The Pennsylvania State University
The Graduate School

EFFICIENT ESTIMATION AND ORDER DETERMINATION FOR SUFFICIENT DIMENSION REDUCTION

A Dissertation in
Statistics,
by
Wei Luo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2014

The dissertation of Wei Luo was reviewed and approved* by the following:

Bing Li
Professor of Statistics
Dissertation Advisor, Chair of Committee

Naomi S. Altman
Professor of Statistics

Runze Li
Distinguished Professor of Statistics and
Professor of Public Health Sciences

Jesse Louis Barlow
Professor of Computer Science and Engineering

Aleksandra Slavkovic
Associate Professor of Statistics and
Associate Head for Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

Sufficient dimension reduction (SDR) has driven intense interest in the recent decades as a solution to deal with high-dimensional data. The goal of SDR is to construct, usually by a linear transformation of the original predictor, a lower-dimensional sufficient statistic that serves as the new predictor in subsequent modeling. An important problem in SDR, is to determine the reduced dimension of the new predictor. In this dissertation, we first propose two order-determination methods that are applicable to a large class of SDR methods, with both of them proved consistent and shown efficient via simulation study and real data examples.

Another part of the dissertation focuses on the development of a new class of efficient estimators of the linear transformation under various SDR assumptions, in a unifying semi-parametric approach. These estimators are expected to outperform their competitors in the literature, which were developed without consideration of semi-parametric efficiency. We derive the efficient score functions that generate these estimators, together with a computationally efficient algorithm. We also conduct the corresponding simulation studies and real data analysis to further show the effectiveness of the estimators in application.

Key Words and Phrases: dimension reduction; order-determination; bootstrap; augmentation; semi-parametric; efficient estimation.

Table of Contents

List of Figures **vi**

List of Tables **vii**

Acknowledgments **viii**

Chapter 1
Introduction **1**
1.1 General introduction 1
1.2 Contributions of the dissertation 3

Chapter 2
Literature Review **6**
2.1 General Introduction to SDR 6
2.2 Inverse regression SDR methods 9
2.3 MAVE-type SDR methods 16
2.4 Other types of SDR methods 21
2.5 Order - determination methods 23

Chapter 3
Order determination for dimension reduction using an alternating pattern of spectral variability **26**
3.1 Introduction 26
3.2 The ladle plot 29
3.3 Theoretical analysis of alternating spectral variation 33
3.4 Consistency of CSE 39
3.5 Comparison with other methods via simulation studies 41

3.6 A real-data example 44
3.7 Discussion 46
3.8 Proofs of Lemmas and Corollaries 47

Chapter 4
Order-determination for dimension reduction using augmentation predictors **51**
4.1 Introduction 51
4.2 Consistency of ASE in a special case 53
4.3 Consistency of ASE under the general setting 59
4.4 Simulation study 64
4.5 Discussion 67

Chapter 5
On efficient dimension reduction with respect to a statistical functional of interest **69**
5.1 Introduction 69
5.2 Dimension reduction for conditional statistical functional 73
5.3 Formulation of the semiparametric problem 75
5.4 Efficient score and efficient information 77
5.5 Linear statistical functionals 81
5.6 Composite linear statistical functionals 83
5.7 Implicit statistical functionals 86
5.8 Effect of estimating the central subspace 90
5.9 Estimation 91
5.10 Simulation comparisons 96
5.11 Application: age of abalones 101
5.12 Discussion 103
5.13 Appendix 104

Bibliography **115**

List of Figures

3.1 The benefit of combining eigenvalues and variability of eigenvectors: the upper-left panel is the plot of bootstrap variability of eigenvectors; the upper-right panel is the scree plot; the lower panel is the ladle plot based on a combination of eigenvalues and eigenvector variability. 28

3.2 The upper-left panel shows the bootstrap eigenvector variability; the upper-right panel is the scree plot of sample eigenvalues; the lower-left panel is the ladle plot; the lower-right panel is the scatter plot of the reduced predictor indexed by cultivars. 45

4.1 The trend of augmentation components in the sample eigenvectors 57

5.1 Comparison of SEE with other estimators of CMS and CVS for the abalone data. In the upper panels, the x-axes are the predictors obtained by RMAVE, Zhu-Dong-Li, and SEE estimates for the CMS; the y-axes are the abalones' ages. In the lower panels, the x-axes are the predictors derived from Zhu-Zhu, Zhu-Dong-Li, and SEE estimates of the CVS; the y-axes are the estimated absolute residuals. 102

