets (*negative datasets*). For each dependence measure, we obtain 500 dependence values from the *positive datasets*, and another 500 dependence values from the *negetive datasets*. Denote the 95 percent quantile of those values obtained from *negetive datasets* by m_1. In spirit of Simon and Tibshirani (2014), we have an empirical evaluation of *power* as "the proportion of those 500 values from *positive datasets* exceeding m_1". We repeated the above procedures for every $L \in \{1, 2, \ldots, 30\}$. Figure 4.1 shows results of power curves for each relationship type, as the noise level L increases. In most of the relationships, BMC and T-BMC consistently achieve higher power or the best power in detecting the dependence, especially for Quadratic, Cubic, Circle, Sinusoidal (both high- and low-frequency) types.

4.5.2 Simulation Results for MBMC/T-MBMC

To test the efficiency of MBMC and T-MBMC (with additive structure) in measing dependence for high dimensional data, we adopt the same eight experiment settings as in (Jiang and Ding, 2014).

Example 4.5.2 *Consider the following models:*

1. $y_1 = x_1x_2$, $y_2 = x_2x_3$, $y_3 = x_3x_1$

2. $y = x_2x_1 + \log(x_3^2)x_2^2 + \sin(x_1)(x_3 - 5)^2$

3. $y_1 = \log(x_1^2)x_2 + x_3$, $y_2 = \log(x_2^2)\sin(x_1) + x_1^2$, $y_3 = \log(x_3^2)x_1$

4. $y_1 = \cos(x_2(1 + x_1)x_3)$, $y_2 = \sin(6\pi x_2^2)$, $y_3 = \sin(x_2)\cos(x_3(1 + x_2))$

5. $y_1 = \cos(x_1)\cos(x_2) + x_1x_2$, $y_2 = \sin(x_2)\sin(x_3) + x_2x_3$, $y_3 = \cos(x_3)\sin(x_1) + x_1x_3$

6. $y_1 = x_1$, $y_2 = x_2^2$, $y_3 = x_3^3$

7. $y_1 = \sin(x_2)2^{x_3} + 3x_2x_1^3$, $y_2 = 4x_2\log(x_1^2) + x_1^2$, $y_3 = \sin(x_3)\log(x_1) + 4x_1^2$

