NONLINEAR DIMENSION REDUCTION IN FEATURE SPACE

A Dissertation in
Statistics
by
Yu Wang

© 2008 Yu Wang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2008
The dissertation of Yu Wang was reviewed and approved* by the following:

Bing Li  
Professor of Statistics  
Dissertation Advisor  
**Chair of Committee**

Francesca Chiaromonte  
Associate Professor of Statistics and Health Evaluation Sciences

Bruce G. Lindsay  
Distinguished Professor of Statistics  
Head of Department of Statistics

Hongyuan Zha  
Professor of Computer Science

*Signatures are on file in the Graduate School.
Abstract

In this thesis I introduce an idea for applying dimension reduction methods to feature spaces. Three main methods will be used to estimate dimension reduction directions on feature spaces.

Feature spaces are manifolds embedded in extended Euclidean spaces and do not satisfy the usually required elliptical distribution assumption. However, by applying an asymptotic normality result, we proved that when the original predictors have multivariate normal distribution, we can perform dimension reduction directly on a second degree polynomial feature space.

By linearly approximating the feature manifold locally with tangent spaces, and by a result on local transformation relation between approximate dimension reduction directions on the tangent space and the original space, we can estimate the directions on the feature space by aggregating information from transformed local estimators on the original space.

Finally, the directions in the feature space can be estimated using a method of the Central Solution Space recently introduced by Li and Dong (2008).
# Table of Contents

List of Figures vii

List of Tables ix

Acknowledgments x

Chapter 1

## Preliminaries

1.1 Introduction ................................................................. 1
  1.1.1 The Dimension Reduction Model ............................... 3
  1.1.2 Dimension Reduction Space, Minimum Dimension Reduction Space, and Central Space ......................... 4
  1.1.3 Linear Conditional Mean (LCM) ................................. 6
1.2 Dimension Reduction Methods .......................................... 6
  1.2.1 Ordinary Least Squares ........................................... 7
  1.2.2 Sliced Inverse Regression ....................................... 8
  1.2.3 Sliced Average Variance Estimation ........................... 10
  1.2.4 Principal Hessian Directions .................................. 12
  1.2.5 Iterative Hessian Transformation .............................. 14
  1.2.6 Minimum Average Variance Estimation ....................... 17
  1.2.7 Contour Regression ............................................... 19
  1.2.8 Directional Regression ........................................... 22
1.3 Several Issues in Dimension Reduction .............................. 25
  1.3.1 Global Methods and Local Methods ........................... 25
  1.3.2 Central Subpace and Central Mean Subspace .................. 26
Chapter 2

**Global Dimension Reduction on Feature Space (G-DRFS)**

- **2.1 Idea** ................................................................. 29
  - 2.1.1 Introducing the idea of feature .......................... 29
  - 2.1.2 Non-linear Predictor Space ............................... 34
- **2.2 Approximate Normality of Projected Vector** ....... 35
  - 2.2.1 Condition 1 ....................................................... 38
  - 2.2.2 Condition 2 ....................................................... 42
- **2.3 Global Estimation on the Feature Space** ............ 44
- **2.4 Simulations** ..................................................... 45
  - 2.4.1 Checking Asymptotic Normality of Projected Variables . 45
  - 2.4.2 Checking Linear Conditional Mean Condition .......... 47
  - 2.4.3 Global Method on the Feature Space ................. 48
- **2.5 Application with Baltimore Lead Data** ............... 50

Chapter 3

**Local Dimension Reduction Method on Feature Space (L-DRFS)**

- **3.1 Local Methods vs. Global Methods** ................. 67
- **3.2 Direct Local Dimension Reduction Methods** ......... 67
  - 3.2.1 Nonparametric methods .................................... 67
  - 3.2.2 Dimension Reduction Methods .......................... 69
  - 3.2.3 Invariant Aggregation Dimension Reduction on Feature Space (IA-DRFS) ...................................................... 70
  - 3.2.4 Simulations ....................................................... 70
  - 3.2.5 Application with Baltimore Lead Data ............... 71
- **3.3 Nonlinear Aggregation Dimension Reduction on Feature Space (NA-DRFS)** ...................................................... 72
  - 3.3.1 Local Linear Approximation of the Manifold .......... 72
  - 3.3.2 Nonlinear Aggregation Dimension Reduction on Feature Space (NA-DRFS) ...................................................... 77
  - 3.3.3 Estimation Procedure ........................................ 78
  - 3.3.4 Simulations ....................................................... 79
  - 3.3.5 Application with the Baltimore Lead Data .......... 80
- **3.4 Corrected Local Estimator for SIR** .................... 80
  - 3.4.1 Corrected SIR Estimation Method with Approximate Linear Local Density ...................................................... 81
  - 3.4.2 Simulations ....................................................... 84
Chapter 4
Nonlinear Feature Selection by the Central Solution Space Method
(CSS-DRFS) 92
4.1 The idea ................................................................. 92
4.2 Simulation ............................................................. 95
4.3 Application with the Baltimore Lead Data ......................... 96
4.4 Comparison Among The Feature Space Methods ................. 96

Appendix A
R Code for Generating Features 99

Appendix B
R Code for Nonlinear Aggregation Method 102

References 105
List of Figures

2.1 A separating hyperplane for a two dimensional data set. .................. 29
2.2 Data cannot be separated by a linear function. ....................... 30
2.3 A feature map can simplify the classification task. ............... 30
2.4 3D scatter plots of $Y$ versus attributes $X_1$ and $X_2$ in Example 2.1.1. 33
2.5 3D scatter plots of $Y$ versus features $X_1^2$ and $X_2^2$ in Example 2.1.1. 33
2.6 Histograms for projected random vectors. .......................... 52
2.7 Scatter plots of first dimension of each random vector v.s. projected random vectors. ........................................ 53
2.8 Scatter plots for $\phi_i(X)$ v.s. $\alpha^T \phi(X)$, $i = 1, 2, 3, 4$, $p = 9$, $n = 100$. 54
2.9 Scatter plots for some randomly drawn $\phi_i(X)$ v.s. $\alpha^T \phi(X)$. $p = 90$, $n = 100$. .................................................. 55
2.10 Scatter plots for some randomly drawn $\phi_i(X)$ v.s. $\alpha^T \phi(X)$. $p = 90$, $n = 1000$. .................................................. 56
2.11 Scatter plots of $Y$ versus the six estimated predictors with SIR on $X$. ................................................................. 57
2.12 Scatter plots of $Y$ versus the first six estimated predictors with SIR on $\phi(X)$. ............................................................. 58
2.13 $Y$ versus known predictor and $Y$ versus estimated predictor. .... 59
2.14 Scatter plots of $Y$ versus the six estimated predictors with SIR on $X$ with full circle of the sin function. .............................. 60
2.15 Scatter plots of $Y$ versus the first six estimated predictors with SIR on $\phi(X)$ with full circle of the sin function. ..................... 61
2.16 $Y$ versus known predictor and $Y$ versus estimated predictor with full circle of the sin function. .............................. 62
2.17 2-D Scatter plots of $X$ v.s. $Y$ in Baltimore Lead Data. ........... 62
2.18 3-D Scatter plots of Baltimore Lead Data. ............................ 63
2.19 Scatter plots matrix of Feature Vector on Baltimore Lead Data. .. 64
2.20 Lead level versus the first estimated predictor with global SIR on feature space. .................................................. 65
2.21 Lead level versus the second to the fifth estimated predictor with global SIR on feature space. ............................. 65

3.1 Scatter plots of $Y$ versus the first four estimated predictors with IA-DRFS. ......................................................... 86
3.2 Comparison of the estimators from three methods .................. 87
3.3 Scatter plot of lead level versus the first estimated predictors with IA-DRFS. .......................................................... 88
3.4 Scatter plots of lead level versus the last four estimated predictors with IA-DRFS. ....................................................... 88
3.5 Scatter plots of $Y$ versus the first four estimated predictors of NA-DRFS method. ................................................... 89
3.6 Comparison of the estimators from three methods ................. 90
3.7 Scatter plot of lead level versus the first estimated predictors from the NA-DRFS method. ........................................... 91
3.8 Scatter plots of lead level versus the second to the fifth estimated predictors from the NA-DRFS method. ....................... 91

4.1 Scatter plots of $Y$ versus estimated predictor from CSS-DRFS overlayed with the known predictor. ......................... 95
4.2 Scatter plots of lead level versus estimated predictor by CSS-DRFS method. .......................................................... 96
4.3 Lead level versus the first estimated predictor with G-DRFS method. .......................................................... 97
4.4 Lead level versus the first estimated predictor with IA-DRFS method. .......................................................... 97
4.5 Lead level versus the first estimated predictor with NA-DRFS method. .......................................................... 98
List of Tables

2.1 Comparison of estimated direction vector with known direction vector. 49
I gratefully thank my advisor, Dr. Bing Li. It is Dr. Li who introduced me to this great research topic of Dimension Reduction in Regression, with his knowledge, his insight, his encouragement and his patience, he taught me both the way to learn knowledge and the way to do research. He showed me the beauty of statistics. I appreciate his guidance.

I want to thank my committee members, Dr. Francesca Chiaromonte, Dr. Bruce Lindsay, and Dr. Hongyuan Zha for their inspirational questions, advice and help. Each of them helps me in her or his unique way.

Finally, I thank my mother and father for the way they raised me; I thank my brother for his always there support starting from our earliest memories; I thank my husband for his accompany ever since we first met nineteen years ago; I thank my son for his love and being so lovely.
Chapter 1

Preliminaries

1.1 Introduction

Let $X$ be a $p$-dimensional explanatory predictor, $Y$ be a univariate response variable. Regression analysis is one of the most important ways of studying the dependency of $Y$ on $X$. When there is a parsimonious parametric model, techniques such as maximum likelihood or the least squares methods can be used to estimate the model.

However, in most applications, we do not know the parametric model. Nonparametric regression techniques can be applied in these situations with the advantage of flexible modeling. Assuming only the continuity of the regression function, nonparametric methods explore the regression relationship by local smoothing. Without any descriptive structure for the conditional distribution of $Y|X$, it is possible that each distinct $p$-dimensional vector value of $X$ will index a different and arbitrary conditional distribution of $Y|X$. The performance of local smoothing relies on the presence of sufficiently many data points around each $X$ vector. There are many good smoothing techniques when the predictor is one-dimensional.
However, when the dimension of $X$ increases, standard smoothing methods such as kernel estimates or nearest neighbor estimates break down quickly because of the sparseness of the data points in each local region. In other words, we face the "curse of dimensionality," a phrase coined by Bellman (1961) to describe the problem caused by the rapid decrease of number of observations in each local region when the dimension is increased.

To efficiently estimate the relationship between $Y$ and $X$ while avoiding "the curse of dimensionality," several high dimensional analysis techniques have been proposed. Among them, there are essentially two approaches as summarized in Xia, Tong, Li, and Zhu (2002): the first is largely concerned with function approximation and the second with dimension reduction. Examples of the former are the additive model approach of Hastie and Tibshirani (1986) and the projection pursuit regression proposed by Friedman and Stuetzle (1981); both assume that the regression function is a sum of univariate smooth functions. Examples of the latter are dimension reduction and regression graphics (K.-C. Li, 1991; Cook and Weisberg, 1991). In the first approach, function approximation is of main concern; in dimension reduction, effective dimension reduction directions could be estimated without estimating the response surface and without strong assumptions on the form of relation between $Y$ and $X$. Dimension reduction is an effective premodeling tool, a means of avoiding the curse of dimensionality, and the foundation for regression graphics.

My main effort will be focused on the dimension reduction approach. In the rest of this section, I will introduce the basic dimension reduction model and several important concepts in dimension reduction. In the next section, I will give an overview of the main methods in dimension reduction.
1.1.1 The Dimension Reduction Model

K.-C. Li (1991) proposed a technique for reducing the dimension of the predictor space before applying regression methods to analyze data. In that paper, Li proposed a model assuming an ideal situation that high-dimensional predictor $X$ are retrievable from low-dimensional projections for the purpose of regressing $Y$ on $X$:

$$Y = f(\beta_1^T X, \beta_2^T X, \cdots, \beta_k^T X, \epsilon).$$ (1.1)

Here, $Y$ is the one-dimensional response variable and $X$ is a $p$-dimensional predictor. The $\beta_i$’s are unknown column vectors. We use $\beta \in \mathbb{R}^{p \times k}$, $k \leq p$ to denote the $p \times k$ matrix with columns $\beta_i$, $i = 1, \cdots, k$. In K.-C. Li (1991), $\epsilon$ is independent of $X$; in later methods, $\epsilon$ is allowed to be dependent on $X$, with the requirement that $E(\epsilon | X) = 0$ almost surely. Here $f$ is an arbitrary unknown function from $\mathbb{R}^{k+1}$ to $\mathbb{R}$.

When this model holds, the projection of the $p$-dimensional explanatory variable $X$ onto the $k$ dimensional subspace, $(\beta_1^T X, \cdots, \beta_k^T X)^T$, captures all the regression information. The potential advantages of this model are that no prespecified model for $Y|X$ is required and the curse of dimensionality may be avoided because $k$ is usually very small.

A more general representation of the assumption that $Y$ depends on $X$ only through $\beta^T X$ is the conditional independence

$$Y \indep X | \beta^T X,$$

where $\indep$ indicates independence. (1.2)

The conditional independent model (1.2) is more general than the functional model (1.1) in that it does not separate the role of $f$ from $\epsilon$ in (1.1); see for example
the discussions in B. Li and Yin (2007). From regression graphics point of view, we can replace the full plot \{Y, X\} by the lower dimensional plot \{Y, \beta^T X\}. That is, (1.2) implies that the \((k + 1)\)-dimensional plot \{Y, \beta^T X\} is sufficient for the regression of \(Y\) on \(X\). Model (1.2) will be used throughout this paper. We now introduce several important concepts in dimension reduction.

1.1.2 Dimension Reduction Space, Minimum Dimension Reduction Space, and Central Space

Definition 1.1.1. When model (1.2) holds, the subspace \(S(\beta)\) is called a dimension-reduction subspace (DRS) for \(Y|X\), or equivalently, for the regression of \(Y\) on \(X\).

The short-hand phrase “dimension-reduction subspace” will be used when the response and predictors are clear from context; it will be denoted as \(S_{drs}\).

Since model (1.2) holds trivially for \(\beta = I_p\), the \(p \times p\) identity matrix, dimension reduction subspaces need not result in reduced dimensionality. With the same reasoning, a DRS is not necessarily unique. While the intent behind model (1.2) is to provide a framework for reducing the dimensionality of the predictors, the idea of a smallest subspace will be helpful; it is only through the notion of the smallest dimension reduction subspace that this intent can be clearly pronounced.

Definition 1.1.2. A subspace \(S\) is said to be a minimum DRS for \(Y|X\) if \(S\) is a DRS and \(\text{dim}[S] \leq \text{dim}[S_{drs}]\) for all \(S_{drs}\).

A minimum DRS is not necessarily unique. However, if a regression admits several minimum DRSs, their dimensions must be the same by definition. Such unique dimension is called the structural dimension of the regression and denoted by \(k\).
To avoid the possible nonuniqueness, Cook (1994) introduced the idea of central DRS.

**Definition 1.1.3.** A subspace $S$ is the central dimension reduction subspace for the regression of $Y$ on $X$ if $S$ is a DRS and is the intersection of all DRSs $S_{drs}$. The central dimension reduction subspace will be denoted by $S_{Y|X}$.

A central dimension reduction subspace will often be referred to simply as a central subspace. By definition, a central subspace exists if and only if the intersection $\cap S_{drs}$ of all DRSs is itself a DRS. The intersection $\cap S_{drs}$ is always a subspace but it is not necessarily a DRS. Although a central subspace might not exist, it is unique when it does. The existence of a central subspace can depend on the conditional distribution of $Y|X$ or on the marginal distribution of $X$. Cook (1998), Chiaromonte and Cook (2002) and recently Yin, Li, and Cook (2008) provided very mild conditions under which the central subspace exists.

The following two propositions from Cook (1998) give two properties of central subspaces: uniqueness and invariance.

**Proposition 1.1.1.** If $S_{Y|X}$ is the central subspace for the regression of $Y$ on $X$ then $S_{Y|X}$ is the unique minimum DRS.

**Proposition 1.1.2.** Let $S_{Y|X}$ be the central subspace for the regression of $Y$ on $X$, and let $S_{Y|Z}$ be the central subspace for the regression of $Y$ on $Z = A^T X + b$ where $A$ is a full rank, $p \times p$ matrix. Then $S_{Y|Z} = A^{-1} S_{Y|X}$.

Proposition 1.1.2 suggests that it is equivalent to consider any affine transformation of $X$. Henceforth, we will assume that $E(X) = 0$, and var($X$) = $I_p$.