List of Tables

3.1 Comparison of order-determination methods 43

4.1 Comparison of order-determination methods 66

5.1 Using MAVE to estimate quantities (a) through (h) in efficient score 94
5.2 Comparison of SEE with other estimators for three statistical functionals 99
5.3 Comparison of SEE with other estimators with correlated predictors 100
5.4 Comparison of SEE with AQE at $\tau = 0.75$ 101
5.5 Bootstrap error of the estimators 103

Acknowledgments

I am heartily grateful to my dissertation adviser, Dr. Bing Li, for all his valuable guidance, inspirational encouragement and all the time he spent on me in my academic research. Without his supervision, suggestions and support, it would be impossible for me to make progress in the academic research area and complete this dissertation. There is so much I have learned and much more I should learn from him, within and beyond statistics. I feel myself very lucky to be one of his students.

Also, I wish to express my propound appreciation to Dr. Naomi S. Altman and Dr. Runze Li. They have played extremely important roles in my academic career, not only in statistical learning and research, but also offering me the opportunity to be one of the students in this wonderful department, let alone all the valuable tips and suggestions to my career plan.

I also want to thank Dr. Jesse Louis Barlow for his time and suggestions that significantly improve this dissertation.

Besides, I sincerely want to say thank you to all my friends. Thank you for all your support and happiness brought to me.

Finally, I want to thank my family members - my parents, my wife Xizhen, my aunts and uncles, my cousins, my adorable grandmother and my beloved grandfather. It is because of them that I can have an opportunity to lead this good life and have the courage to look forward to all the challenges and opportunities in the unknown future. I will always love them.

Chapter 1

Introduction

1.1 General introduction

In the recent decades, high-dimensional data have appeared in bioinformatics, image processing, data mining, and many other fields. Typically, when a predictor X is collected to model a response variable Y, people tend to include a large number of variables in X to avoid potential model bias. In contrast, very often only a limited sample size is available. In other words, if we denote the dimension of X as p and the sample size as n, then p is large relative to n. This situation brings much trouble in data analysis, because in most traditional methods, p needs to be small relative to n to ensure enough degrees of freedom in estimating parameters. Otherwise the result will not be trustworthy. Consequently, efforts have been taken to find a new predictor with lower dimension, and this new predictor must be a sufficient statistic so that there is no loss of information in subsequent modeling.

One common way to achieve this goal is sufficient dimension reduction (SDR), in which the new predictor is generated by a lower-dimensional linear transformation of the original predictor X. If we denote the matrix of linear coefficients as β, then the new predictor is $\beta^{\intercal}X$. Note that there is an identifiability problem about β (for more details please see Chapter 2). In Cook (1998), this problem has been solved by assuming that there exists unique $d < p$ such that a β of dimension $p \times d$ exists, and the space spanned by the columns of β is unique. This space is then called the central subspace.

Very often in practice, data analysis focuses only on a specific feature of the

conditional distribution of Y given X. For example, in regression, only the conditional mean $E(Y|X)$ is of interest. In this case, a more efficient method for dimension reduction is to find β such that $\beta^{\mathsf{T}}X$ only preserves full information about $E(Y|X)$. To address the corresponding identifiability problem, the central mean subspace has been defined similarly to the central subspace. These two spaces are often called the central dimension reduction subspace if no further information is provided.

In the literature, a variety of SDR methods have been proposed to estimate the central dimension reduction subspace, such as sliced inverse regression (SIR) by Li (1991), sliced average variance estimation (SAVE) by Cook and Weisburg (1991), contour regression by Li, Zha and Chiaromonte (2005) and directional regression by Li and Wang (2007) that estimate the central subspace, and ordinary least square estimator (OLS) by Li and Duan (1989), principle Hessian directions (pHd) by Li (1992), iterative Hessian transformation (iHt) by Cook and Li (2002), minimum average variance estimation (MAVE) by Xia, Tong, Li and Zhu (2002), and groupwise dimension reduction by Li, Li and Zhu (2010) that estimate the central mean subspace. A large portion of these methods form the family of "inverse regression" methods, in which the estimation procedure can be generally characterized as first using moments of $X|Y$ to construct a positive semi-definite matrix parameter called the candidate matrix, and then taking the column space of its sample estimate as the estimate of the central dimension reduction subspace.