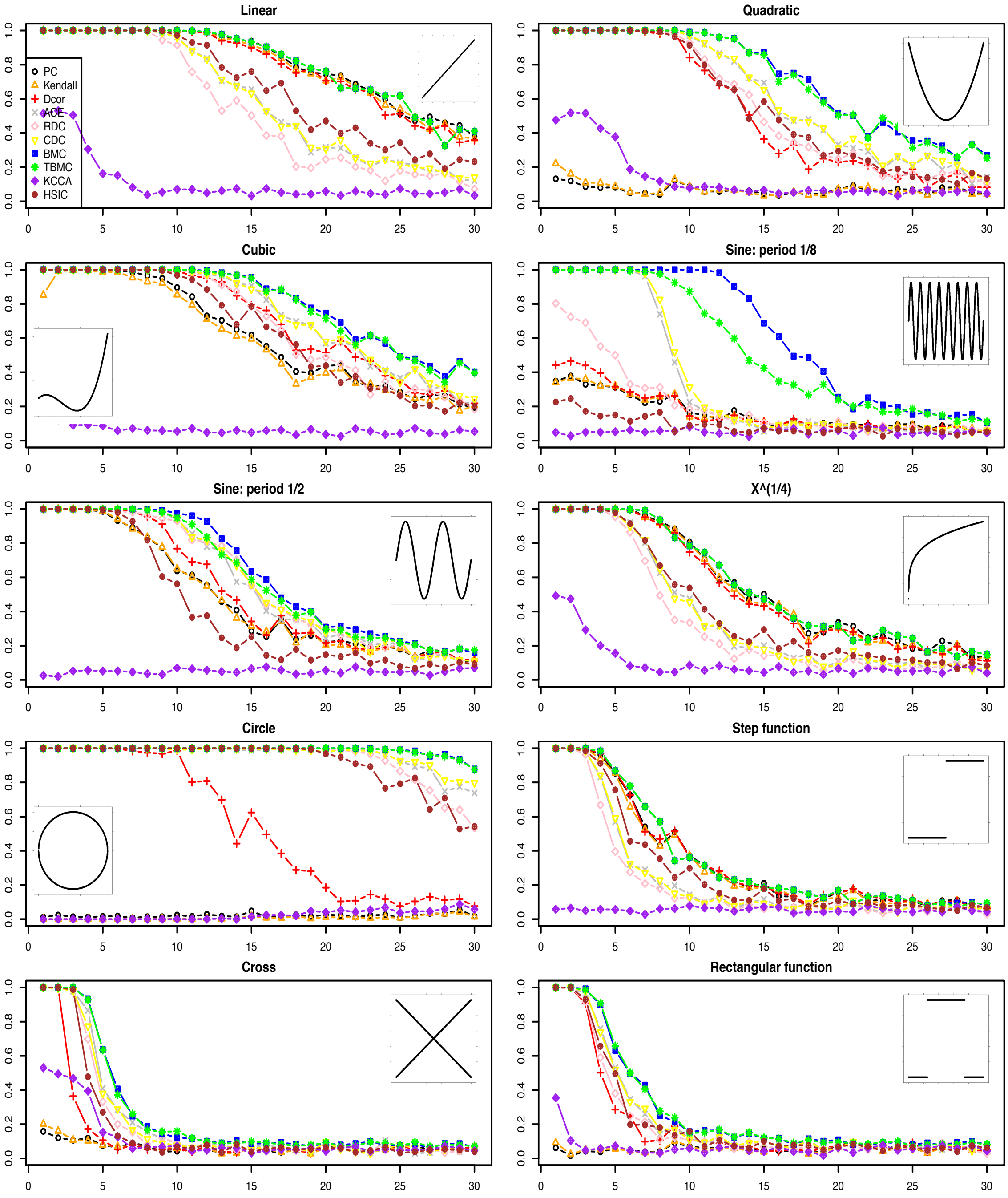

Figure 4.1. Power of different measures on detecting dependence for different bivariate relationships, as noise level increases.

8. $y_1 = 2x_1x_2 + x_1^3 \sin(x_2)$, $y_2 = \cos(x_2) + 5x_2 \log(x_1^2) + x_1^2$, $y_3 = \sin(x_2)\log(x_3) + 5x_2$

where $x_1, x_2, x_3 \sim Unif[0, 1]$, and sample size $n = 320$. We measure the dependence between multivariate variables $y = (y_1, y_2, y_3)$ and $x = (x_1, x_2, x_3)$. Power is calculated using the same procedure as stated in Section 5 in the main paper.

