The Pennsylvania State University
The Graduate School

DIMENSION REDUCTION FOR NON-ELLIPTICALLY DISTRIBUTED PREDICTORS

A Dissertation in
Statistics
by
Yuexiao Dong

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2009

The dissertation of Yuexiao Dong was reviewed and approved* by the following:

Bing Li
Professor of Statistics, Chair of Committee
Dissertation Advisor

Michael G. Akritas
Professor of Statistics

Runze Li
Professor of Statistics, Graduate Program Chair

Dennis K.J. Lin
Distinguished Professor of Supply Chain and Statistics

Bruce G. Lindsay
Willaman Professor of Statistics, Department Head

*Signatures are on file in the Graduate School.

Abstract

Many classical dimension reduction methods require the predictors to have elliptical distributions, or at least to satisfy a linearity condition. In this dissertation we reformulate the commonly used dimension reduction methods to circumvent the requirement of such strong assumptions, while at the same time preserve the desirable properties of the classical methods, such as $\sqrt{n}$-consistency and asymptotic normality. The notion of "central solution space" is introduced first. We then use it to modify essentially all inverse conditional moment based methods under a general framework. Imposing elliptical distributions or even stronger assumptions on predictors is often considered as the necessary tradeoff for overcoming the "curse of dimensionality", but the development of this dissertation shows this need not be the case.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Life is a journey. I am very happy and very thankful that my five year's journey at Penn State has been accompanied by a great advisor, Professor Bing Li. It is very difficult to overstate my gratitude towards him. Not only did he teach me how to work as a researcher, but more importantly, he also taught me how to think as a researcher. Professor Bing Li has become and will always be my role model, as a scholar and as a man.

I am also blessed to have such a wonderful thesis committee. Professor Michael Akritas, Professor Runze Li and Professor Dennis Lin have all been bothered by me constantly, yet they provide me with nothing but encouragement and all kinds of sound suggestions about both my study and my career.

I also want to thank my family. You are the ones that give me strength from the deepest in my heart.

Chapter 1

Introduction

1.1 Central space

Regression or classification problems with high-dimensional predictors are increasingly prevalent in contemporary applications. An important problem that arises from these applications is how to avoid smoothing over high-dimensional space, which seriously hinders the accuracy of statistical inference — a phenomenon commonly known as the curse of dimensionality (Bellman, 1961). Sufficient dimension reduction estimates, especially those based on inverse conditional moments, can avoid the curse of dimensionality because they involve only 1-dimensional smoothing.

Suppose that X is a p-dimensional random vector representing the predictor, and Y is a random variable representing the response. Dimension reduction (K. C. Li, 1991, 1992; Cook & Weisberg, 1991; Cook, 1998b) is aimed at replacing X by its lower dimensional linear combination without losing any information for the regression of Y on X. That is, we seek a matrix β of dimension $p \times d$, with $d < p$, such that Y is independent of X given $\beta^T X$. We write this conditional independence as

$$Y \perp\!\!\!\perp X | \beta^T X. \tag{1.1}$$

Note that, if A is any $d \times d$ non-singular matrix, then $\beta^T X$ and $A^T \beta^T X$ has one-to-one correspondence. Therefore, $Y \perp\!\!\!\perp X|\beta^T X$ if and only if $Y \perp\!\!\!\perp X|(\beta A)^T X$. Note that β and βA have the same column space. Thus we define the column space of β as a dimension reduction subspace (DRS; Cook, 1994, 1996). For a matrix A, we denote by $\mathcal{S}(A)$ the subspace spanned by the columns of A.

Definition 1.1.1. *The central space (CS) for (X, Y) is the intersection of all dimension reduction spaces for (X, Y). This space is written as $\mathcal{S}_{Y|X}$.*

The goal of dimension reduction is to find the central space $\mathcal{S}_{Y|X}$. $\mathcal{S}_{Y|X}$ is also called central subspace, sufficient dimension reduction (SDR) central space or effective dimension reduction (e.d.r.) central space. Without ambiguity, we will just call $\mathcal{S}_{Y|X}$ central space in this thesis. The dimension d of $\mathcal{S}_{Y|X}$ is called the *structure dimension*. Determining the structure dimension is another important topic in dimension reduction. In most part of this thesis, we will focus on recovering the central space assuming structure dimension d is known.

The central space has the following invariance property (Cook, 1998b).

Theorem 1.1.1. *Let $\mathcal{S}_{Y|X}$ be the central space for (X, Y). Let $Z = AX + b$, where A is a $p \times p$ non-singular matrix and $b \in \mathbb{R}^p$,. Then*

$$\mathcal{S}_{Y|Z} = A^{-T} \mathcal{S}_{Y|X}.$$

This invariance property implies that we can always standardize X to Z first, because we can transform $\mathcal{S}_{Y|Z}$ to $\mathcal{S}_{Y|X}$ easily. Thus we can make the following assumption for the sake of simplicity. We will refer to it as the normalizing assumption in later sections.

Assumption 1.1.1. *We assume that*

$$E(X) = 0, Var(X) = I_p.$$

1.2 Inverse conditional moments methods

Methods based on moments or inverse conditional moments are among the most commonly used dimension reduction techniques. A singular advantage of these methods is that, unlike other nonparametric estimates, they do not require multi-dimensional smoothing regardless of the dimension of X, and regardless of the relation between X and Y. They can be divided further into two categories. The first category depends on moments or inverse conditional moments such as

$$E(XY^k),\ E(X|Y),$$

where k is an integer. Methods in this category includes Ordinary Least Squares (OLS; K. C. Li & Duan, 1989), Sliced Inverse Regression (SIR; K. C. Li, 1991), Parametric Inverse Regression (Bura & Cook, 2001), Canonical Correlation (Fung, He, Liu, & Shi, 2002), and Kernel Inverse Regression (L. X. Zhu & Fang, 1996; Ferre & Yao, 2005). Because the moments involved here are only linear functions of X, we will refer to this category as *the first-order methods*.

The second category depends on moments or inverse conditional moments such as

$$E(XY^k),\ E(X|Y),\ E(Y^kXX^T),\ \text{and}\ E(XX^T|Y),$$

where k is an integer. This category includes Principal Hessian Directions (PHD; K. C. Li, 1992 and Cook, 1998a), Sliced Average Variance Estimator (SAVE; Cook & Weisberg, 1991), SIRII (K. C. Li, 1991), Contour Regression (B. Li, Zha, & Chiaromonte, 2005), and Directional Regression (DR; B. Li & Wang, 2007). Because the moments involved here include both linear and quadratic functions of X, we will refer to this category as *the second-order methods.*

It is well known that methods based on inverse conditional moments impose strong assumptions on the predictors. Suppose β is a basis for the central space $\mathcal{S}_{Y|X}$, or $\text{span}\beta = \mathcal{S}_{Y|X}$. The first-order methods then require the *linear conditional mean* assumption. That is,

Assumption 1.2.1. *Let β be a $\mathbb{R}^{p\times d}$ matrix whose columns form a basis in $\mathcal{S}_{Y|X}$. We will assume that $E(X|\beta^T X)$ is a linear function of X; that is, the conditional mean of X given $\beta^T X$ is linear in X.*

The second-order methods require, in addition, the *constant conditional variance* assumption.

Assumption 1.2.2. *Let β be a $\mathbb{R}^{p\times d}$ matrix whose columns form a basis in $\mathcal{S}_{Y|X}$. We assume that the conditional variance*

$$\text{Var}(X|\beta^T X)$$

is a non-random matrix.

1.3 Review of first-order methods

In this section, we will provide a brief review of several common first-order methods in dimension reduction literature.

1.3.1 Ordinary Least Squares

Ordinary Least Squares proposed by K. C. Li and Duan (1989) appears to be the first dimension reduction method. It is based on the following fact.

Theorem 1.3.1. *If Assumption 1.1.1 and 1.2.1 hold, then*

$$E(XY) \in \mathcal{S}_{Y|X}.$$

The geometric implication of the linear conditional mean Assumption 1.2.1 is that the L-2 projection and the Euclidean projection are the same. Suppose $\beta = (\beta_1, ..., \beta_d)$ is a basis of the central space, or span$\beta = \mathcal{S}_{Y|X}$. Let $P_\beta = \beta(\beta^T\beta)^{-1}\beta^T$ be the projection matrix on to spanβ. The key to the proof of Theorem 1.3.1 is that the conditional expectation $E(X|\beta^T X)$ coincides with $P_\beta X$.

Proof. Note that

$$E(XY) = E[E(XY|X)] = E[XE(Y|X)]. \tag{1.2}$$

However, because $Y \perp\!\!\!\perp X|\beta^T X$, we have,

$$E(Y|X) = E(Y|X, \beta^T X) = E(Y|\beta^T X).$$

Therefore, the right hand side of (1.2) is $E[XE(Y|\beta^T X)]$. However, under Assumption 1.2.1, we have

$$E[E(X|\beta^T X)Y] = E[(P_\beta X)Y] = P_\beta E(XY).$$

Thus

$$E(XY) = P_\beta E(XY).$$

In other words, $E(XY)$ equals its projection onto $\mathcal{S}(\beta)$. Therefore $E(XY) \in \mathcal{S}_{Y|X}$. $\square$

Let $W_1, \ldots, W_n$ be independent copies of (X, Y). We write $E_n(W) = n^{-1} \sum_{i=1}^n W_i$. By the same token, we define $\text{cov}_n(W, V) = E_n[(W - E_n(W))(V - E_n(V))^T]$. Based on the population level result, we take the following steps to estimate the central space at the sample level.

1. Compute $\hat{\Sigma} = \text{Var}_n(X)$, $\hat{\mu} = E_n(X)$. Standardize $X_1, \ldots, X_n$ to be $\hat{Z}_i = \hat{\Sigma}^{-1/2}(X_i - \hat{\mu})$.

2. Center $Y_1, \ldots, Y_n$ to $\hat{Y}_i = Y_i - E_n(Y)$.

3. Let $\hat{\gamma}$ be the vector $E_n(\hat{Z}\hat{Y})$. This is an estimator of $E(ZY)$, a vector in $\mathcal{S}_{Y|Z}$.

4. Let $\hat{\beta} = \hat{\Sigma}^{-1/2}\hat{\gamma}$, and this is an estimator of $\mathcal{S}_{Y|X}$.

It is well known that the OLS estimator for $\mathcal{S}_{Y|X}$ is $\sqrt{n}$-consistent. In other words, it converges at $\sqrt{n}$-rate to a vector that belongs to $\mathcal{S}_{Y|X}$. The major limitation of OLS is that it can estimate at most one direction in the central space. OLS can not provide a comprehensive estimation of the central space if its dimension is greater than 1.

1.3.2 Sliced Inverse Regression

Sliced Inverse Regression is the most common dimension reduction technique nowadays. It is easy to implement and have nice asymptotic properties.

Theorem 1.3.2. *If Assumption 1.1.1 and 1.2.1 hold, then for any y,*

$$E(X|Y = y) \in \mathcal{S}_{Y|X}.$$

PROOF. Let β be the $p \times d$ matrix whose columns form a basis of $\mathcal{S}_{Y|X}$. Then

$$E(X|Y) = E[E(X|Y, \beta^T X)|Y].$$

Because $Y \perp\!\!\!\perp X|\beta^T X$, we have that

$$E[E(X|Y, \beta^T X)|Y] = E[E(X|\beta^T X)|Y].$$

However, under Assumption 1.2.1, we have

$$E[E(X|\beta^T X)|Y] = E[(P_\beta X)|Y] = P_\beta E(X|Y).$$

In other words $E(X|Y)$ belongs to $\mathcal{S}(\beta)$, which is the central space. $\square$

From $E(X|Y) = P_\beta E(X|Y)$, we can easily derive that $\mathrm{Var}[E(X|Y = y)] = P_\beta \mathrm{Var}[E(X|Y = y)]P_\beta$. Thus a natural corollary of Theorem 1.3.2 is

Corollary 1.3.1. *If Assumption 1.1.1 and 1.2.1 hold, then the column space of the matrix*

$$\mathrm{Var}[E(X|Y = y)]$$

is a subspace of the central space.

In practice, we will use the discretized version of the above result. Let $I_1, ..., I_k$ be

k intervals that partition Υ, the space of Y. Let $\widetilde{Y}$ be the discretized Y, defined by

$$\widetilde{Y} = i, \text{ if } Y \in I_i, i = 1, ..., k.$$

We will use the sample version of the matrix $S = \text{Var}[E(X|\widetilde{Y})]$ to estimate the central space. This matrix can be written as

$$\sum_{i=1}^{k} Pr(\widetilde{Y} = i)E(X|\widetilde{Y} = i)E(X^T|\widetilde{Y} = i).$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the sample and we have the following algorithm.

1. Standardize X_i to $\hat{Z}_i$ and centralize Y_i to $\hat{Y}_i$ as before.

2. Partition the interval $[min\{\hat{Y}_1, \ldots, \hat{Y}_n\}, max\{\hat{Y}_1, \ldots, \hat{Y}_n\}]$ into k intervals, say $I_1, \ldots, I_k$, and compute the mean of $\hat{Z}$ within each slice; that is

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j \in I_i} \hat{Z}_j,$$

where n_i is the number of $\hat{Y}$ that are within slice I_i.

3. Construct the SIR matrix
$$\hat{S} = \sum_{i=1}^{k} \frac{n_i}{n} \hat{\mu}_i \hat{\mu}_i^T.$$

4. Assuming d is known. Let $v_1, \ldots, v_d$ be the eigenvectors of $\hat{S}$ corresponding to the largest d eigenvalues. This is used to estimate $\mathcal{S}_{Y|Z}$.

5. Let $w_i = \Sigma^{-1/2} v_i, i = 1, \ldots, d$, These will be used as the estimator of $\mathcal{S}_{Y|X}$.

It is well known that the SIR estimator for $\mathcal{S}_{Y|X}$ is $\sqrt{n}$-consistent. Another nice property of SIR is that we can determine the structure dimension d of the central space $\mathcal{S}_{Y|X}$ using a sequential hypothesis testing scheme. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p$ be the eigenvalues of the matrix $n\widehat{S}$, where $\widehat{S}$ is defined in the above sample estimation algorithm, then we have the following Theorem (K. C. Li, 1991).

Theorem 1.3.3. *Suppose that*

1. $d > j + 1$.

2. *The column space of S exhausts the central space; that is span$S = \mathcal{S}_{Y|X}$.*

3. *Assumption 1.1.1 and 1.2.1 hold.*

Then, under the null hypothesis

$$H_0 : \lambda_{p-j+1} = ... = \lambda_p = 0,$$

the test statistic $\sum_{i=p-j+1}^{p} \widehat{\lambda}_i$ converges in distribution to a χ^2 distribution with $(p-j)(p-j-1)/2$ degrees of freedom.

SIR can estimate at most one direction in the central space when the response is a binary variable. Another limitation is that sometimes it can only estimate a proper subspace of the central space $S_{Y|X}$. It is well known that if the response is U-shaped, then SIR is going to miss the corresponding direction in the central space.

1.3.3 Kernel Inverse Regression

Instead of using k-slice method for estimating $\mathrm{Var}[E(X|Y)]$, Kernel Inverse Regression uses kernel method for the same purpose. Asymptotic normality and $\sqrt{n}$-consistency are established for the KIR estimators. We use the following vector notation for the ease of demonstration. Let $\mathbf{X}_j = (X_{1j}, \ldots, X_{pj})^T$, where $j = 1, \ldots, n$, be a sample of $\mathbf{X} = (X_1, \ldots, X_p)^T$. Our objective is to estimate $\mathrm{Var}[E(\mathbf{X}|Y)]$, its eigenvalues and the corresponding eigenvectors, based on an independent sample of $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. Define

$$\mathbf{R}(Y) = \{R_1(Y), \ldots, R_p(Y)\}^T = \{E(X_1|Y), \ldots, E(X_p|Y)\}^T = E(\mathbf{X}|Y),$$

$$\mathbf{g}(Y) = \{g_1(Y), \ldots, g_p(Y)\}^T = \{R_1(Y)f(Y), \ldots, R_p(Y)f(Y)\}^T,$$

and

$$\Lambda_1 = \mathrm{Var}[\mathbf{R}(Y)] = \mathrm{Var}[E(\mathbf{X}|Y)].$$

We may estimate $\mathbf{R}(Y)$ by a kernel method

$$\hat{g}_i(Y) = \frac{1}{nh}\sum_{j=1}^{n} X_{ij} K((Y - Y_j)h^{-1}),$$

$$\hat{\mathbf{g}}(Y) = \{\hat{g}_1(Y), \ldots, \hat{g}_p(Y)\}^T,$$

$$\hat{f}(Y) = \frac{1}{nh}\sum_{j=1}^{n} K((Y - Y_j)h^{-1}),$$

$$\hat{\mathbf{R}}(Y) = \hat{\mathbf{g}}(Y)/\hat{f}(Y),$$

where h is a bandwidth and $K(\cdot)$ is a kernel function.

Without loss of generality, let $E(\mathbf{X}) = \mathbf{0}$. Then an estimate of Λ_1 can be constructed as follows:

$$\Lambda_{1n} = \frac{1}{n}\sum_{j=1}^{n} \hat{\mathbf{R}}(Y_j)\hat{\mathbf{R}}^T(Y_j).$$

Under certain regularity conditions, L. X. Zhu and Fang (1996) have proved that $\sqrt{n}(\Lambda_{1n} - \Lambda_1) \Rightarrow \mathbf{H}$ as $n \to \infty$, where $\Rightarrow$ stands for convergence in distribution and $\lambda^T Vech(\mathbf{H})$ has normal distribution $N(0, \sigma_\lambda^2)$ for any $\lambda \neq 0$. This implies the $\sqrt{n}$-consistency and asymptotic normality of KIR estimators. Assuming structure dimension d is known. The eigenvectors of Λ_{1n} corresponding to the largest d eigenvalues are used to estimate $\mathcal{S}_{Y|X}$.

1.3.4 Canonical Correlation

Suppose that $\mathbf{W}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ is a sample of size n, where Y_i is real-valued and $\mathbf{X}_i \in \mathbb{R}^p$. Without loss of generality, suppose $\mathbf{X}_i$ is already centralized and let $E(\mathbf{X}_i) = \mathbf{0}$. We assume that the response variable Y is supported on a finite interval $[a, b]$. The basic idea of Canonical Correlation is to build a B-spline basis for the Y variable and find their correlations with other variables $\mathbf{X}$. Consider a partition $a = t_0 < t_1 \cdots < t_H < t_{H+1} = b$ and let $\pi(Y) \in \mathbb{R}^{H+m}$ be the set of normalized B-spline basis functions of order m associated with this partition. H is the number of internal knots. For linear splines, $m = 2$; for quadratic splines, $m = 3$; for cubic splines, $m = 4$. The minimum size of partition H should be chosen such that $H + m > d$, where d is the number of effective dimensions being sought. Let $\Pi = \{\pi(Y_1), \ldots, \pi(Y_n)\}^T$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$. To estimate $\Lambda_1 = \mathrm{Var}[E(\mathbf{X}_1|Y_1)]$, CANCOR uses the following estimator,

$$\Lambda_{1n} = n^{-1}\mathbf{X}^T\Pi(\Pi^T\Pi)^{-1}\Pi^T\mathbf{X}.$$

Here we have a spline-based estimate. The idea of using splines to estimate $\Lambda_1 = \mathrm{Var}[E(\mathbf{X}_1|Y_1)]$ was first mentioned in the discussion of K. C. Li (1991) by Kent. The relationship between SIR and canonical correlation was explored by Chen and Li (1998). CANCOR was studied in detail by Fung, He, Liu, and Shi (2002). Just like KIR described in the previous section, CANCOR is a variation of SIR. It is proved by Fung, He, Liu, and Shi (2002) that CANCOR estimators maintains the same asymptotic properties as SIR.

1.3.5 Parametric Inverse Regression

We have seen several different methods to estimate $\mathrm{Var}[E(X|Y))]$ for the purpose of finding the dimension reduction central space. SIR uses slicing; KIR uses kernel function; CANCOR uses B-spline basis. Parametric Inverse Regression suggested using parametric model to fit inverse regression of X on Y. The original paper allows Y to be a vector and proposes a multivariate linear model to fit X and Y. We are going to deal with the case that Y is a scalar.

We have a sample of size n, $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$. Let $h_1, \ldots, h_s$ be square integrable functions from Ω_Y to $\mathbb{R}$, one of which (say h_1) must be taken to be 1 if Y is not centered. Let $H(Y) = \{h_1(Y), \ldots, h_s(Y)\}^T$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$.

To estimate $\Lambda_1 = \mathrm{Var}[E(\mathbf{X}_1|Y_1)]$, PIR uses the following estimator,

$$\Lambda_{1n} = n^{-1}\mathbf{X}^T H(H^T H)^{-1} H^T \mathbf{X}.$$

In practise, $H(Y)$ is often set to be monomial functions of Y.

1.4 Review of second-order methods

In the previous section, we have seen dimension reduction methods based on first order inverse conditional moments. In this section, we will focus on methods based on second order inverse conditional moments.

1.4.1 Sliced Average Variance Estimator

Sliced Average Variance Estimator is commonly known as SAVE. It is another method of estimating the central subspace based on slicing the response Y. Instead of calculating the mean within each slice as in SIR, SAVE averages the within slice variance.