The main goal of dimension reduction is to estimate the central subspace or at least a portion of it.
1.1.3 Linear Conditional Mean (LCM)

Over the past two decades or so, both graphical and numerical methods have been developed for estimating central subspaces, each with its own requirements, limitations, and advantages. Sometimes, new methods stem from old methods; sometimes, they are proposed to improve weakness of the previous methods.

The following assumption, called the linear conditional mean assumption, is a key assumption for many dimension reduction methods.

Assumption 1.1.1. For any \( b \) in \( \mathbb{R}^p \), the conditional expectation

\[
E(bX | \beta_1^T X, \cdots, \beta_k^T X)
\]

is linear in \( \beta_1^T X, \cdots, \beta_k^T X \); that is, for some constants \( c_0, c_1, \cdots, c_k \),

\[
E(bX | \beta_1^T X, \cdots, \beta_k^T X) = c_0 + c_1 \beta_1 X + \cdots + c_k \beta_k X. \tag{1.3}
\]

Although the assumption is required to hold only for the basis matrix \( \beta \), since \( \beta \) is unknown, in practice, we require that it hold for all possible \( \beta \). This is satisfied if and only if the distribution of \( X \) is elliptically symmetric (Eaton, 1986).

1.2 Dimension Reduction Methods

Aside from the semiparametric method proposed by Xia et al. (2002), the dimension reduction methods could be divided again into two categories: classical methods and convex combination of the first two conditional moments. I will try to follow chronically the order in which they are proposed. For each method, I will state the requirements for the method, the population development for the
method, the estimating procedure at the sample level, the asymptotic results of
the estimator, and finally the advantages and disadvantages of the method.

1.2.1 Ordinary Least Squares

The theoretical background for Ordinary Least Squares (OLS) is based on K.-C.
Li and Duan (1989). Let $X$ and $Y$ be defined as before. The OLS vector is the
minimizer of the loss function

$$L(\alpha, \beta) = E(Y - \alpha - \beta^T X)^2.$$ 

It is easy to see that

$$\beta = [E(X - EX)(X - EX)^T]^{-1} E(X - EX)(Y - EY) = [\text{var}(X)]^{-1} \text{cov}(X, Y).$$

When $E(X) = 0$ and $\text{var}(X) = I_p$, this reduces to

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

The central result of Li and Duan (1989) is the following theorem.

**Theorem 1.2.1.** Suppose that $E(X) = 0$ and $\text{var}(X) = I_p$, and that Assumption
1.1.1 holds. Then

$$E(XY) \in S_{Y|X}.$$

The sample estimating procedure can be described as follows.

1. Standardization of $X_i$: $Z_i = \hat{\Sigma}^{-1/2}(X_i - \hat{\mu}), i = 1, \ldots, n$, where $\hat{\mu}$ =
\[ n^{-1} \sum_{i=1}^{n} X_i, \hat{\Sigma} = n^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T. \]

2. Center \( Y_i \): \( \hat{Y}_i = Y_i - n^{-1} \sum_{i=1}^{n} Y_i. \)

3. Calculate \( \hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} Z_i \hat{Y}_i \): This is an estimator of \( E(ZY) \), a vector in \( S_{Y|Z} \).

4. Estimation of the dimension-reduction direction \( \beta \): Let \( \hat{\beta} = \hat{\Sigma}^{-1/2} \hat{\gamma}. \)

The OLS estimator has a parametric convergence rate, regardless of the relation between \( Y \) and \( X \).

**Theorem 1.2.2.** The OLS estimator for \( S_{Y|X} \) is \( \sqrt{n} \)-consistent. In other words, it converges at \( \sqrt{n} \)-rate to a vector that belongs to \( S_{Y|X} \).

The disadvantages of the OLS method are that it can only estimate at most one direction in the central subspace, and it cannot find a U-shaped trend. The first limitation implies that OLS cannot provide a comprehensive estimate in case the dimension of the central subspace is more than one.

### 1.2.2 Sliced Inverse Regression

Sliced Inverse Regression (SIR) was proposed by K.-C. Li (1991). It is one of the most widely used dimension reduction methods. The main idea is that under the LCM assumption (Assumption 1.1.1), the centered inverse regression curve \( E(X|Y) - E(X) \) falls almost surely within the subspace spanned by \( \Sigma \beta_1, \ldots, \Sigma \beta_k \), where \( \Sigma = \text{var}(X) \). So, for the standardized version of \( X \), \( \text{Span(\text{var}(E(Z|Y)))} \subseteq S_{Y|Z} \). Here, for a matrix \( A \), we use \( \text{Span}(A) \) to denote the space spanned by the columns of \( A \). When \( k \) is known, a spectral decomposition will give a consistent estimate of a subspace of \( S_{Y|Z} \).

The following theorems are the key theoretical results for Sliced Inverse Regression.
Theorem 1.2.3. Suppose that Assumption 1.1.1 holds, and that $X$ is standardized, so that $E(X) = 0$ and $\text{var}(X) = I_p$. Then, $E(X|Y) \in S_{Y|X}$ almost surely. And the column space of the matrix $\text{var}(E(X|Y))$ is a subspace of $S_{Y|X}$.

In practice, we have to use the discretized version of $Y$ unless $Y$ itself is a discrete random variable. Let $I_1, \ldots, I_h$ be the $h$ intervals that partition $\Psi$, the sample space of $Y$. Let $\tilde{Y}$ be the discretized $Y$, defined by

$$\tilde{Y} = i, \text{ if } Y \in I_i, \ i = 1, \ldots, h.$$  

We have the following theorem.

Theorem 1.2.4. Suppose that Assumption (1.1.1) holds, and that $X$ is standardized. Then, for any $i = 1, \cdots, h$, $E(X|\tilde{Y} = i) \in S_{Y|X}$. Consequently, the column space of the matrix $\text{var}(E(X|\tilde{Y}))$ is a subspace of $S_{Y|X}$.

The sample estimating procedure can be described as follows.

1. Standardization: Standardize $X_i$ to $Z_i$ as in OLS.

2. Slicing: Distribute $Z_1, \cdots, Z_n$ evenly into $H$ slices $I_h$, $h = 1, \cdots, H$, according to the order of the respective values of $Y_1, \cdots, Y_n$, each slice has $n_h = n/H$ observations.

3. Calculating slice means: $\hat{m}_h = \sum_{I_h} Z_i / n_h$.

4. Constructing SIR: $\widehat{\text{SIR}} = \sum_{h=1}^H \hat{m}_h \hat{m}_h^T / H$.

5. Performing principal component analysis on $\widehat{\text{SIR}}$: Assume $k$ is known, let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_k$ be the largest $k$ eigenvalues, and $\hat{\eta}_1, \cdots, \hat{\eta}_k$ be the corresponding normalized eigenvectors.
6. Estimating the dimension reduction directions $\beta_j$: 
\[ \hat{\beta}_j = \hat{\Sigma}^{-1/2} \hat{\eta}_j, \quad j = 1, \ldots, k. \]

Like OLS, the SIR estimator converges at $\sqrt{n}$-rate to a set of vectors in $S_{Y|X}$.

In the estimating procedure, we suppose that $k$ is known. K.-C. Li (1991) developed a Chi-square sequential test to determine $k$ when $X$ is normally distributed.

**Theorem 1.2.5.** If $X$ has a $p$-dimensional multivariate normal distribution, and $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ are the eigenvalues of $\hat{S}_{\text{IR}}$, then under the null hypothesis $H_0: \lambda_{p-j+1} = \cdots = \lambda_p = 0$,
\[
\sum_{i=p-j+1}^p \hat{\lambda}_i
\]
converges in distribution to a $\chi^2$ distribution with $(p - j)(p - j + 1)/2$ degrees of freedom.

**1.2.3 Sliced Average Variance Estimation**

Sliced Average Variance Estimation (SAVE) was proposed by Cook and Weisberg (1991). Further theoretical results are given in Cook and Lee (1999). Instead of calculating the mean within each slice as in SIR, we compute the within-slice variance for SAVE.

Besides the LCM assumption (1.1.1), we also need the following Constant Conditional Variance (CCV) condition.

**Assumption 1.2.1.** The conditional variance $\text{var}(X|\beta^T X)$ is a nonrandom matrix.

The theoretical basis for SAVE is the following theorem.
Theorem 1.2.6. Suppose that LCM assumption 1.1.1 and CCV assumption 1.2.1 hold and that \( E(X) = 0 \) and \( \text{var}(X) = I_p \). Then the columns of the random matrix

\[
I_p - \text{var}(X|Y)
\]

is a subspace of the central space \( S_{Y|X} \) almost surely. Consequently, the column space of the matrix

\[
E[I_p - \text{var}(X|Y)]^2
\]  

(1.4)

is a subspace of \( S_{Y|X} \).

Similar to SIR, in case that \( Y \) is continuous we replace it with a discrete sliced response \( \tilde{Y} \). The conditional covariance matrices are then based on the sliced response leading to a discretized version of the theorem above:

Theorem 1.2.7. Suppose that LCM assumption 1.1.1 and CCV assumption 1.2.1 hold and that \( E(X) = 0 \) and \( \text{var}(X) = I_p \). Then, for any \( i = 1, \ldots, h \), the column space of the matrix

\[
I_p - \text{var}(X|\tilde{Y} = i)
\]

is a subspace of the central space \( S_{Y|X} \). Consequently, the column space of the matrix

\[
E[I_p - \text{var}(X|\tilde{Y})]^2
\]

is a subspace of \( S_{Y|X} \).
The sample estimating procedure can be described as follows.

1. Standardization of $X_i$ to $Z_i$ and slicing of $Y$ will be the same as in SIR.

2. Calculating slice covariance matrices: $\hat{V}_h = \text{var}(Z|\tilde{Y}_h)$.

3. Constructing $\widehat{SAVE}$: $\widehat{SAVE} = \frac{1}{H} \sum_{h=1}^{H} (I - \hat{V}_h)^2$.

4. Performing principal component analysis on $\widehat{SAVE}$: Assume $k$ is known, let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_k$ be the largest $k$ eigenvalues, and $\hat{\eta}_1, \cdots, \hat{\eta}_k$ be the corresponding normalized eigenvectors.

5. Estimating the dimension reduction directions $\beta_j$: $\hat{\beta}_j = \hat{\Sigma}^{-1/2} \hat{\eta}_j, \ j = 1, \cdots, k$.

### 1.2.4 Principal Hessian Directions

Principal Hessian Directions (pHd) was proposed by K.-C. Li (1992), and extended by Cook (1998). The main idea is: Let $H(X)$ be the $p \times p$ Hessian matrix of the forward regression function,

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

where $\beta$ is a basis for the central subspace $S_{Y|X}$. By the above representation, $H(X)$ is degenerate along directions that are orthogonal to $S_{Y|X}$. Also, by assuming normally distributed predictors, we can apply Stein’s lemma on the expectation of the derivative of a function of a normal random variable. In K.-C. Li (1992), he based the procedure on an estimate of $E[H(X)]$.

There are two forms of Hessian matrix. Let $\alpha$ be the OLS vector $E(XY)$. Let $e = Y - \alpha^T X$. 
Definition 1.2.1. The matrix $H_1 = E(YXX^T)$ is called the $y$-based Hessian matrix, the matrix $H_2 = E(eXX^T)$ is called the $e$-based Hessian matrix.

The following theorem is proved in K.-C. Li (1992) and Cook (1998).

Theorem 1.2.8. Suppose that $X$ is multivariate normal with mean 0 and covariance matrix $I_p$. Suppose that Assumption 1.1.1 and 1.2.1 hold. Then the column space of $y$-based Hessian matrix $H_1 = E(YXX^T)$ is a subspace of $S_Y|X$; and the column space of $e$-based Hessian matrix $H_2 = E(eXX^T)$ is also a subspace of $S_Y|X$.

The sample estimating procedure can be described as follows.

1. Standardize $X_i$ to $Z_i$ and center $Y_i$ to $\hat{Y}_i$ as in the algorithm for OLS.

2. Compute the OLS vector $\hat{\alpha}$ of centered $\hat{Y}$ versus $Z$, and obtain $e_i = \hat{Y}_i - \hat{\alpha}^T Z_i$.

3. Construct the $y$-based and $e$-based Hessian matrix $\hat{H}_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i Z_i Z_i^T$, and $\hat{H}_2 = \frac{1}{n} \sum_{i=1}^{n} e_i Z_i Z_i^T$.

4. Assume the structural dimension $k$ is known. Let $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$ be the $k$ eigenvectors corresponding to the $k$ largest eigenvalues of $\hat{H}_1 \hat{H}_1^T$; and let $\hat{\delta}_1, \ldots, \hat{\delta}_k$ be the $k$ eigenvectors corresponding to the $k$ largest eigenvalues of $\hat{H}_2 \hat{H}_2^T$.

We use $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$ or $\hat{\delta}_1, \ldots, \hat{\delta}_k$ as the estimator of $S_Y|Z$.

5. Let $\hat{\beta}_i = \hat{\Sigma}^{-1/2} \gamma_i$ and $\hat{\eta}_i = \hat{\Sigma}^{-1/2} \delta_i$. Then, $\hat{\beta}_1, \ldots, \hat{\beta}_k$ or $\hat{\eta}_1, \ldots, \hat{\eta}_k$ will be the estimators of $S_Y|X$.

For determining $k$, the idea is similar to the SIR case. See K.-C. Li (1991).

The disadvantage of pHd is that it is unable to find linear trends.
1.2.5 Iterative Hessian Transformation

Cook and Li (2002) introduced the concept of the Central Mean Subspace (CMS), a natural inferential object for dimension reduction when the mean function $E(Y|X)$ is of interest. CMS is a subspace of CS that captures all the information in the conditional mean function $E(Y|X)$. In this paper, they also studied the properties of CMS and proposed methods to estimate CMS. Iterative Hessian Transformation (IHT) is one of them. Cook and Li (2004) give an asymptotic analysis of IHT and provide step-wise asymptotic hypothesis tests to determine the dimension of CMS, as estimated by IHT.

Definition 1.2.2. If $Y \perp E(Y|X)|\alpha^TX$, then Span($\alpha$) is a mean dimension reduction subspace for the regression of $Y$ on $X$.

Definition 1.2.3. Let $S_{E(Y|X)} = \bigcap S_m$ where intersection is over all mean dimension reduction subspaces $S_m$. If $S_{E(Y|X)}$ is itself a mean dimension reduction subspace, it is called the central mean dimension reduction subspace, or simply the Central Mean Subspace (CMS).

Similar to CS, CMS also has an invariance property, by which we can work with the standardized predictor in estimation without loss of generality.

Proposition 1.2.1. Let $S_{E(Y|X)}$ be the CMS for the regression of $Y$ on $X$, and let $S_{E(Y|Z)}$ be the CMS for the regression of $Y$ on $Z = A^TX + b$ where $A$ is a nonsingular, $p \times p$ matrix. Then $S_{E(Y|Z)} = A^{-T}S_{E(Y|X)}$.

By surveying the then available methods which were originally proposed to estimate CS, Cook and Li (2002) find that OLS and related convex objective functions based methods, and pHd are actually methods to estimate CMS, while SIR and
SAVE are methods to estimate CS. There is a detailed discussion differentiating estimators for CMS and CS in the non-published lecture notes B. Li (2003).

The theoretical basis for the IHT method is the following theorem.

**Theorem 1.2.9.** Suppose the linear conditional mean assumption holds and that $X$ is standardized. The CMS is an invariant subspace of the linear transformation $v \mapsto Hv$ (where $H$ can be either $H_1$ or $H_2$). In symbols,

$$H_1S_{E(Y|X)} \subseteq S_{E(Y|X)}$$
$$H_2S_{E(Y|X)} \subseteq S_{E(Y|X)}$$

Since this result only needs the LCM assumption for $S_{E(Y|X)}$ to be an invariant subspace, if we can find a “seed vector” in $S_{E(Y|X)}$ under LCM assumption, then we can derive other vectors in $S_{E(Y|X)}$ without the CCV assumption. As the OLS vector $\alpha = E(XY)$ belongs to $S_{E(Y|X)}$, we can use it as the “seed vector”.

**Corollary 1.2.1.** Under the LCM assumption,

$$\text{Span}\{H_1^j\alpha : j = 0, 1, \ldots\} \subseteq S_{E(Y|X)}$$
$$\text{Span}\{H_2^j\alpha : j = 0, 1, \ldots\} \subseteq S_{E(Y|X)}$$

For choosing the size of $j$ for the first $j$ vectors to exhaust all possible vectors in the sequence, Cook and Li (2002) give the following proposition.

**Proposition 1.2.2.** Let $A$ be a $p \times p$ matrix and $\alpha$ be a $p$-dimensional vector. If $A^j\alpha$ belongs to the subspace spanned by $\alpha, \ldots, A^{j-1}\alpha$, then so does $A^s\alpha$ for any $s > j$.