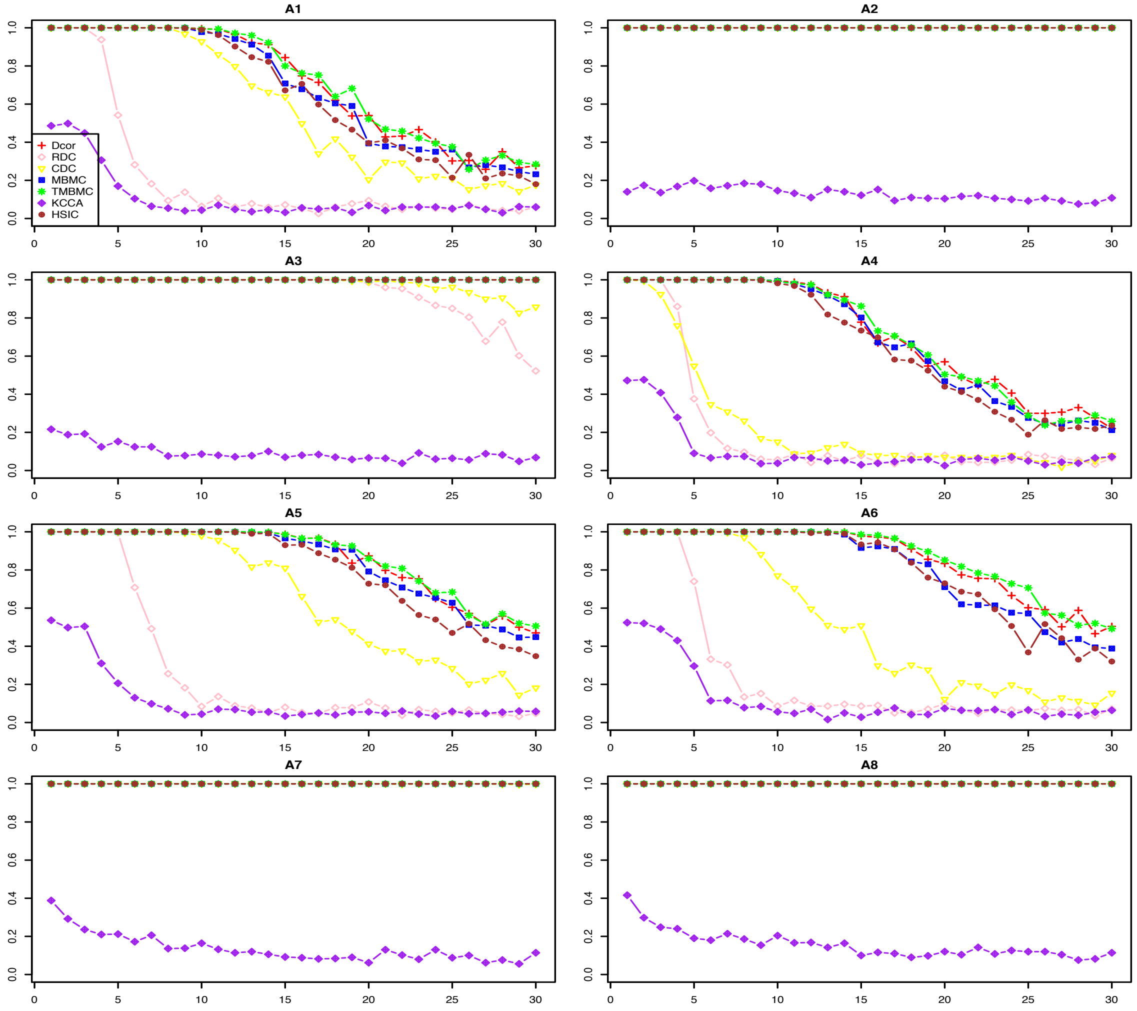

Figure 4.2. Power of different measures on detecting dependence for different multivariate relationships, as noise level increases.

4.6 Discussions

4.6.1 On Dependence Measures

We introduced four dependence measures based on B-splines and maximum correlation. For univariate random variables, BMC and T-BMC are introduced, and their asymptotic convergence rates are investigated. Multivariate counterparts (MBMC and T-MBMC, respectively) to BMC and T-BMC are also introduced, by using tensor product B-splines. Special cases for MBMC and T-MBMC under additive assumptions are discussed.

As mentioned in Section 4.3.1 when defining T-BMC, the development of T-BMC from BMC follows the same idea of developing HSIC from COCO, aiming at constructing a robust indication of dependence by making use of the full spectrum (all singular values) rather than only the largest singular value. As population versions of HSIC and COCO are both well-defined in corresponding RKHS, and population counterpart of BMC in L_2 space is the maximum correlation, we explicitly constructed the counterpart of T-BMC in general L_2 space in Section 4.3.1.

Here, we point out several interesting relations between commonly used measures. For univariate cases, the proposed dependence measure BMC has essential connections with KCCA and RDC. In fact, they all can be represented using the following form,

$$\sup_{\alpha,\beta\in\mathbf{R}^m} \rho\left(\alpha^T\mathbf{B}(Y), \beta^T\mathbf{B}(X)\right) \tag{4.10}$$

The differences among BMC, KCCA and RDC are due to different choices of structure and size for basis functions, as showed in expression (4.10). When samples are given, KCCA utilizes the basis (it can be showed that $\mathbf{B}(Y)$ and $\mathbf{B}(X)$ are just Gram matrix of the corresponding samples) of the same length with the sample size, that is, $m = n$ in (4.10). Due to this choice, regularization is needed in calculating KCCA to avoid trivial solutions. RDC, on the other hand, is flexible in choosing the basis size. However, as RDC is using random projections, structure on its basis is not well-studied. From this point of view, BMC is a better choice in making use of a both flexible and well-structured basis.

