On Sufficient Dimension Reduction via Asymmetric Least Squares

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fullfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Abdul-Nasah Soale
May 2021

Examining Committee Members:

Dr. Yuexiao Dong, Advisory Chair, Department of Statistical Science, Temple University

Dr. Cheng Yong Tang, Department of Statistical Science, Temple University

Dr. Kuang-Yao Lee, Department of Statistical Science, Temple University

Dr. Cencheng Shen, Department of Applied Economics and Statistics, University of Delaware

Committee

The committee members, hereby, certify they have read the thesis presented by Abdul-Nasah Soale, and that it is fully adequate in scope and quality as a partial requirement for the degree of Doctor of Philosophy in Statistics.

Dr. Yuexiao Dong
Temple University
Principal Advisor

Dr. Cheng Yong Tang
Temple University
Committee Member

Dr. Kuang-Yao Lee
Temple University
Committee Member

Dr. Cencheng Shen
University of Delaware
Committee Member

Declaration of Authorship

I, Abdul-Nasah Soale, declare that this thesis titled “On Sufficient Dimension Reduction Via Asymmetric Least Squares” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with my thesis advisor, I have made clear exactly what was done by others and what I have contributed myself.

Abdul-Nasah Soale

March 11, 2021

ABSTRACT

On Sufficient Dimension Reduction Via Asymmetric Least Squares

Abdul-Nasah Soale
DOCTOR OF PHILOSOPHY
Temple University, March 2021
Dr. Yuexiao Dong, Advisory Chair

Accompanying the advances in computer technology is an increase collection of high dimensional data in many scientific and social studies. Sufficient dimension reduction (SDR) is a statistical method that enable us to reduce the dimension of predictors without loss of regression information. In this dissertation, we introduce principal asymmetric least squares (PALS) as a unified framework for linear and nonlinear sufficient dimension reduction. Classical methods such as sliced inverse regression (Li, 1991) and principal support vector machines (Li, Artemiou and Li, 2011) often do not perform well in the presence of heteroscedastic error, while our proposal addresses this limitation by synthesizing different expectile levels. Through extensive numerical studies, we demonstrate the superior performance of PALS in terms of both computation time and estimation accuracy. For the asymptotic analysis of PALS for linear sufficient dimension reduction, we develop new tools to compute the derivative of an expectation of a non-Lipschitz function.

PALS is not designed to handle symmetric link function between the response and the predictors. As a remedy, we develop expectile-assisted inverse regression estimation (EA-IRE) as a unified framework for moment-based inverse regression. We propose to first estimate the expectiles through kernel expectile regression, and then carry out dimension reduction based on random projections of the regression expectiles. Several popular inverse regression methods in the literature including slice inverse regression, slice average variance estimation, and directional regression are extended under this general framework. The proposed expectile-assisted methods outperform existing moment-based dimension reduction methods in both numerical studies and an analysis of the Big Mac data.

DEDICATION

To my family and friends and the memory of Lisa Fitch.

ACKNOWLEDGEMENT

All praise and thanks are due to Allah for every moment and breath we take, and for all the blessings that can never be counted. There are many people who have made this PhD journey possible, and words cannot express my gratitude to all of them for their contributions. To my dear advisor Dr. Yuexiao Dong, I am forever in your debt for your encouragement, patience, and continuous support of my PhD study and beyond. Under your supervision, I have learned not just how to conduct scientific research in statistics, but the direction I want to go with my degree after Temple. I have learned what it means to truly be a mentor, and how to invest the necessary time, energy, and passion into seeing another person succeed. It has been a great honor and a privilege to pursue a PhD under your supervision.

I would also like to express my sincere gratitute to the members of my committee, Dr. Cheng Yong Tang and Dr. Kuang-Yao Lee for their insightful comments and questions which made my research richer. Your advice and recommendations for my job search are deeply appreciated. To Dr. Cencheng Shen, thank you very much for agreeing to be my external reader. My special thanks goes to Dr. Edoardo M. Airoldi for introducing me to other areas of research and for forwarding every good job ad. To Dr. Pallavi Chitturi, thank you very much for all the timely recommendation letters and the valuable teaching lessons. I would also like to thank Dr. Sanat K. Sarkar and the other faculty members and staff of the Department of Statistical Science who have taught and assisted me in various ways during my studies.

Dr. Scott A. Bruce at the Department of Statistics at George Mason University and my fellow graduate students deserve acknowledgment as well. A big thanks to all my friends; I am very glad I met you all. Last but not the least, I would like to thank my family: Mr & Mrs. J.M. Soale and my siblings for their love and support throughout my life. I appreciate all what you have sacrificed and continue to endure to help me get here and beyond.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The discipline of statistics broadly deals with summarizing data for easy interpretation and visualization. Thus, reducing data into a statistic, a function of the sample, is at the core of statistical analysis. For a univariate variable, a statistic such as the mean, median, or variance may provide a sufficient reduction of the data without discarding any information about the parameter that generated the data. Such a statistic is called a sufficient statistic. Sufficient dimension reduction extends this idea of sufficiency to the regression setting where multiple variables are reduced into a few sufficient variables which capture all the regression information.

1.1 Sufficient Dimension Reduction

Regression analysis is a widely used statistical modeling technique that seeks to estimate the relationship between two sets of quantities: the dependent variable(s) called the response and the independent variable(s) referred to as the predictor. The central goal of regression analysis is dimension reduction, which means reducing many variables into a few principal variables.

The most popular dimension reduction method is principal component analysis (PCA). While PCA reduces the predictor dimensionality, the reduced space does not guarantee to identify the relationship between the predictor and the response. This is so because PCA is unsupervised, as it does not incorporate any information about the response. Sufficient dimension reduction (SDR) overcomes this handicap by combining dimension reduction and the sufficiency principle. Because SDR involves the response variable, we say, it is a supervised reduction method. The reduced predictor obtained from SDR is suffcient to predict the relationship between the response and the predictor.

To illustrate, let $\boldsymbol{Y} \in \mathbb{R}^q$ be a response variable and $\mathbf{X} \in \mathbb{R}^p$ be the predictor variable, where $q \geqslant 1$ and $p \geqslant 2$. A reduction $R(\mathbf{X})$ is sufficient if $\boldsymbol{Y} \perp\!\!\!\perp \mathbf{X} \mid R(\mathbf{X})$, where the function $R(.)$ maps $\mathbf{X}$ from $\mathbb{R}^p$ to $\mathbb{R}^d$ such that $d \leqslant p$, and " $\perp\!\!\!\perp$ " means statistical independence. Trivially, this holds when $R(\mathbf{X}) = \mathbf{X}$. Moreover, $R(\mathbf{X})$ can take a linear or a nonlinear form. Details of linear and nonlinear sufficient dimension reduction are given in Chapter 2.

Before we get into the details of sufficient dimension reduction, it is important to distinguish SDR from variable selection. Unlike SDR which finds a few linear combinations of the predictor, variable selection believes that only a few of the original predictors can explain the response. Thus, variable selection discard some of the predictor variables whereas SDR uses all the variables to create new set of variables.

1.2 Asymmetric Least Squares Loss

One of the most popular SDR methods is the ordinary least squares (OLS) regression. OLS finds the basis of the central mean space by minimizing the quadratic loss function. This loss function is the most popular in the literature, and for good reasons. First, it is mathematically more flexible than other loss functions. Moreover, the variance of the estimates from the quadratic loss have very good properties.

However, the quadratic loss is not robust to outliers. Hence, in the presence of heteroscedastic error, it may not be suitable. Fortunately, we can tilt the loss function by giving unequal weights to the positive and negative inputs, resulting in a new function called the asymmetric least squares loss (ALS). Newey and Powell (1987) introduced the asymmetric least squares as a generalization of the quadratic loss. ALS is a close surrogate of the check loss for quantiles proposed by Koenker and Bassett (1978). Thus, ALS is also called expectiles. Although, ALS is very similar to quantiles, it has some advantages over quantiles, one of which is computational efficiency. For more on the advantages of ALS over quantiles, see Newey and Powell (1987).

The asymmetric least squares loss function is defined as

$$\rho_\tau(c) = \begin{cases} (1-\tau)c^2 & \text{if } c \leqslant 0, \\ \tau c^2 & \text{if } c > 0, \end{cases} \tag{1.2.1}$$

where $\tau \in (0,1)$. We illustrate the behavior of this function in Figure 1.1 below. The black solid line is the asymmetric loss for $\tau = 0.50$ which is the same as the regular quadratic loss. For $\tau = 0.30$, indicated by the red dash line, more weight is given to the negative residuals. Similarly, for $\tau = 0.70$ which is illustrated by the blue dash line, the positive inputs are weighed more. Thus, the asymmetric least squares weighs the contribution of the inputs in determining the target depending on whether it is above or below a given percentile. This makes it possible to track extreme behavior at the tails. Thus, in a regression setting, by synthesizing information across different percentile levels, we are able to obtain complete information about the conditional distribution of the response given the predictor. Further discussions about the asymmetric least squares regression are given in Chapters 2 and 3.

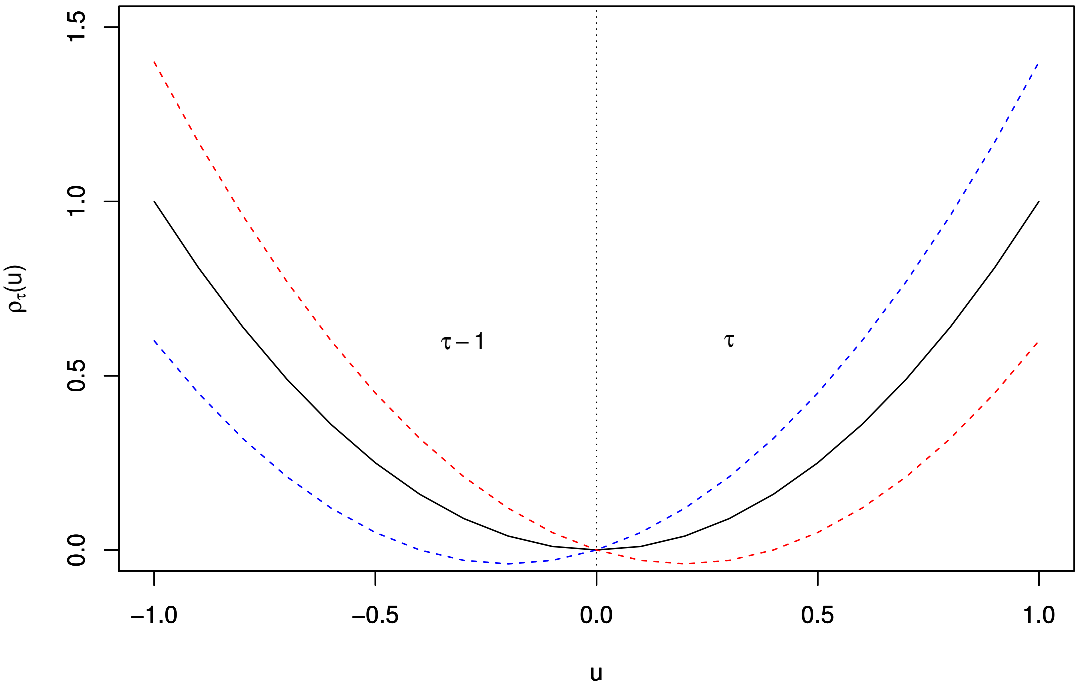

Figure 1.1: Asymmetric least squares loss functions for $\tau = 0.30, 0.50$, and 0.70.

CHAPTER 2

ON SUFFICIENT DIMENSION REDUCTION VIA PRINCIPAL ASYMMETRIC LEAST SQUARES

2.1 Introduction

For univariate response Y and multivariate predictor $\mathbf{X} \in \mathbb{R}^p$, sufficient dimension reduction (Li, 1991; Cook, 1998) aims to find $\mathbf{B} \in \mathbb{R}^{p \times d}$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^\top \mathbf{X}, \tag{2.1.1}$$

where " $\perp\!\!\!\perp$″ means statistical independence. Under (2.1.1), the conditional distribution of Y given $\mathbf{X}$ is the same as the conditional distribution of Y given $\mathbf{B}^\top\mathbf{X}$. If $\mathbf{B}$ satisfies (2.1.1), the column space of $\mathbf{B}$ is called a dimension reduction space. Under very general conditions as discussed in Yin, Li and Cook (2008), the intersection of all dimension reduction spaces is also a dimension reduction space. We refer to this minimum dimension reduction space as the central space for the regression between Y and $\mathbf{X}$, and we denote it as $\mathcal{S}_{Y|\mathbf{X}}$. The dimensionality of the central space is known as the structural dimension.

Moment-based methods such as sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), and directional regression (Li and Wang, 2007) are among the most popular sufficient dimension reduction methods. These moment-based methods are easy to implement in practice, and their extensions include sparse sufficient dimension reduction (Li, 2007), dimension reduction with matrix-valued predictors (Li, Kim and Altman, 2010), and dimension reduction for functional data (Li and Song, 2017). For an excellent review, please refer to Li (2018). More recently, Li, Artemiou and Li (2011) proposed the principal support vector machine (PSVM), which applies a modified support vector machine to find the optimal separating hyperplanes of the discretized response. It is shown that the normal vector of the separating hyperplanes can be used to recover $\mathcal{S}_{Y|\mathbf{X}}$. Extensions of PSVM include ℓq PSVM (Artemiou and Dong, 2016), principal logistic regression (Shin and Artemiou, 2017), and weighted PSVM (Shin et al., 2017).

A well-known limitation of the moment-based sufficient dimension reduction methods as well as PSVM is that they often do not perform well in the presence of heteroscedastic error. In this paper, we propose to replace the hinge loss in PSVM with the asymmetric least squares loss, and we refer to the new proposal as principal asymmetric least squares (PALS). By synthesizing different expectile levels, PALS can improve the performance of PSVM when the error is heteroscedastic. We implement the sample level estimation of PALS through quadratic programming, provide the asymptotic normality of the sample PALS estimator, and extend PALS for nonlinear sufficient dimension reduction. Wang, Shin, and Wu (2018) proposed principal quantile regression (PQR), where quantile regression was used instead of the expectile regression in our proposal. We note that the check loss from the PQR objective function is not smooth, while PALS utilizes a smooth objective function. As a result, PALS leads to more accurate estimation with much improved computational speed compared to PQR.

The rest of the paper is organized as follows. The population level development and the sample level estimation of PALS are studied in Section 2 and Section 3, respectively. Extensions to nonlinear sufficient dimension reduction are examined in Section 4. Extensive simulation studies are reported in Section 5 and we provide a real data analysis in Section 6. Section 7 concludes the paper with some discussions. All the proofs are relegated to the final section.

2.2 Population level development

Let $\boldsymbol{\mu} = E(\mathbf{X})$ and $\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{X})$. The population τ-th level objective function of PALS is

$$L_\tau(\alpha, \boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + \lambda E\left[\rho_\tau\left\{Y - \alpha - \boldsymbol{\beta}^\top(\mathbf{X} - \boldsymbol{\mu})\right\}\right]. \tag{2.2.1}$$

Here $\tau \in (0,1)$ denotes the expectile level, $\lambda > 0$ is a tuning parameter, and ρ_τ is the asymmetric least squares loss function (Newey and Powell, 1987) defined as follows

$$\rho_\tau(c) = \begin{cases} (1-\tau)c^2 & \text{if } c \leqslant 0, \\ \tau c^2 & \text{if } c > 0. \end{cases} \tag{2.2.2}$$

The asymmetric least squares loss was originally designed to recover the regression expectiles, which is known be closely related to regression quantiles (Abdous and Remillard, 1995). The objective function (2.2.1) is linked to sufficient dimension reduction through the next result.

Theorem 1. Suppose $E(\mathbf{X}|\mathbf{B}^\top\mathbf{X})$ is linear in $\mathbf{B}^\top\mathbf{X}$, where $\mathbf{B} \in \mathbb{R}^{p\times d}$ is a basis of

$\mathcal{S}_{Y|\mathbf{X}}$. Let

$$(\alpha_{0,\tau}, \boldsymbol{\beta}_{0,\tau}) = \underset{\alpha \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \ L_\tau(\alpha, \boldsymbol{\beta}).$$

Then $\boldsymbol{\beta}_{0,\tau} \in \mathcal{S}_{Y|\mathbf{X}}$.

The assumption about $E(\mathbf{X}|\mathbf{B}^\top\mathbf{X})$ is known as the linear conditional mean condition, and is common in the sufficient dimension reduction literature. As a result of Theorem 1, we have

Corollary 1. Suppose $E(\mathbf{X}|\mathbf{B}^\top\mathbf{X})$ is linear in $\mathbf{B}^\top\mathbf{X}$, where $\mathbf{B} \in \mathbb{R}^{p\times d}$ is a basis of $\mathcal{S}_{Y|\mathbf{X}}$. Let $0 < \tau_1 < \ldots < \tau_K < 1$ and

$$(\alpha_{0,\tau_k}, \boldsymbol{\beta}_{0,\tau_k}) = \underset{\alpha \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \ L_{\tau_k}(\alpha, \boldsymbol{\beta}) \text{ for } k = 1, \ldots, K.$$

Then $\operatorname{span}(\boldsymbol{\Lambda}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$, where $\boldsymbol{\Lambda} = \sum_{k=1}^K \boldsymbol{\beta}_{0,\tau_k}\boldsymbol{\beta}_{0,\tau_k}^\top$.

Here span denotes the column space. Corollary 1 suggests that we can recover the central space by optimizing the PALS objective function (2.2.1) at multiple expectile levels.

2.3 Sample level estimation

Given an i.i.d sample $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$, the sample version of (2.2.1) becomes

$$\hat{L}_\tau(\alpha, \boldsymbol{\beta}) = \boldsymbol{\beta}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + \frac{\lambda}{n}\sum_{i=1}^n \rho_\tau \left\{Y_i - \alpha - \boldsymbol{\beta}^\top\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\right\}, \tag{2.3.1}$$

where $\bar{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i$ and $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$. Denote $\tilde{\lambda} = n^{-1}\lambda$, $\mathbf{Z}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\beta}$. (2.3.1) reduces to

$$\tilde{L}_\tau(\alpha, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top\boldsymbol{\theta} + \tilde{\lambda}\sum_{i=1}^{n}\rho_\tau(Y_i - \alpha - \boldsymbol{\theta}^\top\mathbf{Z}_i). \tag{2.3.2}$$

Let $c_+ = max(0, c)$. We now introduce

$$\xi_{i+} = (Y_i - \alpha - \boldsymbol{\theta}^\top\mathbf{Z}_i)_+ \text{ and } \xi_{i-} = (\alpha + \boldsymbol{\theta}^\top\mathbf{Z}_i - Y_i)_+.$$

From the definition of ρ_τ in (3.2.1), (2.3.2) leads to the following primal optimization problem

$$(\hat{\alpha}_{0,\tau}, \hat{\boldsymbol{\theta}}_{0,\tau}) = \underset{\alpha\in\mathbb{R},\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \ \boldsymbol{\theta}^\top\boldsymbol{\theta} + \tilde{\lambda}\tau\sum_{i=1}^{n}\xi_{i+}^2 + \tilde{\lambda}(1-\tau)\sum_{i=1}^{n}\xi_{i-}^2 \tag{2.3.3}$$

subject to $\xi_{i+} \geqslant 0$, $\xi_{i-} \geqslant 0$, $\xi_{i+} \geqslant Y_i - \alpha - \boldsymbol{\theta}^\top\mathbf{Z}_i$, and $\xi_{i-} \geqslant \alpha + \boldsymbol{\theta}^\top\mathbf{Z}_i - Y_i$.