Unlike SIR, which will miss the central space when the response is symmetric, SAVE can exhaustively estimate the central space $\mathcal{S}_{Y|X}$. SAVE requires both the linear conditional mean and the constant conditional variance assumption.

Theorem 1.4.1. *If Assumption 1.1.1, 1.2.1 and 1.2.2 hold, then for any value of y, the column space of the matrix*

$$I_p - \mathrm{Var}(X|Y = y)$$

is a subspace of the central space. Consequently, the column space of the matrix

$$E[I_p - \mathrm{Var}(X|Y = y)]^2$$

is a subspace of the central space $\mathcal{S}_{Y|X}$.

Proof. Let β be the $p \times d$ matrix whose columns form a basis of $\mathcal{S}_{Y|X}$. Let $P_\beta = \beta(\beta^T\beta)^{-1}\beta^T$ be the projection matrix on to span(β), and let $Q_\beta = I_p - P_\beta$ be the projection matrix on to the orthogonal complement of span(β), where I_p is the identity matrix of dimension p. We have

$$\begin{aligned}\mathrm{Var}(X|Y) =& E[\mathrm{Var}(X|Y, \beta^T X)|Y] + \mathrm{Var}[E(X|Y, \beta^T X)|Y] \\ =& E[\mathrm{Var}(X|\beta^T X)|Y] + \mathrm{Var}[E(X|\beta^T X)|Y] \\ =& Q_\beta + P_\beta \mathrm{Var}(X|Y) P_\beta.\end{aligned}$$

Hence $I_p - \mathrm{Var}(X|Y) = P_\beta[I_p - \mathrm{Var}(X|Y)]P_\beta$. The column space of the matrix $I_p - \mathrm{Var}(X|Y)$ is thus contained in the range space of the projection operator P_β, which is $\mathcal{S}_{Y|X}$. Consequently, the column space of the matrix $E[I_p - \mathrm{Var}(X|Y)]^2$ is a subspace of $\mathcal{S}_{Y|X}$. $\square$

In practice, we will use the discretized version of the above result. The sample estimator of SAVE can be calculated using a similar algorithm to SIR as in Section 1.3.2. The only difference is in Step 3, where we calculate the SAVE matrix

$$SAVE = \sum_{i=1}^{k} \frac{n_i}{n}[I_p - \mathrm{Var}_n(\hat{Z}|\hat{Y} \in I_i)].$$

One limitation of SAVE is its inefficiency in estimating monotone trend for small to moderate sample sizes.

1.4.2 SIRII

To estimate the central space, SIRII uses the eigenvalue decomposition of the following matrix,

$$SIRII = E\{\mathrm{Var}(X|Y) - E[\mathrm{Var}(X|Y)]\}^2.$$

Theorem 1.4.2. *If Assumption 1.1.1, 1.2.1 and 1.2.2 hold for X, then the column space of SIRII is contained in $\mathcal{S}_{Y|X}$.*

Proof. We have seen before in the proof of Theorem 1.4.1 that

$$I_p - \mathrm{Var}(X|Y) = P_\beta[I_p - \mathrm{Var}(X|Y)]P_\beta.$$

Thus we have

$$\mathrm{Var}(X|Y) - E[\mathrm{Var}(X|Y)] = P_\beta\{\mathrm{Var}(X|Y) - E[\mathrm{Var}(X|Y)]\}P_\beta.$$

Consequently the column space of SIRII is contained in $\mathcal{S}_{Y|X}$. □

SIRII has close relationship with SIR and SAVE. Remember that we have the following matrices for SIR and SAVE respectively,

$$SIR = \mathrm{Var}[E(X|Y)],$$

$$SAVE = E[I_p - \mathrm{Var}(X|Y)]^2.$$

As a result, SAVE can be seen as a convex combination of SIR and SIRII:

$$SAVE = SIR^2 + SIRII.$$

1.4.3 Principle Hessian Directions

The motivation of Principal Hessian Directions is the observation that the Hessian matrix

$$H(X) = \frac{\partial^2 E(Y|X)}{\partial X \partial X^T}$$

is degenerate along any directions that are orthogonal to $\mathcal{S}_{Y|X}$. Let α be the OLS vector $E(XY)$ and let e be the residual from the simple linear regression; that is

$$e = Y - \alpha^T X.$$

There are two forms Hessian matrices. The matrix $H_1 = E(YXX^T)$ is called the y-based Hessian matrix and the matrix $H_2 = E(eXX^T)$ is called the e-based Hessian matrix. The following theorem is provided by K. C. Li (1992) and Cook (1998a).

Theorem 1.4.3. *If Assumption 1.1.1, 1.2.1 and 1.2.2 hold, then the column space of H_1 (H_2) is a subspace of $\mathcal{S}_{Y|X}$.*

PROOF. Let β be the $p \times d$ matrix whose columns form a basis in $\mathcal{S}_{Y|X}$. Then

$$\begin{aligned} E(YXX^T) =& E[E(YXX^T|X)] = E[E(Y|X)XX^T] \\ =& E[E(Y|\beta^T X)XX^T] = E[YE(XX^T|\beta^T X)]. \end{aligned}$$

By Assupmtion 1.2.1 and 1.2.2, we have

$$\begin{aligned} E(XX^T|\beta^T X) = & \mathrm{Var}(X|\beta^T X) + E(X|\beta^T X)E(X^T|\beta^T X) \\ =& Q_\beta + P_\beta XX^T P_\beta. \end{aligned}$$

Thus

$$\begin{aligned} H_1 =& E[YE(XX^T|\beta^T X)] \\ =& E[Y(Q_\beta + P_\beta XX^T P_\beta)] \\ =& E(Y)Q_\beta + P_\beta E(YXX^T)P_\beta \\ =& P_\beta H_1 P_\beta. \end{aligned}$$

This means the column space of $H_1 = E(YXX^T)$ is a subspace of the central space $\mathcal{S}_{Y|X}$. The same thing can be said about e-based PHD estimator $H_2 = E(eXX^T)$. The proof is similar and thus omitted. □

1.4.4 Contour Regression

As a dimension reduction method, Contour Regression target directly at the contour directions of the response surface. Contour directions are those along which the response has small variation; they span the orthogonal complement of the central space. They can be extracted according to two measures of variation in the response, leading to two methods: simple and general Contour Regression (SCR and GCR).

Let $(\widetilde{X}, \widetilde{Y})$ be an independent copy of (X, Y) and suppose that the central subspace $\mathcal{S}_{Y|X}$ for the regression of Y on X is spanned by the column space of a $p \times d$ matrix β with $d < p$. In simple contour regression, we consider the matrix

$$K(c) = E\left[(X - \widetilde{X})(X - \widetilde{X})^T \Big| \mid Y - \widetilde{Y} \mid \leq c\right].$$

In addition to Assumption 1.1.1, 1.2.1 and 1.2.2, SCR needs the following assumption.

Assumption 1.4.1. *For any choice of vectors $v \in \mathcal{S}_{Y|X}$ and $w \in (\mathcal{S}_{Y|X})^{\perp}$ such that $\|v\| = \|w\| = 1$, and any sufficiently small $c > 0$, we have*

$$\mathrm{Var}\left[w^T(\widetilde{X} - X)\Big| \mid \widetilde{Y} - Y \mid \leq c\right] > \mathrm{Var}\left[v^T(\widetilde{X} - X)\Big| \mid \widetilde{Y} - Y \mid \leq c\right].$$

The next theorem is provided by B. Li, Zha, and Chiaromonte (2005).

Theorem 1.4.4. *Suppose that X has an elliptical distribution with $E(X) = 0$ and $\mathrm{Var}(X) = I_p$. If Assumption 1.4.1 holds, then, for a sufficiently small c, the eigenvectors of $K(c)$ corresponding to its smallest d eigenvalues span the central subspace $S_{Y|X}$.*

In general contour regression, we consider the variance of Y along the line through x_i and x_j. Formally, let $l(t; x_i, x_j) = (1 - t)x_i + tx_j, t \in R$, be the straight line that goes through x_i and x_j, and define

$$V(x_i, x_j) = \mathrm{Var}[Y|X = l(t; x_i, x_j) \text{ for some t}].$$

Consider the matrix

$$G(c) = E\left[(\widetilde{X} - X)(\widetilde{X} - X)^T \Big| V(\widetilde{X}, X) \le c\right].$$

We make the following assumption parallel to Assumption 1.4.1.

Assumption 1.4.2. *For any choice of vectors $v \in \mathcal{S}_{Y|X}$ and $w \in (\mathcal{S}_{Y|X})^{\perp}$ such that $\|v\| = \|w\| = 1$, and any sufficiently small $c > 0$, we have*

$$\text{Var}\left[w^T(\widetilde{X} - X)\Big| V(\widetilde{X}, X) \le c\right] > \text{Var}\left[v^T(\widetilde{X} - X)\Big| V(\widetilde{X}, X) \le c\right].$$

Theorem 1.4.5. *Suppose that X has an elliptical distribution with $E(X) = 0$ and $var(X) = I_p$. If Assumption 1.4.2 holds, then, for a sufficiently small $c > 0$, the eigenvectors of $G(c)$ corresponding to its smallest d eigenvalues span the central subspace $\mathcal{S}_{Y|X}$.*

1.4.5 Directional Regression

Sliced Inverse Regression and Sliced Average Variance Estimator are both based on the first two conditional moments $E(X|Y)$ and $E(XX^T|Y)$. Directional Regression is a dimensional reduction method that synthesizes the first two conditional moments more efficiently than either SIR or SAVE. Let $(\widetilde{X}, \widetilde{Y})$ be an independent copy of (X, Y), and let

$$A(Y, \widetilde{Y}) = E[(X - \widetilde{X})(X - \widetilde{X})^T | Y, \widetilde{Y}].$$

Theorem 1.4.6. *If Assumption 1.1.1, 1.2.1 and 1.2.2 hold for X, then, for any $(Y, \widetilde{Y})$, the column space of $2I_p - A(Y, \widetilde{Y})$ is contained in $\mathcal{S}_{Y|X}$.*

1.5 Central mean space

In previous sections, we introduce the definition of central space and various methods for estimation of central space. Estimation of central space has been the main focus in dimension reduction, however, in many situations, regression analysis is mostly concerned about inferring the conditional mean of the response given the predictors. That is why we introduce the notion of Central Mean Space (CMS; Cook & Li, 2002).

Definition 1.5.1. *If*

$$Y \perp\!\!\!\perp E(Y|X)|\alpha^T X,$$

then $\mathcal{S}(\alpha)$ is a mean dimension reduction space for the regression of Y on X.

It follows from this definition that a dimension-reduction space is necessarily a mean dimension reduction space, because $Y \perp\!\!\!\perp X|\alpha^T X$ implies $Y \perp\!\!\!\perp E(Y|X)|\alpha^T X$. Parallel to the development of central spaces, we would like the smallest dimension reduction space, as formalized in the next definition.

Definition 1.5.2. *Let $\mathcal{S}_{E(Y|X)} = \cap \mathcal{S}_m$ where intersection is over all mean dimension reduction spaces $\mathcal{S}_m$. If $\mathcal{S}_{E(Y|X)}$ is itself a mean dimension reduction space, it is called the central mean space.*

The central mean space also has the invariance property as does the central space.

1.6 Two adaptive estimators

Central mean space provides more insight into dimension reduction. Through the notion of central mean space, a distinction among the dimension reduction methods introduced earlier can be discovered. For example, OLS estimators and PHD estimators not only belong to central space (CS), but also belong to central mean space

(CMS). SIR and SAVE estimators do not have this property. This can be shown in the table below:

	LCM	CCV
CMS	OLS	PHD
CS	SIR	SAVE

Table 1.1. Relationship between common CS and CMS estimators .

Actually, we know the central mean space is always contained in the central space, since the former is an intersecton of larger collectino of subspaces. In this section, we are going to introduce two adaptive methods, which is targeted directly at estimating the central mean space.

1.6.1 Outer Product of Gradients

Outer Product of Gradients (OPG; Xia, Tong, Li, & Zhu, 2002) is related to the Average Derivative Estimation (ADE; Powell, Stock, & Stoker, 1989, Hardle & Stoker, 1989). ADE is based on the fact that the derivative or gradient of the regression funtion at any point falls in the central mean space. In symbols

$$\nabla E(Y|X=x) = \frac{\partial E(Y|X=x)}{\partial x} \in \mathcal{S}_{E(Y|X)}.$$

ADE thus uses $E[\nabla E(Y|X=x)]$ to estimate the central mean space. However, estimating $\nabla E(Y|X=x)$ invovles high-order kernel smoothers and still suffers curse of dimensionality. Another disavantage of ADE is that when $\nabla E(Y|X=x)$ is symmetric we may have $E[\nabla E(Y|X=x)] = 0$.

OPG overcomes these drawbacks by considering

$$E[\nabla E(Y|X=x)\nabla^T E(Y|X=x)].$$

First we minimize the following objective function,

$$\sum_{i=1}^{n}\{Y_i - [a_j + b_j^T(X_i - X_j)]\}^2 w_{ij}$$

and get minimizers a_j, b_j. Then we estimate outer product of gradients by

$$\hat{\Sigma} = n^{-1}\sum_{j=1}^{n}\hat{b}_j\hat{b}_j^T.$$

Finally, use the d eigenvectors of $\hat{\Sigma}$ corresponding to the largest d eigenvalues as estimator for the central mean subspace.

1.6.2 Minimized Average Variance Estimator

Minimized Average Variance Estimator (MAVE; Xia, Tong, Li, & Zhu, 2002) is based on the minimization of

$$\begin{aligned} E[Y - E(Y|\beta^T X)]^2 &= E\{E\{[Y - E(Y|\beta^T X)]^2|\beta^T X\}\} \\ &= E[\mathrm{Var}(Y|\beta^T X)], \end{aligned}$$

among all $\beta \in \mathbb{R}^{p\times d}(d < p)$ with $\beta^T\beta = I_d$.

Let $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$ be a sample of (X, Y). For any given X_0, a local linear expansion of $E(Y_i|\beta^T X_i)$ at X_0 is

$$E(Y_i|\beta^T X_i) \approx a + b^T\beta^T(X_i - X_0).$$

The kernel estimator of $E\{[Y - E(Y|\beta^T X)]^2|\beta^T X_0\}$ is

$$\sum_{i=1}^{n}[Y_i - E(Y_i|\beta^T X_i)]^2 w_{i0} \approx \sum_{i=1}^{n}\{Y_i - [a + b^T\beta^T(X_i - X_0)]\}^2 w_{i0},$$

where w_{i0} are the kernel weights.

Then we estimate $E[\mathrm{Var}(Y|\beta^T X)]$ by averaging over $X_0 = X_1, \ldots, X_n$, that is,

$$n^{-1}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - [a_j + b_j^T\beta^T(X_i - X_j)]\}^2 w_{ij}. \tag{1.3}$$

Under the constraint $\beta^T\beta = I_d$, we look for $\beta \in \mathbb{R}^{p\times d}(d < p)$ that minimizes the above objective function (1.3) as our central mean subspace estimator. The choice of the weights w_{ij} plays a key role. In direct MAVE method, we use

$$w_{ij} = K_h(X_i - X_j)/\sum_{l=1}^{n} K_h(X_l - X_j),$$

where $K_h(\cdot) = h^d K(\cdot/h)$ and d is the dimension of $K(\cdot)$.

In refined MAVE method (RMAVE; Xia, Tong, Li, & Zhu, 2002), we used kernel function defined on lower dimensional space to improve the accuracy of the estimation. Start with estimator β_i, minimizing objective function (1.3) using kernel weights

$$\widetilde{w}_{ij} = K_h(\beta_i(X_i - X_j))/\sum_{l=1}^{n} K_h(\beta_i(X_l - X_j))$$

Then we get the updated estimator β_{i+1}. Repeat this procedure until β_i converges. The choice of bandwidth h might effect the performance of the estimators. We should try different bandwidth and choose the optimal one in practice.

MAVE and OPG are inherently different from the inverse moments methods we introduced in earlier sections. They are sometimes referred to adaptive methods since they require kernel estimation and their efficacy depends on the choice of weights. They don't require Assupmtion 1.2.1 or Assupmtion 1.2.2, but they involve high-dimensional kernel estimation that may not be desirable in practice.

Chapter 2

First-order central solution space methods

2.1 Assumptions for dimension reduction

Dimension reduction for regression is aimed at finding a lower dimensional vector of linear combinations of the predictors which retains as much as possible the information in the relationship between the response and the original predictors. Most of the current dimension reduction methods, however, require strong conditions on the joint distribution of the predictors, such as elliptical symmetry or even multivariate normality. The purpose of this thesis is to remove such strong conditions while preserving the desirable properties of these estimators.

As we have seen, the first-order methods, such as OLS and SIR, all require the linear conditional mean Assumption 1.2.1. The second-order methods, such as SAVE and DR, all require the constant conditional variance Assumption 1.2.2 in addition to Assumption 1.2.1. Although these conditions only need to be true at true β which satisfies span$\beta = \mathcal{S}_{Y|X}$, these conditions are assumed to hold for all possible β since β

is unknown in practice. If Assumption 1.2.1 holds for all β, then X has an elliptically-contoured distribution (Eaton, 1986); if both conditions hold for all β, then X has a multivariate normal distribution. Thus, in effect, either elliptically-contoured or multivariate normal distribution has to be assumed when applying these methods. While it is true that, as the dimension p tends to infinity, the projection of X tends to behave like a normal random variable (Diaconis & Freedman, 1984; Hall & Li, 1993), nevertheless in many practical situations non-ellipticity is commonly seen, and the classical dimension reduction methods are often criticized for lacking an automatic mechanism to take it into account.

If the actual predictors do not satisfy these conditions, current practice often relies on transformation — that is, transform the p components of X, $(X_1, \ldots, X_p)$, to $(h_1(X_1), \ldots, h_p(X_p))$ by some functions $h_1, \ldots h_p$, so that the scatter plot matrix of the transformed predictors resembles that of a multivariate normal distribution. While transformation is a pragmatic — and often effective — strategy, it has both theoretical and practical difficulties. Theoretically, such transformations are intrinsically marginal. It targets the marginal distributions of $X_1, \ldots, X_p$, and as such does not guarantee that $E(X|\beta^T X)$ has desired linearity when $\beta^T X$ is not a set of X_i's. Indeed, there can be hidden nonlinearity among the predictors even if their scatter plot matrix looks perfectly linear. On the other hand marginal transformations may also be excessive: that $E(X|\beta^T X)$ is linear in X does not require every component of X to be linear against every other component. Practically, whether a transformation has succeeded in transforming a set of observed predictors to an elliptical shape often relies entirely on subjective judgement. Moreover, transforming a high dimensional predictor may be tedious or even infeasible. Another way of dealing with non-ellipticity is reweighting (Cook & Nachtsheim, 1994). However, like transformation, it is not focused on that part of the nonlinearity in the predictors that is relevant to dimension reduction. It is also computationally intensive — especially if the dimension p is high.

When the conditions LCM and/or CCV are satisfied, however, the above-mentioned methods share properties that make them uniquely desirable among nonparametric methods. First, the slicing (or smoothing) involved in these estimators is over the response Y, which is always one dimensional regardless of the dimension of X. It is well known that smoothing over a high dimensional vector space is undesirable, because the data points within a slice (or a region covered by a smoothing kernel) become sparse at an exponential rate as the dimension increases — a phenomenon often referred to as the "curse of dimensionality" (Bellman, 1961). Second, the size of the slice (or bandwidth of the kernel) for the above methods need not decrease with the sample size for consistency. These properties make the above methods resemble parametric estimators — they are $\sqrt{n}$-consistent regardless of the dimension of X and have simple asymptotic structure — even though the problems they tackle are in fact nonparametric, in the sense that virtually no assumption is imposed on the conditional distribution of $Y|X$.

In the following sections, we will introduce a method that does not require LCM or CCV, while at the same time preserves all the desirable properties described in the foregoing paragraph. We will focus on the first-order methods in this chapter. The new method is akin to inverse regression, but it is adapted in an automated fashion to the nonlinearity in the predictors — and only that part of the nonlinearity relevant for dimension reduction.

2.2 Central solution space for SIR

The best way to explain the central idea of this new dimension reduction method is to explain it in comparison with Sliced Inverse Regression. Assume, without loss of generality, that $E(X) = 0$, $\text{Var}(X) = I_p$ and $E(Y) = 0$. Suppose the Central Space has dimension d, and let β be $p \times d$ matrix whose columns form a basis in $\mathcal{S}_{Y|X}$. Sliced Inverse Regression is based on the following fact. If Assumption 1.2.1 holds for β, or

$$E(X|\beta^T X) \text{ is linear in } X \text{ (LCM)}, \tag{2.1}$$

then the random vector $E(X|Y)$ belongs to $\mathcal{S}_{Y|X}$ almost surely. To see this, let P_β be the projection on to $\mathcal{S}_{Y|X}$ with respect to the inner product $\langle a, b \rangle = a^T b$; that is, $P_\beta = \beta(\beta^T\beta)^{-1}\beta^T$. Condition (2.1) implies $E(X|\beta^T X) = P_\beta X$. Hence

$$\begin{aligned} E(X|Y) &= E[E(X|\beta^T X, Y)|Y] \\ &= E[E(X|\beta^T X)|Y] = P_\beta E(X|Y) = P_\beta E(X|Y). \end{aligned} \tag{2.2}$$

Thus the random vector $E(X|Y)$ belongs to the range of the projection operator P_β, which is $\mathcal{S}_{Y|X}$. Consequently, the column space of the matrix

$$\text{Var}[E(X|Y)] = P_\beta \text{Var}[E(X|Y)] P_\beta \tag{2.3}$$

is a subspace of $\mathcal{S}_{Y|X}$. This column space will be called the inverse regression space, and written as $\mathcal{S}_{\text{IR}}$; the matrix (2.3) will be written as A_{IR}.