4.6.2 On Application to Sufficient Dimension Reduction

Recall that in linear dimension reduction, one aims to find a few linear combinations $\beta_1^\top \mathbf{X}, \ldots, \beta_d^\top \mathbf{X}$, so that

$$Y \perp\!\!\!\perp \mathbf{X} | \{\beta_1^\top \mathbf{X}, \ldots, \beta_d^\top \mathbf{X}\}.$$

We consider to extend the linear combinations of $\mathbf{X}$ to nonlinear cases, where we hope to find additive functional components $\mathbf{f}_1, \ldots, \mathbf{f}_d$ satisfying

$$Y \perp\!\!\!\perp \mathbf{X} | \{\mathbf{f}_1(\mathbf{X}), \ldots, \mathbf{f}_d(\mathbf{X})\},$$

where $\mathbf{f}_i(\mathbf{X}) = \sum_{j=1}^p f_{ij}((X_j)$.

Similar to the SIR procedure described in (1.10) and (1.11), which recovers the space spanned by $\beta_1, \ldots, \beta_d$, we propose to use optimal transformations to recover $\mathbf{f}_1, \ldots, \mathbf{f}_d$. Let $\mathbf{g}_1$ be the sum of optimal transformations of the predictor variables, we solve successive transformations by

$$\begin{aligned} & e^2(h_i, \mathbf{g}_i) = \min_{h,\mathbf{g}} \quad e^2(h, \mathbf{g}) \\ & \text{Cov}(h_i, h_j) = 0, \\ & \text{Cov}(\mathbf{g}_i, \mathbf{g}_j) = 0, \text{ for } j = 1, \ldots, i-1; \end{aligned} \tag{4.11}$$

where

$$\begin{aligned} \min_{h,\mathbf{g}} \qquad & e^2(h, \mathbf{g}) = \mathrm{E}\Big[\{h(Y) - \sum_{j=1}^p g_j(X_j)\}^2\Big], \\ \text{s.t.} \qquad & \mathrm{E}[h(Y)] = \mathrm{E}[g_j(X_j)] = 0; \\ & \mathrm{E}[h^2(Y)] = 1, \mathrm{E}[g_j^2(X_j)] < \infty. \end{aligned}$$

and $\mathbf{g}(\mathbf{X}) = \sum_{j=1}^p g_j((X_j)$.

It is known that the resulting directions $\{b_1, \ldots, b_d\}$ obtained by SIR in (1.10) and (1.11) may not be exactly $\beta_1, \ldots, \beta_d$, but their column spaces are equivalent. Similarly, we do not expect $\mathbf{g}_1, \ldots, \mathbf{g}_d$ from the procedure above being equal to $\mathbf{f}_1, \ldots, \mathbf{f}_d$. It will be an interesting future research to explore their relationships.

4.7 Technical Proofs

4.7.1 Proof of Theorem 4.3.1

Proof If X and Y are independent, it is known that $cov(f(Y), g(X)) = 0$ for each pair of (f, g) of bounded continious functions. Therefore, $cov(\theta_n(Y), \phi_n(X)) = 0$ for spline functions (θ_n, ϕ_n). As variances of θ_n, ϕ_n are restricted to be positive, we have $\rho(\theta_n(Y), \phi_n(X)) = 0$. Therefore, $\lambda_1 = 0$.

When $\lambda_1 = 0$, we have $\mathrm{E}(\phi_n^{*2}) = 0$ since $\lambda_1 = \mathrm{E}(\phi_n^{*2})$. According to equation (2b), $E(\phi^{*2}) = \rho^{*2}$. From Burman (1991), we have $\mathrm{E}\{(\phi^* - \phi_n^*)^2\} \leq c_1 k^{-w}$ for constant $c_1 > 0$ and $w > 1/2$. Then, $\rho^{*2} = E(\phi^{*2}) \leq 2\mathrm{E}\{(\phi^* - \phi_n^*)^2\} + 2\mathrm{E}(\phi_n^{*2}) \leq 2c_1 k^{-w}$. ■