Theorem 2. Let $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbb{Z} = (\mathbf{Z}_1^\top, \ldots, \mathbf{Z}_n^\top)^\top$. The dual optimization problem of (2.3.3) is

$$\begin{aligned}(\hat{\boldsymbol{a}}_{0,\tau}, \hat{\boldsymbol{\eta}}_{0,\tau}) = \underset{\boldsymbol{a}\in\mathbb{R}^n,\boldsymbol{\eta}\in\mathbb{R}^n}{\operatorname{argmax}} \ & (\boldsymbol{a}-\boldsymbol{\eta})^\top\mathbb{Y} - \frac{1}{4}(\boldsymbol{a}-\boldsymbol{\eta})^\top\mathbb{Z}\mathbb{Z}^\top(\boldsymbol{a}-\boldsymbol{\eta}) - \frac{1}{4\tilde{\lambda}\tau}\boldsymbol{a}^\top\boldsymbol{a} \\ & - \frac{1}{4\tilde{\lambda}(1-\tau)}\boldsymbol{\eta}^\top\boldsymbol{\eta}\end{aligned} \tag{2.3.4}$$

subject to $\boldsymbol{a} \geqslant \mathbf{0}_n$, $\boldsymbol{\eta} \geqslant \mathbf{0}_n$, and $(\boldsymbol{a}-\boldsymbol{\eta})^\top\mathbf{1}_n = 0$. Furthermore, we have

$$\hat{\boldsymbol{\theta}}_{0,\tau} = \frac{1}{2}\mathbb{Z}^\top(\hat{\boldsymbol{a}}_{0,\tau} - \hat{\boldsymbol{\eta}}_{0,\tau}). \tag{2.3.5}$$

Consider expectile levels $0 < \tau_1 < \ldots < \tau_K < 1$. For a given τ_k, the dual problem (2.3.4) can be solved through standard quadratic programming to get $\hat{\boldsymbol{a}}_{0,\tau_k}$

and $\hat{\boldsymbol{\eta}}_{0,\tau_k}$. After computing $\hat{\boldsymbol{\theta}}_{0,\tau_k}$ from (2.3.5), we get the minimizer of $\hat{L}_{\tau_k}$ in (2.3.1) as $\hat{\boldsymbol{\beta}}_{0,\tau_k} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\theta}}_{0,\tau_k}$. Based on Corollary 1, we get the estimator of $\boldsymbol{\Lambda} = \sum_{k=1}^{K} \boldsymbol{\beta}_{0,\tau_k}\boldsymbol{\beta}_{0,\tau_k}^{\top}$ as $\hat{\boldsymbol{\Lambda}} = \sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_{0,\tau_k}\hat{\boldsymbol{\beta}}_{0,\tau_k}^{\top}$. Recall that d denotes the structural dimension of $\mathcal{S}_{Y|\mathbf{X}}$. The eigenvectors corresponding to the d largest eigenvalues of $\hat{\boldsymbol{\Lambda}}$ then consist the final PALS estimator to recover the central space.

We conclude this section with the asymptotic normality of Vec$(\hat{\boldsymbol{\Lambda}})$, where Vec means vectorization. The details are provided in the Appendix. In order to compute the derivative of an expectation of a non-Lipschitz function, we extend the theoretical development of PSVM (Li, Artemiou and Li, 2011). This extension is necessary as PSVM deals with discretized response while PALS applies to the continuous response without discretization.

Theorem 3. Suppose the regularity conditions in Theorem 4 and Theorem 5 from the Appendix are satisfied. Then we have

$$\sqrt{n}\left\{\mathrm{Vec}(\boldsymbol{\Lambda}) - \mathrm{Vec}(\hat{\boldsymbol{\Lambda}})\right\} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Omega})$$

as $n \to \infty$, where "$\xrightarrow{D}$" means converge in distribution and $\boldsymbol{\Omega}$ is specified in the Appendix.

2.4 Nonlinear sufficient dimension reduction

Suppose $\boldsymbol{\varphi} : \mathbb{R}^p \mapsto \mathbb{R}^d$ with $d < q$ are nonlinear functions satisfying

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\varphi}(\mathbf{X}), \tag{2.4.1}$$

where $\boldsymbol{\varphi}(\mathbf{X}) = \{\varphi_1(\mathbf{X}), \ldots, \varphi_d(\mathbf{X})\}$. Then the conditional distribution of Y given $\mathbf{X}$ is the same as the conditional distribution of Y given $\boldsymbol{\varphi}(\mathbf{X})$, and identifying $\boldsymbol{\varphi}(\mathbf{X})$ is known as nonlinear sufficient dimension reduction. Let $\mathcal{H}$ be a reproducing kernel

Hilbert space of the functions of $\mathbf{X}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Let $\Sigma : \mathcal{H} \mapsto \mathcal{H}$ be the covariance operator such that $\langle f_1, \Sigma f_2\rangle_{\mathcal{H}} = \mathrm{Cov}\{f_1(\mathbf{X}), f_2(\mathbf{X})\}$ for any $f_1, f_2 \in \mathcal{H}$. Consider objective function

$$\Pi_\tau(\alpha, \varphi) = \langle\varphi, \Sigma\varphi\rangle_{\mathcal{H}} + \lambda E[\rho_\tau\{Y - \alpha - \varphi(\mathbf{X})\}]. \tag{2.4.2}$$

Compared with (2.2.1), we see that $\Pi_\tau(\alpha, \varphi)$ is a generalization of $L_\tau(\alpha, \boldsymbol{\beta})$ with the matrix $\boldsymbol{\Sigma}$ replaced by the operator Σ, the linear function $\boldsymbol{\beta}^\top\mathbf{X}$ replaced by the nonlinear function $\varphi(\mathbf{X})$, and the inner product in $\mathbb{R}^p$ replaced by the inner product in $\mathcal{H}$. Let $(\alpha_{0,\tau}, \varphi_{0,\tau})$ be the minimizer of $\Pi_\tau(\alpha, \varphi)$ over $\alpha \in \mathbb{R}$ and $\varphi \in \mathcal{H}$. Under proper conditions, it can be shown that $\varphi_{0,\tau}$ is a function of $\varphi_1, \ldots, \varphi_d$. See, for example, Theorem 2 of Li, Artemiou and Li (2011).

Based on an i.i.d. sample $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$, we now describe the implementation of nonlinear dimension reduction through PALS. Suppose $\mathcal{H}$ can be spanned by $\{\psi_1, \ldots, \psi_m\}$. Then any function $\varphi \in \mathcal{H}$ becomes $\varphi(\mathbf{X}) = \boldsymbol{\gamma}^\top\boldsymbol{\psi}(\mathbf{X})$, where $\boldsymbol{\gamma} \in \mathbb{R}^m$ and $\boldsymbol{\psi}(\mathbf{X}) = \{\psi_1(\mathbf{X}), \ldots, \psi_m(\mathbf{X})\}^\top$. The sample version of (2.4.2) thus becomes

$$\hat{\Pi}_\tau(\alpha, \boldsymbol{\gamma}) = \frac{1}{n}\boldsymbol{\gamma}^\top\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\boldsymbol{\gamma} + \frac{\lambda}{n}\sum_{i=1}^{n}\rho_\tau\{Y_i - \alpha - \boldsymbol{\gamma}^\top\boldsymbol{\psi}(\mathbf{X}_i)\}, \tag{2.4.3}$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{n\times m}$, and the ith row of $\boldsymbol{\Psi}$ is $\boldsymbol{\psi}^\top(\mathbf{X}_i)$. $\hat{\Pi}_\tau(\alpha, \boldsymbol{\gamma})$ has the same form as $\hat{L}_\tau(\alpha, \boldsymbol{\beta})$ in (2.3.1), and can be minimized in a similar fashion. Denote the minimizer of $\hat{\Pi}_\tau(\alpha, \boldsymbol{\gamma})$ as $(\hat{\alpha}_{0,\tau}, \hat{\boldsymbol{\gamma}}_{0,\tau})$. We then estimate $\varphi_{0,\tau}(\mathbf{X})$ by $\hat{\varphi}_{0,\tau}(\mathbf{X}) = \hat{\boldsymbol{\gamma}}_{0,\tau}^\top\boldsymbol{\psi}(\mathbf{X})$. To synthesize multiple expectile levels, consider expectile levels $0 < \tau_1 < \ldots < \tau_K < 1$. For a given τ_k, we get $\hat{\boldsymbol{\gamma}}_{0,\tau_k}$ from minimizing $\hat{\Pi}_{\tau_k}(\alpha, \boldsymbol{\gamma})$. Denote $\hat{\boldsymbol{\Gamma}} = \sum_{k=1}^{K}\hat{\boldsymbol{\gamma}}_{0,\tau_k}\hat{\boldsymbol{\gamma}}_{0,\tau_k}^\top$ with d leading eigenvectors as $\hat{\boldsymbol{\nu}}_1, \ldots, \hat{\boldsymbol{\nu}}_d$. The final estimator of $\boldsymbol{\varphi}(\mathbf{X})$ in (2.4.1) is $\{\hat{\boldsymbol{\nu}}_1^\top\boldsymbol{\psi}(\mathbf{X}), \ldots, \hat{\boldsymbol{\nu}}_d^\top\boldsymbol{\psi}(\mathbf{X})\}$.

It remains to choose a proper basis $\{\psi_1, \ldots, \psi_m\}$ for $\mathcal{H}$. Define kernel matrix

$\mathbf{K}_n \in \mathbb{R}^{n\times n}$, with the element in the ith row and jth column as

$$\kappa(\mathbf{X}_i, \mathbf{X}_j) = \exp(-r\|\mathbf{X}_i - \mathbf{X}_j\|^2). \tag{2.4.4}$$

Here r is a tuning parameter and $\|\cdot\|$ denotes the Euclidean norm. Define $\mathbf{Q}_n = \mathbf{I}_n - \mathbf{J}_n/n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{J}_n$ is the $n \times n$ matrix whose entries are 1. For $j = 1, \dots, m$, let $\boldsymbol{w}_j = (w_{j1}, \dots, w_{jn})^\top$ be the eigenvector corresponding to the j-th largest eigenvalue of $\mathbf{Q}_n\mathbf{K}_n\mathbf{Q}_n$. Then $\psi_j(\mathbf{X}) = \sum_{\ell=1}^n \kappa(\mathbf{X}, \mathbf{X}_\ell)w_{j\ell}$ following Proposition 2 of Li, Artemiou and Li (2011). In our simulations, we choose $m = n/2$, and use the sample version of $E^{-1/2}(\|\mathbf{X} - \mathbf{X}'\|)$ for r in (3.3.1), where $\mathbf{X}$ and $\mathbf{X}'$ are independent $N(\mathbf{0}, \mathbf{I_p})$.

2.5 Simulation studies

2.5.1 Linear sufficient dimension reduction

We evaluate the performance of PALS for linear sufficient dimension reduction in this section. The following models are considered:

$$\begin{aligned}
\text{I: } Y &= \frac{X_1}{0.5 + (X_2 + 1.5)^2} + \varepsilon;\\
\text{II: } Y &= 3\sin\{0.25(X_1 + X_2)\} + 3\sin\{0.25(X_3 + X_4)\} + \varepsilon;\\
\text{III: } Y &= X_1 + 0.5\big(e^{0.15X_2}\big)\varepsilon,
\end{aligned}$$

where $\varepsilon \sim N(0, 1)$ and ε is independent of $\mathbf{X} = (X_1, \dots, X_p)^\top$. The distribution of $\mathbf{X}$ will be specified later. Let $\boldsymbol{\beta}_1 = (1, 0, \dots, 0)^\top$, $\boldsymbol{\beta}_2 = (0, 1, 0, \dots, 0)^\top$, $\boldsymbol{\beta}_3 = (1, 1, 0, \dots, 0)^\top$, and $\boldsymbol{\beta}_4 = (0, 0, 1, 1, 0, \dots, 0)^\top$. Denote $\mathbf{B}$ as the basis of the central space $\mathcal{S}_{Y|\mathbf{X}}$. Then $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ for models I and III, while $\mathbf{B} = (\boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$ for model II.

We compare PALS with five existing methods in the literature: SIR, SAVE, directional regression (DR), PSVM, and PQR. The number of slices for SIR is set as 10, and we use 4 slices for SAVE and DR. Note that SIR is generally not sensitive to the choice of slice numbers, while SAVE and DR work better with fewer slices. For PSVM, the number of dividing points is set as 9, as Li, Artemiou and Li (2011) recommend a larger number is preferable. For a given set of dividing points, two ways to dichotomize the response are considered in Li, Artemiou and Li (2011), "left versus right" (LVR) and "one versus another". We adopt the LVR scheme in our simulations. For PQR, we follow Wang, Shin and Wu (2018) and set the number of quantile levels to be 9, which leads to 10 slices. For PALS, we set $\tau_k = k/10$ for $k = 1, \dots, 9$. To evaluate the performance of each estimator $\hat{\mathbf{B}}$, we report

$$\Delta = \|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\hat{\mathbf{B}}}\|_F, \tag{2.5.1}$$

where $\mathbf{P}_{\mathbf{A}}$ denotes the orthogonal projection onto $\text{span}(\mathbf{A})$, and $\|\cdot\|_F$ is the matrix Frobenius norm. Smaller Δ value means more accurate estimation.

For the choice of the tuning parameter λ, PSVM, PQR and PALS seem to be not overly sensitive. We try $\lambda = 0.1, 1, 10, 100$, and report the best results that a fixed λ can achieve. In addition, we propose a variable λ scheme for PALS so that one can use different λ values across repetitions. Specifically, denote $\hat{\mathbf{B}}_\lambda$ as the PALS estimator for a specific λ. We choose λ such that the squared sample distance correlation (Székely, Rizzo and Bakirov, 2007) between Y and $\hat{\mathbf{B}}_\lambda^\top \mathbf{X}$ is maximized. We refer to this method as DC-PALS.

First, we set $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma})$, where the element in the ith row and jth column of $\mathbf{\Sigma}$ is $\sigma_{i,j} = 0.5^{|i-j|}$ for $i, j = 1, \dots, p$. We fix $n = 100$, and consider $p = 10, 15, 20$. The results based on 100 repetitions are summarized in Table 1. We report the average of Δ in (2.5.1) and include its standard error in the parenthesis. We see that PALS with fixed λ leads to the best result across all three models. PSVM is not as good as

model	p	SIR	SAVE	DR	PSVM	PQR	PALS	DC-PALS
I	10	1.552	1.779	1.702	1.482	1.530	1.424	1.454
		(0.016)	(0.013)	(0.017)	(0.019)	(0.018)	(0.020)	(0.018)
	15	1.698	1.854	1.800	1.643	1.643	1.599	1.606
		(0.013)	(0.010)	(0.011)	(0.011)	(0.012)	(0.014)	(0.014)
	20	1.770	1.914	1.880	1.712	1.722	1.672	1.681
		(0.010)	(0.006)	(0.008)	(0.011)	(0.010)	(0.012)	(0.011)
II	10	1.427	1.617	1.445	1.439	1.424	1.410	1.424
		(0.006)	(0.014)	(0.007)	(0.005)	(0.007)	(0.008)	(0.006)
	15	1.494	1.916	1.526	1.501	1.480	1.467	1.470
		(0.005)	(0.007)	(0.007)	(0.005)	(0.005)	(0.005)	(0.005)
	20	1.521	1.947	1.584	1.538	1.511	1.505	1.503
		(0.005)	(0.004)	(0.007)	(0.006)	(0.005)	(0.005)	(0.006)
III	10	1.331	1.487	1.383	1.335	1.306	1.266	1.299
		(0.013)	(0.011)	(0.009)	(0.013)	(0.016)	(0.016)	(0.017)
	15	1.413	1.843	1.429	1.420	1.360	1.331	1.378
		(0.008)	(0.012)	(0.007)	(0.008)	(0.011)	(0.014)	(0.009)
	20	1.452	1.924	1.485	1.458	1.416	1.408	1.419
		(0.006)	(0.006)	(0.006)	(0.007)	(0.008)	(0.008)	(0.007)

Table 2.1: Results for linear sufficient dimension reduction with different p. The average of Δ in (2.5.1) and its standard error (in parenthesis) are reported based on 100 repetitions.

classical method such as SIR in model III, where heteroscedasticity is present. PQR is very competitive in models II and III, but does not perform as well as PSVM and PALS in model I. Furthermore, DC-PALS with variable λ has the second best overall performance, and it is only slightly worse than PALS with fixed λ. As p increases, all methods deteriorate, while PALS and DC-PALS maintain their advantage over the other methods.

Next, we fix $n = 100$, $p = 10$, and consider three cases for the distribution of $\mathbf{X}$: case (i), $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I_p})$; case (ii), $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\sigma_{i,j} = 0.5^{|i-j|}$; and case (iii), $X_j \sim \text{Uniform}(-1, 1)$, $j = 1, \dots, p$, where the components of $\mathbf{X}$ are independent. The linear conditional mean assumption holds for cases (i) and (ii), and is no longer satisfied for case (iii). The results based on 100 repetitions are summarized in Table 2. Compared to PALS, PQR does not work as well for model I, and PSVM is significantly worse for model III when $\mathbf{X}$ is normal. For cases (i) and (ii), all the estimators become worse when the correlation between the normal predictors increase. For cases (i) and

model	case	SIR	SAVE	DR	PSVM	PQR	PALS	DC-PALS
I	(i)	1.480	1.751	1.642	1.370	1.433	1.283	1.316
		(0.019)	(0.014)	(0.017)	(0.023)	(0.018)	(0.021)	(0.022)
	(ii)	1.552	1.779	1.702	1.482	1.530	1.424	1.454
		(0.016)	(0.013)	(0.017)	(0.019)	(0.018)	(0.020)	(0.018)
	(iii)	1.686	1.759	1.739	1.627	1.623	1.613	1.620
		(0.015)	(0.013)	(0.014)	(0.015)	(0.013)	(0.014)	(0.013)
II	(i)	1.383	1.565	1.364	1.361	1.358	1.350	1.357
		(0.009)	(0.019)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
	(ii)	1.427	1.617	1.445	1.439	1.424	1.410	1.424
		(0.006)	(0.014)	(0.007)	(0.005)	(0.007)	(0.008)	(0.006)
	(iii)	1.404	1.747	1.447	1.402	1.400	1.394	1.393
		(0.012)	(0.016)	(0.014)	(0.013)	(0.010)	(0.012)	(0.012)
III	(i)	1.355	1.451	1.360	1.330	1.261	1.203	1.264
		(0.010)	(0.014)	(0.011)	(0.012)	(0.016)	(0.017)	(0.016)
	(ii)	1.331	1.487	1.383	1.335	1.306	1.266	1.299
		(0.013)	(0.011)	(0.009)	(0.013)	(0.016)	(0.016)	(0.017)
	(iii)	1.384	1.581	1.391	1.365	1.358	1.355	1.344
		(0.009)	(0.016)	(0.008)	(0.012)	(0.010)	(0.012)	(0.012)

Table 2.2: Results for linear sufficient dimension reduction with different predictor distribution. The average of Δ in (2.5.1) and its standard error (in parenthesis) are reported based on 100 repetitions.