At the first sight, LCM seems crucial in the foregoing argument. However, note that it is the second equality in (3.2) that reflects the conditional independence $Y \perp\!\!\!\perp X|\beta^T X$, and it requires virtually no condition. The next two equalities in (3.2), which require LCM, merely serve to make $E(X|Y)$ an explicit vector in $\mathcal{S}_{Y|X}$. This leads us to pay special attention to the equation

$$E(X|Y) = E[E(X|\beta^T X)|Y] \quad \text{a.s.} \tag{2.4}$$

That is, the inverse (L_2-) regression of X on Y is the same as the double (L_2-) regressions of X on $\beta^T X$ and then on Y. Because of the importance of this equation we will call it the inverse regression equation. Note that if β solves this equation, then so does βA for any $d \times d$ nonsingular matrix A. That is, the above equation is identified only up to the column space of β.

Definition 2.2.1. *If β is a matrix of p rows that satisfies the inverse regression equation (2.4), then* $\text{span}(\beta)$ *is called a solution space of inverse regression equation.*

It is easy to see that if β_1 satisfies (2.4) and β_2 is another matrix such that $\text{span}(\beta_1) \subseteq \text{span}(\beta_2)$, then β_2 also satisfies (2.4). For maximum dimension reduction we would like to seek β of lowest rank. This leads to the notion of *central solution space* (CSS).

Definition 2.2.2. *If the intersection of any two solution spaces of (2.4) is itself a solution space of (2.4), then the intersection of all such spaces will be called the central solution space of the inverse regression equation, and is written as $\mathcal{S}_{\text{CSS}}$.*

Central solution space is defined under the premise that the intersection of two solution spaces of (2.4) is again a solution space of (2.4). The similar premise also underlies the construction of the central space, which was recently proved under very weak assumptions by Yin, Li, and Cook (2008) in that context. The proof in our context is similar, and is omitted.

The next proposition reveals the relation among $\mathcal{S}_{\text{CSS}}$, $\mathcal{S}_{\text{IR}}$, and $\mathcal{S}_{Y|X}$, which is the theoretical foundation of our method. We will say that condition (2.1) holds for a subspace $\mathcal{S}$ of $\mathbb{R}^p$ if it holds for a matrix η whose columns form a basis in $\mathcal{S}$. Henceforth P_η will denote the orthogonal projection on to span(η) with respect to the regular inner product.

Theorem 2.2.1. *Suppose that Y and the elements of X are square integrable and $E(X) = 0$. Then*

1. $\mathcal{S}_{\text{CSS}} \subseteq \mathcal{S}_{Y|X}$.
2. *If, in addition, condition (2.1) holds for both $\mathcal{S}_{\text{CSS}}$ and $\mathcal{S}_{\text{IR}}$, then $\mathcal{S}_{\text{IR}} = \mathcal{S}_{\text{CSS}}$.*

PROOF.

1. Let β be p-row matrix such that span$(\beta) = \mathcal{S}_{Y|X}$. Then $Y \perp\!\!\!\perp X|\beta^T X$, which, by (3.2), implies (2.4). Thus $\mathcal{S}_{Y|X}$ is a solution space of (2.4), and assertion 1 follows.

2. Let η be a p-row matrix whose columns form a basis in $\mathcal{S}_{\text{CSS}}$. If condition (2.1) holds for η, then

$$E(X|Y) = E[E(X|\eta^T X)|Y] = P_\eta E(X|Y).$$

   Consequently

$$\text{Var}[E(X|Y)] = P_\eta \text{Var}[E(X|Y)] P_\eta. \tag{2.5}$$

   Thus we have $\mathcal{S}_{\text{IR}} \subseteq \mathcal{S}_{\text{CSS}}$.

Conversely, let ξ is a p-row matrix whose columns form a basis in $\mathcal{S}_{\mathrm{IR}}$. Then

$$\begin{aligned} & E\|E(X|Y) - P_\xi E(X|Y)\|^2 \\ & \qquad = \mathrm{trace}(A_{\mathrm{IR}}) - \mathrm{trace}(A_{\mathrm{IR}} P_\xi) - \mathrm{trace}(P_\xi A_{\mathrm{IR}}) + \mathrm{trace}(P_\xi A_{\mathrm{IR}} P_\xi), \end{aligned}$$

where A_{IR} is as defined in (2.3). Because $\mathrm{span}(A_{\mathrm{IR}}) = \mathrm{span}(\xi)$ and because A_{IR} is symmetric, the last three terms on the right (without sign) all reduce to $\mathrm{trace}(A_{\mathrm{IR}})$. Consequently, the above quantity is 0, implying

$$E(X|Y) = P_\xi E(X|Y) \text{ a.s.},$$

Because $E(X|\xi^T X)$ is linear in ξ, the right hand side is

$$P_\xi E(X|Y) = E[E(X|\xi^T X)|Y].$$

Hence $\mathcal{S}_{\mathrm{CSS}} \subseteq \mathcal{S}_{\mathrm{IR}}$. □

Observe that part 1 of the theorem holds without any assumption except the existence of moments; the linearity assumption is required only when $\mathcal{S}_{\mathrm{IR}}$ enters the picture. Thus if we target $\mathcal{S}_{\mathrm{CSS}}$ instead of $\mathcal{S}_{\mathrm{IR}}$, then we can avoid the linearity assumption.

2.3 A general formulation of first-order methods

Several important dimension reduction methods are directly or indirectly related to the fundamental fact that $\mathcal{S}_{\mathrm{IR}} \subseteq \mathcal{S}_{Y|X}$ under condition (2.1). These include Ordinary Least Squares, Sliced Inverse Regression, Parametric Inverse Regression, Canonical

Correlation, and Kernel Inverse Regression. All these methods rely on the condition (2.1) for their consistency. The original form of PIR (Bura & Cook, 2001) was introduced under the assumption that an inverse parametric regression model is true, and under that assumption no restriction needs to be imposed on X. However, PIR is in fact consistent when the parametric inverse model is not true, and in this case condition (2.1) is needed for its consistency. This fact is noted in Fung, He, Liu, and Shi (2002) in a different context. The goal of this paper is to use the general mechanism of the central solution space (CSS) to extend these methods so that their consistency does not rely on condition (2.1). For this purpose we now give a brief outline of the construction of these estimators, and synthesize them into a common form. Suppose the normalizing Assumption 1.1.1 is met. Or we can standardize X first and transform the central space back to the original scale later.

The OLS estimator is based on the following matrix

$$A_{\mathrm{OLS}} = E(YX)E(YX^T).$$

Let $\{J_1, \ldots, J_k\}$ be a (measurable) partition of Ω_Y, the sample space of Y, and define the discretized version of Y as

$$\delta(Y) = \sum_{\ell=1}^{k} \ell I(Y \in J_\ell).$$

The SIR estimator is based on the following matrix:

$$A_{\mathrm{SIR}} = \mathrm{Var}\left[E(X|\delta(Y))\right].$$

Let $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ be a probability density function, $h > 0$, and $y \in \Omega_Y$. Let

$$\kappa(y, \tilde{y}) = \psi\left(h^{-1}|y - \tilde{y}|\right) / \, E\left[\psi\left(h^{-1}|Y - \tilde{y}|\right)\right]. \tag{2.6}$$

Because h will be treated as fixed throughout the paper, we suppress the dependence on h from the notation. Let $\tilde{Y}$ be a random variable having the same distribution as Y with $\tilde{Y} \perp\!\!\!\perp (X, Y)$. The KIR estimator is based on the following matrix:

$$A_{\text{KIR}} = E\left\{E[X\kappa(Y, \tilde{Y})|\tilde{Y}]E[X^T\kappa(Y, \tilde{Y})|\tilde{Y}]\right\}. \tag{2.7}$$

Finally, let $h_1, \ldots, h_s$ be square integrable functions from Ω_Y to $\mathbb{R}$, one of which (say h_1) must be taken to be 1 if Y is not centered. Let $H(y) = (h_1(y), \ldots, h_s(y))^T$. Let

$$\rho(y, \tilde{y}) = H^T(y)E[H(Y)H^T(Y)]^{-1}H(\tilde{y}). \tag{2.8}$$

The following matrix

$$A_{\text{PIR}} = E\left\{E[X\rho(Y, \tilde{Y})|\tilde{Y}]E[X^T\rho(Y, \tilde{Y})|\tilde{Y}]\right\}.$$

is sufficiently general to accommodate (the population versions of) both PIR and CANCOR — though their original forms were quite different. We also note that both PIR and CANCOR allow Y to be a vector, but this is not considered in this thesis.

It turns out all four matrices can be written in the same form, which will greatly simplify the subsequent development and provide insights into the relationship among these methods. Henceforth, for two random elements U and V, $U \overset{\mathcal{D}}{=} V$ means that they have the same distribution.

Theorem 2.3.1. *The matrices A_{OLS}, A_{SIR}, A_{KIR}, A_{PIR} can be written in the following* *form*

$$E\left\{E[Xg(Y,\tilde{Y})|\tilde{Y}]E[X^Tg(Y,\tilde{Y})|\tilde{Y}]\right\}, \tag{2.9}$$

where $g : \Omega_Y \times \Omega_Y \to \mathbb{R}$, $\tilde{Y} \perp\!\!\!\perp (X,Y)$, *and* $\tilde{Y} \overset{\mathcal{D}}{=} Y$.

PROOF. That A_{KIR} and A_{PIR} have the form (2.9) follows from their definitions. Also, if we let $g(y,\tilde{y}) = y$, then

$$E(XY) = E(XY|\tilde{Y}) = E[Xg(Y,\tilde{Y})|\tilde{Y}].$$

Thus A_{OLS} conforms to (2.9).

For A_{SIR}, note that for any $j \in \{1,\ldots,k\}$,

$$E[X|\delta(Y)=j] = \frac{E[XI(\delta(Y)=j)]}{P(\delta(Y)=j)} = \frac{E[XI(\delta(Y)=\delta(\tilde{Y}))|\delta(\tilde{Y})=j]}{P(\delta(Y)=\delta(\tilde{Y})|\delta(\tilde{Y})=j)}.$$

Because $Y \overset{\mathcal{D}}{=} \tilde{Y}$, the above equality implies that

$$E[X|\delta(Y)] \overset{\mathcal{D}}{=} E[XI(\delta(Y)=\delta(\tilde{Y}))|\delta(\tilde{Y})]/P[\delta(Y)=\delta(\tilde{Y})|\delta(\tilde{Y})].$$

Let $g(Y,\tilde{Y}) = I(\delta(Y)=\delta(\tilde{Y}))/P[\delta(Y)=\delta(\tilde{Y})|\delta(\tilde{Y})]$. Then

$$E[X|\delta(Y)] \overset{\mathcal{D}}{=} E[Xg(Y,\tilde{Y})|\delta(\tilde{Y})]. \tag{2.10}$$

In the meantime,

$$\tilde{Y} \perp\!\!\!\perp (X,Y) \Rightarrow (\tilde{Y},\delta(\tilde{Y})) \perp\!\!\!\perp (X,Y) \Rightarrow (X,Y) \perp\!\!\!\perp \tilde{Y}|\delta(\tilde{Y}) \Rightarrow (X,Y) \perp\!\!\!\perp \tilde{Y}|\{\delta(\tilde{Y}),\delta(\tilde{Y})\},$$

which, together with $\delta(\tilde{Y}) \perp\!\!\!\perp \tilde{Y}|\delta(\tilde{Y})$, implies that

$$(X, Y, \delta(\tilde{Y})) \perp\!\!\!\perp \tilde{Y}|\delta(\tilde{Y}).$$

See, for example, Dawid (1979) and Cook (1998b). Hence the right hand side of (2.10) reduces to $E[Xg(Y, \tilde{Y})|\tilde{Y}]$, and equality (2.10) reduces to

$$E[X|\delta(Y)] \overset{\mathcal{D}}{=} E[Xg(Y, \tilde{Y})|\tilde{Y}].$$

Thus A_{SIR} also has the form (2.9). □

2.4 Extension to nonlinear predictor cases

The synthesis of the last section provides us a platform on which to extend the five methods to situations where the LCM condition (2.1) does not hold. We now carry out this extension.

While Theorem 2.2.1 lays out the basic principle of central solution space, as we have noticed in Section 2.3, various versions of inverse regressions do not take the exact form Var$[E(X|Y)]$. We now extend Theorem 2.2.1 to accommodate the various forms of inverse regressions, as synthesized in Section 2.3.

Denote the matrix (2.9) by $A_{\text{IR}}(g)$ and its column space by $\mathcal{S}_{\text{IR}}(g)$, where g stands for the function $g(Y, \tilde{Y})$ in Theorem 2.3.1. Consider the following equation

$$E[Xg(Y, \tilde{Y})|\tilde{Y}] = E[E(X|\beta^T X)g(Y, \tilde{Y})|\tilde{Y}], \quad \text{a.s.} \tag{2.11}$$

where, recall that $\tilde{Y} \overset{\mathcal{D}}{=} Y$ and $\tilde{Y} \perp\!\!\!\perp (X, Y)$. Let $\mathcal{S}_{\text{CSS}}(g)$ be the central solution space of this equation.

Theorem 2.4.1. *Suppose that $g : \Omega_Y \times \Omega_Y \to \mathbb{R}$ is a measurable function such that the elements of $Xg(Y, \tilde{Y})$ are square integrable. Suppose Y and the elements of X are square integrable with $E(X) = 0$ and $E(Y) = 0$. Then*

1. $\mathcal{S}_{\mathrm{CSS}}(g) \subseteq \mathcal{S}_{Y|X}$.

2. *If, in addition, condition (2.1) holds for both $\mathcal{S}_{\mathrm{CSS}}(g)$ and $\mathcal{S}_{\mathrm{IR}}(g)$, then $\mathcal{S}_{\mathrm{IR}}(g) = \mathcal{S}_{\mathrm{CSS}}(g)$.*

Proof. 1. Let β be a matrix such that $\mathrm{span}(\beta) = \mathcal{S}_{Y|X}$. Because $\tilde{Y} \perp\!\!\!\perp (X, Y)$, we have

$$\tilde{Y} \perp\!\!\!\perp (X, Y, \beta^T X) \Rightarrow \tilde{Y} \perp\!\!\!\perp (X, Y)|\beta^T X.$$

The expression on the right-hand side, together with $Y \perp\!\!\!\perp X|\beta^T X$, implies that $X \perp\!\!\!\perp Y \perp\!\!\!\perp \tilde{Y}|\beta^T X$, and hence that $X \perp\!\!\!\perp (Y, \tilde{Y})|\beta^T X$. It follows that $E(X|Y, \tilde{Y}, \beta^T X) = E(X|\beta^T X)$, and consequently

$$\begin{aligned} E[Xg(Y, \tilde{Y})|\tilde{Y}] =& E[E(X|\beta^T X, Y, \tilde{Y})g(Y, \tilde{Y})|\tilde{Y}] \\ =& E[E(X|\beta^T X)g(Y, \tilde{Y})|\tilde{Y}]. \end{aligned} \tag{2.12}$$

Thus $\mathcal{S}_{Y|X}$ is a solution space of (2.11), and assertion 1 follows.

2. Let η be a matrix such that $\mathrm{span}(\eta) = \mathcal{S}_{\mathrm{CSS}}(g)$. Since (2.1) holds for η, we have $E(X|\eta^T X) = P_\eta X$, and so (2.11) becomes

$$E[Xg(Y, \tilde{Y})|\tilde{Y}] = E[E(X|\eta^T X)g(Y, \tilde{Y})|\tilde{Y}] = P_\eta E[Xg(Y, \tilde{Y})|\tilde{Y}],$$

which implies $A_{\mathrm{IR}}(g) = P_\eta A_{\mathrm{IR}}(g) P_\eta$. Hence $\mathcal{S}_{\mathrm{IR}}(g) \subseteq \mathcal{S}_{\mathrm{CSS}}(g)$. Conversely, let ξ be a matrix such that $\mathrm{span}(\xi) = \mathcal{S}_{\mathrm{IR}}(g)$. Then

$$E\|E[(Xg(Y,\tilde{Y})|\tilde{Y}] - P_\xi E[(Xg(Y,\tilde{Y})|\tilde{Y}]\|^2$$
$$= \mathrm{trace}[A_{\mathrm{IR}}(g)] - \mathrm{trace}[A_{\mathrm{IR}}(g)P_\xi] - \mathrm{trace}[P_\xi A_{\mathrm{IR}}(g)] + \mathrm{trace}[P_\xi A_{\mathrm{IR}}(g)P_\xi].$$

Because $\mathrm{span}[A_{\mathrm{IR}}(g)] = \mathrm{span}(\xi)$ and because $A_{\mathrm{IR}}(g)$ is symmetric, the last three terms on the right (without sign) all reduce to $\mathrm{trace}[A_{\mathrm{IR}}(g)]$. Consequently, the above quantity is 0, implying

$$E[(Xg(Y,\tilde{Y})|\tilde{Y}] = P_\xi E[(Xg(Y,\tilde{Y})|\tilde{Y}] \quad \text{a.s.}$$

Because $E(X|\xi^T X)$ is linear in X, the right hand side reduces to

$$P_\xi E[(Xg(Y,\tilde{Y})|\tilde{Y}] = E[E(X|\xi^T X)g(Y,\tilde{Y})|\tilde{Y}].$$

Hence $\mathcal{S}_{\mathrm{CSS}}(g) \subseteq \mathcal{S}_{\mathrm{IR}}(g)$. □

Let $\mathcal{S}_{\mathrm{OLS}}$, $\mathcal{S}_{\mathrm{SIR}}$, $\mathcal{S}_{\mathrm{KIR}}$, $\mathcal{S}_{\mathrm{PIR}}$ be the columns spaces of A_{OLS}, A_{SIR}, A_{KIR}, A_{PIR}. Let $\mathcal{S}_{\mathrm{CSS-OLS}}$, $\mathcal{S}_{\mathrm{CSS-SIR}}$, $\mathcal{S}_{\mathrm{CSS-KIR}}$, $\mathcal{S}_{\mathrm{CSS-PIR}}$ be the column spaces of $A_{\mathrm{IR}}(g)$ with g taken to be the four $g(Y,\tilde{Y})$ functions described in the proof of Theorem 2.3.1. The following corollary follows immediately from Theorem 2.4.1.

Corollary 2.4.1. *Suppose all the moments involved in the definitions of $\mathcal{S}_{\mathrm{OLS}}$, ..., $\mathcal{S}_{\mathrm{PIR}}$ and $\mathcal{S}_{\mathrm{CSS-OLS}}$, ..., $\mathcal{S}_{\mathrm{CSS-PIR}}$ are finite. Then*

1. $\mathcal{S}_{\mathrm{CSS-OLS}} \subseteq \mathcal{S}_{Y|X}$, $\mathcal{S}_{\mathrm{CSS-SIR}} \subseteq \mathcal{S}_{Y|X}$, $\mathcal{S}_{\mathrm{CSS-KIR}} \subseteq \mathcal{S}_{Y|X}$, $\mathcal{S}_{\mathrm{CSS-PIR}} \subseteq \mathcal{S}_{Y|X}$.
2. *If (2.1) holds for $\mathcal{S}_{\mathrm{OLS}}$, ..., $\mathcal{S}_{\mathrm{PIR}}$ and $\mathcal{S}_{\mathrm{CSS-OLS}}$, ..., $\mathcal{S}_{\mathrm{CSS-PIR}}$, then*

$$\mathcal{S}_{\mathrm{OLS}} = \mathcal{S}_{\mathrm{CSS-OLS}}, \quad \mathcal{S}_{\mathrm{SIR}} = \mathcal{S}_{\mathrm{CSS-SIR}},$$

$$\mathcal{S}_{\text{KIR}} = \mathcal{S}_{\text{CSS}-\text{KIR}}, \quad \mathcal{S}_{\text{PIR}} = \mathcal{S}_{\text{CSS}-\text{PIR}}.$$

Again, note that inclusions in part 1 hold without linearity condition (2.1). Part 2 says that when condition (2.1) does hold, using central solution space based methods will not lose information as compared the inverse regression based methods.

2.5 Objective functions

We now introduce a population-level objective function whose minimizer yields the solution to (2.11) for each given g. We will also describe how it can be estimated based on an i.i.d. sample of (X, Y). The next theorem will provide a guiding principle for defining the objective function.

Theorem 2.5.1. *Suppose that $\mathcal{S}_{\text{CSS}}(g)$ has dimension $d \leq p$ and let β be a $p \times d$ matrix whose columns form a basis in $\mathcal{S}_{\text{CSS}}(g)$. Let $f(\eta^T X)$ be a square-integrable function such that, whenever* $\text{span}(\eta) = \text{span}(\beta)$*, $f(\eta^T X) = E(X|\beta^T X)$, and whenever* $\text{span}(\eta) \neq \text{span}(\beta)$,

$$P\left\{E[f(\eta^T X)g(Y, \tilde{Y})|\tilde{Y}] \neq E[f(\beta^T X)g(Y, \tilde{Y})|\tilde{Y}]\right\} > 0. \tag{2.13}$$

Let $\eta_0 \in \mathbb{R}^{p \times d}$ be the minimizer of

$$L(\eta) = E\left\|E\{[X - f(\eta^T X)]g(Y, \tilde{Y})|\tilde{Y}\}\right\|^2 \tag{2.14}$$

over $\mathbb{R}^{p \times d}$. Then $\text{span}(\eta_0) = \mathcal{S}_{\text{CSS}}(g)$.