4.7.2 Proof of Theorem 4.3.2

Proof With $\lambda_1 = \mathrm{E}(\phi_n^*)^2$, $\rho^{*2} = \mathrm{E}(\phi^{*2})$, $\mathrm{E}\{(\phi^* - \phi_n^*)^2\} \leq c_1 k^{-w}$, we have

$$\begin{aligned}
|\mathrm{E}(\phi_n^*)^2 - \mathrm{E}(\phi^{*2})| &= |\mathrm{E}(\phi_n^* - \phi^* + \phi^*)^2 - \mathrm{E}(\phi^*)^2| \\
&= |\mathrm{E}(\phi_n^* - \phi^*)^2 + 2\mathrm{E}\{(\phi_n^* - \phi^*)\phi^*\}| \\
&\leq |\mathrm{E}(\phi_n^* - \phi^*)^2 + 2\sqrt{\mathrm{E}(\phi_n^* - \phi^*)^2 \mathrm{E}(\phi^{*2})} \\
&\leq \mathrm{E}(\phi_n^* - \phi^*)^2 + 2\sqrt{\mathrm{E}(\phi_n^* - \phi^*)^2} \\
&\leq c_1 k^{-w} + 2\sqrt{c_1 k^{-w}}
\end{aligned}$$

■

4.7.3 Proof of Theorem 4.3.6

Proof Similar to the proof of Theorem 2.2.2 in Chapter 2, we can easily generalize the consistency result for each of the eigenvalue $\widehat{\lambda}_i$ to λ_i. In fact, the result in Theorem 2.2.2 applies to any of the eigenvalues, which is,

$$\Pr\left(\max_{1\leq i\leq d_n} |\widehat{\lambda}_i - \lambda_i| \geq c_2 d_n n^{-2\kappa}\right) \leq \mathcal{O}\left(\zeta(d_n, n)\right).$$

Theorem 6 follows by combining it and the fact that $|\widehat{\eta} - \eta| \leq d_n \max_{1\leq i\leq d_n} |\widehat{\lambda}_i - \lambda_i|$. ■

5. SUMMARY

In high dimensional data analysis, noises accumulated by a large number of spurious predictor variables can make the real signals difficult to be discovered, resulting in model inaccuracy and poor prediction capability. Effective variable screening and variable selection methods are important tools to reduce the size of predictor variables, which aid in efficient model building. In general, a sparse model can lead to higher prediction accuracy by reducing the number of spurious variables.

In this thesis, we first present a screening method, MC-SIS, to reduce the dimensionality from ultrahigh to relatively high dimension prior to model building. MC-SIS ranks all predictor variables according to their marginal maximum correlations with the response and selects the top predictors with relatively large maximum correlation values. It is theoretically justified that MC-SIS is a model-free sure screening procedure, which enjoys the sure screening property without imposing any specific model assumptions. Numerical experiments further show that MC-SIS can outperform other existing screening methods when their model assumptions are violated, and remain competitive when the model assumptions are satisfied. Another method, SPOT, is introduced to simultaneously select important variables and explore relationships between the response and predictor variables in high dimensional nonparametric regression analysis. SPOT combines the advantages of optimal transformations in producing the best-fitting additive models, and SICA penalty functions in selecting the important variables. SPOT can also be used for response prediction due to the monotone constraint on the response transformation. Numerical experiments demonstrate that SPOT achieves better variable selection performance and higher prediction accuracy. Therefore, it can serve as an effective tool in both variable selection and exploratory regression analysis.

Both MC-SIS and SPOT are developed under the framework of optimal transformations. MC-SIS makes use of the maximum correlation which has an equivalent form by using optimal transformations in bivariate case. SPOT is a sparse and constrained version of optimal transformations.

Many useful methodologies can be developed under the same framework of optimal transformation besides the ones proposed for variable screening and selection. In the thesis, we also consider applying optimal transformations to develop novel methods for dependence measures and nonlinear sufficient dimension reduction. In developing dependence measures, we notice that using additional transformations besides optimal transformations would provide more comprehensive understandings of dependence between the response and predictor variables. The same strategy could be potentially applied to develop meaningful tools for nonlinear sufficient dimension reduction.

Another interesting research direction is to consider shape constraints in optimal transformations, where each transformation can be restricted to certain classes of functions, such as monotone, concave/convex, linear/nonlinear, etc.

REFERENCES

REFERENCES

Akeike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory, pages 267–281. Akademinai Kiado.