By the above proposition, in estimation, we can just focus on the vectors
\[ \alpha, \ldots, H^{p-1}\alpha \] without missing any vector in the subsequent iteration.

The sample estimating procedure can be described as follows.

1. Standardize and centerize to get \( Z \)'s and \( \hat{Y} \)'s.

2. Construct the OLS estimator \( \hat{\alpha} \) based on \( Z \) and \( \hat{Y} \).

3. Construct Hessian matrix \( \hat{H} \) based on \( Z \) and \( \hat{Y} \). Here, \( \hat{H} \) can be either \( \hat{H}_1 \) or \( \hat{H}_2 \).

4. Let

\[ \hat{B} = (\hat{\alpha}, \ldots, \hat{H}^{p-1} \hat{\alpha}). \]

Assume \( k \) is known. Let \( \eta_1, \ldots, \eta_q \) be the eigenvectors of \( \hat{B}\hat{B}^T \) corresponding to the \( k \) largest eigenvalues. This is used to estimate \( S_{E(Y|Z)} \).

5. Transform them back, as described before, to estimate \( S_{E(Y|X)} \).

In Cook and Li (2004), the authors developed two asymptotic tests for the dimension \( k \) of the IHT subspace, \( S_{IHT} \). The tests use the same statistic \( T_j = \hat{C}_2^{-1} \sum_{i=j+1}^{p} \hat{\lambda}_i \) for the hypothesis \( \text{rank}(B) = j \), but have different reference distributions depending on characteristics of the regression.

The availability of these tests means that IHT is a fully functioning methodology like pHd, while it does not need the constant conditional variance condition for either estimation or testing. When pHd is applicable the IHT method automatically combines pHd with OLS, taking advantage of the ability of OLS to find a linear trend in the mean function, and the ability of pHd to find nonlinear trends.
1.2.6 Minimum Average Variance Estimation

While a condition like LCM of the classical methods is not necessarily restrictive for many data sets, it is for time series data. Also, that will affect the corresponding dimension determination. To avoid this strong assumption on the predictor $X$, Xia et al. (2002) proposed a method to estimate the dimension reduction directions. They call it the (conditional) Minimum Average Variance Estimation (MAVE) method. This approach is inspired by the SIR method, the Average Derivative Estimation (ADE) method (Hädle and Stoker, 1991) and the idea of local linear smoothings. It is easy to implement and needs no strong assumptions on the probabilistic structure of $X$. However, a drawback is that it employs multivariate kernel, which make it susceptible to the curse of dimensionality.

While the above methods do not need to estimate the regression relation between $Y$ and $X$, MAVE determines the relationship between $X$ and $Y$ by of local linear smoothing estimation. Minimization of the following expression can be seen as a combination of nonparametric function estimation and direction estimation, which is executed simultaneously with respect to the directions and nonparametric link function:

$$
\min_{B:B^TB=I} \left\{ \sum_{j=1}^{n} \hat{\sigma}^2_B(B^TX_j) \right\} = \min_{B:B^TB=I} \left( \sum_{j=1}^{n} \sum_{i=1}^{n} [y_i - \{a_j + b_j^TB^T(X_i - X_j)\}]^2 \omega_{ij} \right). \tag{1.5}
$$

Xia et al. (2002) gives two methods to obtain the weights $\omega_{ij}$ in the above expression. Also, they extended the cross-validation method to determine the dimension of the central space. Details related with semiparametrics such as bandwidth
and algorithm are also mentioned.

The estimating procedure could be described as follows.

Step 1 (directions): for each $d$, $1 \leq d \leq p$, we search for the $d$ directions as follows.

(a) Initial value: use the multidimensional kernel weight to obtain an initial estimate of possible EDR directions $\beta_1, \beta_2, \ldots, \beta_d$ by minimizing the expression.

(b) Refined estimation: let $\hat{B} = (\beta_1, \beta_2, \ldots, \beta_d)$ constitute the latest estimator of $B$. Therefore we obtain refined kernel weights by using equation (2.8) in Xia et al. (2002). We refine the estimator via the goal expression using the refined kernel weights. Continue this procedure until convergence. The $CV(d)$ values can be obtained by using the final estimators of the directions.

Step 2 (dimension and output results): compare the $CV(d)$, $0 \leq d \leq p$. The $d$ with the smallest $CV(d)$ value is the estimated dimension. The corresponding estimator of $B$ in step 1(b) gives the estimated EDR directions.

Let $\hat{B}_a$ and $\hat{B}_b$ be the estimation of $B$ in two adjacent iterations in step 1(b). A suggested stopping rule for step 1(b) is when the distances between the subspaces $\mathcal{S}(\hat{B}_a)$ and $\mathcal{S}(\hat{B}_b)$ as measured by $m(\hat{B}_a, \hat{B}_b)$ (given in the paper) in several adjacent iterations are each less than a pre-set tolerance.

Provided that the dimension is chosen correctly, the rate of consistency for $\hat{B}$ is $O_p[h_{opt}^3 \log(n)]$ if one uses the optimal bandwidth $h_{opt}$ of the regression function estimation in the sense of minimizing the mean integrated squared errors. This is faster than the rate that is achieved by other semiparametric methods, which is $O_p[h_{opt}^2]$. Unlike the SIR method, the MAVE method is well adapted to time series and is robust against outliers. As argued in Xia et al. (2002), one possible explanation is that SIR uses $Y$ to produce the kernel weight, so its efficiency will suffer from
fluctuations in the link function. That is, there are double effects by applying
the \( y \)-based one-dimensional kernel in SIR. The gain in efficiency is well known,
however, with fluctuations in the link function, SIR may loss in efficiency. The
balance of these two effects will determine the behavior of SIR.

1.2.7 Contour Regression

Though all the methods above except MAVE have the advantages of \( \sqrt{n} \)-consistency
and computational ease, they also have common limitations. First, they all require
the LCM condition, as \( \beta \) is unknown, they need to satisfy the ellipticity condition
which guarantees LCM. Second, none of them guarantees exhaustive estimation
of the central space. A typical example is the heavy reliance of OLS and SIR on
linear trend in the dependence of \( Y \) on \( X \).

B. Li, Zha, and Chiaromonte (2005) proposed a novel approach based on esti-
mating contour directions of small variation in the response. The contour directions
span the orthogonal complement of the minimal space relevant for the regression.
Two methods are proposed to extract the contour directions, Simple and General
Contour Regressions (SCR and GCR) according to two measures of variation in
the response.

The contour-based methodology guarantees exhaustiveness under ellipticity
and other mild additional assumptions, while maintaining \( \sqrt{n} \)-consistency and
computational ease. Also, it proves robust to departures from ellipticity.

First, I will introduce simple contour directions.

To facilitate communication, I will review several terms. Let \((X_1, Y_1), \ldots, (X_n, Y_n)\)
be independent copies of the random pair \((X, Y)\), where \(X \in \mathbb{R}^p\) and \(Y \in \mathbb{R}\). Let
\(F_{XY}\) be the joint distribution of \((X, Y)\), let \(F_n\) be the corresponding empirical
distribution of the data. Let $T(F_{XY})$ be a dimension reduction estimator on the population level. If the columns of $T(F_{XY})$ belongs to $S_{Y|X}$, then we say that $T(F_n)$ is unbiased at the population level. If the columns of $T(F_{XY})$ actually span $S_{Y|X}$, then we say that $T(F_n)$ is exhaustive at the population level. If $T(F_n)$ converges at $\sqrt{n}$ rate to $T(F_{XY})$ in the first case, then we say that it is $\sqrt{n}$-consistent. If the $\sqrt{n}$ convergence holds in the second case, then we say $T(F_n)$ is $\sqrt{n}$-exhaustive.

Let $(\tilde{X}, \tilde{Y})$ be an independent copy of $(X, Y)$, we need an assumption before the main result.

**Assumption 1.2.2.** For any choice of vectors $v \in S_{Y|X}$ and $w \in (S_{Y|X})^\perp$ such that $\|v\| = \|w\| = 1$, and any sufficient small $c < 0$, we have

$$\text{var}[w^T (\tilde{X} - X) | |\tilde{Y} - Y| \leq c] > \text{var}[v^T (\tilde{X} - X) | |\tilde{Y} - Y| \leq c]$$

We believe this assumption is a reasonable one; the authors checked numerically numerous $f(\cdot)$ and never encountered a violation of it.

Let matrix $K(c)$ be defined as,

$$K(c) = E[(\tilde{X} - X)(\tilde{X} - X)^T | |\tilde{Y} - Y| \leq c],$$

**Theorem 1.2.10.** Suppose that $X$ has an elliptical distribution with $E(X) = 0$ and $\text{var}(X) = I_p$. If Assumption 1.2.2 holds, then, for a sufficiently small $c$, the eigenvectors of the matrix $K(c)$ corresponding to its smallest $k$ eigenvalues span the central subspace $S_{Y|X}$.

The sample estimating procedure can be described as follows.

1. Compute sample mean and variance matrix of the predictor $X$. 
2. Compute the matrix-valued U-statistic

\[ \hat{H}(c) = \frac{1}{\binom{n}{2}} \sum_{(i,j) \in N} (X_j - X_i)(X_j - X_i)^T I(|Y_j - Y_i| \leq c), \]

where \( N \) is the index set \( \{(i,j): i = 2, \ldots, n; j = 1, \ldots, i - 1\} \).

3. Perform principal component analysis on

\[ \hat{\Sigma}^{-1/2} \hat{H}(c) \hat{\Sigma}^{-1/2} \]

and let \( \hat{\gamma}_{p-k+1}, \ldots, \hat{\gamma}_p \) be the eigenvectors corresponding to the smallest \( k \) eigenvalues.

4. The span of these eigenvectors estimates \( S_{Y|Z} \), where \( Z \) is the standardized version of \( X \). Thus, the estimation of the central space is

\[ \hat{S}_{Y|X} = \text{Span}(\hat{\Sigma}^{-1/2}\hat{\gamma}_{p-k+1}, \ldots, \hat{\Sigma}^{-1/2}\hat{\gamma}_p). \]

The contour regression method provides the estimates which are \( \sqrt{n} \)-consistent for vectors in \( S_{Y|X} \), and these vectors span the whole central space. Therefore, the estimator is a \( \sqrt{n} \)-exhaustive estimator of \( S_{Y|X} \). That is,

\[ \hat{\Sigma}^{-1/2}\hat{\gamma}_{p-q+1}, \ldots, \hat{\Sigma}^{-1/2}\hat{\gamma}_p \]

is a set of \( \sqrt{n} \)-exhaustive estimators.

Second, General Contour Regression.

The idea underlying SCR is to use inequality \( |Y - \widehat{Y}| \leq c \) to identify vectors aligned with the contour directions. When the regression function is non-monotone, this inequality also picks up other directions. The inequality \( |Y - \widehat{Y}| \leq c \)
is not a very sensitive contour identifier for non-monotone functions. To overcome this drawback, the authors proposed another more sensitive contour identifier.

Let \( l(t; x_i, x_j) = (1 - t)x_i + tx_j, \ t \in \mathbb{R} \) be the straight line that goes through \( x_i \) and \( x_j \), and define

\[
V(x_i, x_j) = \text{var}(Y|X = l(t; x_i, x_j) \text{ for some } t \in \mathbb{R}),
\]

By replacing the \(|Y - \bar{Y}| \leq c\) criterion with \( V(x_i, x_j) \leq c' \), contour vectors can be more clearly identified with much noise screened out.

As for estimation with a sample, each line is replaced with a tube of radius \( \rho \) connecting \( X_i \) and \( X_j \). For a fixed \( \rho \), a corresponding estimator \( V(x_i, x_j; \rho) \) is calculated to estimate the variance along this tube. The other steps of estimation are similar with SCD. GCD also has the \( \sqrt{n} \)-exhaustive property.

As pointed in B. Li et al. (2005), there is a significant difference between contour regression and other global methods that gather directional information by slicing the response and processing predictor observations separately within each slice. The inter-slice directional information relevant to the regression is lost by slicing, while in contour regression, empirical directions \( (X_i - X_j) \) can “cut across” response slices to exploit inter-slice information. There is also a very good comparison between contour regression and SAVE in section 4.3 of B. Li et al. (2005).

1.2.8 Directional Regression

Classical dimension reduction methods such as SIR and SAVE have well known limitations due to their constructions. Based on the first conditional moments \( E(X|Y) \), SIR fails when the response surface is symmetric about the origin; based
on the second conditional moments $E(XX^T|Y)$, SAVE is not very efficient in estimating monotone trend for small to moderate sample sizes.

Several authors have proposed to combine various forms of first two conditional moments by convex combination, such as the convex combinations of SIR and SIR-II, of SIR and PHD of SIR and SAVE. See K.-C. Li (1991), Gannoun and Saracco (2003), Ye and Weiss (2003), and Zhu, Ohtaki, and Li (2005). Furthermore, Ye and Weiss (2003) proposed a bootstrap method that can be used to select the coefficient in the convex combination methods.

B. Li and Wang (2006) introduced Directional Regression (DR) as a method for dimension reduction. Like Contour Regression, it is derived from empirical directions, but achieves higher accuracy and requires substantially less computation. Directional Regression naturally synthesizes the dimension reduction estimators based on conditional moments, such as SIR and SAVE, and in doing so combines the advantages of both methods. Under mild conditions, it provides exhaustive and $\sqrt{n}$-consistent estimate of the dimension reduction space. They also develop the asymptotic distribution of the Directional Regression estimator, and therefore a sequential test procedure to determine the dimension of the Central Subspace.

Let $(\tilde{Z}, \tilde{Y})$ be an independent copy of $(Z,Y)$, by regressing the direction $Z - \tilde{Z}$, as represented by $(Z - \tilde{Z})(Z - \tilde{Z})^T$, onto the space of functions of $(Y, \tilde{Y})$ that are squared integrable with respect to $L_2(P_Y)$

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

The following theorem gives the theoretical background for this method.

**Theorem 1.2.11.** Suppose the LCM and CCV conditions are satisfied, then, $2I_p - A(Y, \tilde{Y}) = P[2I_p - A(Y, \tilde{Y})]P$. 
The above theorem suggests the following object

$$ G = E[2I_p - A(Y, \tilde{Y})]^2 $$

as the population version of the estimate of $S_{Y|Z}$.

By extending the expression of $G$, $(Z, Y)$ and $(\tilde{Z}, \tilde{Y})$ in $G$ could be separated. $G$ could be expressed as a nonlinear functional of conditional moments of $Z$ given $Y$. Compared with $O(n^2)$ operations required by SCR and $O(n^3)$ operations required by GCR, DR only need $O(n)$ operation. This re-expression can save computing time and simplify the asymptotic development.

Under the same mild conditions, $S_{DR}$ is exhaustive and the following theorem gives a result that gives us some illumination about the status of DR on dimension reduction.

**Theorem 1.2.12.** Suppose that the moments involved in the definition of SAVE and DR are finite. Then $S_{DR} = S_{SAVE}$.

Being estimating the same space, the differences in efficiency and accuracy between SAVE and DR may comes from the latter one’s coordination of the first two conditional moments through the efficient arrangement of the empirical directions. The interaction terms between first and second conditional moments reflect the difference between arranging the predictor $X$ according to $Y$ and arranging the empirical directions according to $Y$ and $\tilde{Y}$.

The sample estimation procedure is similar with other method, using the sample estimate of the discretized version of $G$ to estimate $S_{Y|Z}$. They also proposed a sequential test to determine the dimension.

Simulation results are very promising. And in the identification of hand-written digits example, DR is shown to be a very accurate identifier compared with classical
methods.

DR, like SAVE, SIR-II, PHD and SCR, is a second-moment based method. Hence, for the situation where the regression surfaces have highly fluctuating shapes like high frequency trigonometric functions, methods such as GCR may perform better.

1.3 Several Issues in Dimension Reduction

Different methods have different advantages and drawbacks. By categorizing estimation methods from different point views, we want to draw conclusions on main characters of methods.

1.3.1 Global Methods and Local Methods

Most of the methods we reviewed are performed on the predictor space globally. These methods include Ordinary Least Squares (OLS), Sliced Inverse Regression (SIR), Slice Average Variance Estimation (SAVE), Principle Hessian Directions (pHd), Simple Contour Regression (SCR), and Directional Regression (DR). These methods share the common features such as simple estimation, computational speed, $\sqrt{n}$-consistency estimator. But they are also restrictive with distribution of predictor, while OLS, SIR only require linear conditional mean assumption, SAVE, pHd, SCR and DR all based on both linear conditional mean and constant conditional variance assumptions. Also, global methods do not guarantee to recover the whole central subspace. For example, OLS can only estimate one direction, SIR cannot estimate functions symmetric about the origin, pHd cannot find linear trends.