Proof. If $\mathrm{span}(\eta) = \mathrm{span}(\beta)$, then

$$E[f(\eta^T X)g(Y, \tilde{Y})|\tilde{Y}] = E[E(X|\beta^T X)g(Y, \tilde{Y})|\tilde{Y}] = E(Xg(Y, \tilde{Y})|\tilde{Y}) \text{ a.s.}$$

Hence $L(\eta) = 0$. If $\mathrm{span}(\eta) \neq \mathrm{span}(\beta)$, then by assumption (2.13),

$$E\|E\{[f(\eta^T X) - f(\beta^T X)]g(Y, \tilde{Y})|\tilde{Y}\}\|^2 > 0.$$

In the meantime,

$$\begin{aligned}
&L(\eta) = E\|E\{[X - f(\beta^T X)]g(Y, \tilde{Y})|\tilde{Y}\}\|^2 \\
&+ E\|E\{[f(\beta^T X) - f(\eta^T X)]g(Y, \tilde{Y})|\tilde{Y}\}\|^2 \\
&+ 2E\left(E\{[X - f(\beta^T X)]g(Y, \tilde{Y})|\tilde{Y}\}^T E\{[f(\beta^T X) - f(\eta^T X)]g(Y, \tilde{Y})|\tilde{Y}\}\right).
\end{aligned}$$

Because $\mathrm{span}(\beta) = \mathcal{S}_{\mathrm{CSS}}(g)$, the last term is 0. Therefore

$$L(\eta) \geq E\|E\{[f(\beta^T X) - f(\eta^T X)]g(Y, \tilde{Y})|\tilde{Y}\}\|^2 > 0.$$

Hence the minimizer of $L(\eta)$ must satisfy $\mathrm{span}(\eta) = \mathrm{span}(\beta)$. □

Rather than assuming $E(X|\beta^T X)$ to be linear in $\beta^T X$ at the outset, as we do for classical methods such as SIR, here we model $E(X|\beta^T X)$ parametrically. Let $f_1, \ldots, f_k$ be functions from $\mathbb{R}^d$ to $\mathbb{R}$. We will assume that $E(X|\beta^T X)$ lies in the space spanned by $f_1(\beta^T X), \ldots, f_k(\beta^T X)$. That is, each component of $E(X|\beta^T X)$ is a linear combination of $f_1(\beta^T X), \ldots, f_k(\beta^T X)$. Under this assumption the conditional expectation $E(X|\beta^T X)$ can be expressed explicitly as

$$E(X|\beta^T X) = E[XG^T(\beta^T X)] \left\{E\left[G(\beta^T X)G^T(\beta^T X)\right]\right\}^{-1} G(\beta^T X),$$

where

$$G(\beta^T X) = \left(f_1(\beta^T X), \ldots, f_k(\beta^T X)\right)^T. \tag{2.15}$$

Note that we are not assuming — and we do not need to assume — that $E(X|\eta^T X)$ is a linear function of $f_1(\eta^T X), \ldots, f_k(\eta^T X)$ for every η in $\mathbb{R}^{p\times d}$. All we need is that this holds at the true β. We use the function

$$E[XG^T(\eta^T X)] \left\{E\left[G(\eta^T X)G^T(\eta^T X)\right]\right\}^{-1} G(\eta^T X). \tag{2.16}$$

as the $f(\eta^T X)$ in the definition (2.14) of the objective function $L(\eta)$.

2.6 Sample estimation algorithm

We now construct the sample estimate $L_n(\eta)$ of $L(\eta)$. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample of (X, Y). For a function $r(X, Y)$, let $E_n r(X, Y)$ denote the sample average $n^{-1}\sum_{i=1}^n r(X_i, Y_i)$.

1. Center $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_n$ as

$$\hat{Y}_i = Y_i - E_n(Y), \quad \hat{X}_i = X_i - E_n(X).$$

2. Select $\{f_1, \ldots, f_k\}$ that we deem sufficiently flexible to describe the conditional mean $E(X|\beta^T X)$. For example, based on our experience it often suffices to

include linear and quadratic functions of $\beta^T X$. In this case, the set $\{f_1, \ldots, f_k\}$ includes the following $d(d+3)/2+1$ functions

$$\{1\} \cup \{\eta_i^T X : i = 1, \ldots, d\} \cup \{\eta_j^T X \, \eta_k^T X : \ 1 \le j \le k \le d\},$$

where $\eta_1, \ldots, \eta_d$ are columns η. Let

$$\hat{f}(\eta^T \hat{X}) = E_n[\hat{X} G^T(\eta^T \hat{X})]\{E_n[G(\eta^T \hat{X}) G^T(\eta^T \hat{X})]\}^{-1} G(\eta^T \hat{X})$$

3. If using OLS, define $L_n(\eta)$ as

$$E_n \|(\hat{X} - \hat{f}(\eta^T \hat{X}))\hat{Y}\|^2$$

If using SIR, define $L_n(\eta)$ as

$$\frac{1}{n} \sum_{\ell=1}^{k} E_n[I(\hat{Y} \in J_\ell)] \| E_n[(\hat{X} - \hat{f}(\eta^T \hat{X})) | \hat{Y} \in J_\ell] \|^2,$$

where

$$E_n[(\hat{X} - \hat{f}(\eta^T \hat{X})) | \hat{Y} \in J_\ell] = E_n[(\hat{X} - \hat{f}(\eta^T \hat{X})) I(\hat{Y} \in J_\ell)] / E_n[I(\hat{Y} \in J_\ell)].$$

If using KIR, PIR, or CANCOR, define $L_n(\eta)$ as

$$n^{-1} \sum_{j=1}^{n} \left\| n^{-1} \sum_{i=1}^{n} \left\{ \left[\hat{X}_i - \hat{f}(\eta^T \hat{X}_i) \right] g(\hat{Y}_i, \hat{Y}_j) \right\} \right\|^2,$$

where g is either the function κ defined in (2.6) or the function ρ defined in (2.8). Note that for PIR and CANCOR, $g(\hat{Y}_i, \hat{Y}_j)$ can be factorized into functions of

$\hat{Y}_i$ and $\hat{Y}_j$, and thus the above double sum can be simplified as a single sum.

Chapter 3

Second-order central solution space methods

In Chapter 2, we use the notion of the central solution space (CSS) to modify the first-order methods, so that they do not rely on the linear conditional mean Assumption 1.2.1 but at the same time preserve the desirable properties enjoyed by the first-order estimates, such as $\sqrt{n}$-consistency and asymptotic normality. We will study second-order central solution space methods in this chapter. Again, the linear conditional mean assumption can be removed.

3.1 Central solution space for SAVE

Sliced Average Variance Estimator requires both the linear conditional mean Assumption 1.2.1 and the constant conditional variance Assumption 1.2.2. These assumptions are satisfied when the predictor X has a multivariate normal distribution. Using the mechanism of the central solution space, we can loosen both assumptions or the linear conditional mean assumption alone.

3.1.1 Derivation of $\mathcal{S}_{\text{CSS}-\text{SAVE}}$

To understand the core issue, we first outline what makes SAVE work in the classical setting. In the following, for a $p \times d$ matrix β, $\text{span}(\beta)$ denotes the subspace of $\mathbb{R}^p$ spanned by the column vectors of β. Let $P_\beta = \beta(\beta^T\beta)^{-1}\beta^T$ be the projection matrix on to $\text{span}(\beta)$, and let $Q_\beta = I_p - P_\beta$ be the projection matrix on to the orthogonal complement of $\text{span}(\beta)$. Under Assumption 1.1.1, 1.2.1 and 1.2.2, it can be deduced that

$$E(X|\beta^T X) = P_\beta X \text{ and } \text{Var}(X|\beta^T X) = Q_\beta. \tag{3.1}$$

The idea of SAVE is based on the following equalities:

$$\begin{aligned} \text{Var}(X|Y) =& E[\text{Var}(X|Y, \beta^T X)|Y] + \text{Var}[E(X|Y, \beta^T X)|Y] \\ =& E[\text{Var}(X|\beta^T X)|Y] + \text{Var}[E(X|\beta^T X)|Y] \\ =& Q_\beta + P_\beta \text{Var}(X|Y) P_\beta. \end{aligned} \tag{3.2}$$

Hence $I_p - \text{Var}(X|Y) = P_\beta[I_p - \text{Var}(X|Y)]P_\beta$. The column space of the matrix $I_p - \text{Var}(X|Y)$ is thus contained in the range space of the projection operator P_β, which is $\mathcal{S}_{Y|X}$. Consequently, the column space of the matrix $E[I_p - \text{Var}(X|Y)]^2$ is a subspace of $\mathcal{S}_{Y|X}$. This column space will be written as $\mathcal{S}_{\text{SAVE}}$.

Note that, in the derivation of (3.2), Assumption 1.2.1 and 1.2.2 are used only to obtain the last equality, which has nothing to do with (1.1), or the conditional independence $Y \perp\!\!\!\perp X|\beta^T X$. It is the second equality that reflects this conditional independence. In other words, if we derive the estimate of β through the equation

$$\text{Var}(X|Y) = E[\text{Var}(X|\beta^T X)|Y] + \text{Var}[E(X|\beta^T X)|Y], \tag{3.3}$$

then Assumption 1.2.1 and 1.2.2 can be circumvented.

In principle, we can parameterize both $E(X|\beta^T X)$ and $\mathrm{Var}(X|\beta^T X)$ to circumvent Assumption 1.2.1 and 1.2.2 simultaneously. However, since we are primarily concerned about the nonlinearity in X, and since the conditional variance $\mathrm{Var}(X|\beta^T X)$ is more complicated to parameterize, we adopt the less ambitious approach of still assuming the matrix $\mathrm{Var}(X|\beta^T X)$ to be nonrandom.

With the constant conditional variance assumption, equation (3.3) can be further simplified. By assumption (1.2.2), the first term on the right-hand side of equation (3.3) is $E[\mathrm{Var}(X|\beta^T X)]$, which is the same as $\mathrm{Var}(X) - \mathrm{Var}[E(X|\beta^T X)]$. Substitute this into (3.3) to obtain

$$\mathrm{Var}(X) - \mathrm{Var}(X|Y) = \mathrm{Var}(U_\beta) - \mathrm{Var}(U_\beta|Y), \quad \text{a.s.} \tag{3.4}$$

where U_β stands for the random vector $E(X|\beta^T X)$. Using this and the equalities in (3.1), it is easy to see that if the normalizing assumption and the linear conditional mean assumption are satisfied, then (3.4) reduces to

$$I_p - \mathrm{Var}(X|Y) = P_\beta[I_p - \mathrm{Var}(X|Y)]P_\beta. \tag{3.5}$$

which is the basis for the classical SAVE. Thus SAVE is a special case of its CSS modification when the linear conditional mean assumption holds.

We will call equation (3.4) the SAVE equation. Notice that if $\mathrm{span}(\beta) = \mathrm{span}(\eta)$, then β satisfying (3.4) will imply η satisfying (3.4), and vice versa. Thus the SAVE equation is identified only up to the column space of β. This leads to the following definition.

Definition 3.1.1. *If β is a p-row matrix that satisfies equation (3.4), then* $\mathrm{span}(\beta)$ *is* *called a solution space of the SAVE equation. If the intersection of any two solution spaces of (3.4) is itself a solution space of (3.4), then the intersection of all solution spaces of (3.4) will be called the central solution space for the SAVE equation, written as* $\mathcal{S}_{\text{CSS}-\text{SAVE}}$.

Central solution space is defined under the premise that the intersection of two solution spaces is again a solution space. A similar premise also underlies the construction of the central space, which was recently proved under very weak assumptions by Yin et al. (2008) in that context.

The next theorem reveals the relations among $\mathcal{S}_{\text{CSS}-\text{SAVE}}$, $\mathcal{S}_{\text{SAVE}}$ and $\mathcal{S}_{Y|X}$. It provides a theoretical justification of CSS-SAVE as an estimator of the central space.

Theorem 3.1.1. *Suppose that* Y *and the elements of* X *are square integrable and* $E(X) = 0, \text{Var}(X) = I_p$. *In addition, suppose* $\text{Var}(X|\beta^T X)$ *is non-random, where* β *is a matrix such that* $\text{span}(\beta) = \mathcal{S}_{Y|X}$. *Then*

1. $\mathcal{S}_{\text{CSS}-\text{SAVE}} \subseteq \mathcal{S}_{Y|X}$.
2. *If the linear condition mean assumption and the constant conditional variance assumption hold for both* $\mathcal{S}_{\text{CSS}-\text{SAVE}}$ *and* $\mathcal{S}_{\text{SAVE}}$, *then*
   $\mathcal{S}_{\text{CSS}-\text{SAVE}} = \mathcal{S}_{\text{SAVE}}$.

PROOF. 1. Let β be a p-row matrix such that $\text{span}(\beta) = \mathcal{S}_{Y|X}$. Then, as we have shown in Section 2, β satisfies equation (3.4). Thus $\mathcal{S}_{Y|X}$ is a solution space of (3.4), and consequently contains $\mathcal{S}_{\text{CSS}-\text{SAVE}}$.

2. Let β be a p-row matrix such that $\text{span}(\beta) = \mathcal{S}_{\text{CSS}-\text{SAVE}}$. Then it satisfies (3.4), which, as we argued in Section 2, implies that it also satisfies (3.5). Consequently $\text{span}[I_p - \text{Var}(X|Y)] \subseteq \text{span}(\beta) = \mathcal{S}_{\text{CSS}-\text{SAVE}}$ almost surely, implying $\mathcal{S}_{\text{SAVE}} = \text{span}\{E[I_p - \text{Var}(X|Y)]^2\} \subseteq \text{span}(\beta) = \mathcal{S}_{\text{CSS}-\text{SAVE}}$.

Conversely, let ξ be a p-row matrix such that $\text{span}(\xi) = \mathcal{S}_{\text{SAVE}}$. By definition, $\mathcal{S}_{\text{SAVE}} = \text{span}[E(K^2)]$, where $K = I_p - \text{Var}(X|Y)$. Meanwhile,

$$E\|K - P_\xi K\|^2 = \text{trace}[E(K^2)] - 2\text{trace}[P_\xi E(K^2)] + \text{trace}[P_\xi E(K^2)P_\xi].$$

Because $\text{span}[E(K^2)] = \text{span}(\xi)$ and $E(K^2)$ is symmetric, we have $E(K^2) = P_\xi E(K^2) = P_\xi E(K^2)P_\xi$. Hence the right hand side of the above equality is 0. This implies that $K = P_\xi K$ almost surely, or equivalently, $K = P_\xi K P_\xi$ almost surely. In other words

$$I_p - \text{Var}(X|Y) = P_\xi - P_\xi \text{Var}(X|Y)P_\xi, \quad \text{a.s.}$$

However, by the linear condition mean assumption, $P_\xi X = E(X|P_\xi X)$. Thus the right hand side of the above equality reduces to

$$\text{Var}[E(X|\xi^T X)] - \text{Var}[E(X|\xi^T X)|Y],$$

and ξ satisfies (3.4). This means that $\mathcal{S}_{\text{SAVE}}$ is a solution space of (3.4), and consequently $\mathcal{S}_{\text{CSS-SAVE}} \subseteq \mathcal{S}_{\text{SAVE}}$. $\square$

Part 1 of this theorem says $\mathcal{S}_{\text{CSS-SAVE}}$ falls in the central space $\mathcal{S}_{Y|X}$ under the constant conditional variance assumption alone. Part 2 says that when the linear conditional mean assumption does hold, CSS-SAVE and SAVE are the same subspace.

3.1.2 Estimation of $\mathcal{S}_{\text{CSS-SAVE}}$

The next theorem provides a guiding principle for the construction the objective function for CSS-SAVE.

Theorem 3.1.2. *Suppose that $\mathcal{S}_{\text{CSS-SAVE}}$ has dimension $d \leq p$ and let β be a $p \times d$*

matrix whose columns form a basis in $\mathcal{S}_{\mathrm{CSS-SAVE}}$. *Let* U_η *be an* $\mathbb{R}^p$*-valued square-integrable function of* $\eta^T X$ *such that, whenever* $\mathrm{span}(\eta) = \mathrm{span}(\beta)$, $U_\eta = E(X|\beta^T X)$, *and whenever* $\mathrm{span}(\eta) \neq \mathrm{span}(\beta)$,

$$P\left[\mathrm{Var}(U_\eta|Y) - \mathrm{Var}(U_\eta) \neq \mathrm{Var}(U_\beta|Y) - \mathrm{Var}(U_\beta)\right] > 0. \tag{3.6}$$

Let $\tilde{\eta} \in \mathbb{R}^{p\times d}$ *be the minimizer of*

$$L_0(\eta) = E\|[\mathrm{Var}(X|Y) - \mathrm{Var}(X)] - [\mathrm{Var}(U_\eta|Y) - \mathrm{Var}(U_\eta)]\|^2$$

over $\mathbb{R}^{p\times d}$. *Then* $\mathrm{span}(\tilde{\eta}) = \mathcal{S}_{\mathrm{CSS-SAVE}}$.

PROOF. If $\mathrm{span}(\eta) = \mathrm{span}(\beta)$, then $U_\eta = E(X|\beta^T X) = U_\beta$. Hence, with probability one,

$$\mathrm{Var}(U_\eta|Y) - \mathrm{Var}(U_\eta) = \mathrm{Var}(U_\beta|Y) - \mathrm{Var}(U_\beta) = \mathrm{Var}(X|Y) - \mathrm{Var}(X),$$

where the last equality holds because $\mathrm{span}(\beta) = \mathcal{S}_{\mathrm{CSS-SAVE}}$. It follows then $L_0(\eta) = 0$.

Let $A = \mathrm{Var}(X|Y) - \mathrm{Var}(X)$ and $B(\eta) = \mathrm{Var}(U_\eta|Y) - \mathrm{Var}(U_\eta)$. When $\mathrm{span}(\eta) \neq \mathrm{span}(\beta)$, we have $E\|B(\beta) - B(\eta)\|^2 > 0$ by assumption (3.6). In the mean time,

$$\begin{aligned}
L_0(\eta) &= E\|A - B(\eta)\|^2 \\
&= E\|[A - B(\beta)] + [B(\beta) - B(\eta)]\|^2 \\
&= E\|A - B(\beta)\|^2 + E\|B(\beta) - B(\eta)\|^2 + 2E[A - B(\beta)]^T[B(\beta) - B(\eta)].
\end{aligned}$$

Because $\mathrm{span}(\beta) = \mathcal{S}_{\mathrm{CSS-SAVE}}$, the first and the last terms on the right are 0. Therefore

$$L_0(\eta) = E\|B(\beta) - B(\eta)\|^2 > 0.$$

Hence the minimizer of $L_0(\eta)$ must satisfy $\text{span}(\eta) = \text{span}(\beta) = \mathcal{S}_{\text{CSS}-\text{SAVE}}$. $\square$

Let Ω_Y be the sample space of Y and $\{J_1, \ldots, J_h\}$ be a partition of Ω_Y. Then $L(\eta)$, the discretized version of $L_0(\eta)$, is defined as follows:

$$L(\eta) = \text{trace}\left[\sum_{k=1}^{h} p_k(A_k - B_k(\eta))(A_k - B_k(\eta))^T\right], \tag{3.7}$$

where $p_k = P(Y \in J_k)$,

$$\begin{aligned} A_k &= \text{Var}(X|Y \in J_k) - \text{Var}(X), \text{ and} \\ B_k(\eta) &= \text{Var}(U_\eta|Y \in J_k) - \text{Var}(U_\eta). \end{aligned} \tag{3.8}$$

Given an iid sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), we construct the sample estimate $L_n(\eta)$ of $L(\eta)$ as follows.

1. Center $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ as $\hat{X}_i = X_i - E_n(X)$, and $\hat{Y}_i = Y_i - E_n(Y)$.