(iii) with uncorrelated predictors, we see that all the methods perform worse when the linear conditional mean assumption is violated. PALS and DC-PALS again have the best overall performances.

Last but not least, we list the computation time of 100 repetitions in Table 3 for PSVM, PQR and PALS when we fix $\lambda = 1$ and $n = 100$. We only report the results for model III. The other two models lead to similar results and are omitted. We see that the computation time generally increases when p increases, although the increase does not seem to be significant. The predictor distribution does not seem to affect the computation time. PSVM costs the least computation time among all three methods. Although not as fast as PSVM, PALS is almost four times faster than PQR across all settings.

model	case	p	PSVM	PQR	PALS
III	(i)	10	3.98	20.91	5.46
		15	4.34	21.58	5.66
		20	4.81	20.79	5.91
	(ii)	10	3.86	20.88	5.57
		15	4.36	21.25	5.62
		20	4.85	21.57	5.88
	(iii)	10	3.80	20.89	5.40
		15	4.34	21.10	5.57
		20	4.84	22.09	5.70

Table 2.3: Computation time in seconds for 100 repetitions with $\lambda = 1$.

2.5.2 Nonlinear sufficient dimension reduction

For nonlinear sufficient dimension reduction, we consider the following models:

$$\begin{aligned} &\text{IV: } Y = \sqrt{\varphi_1(\mathbf{X})}\log\left\{\sqrt{\varphi_1(\mathbf{X})}\right\} + 0.5\varepsilon; \\ &\text{V: } Y = \varphi_1^2(\mathbf{X}) + 0.5\varphi_2(\mathbf{X})\varepsilon, \end{aligned}$$

where $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I_p})$, $\varphi_1(\mathbf{X}) = \sqrt{X_1^2 + X_2^2}$, $\varphi_2(\mathbf{X}) = \sin(X_2)$, $\varepsilon \sim N(0, 0.2)$, and ε is independent of $\mathbf{X}$. Denote $\boldsymbol{\varphi}(\mathbf{X})$ as the basis for nonlinear sufficient dimension reduction such that $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\varphi}(\mathbf{X})$. Then $\boldsymbol{\varphi}(\mathbf{X}) = \varphi_1(\mathbf{X})$ for model IV, and $\boldsymbol{\varphi}(\mathbf{X}) = \{\varphi_1(\mathbf{X}), \varphi_2(\mathbf{X})\}$ for model V.

We denote our proposal in Section 4 as kernel PALS (kPALS), and we compare it with kernel SIR (kSIR) (Wu, 2008), kernel PSVM (kPSVM) (Li, Artemiou and Li, 2011), and kernel PQR (kPQR) (Wang, Shin and Wu, 2018). For estimator $\hat{\boldsymbol{\varphi}}(\mathbf{X})$, we measure its performance by the squared sample distance correlation between $\boldsymbol{\varphi}(\mathbf{X})$ and $\hat{\boldsymbol{\varphi}}(\mathbf{X})$ as

$$\Upsilon = \text{dCor}^2\left\{\boldsymbol{\varphi}(\mathbf{X}), \hat{\boldsymbol{\varphi}}(\mathbf{X})\right\}. \tag{2.5.2}$$

model	n	kSIR	kPSVM	kPQR	kPALS	DC-kPALS
IV	100	0.523 (0.016)	0.742 (0.004)	0.750 (0.004)	0.750 (0.004)	0.750 (0.004)
	150	0.616 (0.014)	0.749 (0.003)	0.762 (0.003)	0.763 (0.003)	0.763 (0.003)
	200	0.678 (0.010)	0.756 (0.003)	0.770 (0.003)	0.772 (0.003)	0.772 (0.003)
V	100	0.496 (0.008)	0.578 (0.003)	0.583 (0.004)	0.605 (0.004)	0.599 (0.003)
	150	0.537 (0.005)	0.580 (0.003)	0.581 (0.003)	0.606 (0.003)	0.602 (0.003)
	200	0.554 (0.004)	0.579 (0.002)	0.582 (0.003)	0.605 (0.002)	0.598 (0.002)

Table 2.4: Results for nonlinear sufficient dimension reduction with different n. The average of Υ in (2.5.2) and its standard error (in parenthesis) are reported based on 100 repetitions.

Larger values of Υ mean better estimation. Similar to PSVM, PQR and PALS for linear sufficient dimension reduction, their kernel counterparts require a choice of λ. See, for example, λ for kPALS in (2.4.3). For $\lambda = 0.1, 1, 10, 100$, we report the results based on the best λ. Parallel to DC-PALS, we also include DC-kPALS, where λ is chosen such that the squared sample distance correlation between $\hat{\boldsymbol{\varphi}}_\lambda(\mathbf{X})$ and Y is maximized. We fix $p = 10$ and set $n = 100, 150, 200$. From Table 4, we see that kPALS has the best performance, and DC-kPALS is a close second. All methods improve as n increases for model IV, and only kSIR improves as n increases for model V. Together with previous simulation studies, we conclude that distance correlation can be a useful tool to select λ for PALS in both linear and nonlinear sufficient dimension reduction.

2.6 Analysis of the Boston housing data

We consider Boston housing data for the real data analysis. The data is originally studied in Harrison and Rubinfeld (1978). There are a total of 506 observations with

13 attribute variables. After removing the categorical variable and excluding the cases where the census tract bounds the Charles river, we end up with 12 predictors and 471 observations. The response (Y) is the median value of owner-occupied homes in each census tract. A complete list of the predictors ($\mathbf{X}$) is given in Table 2.5.

variable	description
medv	median value of owner-occupied homes in \$1000's
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances of five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per \$10,000
ptratio	pupil-teacher ratio by town
b	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
lstat	% lower status of the population

Table 2.5: *Boston housing data variables*

As suggested by Wang, Shin and Wu (2018), we set the structural dimension to be $d = 1$ and denote the estimator as $\hat{\boldsymbol{\beta}}$. We apply PSVM, PQR and PALS to this data set, and report the squared sample distance correlation between Y and $\hat{\boldsymbol{\beta}}^\top\mathbf{X}$ for different λ. From Table 5, we see that PQR and PALS perform similarly, and both are better than PSVM. Furthermore, PQR and PALS are less sensitive to the choice of λ than PSVM.

	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$	$\lambda = 100$
PSVM	0.831	0.812	0.721	0.711
PQR	0.866	0.864	0.864	0.864
PALS	0.863	0.863	0.863	0.864

Table 2.6: The squared sample distance correlation between Y and $\hat{\boldsymbol{\beta}}^\top\mathbf{X}$ for the Boston housing data.

The scatter plots of Y against the sufficient predictor $\hat{\boldsymbol{\beta}}_1^\top \mathbf{X}$ for PALS, PQR, and SIR is given in Figure 3.2.

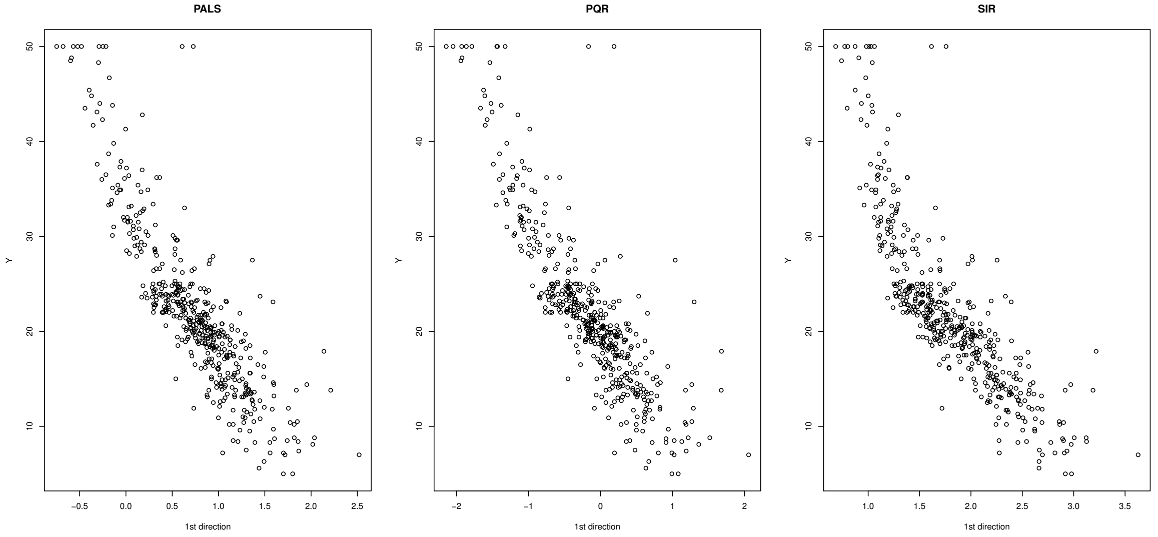

Figure 2.1: Scatter plot of response and first sufficient predictors for PALS, PQR, and SIR.

2.7 Discussion

We propose PALS for linear and nonlinear sufficient dimension reduction in this paper. Our proposed method is very competitive with existing methods in the literature. On one hand, our proposal enjoys better estimation accuracy than SIR and PSVM, especially in the presence of heteroscedasticity. On the other hand, our proposal is computationally more efficient compared to PQR. Unlike PSVM where the response is dichotomized, both PQR and PALS deal with continuous response directly. We develop new tools for the asymptotic analysis of PALS. Specifically, Lemma 3 of Li, Artemiou and Li (2011) provides a tool to compute the derivative of an expectation of a non-Lipschitz function, and Theorem 3 of Wang, Shin and Wu (2018) applied this Lemma directly without considering the continuous support of the response in PQR. This limitation is addressed in Lemma 1 and Theorem 5 of the Appendix, where

Lemma 3 of Li, Artemiou and Li (2011) is adapted for continuous response.

We consider a fixed set of expectile levels in this paper. Although our experience indicates that the performance of PALS is not sensitive to the choice of expectile levels, choosing an optimal set of expectile levels is worth further investigation. Kim, Wu and Shin (2019) develop quantile-slicing for sufficient dimension reduction, and expectile-slicing for sufficient dimension reduction may be an interesting research direction.

2.8 Proofs

Proof of Theorem 1. We assume without loss of generality that $E(\mathbf{X}) = \mathbf{0}$. Note that $\mathrm{Var}(\boldsymbol{\beta}^\top \mathbf{X}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$. Then (2.2.1) becomes

$$L_\tau(\alpha, \boldsymbol{\beta}) = \mathrm{Var}(\boldsymbol{\beta}^\top \mathbf{X}) + \lambda E\{\rho_\tau(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X})\}. \tag{2.8.1}$$

The first term on the right hand side of (2.8.1) satisfies

$$\mathrm{Var}(\boldsymbol{\beta}^\top \mathbf{X}) \geqslant \mathrm{Var}\{E(\boldsymbol{\beta}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{X})\}. \tag{2.8.2}$$

The second term on the right hand side of (2.8.1) satisfies

$$\begin{aligned} E\{\rho_\tau(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X})\} &= E[E\{\rho_\tau(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X}) | \mathbf{B}^\top \mathbf{X}, Y\}] \\ &\geqslant E[\rho_\tau\{E(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X}) | \mathbf{B}^\top \mathbf{X}, Y\}] \\ &= E[\rho_\tau\{Y - \alpha - E(\boldsymbol{\beta}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{X})\}], \end{aligned} \tag{2.8.3}$$

where the inequlality is due to the convexity of ρ_τ, and the last equality is due to the conditional independence (2.1.1). The assumption that $E(\mathbf{X}|\mathbf{B}^\top\mathbf{X})$ is linear in $\mathbf{B}^\top\mathbf{X}$

implies that

$$E(\mathbf{X}|\mathbf{B}^\top\mathbf{X}) = \boldsymbol{\Sigma}(\mathbf{B}^\top\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{X}. \tag{2.8.4}$$

(2.8.1), (2.8.2), (2.8.3) and (2.8.4) together imply that

$$L_\tau(\alpha, \boldsymbol{\beta}) \geqslant L_\tau(\alpha, \tilde{\boldsymbol{\beta}}) \text{ with } \tilde{\boldsymbol{\beta}} = \mathbf{B}(\mathbf{B}^\top\boldsymbol{\Sigma}\mathbf{B})^{-1}\boldsymbol{\Sigma}\boldsymbol{\beta}.$$

Thus the minimizer $\boldsymbol{\beta}_{0,\tau}$ must satisfy $\boldsymbol{\beta}_{0,\tau} = \mathbf{B}(\mathbf{B}^\top\boldsymbol{\Sigma}\mathbf{B})^{-1}\boldsymbol{\Sigma}\boldsymbol{\beta}_{0,\tau} \in \text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{X}}$. □

Proof of Corollary 1. The proof follows directly from Theorem 1 and is omitted. □

Proof of Theorem 2. Denote $\boldsymbol{\xi}_+ = (\xi_{1+}, \ldots, \xi_{n+})^\top$, $\boldsymbol{\xi}_- = (\xi_{1-}, \ldots, \xi_{n-})^\top$, $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$, $\boldsymbol{v} = (v_1, \ldots, v_n)^\top$, $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$. Denote $L^*(\alpha, \boldsymbol{\theta}, \boldsymbol{\xi}_+, \boldsymbol{\xi}_-, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{a}, \boldsymbol{\eta})$ as the Lagrangian of the primal optimization problem (2.3.3) and abbreviate it as L^*. Then we have

$$\begin{aligned} L^* = \ & \boldsymbol{\theta}^\top\boldsymbol{\theta} + \tilde{\lambda}\tau\sum_{i=1}^n \xi_{i+}^2 + \tilde{\lambda}(1-\tau)\sum_{i=1}^n \xi_{i-}^2 - \sum_{i=1}^n u_i\xi_{i+} - \sum_{i=1}^n v_i\xi_{i-} \\ & + \sum_{i=1}^n a_i(Y_i - \alpha - \boldsymbol{\theta}^\top\mathbf{Z}_i - \xi_{i+}) + \sum_{i=1}^n \eta_i(-Y_i + \alpha + \boldsymbol{\theta}^\top\mathbf{Z}_i - \xi_{i-}), \end{aligned} \tag{2.8.5}$$

where $u_i \geqslant 0$, $v_i \geqslant 0$, $a_i \geqslant 0$, and $\eta_i \geqslant 0$ for all i. Take partial derivatives of (2.8.5) and set them to be zero. We get

$$\begin{cases} \partial L^*/\partial\boldsymbol{\theta} = 2\boldsymbol{\theta} - \sum_{i=1}^n (a_i - \eta_i)\mathbf{Z}_i = \mathbf{0} \\ \partial L^*/\partial\alpha = \sum_{i=1}^n (\eta_i - a_i) = 0 \\ \partial L^*/\partial\xi_{i+} = 2\tilde{\lambda}\tau\xi_{i+} - u_i - a_i = 0 \\ \partial L^*/\partial\xi_{i-} = 2\tilde{\lambda}(1-\tau)\xi_{i-} - v_i - \eta_i = 0 \end{cases} \tag{2.8.6}$$

Assume $u_i > 0$ for a particular i. The Karush Kuhn Tucker (KKT) conditions state that $u_i\xi_{i+} = 0$ for all i. Then we must have $\xi_{i+} = 0$ from KKT. On the other hand, we have $\xi_{i+} = (u_i + a_i)/(2\tilde{\lambda}\tau)$ from the third equation of (2.8.6), which leads to $\xi_{i+} > 0$ because $u_i > 0$, $a_i \geqslant 0$, $\tilde{\lambda} > 0$ and $\tau > 0$. This contradiction guarantees that $u_i = 0$ for all i. Thus we have

$$\xi_{i+} = \frac{a_i}{2\tilde{\lambda}\tau} \text{ for all } i. \tag{2.8.7}$$

Similarly, from the fourth equation of (2.8.6) and the KKT condition, we have $v_i = 0$ for all i and

$$\xi_{i-} = \frac{\eta_i}{2\tilde{\lambda}(1-\tau)} \text{ for all } i. \tag{2.8.8}$$

Furthermore, the first equation of (2.8.6) leads to

$$\boldsymbol{\theta} = \frac{1}{2}\sum_{i=1}^{n}(a_i - \eta_i)\mathbf{Z}_i \tag{2.8.9}$$

By complementary slackness, we have

$$\begin{aligned} &u_i\xi_{i+} = 0, a_i(Y_i - \alpha - \boldsymbol{\theta}^\top\mathbf{Z}_i - \xi_{i+}) = 0, \\ &v_i\xi_{i-} = 0, \eta_i(-Y_i + \alpha + \boldsymbol{\theta}^\top\mathbf{Z}_i - \xi_{i-}) = 0 \text{ for all } i. \end{aligned} \tag{2.8.10}$$

Plug (2.8.7), (2.8.8), (2.8.9) and (2.8.10) into (2.8.5), and we get the objective function in the dual optimization problem (2.3.4). The constraints for the dual problem are $a_i \geqslant 0$ and $\eta_i \geqslant 0$ for all i. The second equation of (2.8.6) leads to the constraint that $a_i = \eta_i$ for all i. Equation (2.8.9) leads to (2.3.5), which connects the solution of the dual problem to the primal problem. □

We provide the proof of Theorem 3 in this section. The following notations are needed. Without loss of generality, assume $E(\mathbf{X}) = \mathbf{0}$. Denote $\boldsymbol{\Xi} = (\mathbf{X}^\top, Y)^\top$,

$\tilde{\mathbf{X}} = (1, \mathbf{X}^\top)^\top$ and $\tilde{\boldsymbol{\beta}} = (\alpha, \boldsymbol{\beta}^\top)^\top$. Let $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{(p+1)\times(p+1)}$ be a block diagonal matrix such that the block diagonal elements of $\tilde{\boldsymbol{\Sigma}}$ are 0 and $\boldsymbol{\Sigma}$. Then $L_\tau(\alpha, \boldsymbol{\beta})$ in (2.2.1) becomes $E\{\ell_\tau(\tilde{\boldsymbol{\beta}}, \Xi)\}$, where

$$\ell_\tau(\tilde{\boldsymbol{\beta}}, \Xi) = \tilde{\boldsymbol{\beta}}^\top \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\beta}} + \lambda \rho_\tau(Y - \tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{X}}). \tag{2.8.11}$$

Let $D_{\tilde{\boldsymbol{\beta}}}$ be the $(p+1)$-dimensional column vector of differential operators $(\partial/\partial\alpha, \partial/\partial\beta_1, \ldots, \partial/\partial\beta_p)^\top$. The next result gives the gradient of $E\{\ell_\tau(\tilde{\boldsymbol{\beta}}, \Xi)\}$.