In all these SDR methods, the dimension d of the central dimension reduction subspace is assumed to be known as a priori. However, this is not the case in practice. Hence the estimation of d becomes an important and relatively separate problem. In the literature, several order-determination methods have been proposed including sequential tests (Li 1991; Bura and Yang 2009), cross validation (Xia, Tong, Li and Zhu 2002), the bootstrap method (Ye and Weiss 2003), and BIC-type criterion (Zhu, Miao and Peng 2006). Most of these methods are designed for inverse regression SDR methods, in the sense that they are indeed estimating the rank of the candidate matrix, which is also equal to d. In this way the order-determination problem is connected to the rich literature in matrix algebra. In this dissertation, we have proposed two order-determination methods that are also designed for this family of SDR methods.

On the other hand, although various SDR methods have been proposed in the

literature, systematic connections and comparisons between these methods have only been explored within inverse regression methods such as SIR, SAVE and directional regression. In Ma and Zhu (2012a), a semi-parametric family of estimating equations has been proposed for both the central subspace and the central mean subspace. Their approach is unifying, in the sense that all the SDR methods mentioned above can be shown equivalent to solving the corresponding estimating equations in this family (Li and Dong 2009, Dong and Li 2010). Motivated by their work, in this dissertation we have also focused our interest on SDR with respect to a general statistical functional, and we have derived in a unifying approach, the corresponding family of estimating equations, the semiparametrically efficient estimator and a computationally efficient algorithm. Special cases of this functional include conditional mean (Cook and Li 2002), conditional variance (Zhu and Zhu 2009), and conditional median (Kong and Xia 2012).

1.2 Contributions of the dissertation

Three projects are included in this dissertation. The first two are joint work with my dissertation adviser Dr. Bing Li, and the third one is joint work with Dr. Bing Li and Dr. Xiangrong Yin. In these projects, we aim to address two important problems as mentioned above. Namely, we provide novel approaches to estimate the dimension d in the first two projects, and we derive the semiparametric efficient estimator for a variety of SDR problems in the third project. The manuscript of the first project has been under review for *Biometrika* (Luo and Li 2014), and the paper of the third project has been accepted by *Annals of Statistics* (Luo, Li and Yin 2014a, 2014b).

Both of the two order-determination methods in this dissertation are designed for inverse regression SDR methods, and they are based on the same framework that estimates the rank of candidate matrix by the combination of information from both its eigenvectors and its eigenvalues. For this reason, both methods significantly outperform their competitors in the literature. Depending on different ways of extracting information, these two methods have their own pros and cons when compared to each other.

Our first order-determination method extracts information from eigenvectors using the bootstrap method in Ye and Weiss (2003). Their method has been popular

in the literature, however, it is heuristically defined and its asymptotic behavior has never been rigorously explored. In our work we have first proved its asymptotic properties. Then by combining it with the level of eigenvalues in an innovative way, we have constructed a new method and shown its consistency. Compared to the bootstrap method, it is defined in a clear way; compared to the sequential tests, it does not involve case-by-case asymptotic derivation; compared to BIC-type criterion and cross-validation method, it employs an intrinsic penalty function of dimension and does not involve parameter tuning. Simulation study provides clear evidence that our method is more efficient than these competitors in finite sample cases.

Our second order-determination method extracts information from eigenvectors using augmentation random variables. Compared to the variable selection methods that use these variables to generate baseline distribution, this method employs them in a distinct way to detect specific pattern of eigenvectors. Consequently it is consistent under fairly general conditions. On the other hand, compared to the other order-determination methods, which treat the candidate matrix as a general matrix parameter, this method effectively makes use of the SDR assumption and its interaction with the candidate matrix, therefore it substantially outperforms them as shown in simulation study. Compared to our first method, it is computationally more efficient since no bootstrap re-sampling is involved.