2. Select $\{f_1, \ldots, f_k\}$ that are sufficiently flexible to describe the conditional mean $E(X|\beta^T X)$. It often suffices to include linear and quadratic functions of $\eta^T X$. That is, let $G(\eta^T \hat{X})$ defined by (2.15) include the following $d(d+3)/2 + 1$ functions

$$\{1\} \cup \{\eta_i^T \hat{X} : i = 1, \ldots, d\} \cup \{\eta_j^T \hat{X}\, \eta_k^T \hat{X} : 1 \le j \le k \le d\},$$

where $\eta_1, \ldots, \eta_d$ are columns η. Let

$$\hat{U}_\eta = E_n[\hat{X}G^T(\eta^T\hat{X})]\{E_n[G(\eta^T\hat{X})G^T(\eta^T\hat{X})]\}^{-1}G(\eta^T\hat{X}).$$

3. Let $\{J_1, \ldots, J_k\}$ be a partition of the interval $[\min(\hat{Y}_i), \max(\hat{Y}_i)]$. Compute the sample estimate of the discretized $L(\eta)$ given by (3.7) as follows

$$L_n(\eta) = \text{trace}\left[\sum_{k=1}^{h} \hat{p}_k(\hat{A}_k - \hat{B}_k(\eta))(\hat{A}_k - \hat{B}_k(\eta))^T\right],$$

where $\hat{p}_k = E_n[I(\hat{Y} \in J_k)]$,

$$\begin{aligned}\hat{A}_k &= E_n(\hat{X}\hat{X}^T|\hat{Y} \in J_k) \\ &\quad - E_n(\hat{X}|\hat{Y} \in J_k)E_n(\hat{X}^T|\hat{Y} \in J_k) - E_n(\hat{X}\hat{X}^T), \text{ and} \\ \hat{B}_k(\eta) &= E_n(\hat{U}_\eta\hat{U}_\eta^T|\hat{Y} \in J_k) \\ &\quad - E_n(\hat{U}_\eta|\hat{Y} \in J_k)E_n(\hat{U}_\eta^T|\hat{Y} \in J_k) - E_n(\hat{U}_\eta\hat{U}_\eta^T).\end{aligned}$$

The empirical conditional expectations above, such as $E_n(\hat{X}|\hat{Y} \in J_k)$, are defined by ratios such as

$$E_n(\hat{X}|\hat{Y} \in J_k) = E_n[\hat{X}I(\hat{Y} \in J_k)]/E_n[I(\hat{Y} \in J_k)].$$

Many softwares are available for minimizing functions such as $L_n(\eta)$. We use the OPTIM function in R. All it requires is a subroutine that evaluates the objective function and an initial value of η. The column space spanned by the minimizer $\hat{\eta}$ of $L_n(\eta)$ is the CSS-SAVE estimator for the central space $\mathcal{S}_{Y|X}$.

3.2 Central solution space for DR

Like SAVE, Directional Regression also requires both the linear conditional mean Assumption 1.2.1 and the constant conditional variance Assumption 1.2.2. Again, we can loosen both assumptions or the linear conditional mean assumption alone through central solution space.

3.2.1 Derivation of $\mathcal{S}_{\mathrm{CSS-DR}}$

Directional Regression synthesizes the dimension reduction methods based on first two conditional moments. B. Li and Wang (2007) provided strong empirical evidence that DR is more accurate in estimating the central space than using SIR or SAVE alone, and that it is comparable with and even superior to the bootstrapped optimal convex combination methods (Ye & Weiss, 2003), even though the former requires substantially less computing time. Let $(\widetilde{X}, \widetilde{Y})$ be an independent copy of (X, Y), and let

$$A(Y, \widetilde{Y}) = E[(X - \widetilde{X})(X - \widetilde{X})^T | Y, \widetilde{Y}].$$

As shown in Li and Wang (2007), if Assumption 1.1.1, 1.2.1 and 1.2.2 are satisfied, then the column space of $2I_p - A(Y, \widetilde{Y})$ is contained in $\mathcal{S}_{Y|X}$ almost surely. Consequently, the column space of the matrix $E[2I_p - A(Y, \widetilde{Y})]^2$ is a subspace of $\mathcal{S}_{Y|X}$. This column space is called the Directional Regression subspace, and will be written as $\mathcal{S}_{\mathrm{DR}}$.

The equation for CSS-DR can be derived similarly to CSS-SAVE. First note that

$$\begin{aligned} &E[(X-\widetilde{X})(X-\widetilde{X})^T|Y,\widetilde{Y}] \\ &\qquad = E\{E[(X-\widetilde{X})(X-\widetilde{X})^T|\beta^T X, \beta^T \widetilde{X}]|Y,\widetilde{Y}\}. \end{aligned} \tag{3.9}$$

Using CCV Assumption 1.2.2, it can be shown that equation (3.9) reduces to

$$\begin{aligned} &2E(XX^T) - E[(X-\widetilde{X})(X-\widetilde{X})^T|Y,\widetilde{Y}] \\ &\qquad = 2E(U_\beta U_\beta^T) - E[(U_\beta-\widetilde{U}_\beta)(U_\beta-\widetilde{U}_\beta)^T|Y,\widetilde{Y}], \end{aligned} \tag{3.10}$$

where $U_\beta = E(X|\beta^T X)$ and $\widetilde{U}_\beta = E(\widetilde{X}|\beta^T\widetilde{X})$. We will call equation (3.10) the Directional Regression equation, or the DR equation.

Definition 3.2.1. *If β is a p-row matrix that satisfies the DR equation (3.10), then* $\mathrm{span}(\beta)$ *is called a solution space of the DR equation. If the intersection of any two solution spaces of (3.10) is itself a solution space of (3.10), then the intersection of all solution spaces of (3.10) will be called the central solution space for the DR equation, written as* $\mathcal{S}_{\mathrm{CSS-DR}}$.

As the counterpart of Theorem 3.1.1, the following theorem describes the relations among $\mathcal{S}_{\mathrm{CSS-DR}}$, $\mathcal{S}_{\mathrm{DR}}$, and $\mathcal{S}_{Y|X}$.

Theorem 3.2.1. *Suppose that Y and the elements of X are square integrable and* $E(X)=0, \mathrm{Var}(X)=I_p$. *In addition, suppose* $\mathrm{Var}(X|\beta^T X)$ *is non-random, where β is a matrix such that* $\mathrm{span}(\beta)=\mathcal{S}_{Y|X}$. *Then*

1. $\mathcal{S}_{\mathrm{CSS-DR}} \subseteq \mathcal{S}_{Y|X}$.

2. *If the linear condition mean assumption and the constant conditional variance assumption hold for both $\mathcal{S}_{\mathrm{CSS-DR}}$ and $\mathcal{S}_{\mathrm{DR}}$, then*
   $\mathcal{S}_{\mathrm{CSS-DR}} = \mathcal{S}_{\mathrm{DR}}$.

The proof of this theorem is similar to Theorem 3.1.1 and is omitted. Again, as long as the constant conditional variance assumption holds, $\mathcal{S}_{\text{CSS}-\text{DR}}$ is a subspace of central space $\mathcal{S}_{Y|X}$. When both Assumption 1.2.1 and 1.2.2 are satisfied, $\mathcal{S}_{\text{CSS}-\text{DR}}$ becomes the same as $\mathcal{S}_{\text{DR}}$ and thus enjoys all the desirable properties of DR, such as high accuracy and exhaustiveness.

3.2.2 Estimation of $\mathcal{S}_{\text{CSS}-\text{DR}}$

Let $(\widetilde{X}, \widetilde{Y})$ be an independent copy of (X, Y). Let

$$A = 2E(XX^T) - E[(X - \widetilde{X})(X - \widetilde{X})^T | Y, \widetilde{Y}],$$

$$B(\eta) = 2E(U_\eta U_\eta^T) - E[(U_\eta - \widetilde{U}_\eta)(U_\eta - \widetilde{U}_\eta)^T | Y, \widetilde{Y}],$$

where $U_\eta = E(X|\eta^T X)$ and $\widetilde{U}_\eta = E(\widetilde{X}|\eta^T \widetilde{X})$. Then the objective function for CSS-DR becomes

$$L_0(\eta) = E\|A - B(\eta)\|^2 = \text{trace}\{E[(A - B(\eta))(A - B(\eta))^T]\}.$$

Parallel to the development for CSS-SAVE, the discretized version of $L_0(\eta)$ is the following

$$L(\eta) = \text{trace}\left[\sum_{i<j} p_{i,j}(A_{i,j} - B_{i,j}(\eta))(A_{i,j} - B_{i,j}(\eta))^T\right],$$

where $p_{i,j} = P(Y \in J_i, \widetilde{Y} \in J_j)$,

$$A_{i,j} = 2E(XX^T) - E[(X - \widetilde{X})(X - \widetilde{X})^T | Y \in J_i, \widetilde{Y} \in J_j], \text{ and}$$
$$B_{i,j}(\eta) = 2E(U_\eta U_\eta^T) - E[(U_\eta - \widetilde{U}_\eta)(U_\eta - \widetilde{U}_\eta)^T | Y \in J_i, \widetilde{Y} \in J_j].$$

As described in Li and Wang (2007), one of the advantages of Directional Regression over Contour Regression is that we can rewrite the relevant formula in a form that does not rely on $(\tilde{X}, \tilde{Y})$, so that the amount of computation is reduced from $O(n^2)$ to $O(n)$. The same reduction also applies to CSS-DR.

By definition $L_0(\eta)$ can be written as

$$L_0(\eta) = \text{trace}\left[E(B(\eta)^T B(\eta))\right] - 2\text{trace}\left[E(A^T B(\eta))\right] + \text{trace}\left[E(A^T A)\right]. \quad (3.11)$$

The next theorem gives the specific expressions of the three terms in (3.11) that do not depend on $(\tilde{X}, \tilde{Y})$.

Theorem 3.2.2. *The three terms in (3.11) can be re-expressed as*

$$\begin{aligned}\text{trace}\left[E(B(\eta)^T B(\eta))\right] &= 2\text{trace}\left\{E[E(U_\eta^T|Y)E(U_\eta|Y)]E[E(U_\eta^T|Y)E(U_\eta|Y)]\right\} \\ &+ 2\text{trace}\left\{E[E(U_\eta|Y)E(U_\eta^T|Y)]E[E(U_\eta|Y)E(U_\eta^T|Y)]\right\} \\ &+ 4\text{trace}E[E(U_\eta U_\eta^T|Y)E(U_\eta U_\eta^T|Y)] - 4\text{trace}\left[E(U_\eta U_\eta^T)E(U_\eta U_\eta^T)\right] \\ &+ 8\text{trace}\left[\mathrm{E}(U_\eta U_\eta^T)E(U_\eta)E(U_\eta^T)\right] - 8\text{trace}\left\{E[E(U_\eta U_\eta^T|Y)E(U_\eta|Y)]E(U_\eta^T)\right\},\end{aligned}$$

$$\begin{aligned}\text{trace}\left[E(A^T A)\right] &= 2\text{trace}\left\{E[E(X^T|Y)E(X|Y)]E[E(X^T|Y)E(X|Y)]\right\} \\ &+ 2\text{trace}\left\{E[E(X|Y)E(X^T|Y)]E[E(X|Y)E(X^T|Y)]\right\} \\ &+ 4\text{trace}\left\{E[E(XX^T|Y)E(XX^T|Y)]\right\} - 4\text{trace}\left[E(XX^T)E(XX^T)\right] \\ &+ 8\text{trace}\left[E(XX^T)E(X)E(X^T)\right] - 8\text{trace}\left\{E[E(XX^T|Y)E(X|Y)]E(X^T)\right\},\end{aligned}$$

and

$$\text{trace}\left[E(A^T B(\eta))\right] = 2\text{trace}\left\{E[E(X|Y)E(U_\eta^T|Y)]E[E(X|Y)E(U_\eta^T|Y)]\right\}$$

$$+ 2\text{trace}\left\{E[E(U_\eta^T|Y)E(X|Y)]E[E(X^T|Y)E(U_\eta|Y)]\right\}$$
$$+ 4\text{trace}\left\{E[E(XX^T|Y)E(U_\eta U_\eta^T|Y)]\right\} + 4\text{trace}\left[E(U_\eta)E(U_\eta^T)\right]$$
$$- 4\text{trace}\left\{E[E(XX^T|Y)E(U_\eta|Y)]E(U_\eta^T)\right\} - 4\text{trace}\left[E(U_\eta U_\eta^T)\right].$$

The discretization and estimation of $L(\eta)$ are parallel to those for CSS-SAVE, and will be omitted.

3.3 Other second-order CSS methods

In this section, we briefly describe the further extensions to other second-order methods, such as PHD, Contour Regression, and SIRII. Following the same strategy as the previous sections, we will relax the linear conditional mean assumption while leaving intact the constant variance assumption.

One form of PHD (Li, 1992) is based on the fact that if Assumption 1.1.1, 1.2.1 and 1.2.2 are satisfied, then

$$E(YXX^T) = P_\beta E(YXX^T)P_\beta.$$

Following the same arguments that lead to equations (3.4) and (3.10) for SAVE and DR, employing the constant conditional variance assumption in the similar fashion, we derive the equation for PHD as

$$E(YXX^T) = E(YU_\beta U_\beta^T).$$

The central solution space of this equation will be denoted by $\mathcal{S}_{\text{CSS}-\text{PHD}}$.

SIRII is based on the fact that, under Assumption 1.1.1, 1.2.1 and 1.2.2,

$$\mathrm{Var}(X|Y) - E[\mathrm{Var}(X|Y)] = P_\beta\{\mathrm{Var}(X|Y) - E[\mathrm{Var}(X|Y)]\}P_\beta.$$

The corresponding central solution space, to be written as $\mathcal{S}_{\mathrm{CSS-SIRII}}$, is defined through the relation

$$\mathrm{Var}(X|Y) - E[\mathrm{Var}(X|Y)] = \mathrm{Var}(U_\beta|Y) - E[\mathrm{Var}(U_\beta|Y)].$$

Turning now to Simple Contour Regression (SCR) of Li et al. (2005), let $(\widetilde{X}, \widetilde{Y})$ be an independent copy of (X, Y) and, for some constant $c > 0$, let

$$K(c) = E\left[(X - \widetilde{X})(X - \widetilde{X})^T \Big| \mid Y - \widetilde{Y} \mid \leq c\right].$$

Suppose X has an elliptical distribution, $E(X) = 0$, and $\mathrm{Var}(X) = I_p$. Then, under mild additional conditions, the eigenvectors of $K(c)$ corresponding to the smallest d eigenvalues span the central space $\mathcal{S}_{Y|X}$, where d is the dimension of $\mathcal{S}_{Y|X}$. The corresponding central solution space, to be written as $\mathcal{S}_{\mathrm{CSS-SCR}}$, is defined through the following equation:

$$\begin{aligned} &2E(XX^T) - E\left[(X - \widetilde{X})(X - \widetilde{X})^T \Big| \mid Y - \widetilde{Y} \mid \leq c\right] \\ &\quad = 2E(U_\beta U_\beta^T) - E\left[(U_\beta - \widetilde{U}_\beta)(U_\beta - \widetilde{U}_\beta)^T \Big| \mid Y - \widetilde{Y} \mid \leq c\right]. \end{aligned}$$

As in the cases of $\mathcal{S}_{\mathrm{CSS-SAVE}}$ and $\mathcal{S}_{\mathrm{CSS-DR}}$, if we write $\mathcal{S}_{\mathrm{CSS-A}}$, where A indicates any of the above second-order methods such as PHD, SIRII, and SCR, then, under the constant variance assumption alone, we have $\mathcal{S}_{\mathrm{CSS-A}} \subseteq \mathcal{S}_{Y|X}$. Under both the linear conditional mean assumption and the constant variance assumption, we have

$\mathcal{S}_{\mathrm{CSS-A}} = \mathcal{S}_A$.

All the dimension reduction equations described above are of the form

$$g(\beta, Y) = 0 \text{ almost surely,}$$

where $g(\beta, Y)$ is a $p \times p$ random matrix. This equation is equivalent to

$$L_0(\beta) = E\|g(\beta, Y)\|^2 = 0,$$

where $\|\cdot\|$ is the Frobenious norm of a matrix. We estimate β by minimizing the sample version of $L_0(\eta)$ like before. Similar procedures to estimate $\mathcal{S}_{\mathrm{CSS-SAVE}}$ and $\mathcal{S}_{\mathrm{CSS-DR}}$ can be developed for all other second-order approaches, and thus omitted.

Chapter 4

Asymptotic analysis

In previous chapters, we have used the idea of central solution space to generalize both first-order and second-order dimension reduction methods. CSS-based estimators can circumvent the linear conditional mean assumption by minimizing objective functions

$$L(\eta) = E\|g(\eta, Y)\|^2,$$

where $g(\eta, Y)$ takes different forms for different dimension reduction methods. Based on an iid sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), we construct the sample estimate $L_n(\eta)$ of $L(\eta)$, and then use the minimizer $\hat{\eta}$ of $L_n(\eta)$ to estimate the central space $\mathcal{S}_{Y|X}$. We will derive the asymptotic distributions of $\hat{\eta}$ in this chapter. We are going to see that $\hat{\eta}$ is $\sqrt{n}$-consistent and asymptotic normal.

Without loss of generality, we will only tackle the asymptotic analysis of CSS-SAVE, and only consider the case when the structure dimension $d = 1$. The development for other estimators can be derived in a similar fashion. For general d, the derivation is fundamentally similar to this special case, but would require rather tedious notation. The original CSS-SAVE equation (3.4) is based on the assumption that $E(X) = 0$. For sample estimation, this is not a problem since we can always centralize X to $\hat{X} = X - E_n(X)$. But for the asymptotic analysis, this additional assumption disrupt the symmetry, and hence the intrinsic simplicity, of the problem. For this reason we will not make this assumption in this chapter. Recall that $U_\beta = E(X|\beta^T X)$. From the fact that $E(X) = E(U_\beta)$, it can be shown easily that

$$\mathrm{Var}(X|Y) - \mathrm{Var}(X) = \mathrm{Var}(U_\beta|Y) - \mathrm{Var}(U_\beta) \tag{4.1}$$

is equivalent to CSS-SAVE equation (3.4) with or without the normalizing assumption.

4.1 Asymptotic normality of CSS-SAVE estimator

We will use Frechet derivatives to derive the desired asymptotic expansion. See Fernholz (1983), Chapter 2, for more details about Frechet derivatives. In particular, we will assume that all estimators of the form $T(F_n)$, where F_n is the empirical distribution based on the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, are asymptotic linear, in the sense that

$$T(F_n) = T(F_0) + E_n\phi(X, Y) + o_P(n^{-1/2}),$$

where $E\phi(X, Y) = 0$ and the elements of $\phi(X, Y)$ have finite variances. This assumption is a mild one, given that the estimators considered here are essentially moment estimators. The quantity $\phi(X, Y)$ is the influence function of T, or (the representation of) the Frechet derivative of T. We will use $T^*(F)$ to denote the Frechet derivative $\phi(X, Y)$. For example, $E_F^*(X) = X - E_F(X)$.

Let $\mathcal{F}$ be a convex set of distributions of (X, Y) which contains the true distribution F_0 and all empirical distributions. Let $E_F(\cdot)$ denote the expectation under

F, and $E(\cdot)$ denote the expectation under F_0. For ease of exposition, we first write $L(\eta)$ and $L_n(\eta)$ as explicit functionals of (η, F). Let $\ell : \mathbb{R}^{p\times d} \times \mathcal{F} \mapsto \mathbb{R}$ denote the functional

$$\ell(\eta, F) = \text{trace}\left\{\sum_{k=1}^{h} p_k(F)[A_k(F) - B_k(\eta, F)][A_k(F) - B_k(\eta, F)]^T\right\}, \tag{4.2}$$

where

$$p_k(F) = E_F I(Y \in J_k),\ I(\cdot) \text{ being the indicator function,}$$
$$A_k(F) = \text{Var}_F(X|Y \in J_k) - \text{Var}_F(X),\ \text{and}$$
$$B_k(\eta, F) = \text{Var}_F(U_\eta|Y \in J_k) - \text{Var}_F(U_\eta).$$

In this notation, $L(\eta)$ and $L_n(\eta)$ in Section 3.1.2 become $\ell(\eta, F_0)$ and $\ell(\eta, F_n)$.

Since η is not uniquely determined by the subspace span(η), the Hessian matrix of $\ell(\eta, F_0)$ is singular. Let

$$g(\eta_0, F) = \left[\frac{\partial \ell(\eta, F)}{\partial \eta}\right]_{\eta=\eta_0},\quad W = W(\eta_0, F_0) = \left[\frac{\partial^2 \ell(\eta, F_0)}{\partial \eta \partial \eta^T}\right]_{\eta=\eta_0}.$$

Let P_W be the projection on to the column space of W, and let $Q_W = I_p - P_W$. By Taylor expansion, it is easy to see that

$$\ell(\eta_0 + n^{-1/2}\delta, F_0) = n^{-1}\delta^T W \delta + o(n^{-1}),$$
$$\ell(\eta_0 + n^{-1/2}P_W\delta, F_0) = n^{-1}\delta^T W \delta + o(n^{-1}).$$

That is, in a contiguity neighborhood of η_0, $\ell(\cdot, F_0)$ is unaffected by the component $Q_W\delta$ of the parameter. In other words, locally at η_0, it is $P_W\delta$ that parameterizes the subspace span$(\eta_0 + n^{-1/2}P_W\delta)$, and the component $Q_W\delta$ has no effect on this subspace. Similarly, at the sample level, it can be shown that

$$\ell(\hat{\eta}, F_n) = \ell(\eta_0 + P_W(\hat{\eta} - \eta_0), F_n) + o_p(n^{-1}), \quad \ell(\hat{\eta}, F_n) = O_p(n^{-1}).$$

Thus $Q_W(\hat{\eta} - \eta_0)$ has no effect on the sample objective function $\ell(\cdot, F_n)$. For this reason, the relevant asymptotic distribution is that of $\sqrt{n}P_W(\hat{\eta} - \eta_0)$, rather than that of $\sqrt{n}(\hat{\eta} - \eta_0)$. The next theorem gives the asymptotic expansion of the former.