Theorem 4. Suppose for any y, the distribution of $\mathbf{X}|Y = y$ is dominated by the Lebesgue measure, $E(Y^2) < \infty$ and $E(\|\mathbf{X}\|^2) < \infty$. Then

$$D_{\tilde{\boldsymbol{\beta}}} E\{\ell_\tau(\tilde{\boldsymbol{\beta}}, \Xi)\} = (0, 2\boldsymbol{\beta}^\top \boldsymbol{\Sigma})^\top - 2\lambda E\{\tau\xi_+ \tilde{\mathbf{X}} - (\tau\xi + \xi_-)\mathbb{I}(\xi < 0)\tilde{\mathbf{X}}\}, \tag{2.8.12}$$

where $\xi = Y - \tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{X}}$, $\xi_+ = max(\xi, 0)$, $\xi_- = max(-\xi, 0)$, and $\mathbb{I}(\cdot)$ is the indicator function.

Proof. Denote $\ell_\tau^*(\tilde{\boldsymbol{\beta}}, \Xi) = \rho_\tau(Y - \tilde{\boldsymbol{\beta}}^\top \tilde{\mathbf{X}})$. It is easy to check that

$$\ell_\tau^*(\tilde{\boldsymbol{\beta}}, \Xi) = \tau\xi_+^2 + (1 - \tau)\xi_-^2. \tag{2.8.13}$$

Note that $D_{\tilde{\boldsymbol{\beta}}}\xi_+ = -\{1 - \mathbb{I}(\xi < 0)\}\tilde{\mathbf{X}}$. It follows that

$$D_{\tilde{\boldsymbol{\beta}}}\xi_+^2 = -2\xi_+\{1 - \mathbb{I}(\xi < 0)\}\tilde{\mathbf{X}}. \tag{2.8.14}$$

Similarly from $D_{\tilde{\boldsymbol{\beta}}}\xi_- = \mathbb{I}(\xi < 0)\tilde{\mathbf{X}}$, we have

$$D_{\tilde{\boldsymbol{\beta}}}\xi_-^2 = 2\xi_-\mathbb{I}(\xi < 0)\tilde{\mathbf{X}}. \tag{2.8.15}$$

Plug (2.8.14) and (2.8.15) into (2.8.13). After taking derivatives, we get

$$\begin{aligned} D_{\tilde{\boldsymbol{\beta}}}\ell^*_\tau(\tilde{\boldsymbol{\beta}},\boldsymbol{\Xi}) &= -2\tau\xi_+\{1-\mathbb{I}(\xi<0)\}\tilde{\mathbf{X}} + 2(1-\tau)\xi_-\mathbb{I}(\xi<0)\tilde{\mathbf{X}} \\ &= -2\tau\xi_+\tilde{\mathbf{X}} + 2\tau(\xi_+ - \xi_-)\mathbb{I}(\xi<0)\tilde{\mathbf{X}} + 2\xi_-\mathbb{I}(\xi<0)\tilde{\mathbf{X}} \\ &= -2\tau\xi_+\tilde{\mathbf{X}} + 2(\tau\xi + \xi_-)\mathbb{I}(\xi<0)\tilde{\mathbf{X}}. \end{aligned} \tag{2.8.16}$$

(2.8.16) and (2.8.11) together imply that

$$E\{D_{\tilde{\boldsymbol{\beta}}}\ell_\tau(\tilde{\boldsymbol{\beta}},\boldsymbol{\Xi})\} = (0, 2\boldsymbol{\beta}^\top\boldsymbol{\Sigma})^\top - 2\lambda E\{\tau\xi_+\tilde{\mathbf{X}} - (\tau\xi+\xi_-)\mathbb{I}(\xi<0)\tilde{\mathbf{X}}\}. \tag{2.8.17}$$

Let Θ be the support of $\tilde{\boldsymbol{\beta}}$. For $\tilde{\boldsymbol{\beta}}_1 = (\alpha_1, \boldsymbol{\beta}_1^\top)^\top \in \Theta$ and $\tilde{\boldsymbol{\beta}}_2 = (\alpha_2, \boldsymbol{\beta}_2^\top)^\top \in \Theta$, we have

$$\begin{aligned} \ell^*_\tau(\tilde{\boldsymbol{\beta}}_1,\boldsymbol{\Xi}) - \ell^*_\tau(\tilde{\boldsymbol{\beta}}_2,\boldsymbol{\Xi}) = \tau\{&(Y-\tilde{\boldsymbol{\beta}}_1^\top\tilde{\mathbf{X}})_+^2 - (Y-\tilde{\boldsymbol{\beta}}_2^\top\tilde{\mathbf{X}})_+^2\} \\ &+ (1-\tau)\{(\tilde{\boldsymbol{\beta}}_1^\top\tilde{\mathbf{X}}-Y)_+^2 - (\tilde{\boldsymbol{\beta}}_2^\top\tilde{\mathbf{X}}-Y)_+^2\}. \end{aligned} \tag{2.8.18}$$

Note that $u_+ - v_+ \leqslant |u-v|$ and $u_+ + v_+ \leqslant |u| + |v|$. Then

$$\begin{aligned} &(Y-\tilde{\boldsymbol{\beta}}_1^\top\tilde{\mathbf{X}})_+^2 - (Y-\tilde{\boldsymbol{\beta}}_2^\top\tilde{\mathbf{X}})_+^2 \leqslant |(\tilde{\boldsymbol{\beta}}_1-\tilde{\boldsymbol{\beta}}_2)^\top\tilde{\mathbf{X}}|(|Y-\tilde{\boldsymbol{\beta}}_1^\top\tilde{\mathbf{X}}| + |Y-\tilde{\boldsymbol{\beta}}_2^\top\tilde{\mathbf{X}}|) \\ &\leqslant (1+\|\mathbf{X}^2\|)^{1/2}(|Y-\tilde{\boldsymbol{\beta}}_1^\top\tilde{\mathbf{X}}| + |Y-\tilde{\boldsymbol{\beta}}_2^\top\tilde{\mathbf{X}}|)\|\tilde{\boldsymbol{\beta}}_1-\tilde{\boldsymbol{\beta}}_2\| < c\|\tilde{\boldsymbol{\beta}}_1-\tilde{\boldsymbol{\beta}}_2\| \end{aligned}$$

for some constant $c < \infty$. The last inequality is due to the assumption that $E(Y^2) < \infty$ and $E(\|\mathbf{X}\|^2) < \infty$. Thus the first term on the right hand side of (2.8.18) satisfies the Lipschitz condition with respect to $\tilde{\boldsymbol{\beta}}$. Similarly, one can show the second term on the right hand side of (2.8.18) also satisfies the Lipschitz condition. Together, we know $\ell_\tau(\tilde{\boldsymbol{\beta}},\boldsymbol{\Xi})$ satisfies the Lipschitz condition with respect to $\tilde{\boldsymbol{\beta}}$. From Lemma 2 of Li, Artemiou and Li (2011), we have

$$D_{\tilde{\boldsymbol{\beta}}}E\{\ell_\tau(\tilde{\boldsymbol{\beta}},\boldsymbol{\Xi})\} = E\{D_{\tilde{\boldsymbol{\beta}}}\ell_\tau(\tilde{\boldsymbol{\beta}},\boldsymbol{\Xi})\}. \tag{2.8.19}$$

(2.8.17) and (2.8.19) together lead to the desired result. □

The next lemma is used to compute the derivative of an expectation of a non-Lipschitz function. Let $D_{\epsilon=0}$ denote the operation of first taking derivative with respect to ϵ and then evaluating the derivative at $\epsilon = 0$.

Lemma 1. Suppose U, V and W are random variables, and $\mathbf{h}(u, v, w)$ is a measurable $\mathbb{R}^k$-valued function. Suppose, moreover,

1. the joint distribution of (U, V, W) is dominated by the Lebesgue measure;

2. for any (v, w), the function $u \mapsto \mathbf{h}(u, v, w) f_{U|V,W}(u|v, w)$ is continuous, where $f_{U|V,W}$ denotes the conditional density of U given (V, W);

3. for each component $h_i(u, v, w)$ of $\mathbf{h}(u, v, w)$, there is a function $c_i(v, w) \geqslant 0$ such that
$$h_i(u, v, w) f_{U|V,W}(u|v, w) \leqslant c_i(v, w) \text{ and } E\{c_i(V, W)\} < \infty.$$

Then, for any constant a, the function

$$\epsilon \mapsto E\{\mathbf{h}(U, V, W)\mathbb{I}(W + U + \epsilon V < a + \epsilon\eta)\}$$

is differentiable at $\epsilon = 0$ with derivative

$$\begin{aligned} &D_{\epsilon=0} E\{\mathbf{h}(U, V, W)\mathbb{I}(W + U + \epsilon V < a + \epsilon\eta)\} \\ &= E_W[f_{U|W}(a - W|W) E_V\{(\eta - V) h_i(a - W, V, W)|U = a - W, W\}], \end{aligned} \tag{2.8.20}$$

where $E(\cdot)$ is with respect to the joint distribution of (U, V, W), $E_W(\cdot)$ is with respect to the marginal distribution of W, and $E_V(\cdot|U = a - W, W)$ is with respect to the conditional distribution $f_{V|U,W}(v|a - w, w)$.

Proof. By the mean value theorem and assumptions 2 and 3, there is a $\delta \in (0, \epsilon)$

such that

$$\left|\epsilon^{-1}\int_{a-w}^{a-w+\epsilon(\eta-v)} h_i(u,v,w)f_{U|V,W}(u|v,w)du\right|$$
$$= |h_i(a-w+\delta(\eta-v),v,w)f_{U|V,W}(a-w+\delta(\eta-v)|v,w)| \leqslant c_i(v,w)$$

By the dominated convergence theorem, we have

$$\begin{aligned}
&\lim_{\epsilon\to 0}\int\int\left\{\epsilon^{-1}\int_{a-w}^{a-w+\epsilon(\eta-v)} h_i(u,v,w)f_{U|V,W}(u|v,w)du\right\} f_{V,W}(v,w)dvdw \\
&= \int\int \lim_{\epsilon\to 0}\left\{\epsilon^{-1}\int_{a-w}^{a-w+\epsilon(\eta-v)} h_i(u,v,w)f_{U|V,W}(u|v,w)du\right\} f_{V,W}(v,w)dvdw \\
&= \int\int (\eta-v)h_i(a-w,v,w)f_{U|V,W}(a-w|v,w)f_{V,W}(v,w)dvdw \\
&= \int\left\{f_W(w)f_{U|W}(a-w|w)\int(\eta-v)h_i(a-w,v,w)f_{V|U,W}(v|a-w,w)dv\right\}dw \\
&= E_W[f_{U|W}(a-W|W)E_V\{(\eta-V)h_i(a-W,V,W)|U=a-W,W\}].
\end{aligned}$$

Here $f_{V|U,W}$ and $f_{U|W}$ are conditional density functions, and f_W denotes the marginal density of W. □

We now present the hessian matrix of the PALS objective function (2.2.1) in the next Theorem.

Theorem 5. Suppose $\mathbf{X}$ has a convex and open support and its conditional distribution given $Y = y$ for any $y \in \mathbb{R}$ is dominated by the Lebesgue measure. Suppose, moreover,

1. for any linearly independent $\boldsymbol{\beta}, \boldsymbol{\delta} \in \mathbb{R}^p$ and any $y \in \mathbb{R}$, the following function is continuous

$$u \mapsto E\{\tilde{\mathbf{X}}(\tau\xi+\xi_-)| -\boldsymbol{\beta}^\top\mathbf{X} = u, \boldsymbol{\delta}^\top\mathbf{X} = v, Y = y\}f_{-\boldsymbol{\beta}^\top\mathbf{X}|\boldsymbol{\delta}^\top\mathbf{X},Y}(u|v,y);$$

2. for any $i = 1, \ldots, p$ and any $y \in \mathbb{R}$, there is a nonnegative function $c_i(v, y)$ with $E\{c_i(V, Y)\} < \infty$ such that

$$E\{X_i(\tau\xi + \xi_-)| - \boldsymbol{\beta}^\top \mathbf{X} = u, \boldsymbol{\delta}^\top \mathbf{X} = v, Y = y\} f_{-\boldsymbol{\beta}^\top \mathbf{X}|\boldsymbol{\delta}^\top \mathbf{X}, Y}(u|v, y) \leqslant c_i(v, y);$$

3. there is a nonnegative function $c_0(v, y)$ with $E\{c_0(V, Y)\} < \infty$ such that

$$f_{-\boldsymbol{\beta}^\top \mathbf{X}|\boldsymbol{\delta}^\top \mathbf{X}, Y}(u|v, y) \leqslant c_0(v, y).$$

Then the function $\tilde{\boldsymbol{\beta}} \mapsto D_{\tilde{\boldsymbol{\beta}}} E\{\ell_\tau(\tilde{\boldsymbol{\beta}}, \Xi)\}$ is differential in all directions with derivative matrix

$$\begin{aligned}\mathbf{H}_\tau =& 2\mathrm{diag}(0, \boldsymbol{\Sigma}) + 2\lambda\tau E[\{1 - \mathbb{I}(\xi < 0)\}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top] \\ &+ 2\lambda E_Y[f_{-\boldsymbol{\beta}^\top \mathbf{X}|Y}(\alpha - Y|Y) E_{\mathbf{X}}\{(\tau\xi + \xi_-)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top| - \boldsymbol{\beta}^\top \mathbf{X} = \alpha - Y, Y\}],\end{aligned}$$

where $E(\cdot)$ is with respect to the joint distribution of $(\mathbf{X}, Y)$, $E_Y(\cdot)$ is with respect to the marginal distribution of Y, and $E_{\mathbf{X}}(\cdot| - \boldsymbol{\beta}^\top \mathbf{X} = \alpha - Y, Y)$ is with respect to the conditional distribution $f_{\mathbf{X}|-\boldsymbol{\beta}^\top \mathbf{X}, Y}(\boldsymbol{x}|\alpha - y, y)$. Furthermore, if the function $\tilde{\boldsymbol{\beta}} \mapsto \tau E[\{1 - \mathbb{I}(\xi < 0)\}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top] + E_Y[f_{-\boldsymbol{\beta}^\top \mathbf{X}|Y}(\alpha - Y|Y) E_{\mathbf{X}}\{(\tau\xi + \xi_-)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top| - \boldsymbol{\beta}^\top \mathbf{X} = \alpha - Y, Y\}]$ is continuous, then $D_{\tilde{\boldsymbol{\beta}}} E\{\ell_\tau(\tilde{\boldsymbol{\beta}}, \Xi)\}$ is jointly differentiable with respect to $\tilde{\boldsymbol{\beta}}$.

Proof. Recall that $\xi = Y - \boldsymbol{\beta}^\top \mathbf{X} - \alpha$ and $\xi_- = max(-\xi, 0)$. First, we verify the directional differentiability of the function $\tilde{\boldsymbol{\beta}} \mapsto E\{(\tau\xi + \xi_-)\mathbb{I}(\xi < 0)\tilde{\mathbf{X}}\}$. For $\boldsymbol{\delta} \in \mathbb{R}^p$ and $\eta \in \mathbb{R}$, the directional derivative along $(\eta, \boldsymbol{\delta}^\top)^\top$ is the derivative of the following function with respect to ϵ at $\epsilon = 0$,

$$
\begin{aligned}
&E\{\tilde{\mathbf{X}}(\tau\xi + \xi_-)\mathbb{I}(Y - \boldsymbol{\beta}^\top\mathbf{X} - \alpha + \epsilon\boldsymbol{\delta}^\top\mathbf{X} < \epsilon\eta)\} \\
&\quad = E[E\{\tilde{\mathbf{X}}(\tau\xi + \xi_-)|Y, \boldsymbol{\beta}^\top\mathbf{X}, \boldsymbol{\delta}^\top\mathbf{X}\}\mathbb{I}(Y - \boldsymbol{\beta}^\top\mathbf{X} + \epsilon\boldsymbol{\delta}^\top\mathbf{X} < \alpha + \epsilon\eta)]
\end{aligned}
$$

Let $W = Y$, $U = -\boldsymbol{\beta}^\top\mathbf{X}$, $V = \boldsymbol{\delta}^\top\mathbf{X}$, $\mathbf{h}(U, V, W) = E\{\tilde{\mathbf{X}}(\tau\xi + \xi_-)|U, V, W\}$, and $a = \alpha$. By (2.8.20) in Lemma 1, the derivative above is

$$
E_Y[f_{-\boldsymbol{\beta}^\top\mathbf{X}|Y}(\alpha - Y|Y)E_{\boldsymbol{\delta}^\top\mathbf{X}}\{\tilde{\mathbf{X}}(\tau\xi + \xi_-)(\eta - \boldsymbol{\delta}^\top\mathbf{X})| - \boldsymbol{\beta}^\top\mathbf{X} = \alpha - Y, Y\}].
$$

Since this holds for all $(\eta, \boldsymbol{\delta}^\top)^\top$, the function $\tilde{\boldsymbol{\beta}} \mapsto E\{(\tau\xi + \xi_-)\mathbb{I}(\xi < 0)\tilde{\mathbf{X}}\}$ is directionally differentiable with derivative matrix

$$
E_Y[f_{-\boldsymbol{\beta}^\top\mathbf{X}|Y}(\alpha - Y|Y)E_{\mathbf{X}}\{(\tau\xi + \xi_-)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top| - \boldsymbol{\beta}^\top\mathbf{X} = \alpha - Y, Y\}]. \tag{2.8.21}
$$

On the other hand, it is easy to see

$$
D_{\tilde{\boldsymbol{\beta}}}(0, \boldsymbol{\beta}^\top\boldsymbol{\Sigma})^\top = \mathrm{diag}(0, \boldsymbol{\Sigma}) \tag{2.8.22}
$$

and

$$
D_{\tilde{\boldsymbol{\beta}}}E(\xi_+\tilde{\mathbf{X}}) = -E[\{1 - \mathbb{I}(\xi < 0)\}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top]. \tag{2.8.23}
$$

Plug (2.8.21), (2.8.22) and (2.8.23) into (2.8.12), and we get the desired result. □

Let $\tilde{\boldsymbol{\beta}}_{0,\tau} = (\alpha_{0,\tau}, \boldsymbol{\beta}_{0,\tau}^\top)^\top$ from minimizing $L_\tau(\alpha, \boldsymbol{\beta})$ in (2.2.1). Let $\hat{\tilde{\boldsymbol{\beta}}}_{0,\tau} = (\hat{\alpha}_{0,\tau}, \hat{\boldsymbol{\beta}}_{0,\tau}^\top)^\top$ from minimizing $\hat{L}_\tau(\alpha, \boldsymbol{\beta})$ in (2.3.1). The next result gives the influence function of PALS. Its proof follows Theorem 5.23 of Van der Vaart (2000), and is thus omitted.