In the third project, we have focused our interest on developing efficient dimension reduction methods with respect to statistical functionals. This unifying approach can be applied to special cases including conditional mean and conditional variance which have been discussed in the SDR literature, and conditional quantile which has been recently popular in data analysis but yet been discussed in the SDR literature. As mentioned before, we have derived the general form of the estimating equation family, and the semi-parametric efficient estimator, and an efficient algorithm. We have also discussed the special forms of our approach for linear functionals, composite linear functionals and implicit functionals, which include conditional mean, conditional variance and conditional quantiles as special cases. As a result, we have found that our approach not only explores new research directions when applied to conditional quantiles, but also discovers interesting issues that has not been noticed in the literature when applied to conditional mean. These issues certainly deserve further investigations, and we expect a series

of subsequent research projects in the future.

Chapter 2

Literature Review

2.1 General Introduction to SDR

Since in SDR we seek sufficient statistics, these methods are expected to be supervised learning procedures. However, before systematic SDR methods were proposed, unsupervised dimension reduction had been used for a long time. One of the earliest types can be dated back to year 1901 when Karl Pearson discussed Principal Component Analysis (PCA). PCA computes the covariance matrix of the predictor X and uses its eigenvectors with significant eigenvalues, to construct the new predictor using the linear combinations of X induced by these eigenvectors. Very often in practice, X has significant variation only in certain directions, so that the number of significant eigenvalues in the covariance matrix - the dimension of the new predictor - is less than p. In this way the dimension is reduced, which makes subsequent data analysis available.

The method is very handy in computation, though it has several drawbacks, among which the most crucial one is its nature of being unsupervised. This can cause failures of PCA in that a linear combination of X that contains unique information about Y could be dropped due to its relatively sample variation, leading significant loss of information in the modified data. This drawback can not be avoided unless Y is considered in the procedure, that is, the dimension reduction procedure becomes supervised. In the recent years, PCA has been generalized to become sufficient under specific settings (Cook 2007; Cook and Li 2009). Nevertheless, this generalization is indeed a special case of supervised sufficient dimen-

sion reduction.

Once the dimension reduction procedure becomes supervised, a desired property is that the new predictor, constructed by a linear transformation of X, is a sufficient statistic. Denote the response by Y and the new predictor by $\beta^{\mathsf{T}}X$ in which β is a $p \times d$ matrix with $d \leq p$, then this property is equivalent to:

$$Y \perp\!\!\!\perp X \mid \beta^{\mathsf{T}}X \tag{2.1}$$

in which $\perp\!\!\!\perp$ means independence between two random variables. Theoretically the dimension of Y can be any positive integer. However, in practice Y is mostly univariate. In the rest of the article, we will suppose that Y is univariate unless explicitly specified.

To estimate the sufficient statistic in (2.1), an identifiability problem naturally arises that β in (2.1) is not unique. For example, a trivial case occurs when $d = p$ and β is any invertible matrix. This problem is indeed two-fold: first, for any β satisfying (2.1), any invertible column linear transformation of β also satisfies (2.1), thus an alternative parameter is the column space of β; second, for any column space of β satisfying (2.1), any space that includes it as a subspace also satisfies (2.1), thus we should define d carefully so that we are estimating the unique "smallest" space. This problem has been addressed in Cook (1998), which defines the central subspace as the unique space with the smallest dimension, and proved its existence under certain regularity conditions. Related discussion can be also found in Cook and Li (2002), and Yin, Li and Cook (2008).

In many cases, people are only interested in a specific aspect of the conditional distribution of Y given X, such as $E(Y|X)$ in regression, $Var(Y|X)$ in volatility analysis, and conditional median $M(Y|X)$ in quantile regression (Zou and Yuan 2008; Kong and Xia 2012). Correspondingly, the SDR assumption (2.1) is loosen to

$$E(Y|X) \perp\!\!\!\perp Y \,|\, \beta^{\mathsf{T}}X \tag{2.2}$$

$$\mathrm{Var}(Y|X) \perp\!\!\!\perp Y \,|\, \beta^{\mathsf{T}}X. \tag{2.3}$$

Note that they are equivalent to $E(Y|X) = E(Y|\beta^{\mathsf{T}}X)$ (Cook and Li 2002) and $\mathrm{Var}(Y|X) = E[\mathrm{Var}(Y|X)|\beta^{\mathsf{T}}X]$ (Zhu and Zhu 2009) correspondingly. Similarly as in (2.1), in these two assumptions the identifiability problem arises again,

and has been addressed in Cook and Li (2002) and Zhu and Zhu (2009). The corresponding space is called the central mean subspace and the central variance subspace. Note that the central mean subspace (central variance subspace) will always be a subspace of the central subspace. This is a theoretical foundation of the estimation procedure in our third project presented in Chapter 5.