Theorem 4.1.1. *Let $W^\dagger$ be the Moore-Penrose inverse of W. Then, under regularity conditions,*

$$P_W(\hat{\eta} - \eta_0) = -W^\dagger E_n g^*(X, Y, \eta_0, F_0) + o_p(n^{-1/2}), \tag{4.3}$$

where W is given by (4.18), and $g^*(X, Y, \eta_0, F_0)$ is given by expressions (4.5) through (4.17) in the next section.

From expansion (4.3) we can easily derive the asymptotic distributions of $\sqrt{n}P_W(\hat{\eta} - \eta_0)$, as follows.

Corollary 4.1.1. *Under regularity conditions,*

$$\sqrt{n}P_W(\hat{\eta} - \eta_0) \xrightarrow{\mathcal{D}} N(0, \Lambda(\eta_0, F_0)),$$

where $\Lambda(\eta_0, F_0) = W^\dagger E\{g^*(X, Y, \eta_0, F_0)[g^*(X, Y, \eta_0, F_0)]^T\}W^\dagger$.

The regularity conditions for Theorem 4.1.1 and Corollary 4.1.1 are quite mild. Essentially they amount to Frechet differentiability. To estimate $\Lambda(\eta_0, F_0)$, we simply replace W by its sample estimate $W(\hat{\eta}, F_n)$ and replace

$$E\{g^*(X, Y, \eta_0, F_0)[g^*(X, Y, \eta_0, F_0)]^T\}$$

by $n^{-1}\sum_{i=1}^{n}\{g^*(X_i,Y_i,\hat{\eta},F_n)[g^*(X_i,Y_i,\hat{\eta},F_n)]^T\}$.

4.2 Proof of Theorem 4.1.1

Let $\eta(F)$ be the minimizer of $\ell(\eta, F)$ as defined in (4.2). We will abbreviate the partial differentiation $\partial/\partial\eta$ by ∂_η. Recalling that $g(\eta, F) = \partial_\eta \ell(\eta, F)$, we have

$$g(\eta(F), F) = 0$$

for all $F \in \mathcal{F}$. Take Frechet derivative on both sides to obtain,

$$W\eta^*(F_0) + g^*(\eta_0, F_0) = 0,$$

where W is as defined above Theorem 4.1.1. Here, $g^*(\eta_0, F_0)$ is to be understood as the Frechet derivative of $F \to g(\eta_0, F)$ at $F = F_0$ (more precisely, $g^*(\eta_0, F_0)$ should be $g^*(X, Y, \eta_0, F_0)$; we suppress the dependence on X, Y in the proof for simplicity). Multiply both sides of the above equality by $W^\dagger$, and use the relation $W^\dagger W = P_W$, to obtain

$$P_W\eta^*(F_0) = -W^\dagger g^*(\eta_0, F_0). \tag{4.4}$$

It remains to compute the $p\times p$ nonrandom matrix W and the p-dimensional random vector $g^*(\eta_0, F_0)$.

By definition, the tth component of $g(\eta_0, F) = \partial_\eta \ell(\eta_0, F)$ is

$$g_t(\eta_0, F) = -2\text{trace}\left\{\sum_{k=1}^{h} p_k(F)[\partial_{\eta_t} B_k(\eta_0, F)][A_k(F) - B_k(\eta_0, F)]^T\right\}.$$

Since $A_k(F) = B_k(\eta, F)$ at $(\eta, F) = (\eta_0, F_0)$, the tth component of $g^*(\eta_0, F_0)$ is

$$\begin{aligned} &g_t^*(\eta_0, F_0) \\ &= -2\text{trace}\left\{\sum_{k=1}^{h} p_k(F_0)[\partial_{\eta_t} B_k(\eta_0, F_0)][A_k^*(F_0) - B_k^*(\eta_0, F_0)]^T\right\}. \end{aligned} \tag{4.5}$$

We now derive the explicit expressions for $A_k^*(F_0)$, $B_k^*(\eta_0, F_0)$, and $\partial_{\eta_t} B_k(\eta_0, F_0)$, in that order.

Let $R_k = I(Y \in J_k)$. Then $p_k = p_k(F_0) = ER_k$ and terms such as $E_F(X|Y \in J_k)$ become $E_F(XR_k)/E_F R_k$. The term $A_k^*(F_0)$ in (4.5) is:

$$\begin{aligned} A_k^*(F_0) &= E^*(XX^T|Y \in J_k) - [E(X|Y \in J_k)E^T(X|Y \in J_k)]^* \\ &- E^*(XX^T) + [E(X)E^T(X)]^*. \end{aligned} \tag{4.6}$$

The first signed term on the right hand side of (4.6) is

$$\begin{aligned} E^*(XX^T|Y \in J_k) &= E^*(XX^T R_k)/ER_k + E(XX^T R_k)(1/ER_k)^* \\ &= XX^T R_k/p_k - E(XX^T R_k)/p_k - E(XX^T R_k)(R_k - p_k)/p_k^2. \end{aligned}$$

The second signed term is

$$\begin{aligned} &- [E(X|Y \in J_k)E^T(X|Y \in J_k)]^* = -[E^*(X|Y \in J_k)E^T(X|Y \in J_k) \\ &+ E(X|Y \in J_k)E^*(X^T|Y \in J_k)], \end{aligned}$$

where

$$E^*(X|Y \in J_k) = XR_k/p_k - E(XR_k)/p_k - E(XR_k)(R_k - p_k)/p_k^2.$$

The last two signed terms are

$$-[XX^T - E(XX^T)] + [XE^T(X) + E(X)X^T - 2E(X)E^T(X)].$$

To compute the term $B_k^*(\eta_0, F_0)$ in (4.5), let G abbreviate the random vector $G(\eta_0^T X)$ defined in (2.15). Then $B_k(\eta_0, F)$ can be expressed as

$$B_{1k}(\eta_0, F) - B_{2k}(\eta_0, F)B_{2k}^T(\eta_0, F) - B_{3k}(\eta_0, F) + B_{4k}(\eta_0, F)B_{4k}^T(\eta_0, F),$$

where

$$\begin{aligned}
B_{1k}(\eta_0, F) &= E_F(XG^T)E_F^{-1}(GG^T)E_F(GG^T|Y \in J_k)E_F^{-1}(GG^T)E_F(GX^T),\\
B_{2k}(\eta_0, F) &= E_F(XG^T)E_F^{-1}(GG^T)E_F(G|Y \in J_k),\\
B_{3k}(\eta_0, F) &= E_F(XG^T)E_F^{-1}(GG^T)E_F(GX^T),\\
B_{4k}(\eta_0, F) &= E_F(XG^T)E_F^{-1}(GG^T)E_F(G).
\end{aligned}$$

Taking Frechet derivative, we have

$$\begin{aligned}
B_k^*(\eta_0, F_0) = {} & B_{1k}^*(\eta_0, F_0) - B_{2k}^*(\eta_0, F_0)B_{2k}^T(\eta_0, F_0)\\
& - B_{2k}(\eta_0, F_0)[B_{2k}^*(\eta_0, F_0)]^T - B_{3k}^*(\eta_0, F_0)\\
& + B_{4k}^*(\eta_0, F_0)B_{4k}^T(\eta_0, F_0) + B_{4k}(\eta_0, F_0)[B_{4k}^*(\eta_0, F_0)]^T.
\end{aligned} \tag{4.7}$$

The Frechet derivatives of B_{1k}, B_{2k}, B_{3k} and B_{4k} in (4.7) are

$$
\begin{aligned}
B^*_{1k}(\eta_0, F_0) &= E^*(XG^T)E^{-1}(GG^T)E(GG^T|Y \in J_k)E^{-1}(GG^T)E(GX^T)\\
&+ E(XG^T)[E^{-1}(GG^T)]^*E(GG^T|Y \in J_k)E^{-1}(GG^T)E(GX^T)\\
&+ E(XG^T)E^{-1}(GG^T)E^*(GG^T|Y \in J_k)E^{-1}(GG^T)E(GX^T)\\
&+ E(XG^T)E^{-1}(GG^T)E(GG^T|Y \in J_k)[E^{-1}(GG^T)]^*E(GX^T)\\
&+ E(XG^T)E^{-1}(GG^T)E(GG^T|Y \in J_k)E^{-1}(GG^T)E^*(GX^T),
\end{aligned}
\tag{4.8}
$$

$$
\begin{aligned}
B^*_{2k}(\eta_0, F_0) &= E^*(XG^T)E^{-1}(GG^T)E(G|Y \in J_k)\\
&+ E(XG^T)[E^{-1}(GG^T)]^*E(G|Y \in J_k)\\
&+ E(XG^T)E^{-1}(GG^T)E^*(G|Y \in J_k),
\end{aligned}
\tag{4.9}
$$

$$
\begin{aligned}
B^*_{3k}(\eta_0, F_0) &= E^*(XG^T)E^{-1}(GG^T)E(GX^T)\\
&+ E^*(XG^T)[E^{-1}(GG^T)]^*E(GX^T)\\
&+ E(XG^T)E^{-1}(GG^T)E^*(GX^T),
\end{aligned}
\tag{4.10}
$$

and

$$
\begin{aligned}
B^*_{4k}(\eta_0, F_0) &= E^*(XG^T)E^{-1}(GG^T)E(G)\\
&+ E(XG^T)[E^{-1}(GG^T)]^*E(G)\\
&+ E(XG^T)E^{-1}(GG^T)E^*(G).
\end{aligned}
\tag{4.11}
$$

The Frechet derivatives involved in (4.8) through (4.11) are

$$
\begin{aligned}
&E^*(G) = G - E(G),\\
&E^*(XG^T) = XG^T - E(XG^T),\\
&E^*(GG^T) = GG^T - E(GG^T),\\
&[E^{-1}(GG^T)]^* = -E^{-1}(GG^T)E^*(GG^T)E^{-1}(GG^T),\\
&E^*(G|Y \in J_k) = E^*(GR_k)/ER_k + E(GR_k)(1/ER_k)^*\\
&\quad = GR_k/p_k - E(GR_k)/p_k - E(GR_k)(R_k - p_k)/p_k^2,\\
&E^*(GG^T|Y \in J_k) = E^*(GG^TR_k)/ER_k + E(GG^TR_k)(1/ER_k)^*\\
&\quad = GG^TR_k/p_k - E(GG^TR_k)/p_k - E(GG^TR_k)(R_k - p_k)/p_k^2.
\end{aligned}
$$

For $\partial_{\eta t} B_k(\eta_0, F_0)$ in (4.5), we have

$$
\begin{aligned}
\partial_{\eta t} B_k(\eta_0, F_0) = {} & \partial_{\eta t} B_{1k}(\eta_0, F_0) - [\partial_{\eta t} B_{2k}(\eta_0, F_0)]B_{2k}^T(\eta_0, F_0)\\
& - B_{2k}(\eta_0, F_0)[\partial_{\eta t} B_{2k}(\eta_0, F_0)]^T - \partial_{\eta t} B_{3k}(\eta_0, F_0)\\
& + [\partial_{\eta t} B_{4k}(\eta_0, F_0)]B_{4k}^T(\eta_0, F_0) + B_{4k}(\eta_0, F_0)[\partial_{\eta t} B_{4k}(\eta_0, F_0)]^T,
\end{aligned}
\tag{4.12}
$$

where

$$
\begin{aligned}
\partial_{\eta t} B_{1k}(\eta_0, F_0) = {} & [\partial_{\eta t} E(XG^T)]E^{-1}(GG^T)E(GG^T|Y \in J_k)E^{-1}(GG^T)E(GX^T)\\
& + E(XG^T)[\partial_{\eta t} E^{-1}(GG^T)]E(GG^T|Y \in J_k)E^{-1}(GG^T)E(GX^T)\\
& + E(XG^T)E^{-1}(GG^T)[\partial_{\eta t} E(GG^T|Y \in J_k)]E^{-1}(GG^T)E(GX^T)\\
& + E(XG^T)E^{-1}(GG^T)E(GG^T|Y \in J_k)[\partial_{\eta t} E^{-1}(GG^T)]E(GX^T)\\
& + E(XG^T)E^{-1}(GG^T)E(GG^T|Y \in J_k)E^{-1}(GG^T)[\partial_{\eta t} E(GX^T)],
\end{aligned}
\tag{4.13}
$$

$$\begin{aligned}\partial_{\eta_t} B_{2k}(\eta_0, F_0) &= [\partial_{\eta_t} E(XG^T)]E^{-1}(GG^T)E(G|Y \in J_k) \\ &+ E(XG^T)[\partial_{\eta_t} E^{-1}(GG^T)]E(G|Y \in J_k) \\ &+ E(XG^T)E^{-1}(GG^T)[\partial_{\eta_t} E(G|Y \in J_k)],\end{aligned} \tag{4.14}$$

$$\begin{aligned}\partial_{\eta_t} B_{3k}(\eta_0, F_0) &= [\partial_{\eta_t} E(XG^T)]E^{-1}(GG^T)E(GX^T) \\ &+ E(XG^T)[\partial_{\eta_t} E^{-1}(GG^T)]E(GX^T) \\ &+ E(XG^T)E^{-1}(GG^T)[\partial_{\eta_t} E(GX^T)],\end{aligned} \tag{4.15}$$

and

$$\begin{aligned}\partial_{\eta_t} B_{4k}(\eta_0, F_0) &= [\partial_{\eta_t} E(XG^T)]E^{-1}(GG^T)E(G|Y) \\ &+ E(XG^T)[\partial_{\eta_t} E^{-1}(GG^T)]E(G) + E(XG^T)E^{-1}(GG^T)[\partial_{\eta_t} E(G)].\end{aligned} \tag{4.16}$$

The partial derivatives involved in (4.13) through (4.16) can be easily calculated for $d = 1$. For general d we can use vec(η) in place of η and follow the similar steps.

When $d = 1$, $G(\eta_0^T X)$ becomes $\left(1, \eta_0^T X, (\eta_0^T X)^2\right)^T$, where $\eta_0 = (\eta_1, \ldots, \eta_p)^T$ and $X = (X_1, \ldots, X_p)^T$ are both $p \times 1$ vectors. We have

$$\begin{aligned}&\partial_{\eta_t} G = \left(0, X_t, 2X_t(\eta_0^T X)\right)^T, \\ &\partial_{\eta_t} E(XG^T) = E[X(\partial_{\eta_t} G)^T], \\ &\partial_{\eta_t}(GG^T) = (\partial_{\eta_t} G)G^T + G(\partial_{\eta_t} G)^T, \\ &\partial_{\eta_t} E^{-1}(GG^T) = -E^{-1}(GG^T)E[\partial_{\eta_t}(GG^T)]E^{-1}(GG^T), \\ &\partial_{\eta_t} E(G|Y \in J_k) = E[(\partial_{\eta_t} G)R_k]/ER_k, \\ &\partial_{\eta_t} E(GG^T|Y \in J_k) = E[\partial_{\eta_t}(GG^T)R_k]/ER_k.\end{aligned} \tag{4.17}$$

Note that each formula may involve the expressions in the previous formulas.

We now calculate the Hessian matrix W. Because $A_k(F) = B_k(\eta, F)$ at $(\eta, F) = (\eta_0, F_0)$, and because $A_k(F)$ does not depend on η, the (t, u)-th element of $\partial_\eta^2 \ell(\eta_0, F_0)$ is

$$W_{tu} = 2\text{trace}\left\{ \sum_{k=1}^{h} p_k [\partial_{\eta_t} B_k(\eta_0, F_0)][\partial_{\eta_u} B_k(\eta_0, F_0)]^T \right\}, \tag{4.18}$$

where $\partial_{\eta_t} B_k(\eta_0, F_0)$ and $\partial_{\eta_u} B_k(\eta_0, F_0)$ can be calculated from (4.12) through (4.17).
$\square$

Chapter 5

Simulation study

In this chapter, extensive simulation studies are carried out to compare the performances of CSS-based estimators with inverse conditional moments estimators as well as two adaptive estimators, MAVE and OPG. We can see that CSS-based estimators inherit the properties of their classical counterparts, while CSS-based estimators improve the estimation accuracy a lot in the presence of non-elliptical predictors.

5.1 Simulation study of first-order methods

In this section, we will compare the first-order CSS-based methods with their classical counterparts as well as two adaptive estimators. Predictor X will have a non-elliptical distribution so the inverse conditional moments methods will no longer work.

5.1.1 When we know exact function forms for $E(X|\beta^T X)$

We will first consider the following three models:

$$\text{Model 1:} \quad Y = e^{X_1} + (X_2 + 1.5)^2 + \epsilon,$$

$$\text{Model 2:} \quad Y = 0.4X_1^2 + 3\sin(X_2/4) + 0.5\epsilon,$$

$$\text{Model 3:} \quad Y = X_1/[0.5 + (X_2 + 1.5)^2] + 0.1\epsilon,$$

where $\epsilon \sim N(0,1)$ and $\epsilon \perp\!\!\!\perp X$. We first take the sample size to be $n = 100$. The dimensions of X are chosen to be $p = 4, 6, 8$. Note that in all three models $d = 2$, and $\mathcal{S}_{Y|X}$ is spanned by $(1,0,0,\ldots,0)^T$ and $(0,1,0,0,\ldots,0)^T$.

We introduce nonlinearity in the predictor as follows: $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$,

$$\begin{aligned} X_3 =& 0.2X_1 + 0.2(X_2 + 2)^2 + 0.2\delta, \\ X_4 =& 0.1 + 0.1(X_1 + X_2) + 0.3(X_1 + 1.5)^2 + 0.2\delta, \end{aligned} \tag{5.1}$$

where $\delta \perp\!\!\!\perp (X, Y)$ and $\delta \sim N(0,1)$. When $p = 6, 8$, X_5 through X_8 are taken to be independent $N(0,1)$, and to be independent of $(X_1, \ldots, X_4)$. Figure 5.1 shows the scatter plot matrix of $X_1, \ldots, X_4$. Predictors of this type are very common in practice. As we will see later in Section 6.1.

We apply three methods based on central solution space, CSS-SIR, CSS-PIR, CSS-KIR, as well as their classical counterparts, SIR, PIR, KIR, to the three models. Because CSS-OLS and OLS can only estimate one-dimensional central spaces ($d = 1$), we do not include them in the comparison. We also compare with Outer Product Gradient estimator and the Minimum Averaged Variance Estimator. The simulation sample size is $N = 200$. For SIR and CSS-SIR, the number of slices is taken to be 10, with each slice having equal number of observations. For PIR and CSS-PIR, the function $H(Y)$ described in Section 1.3.5 is

$$H(Y) = (1, Y, Y^2).$$

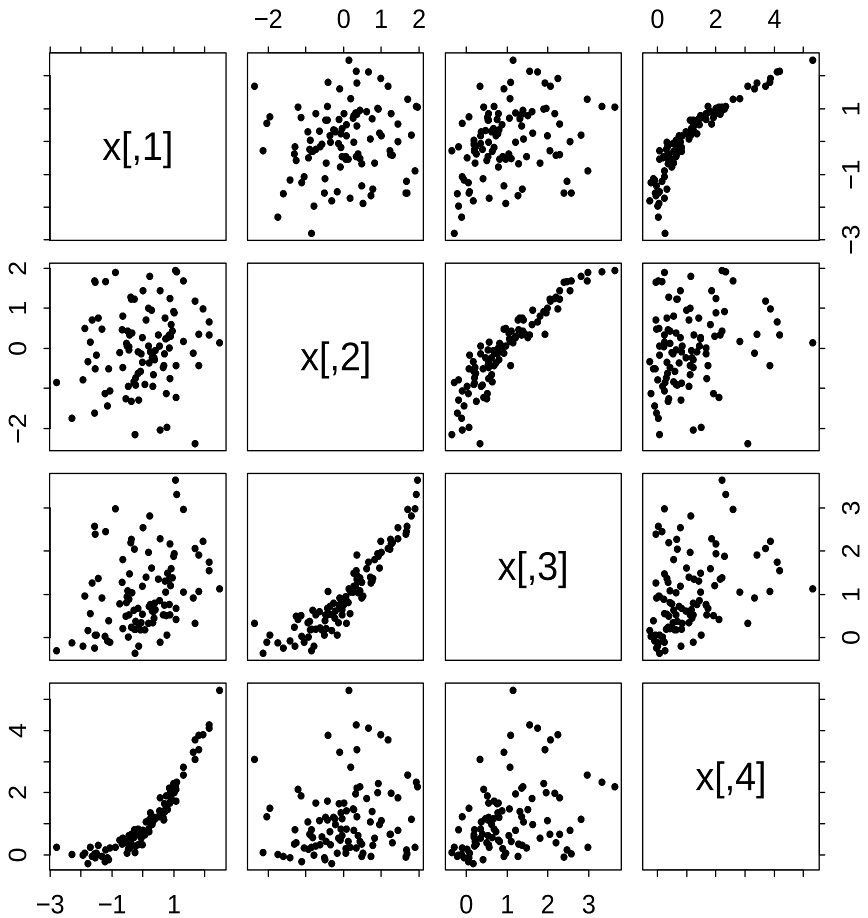


Figure 5.1. Scatter plot matrix for the 4-dimensional nonelliptically distributed predictor X.

For all three CSS methods, the function $G(\eta^T X)$, which is first defined by (2.15) in Section 2.5, is taken to be

$$G(\eta^T X) = \left(1, \eta_1^T X, \eta_2^T X, (\eta_1^T X)^2, (\eta_1^T X)(\eta_2^T X), (\eta_2^T X)^2\right). \tag{5.2}$$

For the KIR and CSS-KIR, the function ψ in (2.6) is taken to be the standard normal density, and the bandwidth h in (2.6) is taken to be 0.4. The kernel function for OPG and MAVE is taken to be the normal density, with standard deviation (kernel width) taken to be 0.7 for $p = 4, 6$ and 0.8 for $p = 8$. These parameters perform reasonably well in several pilot trial runs.