Most local methods, including Minimum Average Variance Estimation, are
nonparametric methods that do not have much restriction on the design distribution. These methods rely both on dimension reduction ideas which give a parsimonious characterization of the data and on the computation power of nonparametric smoothing methods to exploit local dependence of $Y$ on $X$. However, the advantages are balanced by slow convergence rate of estimators and computational complexity.

1.3.2 Central Subpace and Central Mean Subspace

While SIR and SAVE estimators both belong to CS, they may not belong to CMS. This is determined by their reliance on the conditional independence assumption

$$Y \perp X | \beta^T X,$$  \hspace{1cm} (1.6)

while deductions of CMS estimators only need

$$Y \perp E(Y|X) | \alpha^T X.$$  \hspace{1cm} (1.7)

For example, in proving that OLS is in CMS, we used

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

In the fourth step, while both 1.7 and 1.6 are sufficient, only 1.7 is necessary.

In proving that SIR is in CS, we used

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

where the second step can only be derived with 1.6.

If, in a regression problem, we are mainly interested in the estimation of regression function, we can apply OLS and pHd which are estimators of CMS.
Global Dimension Reduction on
Feature Space (G-DRFS)

In traditional dimension reduction, we make the assumption that $X$ and $Y$ being independent given $\beta^T X$; that is,

$$Y \perp X | \beta^T X.$$  \hfill (2.1)

However, this model may not be suitable for some regression relation or it may lose power in some case for reducing dimension. For example, when we have interaction terms in the model, each unique factor in the product results in one more direction in the dimension reduction space. Though we still can apply dimension reduction methods to this situation, we may lose power of detection because too many linear directions are needed to capture the interactions, which are nonlinear in nature. For illustration, below is an extreme example of this type.

**Example 2.0.1.** Suppose $X \sim N(0, I_6)$, $\varepsilon \sim N(0, 0.1^2)$, $\varepsilon$ is independent of $X$, and $Y = X_1X_2 + X_3X_4 + X_5X_6 + \varepsilon$. Then the dimension reduction space as defined
by relation (2.1), is the whole space $\mathbb{R}^6$. There is no dimension reduction at all.

How to tackle this kind of problem? We want to borrow the idea of the feature space used in Support Vector Machines to solve our problem.

2.1 Idea

2.1.1 Introducing the idea of feature

Support Vector Machines (SVMs) (Cristianini and Shawe-Taylor, 2000) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers.

In a typical two-category classification problem, by viewing each data point as a $p$-dimensional vector, we have two clouds of data points in $\mathbb{R}^p$. SVM will construct a $(p - 1)$-dimensional separating hyperplane, which is called a linear classifier, by...
Figure 2.2. Data cannot be separated by a linear function.

Figure 2.3. A feature map can simplify the classification task.
maximizing the margin between the two data clouds. By maximizing the margin, SVM aims at decreasing the error rate for applying the constructed classifier to new data points. Figure 2.1 is a simplified illustration of this idea. In the two dimensional data space, we have two clusters of data points. The line acts as the separating hyperplane; it is chosen to maximize the distance between the two clusters.

However, most often, real-world applications cannot be expressed as simple linear combinations of the original data. Figures 2.2 and 2.3 together illustrate this point. In Figure 2.2, there are two categories of two-dimensional data points, O and X, respectively. They could not be separated by a hyperplane, which should be a line since \( p = 2 \). Instead, in this case, they could be differentiated by a curve. In Figure 2.3, by changing the representation of the points to some function of them, that is, by mapping the data to another space, we can find a hyperplane in the target space that clearly separates the two transformed sets.

The process of changing the representation of the data could be denoted by the following mapping:

\[
X = (X_1, X_2, \ldots, X_p) \mapsto \phi(X) = (\phi_1(X), \ldots, \phi_r(X)).
\]

This idea has been in machine learning for a long time. In SVM terms, the quantities introduced to describe the data are usually called features, while the original quantities are sometimes called attributes. The task of choosing the most suitable representation is known as feature selection. The original space \( X \) is referred to as the input space, while the space \( F = \{\phi(X) : X \in X\} \) is called the feature space.

To adapt this idea into sufficient dimension reduction for regression, let \( X \) be
the random predictor vector in $\mathbb{R}^p$. Let $Y$ be a random response variable in $\mathbb{R}$. Let $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^r$ be the feature function. Here, usually, $r \geq p$. For example, when we extend the predictor by including features of $X$ up to degree 2 polynomials, we can obtain,

$$
\phi(X)^T = (X_1, \ldots, X_p, X_1^2, X_2^2, \ldots, X_p^2, X_1X_2, X_1X_3, \ldots, X_{p-1}X_p)
$$

(2.2)

We can make further extension when necessary. And we assume $\phi(X)$ and $Y$, or equivalently $X$ and $Y$ to be independent given $\alpha^T\phi(X)$,

$$
Y \perp \phi(X)|\alpha^T\phi(X) \quad \text{or} \quad Y \perp X|\alpha^T\phi(X). \quad (2.3)
$$

We use $\alpha \in \mathbb{R}^{r \times d}$, $d \leq r$ to denote the $r \times d$ matrix with columns $\alpha_i$, $i = 1, \cdots, d$. Our goal is to estimate $\alpha$. The direct benefit is that we can partially solve the question raised in Example 2.0.1. Consider, in the attribute space (in SVM terms) of dimension $p = 6$, we will have a $S_{Y|X}$ of dimension $k = 6$. In view of dimensionality, there is no reduction at all. But when we think about the problem in terms of the features, suppose the feature mapping includes terms up to degree 2 polynomials as defined in (2.2), while the random feature vector has length $r = 27$, we have a dimension reduction subspace with basis matrix $\alpha$ of dimension $d = 1$, and

$$
\alpha = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1)^T.
$$

To illustrate this reduction from $k = \dim(S_{Y|X})$ to $d = \dim(S_{Y|\phi(X)})$ when we change the predictor space from $\mathbb{R}^p$ to the extended Euclidean space $\mathbb{R}^r$, limited by the 3D display ability, we use the following example by letting $p = r = 2$. 

Figure 2.4. 3D scatter plots of $Y$ versus attributes $X_1$ and $X_2$ in Example 2.1.1.

Figure 2.5. 3D scatter plots of $Y$ versus features $X_1^2$ and $X_2^2$ in Example 2.1.1.
Example 2.1.1. Let $p = 2$, and $\phi(X) = (X_1^2, X_2^2)$. Suppose $X = (X_1, X_2)$, where $X_1 \sim \text{Uniform} (-1,1)$, $X_1 \sim \text{Uniform} (-1,1)$, $\varepsilon \sim \text{Normal} (0,0.01)$, $Y = X_1^2 + X_2^2 + \varepsilon$, sample size $n = 1000$.

Figure 2.4 is the 3D plot of $Y$ versus $X$; Figure 2.5 is the 3D plot of $Y$ versus $\phi(X)$. By comparing the two plots, we can see with the attributes plot we can detect no direction for dimension reduction; with the feature plot, we can see a clear trend of $Y$ against a linear combination of the features.

2.1.2 Non-linear Predictor Space

While the reduction of structural dimension on the model basis looks promising, we face problems with the nonlinearity of feature spaces. That is, for a given feature mapping $\phi(X)$ with length $r$, the feature space is a manifold embedded in the $\mathbb{R}^r$ space. In this situation, it is not possible for the predictor space to have the elliptical contoured predictor distribution required by most existing dimension reduction methods. That is also the reason why we say by applying dimension reduction to feature spaces, we have only partially solved the problem, since there are still difficulties to tackle.

We here propose five estimation methods to estimate dimension reduction subspace in feature space. Each of these methods is trying to solve this nonlinearity problem, in some respect and to some extent. They are,

1. Global dimension reduction on the feature space;

2. Invariant Aggregation (IA) dimension reduction on the feature space;

3. Nonlinear Aggregation (NA) dimension reduction on the feature space;

4. Corrected NA dimension reduction on the feature space;
5. Central Solution Space (CSS) dimension reduction on the feature space.

In the rest of this chapter, we will present the first method. That is, applying existing dimension reduction methods directly on the feature space. Before applying any of these methods, we need to check the key assumptions of them.

### 2.2 Approximate Normality of Projected Vector

When the predictor is $X$, while $\beta$ is unknown, ellipticity guarantees that LCM holds for all $\beta \in \mathbb{R}^p$, and multivariate normality guarantees that both LCM and CCV hold. Since $\phi(X)$ is a nonlinear manifold in the extended Euclidean space, it does not have an elliptical distribution. However, Diaconis and Freedman (1984) show that when the dimension is reasonably high, “most” projections of a random vector are approximately normal.

In this section, we prove that the asymptotic normality result from Diaconis and Freedman (1984) could be applied to specific feature vectors. That is, for degree two random polynomial feature vectors generated from multivariate normal distributions, we will prove that low dimensional projections of the feature vectors have an asymptotic multivariate normal distribution.

We first introduce this asymptotic result from Diaconis and Freedman (1984) in terms of an empirical distribution. Let $x_1, x_2, \ldots, x_n$ be non-random vectors in $\mathbb{R}^p$. Suppose that the size of the sample $n$, and the length of the vectors $p$, depend on a hidden index $\nu$ and will tend to infinity as $\nu$ does. Suppose that for $\sigma^2$ positive and finite, for any $\varepsilon > 0$, as $\nu \to \infty$,

$$\frac{1}{n} \text{card}\{j \leq n : \|x_j\|^2 - \sigma^2 p > \varepsilon p\} \to 0 \quad (2.4)$$
\[
(1/n^2) \text{card}\{1 \leq j, k \leq n : |x_j \cdot x_k| > \varepsilon p\} \to 0 \quad (2.5)
\]

The first condition says that most vectors have length near \(\sigma^2 p\). The second condition says that most vectors are nearly orthogonal.

Let \(S_{p-1}\) be the unit sphere in \(\mathbb{R}^p\). Let \(\gamma\) have a uniform distribution on \(S_{p-1}\).

The projected data in direction \(\gamma\) is:

\[
\gamma^T x_1, \gamma^T x_2, \cdots, \gamma^T x_n. \quad (2.6)
\]

Let \(\theta_\nu(\gamma)\) be the empirical distribution of this sequence. The following theorem says that \(\theta_\nu(\gamma)\) is close to \(N(0, \sigma^2)\) for most \(\gamma\), for large \(\nu\).

**Theorem 2.2.1.** Under conditions 2.4 and 2.5, as \(\nu \to \infty\), the empirical distribution \(\theta_\nu\) tends to \(N(0, \sigma^2)\) weakly in probability.

By weakly in probability, it means the sequence of conditional characteristic functions \(E(e^{it\theta_\nu(\gamma)}|\gamma)\) converge (pointwise) in probability to a normal characteristic function. Intuitively, this means when \(\nu\) is large, the distribution of \(\theta_\nu(\gamma)\) is nearly normal for most of \(\gamma\)'s. Here, the parameter \(\gamma\) is treated as random to facilitate the notion of “most of \(\gamma\)'s”.

We wish to apply this result to the feature predictor \(\phi(X)\). In this paper, we only consider the case that the original predictor \(X\) has a multivariate normal distribution and the feature vectors only include features of \(X\) up to degree 2 polynomials. Here dimension \(p\) is used to index the sequence of random vector \(X\) as its dimension \(p \to \infty\). In symbols, it could be written as,

\[
X_p \sim N(0_p, I_p).
\]
The corresponding degree 2 polynomial feature random vector is

$$\phi^T(X_p) = (X_1^p, \ldots, X_p^p, X_p^{12}, X_p^{22}, \ldots, X_p^{p2}, X_p^{1}X_p^{2}, X_p^{1}X_p^{3}, \ldots, X_p^{p-1}X_p^{p})$$

In later derivations, when it is obvious, we will omit the dimension index $p$ to simplify notations.

Our goal is to establish that, when dimension $p$ is reasonably large, low dimensional projections of the feature vector $\phi(X_p)$ are approximately multivariate normal as $p \to \infty$.

Hall and Li (1993) has a closely related result for dimension reduction. Let $\alpha$ be uniformly distributed random vectors on the unit sphere $S_{p-1}$. Let the measure of non-linearity be

$$\|E(X|\alpha^T X = t)\|^2 - t^2.$$

They prove that the LCM condition holds for “most” of $\alpha$’s. That is, under certain conditions,

$$\|E(X|\alpha^T X = t)\|^2 - t^2 \xrightarrow{\text{i.p.}} 0 \text{ as } p \to \infty.$$

B. Li and Yin (2007) also used the results of Diaconis and Freedman (1984) and Hall and Li (1993). Parallel assumptions and conclusions have been drawn in their context. Translated to the feature vector scenario, it is equivalent to check the following two assumptions to get the approximate multivariate normality and consequently linearity results.

$$p^{-2}\phi^T(X_p)\phi(X_p) = E[p^{-2}\phi^T(X_p)\phi(X_p)] + o_p(1), \quad (2.7)$$

$$p^{-2}\phi^T(X_p)\phi(\tilde{X}_p) = o_p(1). \quad (2.8)$$
Note that the denominators in conditions 2.7 and 2.8 are \( p^2 \) instead of \( p \). In Diaconis and Freedman (1984) the dimension of the vector \( X \) is \( p \), in our extended feature vector case, \( p \) is the dimension of the original predictor. For degree 2 polynomial feature, the length of \( \phi(X) \) is \( \frac{1}{2}p^2 + \frac{3}{2}p \), which is of order \( p^2 \).

We have the following Lemma to conclude our above reasoning.

**Lemma 2.2.1.** Let \( X_p \) be a sequence of multivariate normal random vectors in \( \mathbb{R}^p \). Let \( \phi : \mathbb{R}^p \to \mathbb{R}^r \) be defined as in 2.2. Let \( \alpha_r \) be a sequence of uniform random vectors on \( S_{r-1} \). Then, when assumptions 2.7 and 2.8 are satisfied, the conditional distribution \( \alpha_r^T \phi(X_p)|\alpha_r \) converges weakly in probability to a normal distribution as \( p \to \infty \).

In the following subsections, we will check the two assumptions to prove the lemma.

### 2.2.1 Condition 1

To check condition 2.7, first, by Chebychev’s inequality,

\[
P(p^{-2}|\phi^T(X_p)\phi(X_p) - E(\phi^T(X_p)\phi(X_p))| > \varepsilon) \leq \frac{1}{\varepsilon^2 p^4} E[|\phi^T(X_p)\phi(X_p) - E(\phi^T(X_p)\phi(X_p))|^2]
\]

(2.9)

The expectation on the RHS could be expanded as

\[
E[\phi^T(X_p)\phi(X_p) - E(\phi^T(X_p)\phi(X_p))]^2
\]

(2.10)

\[
=E(\phi^T(X_p)\phi(X_p))^2 - 2[E(\phi^T(X_p)\phi(X_p))]^2 + [E(\phi^T(X_p)\phi(X_p))]^2
\]

(2.11)

\[
=E(\phi^T(X_p)\phi(X_p))^2 - [E(\phi^T(X_p)\phi(X_p))]^2
\]

(2.12)
The second term on expression (2.12) is easier to calculate

\[-[E(\phi^T(X_p)\phi(X_p))]^2 = -\frac{\sigma^8}{4}p^4 + o(p^4)\]  \hspace{1cm} (2.13)

The first term on expression (2.12) is

\[E[\phi^T(X_p)\phi(X_p)]^2 = E[\sum_{i=1}^{1+\frac{1}{2}p^2+\frac{3}{2}p} \phi(X_p)]^2\]  \hspace{1cm} (2.14)