Theorem 6. If the conditions in Theorem 4 and Theorem 5 are satisfied, then

$$\begin{aligned}\hat{\tilde{\boldsymbol{\beta}}}_{0,\tau} =&\tilde{\boldsymbol{\beta}}_{0,\tau} - \mathbf{H}_{0,\tau}^{-1}\{(0, 2\boldsymbol{\beta}_{0,\tau}^{\top}\boldsymbol{\Sigma})^{\top} - 2\lambda\tau E_n(\xi_{0,\tau}^{+}\tilde{\mathbf{X}}) \\ &+ 2\lambda E_n\{\tilde{\mathbf{X}}(\tau\xi_{0,\tau} + \xi_{0,\tau}^{-})\mathbb{I}(\xi_{0,\tau} < 0)\} + o_P(n^{-1/2}),\end{aligned}$$

where $\xi_{0,\tau} = Y - \tilde{\boldsymbol{\beta}}_{0,\tau}^{\top}\tilde{\mathbf{X}}$, $\xi_{0,\tau}^{+} = max(\xi_{0,\tau}, 0)$, $\xi_{0,\tau}^{-} = max(-\xi_{0,\tau}, 0)$, $E_n(\cdot)$ is with respect to the empirical distribution of $(\mathbf{X}, Y)$, and

$$\begin{aligned}\mathbf{H}_{0,\tau} =& \, 2\text{diag}(0, \boldsymbol{\Sigma}) + 2\lambda\tau E[\{1 - \mathbb{I}(\xi_{0,\tau} < 0)\}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\top}] \\ &+ 2\lambda E_Y[f_{-\boldsymbol{\beta}_{0,\tau}^{\top}\mathbf{X}|Y}(\alpha - Y|Y)E_{\mathbf{X}}\{(\tau\xi_{0,\tau} + \xi_{0,\tau}^{-})\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\top}| - \boldsymbol{\beta}_{0,\tau}^{\top}\mathbf{X} = \alpha - Y, Y\}].\end{aligned}$$

Proof of Theorem 3. Consider $0 < \tau_1 < \ldots < \tau_K < 1$. For any $k = 1, \ldots, K$, let $\mathbf{F}_k$ be the last p rows of $\mathbf{H}_{0,\tau_k}^{-1}$. Denote

$$\mathbf{s}_k(\tilde{\boldsymbol{\beta}}_{0,\tau_k}, \boldsymbol{\Xi}) = \mathbf{F}_k\{(0, 2\boldsymbol{\beta}_{0,\tau_k}^{\top}\boldsymbol{\Sigma})^{\top} - 2\lambda\tau_k\xi_{0,\tau_k}^{+}\tilde{\mathbf{X}} + 2\lambda\tilde{\mathbf{X}}(\tau_k\xi_{0,\tau_k} + \xi_{0,\tau_k}^{-})\mathbb{I}(\xi_{0,\tau_k} < 0),$$

CHAPTER 3

ON EXPECTILE-ASSISTED INVERSE REGRESSION ESTIMATION FOR SUFFICIENT DIMENSION REDUCTION

3.1 Introduction

Since its inception about three decades ago, sufficient dimension reduction (Li, 1991; Cook, 1998a) has become a very important tool for modern multivariate analysis. For predictor $\boldsymbol{X} \in \mathbb{R}^p$ and response $Y \in \mathbb{R}$, the goal of sufficient dimension reduction is to find $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ with $d \leqslant p$ such that

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^\top\mathbf{X}, \tag{3.1.1}$$

where $\perp\!\!\!\perp$ means statistical independence. The column space of $\boldsymbol{B}$ satisfying (3.1.1) is known as a dimension reduction space. Under mild conditions, Yin, Li and Cook (2008) showed that the intersection of all dimension reduction spaces is still a dimension reduction space, and it is referred to as the central space for the regression Y on $\boldsymbol{X}$. We denote the central space by $\mathcal{S}_{Y|\boldsymbol{X}}$. The dimension of the central space is known as the structural dimension.

There are many sufficient dimension reduction methods in the literature. Moment-based estimators include sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), principal Hessian directions (Li, 1992; Cook, 1998b), sliced average third-moment estimation (Yin and Cook, 2003), and SIR-α (Saracco, 2005). Semiparametric estimators include minimum average variance estimation (MAVE) (Xia et al., 2002), and semiparametric dimension reduction (Ma and Zhu, 2012; Luo, Li and Yin, 2014). Sparse dimension reduction estimators include sparse SIR (Li, 2007; Tan, Shi and Yu, 2020), sparse MAVE (Wang and Yin, 2008), coordinate-independent sparse estimation (Chen, Zou and Cook, 2010), and sparse semiparametric estimation (Yu et al., 2013). Other sufficient dimension reduction methods include ensemble sufficient dimension reduction (Yin and Li, 2011), nonlinear sufficient dimension reduction (Li, Artemiou and Li, 2011; Lee, Li and Chiaromonte, 2013), groupwise sufficient dimension reduction (Li, Li and Zhu, 2010; Guo et al., 2015), post dimension reduction inference (Kim et al., 2020), and online sufficient dimension reduction (Chavent et al., 2014; Cai, Li and Zhu, 2020). For general reviews, one can refer to Cook (2007), Ma and Zhu (2013), and Dong (2020). An excellent reference is the recent book by Li (2018).

Due to their ease of implementation, SIR and SAVE are two of the most popular sufficient dimension reduction methods. One well-known limitation of SIR and SAVE is that they are not very efficient in the presence of heteroscedasticity. Quantile-based methods are proposed by Wang, Shin and Wu (2018) and Kim, Wu and Shin (2019) to address this limitation, and their proposals work better than SIR or SAVE with heteroscedastic error. However, another well-known limitation of SIR and SAVE is that they may be sensitive to specific link functions between the response and the predictor. In particular, SIR does not work well when the link function is symmetric, and SAVE is not efficient with monotone link functions. Since the quantile-based methods are extensions of SIR and SAVE, they inherit the limitation of their moment-based counterparts and may still have uneven performances with various link functions.

We propose expectile-assisted inverse regression in this paper. Our contribution is two-fold. First, we provide a general framework to extend moment-based dimension reduction methods to their expectile-based counterparts, such as expectile-assisted SIR, expectile-assisted SAVE, and expectile-assisted directional regression. Similar to the quantile-based methods, our expectile-based proposals utilize the information across different levels of the conditional distribution of Y given $\boldsymbol{X}$, and perform better than the corresponding moment-based methods in the presence of heteroscedasticity. Since directional regression (Li and Wang, 2007) is known to perform well for a wide range of link functions, the expectile-assisted directional regression enjoys the additional benefit that it is no longer sensitive to the specific forms of the unknown link functions. Furthermore, to combine the information across different quantile levels, existing quantile-based methods such as quantile-slicing mean estimation and quantile-slicing variance estimation (Kim, Wu and Shin, 2019) rely on intricate weights, and it is not clear how the choice of different weights may affect the final estimation. We propose to combine the information across different expectile levels through random projection, which has roots in the projected resampling approach for multiple response sufficient dimension reduction (Li, Wen and Zhu, 2008). Our proposed expectile-assisted estimators outperform existing methods in both simulation studies and a real data analysis.

The rest of the paper is organized as follows. In Sections 3.2 and 3.3, we provide the population level and the sample level development of expectile-assisted SIR, respectively. Further extensions to expectile-assisted SAVE and expectile-assisted directional regression are described in Section 3.4. Some practical issues such as tuning parameter selection are discussed in Section 3.5. Extensive simulation studies are provided in Section 3.6 and we conclude the dissertation with a real data analysis in Section 3.7. All proofs and additional simulation results are relegated to the Appendix.

3.2 Population level development of expectile-assisted SIR

Expectiles were first introduced by Newey and Powell (1987) in the seminal asymmetric least squares paper. It has gained popularity in finance and risk management for estimating the expected shortfall and value at risk. See, for example, Kim and Lee (2016), Daouia, Girard and Stupfler (2018), and Chen (2018). For $0 < \tau < 1$, denote $f_\tau(\boldsymbol{X})$ as the τ-th expectile of the conditional distribution of Y given $\boldsymbol{X}$. Then

$$f_\tau(\boldsymbol{x}) = \arg\min_a \; E\big\{\phi_\tau(Y-a)|\boldsymbol{X}=\boldsymbol{x}\big\}, \tag{3.2.1}$$

where $\phi_\tau(\cdot)$ is known as the asymmetric loss function and is defined as

$$\phi_\tau(c) = \begin{cases} (1-\tau)c^2, & \text{if } c \leqslant 0, \\ \tau c^2, & \text{if } c > 0. \end{cases} \tag{3.2.2}$$

Proposition 1. For $0 < \tau_1 < \cdots < \tau_k < 1$, let $\boldsymbol{\xi}_{\boldsymbol{X}} = \big(f_{\tau_1}(\boldsymbol{X}), \ldots, f_{\tau_k}(\boldsymbol{X})\big)^\top$. Then $\mathcal{S}_{\boldsymbol{\xi}_{\boldsymbol{X}}|\boldsymbol{X}} \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$.

Proposition 1 suggests that we can recover the central space $\mathcal{S}_{Y|\boldsymbol{X}}$ through estimation of the central space for the regression of $\boldsymbol{\xi}_{\boldsymbol{X}}$ on $\boldsymbol{X}$. We implicitly assume that $f_{\tau_\ell}(\boldsymbol{X})$ from (3.2.1) is well-defined for $\ell = 1, \ldots, k$.

For $\boldsymbol{\xi}_{\boldsymbol{X}} \in \mathbb{R}^k$, the original SIR can not be applied directly due to the multivariate response. Let $\boldsymbol{T} \in \mathbb{R}^k$ be a random vector. We follow Li, Wen, and Zhu (2008) and apply SIR for the regression between $\boldsymbol{\xi}_{\boldsymbol{X}}^{\top}\boldsymbol{T}$ and $\boldsymbol{X}$ instead. Let $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Then $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ denotes the standardized predictor. Let $J_1(\boldsymbol{T}), \ldots, J_H(\boldsymbol{T})$ be the partition of the support of $\boldsymbol{\xi}_{\boldsymbol{X}}^{\top}\boldsymbol{T}$. For $h = 1, \ldots, H$, denote $I_h(\boldsymbol{T})$ as the indicator function of $\boldsymbol{\xi}_{\boldsymbol{X}}^{\top}\boldsymbol{T} \in J_h(\boldsymbol{T})$. Define

$$\boldsymbol{M}(\boldsymbol{T}) = \sum_{h=1}^{H} p_h(\boldsymbol{T})\boldsymbol{\mu}_h(\boldsymbol{T})\boldsymbol{\mu}_h^{\top}(\boldsymbol{T}), \tag{3.2.3}$$

where $p_h(\boldsymbol{T}) = E\{I_h(\boldsymbol{T})\}$ and $\boldsymbol{\mu}_h(\boldsymbol{T}) = E\{\boldsymbol{Z} | \boldsymbol{\xi}_{\boldsymbol{X}}^{\top}\boldsymbol{T} \in J_h(\boldsymbol{T})\}$.

Before we state the next result, we need the following linear conditional mean (LCM) assumption, which is a common assumption in the sufficient dimension reduction literature.

Assumption 3.2.1. *$E(\boldsymbol{X}|\boldsymbol{B}^{\top}\boldsymbol{X})$ is a linear function of $\boldsymbol{B}^{\top}\boldsymbol{X}$, where $\boldsymbol{B}$ is a basis of $\mathcal{S}_{Y|\boldsymbol{X}}$.*

Proposition 2. Let $\boldsymbol{T}$ be a random vector uniformly distributed on the unit sphere $\mathbb{S}^k$. Then under Assumption 3.2.1, $\mathrm{span}\big(\boldsymbol{\Sigma}^{-1/2}E\{\boldsymbol{M}(\boldsymbol{T})\}\big) \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$.

Here $\mathrm{span}(\cdot)$ denotes the column space, and the expectation $E\{\boldsymbol{M}(\boldsymbol{T})\}$ is over the distribution of $\boldsymbol{T}$. We remark that the LCM assumption is not needed in Proposition 1, and it is only needed in Proposition 2 because the classical SIR requires the LCM assumption.

3.3 Sample level algorithm of expectile-assisted SIR

3.3.1 A toy example

We first consider a toy example to fix the idea of expectile-assisted sufficient dimension reduction. Let $\boldsymbol{X} = (X_1, X_2)^{\top}$, where $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, and X_1 is

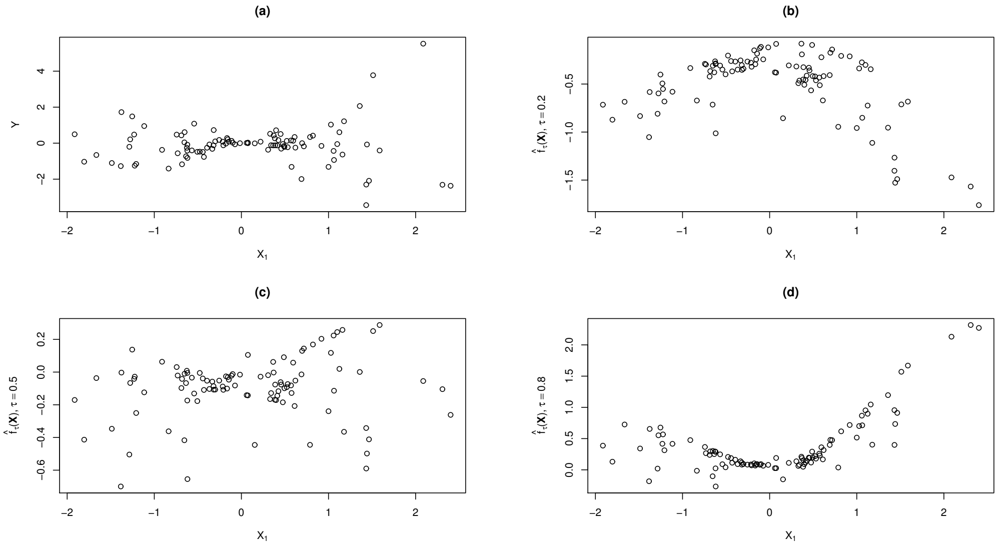

Figure 3.1: A toy example with $Y = X_1\epsilon$. (a) Y versus X_1; (b) $\hat{f}_\tau(\boldsymbol{X})$ versus X_1 with $\tau = 0.2$; (c) $\hat{f}_\tau(\boldsymbol{X})$ versus X_1 with $\tau = 0.5$; (d) $\hat{f}_\tau(\boldsymbol{X})$ versus X_1 with $\tau = 0.8$.

independent of X_2. Let $Y = X_1\epsilon$, where $\epsilon \sim N(0,1)$ and ϵ is independent of $\boldsymbol{X}$. Given 100 data points generated from this model, the scatter plot of Y versus X_1 is provided in panel (a) of Figure 3.1, which shows a clear heteroscedastic trend. While the classical sufficient dimension reduction problem aims to recover the central space $\mathcal{S}_{Y|\boldsymbol{X}}$ through the original response Y, Proposition 1 indicates that we could consider the alternative dimension reduction problem with the expectile $f_\tau(\boldsymbol{X})$ as the new response. In practice, $f_\tau(\boldsymbol{X})$ has to be replaced by its sample level estimator $\hat{f}_\tau(\boldsymbol{X})$, which will be examined in Section 3.2. In panels (b), (c) and (d), the scatter plot of $\hat{f}_\tau(\boldsymbol{X})$ versus X_1 is shown for $\tau = 0.2$, 0.5 and 0.8, respectively. It is obvious from the scatter plots that $\hat{f}_\tau(\boldsymbol{X})$ depends on X_1, and the strength of such dependence may vary when we change the expectile level τ.

3.3.2 Kernel expectile regression

Given an i.i.d. sample $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$, we explain how to estimate the τ-th expectile $f_\tau(\boldsymbol{X})$ of the conditional distribution of Y given $\boldsymbol{X}$ in this section. This step is the same for expectile-assisted SIR and the other expectile-assisted inverse regression methods to be discussed in Section 4. The original estimator in Newey and Powell (1987) focused on expectiles in linear regression. To estimate the conditional expectiles in nonlinear models, Yao and Tong (1996) proposed a local linear polynomial estimator. More recently, Yang, Zhang and Zou (2018) developed a reproducing kernel Hilbert space (RKHS) estimator for flexible expectile regression. We adapt the RKHS estimator with the following Gaussian radial basis kernel

$$K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp(-r\|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2), \tag{3.3.1}$$

where r is a tuning parameter and $\|\cdot\|$ denotes the Euclidean norm. Let $\mathbb{H}_K$ be the RKHS generated from the kernel function (3.3.1). As an element of $\mathbb{H}_K$, $f_\tau(\boldsymbol{X})$ evaluated at $\boldsymbol{X} = \boldsymbol{X}_i$ can be estimated by

$$\hat{f}_\tau(\boldsymbol{X}_i) = \hat{\alpha}_{0,\tau} + \sum_{j=1}^{n} \hat{\alpha}_{j,\tau} K(\boldsymbol{X}_i, \boldsymbol{X}_j). \tag{3.3.2}$$

Let $\hat{\boldsymbol{\alpha}}_\tau = (\hat{\alpha}_{0,\tau}, \hat{\alpha}_{1,\tau}, \ldots, \hat{\alpha}_{n,\tau})$ and $\boldsymbol{\alpha}_\tau = (\alpha_{0,\tau}, \alpha_{1,\tau}, \ldots, \alpha_{n,\tau})$. Then $\hat{\boldsymbol{\alpha}}_\tau$ in (3.3.2) is the minimizer of the regularized empirical risk function on $\mathbb{H}_K$

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_\tau = \underset{\boldsymbol{\alpha}_\tau}{\operatorname{argmin}} \; & \sum_{i=1}^{n} \phi_\tau\Big(Y_i - \alpha_{0,\tau} - \sum_{j=1}^{n} \alpha_{j,\tau} K(\boldsymbol{X}_i, \boldsymbol{X}_j)\Big) \\ & + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{i,\tau}\alpha_{j,\tau} K(\boldsymbol{X}_i, \boldsymbol{X}_j), \end{aligned} \tag{3.3.3}$$

where $\phi_\tau(\cdot)$ is defined in (3.2.1) and λ is a tuning parameter. The optimization (3.3.3) and the evaluation (3.3.2) can be done very efficiently in the KERE package in R. The choices for the tuning parameters r in (3.3.1) and λ in (3.3.3) are discussed in Section 5.

3.3.3 Projective resampling for multiple response SIR

Given an i.i.d. sample $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$, the sample level expectile-assisted SIR algorithm is as follows.