Hereafter, we will call these spaces the central dimension reduction subspace unless otherwise specified. In the literature, there have been a variety of methods to estimate this space, most of which can be classified into one of the two families. The first family contains all the inverse regression SDR methods. In this family, the general framework is to first construct a function of β using sample inverse moments of X given Y, then estimate the central dimension reduction subspace by the linear span of columns of β that maximizes the function. In most common cases, this function is in the form of $\beta^{\intercal}\hat{M}\beta$, in which $\hat{M}$ is a function of sample inverse moments whose limit M is called the candidate matrix. Then the estimate of the central dimension reduction subspace is spanned by the eigenvectors of $\hat{M}$ corresponding to its nonzero eigenvalues. One advantage of this family of methods, is that the implementation is straightforward and computationally fast. One drawback of this family of methods, is that it typically assumes X to be elliptically contoured distributed, or even narrowed down to be normally distributed. This assumption pushes these method away when X has some other distributions, for example, when X is discrete.

The second family of SDR methods is also based on construction of an objective function that reaches its minimum at β. In contrast to the inverse regression methods, it typically involves conditional moments of $Y|X$ rather than the inverse moments, and it typically employs nonparametric techniques to estimate these moments rather than assuming a parametric model for these moments. Consequently, it doesn't invoke any nontrivial distributional assumption on X. However, the nonparametric methods can potentially induce the curse of dimensionality and computational burden. A representative member of this family is MAVE (Xia, Tong, Li and Zhu, 2002). Therefore we call the methods in this family MAVE-type methods.

There are several other methods in the literature that don't clearly belong to either of the above two families, and we also briefly review some of them.

Besides estimating the central dimension reduction subspace itself, another

problem in SDR is to determine its dimension d. In the literature, several order-determination methods have been proposed and are commonly used. We will discuss them together at the end of this chapter.

Also, since we assume that the sufficient statistic is in the form $\beta^{\mathsf{T}}X$, its sigma field will be invariant under invertible linear transformations X, as long as the column space of β is linearly transformed correspondingly. Therefore, without loss of generality we will always assume X to be standardized with mean 0 and variance equal to identity matrix throughout this chapter.

In the rest of this chapter, we will briefly review the literature of inverse regression methods and MAVE-type methods in the first two sections, followed by discussion of some other SDR methods in the third section. Then we will discuss the order-determination methods in the last section.

2.2 Inverse regression SDR methods

The first inverse regression method can be dated back to Li and Duan (1989). In their setting, a single index model is assumed, which is equivalent to assumption (2.2) with $d = 1$. In this paper, the ordinary least squares (OLS) estimator in linear regression is shown to be a consistent estimator of β up to a multiplicative scalar, under the condition that

$$E(X|\beta^{\mathsf{T}}X) = P_{\beta}^{\mathsf{T}}X \tag{2.4}$$

in which $P_{\beta} = \beta(\beta^{\mathsf{T}}\beta)^{-1}\beta^{\mathsf{T}}$. Since β is unknown, in practice this condition is strengthened to be

$$E(X|b^{\mathsf{T}}X) = P_{b}^{\mathsf{T}}X \text{ for } \forall\, b \in \mathbb{R}^{p\times k}, k = 1, \ldots, p. \tag{2.5}$$

so that verification of the condition becomes available - indeed it has been shown that this condition is equivalent to X being elliptically distributed, which is a commonly accepted assumption. For this reason, in the literature the strong version (2.5) is more commonly used than the weak version (2.4), although (2.4) can guarantee the validity of OLS. Moreover, the plausibility of (2.4) is also supported by Hall and Li (1993), which shows that under mild regularity conditions, (2.4) approximately holds as p grows large.

The OLS method belongs to inverse regression type because

$$E(XY) = E[E(X|Y)Y].$$

In the equation above, the left-hand side is the OLS estimator, and the right-hand side can be treated as a weighted average of $E(X|Y)$. Cook and Li (2002) further pointed out that OLS only estimates the central mean subspace, for the following straightforward reason:

$$\begin{aligned} E(XY) &= E[XE(Y|X)] = E[XE(Y|\beta^{\mathsf{T}}X)] = E[E(X|\beta^{\mathsf{T}}X)E(Y|\beta^{\mathsf{T}}X)] \\ &= P_\beta^{\mathsf{T}} E[XE(Y|\beta^{\mathsf{T}}X)] = P_\beta^{\mathsf{T}} E(XY). \end{aligned}$$

Note that besides the elliptical distribution on X, we only need assumption (2.2) to make these equations hold.