The result will be the elementwise summation of the following matrix

\[
\begin{pmatrix}
1 & x_1^2 & \ldots & x_p^2 & x_1^4 & \ldots & x_p^4 & x_1^2x_2 & \ldots & x_p^2x_p \\
x_1^2 & x_1^4 & \ldots & x_1^2x_p^2 & x_1^6 & \ldots & x_1^4x_p^4 & x_1^2x_2x_p & \ldots & x_1^2x_p^2x_p \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_p^2 & x_p^4 & \ldots & x_p^2x_p^2 & x_p^6 & \ldots & x_p^4x_p^4 & x_p^2x_2x_p & \ldots & x_p^2x_p^2x_p \\
x_1^4 & x_1^6 & \ldots & x_1^4x_p^4 & x_1^8 & \ldots & x_1^6x_p^6 & x_1^4x_2x_p^4 & \ldots & x_1^4x_p^2x_p^4 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_p^4 & x_p^6 & \ldots & x_p^4x_p^4 & x_p^8 & \ldots & x_p^6x_p^6 & x_p^4x_2x_p^4 & \ldots & x_p^4x_p^2x_p^4 \\
x_1^2x_2 & x_1^4x_2 & \ldots & x_1^2x_2x_p^2 & x_1^4x_2x_p^4 & \ldots & x_1^2x_2x_p^6 & x_1^4x_2x_2x_p^4 & \ldots & x_1^2x_2x_p^2x_p^4 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_p^2x_p & x_p^4x_p & \ldots & x_p^2x_p^2x_p & x_p^4x_p^4x_p & \ldots & x_p^2x_p^6x_p & x_p^4x_p^4x_2x_p & \ldots & x_p^4x_p^2x_p^4x_p \\
x_1^2x_2^2 & x_1^4x_2^2 & \ldots & x_1^2x_2^2x_p^2 & x_1^4x_2^2x_p^4 & \ldots & x_1^2x_2^2x_p^6 & x_1^4x_2^2x_2x_p^4 & \ldots & x_1^2x_2^2x_p^2x_p^4 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_p^2x_p^2 & x_p^4x_p^2 & \ldots & x_p^2x_p^2x_p & x_p^4x_p^4x_p & \ldots & x_p^2x_p^6x_p & x_p^4x_p^4x_2x_p & \ldots & x_p^4x_p^2x_p^4x_p \\
x_1^2x_2^2x_2 & x_1^4x_2^2x_2 & \ldots & x_1^2x_2^2x_2x_p^2 & x_1^4x_2^2x_2x_p^4 & \ldots & x_1^2x_2^2x_2x_p^6 & x_1^4x_2^2x_2x_2x_p^4 & \ldots & x_1^2x_2^2x_2x_p^2x_p^4 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_p^2x_p^2x_p & x_p^4x_p^2x_p & \ldots & x_p^2x_p^2x_p^2 & x_p^4x_p^4x_p^2 & \ldots & x_p^2x_p^6x_p^2 & x_p^4x_p^4x_2x_p^2x_p & \ldots & x_p^4x_p^2x_p^2x_p^4x_p \\
\end{pmatrix}
\]  \hspace{1cm} (2.15)

To take the expectation, we need formula for moments of normal distribution, the central moments are:

\[\mu_{2n+1} = 0\]  \hspace{1cm} (2.16)

\[\mu_{2n} = \frac{(2n)!}{2^n \cdot n!} \sigma^{2n}\]  \hspace{1cm} (2.17)
We can easily find that while the expectation matrix is of dimension $\frac{p^2+3p+1}{2}$ by $(2.18)$, only the elements from the bottom right block of dimension $\frac{p(p-1)}{2}$ will have sum of order $p^4$. We can take it out and observe it more carefully.

\[
\begin{pmatrix}
1 & \sigma^2 & \ldots & \sigma^2 & 3\sigma^4 & \ldots & 3\sigma^4 & \sigma^4 & \ldots & \sigma^4 \\
\sigma^2 & 3\sigma^4 & \ldots & \sigma^4 & 15\sigma^6 & \ldots & 3\sigma^6 & 3\sigma^6 & \ldots & \sigma^6 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\sigma^2 & \sigma^4 & \ldots & 3\sigma^4 & 3\sigma^6 & \ldots & 15\sigma^6 & \sigma^6 & \ldots & 3\sigma^6 \\
3\sigma^4 & 15\sigma^6 & \ldots & 3\sigma^6 & 105\sigma^8 & \ldots & 9\sigma^8 & 15\sigma^8 & \ldots & 3\sigma^8 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
3\sigma^4 & 3\sigma^6 & \ldots & \sigma^6 & 15\sigma^8 & \ldots & 3\sigma^8 & 9\sigma^8 & \ldots & \sigma^8 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\sigma^4 & \sigma^6 & \ldots & 3\sigma^6 & 3\sigma^8 & \ldots & 15\sigma^8 & \sigma^8 & \ldots & 9\sigma^8 \\
\end{pmatrix}
\]

\[
E \begin{pmatrix}
x_1^4 x_2^4 & x_1^4 x_2^2 x_3^2 & \ldots & x_1^4 x_2^2 x_3 x_4 & x_1^4 x_2^2 x_3^2 & \ldots & x_1^4 x_2^2 x_3 x_4 \\
x_1^4 x_2^2 x_3^2 & x_1^4 x_3^4 & \ldots & x_1^4 x_2 x_3^2 & x_1^4 x_2^2 x_3^2 & \ldots & x_1^4 x_2^2 x_3 x_4 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x_1^4 x_2^2 x_3^2 & x_1^4 x_3^4 & \ldots & x_1^4 x_2 x_3^2 & x_1^4 x_2^2 x_3 x_4 & \ldots & x_1^4 x_2^2 x_3 x_4 \\
x_1^4 x_2 x_3 x_4 & x_1^4 x_3 x_4 & \ldots & x_1^4 x_2^2 x_3 x_4 & x_1^4 x_2^2 x_3 x_4 & \ldots & x_1^4 x_2^2 x_3 x_4 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x_1^4 x_2 x_3 x_4 & x_1^4 x_2^2 x_3 x_4 & \ldots & x_1^4 x_2^2 x_3 x_4 & x_1^4 x_2^2 x_3 x_4 & \ldots & x_1^4 x_2^2 x_3 x_4 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x_1^4 x_2 x_3 x_4 & x_1^4 x_2^2 x_3 x_4 & \ldots & x_1^4 x_2^2 x_3 x_4 & x_1^4 x_2^2 x_3 x_4 & \ldots & x_1^4 x_2^2 x_3 x_4 \\
\end{pmatrix}
\]

\[
(2.19)
\]
While each row has one $9\sigma^8$ term and $(2p - 4)$ $3\sigma^8$ terms, and the rest $\sigma^8$ terms, the terms with $9$ or $3$ as multipliers could not sum to the $p^4$ order. So, we have only $\sigma^8$ terms in each entry sum to the order of $p^4$, with number of entry being

$$\left(\frac{p(p-1)}{2}\right)^2 = \frac{p^4}{4} + o(p^4)$$

That is,

$$[E(\phi^T(X_p)\phi(X_p))]^2 = \frac{\sigma^8}{4}p^4 + o(p^4)$$

Since the derivation of the above result took some trivial calculation. I just leave a note here for bookkeeping.

\[
\begin{pmatrix}
9\sigma^8 & 3\sigma^8 & \ldots & 3\sigma^8 & 3\sigma^8 & \ldots & 3\sigma^8 & \ldots & \sigma^8 \\
3\sigma^8 & 9\sigma^8 & \ldots & 3\sigma^8 & 3\sigma^8 & \ldots & \sigma^8 & \ldots & \sigma^8 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
3\sigma^8 & 3\sigma^8 & \ldots & 9\sigma^8 & \sigma^8 & \ldots & 3\sigma^8 & \ldots & 3\sigma^8 \\
3\sigma^8 & 3\sigma^8 & \ldots & \sigma^8 & 9\sigma^8 & \ldots & 3\sigma^8 & \ldots & \sigma^8 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
3\sigma^8 & \sigma^8 & \ldots & 3\sigma^8 & 3\sigma^8 & \ldots & 9\sigma^8 & \ldots & 3\sigma^8 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\sigma^8 & \sigma^8 & \ldots & 3\sigma^8 & \sigma^8 & \ldots & 3\sigma^8 & \ldots & 9\sigma^8 \\
\end{pmatrix}
\]
The two order \( p^4 \) terms in expression 2.12 cancel out, we are left with a result of \( o(p^4) \). So, the first assumption is satisfied.

### 2.2.2 Condition 2

Still, we suppose, \( X \sim N(0, \sigma^2 I_p) \), By Markov’s inequality,

\[
P(p^{-2}|\phi^T(X_p)\phi(\tilde{X}_p)| > \varepsilon) \leq \frac{1}{\varepsilon^2 p^4} E[\phi^T(X_p)\phi(\tilde{X}_p)]^2
\]

Let \( r = \frac{r^2 + 3p + 1}{2} \). The expectation on the right hand side of (2.23) is

\[
E[\sum_{i=1}^{r} \phi(\tilde{X}_p)_i \phi(\tilde{X}_p)_i]^2
\]

\[
= \sum_{i=1}^{r} \sum_{j=1}^{r} E[\phi(\tilde{X}_p)_i \phi(\tilde{X}_p)_j \phi(\tilde{X}_p)_i \phi(\tilde{X}_p)_j]
\]

\[
= \sum_{i=1}^{r} \sum_{j=1}^{r} E[\phi(\tilde{X}_p)_i \phi(\tilde{X}_p)_j] E[\phi(\tilde{X}_p)_i \phi(\tilde{X}_p)_j]
\]

\[
= ||E[\phi(X_p)\phi^T(X_p)]||_F^2.
\]

Here \( \| \cdot \|_F \) denotes the Frobenius norm. Notice that, since \( E(\phi(X_p)) \neq 0 \), different from proof of Lemma 6.2 in B. Li and Yin (2007), instead of \( \Sigma_{\phi(X_p)} \), we need the moment matrix \( E(\phi(X_p)\phi^T(X_p)) \). That is, the expectation of the following matrix,
\[
\begin{pmatrix}
1 & X_1 & \ldots & X_p & X_1^2 & \ldots & X_p^2 & X_1X_2 & \ldots & X_{p-1}X_p \\
X_1 & X_1^2 & \ldots & X_1X_p & X_1^3 & \ldots & X_1X_p^2 & X_1^2X_2 & \ldots & X_1X_{p-1}X_p \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
X_p & X_1X_p & \ldots & X_p^2 & X_p^3 & \ldots & X_pX_{p-1}X_p & X_1X_p^2 & \ldots & X_1X_pX_{p-1}X_p \\
X_1^2 & X_1^2 & \ldots & X_1^2X_p & X_1^4 & \ldots & X_1^2X_p^2 & X_1^2X_2 & \ldots & X_1^2X_{p-1}X_p \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
X_p^2 & X_1X_p^2 & \ldots & X_p^3 & X_p^4 & \ldots & X_pX_{p-1}X_p^2 & X_1X_pX_p^2 & \ldots & X_1X_{p-1}X_{p}X_p \\
X_1X_2 & X_1^2X_2 & \ldots & X_1X_2X_p & X_1^3X_2 & \ldots & X_1X_2X_p^2 & X_1^2X_2^2 & \ldots & X_1X_2X_{p-1}X_p \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
X_{p-1}X_p & X_1X_{p-1}X_p & \ldots & X_{p-1}X_p^2 & X_1X_{p-1}X_p^2 & \ldots & X_{p-1}X_{p}X_{p-1}X_p & X_1X_{p-1}X_p^2 & \ldots & X_{p-1}X_{p}X_{p-1}X_p \\
\end{pmatrix}
\]

and we get

\[
\begin{pmatrix}
1 & 0^T & \sigma^2 & \sigma^2 & \ldots & \sigma^2 & 0^T \\
0 & \sigma^2I_p & 0 & 0 & \ldots & 0 & 0 \\
\sigma^2 & 0^T & 3\sigma^4 & \sigma^4 & \ldots & \sigma^4 & 0^T \\
\sigma^2 & 0^T & \sigma^4 & 3\sigma^4 & \ldots & \sigma^4 & 0^T \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\sigma^2 & 0^T & \sigma^4 & \sigma^4 & \ldots & 3\sigma^4 & 0^T \\
0 & 0 & 0 & 0 & \ldots & 0 & \sigma^4I_{C^2} \\
\end{pmatrix}
\]

Take the Frobenius norm, we can get

\[
\|E(\phi(X)\phi(X)^T)\|_F^2 = 1 + (3\sigma^4 + \frac{3}{2}\sigma^8)p + (\frac{3}{2}\sigma^8)p^2 = o(p^4)
\] (2.30)

So, when \(X \sim N(0, \sigma^2I_p)\), the second condition is also satisfied.

Since the two assumptions 2.7 and 2.8 are proved to be satisfied, we can apply Lemma 2.2.1 to projected feature vectors. That is, on most of the directions, \(\alpha^T\phi(X)\) will be approximately normal. With this result, we can apply existing
dimension reduction methods on feature space.

Simulations results with asymptotic normality will be found in the simulation section of this chapter.

2.3 Global Estimation on the Feature Space

With the asymptotic normality result from last section, we can apply existing dimension reduction methods on feature space directly. To differentiate this approach from later methods, we call it a global estimation method. With any existing method, we will first extend $X$ to $\phi(X)$, and replace the role of $X$ with $\phi(X)$ in each step of the estimating procedures.

The sample estimating procedure with SIR can be described as follows.

1. Extend the observations $X_i$ to $\phi(X_i)$.

2. Standardization of $\phi(X_i)$: $Z_i = \hat{\Sigma}^{(-1/2)}(\phi(X_i) - \hat{\mu})$, $i = 1, \cdots, n$, where
   \[\hat{\mu} = n^{-1} \sum_{i=1}^{n} \phi(X_i), \quad \hat{\Sigma} = n^{-1} \sum_{i=1}^{n} (\phi(X_i) - \hat{\mu})(\phi(X_i) - \hat{\mu})^T.\]

3. Slicing: Distribute evenly $Z_1, \cdots, Z_n$ into $H$ slices $I_h$, $h = 1, \cdots, H$, according to the order of the respective values of $Y_1, \cdots, Y_n$, each slice has $n_h = n/H$ observations.

4. Calculating slice means: $\hat{m}_h = \sum_{I_h} Z_i / n_h$.

5. Constructing $\hat{\text{SIR}}$: $\hat{\text{SIR}} = \sum_{h=1}^{H} \hat{m}_h \hat{m}_h^T / H$

6. Performing principal component analysis on $\hat{\text{SIR}}$: Assume $d$ is known, let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ be the largest $d$ eigenvalues, and $\hat{\eta}_1, \cdots, \hat{\eta}_d$ be the corresponding normalized eigenvectors.
7. Estimating the dimension reduction directions $\alpha_j$: $\hat{\alpha}_j = \hat{\Sigma}^{-1/2} \hat{\eta}_j$, $j = 1, \cdots, d$.

We can treat the estimated $\alpha$’s just like $\beta$’s as in traditional methods, except we use them with respect to $\phi(X)$’s in further exploration of the model. In the next two sections, we will do simulation studies with constructed data and data analysis with real data.

2.4 Simulations

2.4.1 Checking Asymptotic Normality of Projected Variables

We first check asymptotic normality for one dimensional projected variables with histograms. The procedures are described as following.

1. Generate $n$ observations from a $p$-dimensional standard random vector $X$.

2. Extend $X$’s to $\phi(X)$’s of dimension $r = \frac{p^2 + 3p}{2}$.

3. Let $S_{r-1}$ denote the unit sphere in $\mathbb{R}^r$, and $\text{Unif}(S_{r-1})$ denote the uniform distribution on $S_{r-1}$. To generate a uniform random vector from this unit sphere, we draw a random vector $\alpha$ from standard multivariate normal distribution of dimension $r$, and divided the sample vector by its length to get the uniform unit length random vector, $\alpha^* = \frac{\alpha}{\|\alpha\|}$.

4. Calculate $t_i = \alpha^T \phi(x_i)$, $i = 1, \cdots, n$.

5. Plot the histogram of $t$. 
As argued in the non-published report of Shea (2006), the approximate normality is an asymptotic property with respect to \( p \) for population random vectors. Simulation results are expected to be representative but not exact. Especially, in simulations, behaviors of convergency will also be affected by sample sizes \( n \)’s. On one hand, we generate data sets of different combinations of \( n \) and \( p \) to illustrate the synergetic effects of these two parameters. on the other hand, as we know projections of normal random vectors are still normal variables, we also generated the same combinations of standard normal distributions \( Z \) as reference. That is, by calculating and plotting histograms of \( \alpha^*^T Z \), we can differentiate the behavior of convergence due to simulation (sample size) from dimension. Note the \( p \)’s here are the length of the original predictor \( X \). For the lengths of the standard normal vectors \( Z \), we will use \( r \)’s to make them comparable with \( \phi(X) \)’s.

Since length of feature vector will be functions of the original length of \( X \), we will make special values of \( r \)’s to accommodate this. As the feature vectors here are all degree two polynomial extensions, \( r \) is a function of \( p \), \( r = \frac{p^2 + 3p}{2} \).

1. \( r = 9, n = 100, p = 3 \),

2. \( r = 90, n = 100, p = 12 \)

3. \( r = 90, n = 1000, p = 12 \).

The left column of Figure 2.6 are histograms of projections from standard normal distribution. The right column are histograms of projections from feature vectors. By juxtaposing plots of the same \( n \) and \( p \) combinations together, we can easily judge the convergence behavior of feature vectors with references from normal distributions.

Compare the left panel with the right panel in Figure 2.6, we can see for lower \( p \) (\( p = 9 \)), the projected feature random vector has a very skewed distribution,
while the projected normal vector has pretty good normal distribution; for the combination of \( p = 90 \) and \( n = 100 \), both the two distributions are very flat; for the combination of \( p = 90 \) and \( n = 1000 \), the projected standardized feature distribution is nearly normal.