1. For a given integer k, specify $0 < \tau_1 < \cdots < \tau_k < 1$. For $i = 1, \ldots, n$, calculate $\hat{\boldsymbol{\xi}}_{\boldsymbol{X}_i} = \big(\hat{f}_{\tau_1}(\boldsymbol{X}_i), \cdots, \hat{f}_{\tau_k}(\boldsymbol{X}_i)\big)^\top$, where the ℓ-th component of $\hat{\boldsymbol{\xi}}_{\boldsymbol{X}_i}$ is given by (3.3.2) with $\tau = \tau_\ell$.

2. Let $\hat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$ and $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top$. Calculate standardized predictors $\hat{\boldsymbol{Z}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})$ for $i = 1, \ldots, n$.

3. For a given integer N, generate an i.i.d. sample $\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(N)}$ from the uniform distribution on the unit sphere $\mathbb{S}^k$. For $j = 1, \ldots, N$, let $J_1(\boldsymbol{t}^{(j)}), \ldots,$ $J_H(\boldsymbol{t}^{(j)})$ be the partition of the support of $\hat{\boldsymbol{\xi}}_{\boldsymbol{X}}^\top \boldsymbol{t}^{(j)}$. For $h = 1, \ldots, H$, denote $I_{hi}(\boldsymbol{t}^{(j)})$ as the indicator function of $\hat{\boldsymbol{\xi}}_{\boldsymbol{X}_i}^\top \boldsymbol{t}^{(j)} \in J_h(\boldsymbol{t}^{(j)})$.

4. We now calculate the sample version of $E\{\boldsymbol{M}(\boldsymbol{T})\}$.

   4.1 For $j = 1, \ldots, N$, let $\hat{p}_h(\boldsymbol{t}^{(j)}) = n^{-1}\sum_{i=1}^n I_{hi}(\boldsymbol{t}^{(j)})$ and $\hat{\boldsymbol{\mu}}_h(\boldsymbol{t}^{(j)}) = \{n\hat{p}_h(\boldsymbol{t}^{(j)})\}^{-1}\sum_{i=1}^n \hat{\boldsymbol{Z}}_i I_{hi}(\boldsymbol{t}$
   Then the sample estimator of (3.2.3) with $\boldsymbol{T} = \boldsymbol{t}^{(j)}$ becomes

   $$\hat{\boldsymbol{M}}(\boldsymbol{t}^{(j)}) = \sum_{h=1}^{H} \hat{p}_h(\boldsymbol{t}^{(j)})\hat{\boldsymbol{\mu}}_h(\boldsymbol{t}^{(j)})\hat{\boldsymbol{\mu}}_h^\top(\boldsymbol{t}^{(j)}).$$

   4.2 Calculate $\hat{\boldsymbol{M}}(\boldsymbol{T}) = N^{-1}\sum_{j=1}^N \hat{\boldsymbol{M}}(\boldsymbol{t}^{(j)})$.

5. For a given structural dimension d, let $(\hat{\boldsymbol{v}}_1, .., \hat{\boldsymbol{v}}_d)$ be the eigenvectors corresponding to the d leading eigenvalues of $\hat{\boldsymbol{M}}(\boldsymbol{T})$. The final estimator of $\mathcal{S}_{Y|\boldsymbol{X}}$ is then $\operatorname{span}(\hat{\boldsymbol{B}})$, where $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{v}}_1, .., \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{v}}_d)$.

In the numerical studies in Section 6 and the real data analysis in Section 7, we fix $N = 1000$, $k = 9$, and set $\tau_\ell = 10^{-1}\ell$ for $\ell = 1, \ldots, 9$. The effects of different N and

k are also studied. Our experience suggests that the proposed method is not very sensitive to the choice of N and k.

3.4 Extensions of SAVE and directional regression

Expectile-assisted dimension reduction is a very general framework, and can be readily generalized to other moment-based methods such as SAVE and directional regression. We focus on the population level development of expectile-assisted SAVE and expectile-assisted directional regression in this section.

Recall that $I_h(\boldsymbol{T})$ denotes the indicator function of $\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{T} \in J_h(\boldsymbol{T})$, $p_h(\boldsymbol{T}) = E\{I_h(\boldsymbol{T})\}$, and $\boldsymbol{\mu}_h(\boldsymbol{T}) = E\{\boldsymbol{Z}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{T} \in J_h(\boldsymbol{T})\}$. Define

$$\boldsymbol{G}(\boldsymbol{T}) = \sum_{h=1}^{H} p_h(\boldsymbol{T})\{\boldsymbol{V}_h(\boldsymbol{T}) - \boldsymbol{\mu}_h(\boldsymbol{T})\boldsymbol{\mu}_h^\top(\boldsymbol{T})\}^2,$$

where $\boldsymbol{V}_h(\boldsymbol{T}) = E\{\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{I}_p|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{T} \in J_h(\boldsymbol{T})\}$. In addition to the LCM assumption, we need the constant conditional variance (CCV) assumption as follows

Assumption 3.4.1. $\mathrm{Var}(\boldsymbol{X}|\boldsymbol{B}^\top \boldsymbol{X})$ *is a nonrandom matrix, where* $\boldsymbol{B}$ *is a basis of* $\mathcal{S}_{Y|\boldsymbol{X}}$.

Proposition 3. Let $\boldsymbol{T}$ be a random vector uniformly distributed on the unit sphere $\mathbb{S}^k$. Then under Assumptions 3.2.1 and 3.4.1, $\mathrm{span}\big(\boldsymbol{\Sigma}^{-1/2}E\{\boldsymbol{G}(\boldsymbol{T})\}\big) \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$.

In a similar fashion, define

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{T}) = {} & 2\sum_{h=1}^{H} p_h(\boldsymbol{T})\boldsymbol{V}_h(\boldsymbol{T})\boldsymbol{V}_h(\boldsymbol{T}) + 2\left\{\sum_{h=1}^{H} p_h(\boldsymbol{T})\boldsymbol{\mu}_h(\boldsymbol{T})\boldsymbol{\mu}_h^\top(\boldsymbol{T})\right\}^2 \\ & + 2\left\{\sum_{h=1}^{H} p_h(\boldsymbol{T})\boldsymbol{\mu}_h^\top(\boldsymbol{T})\boldsymbol{\mu}_h(\boldsymbol{T})\right\}\left\{\sum_{h=1}^{H} p_h(\boldsymbol{T})\boldsymbol{\mu}_h(\boldsymbol{T})\boldsymbol{\mu}_h^\top(\boldsymbol{T})\right\},\end{aligned}$$

and we have

Proposition 4. Let $\boldsymbol{T}$ be a random vector uniformly distributed on the unit sphere $\mathbb{S}^k$. Then under Assumptions 3.2.1 and 3.4.1, $\operatorname{span}\big(\boldsymbol{\Sigma}^{-1/2}E\{\boldsymbol{F}(\boldsymbol{T})\}\big) \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$.

Based on Proposition 3 and Proposition 4, we may update step 4 and step 5 of the expectile-assisted SIR algorithm to get the sample estimators of $\boldsymbol{\Sigma}^{-1/2}E\{\boldsymbol{G}(\boldsymbol{T})\}$ and $\boldsymbol{\Sigma}^{-1/2}E\{\boldsymbol{F}(\boldsymbol{T})\}$. We refer to them as the expectile-assisted SAVE estimator and the expectile-assisted directional regression estimator, respectively.

3.5 Additional issues

3.5.1 Selecting tuning parameters

We first discuss the choice of r in (3.3.1). For the Gaussian radial basis kernel, Li, Artemiou and Li (2011) suggested using

$$r = 1/\gamma^2 \text{ with } \gamma = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|, \tag{3.5.1}$$

where $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is an i.i.d. sample. The effect of different choices of r is examined in the Appendix, and it is shown that the proposed methods are not sensitive to the choice of r when it varies around $1/\gamma^2$.

Now we turn our attention to selecting λ in (3.3.3). For a given λ, denote the final estimator from the algorithm in Section 3.3 as $\hat{\boldsymbol{B}}_\lambda$. Distance correlation (Székely, Rizzo and Bakirov, 2007) is a good measure of linear as well as nonlinear dependence. Denote $\rho^2(Y, \boldsymbol{B}^\top\boldsymbol{X})$ as the squared distance correlation between Y and $\boldsymbol{B}^\top\boldsymbol{X}$. It is known that $\rho^2(Y, \boldsymbol{B}^\top\boldsymbol{X}) = 0$ if and only if Y is independent of $\boldsymbol{B}^\top\boldsymbol{X}$.

Given an i.i.d. sample $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$ and $\hat{\boldsymbol{B}}_\lambda$, the sample squared distance correlation between Y and $\hat{\boldsymbol{B}}_\lambda^\top\boldsymbol{X}$ is denoted as $\hat{\rho}^2(Y, \hat{\boldsymbol{B}}_\lambda^\top\boldsymbol{X})$. From a set of candidate values for λ, we choose λ such that the sample squared distance correlation $\hat{\rho}^2(Y, \hat{\boldsymbol{B}}_\lambda^\top\boldsymbol{X})$ is maximized. The comparison between the data-driven approach and using a prespecified λ is provided in the Appendix, which indicates that the data-driven approach works well in practice.

SIR, SAVE, directional regression, and their expectile-assisted counterparts all rely on slicing the continuous response. For the number of slices H, existing sufficient dimension reduction literature suggests that SIR is not very sensitive to H as it only involves intraslice means (Zhu, L. X. and Ng, K. W., 1995). SAVE, on the other hand, involves intraslice variances and is more sensitive to H than SIR (Li, Y. and Zhu, L. X., 2007). We examine the effect of H in the Appendix. It seems that expectile-assisted SIR and expectile-assisted directional regression are not very sensitive to the choice of H, while expectile-assisted SAVE prefers smaller values of H.

In step 5 of the sample level algorithm in Section 3.3, we need the structural dimension d, which has to be estimated in practice. The problem of estimating the structural dimension is known as order determination. Sequential test is a common method for order determination in the sufficient dimension reduction literature. For order determination based on asymptotic sequential test, one may refer to Chapter 9 of Li (2018). Cook and Yin (2001) proposed a permutation sequential test approach for order determination, which can be directly applied to our proposed expectile-assisted methods. The comparison between the asymptotic sequential test and the permutation sequential test is provided in the Section 3.7.

3.5.2 Pooled marginal estimators

In Cook and Setodji (2003), Saracco (2005), Yin and Bura (2006), Barreda, Gannoun and Saracco (2007), Coudret, Girard and Saracco (2014), pooled marginal estimators are proposed for sufficient dimension reduction with multiple responses. Without loss of generality, we propose the pooled marginal expectile-assisted SIR in this section. The extensions to SAVE and directional regression are similar and thus omitted.

Recall that for $\ell = 1, \ldots, k$, $f_{\tau_\ell}(\boldsymbol{X})$ denotes the τ_ℓ-th conditional expectile of

Y given $\boldsymbol{X}$. Let $J_{1,\ell}, \ldots, J_{H,\ell}$ be a partition for the support of $f_{\tau_\ell}(\boldsymbol{X})$. Let $p_{h,\ell} = E\{f_{\tau_\ell}(\boldsymbol{X}) \in J_{h,\ell}\}$, $\boldsymbol{\mu}_{h,\ell} = E\{\boldsymbol{Z} | f_{\tau_\ell}(\boldsymbol{X}) \in J_{h,\ell}\}$, and $\boldsymbol{M}_\ell = \sum_{h=1}^{H} p_{h,\ell} \boldsymbol{\mu}_{h,\ell} \boldsymbol{\mu}_{h,\ell}^\top$. Define $\widetilde{\boldsymbol{M}} = (\boldsymbol{M}_1, \ldots, \boldsymbol{M}_k)$, and we have

Proposition 5. Under Assumption 3.2.1, $\operatorname{span}\big(\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{M}}\big) \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$.

The marginal approach essentially considers k univariate response sufficient dimension reduction problems separately and then assemble the individual estimators for each response to get the final estimator. At the sample level, the estimator of the central space consists of the left singular vectors of the sample version of $\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{M}}$. We refer to it as the pooled marginal expectile-assisted SIR estimator.

3.6 Simulation studies

We examine the empirical performances of our proposals through synthetic examples in this section. The predictor $\boldsymbol{X}$ is generated from $N(\boldsymbol{0}, \boldsymbol{I}_p)$ with $p = 6$ or $p = 20$. The first six components of $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ is $(1,1,1,0,0,0)$, and the first six components of $\boldsymbol{\beta}_2 \in \mathbb{R}^p$ is $(1,0,0,0,1,3)$. The remaining components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are all zero when $p = 20$. The response Y is generated as follows:

$$\text{I} : Y = 0.4(\boldsymbol{\beta}_1^\top \boldsymbol{X})^2 + 3\sin(\boldsymbol{\beta}_2^\top \boldsymbol{X}/4) + \sigma\epsilon,$$

$$\text{II} : Y = 3\sin(\boldsymbol{\beta}_1^\top \boldsymbol{X}/4) + 3\sin(\boldsymbol{\beta}_2^\top \boldsymbol{X}/4) + \sigma\epsilon,$$

$$\text{III} : Y = 0.4(\boldsymbol{\beta}_1^\top \boldsymbol{X})^2 + |\boldsymbol{\beta}_2^\top \boldsymbol{X}|^{1/2} + \sigma\epsilon,$$

$$\text{IV} : Y = 3\sin(\boldsymbol{\beta}_2^\top \boldsymbol{X}/4) + \{1 + (\boldsymbol{\beta}_1^\top \boldsymbol{X})^2\}\sigma\epsilon,$$

$$\text{V} : Y = \boldsymbol{\beta}_1^\top \boldsymbol{X}\epsilon,$$

where $\sigma = 0.2$, $\epsilon \sim N(0,1)$, and ϵ is independent of $\boldsymbol{X}$. Models I through IV are the same models used in Li and Wang (2007), and model V is similar to the toy

Model	SIR	EA-SIR	QUME	SAVE	EA-SAVE	QUVE	DR	EA-DR
I	1.648	1.343	1.512	0.626	0.554	0.939	0.384	0.345
	(0.043)	(0.058)	(0.047)	(0.059)	(0.050)	(0.052)	(0.041)	(0.029)
II	1.521	1.567	1.518	1.565	1.543	1.737	1.492	1.497
	(0.046)	(0.046)	(0.049)	(0.047)	(0.048)	(0.040)	(0.051)	(0.050)
III	2.620	2.308	2.722	0.652	0.547	0.287	0.638	0.543
	(0.063)	(0.060)	(0.067)	(0.050)	(0.046)	(0.034)	(0.049)	(0.048)
IV	1.700	1.396	1.556	1.598	1.247	1.734	1.557	1.177
	(0.034)	(0.054)	(0.047)	(0.046)	(0.056)	(0.047)	(0.046)	(0.056)
V	1.667	1.484	1.513	0.572	0.792	0.967	0.561	0.799
	(0.037)	(0.052)	(0.048)	(0.046)	(0.061)	(0.059)	(0.045)	(0.064)

Table 3.1: Results based on $(n, p) = (100, 6)$. The average of Δ and its standard error (in parentheses) are reported based on 100 repetitions.

example in Section 3.1. Following Li and Wang (2007), two sample size settings are considered. For $n = 100$, we set $p = 6$ and number of slices $H = 5$. For $n = 500$, we set $p = 20$ and $H = 10$.

We compare SIR, SAVE, directional regression (DR), expectile-assisted SIR (EA-SIR), expectile-assisted SAVE (EA-SAVE), and expectile-assisted directional regression (EA-DR). Quantile-slicing mean estimation (QUME) and quantile-slicing variance estimation (QUVE) (Kim, Wu and Shin, 2019) are also included for the comparison. For models I through IV, the basis for the central space is $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, and the central space basis for model V is $\boldsymbol{B} = \boldsymbol{\beta}_1$. For estimator $\hat{\boldsymbol{B}}$, we measure its performance by $\Delta = \|\boldsymbol{P}_{\boldsymbol{B}} - \boldsymbol{P}_{\hat{\boldsymbol{B}}}\|_F$. Here $\boldsymbol{P}_{\boldsymbol{B}} = \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1}\boldsymbol{B}^\top$, $\boldsymbol{P}_{\hat{\boldsymbol{B}}} = \hat{\boldsymbol{B}}(\hat{\boldsymbol{B}}^\top \hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^\top$, and $\| \cdot \|_F$ denotes the matrix Frobenius norm. Smaller values of Δ mean better performances. We fix the number of projections as $N = 1000$ and the number of expectile levels as $k = 9$ for all three expectile-assisted methods. Furthermore, r for the Gaussian radial basis kernel is set as in (3.5.1), λ for the regularization term in (3.3.3) is chosen in a data-driven manner as described in Section 5.1. More simulation studies are provided in the Appendix for different choices of H, N, k, r and λ.

For the $(n, p) = (100, 6)$ setting, we summarize the simulation results based on 100 repetitions in Table 3.1. First we compare SIR with EA-SIR and QUME, which all belong to first-order inverse regression estimators. We see that while QUME improves over SIR with the exception of model III, EA-SIR has the best overall performance among these three methods. Next we compare SAVE with EA-SAVE and QUVE, which are all second-order inverse regression methods. EA-SAVE again has the best overall performance in this group. It is interesting to see that QUVE is worse than both SAVE and EA-SAVE for all models other than model III, where the link functions are symmetric. When we compare SIR-based methods with SAVE-based methods, we see that SIR-based methods do not work as well for models I, III, IV, and V. This is due to the fact that at least one link function in the mean component is symmetric for models I, III and IV, and SIR is known to be ineffective in the presence of symmetric link functions. EA-SIR and QUME inherit this limitation. Although the error term in model V involves an asymmetric linear function of $\boldsymbol{X}$, we have seen in panel (a) of Figure 3.1 that there is still symmetry about the y-axis in this type of heteroscedastic model. We note that EA-SIR improves over SIR in both model IV and V with heteroscedastic error. DR is very competitive across all five models as it is not sensitive to the shape of the link functions. EA-DR further improves over DR in three out of five models and enjoys the best overall performance.

The simulation results for the $(n, p) = (500, 20)$ setting are summarized in Table 3.2. We observe similar results as in Table 3.1: the overall performance of EA-SIR is better than SIR and QUME; the overall performance of EA-SAVE is better than SAVE and QUVE; DR and EA-DR enjoy the best overall performances. Furthermore, SAVE-based methods are significantly worse than their SIR-based counterparts for model II, where both link functions are monotone. It is known in the sufficient dimension reduction literature that SAVE may not be very efficient with monotone link functions (Li and Wang, 2007). EA-SAVE and QUVE also have this limitation.