Following Li and Duan (1989), Duan and Li (1991) discussed another method called slicing regression based on single-index model assumption. Later on, Li (1991) generalized it into Assumption (2.1), and showed that under condition (2.4), the column space of the candidate matrix

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

belongs to the central subspace. The method is called sliced inverse regression (SIR), due to the process of "slicing" Y to estimate $E(X|Y)$ in the sample level. Similar to Li and Duan (1989), condition (2.4) is sufficient, though it is often strengthened to the ellipticity condition (2.5) in practice.

Theoretically, SIR is not exhaustive, that is, it does not guarantee to fully recover the central subspace. In fact, multiple simulation studies and real data applications have shown that it works the best when there is a monotone trend between X and Y, and it loses efficiency when this trend becomes symmetric. However, it is still commonly used in practice. Indeed, it has been widely applied under different names in other research areas. For example, in Meta PCA, if we assign Y to be the indicator of the clusters, then $\mathrm{Var}(X|Y)$ is the covariance matrix of X within the cluster. In this way SIR is the same as a commonly used method in this area. As another example, if we assign Y to be the indicator of treatments, then we can

treat $E[\mathrm{Var}(X|Y)]$ as S_{within} and $\mathrm{Var}[E(X|Y)]$ as $S_{between}$, up to a multiplicative scalar, so that SIR can be connected to one-way ANOVA.

One advantage enjoyed by SIR and other inverse regression methods, is that there is usually an obvious estimate of the inverse moments using sample inverse moments. Namely, in Li (1991) $E(X|Y)$ is estimated by slicing Y and taking the sample mean of X within each slice. However, it can also be estimated using kernel method (Zhu and Fang 1996; Wu, Mukherjee and Liang 2009) or by assuming a parametric model (Bura and Cook 2001). The consistency of SIR is shown in Li (1991) when p is fixed and $n \to \infty$. Later in Zhu, Miao and Peng (2006), its consistency is also proved when p is increasing with n in specific orders.

SIR is the first method explicitly based on inverse regression. After that, a variety of inverse regression methods have been proposed.

Following SIR, Cook and Weisberg (1991) proposed another method called sliced average variance estimation (SAVE) to estimate the central subspace, in which the candidate matrix is

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

under the assumption that

$$\mathrm{Var}(X|\beta^{\intercal}X) = Q_{\beta}^{\intercal}\mathrm{Var}(X)Q_{\beta} \tag{2.6}$$

in which $Q_{\beta} = I_p - P_{\beta}$. Similar to (2.4), since β is unknown, this assumption is strengthened to be

$$\mathrm{Var}(X|b^{\intercal}X) = Q_b^{\intercal}\mathrm{Var}(X)Q_b, \ \ \forall b \in \mathbb{R}^{p\times k}, k = 1, \ldots, p. \tag{2.7}$$

It can be shown that this assumption is equivalent to X having a multivariate normal distribution. In Li and Wang (2007), this assumption has been relaxed back to the elliptical distribution condition (2.5), which still guarantees the validity of SAVE.

Similar to SIR, SAVE can also be connected to PCA. In fact, we can treat this method as a comparison between two covariance matrices of X in each slice: one is the identity matrix which occurs when X and Y are independent, and the other is

the real case in the data. By taking the difference of the two matrices, the directions in X that are marginally independent of Y will be canceled out and what's left is in the central subspace. Note that according to assumption (2.1), the directions in X that are orthogonal to $\beta^{\intercal}X$ are conditionally independent of Y given $\beta^{\intercal}X$, instead of being marginally independent of Y. Referred to the discussions in Yin, Li and Cook (2008), this subtle difference partially explains why normal distribution of X needs to be assumed originally in this method.

Similar to SIR, although SAVE is also a model-free SDR method, it does have its own "preference" on detecting patterns of $Y|X$, and this time it goes the opposite way. That is, when the sample size is limited, SAVE works best when there is a symmetric trend between X and Y, and it loses efficiency when this trend is monotone. Hence in real applications, SIR and SAVE can be used simultaneously to improve the SDR procedure (Cook 1994).