### 2.4.2 Checking Linear Conditional Mean Condition

As argued in Shea (2006), several problems arise in trying to study the linearity measure

\[
\|E(X|\alpha^T X = t)\|^2 - t^2
\]

by simulation. We will use only scatter plots in our simulations. By plotting elements of the random vector \( X \) against \( \beta^T X \), we can roughly check if \( X \) satisfies the LCM condition. Similarly, we can check if the projections of features approximately satisfy LCM by plotting elements of \( \phi(X) \) versus \( \alpha^{*T}\phi(X) \). Note that, if \( X \) satisfies LCM, when \( E(X) = 0, Var(X) = I \) and \( \|\alpha\| = 1 \), the LCM property could be expressed as \( E(X|\alpha^T X = t) = \alpha t \). As we are testing this property on \( \phi(X) \), we first standardize \( \phi(X) \) to \( \tilde{\phi}(X) \); if \( \phi(X) \) satisfies LCM approximately, then the slopes of scatter plots will be roughly the same values of the corresponding coefficients \( \alpha^j, j = 1, \ldots, r \).

The first several steps of simulation will be the same as for normality. The only difference is for the last step, we draw scatter plots instead of histograms.

Similar with the simulations about asymptotic normality, in Figure 2.7, we compare the results from standard normal and feature vectors. From Figure 2.8, Figure 2.9 and Figure 2.10 we can see the linear conditional mean assumption is not violated.
2.4.3 Global Method on the Feature Space

Our model is a simple yet nonlinear function of a linear combination of $X_i$’s and squared terms of $X_i$’s, note that the function is not symmetric about the origin 0.

Let $p = 6$, $X \sim N(0, I_p)$, $\varepsilon \sim N(0, 0.04)$, and $X \perp \varepsilon$. Consider the regression model.

$$y = 3\sin[z \times (X_1^2 + X_2^2 + X_3^2 + X_1^1 + \frac{2}{3}X_2 + \frac{1}{2}X_3)] + \varepsilon$$
$$= 3\sin(z \times [(X_1 - 0.5)^2 + (X_2 - 1/3)^2 + (X_3 - 0.25)^2]) + \varepsilon$$

By tuning $z$, we can control the complexity of the shape of the $\sin$ function.

Let $z = 0.2$, so we can get just half circle of the $\sin$ function. When we apply dimension reduction directly on $X$, the $S_{Y|X}$ space will have $k = 4$. Now we apply dimension reduction methods on the feature vector, we can write it as a model for $\phi(x)$, which has a $S_{Y|\phi(X)}$ with $d = 1$.

$$y = 3\sin[\frac{1}{5}\alpha^T\phi(X) + c] + \varepsilon$$  \hspace{1cm} (2.31)

where

$$\phi(X) = (X_1, \ldots, X_6, X_1^2, \ldots, X_6^2, X_1X_2, \ldots, X_5X_6)^T$$  \hspace{1cm} (2.32)

and

$$\alpha = (-1, -\frac{2}{3}, -\frac{1}{2}, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, \ldots, 0)^T.$$  \hspace{1cm} (2.33)

The sample is of size 500, I used SIR on both $X$ and $\phi(X)$, by plotting $Y$ versus the first 6 sufficient predictors in each case, we can compare the two performances of the two methods. See Figure 2.11 and Figure 2.12.
Also, we can compare the scatter plot of $Y$ versus the known predictor $\alpha^T \phi(x)$ with the plot of $Y$ versus the estimated predictor (Figure 2.13).

We can also see the performance by comparing the known direction with the estimated first direction in Table 2.1.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\alpha_i$</th>
<th>$\hat{\alpha}_i$</th>
<th>$i$</th>
<th>$\alpha_i$</th>
<th>$\hat{\alpha}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-0.420516621</td>
<td>15</td>
<td>0</td>
<td>0.122093279</td>
</tr>
<tr>
<td>2</td>
<td>-2/3</td>
<td>-0.258851704</td>
<td>16</td>
<td>0</td>
<td>0.089802911</td>
</tr>
<tr>
<td>3</td>
<td>-1/2</td>
<td>-0.219378906</td>
<td>17</td>
<td>0</td>
<td>0.025681546</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>-0.007089734</td>
<td>18</td>
<td>0</td>
<td>-0.081891561</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>-0.060701286</td>
<td>19</td>
<td>0</td>
<td>-0.029461987</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.005060340</td>
<td>20</td>
<td>0</td>
<td>0.013165000</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.479873895</td>
<td>21</td>
<td>0</td>
<td>0.080575318</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.463189491</td>
<td>22</td>
<td>0</td>
<td>0.014382039</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.430851617</td>
<td>23</td>
<td>0</td>
<td>0.122358601</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>-0.035582458</td>
<td>24</td>
<td>0</td>
<td>0.047955026</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>-0.061854947</td>
<td>25</td>
<td>0</td>
<td>-0.007790082</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>-0.007092171</td>
<td>26</td>
<td>0</td>
<td>0.021759409</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>-0.057368399</td>
<td>27</td>
<td>0</td>
<td>0.060198338</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>-0.07660486</td>
<td>28</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1. Comparison of estimated direction vector with know direction vector.

While the scatter plots are produced by applying SIR, we can also use other dimension reduction methods. From experience, we know that results from SIR are better than those from DR and SCR methods.

By tuning $z$, we can change the shape of the $\sin$ curve to get a more complex function of a linear combination of $\phi(X)$. Let $z = 0.4$, we will have a full period of the $\sin$ curve, though the points at the later half circle are sparse. While the first panel cannot recover the whole circle, we can easily detect the $\sin$ trend, and the following panels show no pattern at all.

Again, by comparing this set of plots with the plots from applying SIR directly on $X$, we can see the performance of the new method.
2.5 Application with Baltimore Lead Data

Environmental lead pose serious health problems for young children. The U.S. government banned lead paint in 1978. U.S. oil companies began phasing out leaded gasoline in 1975. Leaded gasoline was fully banned in 1996. But lead from air and other sources finally settled into the soil and remains to be an environmental problem. This data is from the Xlisp-Stat program Arc, Cook and Weisberg (1999). While there is no detailed description of the survey, it seems to be related with the data set used by Mielke et al. (1983). 424 soil samples were randomly collected from locations within an area defined by a $45.24\text{km}$ radius from a designated center point (intersection of Baltimore and Charles Streets) of downtown Baltimore.

The original data set only has four variables. The response variable is the lead concentration (ppm), the predictor variables are the geographic coordinates $(X, Y)$ in kilometers of the sampling location with the above mentioned city center as origin. See Figure 2.17 for the 2D plot of $(X, Y)$, and Figure 2.18 for the 3D plot of the lead level versus the geographic coordinates. The data also includes the distance $D = \sqrt{X^2 + Y^2}$ of the sampling locations from the city center.

It is expected that the lead level is a function of the distances of locations from the city center, which is obvious from the 3D scatter plot. But with classical dimension reduction methods, you can only get at best an direction for a linear combination of $X$ and $Y$, that is no different from rotating the 3D plot to an angle with the most sharp result.

Now we extend the predictor $(X, Y)^T$ by adding degree two polynomials of $X$ and $Y$ to get the feature predictor $\phi(X, Y) = (X, Y, X^2, Y^2, XY)^T$. Figure 2.19 is the scatter plot matrix of the elements of $\phi(X, Y)$. We apply SIR on $\phi(X, Y)$ and draw the scatter plot of lead level versus the first estimated predictor. See
Figure 4.3. Also, we draw scatter plots of lead level versus the second to fourth estimated predictors. See Figure 2.21. We can compare Figure 4.3 with any snapshot in rotating the 3D scatter plot.
Figure 2.6. Histograms for projected random vectors. The left three panels are from standard normal distribution. The right three panels are from standardized feature random vectors derived from lower dimensional standard normal distribution. The first row is of $p = 9, n = 100$; the second row is of $p = 90, n = 100$; the third row is of $p = 90, n = 1000$. 
Figure 2.7. Scatter plots of first dimension of each random vector v.s. projected random vectors. The left three panels are from standard normal distribution. The right three panels are from standardized feature random vectors generated from lower dimensional standard normal distribution. The first row is of $p = 9, n = 100$; the second row is of $p = 90, n = 100$; the third row is of $p = 90, n = 1000$. The data is same as those used in the histogram plots.
Figure 2.8. Scatter plots for $\phi_i(X)$ v.s. $\alpha^T\phi(X)$, $i = 1, 2, 3, 4$, $p = 9$, $n = 100$. 
Figure 2.9. Scatter plots for some randomly drawn $\phi_i(X)$ v.s. $\alpha^T\phi(X)$. $p = 90, n = 100$. 
Figure 2.10. Scatter plots for some randomly drawn $\phi_i(X)$ v.s. $\alpha^T\phi(X)$. $p = 90, n = 1000$. 
Figure 2.11. Scatter plots of $Y$ versus the six estimated predictors with SIR on $X$. 
Figure 2.12. Scatter plots of $Y$ versus the first six estimated predictors with SIR on $\phi(X)$. 
Figure 2.13. The plot on the left is the scatter plot of the $Y$ versus the known predictor $\alpha^T \phi(X)$, the plot on the right is $Y$ versus the estimated predictor $\hat{\alpha}^T \phi(X)$. 
Figure 2.14. Scatter plots of $Y$ versus the six estimated predictors with SIR on $X$ with full circle of the sin function.
Figure 2.15. Scatter plots of $Y$ versus the first six estimated predictors with SIR on $\phi(X)$ with full circle of the sin function.
Figure 2.16. The plot on the left is the scatter plot of the $Y$ versus the known predictor $\alpha^T \phi(X)$, the plot on the right is $Y$ versus the estimated predictor $\hat{\alpha}^T \phi(X)$.

Figure 2.17. 2-D Scatter plots of $X$ v.s. $Y$ in Baltimore Lead Data.
Figure 2.18. 3-D Scatter plots of Baltimore Lead Data.
Figure 2.19. Scatter plots matrix of Feature Vector on Baltimore Lead Data.
Figure 2.20. Lead level versus the first estimated predictor with global SIR on feature space.

Figure 2.21. Lead level versus the second to the fifth estimated predictor with global SIR on feature space.
Local Dimension Reduction Method on Feature Space (L-DRFS)

When we first introduced the idea of Dimension Reduction on Feature Spaces (DRFS), we applied established global dimension reduction methods on feature spaces (G-DRFS). While we got results that could not be available from dimension reduction on original predictors in simulations, we can also see its accuracy is limited by the extent to which $\alpha^T \phi(X)$ is elliptical. We attribute the main reason of the low performance to the violation of the LCM assumption of the design matrix when $p$ is reasonably small. While the support of the feature vector is a manifold embedded in the extended Euclidean space, to successfully implement procedures with DRFS, we need to use methods that relax the condition of LCM. We will consider two approaches, local dimension reduction methods and the newly proposed Central Solution Space methods. In this chapter, we explore the first approach, local dimension reduction methods.
3.1 Local Methods vs. Global Methods

All the global methods starting from SIR share the familial trait of simple elegance in their estimating procedures, which is rooted in the assumption of linear conditional mean. Another desirable feature is the $\sqrt{n}$-consistency of the estimators. Contrary to global methods, local methods require less stringent conditions on the distributions of the predictor. However, a trade off is that they typically involve multidimensional smoothing, which slows down the convergence rate, and they are computationally more complex. As discussed in Tang (2007), the local estimators are $\sqrt{n}$-consistent only under restrictive assumptions.

3.2 Direct Local Dimension Reduction Methods

Most of the local dimension reduction methods are related to non-parametric regression, which simultaneously estimating the link function and the dimension reduction directions based on the estimated link function by minimizing a cost function. Among them, there are Structural Adaptive Estimator (SAE; Hristache, Juditsky, Polzehl, and Spokoiny, 2001), and Minimum Average Variance Estimation (MAVE; Xia et al., 2002). In the dimension reduction approach, there are K-mean clustering (Setodji and Cook, 2004) and Invariant Aggregation by the thesis of Tang (2007).

3.2.1 Nonparametric methods

Almost all the non-parametric methods for estimating the dimension reduction directions are derived from Average Derivative Estimation method (ADE), which is based on the simple fact that the expectation of the gradient of the mean function
is a scalar multiple of the $\beta$’s. For illustration, considering the single-index model 
\[ y = g(\beta_1^T X) + \varepsilon, \]
it is easy to see that the expectation of the gradient $\nabla g(X)$ is a scalar multiple of $\beta_1$.

The Minimum Average Variance Estimation method (MAVE) is inspired by the SIR method, the ADE method and idea of local linear smoothers. Suppose the model could be written as

\[ y = g(B_0^T X) + \varepsilon, \tag{3.1} \]

where $g$ is an unknown smooth link function, and $B_0 = (\beta_1, \ldots, \beta_d)$ with $d < p$ and $E(\varepsilon|X) = 0$ almost surely. By solving the minimization problem

\[ \min_{B} [E\{y - E(y|B^T X)\}^2], \tag{3.2} \]

where the function being minimized is the unconditional expectation of the conditional variance given $B^T X$

\[ \sigma^2_B(B^T X) = E\{[y - E(y|B^T X)]^2|B^T X]\}. \tag{3.3} \]

That is,

\[ E\{y - E(y|B^T X)\}^2 = E\{\sigma^2_B(B^T X)\}. \tag{3.4} \]

The minimization process follows the idea of local linear smoothing estimation. The MAVE method combines nonparametric function estimation with dimension reduction direction estimation. The two components are performed simultaneously. The details of the algorithm involves the selection of multidimensional kernel, op-
timal bandwidth and the initial estimates of $\hat{B}$. However, even under the condition that the dimension is correct, the consistency rate for $\hat{B}$ is $O_p\{h_{opt}^3 \log(n)\}$, though $\sqrt{n}$-consistency could be achieved when the idea is extended to higher order local polynomial smoother. The SAE approach is a similar method that shares disadvantages at consistency rate and computational complexity.

### 3.2.2 Dimension Reduction Methods

Motivated by exploiting local features of the dependence of $Y$ on $X$, Tang (2007) proposed an invariant aggregation methods which require weaker assumptions on the distribution of predictor $X$ than the classical global methods such as SIR and SAVE. The IA method requires the joint distribution of $(X, Y)$ to be symmetric about the central space, and the dimension reduction estimators to satisfy the equivariant assumption (Theorem 5.5.2 in Tang, 2007). The aggregation estimator is semi-unbiased. The definition of semi-unbiasedness is related the concept of unbiased estimator $T(F)$. When $T(F)$ is unbiased, all the eigenvectors corresponding to non-zero eigenvalues are in $S_{Y|X}$. When $T(F)$ is semi-unbiased, only a sub-collection of all the eigenvectors corresponding to non-zero eigenvalues are in $S_{Y|X}$. For selecting the eigenvectors in the semi-unbiased $T(F)$ that span $S_{Y|X}$, she proposed to use either MAVE criterion from Xia et al. (2002) or the Information Index criterion from Yin et al. (2008) to reach $\sqrt{n}$-consistency of the dimension reduction direction estimators.

There is also the K-means Inverse Regression methods by Setodji and Cook (2004) that is extension of SIR to multivariate regressions, which I will not discuss here.
3.2.3 Invariant Aggregation Dimension Reduction on Feature Space (IA-DRFS)

Compared with the nonparametric methods MAVE or SAE, the IA methods are computational simple and have faster convergency rate. Here we use a similar local aggregation method on the feature space. Also for simplicity reason, in my simulation studies, only the first estimated eigenvectors will be studied; and I will not apply the direction selection procedure.

As usual, let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be an independent sample from the random distribution \((X, Y)\). Here is the feature space estimating procedures that combine IA and SIR.

1. Extend the observations \(X_i\) to \(\phi(X_i)\).

2. Standardize \(\phi(X_i)\) to \(Z_i\) and center \(Y_i\) to \(\hat{Y}_i\).

3. For each data point \(Z_i\), let \(S_a\) be the set of the data points in the local ball which centered at \(Z_i\) and have radius \(r\); by applying SIR, we can get an estimator \(\hat{T}_{SIR}(Z_i)\).

4. Take the average of \(\hat{T}_{SIR}(Z_i)\). That is, \(\hat{T}^*_{SIR} = \frac{1}{n} \sum \hat{T}_{SIR}(Z_i)\).

5. Perform PCA on \(\hat{T}^*_{SIR}\), assume \(d\) is known, choose the first \(d\) eigenvectors as the direction estimators.

3.2.4 Simulations

For comparison reasons, I will still use the simulated data structure in the previous chapter. Instead of using global dimension reduction methods on feature space, I will use the simplified IA method.
I used the IA method on $\phi(X)$ with $r = \frac{1}{2} = \frac{1}{2} \sigma_{X_i}$. In Figure 3.1, by plotting $Y$ versus the first four sufficient predictors, we can see that the true predictor is recovered with very sharp image, and the other scatter plots show no pattern. This is in accordance with our model assumption that sufficient information for the regression is retained with the lower projection of predictor, here is the extended feature predictor. In Figure 3.2, we can compare the performance of the IA-DRFS method with those of the G-DRFS and dimension reduction on $X$. It is obvious that the IA-DRFS has sharper image than the others.