To assess the accuracy of each method we use the squared multiple correlation coefficient. Specifically, suppose U and V are d dimensional random vectors, and Σ_{UV}, Σ_U and Σ_V are the covariance matrix between U and V, the covariance matrix of U, and the covariance matrix of V, respectively. Then the square multiple correlation coefficient is defined by

$$\rho^2 = \text{trace}\left[\Sigma_U^{-1/2}\Sigma_{UV}\Sigma_V^{-1}\Sigma_{VU}\Sigma_V^{-1/2}\right] = \text{trace}\left[\Sigma_V^{-1/2}\Sigma_{VU}\Sigma_U^{-1}\Sigma_{UV}\Sigma_V^{-1/2}\right]. \qquad (5.3)$$

See Hall and Mathiason (1990). The measure takes maximum value d if U and V have a linear relation, and takes minimum 0 if the components of U and V are uncorrelated. At the sample level, given an estimator $\hat{\beta}$ of β, we use the sample version of the above measure based on

$$\{\hat{\beta}^T X_1, \ldots, \hat{\beta}^T X_n\} \ \text{ and } \ \{\beta^T X_1, \ldots, \beta^T X_n\}.$$

Note that the larger value of this criterion corresponds to a better dimension reduction estimate. This criterion seems to be more reasonable than the projection distance criterion used in Li, Zha, and Chiaromonte (2005), because the former directly reflects the behavior of the estimated predictor and because it takes into account the covariance structure of X.

We compute the errors of estimation by the eight methods, for three models and three choices of p, and across the 200 simulated samples. The results are presented in Table 5.1.

Each entry of Table 5.1 is formatted as $a(b)$, where a is the average of the above

Model	Method	$p=4$	$p=6$	$p=8$
1	PIR	1.053 (0.007)	1.050 (0.006)	1.025 (0.006)
	CSS-PIR	1.784 (0.016)	1.510 (0.020)	1.264 (0.018)
	SIR	1.070 (0.009)	1.053 (0.001)	1.019 (0.008)
	CSS-SIR	1.881 (0.013)	1.401 (0.022)	1.246(0.018)
	KIR	1.114 (0.009)	1.106 (0.008)	1.072 (0.009)
	CSS-KIR	1.956 (0.007)	1.624 (0.020)	1.443 (0.019)
	OPG	1.785 (0.017)	1.462 (0.022)	1.322 (0.020)
	MAVE	1.772 (0.016)	1.602 (0.018)	1.241 (0.018)
2	PIR	0.972 (0.011)	0.946 (0.010)	0.930 (0.009)
	CSS-PIR	1.833 (0.017)	1.560 (0.024)	1.284 (0.025)
	SIR	0.959 (0.013)	0.913 (0.010)	0.884 (0.009)
	CSS-SIR	1.910 (0.011)	1.436 (0.024)	1.222 (0.023)
	KIR	1.004 (0.013)	0.983 (0.012)	0.961 (0.012)
	CSS-KIR	1.937 (0.011)	1.598 (0.023)	1.325 (0.024)
	OPG	1.771 (0.015)	1.567 (0.019)	1.310 (0.019)
	MAVE	1.748 (0.018)	1.471 (0.021)	1.213 (0.021)
3	PIR	1.479 (0.012)	1.422 (0.010)	1.404 (0.009)
	CSS-PIR	1.916 (0.006)	1.849 (0.008)	1.781 (0.008)
	SIR	1.667 (0.007)	1.646 (0.007)	1.613 (0.006)
	CSS-SIR	1.931 (0.004)	1.828 (0.007)	1.805 (0.007)
	KIR	1.530 (0.011)	1.477 (0.010)	1.462 (0.009)
	CSS-KIR	1.960 (0.003)	1.828 (0.008)	1.775 (0.008)
	OPG	1.926 (0.008)	1.893 (0.007)	1.831 (0.009)
	MAVE	1.966 (0.002)	1.942 (0.003)	1.888 (0.005)

Table 5.1. First-order CSS methods with exact $G(\eta^T X)$.

criterion across the 200 simulated samples, and b is the standard error of the average. From the table we see that with fixed sample size n and increasing p, the performance of all estimators deteriorate. However, the CSS-based methods are substantially more accurate than their classical counterparts across all 3 models for different p. OPG and MAVE do not require elliptical distribution assumption and remain competitive. Still, CSS-PIR has similar performance with those two adaptive methods and CSS-KIR outperforms OPG and MAVE most of the cases. The performance of CSS-SIR is not as well, which might be partly due to the fact that slicing is somewhat inefficient without using the intra-slice information — an aspect that cannot be improved by the CSS correction. The loss of intra-slice information by SIR is noticed by Cook and Ni (2005), who proposed a method to reduce it.

5.1.2 When we don't know exact function forms for $E(X|\beta^T X)$

Now we switch the predictors of Y from X_1, X_2 to X_3, X_4, while keeping the same nonlinearity among the components of X. The central subspace is now spanned by $(0, 0, 1, 0, \ldots)$ and $(0, 0, 0, 1, \ldots)$. The regression models remain the same except this switch; namely

$$\begin{aligned}
&\text{Model 4:} \quad Y = e^{X_3} + (X_4 + 1.5)^2 + \epsilon, \\
&\text{Model 5:} \quad Y = 0.4X_3^2 + 3\sin(X_4/4) + 0.5\epsilon, \\
&\text{Model 6:} \quad Y = X_3/[0.5 + (X_4 + 1.5)^2] + 0.1\epsilon,
\end{aligned}$$

where non-ellipticity among predictor X is still set as (5.1).

In this new arrangement, $E(X|\beta^T X)$ no longer lies within the space spanned by the quadratic polynomials of $\beta^T X$; thus the quadratic parameterization of $G(\eta^T X)$ by (5.2) is only approximate. More specifically,

$$E(X|\eta^T X) = E[XG^T(\eta^T X)] \left\{E\left[G(\eta^T X)G^T(\eta^T X)\right]\right\}^{-1} G(\eta^T X)$$

no longer holds true when we set $G(\eta^T X)$ by (5.2). Our CSS-based sample estimators are only approximations and will no longer be consistent. However, from Table 5.2, we can clearly see that the CSS-based methods are still more accurate than their classical counterparts across all 9 cases, indicating that there is much to be gained by correcting the bias caused by non-ellipticity. OPG and MAVE are not as sharp as the CSS-based methods. In particular, the accuracy of CSS-KIR dominates that of OPG and MAVE in all 9 cases by substantial margins (relative to the standard deviations). CSS-PIR also performs better than OPG and MAVE in most (8 out of 9) cases. The performance of CSS-SIR is somewhat similar to OPG and MAVE.

Model	Method	$p=4$	$p=6$	$p=8$
4	PIR	1.366 (0.017)	1.336 (0.017)	1.264 (0.015)
	CSS-PIR	1.658 (0.021)	1.631 (0.017)	1.393 (0.017)
	SIR	1.112 (0.013)	1.100 (0.011)	1.064 (0.007)
	CSS-SIR	1.735 (0.018)	1.423 (0.020)	1.293(0.019)
	KIR	1.701 (0.014)	1.661 (0.015)	1.618 (0.015)
	CSS-KIR	1.832 (0.010)	1.711 (0.014)	1.637 (0.017)
	OPG	1.581 (0.023)	1.377 (0.020)	1.282 (0.016)
	MAVE	1.785 (0.016)	1.602 (0.018)	1.382 (0.017)
5	PIR	1.400 (0.015)	1.346 (0.015)	1.349 (0.013)
	CSS-PIR	1.755 (0.018)	1.558 (0.021)	1.476 (0.021)
	SIR	1.302 (0.022)	1.256 (0.017)	1.208 (0.017)
	CSS-SIR	1.789 (0.013)	1.439 (0.021)	1.333 (0.021)
	KIR	1.514 (0.018)	1.468 (0.016)	1.437 (0.015)
	CSS-KIR	1.794 (0.015)	1.551 (0.022)	1.480 (0.020)
	OPG	1.604 (0.023)	1.406 (0.023)	1.302 (0.020)
	MAVE	1.622 (0.022)	1.397 (0.021)	1.265 (0.018)
6	PIR	1.149 (0.014)	1.115 (0.011)	1.065 (0.009)
	CSS-PIR	1.839 (0.014)	1.694 (0.018)	1.557 (0.020)
	SIR	1.265 (0.020)	1.171 (0.014)	1.116 (0.013)
	CSS-SIR	1.833 (0.008)	1.552 (0.020)	1.454 (0.019)
	KIR	1.146 (0.014)	1.113 (0.011)	1.063 (0.009)
	CSS-KIR	1.862 (0.013)	1.705 (0.019)	1.613 (0.019)
	OPG	1.742 (0.017)	1.584 (0.022)	1.453 (0.020)
	MAVE	1.803 (0.016)	1.584 (0.021)	1.375 (0.019)

Table 5.2. First-order CSS methods with inexact $G(\eta^T X)$.

5.1.3 When sample size increases

For larger sample sizes, the performances of all estimators improve, and MAVE and the CSS-based methods become more similar. Under Model 4, Table 5.3 compares CSS-KIR with KIR, OPG, and MAVE for $p = 6$ and $n = 200, 300, 400, 500$. The kernel width (of X) for OPG and MAVE are taken to be $0.6, 0.5, 0.4, 0.4$, and the kernel width (of Y) for KIR and CSS-KIR are $0.3, 0.2, 0.1, 0.1$. The basis functions in $H(y)$ now include third polynomials, and the basis functions in $G(\eta^T X)$ include fourth polynomials. We see that, while OPG and KIR still trail behind CSS-KIR, MAVE catches up with CSS-KIR at around $n = 400$ and surpasses it at $n = 500$. This is because, as we can see from (5.1), the dependence of X_1 and X_2 on X_3 and X_4 involves the square root function, and as a consequence $E(X|X_3, X_4)$ does not belong to the polynomials of $\eta^T X$.

Method	$n = 200$	$n = 300$	$n = 400$	$n = 500$
KIR	1.704 (0.011)	1.725 (0.010)	1.797 (0.005)	1.781 (0.005)
CSS-KIR	1.816 (0.009)	1.846 (0.005)	1.854 (0.004)	1.861 (0.004)
OPG	1.506 (0.023)	1.614 (0.022)	1.681 (0.020)	1.730 (0.021)
MAVE	1.824 (0.014)	1.885 (0.012)	1.847 (0.013)	1.922 (0.009)

Table 5.3. First-order CSS methods with increasing sample size.

5.2 Simulation study of second-order methods

In this section we compare the second-order CSS methods with their classical counterparts when the distribution of X is non-elliptical. First-order methods as well as two adaptive methods are also included for a more thorough comparison.

5.2.1 When we know exact function forms for $E(X|\beta^T X)$

In the first simulation, the following three models are used:

$$\text{Model 7:} \quad Y = 3\sin(X_1/4) + 3\sin(X_2/4) + \epsilon,$$

$$\text{Model 8:} \quad Y = X_1^2 + 3\sin(X_2/4) + 0.5\epsilon,$$

$$\text{Model 9:} \quad Y = X_1/[0.5 + (X_2 + 1.5)^2] + 0.1\epsilon,$$

where $\epsilon \sim N(0, 1)$ and $\epsilon \perp\!\!\!\perp X$. We first take the sample size to be $n = 200$ for different dimension p, and then increase n for a fixed p. The dimensions of X are chosen to be $p = 4$, 5, 6. Note that in all three models $d = 2$, and $\mathcal{S}_{Y|X}$ is spanned by $(1, 0, \ldots, 0)^T$ and $(0, 1, 0, \ldots, 0)^T$. We introduce nonlinearity in the predictor by (5.1) as in the previous sections.

There are several points worth noticing here. We replace Model 1 with Model 7 because both components in Model 7 are monotone and ideally should favor CSS-SIR. Model 8 is slightly different from Model 2. We only change the coefficient to make both components have comparable variance. Model 8 should favor CSS-SAVE and CSS-DR since both monotone and quadratic components are involved. Model 9 is the same as Model 3. A ratio of two components are involved and it is hard to say which method is favored. Please also notice that, since we have more parameters to estimate for second-order methods, such as CSS-SAVE and CSS-DR, we use larger sample size and smaller p here compared with the simulation in Section 5.1.1.

We apply three methods based on the central solution space, CSS-SIR, CSS-SAVE and CSS-DR, as well as their classical counterparts, SIR, SAVE and DR, to these models. OPG and MAVE are also included in the simulation. The simulation sample size is $N = 200$. For SIR and CSS-SIR, the number of slices is taken to be

Model	Method	$p = 4$	$p = 5$	$p = 6$
7	SIR	1.115(0.012)	1.108(0.012)	1.082(0.009)
	CSS-SIR	1.837(0.018)	1.466(0.022)	1.342(0.021)
	SAVE	1.008(0.015)	0.957(0.012)	0.961(0.014)
	CSS-SAVE	1.957(0.012)	1.812(0.014)	1.723(0.016)
	DR	1.155(0.012)	1.148(0.012)	1.144(0.011)
	CSS-DR	1.933(0.014)	1.763(0.017)	1.614(0.021)
	OPG	1.483(0.021)	1.291(0.021)	1.114(0.018)
	MAVE	1.451(0.021)	1.319(0.020)	1.171(0.022)
8	SIR	0.921(0.007)	0.909(0.007)	0.890(0.008)
	CSS-SIR	1.957(0.005)	1.867(0.010)	1.812(0.014)
	SAVE	1.119(0.011)	1.120(0.011)	1.110(0.011)
	CSS-SAVE	1.987(0.001)	1.879(0.005)	1.842(0.006)
	DR	1.080(0.010)	1.055(0.011)	1.063(0.012)
	CSS-DR	1.979(0.005)	1.920(0.008)	1.890(0.008)
	OPG	1.860(0.014)	1.784(0.017)	1.625(0.018)
	MAVE	1.869(0.011)	1.740(0.017)	1.636(0.017)
9	SIR	1.673(0.005)	1.679(0.005)	1.665(0.005)
	CSS-SIR	1.951(0.003)	1.888(0.006)	1.878(0.005)
	SAVE	1.142(0.015)	1.163(0.017)	1.146(0.016)
	CSS-SAVE	1.989(0.0004)	1.893(0.004)	1.879(0.004)
	DR	1.447(0.009)	1.464(0.010)	1.448(0.009)
	CSS-DR	1.979(0.001)	1.927(0.005)	1.906(0.004)
	OPG	1.953(0.008)	1.950(0.007)	1.923(0.007)
	MAVE	1.971(0.005)	1.925(0.012)	1.921(0.007)

Table 5.4. Second-order CSS methods with exact $G(\eta^T X)$.

10. For SAVE, DR, CSS-SAVE and CSS-DR, the number of slices is taken to be 5. Each slice has the same number of observations. For the CSS methods, the function $G(\eta^T X)$ is set by (5.2) as in Section 5.1.1. In other words, $E(X|\eta^T X)$ indeed lies within the space spanned by the quadratic polynomials of $\eta^T X$ and can be estimated by its moment estimators consistently.

where η_1 and η_2 are p-dimensional vectors. Again, we use the squared multiple correlation coefficient ρ^2 to access the estimation accuracy. A larger value of

$\hat{\rho}^2(\hat{\beta}^T X, \beta^T X)$ indicates a better performance. Table 5.4 gives $\hat{\rho}^2(\hat{\beta}^T X, \beta^T X)$ for the combinations of the eight estimators and the three models. The standard errors of $\hat{\rho}^2$ are given in the parentheses, calculated using the 200 simulated samples.

Not to our surprise, CSS-based methods are substantially more accurate with non-elliptical X. More importantly, Table 5.4 shows that there is consistent improvement of CSS-SAVE and CSS-DR over CSS-SIR (even when the model is more favorable for CSS-SIR), indicating the gain by involving the second-order conditional moments. OPG and MAVE do not need elliptical assumptions, and perform consistently better than classical SIR, SAVE and DR. However, their performance is in most cases dominated by the second-order CSS methods (CSS-SAVE and CSS-DR), which is due to the fact that the CSS methods only involve one-dimensional smoothing, whereas OPG and MAVE involve p-dimensional smoothing.

5.2.2 When we don't know exact function forms for $E(X|\beta^T X)$

Now we switch the predictors of Y from X_1, X_2 to X_3, X_4, while keeping the same nonlinearity among the components of X. The central subspace is now spanned by $(0, 0, 1, 0, \ldots)$ and $(0, 0, 0, 1, \ldots)$. In the new arrangement $E(X|\beta^T X)$ no longer lies within the space spanned by the quadratic polynomials of $\beta^T X$; thus the quadratic parameterization of $G(\eta^T X)$ is only approximate. The regression models remain the same except this switch; namely

$$\begin{aligned} &\text{Model 10:} \quad Y = 3\sin(X_3/4) + 3\sin(X_4/4) + \epsilon, \\ &\text{Model 11:} \quad Y = X_3^2 + 3\sin(X_4/4) + 0.5\epsilon, \\ &\text{Model 12:} \quad Y = X_3/[0.5 + (X_4 + 1.5)^2] + 0.1\epsilon. \end{aligned}$$

Model	Method	$p = 4$	$p = 5$	$p = 6$
10	SIR	1.176(0.016)	1.170(0.015)	1.141(0.013)
	CSS-SIR	1.681(0.015)	1.441(0.020)	1.291(0.019)
	SAVE	1.426(0.025)	1.414(0.024)	1.325(0.024)
	CSS-SAVE	1.729(0.005)	1.683(0.011)	1.666(0.011)
	DR	1.581(0.019)	1.575(0.020)	1.498(0.021)
	CSS-DR	1.695(0.010)	1.659(0.012)	1.591(0.013)
	OPG	1.347(0.021)	1.191(0.020)	1.172(0.020)
	MAVE	1.363(0.022)	1.213(0.020)	1.138(0.020)
11	SIR	1.445(0.014)	1.441(0.013)	1.418(0.012)
	CSS-SIR	1.780(0.010)	1.679(0.015)	1.637(0.017)
	SAVE	1.724(0.018)	1.696(0.018)	1.689(0.020)
	CSS-SAVE	1.739(0.004)	1.760(0.008)	1.760(0.008)
	DR	1.754(0.015)	1.736(0.014)	1.721(0.016)
	CSS-DR	1.727(0.006)	1.746(0.007)	1.754(0.008)
	OPG	1.716(0.020)	1.557(0.022)	1.447(0.023)
	MAVE	1.689(0.020)	1.565(0.022)	1.444(0.021)
12	SIR	1.220(0.017)	1.223(0.017)	1.204(0.016)
	CSS-SIR	1.820(0.009)	1.723(0.014)	1.643(0.016)
	SAVE	1.300(0.021)	1.305(0.021)	1.267(0.021)
	CSS-SAVE	1.748(0.004)	1.743(0.008)	1.714(0.010)
	DR	1.325(0.019)	1.315(0.017)	1.320(0.018)
	CSS-DR	1.788(0.008)	1.725(0.011)	1.710(0.011)
	OPG	1.811(0.014)	1.609(0.021)	1.450(0.023)
	MAVE	1.793(0.017)	1.617(0.021)	1.467(0.023)

Table 5.5. Second-order CSS methods with inexact $G(\eta^T X)$.

Table 5.5 shows the similar improvements of CSS methods over classical methods, and improvements of second-order CSS methods over CSS-SIR. As we have already seen in Section 5.1.2, even if the parameterization of $E(X|\beta^T X)$ is not exact, CSS methods work well because they take into consideration of the non-ellipticity.

5.2.3 When sample size increases

Table 5.6 compares the eight methods for Model 8, with $p = 6$ and $n = 200$, 300, 400, 500. The kernel width for OPG and MAVE are taken to be 0.4 for all sample sizes. For larger sample sizes, the performances of all estimators improve. MAVE and the CSS-SIR become more similar. The performance of CSS-SIR catches up with CSS-SAVE as the sample size increases. The performance of CSS-DR inproves the fastest of all, and has a comfortable margin of lead over all the other estimators. As noted in Li and Wang (2007), DR is a combination of SAVE and SIR and typically has better accuracy with elliptical predictors. This advantage seems to have been inherited by CSS-DR.

Method	$n = 200$	$n = 300$	$n = 400$	$n = 500$
SIR	0.890(0.008)	0.902 (0.005)	0.907 (0.005)	0.897 (0.005)
CSS-SIR	1.812(0.014)	1.863 (0.008)	1.892 (0.008)	1.907 (0.007)
SAVE	1.110(0.011)	1.091 (0.008)	1.090 (0.008)	1.079 (0.007)
CSS-SAVE	1.842(0.006)	1.872 (0.005)	1.886 (0.006)	1.905 (0.003)
DR	1.063(0.012)	1.077 (0.009)	1.096 (0.008)	1.101 (0.006)
CSS-DR	1.890(0.008)	1.920 (0.004)	1.944 (0.003)	1.947 (0.006)
OPG	1.625(0.018)	1.771 (0.014)	1.842 (0.012)	1.905 (0.007)
MAVE	1.636(0.017)	1.764 (0.015)	1.848 (0.011)	1.891 (0.010)

Table 5.6. Second-order CSS methods with increasing sample size.