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis. Wiley, New York, NY, second edition.

Balakrishnan, S., Puniyani, K., and Lafferty, J. D. (2012). Sparse additive functional and kernel cca. In Proceedings of the 29th International Conference on Machine Learning (ICML-12), pages 911–918.

Bickel, P. J. and Xu, Y. (2009). Discussion of: Brownian distance covariance. The Annals of Applied Statistics, 3(4):1266–1269.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), pages 211–252.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. The Annals of Statistics, pages 2350–2383.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple-regression and correlation. Journal of the American Statistical Association, 80(391):580–598.

Bryc, W. and Dembo, A. (2005). On the maximum correlation coefficient. Theory of Probability & Its Applications, 49(1):132–138.

Burman, P. (1991). Rates of convergence for the estimates of the optimal transformations of variables. Annals of Statistics, 19(2):702–723.

Chen, C.-H. and Li, K.-C. (1998). Can sir be as popular as multiple linear regression? Statistica Sinica, 8(2):289–316.

Chiappori, P.-A., Komunjer, I., and Kristensen, D. (2015). Nonparametric identification and estimation of transformation models. Journal of Econometrics, 188(1):22–39.

De Boor, C. (2001). A practical guide to splines, revised edition, vol. 27 of applied mathematical sciences.

Dembo, A., Kagan, A., and Shepp, L. A. (2001). Remarks on the maximum correlation coefficient. Bernoulli, 7(2):343–350.

Efromovich, S. (2007). Conditional density estimation in a regression setting. The Annals of Statistics, pages 2504–2535.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. The Annals of statistics, 32(2):407–499.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. Journal of the American Statistical Association, 106(494):544–557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society, Series B, 70(5):849–911.

Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. Journal of the American Statistical Association, 109(507):1270–1284.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. The Journal of Machine Learning Research, 10:2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. Annals of Statistics, 38(6):3567–3604.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika, 83(1):189–206.

Fan, Y., James, G. M., and Radchenko, P. (2015). Functional additive regression. The Annals of Statistics, 43(5):2296–2325.

Faouzi, E., Eddin, N., et al. (1999). Rates of convergence for spline estimates of additive principal components. Journal of Multivariate Analysis, 68(1):120–137.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. Technometrics, 35(2):109–135.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. The Annals of Applied Statistics, 1(2):302–332.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. Journal of computational and graphical statistics, 7(3):397–416.

Gebelein, H. (1941). Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. Zeitschrift für Angewandte Mathematik und Mechanik, 21(6):364–379.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a). Measuring statistical dependence with hilbert-schmidt norms. In Algorithmic learning theory, pages 63–77. Springer.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b). Kernel methods for measuring independence. The Journal of Machine Learning Research, 6:2075–2129.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Schölkopf, B., and Logothetis, N. (2004). Behaviour and convergence of the constrained covariance.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. Journal of Computational and Graphical Statistics, 18(3):533–550.

Hall, P. and Miller, H. (2011). Determining and depicting relationships among components in high-dimensional variable selection. Journal of Computational and Graphical Statistics, 20(4):988–1006.

Hastie, T. J. and Tibshirani, R. J. (1990). Generalized additive models, volume 43. CRC Press.

Heller, R., Heller, Y., and Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. Biometrika, 100(2):503–510.

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

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. Statistical science: a review journal of the Institute of Mathematical Statistics, 27(4).

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. Annals of statistics, 38(4):2282.

Jacho-Chávez, D., Lewbel, A., and Linton, O. (2010). Identification and nonparametric estimation of a transformed additively separable model. Journal of Econometrics, 156(2):392–407.

Jiang, H. and Ding, Y. (2014). Dependence measure for non-additive model. stat, 1050:14.

Johnson, R. A. and Wichern, D. W. (2007). Applied Multivariate Statistical Analysis (6th Edition). Pearson.

Katō, T. (1995). Perturbation theory for linear operators, volume 132. Springer Verlag.

Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. Biometrika, 44(1-2):289–292.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. Ann. Statist., 40(3):1846–1877.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414):316–327.

Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via distance correlation learning. Journal of the American Statistical Association, 107(499):1129–1139.

Lin, W. and Lv, J. (2013). High-dimensional sparse additive hazards regression. Journal of the American Statistical Association, 108(501):247–264.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. The Annals of Statistics, 34(5):2272–2297.