From the scatter plot of comparison (Figure 3.2) we can compare the performance of the three estimation methods, as the images of the $\sin$ curves gets sharper in the order of our development. We can conclude the IA-DRFS method performs better then global methods. However, the difference is not so significant. We also tried different combination of radius and slice number, while there are subtle differences, they are not so significant compared with differences among methods.

For the setting of radius of local balls, when we choose the radius big enough, every local ball will cover all the data points, so the IA method will give the same result as the global method when the radius is large enough. By comparison, we can detect a trend of the images that first gets sharper then gets dull as we steadily increase the radius.

### 3.2.5 Application with Baltimore Lead Data

Again, we apply the IA-DRFS method on the Baltimore Lead Data. From the scatter plot of $Y$ versus the first estimated sufficient predictor (Figure 3.3), we can see the relationship is very sharp compared with its counterpart with global method. However, the plots of the last four estimated predictors (Figure 3.4) still
show strong patterns, which could only be attributed to the non-conformity of the Baltimore lead data to our basic dimension reduction model assumption (Equation 2.4).

3.3 Nonlinear Aggregation Dimension Reduction on Feature Space (NA-DRFS)

By comparing the performances of estimation methods on feature spaces previously, from Global Dimension Reduction in the Attribute Space (G-DRAS) to G-DRFS, to IA-DRFS, the estimators are giving sharper images in each step. But, we still want to take into consideration the non-linearity and non-ellipticity of feature spaces. Since most of the time, when we encounter low performances of estimators, we can only attribute them to violation of assumptions under the applied methods. In this chapter, we will propose a method that does not depend on the linearity or ellipticity condition. Also, we add some correction procedure to account for the non-ellipticity of the local predictor space in this method. That is, we are proposing a method that based on local linearizations which is characterized by two linear approximations, one transformation and one correction.

3.3.1 Local Linear Approximation of the Manifold

Since \( \{ \phi(X) : X \in \Omega_X \} \) is a manifold embedded in the extended Euclidean space, \( \phi(X) \) does not have elliptical contoured distributions required by established dimension reduction methods. Our idea is to approximate the manifold locally by tangent spaces at each point of the sample. Suppose that \( E(X) = 0 \) and
\[ \text{var}(X) = I_p. \] Then, for any arbitrary point \( a \in \Omega_X \), near \( x = a \), we have
\[
\phi(x) \approx \phi(a) + \frac{\partial \phi(a)}{\partial x^T}(x - a) = \phi(a) - \dot{\phi}(a)a + \ddot{\phi}(a)x. \tag{3.5}
\]

Suppose the dimension reduction model is,
\[
Y \perp X|\eta^T \phi(X). \tag{3.6}
\]

Then, in a neighborhood of \( a \), we have, approximately,
\[
Y \perp X|\eta^T \dot{\phi}(a)X. \tag{3.7}
\]

In this relation, \( \eta \) is the global characteristic represented by (3.6), whereas \( \dot{\phi}(a) \) represents local linear approximation. Because \( \dot{\phi}(a) \) is known, \( \eta \) can be recovered from local linearizations point by point at different \( a \)'s.

That is, while the dimension reduction relationship goes unchanged from global to local, we can change the non-linearity of the predictor space piece by piece by approximating the manifold with tangent spaces. Consequently, at each point of the data, we want to find the dimension reduction estimator on a degenerated but linear predictor space centered at that point; at each point \( a \) we want to find \( S_{Y|\phi(a),X} \). Though we have different local distribution of the tangent spaces at each data point, we are estimating the same dimension reduction space, as represented by \( \eta \). By putting the information from each piece together, we aim at recovering the global relationship \( Y \perp X|\eta^T \phi(X) \).

Let \( X(a; \rho) \) be the localized \( X \) near \( a \). One way to define it is
\[
X(a; \rho) = XI(\|X - a\| \leq \rho). \tag{3.8}
\]
Let $M(a; \rho)$ be a symmetric positive semi-definite matrix whose columns span $S_{Y|\phi(a)X(a; \rho)}$. Then we target the population object

$$E[M(X, \rho)]$$

(3.9)

However, in this process, each local tangent space defined by $\dot{\phi}(a)X$ has a singular distribution; we cannot apply traditional dimension reductions method on them. We are faced with the question: how to estimate $\eta$?

Note that, by assuming $Y \perp X|\eta^T\phi(X)$, we no longer have any kind of global relationship of the form $Y \perp X|\beta^TX$. Instead, since we have the approximate local relation

$$Y \perp X|\eta^T\dot{\phi}(a)X.$$  

(3.10)

and $\dot{\phi}(a)X$ is a known linear function of $X$, we have an approximate local conditional independence with respect to $X$,

$$Y \perp X(a; \rho)|\beta(a; \rho)^TX(a; \rho).$$  

(3.11)

The notation $\beta(a; \rho)$ is adopted to emphasize the fact that the $\beta$’s here depend on $a$, and as such are different from point to point. With the following Lemmas, we can find the relationship between $S_{Y|\phi(a)X(a; \rho)}$ and $S_{Y|X(a; \rho)}$ of the local approximate relationship, which is estimable. More generally, let $U$ be a generic $p$-dimensional random vector, $A$ be a $r \times p$ matrix with $r \geq p$. We can find the relation between $S_{V|U}$ and $S_{V|AU}$.

**Lemma 3.3.1.** $S_{V|AU} \subseteq \text{Span}(A)$. 
Proof. Let $P_A$ be the projection onto $\text{Span}(A)$ with respect to the usual inner product; that is, $P_\eta = \eta(\eta^T\eta)^{-1}\eta^T$. Let $Q_A = I_r - P_A$. Let $\eta$ be a matrix whose columns span $S_{Y|AU}$. Then, by definition of projection, we have $\eta^T A = \eta^T (P_A A) = (P_A \eta)^T A$. By the definition of central subspace, we have $Y \perp AU | \eta^T AU$. Hence

$$Y \perp AU | (P_A \eta)^T AU,$$

which implies that $\text{Span}(P_A \eta)$ is also a dimension reduction space of $Y|AU$. Hence

$$S_{Y|AU} \subseteq \text{Span}(P_A \eta) = P_A S_{Y|AU} \subseteq \text{Span}(A). \quad (3.13)$$

Lemma 3.3.2. Suppose that $A$ has full column rank. Then

$$(AA^T)^\dagger A S_{Y|U} \subseteq S_{Y|AU} \subseteq A(A^T A)^{-1} S_{Y|U}. \quad (3.14)$$

where $\dagger$ indicates Moore-Penrose inverse of a matrix.

Proof. Let $\eta$ be a matrix such that $\text{Span}(\eta) = S_{Y|AU}$. Then

$$Y \perp AU | \eta^T AU \Rightarrow Y \perp U | \eta^T AU \Rightarrow Y \perp U | (A^T \eta)^T U \quad (3.15)$$

where the first implication holds because $A$ has full column rank. It follows that $\text{Span}(A^T \eta)$ is a dimension reduction space for $Y|U$. Hence

$$S_{Y|U} \subseteq A^T S_{Y|AU} \Rightarrow (AA^T)^\dagger A S_{Y|U} \subseteq (AA^T)^\dagger (AA^T) S_{Y|AU}$$

$$\Rightarrow (AA^T)^\dagger A S_{Y|U} \subseteq P_A S_{Y|AU}$$
\[
(AA^T)^\dagger AS_{Y|U} \subseteq S_{Y|AU},
\]

where the last implication follows from Lemma 3.3.1. This proves the first inclusion in Equation 3.14.

Next, let \( \beta \) be a matrix such that \( \text{Span}(\beta) = S_{Y|U} \). Because \( A \) has full column rank and \( r \geq p \), \( A^T A \) is nonsingular. Then

\[
Y \perp U | \beta^T U \Rightarrow Y \perp AU | \beta^T (A^T A)^{-1} A^T AU
\Rightarrow Y \perp AU | (A(A^T A)^{-1} \beta)^T AU,
\]

where the first implication follows from \( (A^T A)^{-1} A^T A = I \). Thus \( \text{Span}(A(A^T A)^{-1} \beta) \) is a dimension reduction space for \( Y|AU \), and consequently

\[
S_{Y|AU} \subseteq \text{Span}(A(A^T A)^{-1} \beta) = A(A^T A)^{-1} \text{Span} \beta = A(A^T A)^{-1} S_{Y|U}.
\]

The second inclusion in Equation (3.14) is proved. \( \square \)

**Corollary 3.3.1.** Suppose that \( A \) has full column rank. Then

\[
(AA^T)^\dagger AS_{Y|U} = S_{Y|AU} = A(A^T A)^{-1} S_{Y|U}. \quad (3.16)
\]

**Proof.** Note that

\[
P_A = (AA^T)^\dagger (AA^T) = A(A^T A)^{-1} A^T
\]

\[
(3.17)
\]
Hence

\[(AA^T)^\dagger AA^T A = A(A^T A)^{-1}A^T A \Rightarrow (AA^T)^\dagger A = A(A^T A)^{-1}, \tag{3.18}\]

Now the corollary follows from 3.14. \(\square\)

### 3.3.2 Nonlinear Aggregation Dimension Reduction on Feature Space (NA-DRFS)

As suggested by the discussion in section 3.3.1, we can first use established dimension reduction methods at each local region of the original predictor \(X\) to get the estimator \(M_0(a; \rho)\). For example, let \(\mu(a; \rho) = EX(a; \rho)\) and \(\Sigma(a; \rho) = \text{var}[X(a; \rho)]\). Let \(Y(a; \rho)\) be the response corresponding to \(X(a; \rho)\). In case of SIR estimator, we have

\[M_0(a; \rho) = \Sigma^{-1}(a; \rho)\text{var}\{E[X(a; \rho) - \mu(a; \rho)|Y(a; \rho)]\} \Sigma^{-1}(a; \rho). \tag{3.19}\]

Then, apply the second equality \(S_{Y|AX} = A(A^T A)^{-1}S_{Y|X}\) from Corollary 3.3.1, we can have the dimension reduction estimator for \(S_{Y|\phi(X)}\),

\[M(a; \rho) = \phi(a)[\phi(a)^T \phi(a)]^{-1}M_0(a; \rho)[\phi(a)^T \phi(a)]^{-1} \phi^T(a). \tag{3.20}\]

By aggregating local information,

\[E[M(a; \rho)] \tag{3.21}\]

we can obtain the estimator for \(S_{Y|\phi(X)}\).
If there is anything unsolved, it is in the step of estimating $\beta(a; \rho)$. We use SIR directly on the local ball of $B(a; \rho)$ without reasoning for the underlying LCM assumption. As of now, we may think of each local distribution of $B(a; \rho)$ as approximately uniform, which satisfies the LCM condition. We will discuss a possible way to solve this problem at the end of this chapter.

### 3.3.3 Estimation Procedure

The sample estimation procedure using the nonlinear aggregation method could be described as follows.

1. For each data point $x_i$, $i = 1, \ldots, n$, construct a local ball $B(x_i; \rho)$ with $x_i$ as the center and radius being $\rho$. The ball consists of points $x_j$ such that for $j = 1, \ldots, n$,

$$\|x_j - x_i\| \leq \rho.$$

2. Perform dimension reduction methods on the local ball $B(x_i; \rho)$. Here we use SIR as an example.

3. Compute the sample mean and variance matrix of the points in $B(X_i; \rho)$. With the ball mean $\mu(x_i; \rho)$ and ball covariance matrix $\Sigma(x_i; \rho)$, we can standardized the points in the ball, denote the standardized predictor as $z_j(i)$’s accordingly.

4. Denote the $y$’s corresponding to the ball $B(x_i; \rho)$ as $y_j(i)$. Do the partition on $y_j(i)$’s and the slicing upon $z_j(i)$’s. Compute the slice means and denote the means as $\mu^k_{z_i}$, where $k = 1, \ldots, h$. Here, $h$ is the number of slice in each ball.
5. Compute the estimator of the local parameter $\beta(x_i; \rho)$, which has the approximate local relation $Y \perp \perp X(a; \rho) | \beta(a; \rho)^T X(a; \rho)$,

$$M_0(x_i; \rho) = \Sigma_{x_i; \rho}^{-\frac{1}{2}} \left\{ \sum_{k=1}^{h} \mu_k^{x_i} \mu_k^{T} \right\} \Sigma_{x_i; \rho}^{-\frac{1}{2}}.$$

6. Compute $\dot{\phi}(x_i)$, then the local dimension reduction estimator for the tangent space is

$$M(x_i; \rho) = \dot{\phi}(x_i)[\dot{\phi}(x_i)^T \dot{\phi}(x_i)]^{-1} M_0(x_i; \rho)[\dot{\phi}(x_i)^T \dot{\phi}(x_i)]^{-1} \dot{\phi}^T(x_i).$$

7. Sum up $M(x_i; \rho)$ to get the final estimator $M$.

8. Perform PCA on the $M$ to get $\hat{\alpha}$.

Since the procedure is more complex compared with previous method, the R code I used for this method is attached in Appendix A.

### 3.3.4 Simulations

For comparison, we still use the $\sin$ function model from Chapter 2. Recall that, in that model, $p = 6$, $X \sim N(0, I_p)$, $\varepsilon \sim N(0, 0.04)$, and $X \perp \perp \varepsilon$. The regression model is

$$y = 3\sin[\frac{1}{5}(X_1^2 + X_2^2 + X_3^2 + X_1 + \frac{2}{3}X_2 + \frac{1}{2}X_3) + c] + \varepsilon \quad (3.22)$$

The sample is of size 500. I used the NA-DRFS method with varying $r \in [1.5, 6]$. With rough comparison among the resulted sufficient plots from different combination of $r$ and slice number $h$, I find when $r = 1.7, h = 2$, we get a more sharp plot. By plotting $Y$ versus the first 4 estimated sufficient predictors (Figure 3.5), we can...
see that there is no pattern in the last three plots, which means, our estimator is the sufficient predictor of the the relationship between $Y$ and $X$. By comparing the sufficient plot with plots from other methods (Figure 3.6), we can see the effectiveness of this method compared with the G-DRFS method, and the IA-DRFS method. While all three methods are DRFS, both the NA method and IA method are better than the G method. However, the differences between themselves may resulted from different combination of parameters used in applying SIR. In NA, we used $r = 1.7, h = 2$; where in IA, we used $r = 8, h = 3$. Note that there is a crucial difference between the two sets of parameters, when we apply IA, the radius of the local ball are far larger than the radius used in NA, which I guess is due to the precision of linear approximation for the NA method.

### 3.3.5 Application with the Baltimore Lead Data

Again, we apply the NA-DRFS method on the Baltimore Lead Data. From the scatter plot of $Y$ versus the first estimated sufficient predictor (Figure 3.7), we can see the relationship is very sharp compared with its counterparts from both G-DRFS and IA-DRFS methods. However, the plots of the following four estimators (Figure 3.8) still show strong pattern, which we could only explain with the non-conformity of the Baltimore lead data to our basic dimension reduction conditional independence model (Equation 2.4).

### 3.4 Corrected Local Estimator for SIR

The performance of the NA-DRFS method depends on many elements in the whole procedure. Besides the radius of local balls and possibly slice numbers in each ball, there is a very crucial one: the correct finding of the $\beta(a; \rho)$’s of the approximate
local relationship between $Y$ and $X$. For the local estimations, we need to tackle the violation of the basic assumptions of classical dimension reduction methods, that is, local balls are not likely to be elliptical contoured distributed as the whole population usually does. In this section, we propose a method to correctly estimate $\beta(a; \rho)$.

The reason we can have very simple and nice SIR estimator is because of the linear conditional mean property of the distribution of $X$. When this requirement is not satisfied, we have to construct a totally different estimator from SIR estimator.

### 3.4.1 Corrected SIR Estimation Method with Approximate Linear Local Density

With the above suggestion, as we are trying to find $\eta$ being the direction parameter for relation between $Y$ and $\phi(X)$, we met problem of non-linear feature space, not to mention elliptical contoured density. We first detour to approximate the manifold of feature space by local tangent spaces. We then estimate the local dimension reduction directions on tangent spaces by transformation from approximate dimension reduction directions on local original predictors. Then, we aggregate the local findings to get the global estimator. But we can see, the non-elliptical problem is always lingering around, even for the local estimation of SIR on the original predictor space.