Next we compare our proposed expectile-assisted methods based on random projections with the pooled marginal estimators described in Section 5.2. We denote the pooled marginal expectile-assisted SIR as mEA-SIR. Similarly, mEA-SAVE and

Model	SIR	EA-SIR	QUME	SAVE	EA-SAVE	QUVE	DR	EA-DR
I	1.845 (0.026)	1.707 (0.037)	1.724 (0.035)	1.114 (0.061)	0.454 (0.017)	0.846 (0.040)	0.245 (0.007)	0.265 (0.009)
II	1.564 (0.036)	1.685 (0.033)	1.871 (0.016)	1.796 (0.023)	1.845 (0.019)	2.017 (0.011)	1.710 (0.029)	1.735 (0.032)
III	3.594 (0.028)	3.447 (0.036)	3.534 (0.031)	0.451 (0.016)	0.364 (0.012)	0.710 (0.050)	0.443 (0.015)	0.355 (0.013)
IV	1.908 (0.016)	1.832 (0.027)	1.904 (0.015)	1.747 (0.040)	1.524 (0.042)	2.053 (0.013)	1.584 (0.041)	1.473 (0.045)
V	1.915 (0.011)	1.850 (0.019)	1.842 (0.018)	0.283 (0.011)	0.388 (0.044)	0.313 (0.022)	0.280 (0.011)	0.373 (0.043)

Table 3.2: Results based on $(n, p) = (500, 20)$. The average of Δ and its standard error (in parentheses) are reported based on 100 repetitions.

Model	n	EA-SIR	mEA-SIR	EA-SAVE	mEA-SAVE	EA-DR	mEA-DR
I	50	1.531 (0.057)	1.481 (0.059)	1.772 (0.072)	2.050 (0.065)	0.893 (0.057)	0.934 (0.059)
	100	1.343 (0.058)	1.302 (0.059)	0.554 (0.050)	0.678 (0.061)	0.345 (0.029)	0.368 (0.032)
	150	1.274 (0.062)	1.160 (0.057)	0.270 (0.033)	0.270 (0.033)	0.174 (0.012)	0.190 (0.015)
II	50	1.540 (0.050)	1.558 (0.046)	1.698 (0.048)	2.144 (0.071)	1.574 (0.045)	1.571 (0.044)
	100	1.567 (0.046)	1.564 (0.043)	1.543 (0.048)	1.612 (0.043)	1.497 (0.050)	1.521 (0.049)
	150	1.406 (0.055)	1.473 (0.052)	1.426 (0.048)	1.474 (0.050)	1.425 (0.052)	1.460 (0.051)

Table 3.3: Results based on $p = 6$. The average of Δ and its standard error (in parentheses) are reported based on 100 repetitions.

mEA-DR denote the corresponding pooled marginal estimators for SAVE and DR. For this comparison, we fix $p = 6$, $H = 5$, and consider $n = 50$, 100 and 150. The results based on 100 repetitions are summarized in Table 3.3. The pooled marginal estimators have decent performances, which confirms the result of Proposition 5. As sample size increases, the performances of all projective resampling methods as well as the pooled marginal methods improve. The pooled marginal estimators are outperformed by the corresponding projective resampling estimators in 5 out of 9 cases for model I, and in 7 out of 9 cases for model II. This confirms the finding in Li, Wen and Zhu (2008) that projective resampling is more efficient than the pooled marginal estimators.

We present additional simulation results to inspect the effects of H, N, k, r and λ for expectile-assisted estimators. From Table 3.4, we see that EA-SIR and EA-DR are not very sensitive to the choice of number of slices in both model I and model II. EA-SAVE has stable performance in model II, but becomes worse in model I when H increases from 4 to 10.

Model	Method	$H = 4$	$H = 5$	$H = 10$
I	EA-SIR	1.364 (0.060)	1.343 (0.058)	1.217 (0.060)
	EA-SAVE	0.489 (0.044)	0.554 (0.050)	1.228 (0.062)
	EA-DR	0.365 (0.036)	0.345 (0.029)	0.380 (0.035)
II	EA-SIR	1.525 (0.046)	1.567 (0.046)	1.497 (0.047)
	EA-SAVE	1.515 (0.051)	1.543 (0.048)	1.616 (0.041)
	EA-DR	1.496 (0.049)	1.497 (0.050)	1.517 (0.049)

Table 3.4: Effect of H (number of slices). The average of Δ and its standard error (in parentheses) are reported based on 100 repetitions for $(n, p) = (100, 6)$.

Model	Method	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
I	EA-SIR	1.464 (0.052)	1.444 (0.052)	1.431 (0.052)	1.343 (0.058)
	EA-SAVE	0.584 (0.054)	0.546 (0.052)	0.558 (0.053)	0.554 (0.050)
	EA-DR	0.359 (0.033)	0.359 (0.033)	0.356 (0.033)	0.345 (0.029)
II	EA-SIR	1.494 (0.047)	1.519 (0.047)	1.505 (0.048)	1.567 (0.046)
	EA-SAVE	1.514 (0.049)	1.495 (0.048)	1.502 (0.047)	1.543 (0.048)
	EA-DR	1.497 (0.051)	1.472 (0.052)	1.461 (0.053)	1.497 (0.050)

Table 3.5: Effect of N (number of projections). The average of Δ and its standard error (in parentheses) are reported based on 100 repetitions for $(n, p) = (100, 6)$.

From Table 3.5, we see that all three expectile-assisted methods are not overly sensitive to N, which denotes the number of projections. Table 3.6 summarizes the results for different k. For $k = 4$, we set the expectile levels to be 0.2, 0.4, 0.6 and 0.8; for $k = 9$, we set the expectile levels to be $0.1, 0.2, \ldots, 0.9$; for $k = 19$, the expectile

Model	Method	$k = 4$	$k = 9$	$k = 19$
I	EA-SIR	1.455 (0.052)	1.343 (0.058)	1.429 (0.054)
	EA-SAVE	0.567 (0.052)	0.554 (0.050)	0.549 (0.051)
	EA-DR	0.358 (0.034)	0.345 (0.029)	0.368 (0.035)
II	EA-SIR	1.504 (0.049)	1.567 (0.046)	1.512 (0.047)
	EA-SAVE	1.489 (0.050)	1.543 (0.048)	1.487 (0.049)
	EA-DR	1.472 (0.053)	1.497 (0.050)	1.452 (0.053)

Table 3.6: Effect of k (number of expectile levels). The average of Δ and its standard error (in parentheses) are reported based on 100 repetitions for $(n, p) = (100, 6)$.

Model	Method	λ					DD
		0.001	0.01	0.1	1	10	
I	EA-SIR	1.675 (0.042)	1.67 (0.044)	1.634 (0.043)	1.524 (0.050)	**1.339** (0.062)	1.343 (0.058)
	EA-SAVE	0.643 (0.056)	0.619 (0.054)	**0.576** (0.047)	0.927 (0.058)	1.261 (0.058)	0.554 (0.050)
	EA-DR	**0.357** (0.033)	0.376 (0.034)	0.486 (0.041)	0.904 (0.053)	1.114 (0.06)	0.345 (0.029)
II	EA-SIR	**1.521** (0.047)	1.589 (0.041)	1.608 (0.040)	1.576 (0.044)	1.662 (0.044)	1.567 (0.046)
	EA-SAVE	1.573 (0.046)	**1.558** (0.047)	1.574 (0.047)	1.600 (0.045)	1.599 (0.047)	1.543 (0.048)
	EA-DR	**1.492** (0.050)	1.517 (0.047)	1.574 (0.045)	1.645 (0.041)	1.647 (0.042)	1.497 (0.050)

Table 3.7: Fixed λ in (3.3.3) versus data-driven λ (DD). The average of Δ and its standard error (in parentheses) are reported based on 100 repetitions for $(n, p) = (100, 6)$.

levels are $0.05, 0.1, 0.15, \ldots, 0.95$. We see from Table 3.6 that the expectile-assisted methods are not very sensitive to the choice of different expectile levels.

Table 3.8 summarizes the results for different r in the Gaussian radial basis kernel (3.3.1). We see that the results are stable when r varies around $1/\gamma^2$, which is the suggested value in (3.5.1). We compare the choice of fixed λ versus the data-driven λ in Table 3.7. Please refer to Section 5.1 for the details of the data-driven approach to choose λ, which is in the last column of Table 3.7. The best λ that corresponds to the optimal result (boldfaced for easy reference) changes across different models and different methods, and the data-driven λ always achieves a decent result that is very close to the optimal result.

Model	Method	r				
		$1/(4\gamma^2)$	$1/(2\gamma^2)$	$1/\gamma^2$	$2/\gamma^2$	$4/\gamma^2$
I	EA-SIR	1.294	1.308	1.343	1.344	1.431
		(0.058)	(0.058)	(0.058)	(0.056)	(0.053)
	EA-SAVE	0.571	0.544	0.554	0.576	0.541
		(0.045)	(0.047)	(0.050)	(0.051)	(0.051)
	EA-DR	0.463	0.412	0.345	0.368	0.364
		(0.035)	(0.037)	(0.029)	(0.033)	(0.034)
II	EA-SIR	1.601	1.567	1.567	1.507	1.508
		(0.042)	(0.042)	(0.046)	(0.047)	(0.047)
	EA-SAVE	1.562	1.535	1.543	1.556	1.494
		(0.048)	(0.051)	(0.048)	(0.047)	(0.046)
	EA-DR	1.602	1.577	1.497	1.504	1.476
		(0.044)	(0.046)	(0.050)	(0.049)	(0.050)

Table 3.8: Effect of r in (3.3.1). The average of Δ and its standard error (in parentheses) are reported based on 100 repetitions for $(n, p) = (100, 6)$.

3.7 Order Determination

We take a sequential test approach for order determination. Consider

$$H_0^{(m)} : d = m \text{ v.s. } H_a^{(m)} : d > m, \text{ for } m = 0, 1, \ldots, p-1. \tag{3.7.1}$$

Then we estimate the structural dimension d by the smallest m at which $H_0^{(m)}$ in (3.7.1) is accepted. The asymptotic sequential tests for SIR, SAVE and directional regression are well-known in the literature. See, for example, Chapter 9 of Li (2018).

Next, we describe a permutation test approach for order determination based on EA-SIR. Given an i.i.d. sample $\{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$, let $\hat{\boldsymbol{M}}(\boldsymbol{T})$ denote the sample estimator of $E\{\boldsymbol{M}(\boldsymbol{T})\}$. Suppose $\hat{\eta}_1 \geqslant \hat{\eta}_2 \geqslant \ldots \geqslant \hat{\eta}_p$ are the eigenvalues of $\hat{\boldsymbol{M}}(\boldsymbol{T})$, and the corresponding eigenvectors are $\hat{\boldsymbol{v}}_1, .., \hat{\boldsymbol{v}}_p$. Consider test statistic

$$\Lambda_m = n \sum_{j=m+1}^{p} \hat{\eta}_j. \tag{3.7.2}$$

Following Section 3.3 of Cook and Yin (2001), we have the following algorithm

1. Calculate standardized predictors $\hat{\boldsymbol{Z}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})$ for $i = 1, \ldots, n$, where $\hat{\boldsymbol{\mu}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i$ and $\hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top$.

2. Denote $\hat{\boldsymbol{U}}_1 = (\hat{\boldsymbol{v}}_1, .., \hat{\boldsymbol{v}}_m)$ and $\hat{\boldsymbol{U}}_2 = (\hat{\boldsymbol{v}}_{m+1}, .., \hat{\boldsymbol{v}}_p)$. Construct sample principal predictors $\hat{\boldsymbol{W}}_{1i} = \hat{\boldsymbol{U}}_1^\top \hat{\boldsymbol{Z}}_i$ and $\hat{\boldsymbol{W}}_{2i} = \hat{\boldsymbol{U}}_2^\top \hat{\boldsymbol{Z}}_i$, $i = 1, \ldots, n$.

3. For $b = 1, \ldots, B$, randomly permute the indices i of $\hat{\boldsymbol{W}}_{2i}$ to obtain the permuted set $\hat{\boldsymbol{W}}_{2i}^{[b]}$. Construct the test statistic $\Lambda_m^{[b]}$ in (3.7.2) based on $\{(\hat{\boldsymbol{W}}_{1i}, \hat{\boldsymbol{W}}_{2i}^{[b]}, Y_i) : i = 1, \ldots, n\}$.

4. Calculate the p-value as $p^{(m)} = B^{-1} \sum_{b=1}^{B} I(\Lambda_m^{[b]} > \Lambda_m)$. For a prespecified significance level α, reject $H_0^{(m)}$ in (3.7.1) if $p^{(m)} < \alpha$.

A similar permutation test approach can be implemented for EA-SAVE, EA-DR, and their corresponding marginal expectile-assisted methods. When we combine the projection step in the algorithm from Section 3.3 with the permutation step 3 above, the computation becomes very time-consuming. The marginal expectile-assisted methods in Section 5.2 are computationally more efficient as no projection is needed.

Denote $\hat{d}$ as the estimated structural dimension. We report the frequency of $\hat{d}$ in Table 3.9 based on 100 repetitions. We fix $p = 6$ and set $n = 100$ or 300. Directional regression is used for the asymptotic test, and marginal EA-DR with $B = 200$ permutations is used for the permutation test. The significance level is set as $\alpha = 0.1$. The true structural dimension is $d = 2$ for model I and $d = 1$ for model V. As n increases, both the asymptotic test and the permutation test lead to

Model	n	Asymptotic				Permutation			
		$\hat{d}=0$	$\hat{d}=1$	$\hat{d}=2$	$\hat{d}>2$	$\hat{d}=0$	$\hat{d}=1$	$\hat{d}=2$	$\hat{d}>2$
I	100	0	28	59	13	0	17	71	12
	300	0	0	96	4	0	0	89	11
V	100	8	82	9	1	55	38	6	1
	300	0	89	10	1	49	43	5	3

Table 3.9: Sequential test for order determination with $\alpha = 0.1$. The frequency of $\hat{d}$ is reported based on 100 repetitions for $p = 6$.

higher frequency of correct estimation. While both tests work well for model I, the asymptotic test is better than the permutation test for model V.

3.8 Analysis of the Big Mac data

The Big Mac data contains 10 economic variables from 45 cities around the world in 1991. The data can be downloaded at http://www.stat.umn.edu/RegGraph/data/Big-Mac.lsp. The response Y is the minutes of labor needed to buy a Big Mac. The detailed description of the predictors can be found at the above website. Following the discussions in Li (2008) (page 92), we apply the optimal Box-Cox transformation (Box and Cox, 1964) before we compare different dimension reduction methods. Denote the predictors after the Box-Cox transformation as $\boldsymbol{X} = (X_1, \ldots, X_9)^\top$. As suggested in Li (2018) (page 139, Table 9.1), we use structural dimension $d = 1$ for this data.

First, we use leave-one-out cross validation to compare the performances of six methods: SIR, DR, EA-SIR, EA-DR, QUME and QUVE. Consider the training set to be the 44 observations after removing one observation from the entire data. For a given dimension reduction method, denote $\hat{\boldsymbol{\beta}}^{(-i)} \in \mathbb{R}^9$ as the central space estimator based on the i-th training set, where the i-th observation $(\boldsymbol{X}_i, Y_i)$ is removed, $i = 1, \ldots, 45$. For a fixed expectile level τ, we fit the kernel expectile regression between Y and $\boldsymbol{X}^\top \hat{\boldsymbol{\beta}}^{(-i)}$ for the i-th training set, and denote the estimated τ-th conditional expectile function as $\hat{f}_\tau^{(-i)}$. Then we evaluate $\hat{f}_\tau^{(-i)}$ at $\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}^{(-i)}$ and compare it to Y_i

τ	SIR		EA-SIR		DR		EA-DR		QUME	QUVE
	$H=2$	$H=4$	$H=2$	$H=4$	$H=2$	$H=4$	$H=2$	$H=4$		
0.2	1165	831	615	711	1294	1330	1028	951	2009	1332
0.5	858	697	492	592	1073	1122	850	780	1776	1035
0.8	1175	994	699	864	1523	1591	1165	1073	2529	1444

Table 3.10: Big Mac data. Average asymmetric least squares loss δ_τ with leave-one-out cross validation is reported.

through the asymmetric least squares loss function $\phi_\tau\{Y_i - \hat{f}_\tau^{(-i)}(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}^{(-i)})\}$, where ϕ_τ is defined in (3.2.2). Repeat this process for all i, and the average asymmetric least squares loss is defined as

$$\delta_\tau = \frac{1}{45}\sum_{i=1}^{45} \phi_\tau\{Y_i - \hat{f}_\tau^{(-i)}(\boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}^{(-i)})\}.$$

For SIR, DR, EA-SIR and EA-DR, we use $H = 2$ or $H = 4$ to estimate $\hat{\boldsymbol{\beta}}^{(-i)}$. Fix τ to be 0.2, 0.5 and 0.8, and the δ_τ values are summarized in Table 3.10. We see that the expectile-assisted methods improve over their classical counterparts. For each fixed τ, quantile-based methods have the largest δ_τ values, and EA-SIR with 2 slices consistently has the smallest δ_τ.

Next, we use EA-SIR with 2 slices to perform dimension reduction based on the full data set. Denote the resulting central space estimator as $\hat{\boldsymbol{\beta}}$. For a fixed expectile level τ, we fit the kernel expectile regression between Y and $\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}$ based on the entire data, and denote the estimated τ-th conditional expectile function as $\hat{f}_\tau$. For $\tau = 0.2$, 0.5 and 0.8, we plot $\hat{f}_\tau$ versus $\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}$ in Figure 3.2. We see a monotone trend between Y and $\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}$, which explains the superior performance of EA-SIR over EA-DR in Table 3.10. As we have seen in models IV and V of the simulation studies, EA-SIR can improve over SIR in the presence of heteroscedasticity, which can be clearly seen from the estimated conditional expectile functions.

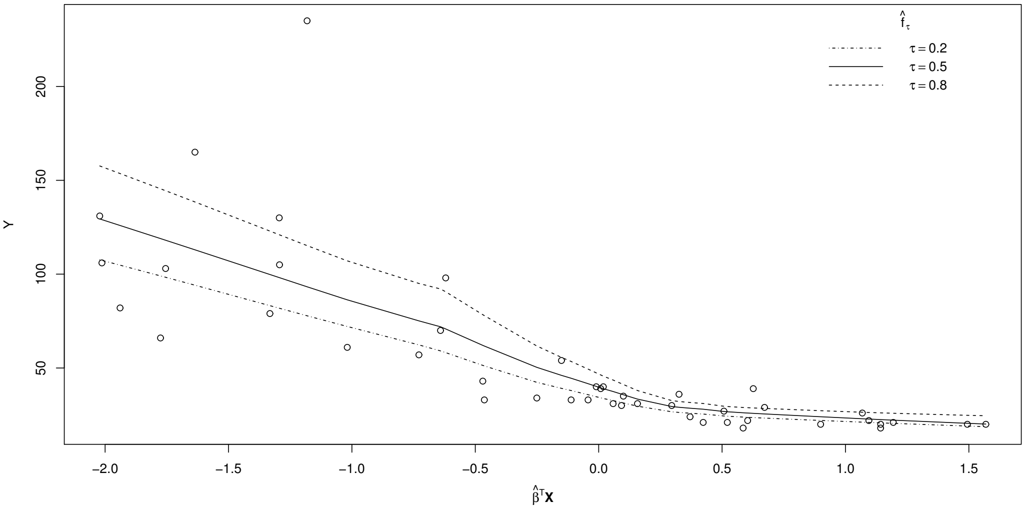

Figure 3.2: Scatter plot of the response Y and the first sufficient direction $\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}$ with fitted $\hat{f}_\tau$ for $\tau = 0.2$, 0.5 and 0.8 respectively.