Linton, O., Sperlich, S., and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. The Annals of Statistics, pages 686–718.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Journal of Machine Learning Research, 16:559–616.

Lopez-Paz, D., Hennig, P., and Schölkopf, B. (2013). The randomized dependence coefficient. In Advances in Neural Information Processing Systems, pages 1–9.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. The Annals of Statistics, pages 3498–3528.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):141–167.

Maldonado, S. and Weber, R. (2010). Feature selection for support vector regression via kernel penalization. In Neural Networks (IJCNN), The 2010 International Joint Conference on, pages 1–7. IEEE.

Mallows, C. L. (1973). Some comments on c p. Technometrics, 15(4):661–675.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). Multivariate Analysis. Academic Press.

Meier, L., Van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. The Annals of Statistics, 37(6B):3779–3821.

Nikolova, M. (2000). Local strong homogeneity of a regularized estimator. SIAM Journal on Applied Mathematics, 61(2):633–658.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2007). Spam: Sparse additive models. In Advances in Neural Information Processing Systems, pages 1201–1208.

Redmond, M. and Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. European Journal of Operational Research, 141(3):660–678.

Rényi, A. (1959). On measures of dependence. Acta Mathematica Hungarica, 10(3):441–451.

Reshef, D., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. Science, 334(6062):1518–1524.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. Multivariate Analysis II, 25:31.

Schumaker, L. (1981). Spline functions: basic theory. Wiley, New York.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464.

Schwarz, G. et al. (1978). Estimating the dimension of a model. The annals of statistics, 6(2):461–464.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003). Regression approaches for microarray data analysis. Journal of Computational Biology, 10(6):961–980.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. Journal of the American Statistical Association, 109(507):1302–1318.

Simon, N. and Tibshirani, R. (2014). Comment on" detecting novel associations in large data sets" by reshef et al, science dec 16, 2011. arXiv preprint arXiv:1401.7645.

Song, L., Xing, E. P., and Parikh, A. P. (2011). Kernel embeddings of latent tree graphical models. In Advances in Neural Information Processing Systems, pages 2708–2716.

Speed, T. (2011). A correlation for the 21st century. Science, 334(6062):1502–1503.

Stone, C. J. (1985). Additive regression and other nonparametric models. The Annals of Statistics, pages 689–705.

Stone, C. J. et al. (1985). Additive regression and other nonparametric models. Annals of Statistics, 13(2):689–705.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., and Okanohara, D. (2010). Conditional density estimation via least-squares density ratio estimation. In International Conference on Artificial Intelligence and Statistics, pages 781–788.

Szekely, G. and Mori, T. (1985). An extremal property of rectangular distributions. Statistics & probability letters, 3(2):107–109.

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. Annals of Statistics, 35(6):2769–2794.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108.

Van der Vaart, A. and Wellner, J. (1996). Weak convergence and empirical processes: with applications to statistics. Springer.

Yin, J., Chen, X., and Xing, E. P. (2012). Group sparse additive models. In Proceedings of the 29th International Conference on Machine Learning (ICML-12), pages 871–878.

Yu, Y. (2008). On the maximal correlation coefficient. Statistics & Probability Letters, 78(9):1072–1075.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. Ann. Statist., 38(2).

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. The Journal of Machine Learning Research, 7:2541–2563.

Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. Annals of Statistics, 26(5):1760–1782.

Zhu, L., Li, L., Li, R., and Zhu, L. (2011). Model-free feature screening for ultrahigh-dimensional data. J. Am. Statist. Assoc., 106(496):1129–1139.

Zou, H. (2006). The adaptive lasso and its oracle properties. J. Am. Statist. Assoc., 101(476).

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67(2):301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. Annals of statistics, 36(4):1509.

VITA

VITA

Qiming Huang was born in Jianli, Hubei, China in 1988. He received a bachelor's degree in Mathematics from Beijing Institute of Technology, China in 2010. He earned his master's degree in Mathematical Statistics in 2010 and a doctoral degree in Statistics in 2016, both from Purdue University, Indiana. His research interests include variable selection, dimension reduction, kernel methods, computational advertising and statistical application in psychometrics.