In a localized ball, the density distribution is unlikely to be elliptical as required by most dimension reduction methods. The related properties and results fail to work. Among them, the sliced inverse means no longer fall on a curve in the central subspace. However, there may be a chance that the biased slice means could be corrected with knowledge of the local distribution of $X(a; \rho)$. 
Here, to simplify calculation, we can approximate the local predictor distribution by linear densities. With further precision requirement, we may need quadratic density and more. But here our deduction is based on linear density.

We can first find the locus of the sliced means in the local ball. As the machinery of SIR is based on the linear condition mean property of elliptical distribution, we don’t have that established correction transformation as in SIR. However, with knowledge of the linear density, we still can estimate the dimension reduction direction.

The basic idea in linear density case is, for given point of slice mean, with knowledge of the density, we can integrate along slices go through this point and in all possible directions. We claim that only in perpendicular to the true e.d.r. direction that the sliced expectation will coincide with the slice mean. In application, we can just find the estimate of true direction by minimize the distance between slice means and estimated means.

The detail of the method could be illustrated by an example of $X$ with dimension 2:

1. Decide the level of feature vectors, which means, what level of polynomials to use, second degree or third or more. For each point in the original sample $a$, derive the tangent space by calculating the transformation matrix $A(a) = \dot{\phi}(a)$.

2. Decide the radius of approximation $\rho$. For this parameter, we may use a measure of fit of the tangent space to the manifold to help decision making.

3. With known radius and for the fixed point $a$, from now on, we are talking on the basis of a local ball of radius $\rho$ centered around $a$. We estimate linear local density by MLE. For example, when $X$ has dimension 2, the density
function will be of the form

\[ \frac{1}{\pi \rho^2} + s(\cos(\theta)x_1 + \sin(\theta)x_2). \]  

(3.23)

We use the centered \( X \), which is \( \bar{X} = X - a \). Two unknown parameters, the slope \( s \) and the angle \( \theta \) will be estimated by MLE in this case.

4. Find slice means on the literal basis, that is, without standardization. In the local ball of, cutting \( \bar{X} \) into \( h \) slices according to the order of \( Y \), calculate the slice means. Upon each slice mean \( \mu_i \), we can build a slice going through this point. with a normal direction. With estimated density function, slice expectation depend will be a function of the normal direction of the slice. By minimizing the distance between the this theoretical mean and the slice mean by SIR, we can find the estimated \( \hat{\beta}_i \).

5. Summarize findings by weighted aggregation of the information. The reason for aggregation with weight is the estimator will be greatly influenced by the distance of the slice from the mean of the ball. For example, in the extreme case, when the sample mean for one slice is coincide with the ball mean, then, integration on every direction with give you 0 distance between the integrated mean and the \( \mu_i \); that is because, the integrated mean will be the ball mean here, without regarding of direction. So, this process will give us a very big error variance of the estimator. To lower the impact of estimators like this, we need to give small weights. We suggest to use the distance between the ball mean and the slices as weight.
6. Weighted Aggregation is performed by

\[ M_0(a, \rho) = w_i \hat{\beta}_1 \hat{\beta}_1^T + \ldots + w_h \hat{\beta}_h \hat{\beta}_h^T \]  \hfill (3.24)

7. Upon each point \( a \) in the sample, we will have

\[ M(a; \rho) = \hat{\phi}(a)[\hat{\phi}(a)^T \hat{\phi}(a)]^{-1} M_0(a; \rho)[\hat{\phi}(a)^T \hat{\phi}(a)]^{-1} \hat{\phi}^T(a) \]  \hfill (3.25)

Summarize the results by finding expectation on \( M(a; \rho) \), and then perform PCA to find \( \hat{\eta} \).

### 3.4.2 Simulations

Restricted by our estimation algorithm, models in simulation are limited to be on original predictors of dimension \( p = 2 \). Our model is a simple yet nonlinear function of linear combination of squared terms of \( X_i \)'s.

Let \( p = 2 \), \( X \sim N(0, 9I_2) \), \( \epsilon \sim N(0, 1) \), and \( X \perp \epsilon \). The sample size is 1000. We make local balls of diameter \( 1/3 \) of the variance of the predictor population.

Consider the regression model,

\[ Y = 0.6X_1^2 + 0.8X_2^2 + \epsilon \]

The second degree feature vector is

\[ \phi(X)^T = (X_1, X_2, X_1^2, X_2^2, X_1X_2) \]  \hfill (3.26)

While the true dimension reduction direction in the feature space being \((0, 0, 0.6, 0.8, 0)\), our PCA result for the targeted covariance matrix is:
> eigen(m220)

$values [1] 6.905774 3.489692 2.913624 1.860747 1.283536

$vectors

[1,] 0.01808315  0.03038965  -0.02464279  0.05479836  0.997566710
[2,] 0.03095195  0.07862843  0.17165334 -0.97997801  0.055116118
[3,] 0.55485986 -0.01383263 -0.82148469 -0.12876208 -0.022856590
[4,] 0.83099178 -0.01505529  0.54298699  0.11971131 -0.007767593
[5,] 0.01726784  0.99623091 -0.01599671  0.07569531 -0.035215237

From this we can see the first vector is very close to the true vector. But the result is not good as we expected. This section is just listed here for completeness of the whole project.
Figure 3.1. Scatter plots of $Y$ versus the first four estimated predictors with IA-DRFS.
Figure 3.2. The top left panel is $Y$ versus the know predictor; the top right panel is $Y$ versus the first estimated predictor from the global SIR on $X$; the bottom left panel is $Y$ versus the first estimated predictor from G-DRFS; the bottom right panel is $Y$ versus the first estimated predictor from IA-DRFS.
Figure 3.3. Scatter plot of lead level versus the first estimated predictors with IA-DRFS.

Figure 3.4. Scatter plots of lead level versus the last four estimated predictors with IA-DRFS.
Figure 3.5. Scatter plots of $Y$ versus the first four estimated predictors of NA-DRFS method.
Figure 3.6. The top left panel is $Y$ versus the known predictor; the top right panel is the resulted plot from the G-DRFS, the bottom left panel is the resulted plot from the IA-DRFS, the bottom right panel is the resulted plot from the NA-DRFS.
Figure 3.7. Scatter plot of lead level versus the first estimated predictors from the NA-DRFS method.

Figure 3.8. Scatter plots of lead level versus the second to the fifth estimated predictors from the NA-DRFS method.
Chapter 4

Nonlinear Feature Selection by the Central Solution Space Method (CSS-DRFS)

In this chapter I will describe a different approach to nonlinear dimension reduction. This part of my dissertation is taken from an ongoing project by Dr. Bing Li, Yuexiao Dong and myself. My contribution to this part is secondary, and more detailed description of this method will be given in Yuexiao Dong’s dissertation and a manuscript by Li, Dong and Wang (2008).

4.1 The idea

Let $X$ be a $p$-dimensional vector and $Y$ is a random variable. Let $\phi$ be a feature function. That is, $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^r$. Typically, $r$ is larger than $p$, and $\phi(x)$ contains
nonlinear functions of $x$. Our goal is to find $\beta \in \mathbb{R}^r$ such that

$$Y \perp X|\beta^T \phi(X).$$ (4.1)

We will use the idea of Central Solution Space (CSS) developed in Li and Dong (2007) for linear dimension reduction.

Note that under (4.1),

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

We define the column space of $\beta$ that satisfies

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

to be a solution space for dimension reduction (4.1), and the intersection of all such spaces the central solution space.

We construct the following objective function

$$L(\eta) = E\|E(X|Y) - E[E(X|\eta^T \phi(X))|Y]\|^2$$

and minimize $\eta$ to find $\beta$.

Now we assume that $E(X|\beta^T \phi(X))$ can be expected in basis functions. Let $g_1, \ldots, g_k$ be functions from $\mathbb{R}^d \rightarrow \mathbb{R}$, where $d$ is the dimension of the CSS. We assume that $E(X|\beta^T \phi(X) = u)$ lies in the space spanned by the functions

$$g_1(u), \ldots, g_k(u).$$
Then $E(X_i|\beta^T\phi(X) = u)$ can be represented as

$$E(X_i|\beta^T\phi(X)) = \langle X_i, g_1(\beta^T\phi) \rangle, \ldots, \langle X_i, g_k(\beta^T\phi) \rangle$$

$$= \begin{bmatrix} \langle g_1(\beta^T\phi), g_1(\beta^T\phi) \rangle & \ldots & \langle g_1(\beta^T\phi), g_k(\beta^T\phi) \rangle \\ \vdots & \ddots & \vdots \\ \langle g_k(\beta^T\phi), g_1(\beta^T\phi) \rangle & \ldots & \langle g_k(\beta^T\phi), g_k(\beta^T\phi) \rangle \end{bmatrix}^{-1} \begin{bmatrix} g_1(\beta^T\phi) \\ \vdots \\ g_k(\beta^T\phi) \end{bmatrix}.$$ 

Write $G$ as

$$G = \begin{bmatrix} g_1(\beta^T\phi) \\ \vdots \\ g_k(\beta^T\phi) \end{bmatrix}.$$ 

Then

$$E(X|U) = E(XG^T)[E(GG^T)]^{-1}G.$$ 

So our objective function (at the population level) becomes

$$L(\eta) = E\|E(X|Y) - E[E(XG^T)(E(GG^T))^{-1}G|Y]\|^2.$$ 

The conditional expectation can be estimated using the techniques similar to KIR, PIR, SIR as in Li and Dong (2007).
Figure 4.1. Scatter plots of $Y$ versus estimated predictor from CSS-DRFS overlayed with the known predictor.

### 4.2 Simulation

For example, we may include interaction in dimension reduction:

$$\phi(x_1, \ldots, x_p) = (x_1, \ldots, x_p, x_1x_2, \ldots, x_{p-1}x_p)^T.$$  

We can use polynomials; trigonometric functions; and so on, for the feature function $\phi$. For comparison, we still use the sin function example used in previous simulations. Figure 4.1 is the sufficient plot produced by Yuexiao Dong with CSS method. The green image correspond to estimated sufficient plot while the red is the known.
We apply CSS method on the Baltimore lead data. Figure 4.2 is the estimated sufficient plot. This plot is produced by Yuexiao Dong.

For comparison, we also list here the sufficient plots produced by global method (Figure 4.3), IA method (Figure 4.4), and NA method (Figure 4.5).

While this is a real data set that may not conform to our assumption, we can compare the performance of the methods with the sharpness of the sufficient plots. Comparatively speaking, NA and CSS both give very sharp images.
**Figure 4.3.** Lead level versus the first estimated predictor with G-DRFS method.

**Figure 4.4.** Lead level versus the first estimated predictor with IA-DRFS method.
Figure 4.5. Lead level versus the first estimated predictor with NA-DRFS method.
# R Code for Generating Features

```
library(MASS)
library(dr)
library(scatterplot3d)

# Construct 2nd degree feature vector
phi2<-function(x){
  c<-2 # 2nd degree polynomial
  n<-nrow(x)
  s<-matrix(0,n,1000)
  p<-ncol(x)
  s[,1:p]<-x
  s[, (p+1):(2*p)]<-x^2
  count<-p*c+1
  for (i in 1:(p-1)){
    for (j in (i+1):p){
      s[,count]<-x[i]*x[j]
  }
}```
count<-count+1}
length<-count-1
phi2<-s[,1:length]
return(phi2)

} # Construct 3rd degree feature vector
phi3<-function(x){
c<-3 #3nd degree polynomial
n<-nrow(x)
s<-matrix(0,n,1000)
p<-ncol(x)
s[,1:p]<-x # First degree terms
s[,,(p+1):(2*p)]<-x^2 # Second degree terms
count<-p*2+1
for (i in 1:(p-1)){
  for (j in (i+1):p){
    s[,count]<-x[i]*x[j]
    count<-count+1}
}s[,count:(count+(p-1))]<-x^3
count<-count+p
for (i in 1:p){
  s[,,(count+(i-1)*(p-1)):(count+(p-2)+(i-1)*(p-1))]<-x[i]^2*x[-i]
  if (p==2) {
    length=count+1
  } else {
    count<-count+p*(p-1)
    for (i in 1:(p-2)){
      for (j in (i+1):(p-1)){
        for (k in (j+1):p){
          s[,count]<-x[i]*x[j]*x[k]
          count<-count+1}}
    length<-count-1}
  }
phi3<-s[,1:length]
return(phi3)
}

# Construct the 2nd degree linearized feature vector
phidot2<-function(x){
p<-length(x)
I<-diag(p)

a<-diag(2*x)
phidot<-rbind(I,a)
for (i in 1:(p-1)){
    phidot<-rbind(phidot,cbind(matrix(0,(p-i),(i-1)),
        x[(i+1):p],diag(x[i],(p-i))))
}
return (phidot)
}

# Construct the 3rd degree linearized feature vector

phidot3<-function(x){
    p<-length(x)
    phidot<-rbind(phidot2(x),diag(3*x^2))
    for (i in 1:p){
        phidot1<-matrix(0,(p-1),p)
        phidot1[,i]<-2*x[i]*x[-i]
        phidot1[-i]<-diag(x[i]^2,(p-1))
        phidot<-rbind(phidot, phidot1)
    }
    if (p>2){
        phidot2<-matrix(0,p*(p-1)*(p-2)/6,p)
        count<-1
        for (i in 1:(p-2)){
            for (j in (i+1):(p-1)){
                for (k in (j+1):p){
                    phidot2[count,i]<-x[j]*x[k]
                    phidot2[count,j]<-x[i]*x[k]
                    phidot2[count,k]<-x[j]*x[i]
                    count<-count+1}
            }
        }
        phidot<-rbind(phidot,phidot2)
        return(phidot)
    }
R Code for Nonlinear Aggregation Method

#########################
##
## Load Packages
##
##
#########################
library(MASS)
library(dr)
library(scatterplot3d)

##########################################
##
## Dimension Reduction Procedures
##
##
##########################################

# Standardization

stand<-function(x){
n<-nrow(x)
p<-ncol(x)
xb <- apply(x, 2, mean)
xb <- t(matrix(xb, p, n))
x1 <- x - xb
sigma <- t(x1) %*% (x1)/n
eva <- eigen(sigma)$values
eve <- eigen(sigma)$vectors
sigmamrt <- eve %*% diag(1/sqrt(eva)) %*% t(eve)
z <- sigmamrt %*% t(x1)
return(t(z))
}

# Sample Variance

svar <- function(x) {
  xm <- apply(x, 2, mean)
  n <- nrow(x)
  svar1 <- (t(x) - xm) %*% t(t(x) - xm) / n
  return(svar1)
}

# Matrix Inverse

nrt <- function(a) {
  evec <- eigen(a)$vectors
  eval <- eigen(a)$values
  b <- evec %*% diag(1/eval) %*% t(evec)
  return(b)
}

# Negative square root of a matrix

nsrt <- function(a) {
  evec <- eigen(a)$vectors
  eval <- eigen(a)$values
  b <- evec %*% diag(1/sqrt(eval)) %*% t(evec)
  return(b)
}

# M0

M0sir <- function(u, y, h) {
  # h is the number of slices
  x <- stand(u)
  p <- ncol(x)
  n <- nrow(x)
  s <- matrix(0, n, h, p)
  ce <- matrix(0, h, p)
for(i in 1:h){
    index<-order(y)[((i-1)*(n/h)+1):(i*(n/h))]
    s<-x[index,
    ce[i,]<-c(apply(s,2,mean)) }
M0sir<-nsrt(svar(u))%*%t(ce)%*%nsrt(svar(u))
return(M0sir)
}

# local feature of sir

lofes<-function (x,y,rho,h){
    px<-ncol(x)
    pphi<-(px^2+3*px)/2
    n<-nrow(x)
    sumc<-mat.or.vec(pphi,pphi)
    for (i in 1:n){
        a<-x[i,]
        d<-t(t(x)-a)
        distn<-sqrt(diag(d%*%t(d)))
        if (sort(distn)[h*px]>rho) next
        ball<-x[distn<=rho,]
        yb<-y[distn<=rho]
        nb<-nrow(ball)
        m0<-M0sir(ball,yb,h)
        m<-phidot2(a)%*%nrt(t(phidot2(a))%*%phidot2(a))
        %*%m0%*%nrt(t(phidot2(a))%*%phidot2(a))%*%t(phidot2(a))
        sumc<-sumc+m
    }
    lofes<-eigen(sumc)$vectors
    return(lofes)
}
References


Multivariate Analysis, 20, 272-276.


Li, B. (2003, June). Dimension reduction and regression analysis. (Lecture Notes)


Li, K.-C., and Duan, N. (1989, September). Regression analysis under link viola-


Vita

Yu Wang

Yu Wang was born in July 1974 in Heilongjiang Provence, China.

She got her Bachelor degree in International Business from Economics and Management School, Beijing University of Technology, Beijing, China in July, 1997.

She started her graduate studies in Pennsylvania State University since August, 2002. She is scheduled to earn her doctoral degree in August, 2008.