Chapter 6

Application of CSS methods

6.1 Massachusetts college data

We consider data collected for Massachusetts four-year colleges in 1995, which are attempted to study how the percentage of freshmen that graduate (Grad) depends on variables measuring quality of incoming students and features of the colleges. The data is provided as an example data set in MINITAB (release 13.32, data directory STUDNT12). We restricted attention to $n = 46$ colleges and $p = 7$ predictors, which are: the percentage of freshmen that were among the top 25% percent in their graduating high school class (Top25), the median mathematics SAT score (MSAT), the median verbal SAT score (VSAT), the percentage of applicants accepted by the college (Accept), the percentage of accepted applicants who enroll (Enroll), the student-to-faculty ratio (SFRatio), and the out-of-state tuition (Tuition).

The scatter-plot matrix in Figure 6.1 reveals nonlinearity among predictors — for example, in the relations between Top25 and Accept, Accept and Tuition, VSAT and tuition. We apply PIR and CSS-PIR to this data set. Figure 6.2 presents the scatter plots of Y (Grad) versus the first predictors obtained from PIR (left panel)

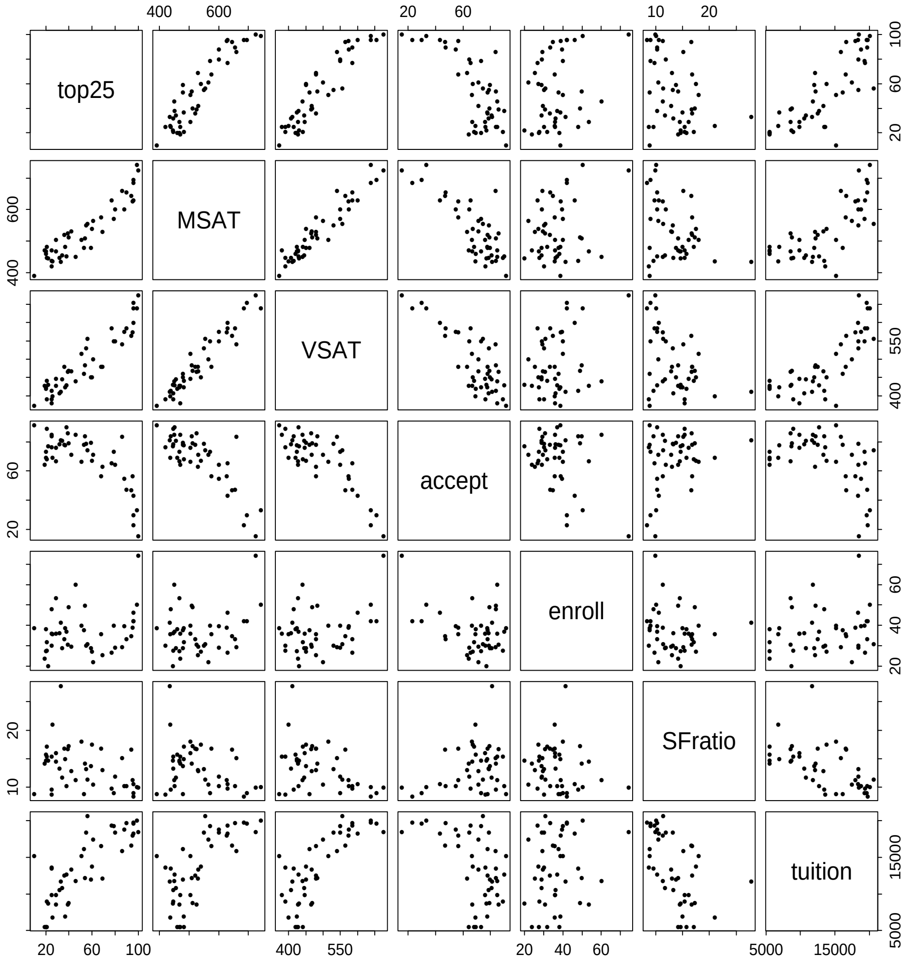
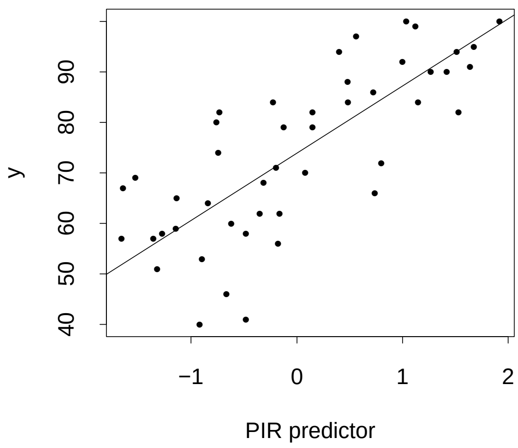


Figure 6.1. Scatter-plot matrix for the seven predictors of the tuition data.

and CSS-PIR (right panel).

Since the true model is unknown, we can no longer use the angle between the estimated and true central spaces to compare the performances of PIR and CSS-PIR, as we did in Section 5.1. We will use instead a leave-one-out cross validation criterion to compare their performances (see, for example, Allen, 1974; Stone, 1974). Let $\tilde{\beta}_{-k}$ and $\hat{\beta}_{-k}$ be the estimated β by PIR and CSS-PIR when (X_k, Y_k) is deleted from the sample. From Figure 6.2 we see that both scatter plots are roughly linear. So for each of the 46 leave-one-out samples, we fit linear models using both the PIR and the CSS-



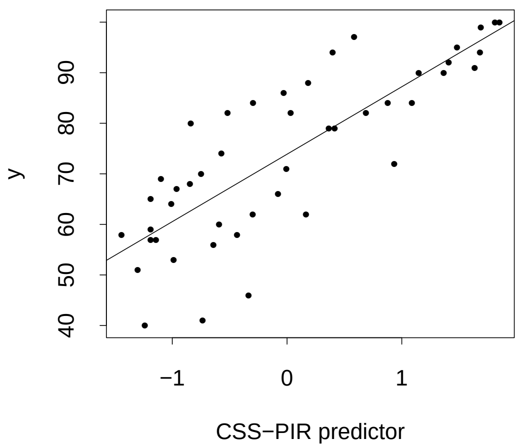


Figure 6.2. Comparison of PIR and CSS-PIR for the tuition data.

PIR predictors, and predict the deleted Y_k by $\tilde{\beta}_{-k}^T X_k$ and $\hat{\beta}_{-k}^T X_k$ using their respective linear models. The sums of squared prediction errors over the 46 samples for PIR and CSS-PIR are, respectively, 6145 and 5203, indicating a respectable improvement by CSS-PIR.

6.2 Handwritten digits data

We now apply the second-order CSS methods to a data set concerning the identification of handwritten digits, which can be found at the UCI machine-learning repository. For earlier studies of this data set, see, for example, M. Zhu and Hastie (2003) and B. Li and Wang (2007). The predictor has 16 components, extracted from the feature of the handwritten digits. The response variable, ranging from 0 to 9, represents the true identification of a digit. The purpose of the data analysis is to classify the digits as 0 through 9 based on the observed predictor. We shall focus on digits 0, 6, and 9 because they are the most difficult to distinguish. This reduced data set consists of 2,219 observations, of which 780 are identified as 0, 720 are identified as 6, and 719 are identified as 9.

Figure 6.3 compares the first three predictors (SAVE1, SAVE2, SAVE3) obtained using SAVE and the first three predictors (CSS-SAVE1, CSS-SAVE2, CSS-SAVE3) obtained using CSS-SAVE. On the top is the perspective plot of SAVE1, SAVE2, SAVE3. The four plots below are the scatter plots for the pairwise CSS-SAVE predictors and the perspective plot of all three CSS-SAVE predictors. The striking feature of this comparison is that while the SAVE predictors are capable of separating the three groups by their variations, they fail to separate them by their locations. In contrast, CSS-SAVE succeeds in separating them by both location and variation.

We have also compared CSS-DR with DR. Since the DR plots for this data set have been presented in B. Li and Wang (2007), and since the CSS-DR plots are similar to the CSS-SAVE plots given in Figure 6.3, the plots for DR and CSS-DR will not be presented here. However, it is worthwhile to mention that, although CSS-SAVE and CSS-DR reach the similar degree of separation as DR, they do seem to gain important additional information. From Figure 3 of B. Li and Wang (2007), we see

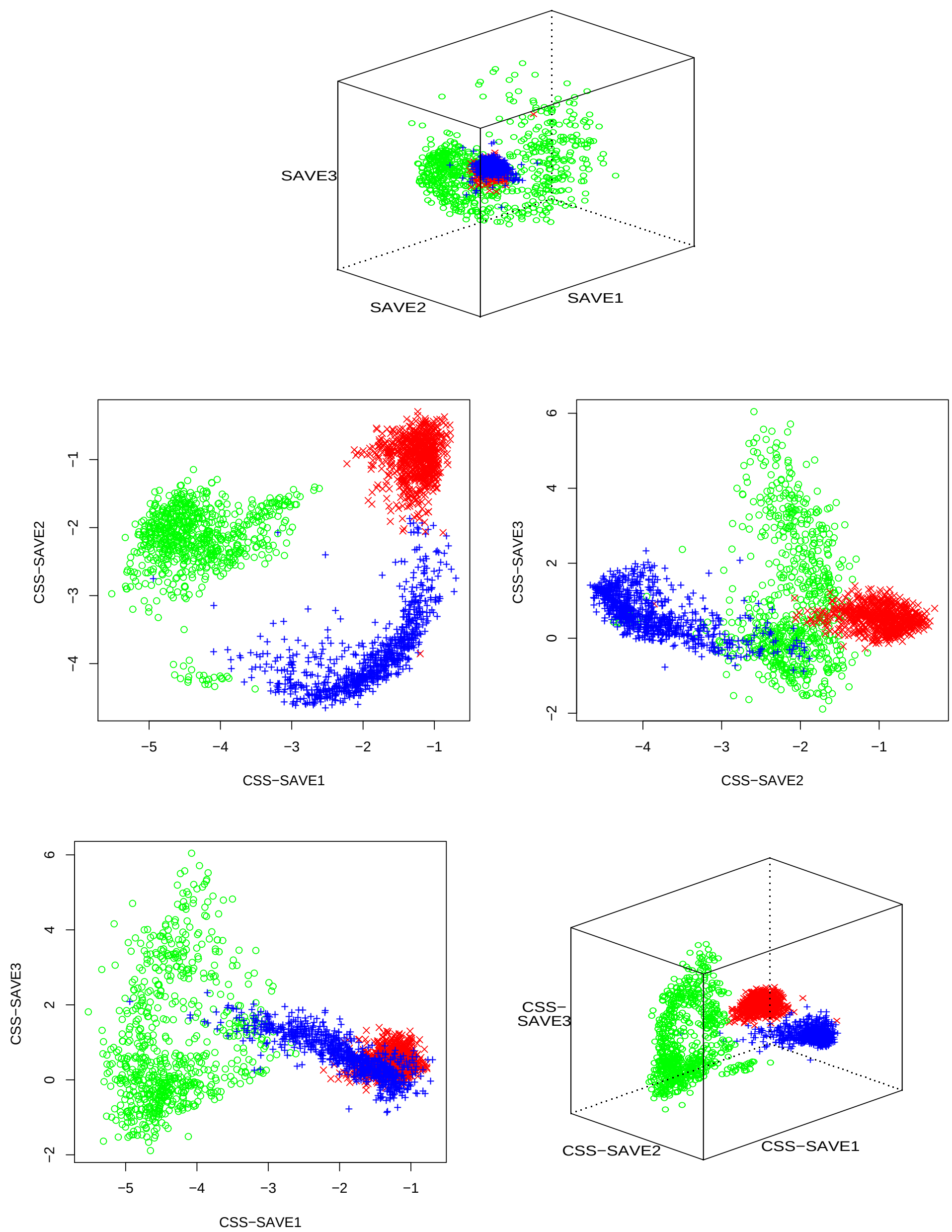


Figure 6.3. Comparison of SAVE and CSS-SAVE for the handwritten digit data set (+,digit 0; ∘, digit 9; ×, digit 6).

that DR succeeded in separating the locations of the three groups, and separating the variations between the 9 group and the other two groups. However, the DR plots did not show substantial difference between the variations of the 0 group and 6 group. In comparison, CSS-SAVE and CSS-DR, in addition to achieving the similar location and variance separations, also capture the difference between the variations of the 0 group (blue) and the 6 group (red). In fact, it is clear that the orientations of the long axes of the 9 group (green) and the 0 group are roughly orthogonal, whereas the 6 group does not show an obvious long axis. The separation of variances by DR of the three groups are not as detailed as or as featured as that achieved by CSS-SAVE and CSS-DR.

Chapter 7

Some extensions and future research directions

Central solution space is an innovative and powerful methodology for dimension reduction. It cures the curse of dimensionality and circumvents the elliptical distribution assumption of classical inverse regression dimension reduction methods at the same time. The work in this thesis has led to two papers B. Li and Dong (2008), Dong and Li (2008). However, the application of central solution space is by no means limited to the methods introduced in this thesis. We will see some extension as well possible future research directions in this chapter.

7.1 Central kth moment solution space

Following the idea of the central mean space, central kth moment dimension reduction space (CKMS; Yin & Cook, 2002) is designed to aim dimension reduction at reducing the mean function, the variance function and up to the kth moment function, leaving the rest of regression Y versus X as the nuisance parameter. We first define

$M^{(k)}(Y|X) = E\{[Y - E(Y|X)]^k|X\}$ for $k \geq 2$ and $M^{(1)}(Y|X) = E(Y|X)$. Then we have the following set of definitions.

Definition 7.1.1. *If*

$$Y \perp\!\!\!\perp \{M^{(1)}(Y|X), \ldots, M^{(k)}(Y|X)\}|\eta^T X,$$

then $\mathcal{S}(\eta) = \mathrm{span}(\eta)$ is a kth moment DRS for the regression of Y versus X.

Definition 7.1.2. *Let $\mathcal{S}^{(k)}_{Y|X} = \cap \mathcal{S}^{(k)}$, where the intersection is over all kth moment DRSs $\mathcal{S}^{(k)}$. If $\mathcal{S}^{(k)}_{Y|X}$ is itself a kth moment DRS, it is called the central kth moment DRS, or CKMS for short.*

If $k = 1$ in the above definitions, they become exactly the same as the definitions of mean dimension reduction space and central mean space. Thus CKMS is a generalization of CMS.

We can also see from the definitions that a DRS is necessarily a kth moment DRS, which must be an ith moment DRS for any $i \leq k$. Just as CMS is always contained in the central space, the CKMS is also contained in the central space, because the former is the intersection of a larger collection of subspaces. These relationships can be summarized as below,

$$\mathcal{S}^{(1)}_{Y|X} \subseteq \cdots \subseteq \mathcal{S}^{(k)}_{Y|X} \subseteq \cdots \subseteq \mathcal{S}_{Y|X}.$$

We can see that if the conditional distribution of Y given X depends only on up to the kth moments of X, then $\mathcal{S}_{Y|X} = \mathcal{S}^{(k)}_{Y|X}$. Furthermore, when the moment-generating function of $Y|X$ exists, we have $\mathcal{S}_{Y|X} = \lim_{k\to\infty}(\mathcal{S}^{(k)}_{Y|X})$.

Let X be a p-dimensional predictor, and Y be a 1-dimensional response. Under normalizing assumption and linear conditional mean asssumption, $E(XY^k) \in \mathcal{S}^{(k)}_{Y|X}$.

Thus $E(ZY^k)$ can be used as an estimate for the central kth moment DRS. In the original CKMS (Yin & Cook, 2002) paper, they propose a population kernel matrix

$$\mathbf{K} = (E(ZY), \ldots, E(ZY^k)),$$

and the subspace spanned by the left singular vectors of $\mathbf{K}$ corresponding to its non-zero singular values is used to span an estimator of $\mathcal{S}^{(k)}_{Y|Z}$. To circumvent the linear conditional mean assumption required here, we can as well use the idea of central solution space.

Definition 7.1.3. *If γ satisfies*

$$E[Zf^k(Y)] = E[E(Z|\gamma^T Z)f^k(Y)] \quad a.s.$$

for any $f^k(Y)$, where $f^k(Y)$ is at most kth degree polynomial of Y, then we call the column space of γ a k-th moment solution space. Furthermore, if the intersection of any two k-th moment solution spaces is itself k-th moment solution space, then the intersection of all such spaces will be called the central k-th moment solution space, and is denoted by $\mathcal{S}^{(k)}_{\text{CKMSS}}$.

More study about this topic can be found in Dong (2008).

7.2 Future work

Instead of $Y \perp\!\!\!\perp X|\alpha^T X$, Cook (2007) formulated a more general sufficient dimension reduction problem as

$$Y \perp\!\!\!\perp X|T(X),$$

where $T(X)$ is a nonlinear function. Wang (2008) introduced *nonlinear feature* as a mapping $\Phi : \mathbb{R}^p \to \mathbb{R}^r$, where $r \geq p$, that augments the original X with additional nonlinear predictors. We thus consider the augmented dimension reduction problem

$$Y \perp\!\!\!\perp X | \beta^T \Phi(X).$$

The nonlinear feature $\Phi(\cdot)$ is assumed to be a known function. For example, if we would like to include interaction in the predictor, we choose Φ such that it consists of the following elements

$$\Phi(X) = [\{X_i : \; i = 1, \ldots, p\} \cup \{X_j \, X_k : \; 1 \leq j < k \leq p\}]^T.$$

We call this new problem nonlinear sufficient dimension reduction. Traditional dimension reduction methods, such as SIR and SAVE, will no longer work for nonlinear dimension reduction, because they make strong assumption about the elliptical distribution of predictor X. Expansion to $\Phi(X)$ from X will easily violate this assumption. Central solution space can circumvent the distribution assumption and should be able to serve the purpose of nonlinear dimension reduction. Extension towards this direction is currently under investigation.

References

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, *16*, 125-127.

Bellman, R. (1961). *Adaptive control processes: A guided tour* (First ed.). New Jersey: Princeton University Press.

Bura, E., & Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, *63*, 393-410.

Chen, C. H., & Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, *8*, 289-316.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, *89*, 177-189.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, *91*, 983-992.

Cook, R. D. (1998a). Principal hessian directions revisited. *Journal of the American Statistical Association*, *93*, 84–100.

Cook, R. D. (1998b). *Regression graphics: Ideas for studying regressions through graphics* (First ed.). New York: Wiley.

Cook, R. D. (2007). Fisher lecture: Dimension reduction for regression (with discussion). *Statistical Science.*, *12*, 1-26.

Cook, R. D., & Li, B. (2002). Dimension reduction for the conditional mean. *The Annals of Statistics*, *30*, 455-474.

Cook, R. D., & Nachtsheim, C. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, *89*, 592-599.

Cook, R. D., & Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, *100*, 410-428.

Cook, R. D., & Weisberg, S. (1991). Discussion of "sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, *86*, 316-342.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*, 1-31.

Diaconis, P., & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *The Annals of Statistics*, *12*, 793-815.

Dong, Y. (2008). Dimension reduction for the conditional kth moment via central solution space. *Master thesis.*

Dong, Y., & Li, B. (2008). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Submitted to Biometrika.*

Eaton, M. L. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis*, *41*, 1-31.

Fernholz, L. T. (1983). *Von mises calculus for statistical functionals* (First ed.). New York: Springer.

Ferre, L., & Yao, A. F. (2005). Smooth function inverse regression. *Statistica Sinica*, *15*, 665-683.

Fung, K. F., He, X., Liu, L., & Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*, *12*, 1093-1113.

Hall, P., & Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, *21*, 867-889.

Hardle, W., & Stoker, T. (1989). Investigating smooth multiple regression by method of average derivatives. *Journal of the American Statistical Association*, *84*, 986-

995.

Li, B., & Dong, Y. (2008). Dimension reduction for non-elliptically distributed predictors. *The Annals of Statistics. To appear.*

Li, B., & Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, *102*, 997-1008.

Li, B., Zha, H., & Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, *33*, 1580-1616.

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

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.

Li, K. C., & Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, *17*, 1009-1052.

Powell, J., Stock, J., & Stoker, T. (1989). Semiparametric estimation of index coefficients. *Econometrica*, *57*, 1403-1430.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*, 111-147.

Wang, Y. (2008). Nonlinear dimension reduction in feature space. *Ph.D. thesis.*

Xia, Y., Tong, H., Li, W. K., & Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace. *Journal of the Royal Statistical Society. Series B (Methodological)*, *64*, 363-410.

Ye, Z., & Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, *98*, 968–979.

Yin, X., & Cook, R. D. (2002). Dimension reduction for the conditional k-th moment in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, *64*, 159-175.

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

Zhu, L. X., & Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, *3*, 1053-1068.

Zhu, M., & Hastie, T. J. (2003). Feature extraction for nonparametric discriminatinant analysis. *Journal of Computational and Graphical Statistics*, *12*, 101–120.

Vita

Yuexiao Dong

Yuexiao Dong was born in September 1983 in Hubei, China. His education started at Danjiangkou Experiment School in 1988. He received a high school diploma from Yuyang High School in 2000. He earned Bachelor's degree of Science in July 2004 from the Mathematics Department at Tsinghua University. Started from August 2004, he studied in the United States to pursue his Ph.D. in statistics at the Pennsylvania State University. He will work as an Assistant Professor for Temple University in July, 2009.