3.9 Proofs

Proof of Proposition 1. Let $\boldsymbol{B}$ be a basis of $\mathcal{S}_{Y|\boldsymbol{X}}$. Then we have

$$E\{\phi_\tau(Y - a)|\boldsymbol{X} = \boldsymbol{x}\} = E\{\phi_\tau(Y - a)|\boldsymbol{B}^\top \boldsymbol{X} = \boldsymbol{B}^\top \boldsymbol{x}\}$$

because $Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^\top \boldsymbol{X}$. By the definition in (3.2.1), $f_\tau(\boldsymbol{X})$ evaluated at $\boldsymbol{x}$ becomes

$$f_\tau(\boldsymbol{x}) = \arg\min_a E\{\phi_\tau(Y - a)|\boldsymbol{B}^\top \boldsymbol{X} = \boldsymbol{B}^\top \boldsymbol{x}\}.$$

This implies that $f_\tau(\boldsymbol{X})$ is a function of $\boldsymbol{B}^\top \boldsymbol{X}$ and $f_\tau(\boldsymbol{X}) \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^\top \boldsymbol{X}$ for any fixed τ. It follows that $\boldsymbol{\xi}_{\boldsymbol{X}} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^\top \boldsymbol{X}$. By the definition of the central space, we have $\mathcal{S}_{\boldsymbol{\xi}_{\boldsymbol{X}}|\boldsymbol{X}} \subseteq \text{span}(\boldsymbol{B}) = \mathcal{S}_{Y|\boldsymbol{X}}$. □

Let $\boldsymbol{t} \in \mathbb{R}^k$ be a realization of $\boldsymbol{T}$. Denote $\boldsymbol{\mu}_h(\boldsymbol{t}) = E\{\boldsymbol{Z}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t} \in J_h(\boldsymbol{t})\}$. The following Lemma is needed before we prove Proposition 2.

Lemma 2. Under Assumption 3.2.1, we have $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_h(\boldsymbol{t}) \in \mathcal{S}_{Y|\boldsymbol{X}}$.

Proof of Lemma 2. Without loss of generality, assume $E(\boldsymbol{X}) = \boldsymbol{0}$. Let $\boldsymbol{B}$ be a basis of $\mathcal{S}_{Y|\boldsymbol{X}}$. The LCM assumption implies that

$$E(\boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X}) = \boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B})\boldsymbol{X}, \tag{3.9.1}$$

where $\boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B}) = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^\top$. From the proof of Proposition 1, we have $\boldsymbol{\xi}_{\boldsymbol{X}} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X}$, which implies that

$$\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X}. \tag{3.9.2}$$

Hence we have

$$E(\boldsymbol{X}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t}) = E\{E(\boldsymbol{X}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t}, \boldsymbol{B}^\top\boldsymbol{X})|\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t}\} = E\{E(\boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X})|\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t}\}, \tag{3.9.3}$$

where the first equality is due to the law of iterative expectations, and the second equality is guaranteed by (3.9.2). Plug (3.9.1) into (3.9.3) and we get

$$E(\boldsymbol{X}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t}) = \boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B})E(\boldsymbol{X}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t}).$$

It follows that

$$\boldsymbol{\Sigma}^{-1}E(\boldsymbol{X}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t}) = \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^\top E(\boldsymbol{X}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top\boldsymbol{t}) \subseteq \mathrm{span}(\boldsymbol{B}) = \mathcal{S}_{Y|\boldsymbol{X}}.$$

Together with the fact that $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$ and the definition of $\boldsymbol{\mu}_h(\boldsymbol{t})$, we have $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_h(\boldsymbol{t}) \in \mathcal{S}_{Y|\boldsymbol{X}}$. □

Proof of Proposition 2. Let $\boldsymbol{t} \in \mathbb{R}^k$ be a realization of $\boldsymbol{T}$. Plug $\boldsymbol{T} = \boldsymbol{t}$ into (3.2.3) and we get $\boldsymbol{M}(\boldsymbol{t}) = \sum_{h=1}^{H} p_h(\boldsymbol{t})\boldsymbol{\mu}_h(\boldsymbol{t})\boldsymbol{\mu}_h^\top(\boldsymbol{t})$. We know from Lemma 2 that $\mathrm{span}\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}(\boldsymbol{t})\} \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$.

Let $\boldsymbol{\omega} \in \mathbb{R}^p$ belong to the orthogonal space of $\mathcal{S}_{Y|\boldsymbol{X}}$. Then it must also belong to the orthogonal space of span$\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}(\boldsymbol{t})\}$. Thus we have $\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}(\boldsymbol{t}) = \mathbf{0}$ for all $\boldsymbol{t}$. It follows that

$$\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1/2}E\{\boldsymbol{M}(\boldsymbol{T})\} = E\{\boldsymbol{\omega}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}(\boldsymbol{T})\} = \mathbf{0}. \tag{3.9.4}$$

Note that (3.9.4) holds for any $\boldsymbol{\omega}$ orthogonal to $\mathcal{S}_{Y|\boldsymbol{X}}$. We have shown that

$$\mathcal{S}^\perp_{Y|\boldsymbol{X}} \subseteq \left\{\text{span}\big(\boldsymbol{\Sigma}^{-1/2}E\{\boldsymbol{M}(\boldsymbol{T})\}\big)\right\}^\perp, \tag{3.9.5}$$

where $\perp$ denotes the orthogonal space. The conclusion follows from (3.9.5). □

Denote $\boldsymbol{V}_h(\boldsymbol{t}) = E\{\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{I}_p|\boldsymbol{\xi}_{\mathbf{X}}^\top\boldsymbol{t} \in J_h(\boldsymbol{t})\}$, where $\boldsymbol{t} \in \mathbb{R}^k$ be a realization of $\boldsymbol{T}$. The following Lemma is needed before we prove Proposition 3 and Proposition 4.

Lemma 3. Under Assumptions 3.2.1 and 3.4.1, we have span$\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}_h(\boldsymbol{t})\} \in \mathcal{S}_{Y|\boldsymbol{X}}$.

Proof of Lemma 3. Without loss of generality, assume $E(\boldsymbol{X}) = \mathbf{0}$. Let $\boldsymbol{B}$ be a basis of $\mathcal{S}_{Y|\boldsymbol{X}}$. The CCV assumption and the EV-VE formula lead to

$$\text{Var}(\boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X}) = E\{\text{Var}(\boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X})\} = \boldsymbol{\Sigma} - \text{Var}\{E(\boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X})\}. \tag{3.9.6}$$

Recall from (3.9.1) that $E(\boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X}) = \boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B})\boldsymbol{X}$ under the LCM assumption. Together with (3.9.6), we have

$$E(\boldsymbol{X}\boldsymbol{X}^\top|\boldsymbol{B}^\top\boldsymbol{X}) - \boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B})\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{P}^\top_{\boldsymbol{\Sigma}}(\boldsymbol{B}) = \boldsymbol{\Sigma} - \boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B})\boldsymbol{\Sigma}\boldsymbol{P}^\top_{\boldsymbol{\Sigma}}(\boldsymbol{B}). \tag{3.9.7}$$

After rearranging the terms of (3.9.7), we have

$$E\{(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})|\boldsymbol{B}^\top\boldsymbol{X}\} = \boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})\boldsymbol{P}^\top_{\boldsymbol{\Sigma}}(\boldsymbol{B}). \tag{3.9.8}$$

Note that

$$\begin{aligned} E\{(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}\} &= E[E\{(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}, \boldsymbol{B}^\top \boldsymbol{X}\}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}] \\ &= E[E\{(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})|\boldsymbol{B}^\top \boldsymbol{X}\}|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}], \end{aligned} \tag{3.9.9}$$

where the first equality is due to the law of iterative expectations, and the second equality is guaranteed by (3.9.2). Plug (3.9.8) into (3.9.9) and we get

$$E\{(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}\} = \boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B}) E\{(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}\} \boldsymbol{P}_{\boldsymbol{\Sigma}}^\top(\boldsymbol{B}). \tag{3.9.10}$$

Recall that $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$ and $\boldsymbol{P}_{\boldsymbol{\Sigma}}(\boldsymbol{B}) = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^\top$. It follows from (3.9.10) that

$$\begin{aligned} &\boldsymbol{\Sigma}^{-1/2} E\{(\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{I}_p)|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}\}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1} E\{(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}\}\boldsymbol{\Sigma}^{-1} \\ &= \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^\top E\{(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\Sigma})|\boldsymbol{\xi}_{\boldsymbol{X}}^\top \boldsymbol{t}\}\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^\top \\ &\subseteq \operatorname{span}(\boldsymbol{B}) = \mathcal{S}_{Y|\boldsymbol{X}}. \end{aligned}$$

Together with the definition of $\boldsymbol{V}_h(\boldsymbol{t})$, we have shown $\operatorname{span}\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V}_h(\boldsymbol{t})\} \in \mathcal{S}_{Y|\boldsymbol{X}}$. □

Proof of Proposition 3. Let $\boldsymbol{G}(\boldsymbol{t})$ be a realization of $\boldsymbol{G}(\boldsymbol{T})$. From Lemma 2, Lemma 3 and the definition of $\boldsymbol{G}(\boldsymbol{T})$, we have $\operatorname{span}\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{G}(\boldsymbol{t})\} \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$. The rest of the proof is exactly parallel to the proof of Proposition 2, and is thus omitted. □

Proof of Proposition 4. Let $\boldsymbol{F}(\boldsymbol{t})$ be a realization of $\boldsymbol{F}(\boldsymbol{T})$. From Lemma 2, Lemma 3 and the definition of $\boldsymbol{F}(\boldsymbol{T})$, we have $\operatorname{span}\{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{F}(\boldsymbol{t})\} \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$. The rest of the proof is exactly parallel to the proof of Proposition 2, and is thus omitted. □

Proof of Proposition 5. Let $\boldsymbol{B}$ be a basis of $\mathcal{S}_{Y|\boldsymbol{X}}$. From the proof of Proposition 1, we have $f_{\tau_\ell}(\boldsymbol{X}) \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X}$. Similar to the proof of Lemma 2, we can show that

$$E\{\boldsymbol{X}|f_{\tau_\ell}(\boldsymbol{X})\} = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^\top E\{\boldsymbol{X}|f_{\tau_\ell}(\boldsymbol{X})\}.$$

Thus we have $\boldsymbol{\Sigma}^{-1}E\{\boldsymbol{X}|f_{\tau_\ell}(\boldsymbol{X})\} \in \text{span}(\boldsymbol{B}) = \mathcal{S}_{Y|\boldsymbol{X}}$. From the definition of $\boldsymbol{M}_\ell$, it follows that $\text{span}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}_\ell) \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$ for $\ell = 1, \ldots, k$. Hence we have $\text{span}(\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{M}}) = \text{span}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}_1, \ldots, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}_k) \subseteq \mathcal{S}_{Y|\boldsymbol{X}}$. □

CHAPTER 4

CONCLUSION

In Chapter 2, we propose principal asymmetric least squares (PALS) as a new sufficient dimension reduction method. PALS utilizes information across different quantiles levels, which makes it very useful in the presence of heteroscedastic error. Like other linear SDR methods such as slice inverse regression, linear PALS requires the linear conditional mean (LCM) assumption. In a situation where this condition cannot be satisfied, we propose the nonlinear PALS which uses the kernel trick to bypass the linearity requirement. Both the linear and nonlinear PALS show superior performance in terms of estimation accuracy and computation time when compared to existing SDR methods in the analysis of synthetic data and Boston Housing data.

A handicap of PALS is that it is not designed to handle a symmetric link function between the response and the predictors. This motivated the development of the expectile-assisted inverse regression estimation (EA-IRE) in Chapter 3. EA-IRE provides a unified framework for moment-based inverse regression such as sliced inverse regression, which may not work well in the presence of heteroscedasticity. We propose to first estimate the expectiles through a kernel expectile regression, and then carry out dimension reduction based on random projections of the regression expectiles. Several popular inverse regression methods in the literature are extended under this general framework. The proposed expectile-assisted methods outperform existing moment-based dimension reduction methods in both numerical studies and an analysis of the Big Mac data.

Lastly, a common limitation of both principal asymmetric least squares and the expectile-assisted extensions is that the response and the predictors be continuous. The restriction on the predictors is due to the fact that both the linear PALS and the expectile-assisted methods require the linearity assumption. This assumption is shown to be satisfied when the distribution of the predictors follows an elliptically contoured distribution, which puts an additional constraint on the predictors. This handicap is well-known in the sufficient dimension reduction literature and not unique to our method.

BIBLIOGRAPHY

[1] Abdous, B. and Remillard, B. (1995). Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics*, ***47***, 371–384.

[2] Artemiou, A. and Dong, Y. (2016). Sufficient dimension reduction via principal ℓ-q support vector machine. *Electronic Journal of Statistics*, ***10***, 783–805.

[3] Barreda, L., Gannoun, A. and Saracco, J. (2007). Some extensions of multivariate sliced inverse regression. *Journal of Statistical Computation and Simulation*, ***77***, 1–17.

[4] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, ***26***, 211–252.

[5] Cai, Z., Li, R. and Zhu, L. (2020). Online sufficient dimension reduction through sliced inverse regression. *Journal of Machine Learning Research*, ***21(10)***, 1–25.

[6] Chavent, M., Girard, S., Kuentz-Simonet, V., Liquet, B., Nguyen, T. and Saracco, J. (2014). A sliced inverse regression approach for data stream. *Computational Statistics*, ***29***, 1129–1152.

[7] Chen, J. (2018). On exactitude in financial regulation: value-at-risk, expected shortfall, and expectiles. *Risks*, ***6***, 61.

[8] Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, ***38***, 3696–3723.

[9] Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

[10] Cook, R. D. (1998a). *Regression graphics: ideas for studying regressions through graphics*. New York: Wiley.

[11] Cook, R. D. (1998b). Principal Hessian directions revisited. *Journal of the American Statistical Association*, ***93***, 84–94.

[12] Cook, R. D. (2007). Fisher lecture: dimension reduction in regression. *Statistical Science*, ***22***, 1–26.

[13] Cook, R. D. and Setodji, M. (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association*, ***98***, 340–351.

[14] Cook, R. D. and Weisberg, S. (1991). Comment on "Sliced inverse regression for dimension reduction". *Journal of American Statistical Association*, ***86***, 28–33.

[15] Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian and New Zealand Journal of Statistics*, ***43***, 147–199.

[16] Coudret, R., Girard, S. and Saracco, J. (2014). A new sliced inverse regression method for multivariate response. Computational Statistics and Data Analysis, ***77***, 285–299.

[17] Daouia, A., Girard, S. and Stupfler, G. (2018). Estimation of tail risk based on extreme expectiles. *Journal of the Royal Statistical Society: Series B*, ***80***, 263–292.

[18] Dong, Y. (2020). A brief review of linear sufficient dimension reduction through optimization. To appear in *Journal of Statistical Planning and Inference.*

[19] Guo, Z., Li, L., Lu, W. and Li, B. (2015). Groupwise dimension reduction via envelope method. *Journal of the American Statistical Association*, ***110***, 1515–1527.

[20] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, ***5***, 81–102.

[21] Kim, H., Wu, Y. and Shin, S. J. (2019). Quantile-slicing estimation for dimension reduction in regression. *Journal of Statistical Planning and Inference*, ***198***, 1–12.

[22] Kim, K., Li, B., Yu, Z. and Li, L. (2020). On post dimension reduction statistical inference. To appear in *The Annals of Statistics.*

[23] Kim, M. and Lee, S. (2016). Nonlinear expectile regression with application to value-at-risk and expected shortfall estimation. *Computational Statistics and Data Analysis*, ***94***, 1–19.

[24] Lee, K. Y., Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: formulation and estimation. *The Annals of Statistics*, ***41***, 221–249.

[25] Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R.* CRC Press.

[26] Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, ***39***, 3182–3210.

[27] Li, B., Kim, M. K. and Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics*, ***38***, 1094–1121.

[28] Li, B. and Song, J. (2017). Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, ***45***, 1059–1095.

[29] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of American Statistical Association*, ***479***, 997–1008.

[30] Li, B., Wen, S. and Zhu, L. X. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of American Statistical Association*, ***103***, 1177–1186.

[31] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, ***86***, 316–342.

[32] Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, ***87***, 1025–1039.

[33] Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, ***94***, 603–613.

[34] Li, L., Li, B. and Zhu, L. X. (2010). Groupwise dimension reduction. *Journal of American Statistical Association*, ***105***, 1188–1201.

[35] Li, Y. and Zhu, L. X. (2007). Asymptotics for sliced inverse variance estimation. *The Annals of Statistics*, ***35***, 41–69.

[36] Luo, W., Li, B. and Yin, X. (2014). On efficient dimension reduction with respect to a statistical functional of interest. *The Annals of Statistics*, ***42***, 382–412.

[37] Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, ***107***, 168–179.

[38] Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistics Review*, ***81***, 134–150.

[39] Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, ***55***, 819–847.

[40] Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis*, ***96***, 117–135.

[41] Shin, S. J. and Artemiou A. (2017). Penalized principal logistic regression for sparse sufficient dimension reduction. *Computational Statistics and Data Analysis*, ***111***, 48–58.

[42] Shin, S. J., Wu, Y., Zhang, H. and Liu, Y. (2017). Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, ***104***, 67–81.

[43] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, ***35***, 2769–2794.

[44] Tan, K., Shi, L. and Yu, Z. (2020). Sparse SIR: optimal rates and adaptive estimation. *The Annals of Statistics*, ***48***, 64–85.

[45] Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

[46] Wang, C., Shin, S. J. and Wu, Y. (2018). Principal quantile regression for sufficient dimension reduction with heteroscedasticity. *Electronic Journal of Statistics*, ***12***, 2114–2140.

[47] Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: sparse MAVE. *Computational Statistics and Data Analysis*, ***52***, 4512–4520.

[48] Wu, H. M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, ***17***, 590–610.

[49] Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace. *Journal of the Royal Statistical Society: Series B*, ***64***, 363–410.

[50] Yang, Y., Zhang, T. and Zou, H. (2018). Flexible expectile regression in reproducing kernel Hilbert spaces. *Technometrics*, ***60***, 26–35.

[51] Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: a nonparametric approach. *Journal of Nonparametric Statistics*, ***6***, 273–292.

[52] Yin, X. and Bura, E. (2006). Moment-based dimension reduction for multivariate response regression. *Journal of Statistical Planning and Inference*, ***136***, 3675–3688.

[53] Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, ***90***, 113–125.

[54] Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, ***39***, 3392–3416.

[55] Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, ***99***, 1733–1757.

[56] Yu, Z., Zhu, L., Peng, H. and Zhu, L. X. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika*, ***100***, 641–654.

[57] Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, ***5***, 727–736.