In 1992, Li (1992) proposed another inverse regression method called principle Hessian directions (pHd), which estimates the central mean subspace under condition (2.7). The idea is that since $E(Y|X)$ is a function of $\beta^{\intercal}X$, if we differentiate it twice with respect to X, by the chain rule it's easy to see that pointwise the Hessian matrix has its columns in the column space of β, as does any of its weighted averages. Using a carefully selected weight, it can be shown by Stein's Lemma that this weighted average has the form

$$E\{XX^{\intercal}[Y - E(Y)]\}$$

for X satisfying (2.7). Using the same trick as in OLS, it's straightforward to show that pHd is in the inverse-regression family.

In Li's original paper, a residual version of pHd is also proposed in which Y is replaced by its residual of linear regression on X. These versions are also discussed in Cook (1998) and Cook and Li (2002). For now let's focus on the original version with the candidate matrix above.

Although the process to derive pHd is very subtle, it's easy to see why the candidate matrix has its columns in the central mean subspace. Note that this

matrix can be re-written as

$$E[XX^{\intercal}E(Y|X)] - E(XX^{\intercal})E[E(Y|X)].$$

Therefore it's easy to see that if a direction in X is independent of $E(Y|X)$, it will be canceled out. Similar to SAVE, the subtle difference between the conditional independence and marginal independence partially explains why the normality assumption of X is needed.

Finally, note that the motivation of pHd is not the inverse regression of X on Y but the regression of Y on X, and that's why it can only recover the central mean subspace. With a similar motivation, Xia, Tong, Li and Zhu (2002) proposed another method to estimate the central mean subspace, called minimum average variance estimation (MAVE), as will be discussed in the next section.

In each of the three methods above, it's possible that only a proper subspace of the central dimension reduction subspace is recovered. Thus none of these three methods are exhaustive (SAVE is later shown to be exhaustive by Li and Wang (2007), though this asymptotic property is not efficiently realized in finite sample cases). One reason is that since inverse regression is involved in these methods, once the slices of Y are given, only the interactions of X's within each slice are considered, leaving "inter-slice" information untouched. This drawback is first discussed and resolved in Li, Zha and Chiaromonte (2005).

In Li, Zha and Chiaromonte (2005), the inverse regression of X is generalized to the regression of directions in X, that is, instead of the conditional moments of a single X, now we consider the conditional moments of a direction $X - \tilde{X}$ in which $\tilde{X}$ is an i.i.d copy of X. This generalization is very natural in the sense that in SDR our interest is the directions in X along which Y varies. In this sense the directions in X along which Y is constant span the orthogonal complement of the central subspace. Based on this idea, the authors proposed contour regression, in which directions in X are regressed within each contour of Y. That is, the following candidate matrix is estimated:

$$E[(X - \tilde{X})(X - \tilde{X})^{\intercal} CI(X, Y, \tilde{X}, \tilde{Y})]$$

in which $(\tilde{X}, \tilde{Y})$ is an i.i.d copy of (X, Y) and $CI(X, Y, \tilde{X}, \tilde{Y})$ is a contour indica-

tor. Then the central subspace is estimated as the null space of this matrix. In most cases, Y is continuous and there is no replication of X for each value of Y in the sample, thus similar to the inverse regression, we need to "define" the contours. However, one important feature that differs from the slicing procedure in inverse regression, is that since the direction $X - \tilde{X}$ is regressed, a contour function will be a function of both Y and $\tilde{Y}$. Hence instead of fixing slices as a partition of the sample space of Y in inverse regression, here different "slices" of Y are constructed as $\tilde{Y}$ varies, and these slices overlap with each other. In this way, the author claimed that the "inter-slice" information has been utilized, which brings benefits to their method such as exhaustiveness under the ellipticity condition (2.4) and mild additional conditions.

Two contour indicators $CI(X, Y, \tilde{X}, \tilde{Y})$ are constructed in their paper. The first one is $I(|Y - \tilde{Y}| < c)$ in which c is a threshold that needs to be determined. Because of the simpleness of this function, the corresponding contour regression is called simple contour regression (SCR). For SCR, asymptotic results have been established and the optimal c has been selected. One potential drawback of SCR is that when there is a symmetric pattern between X and Y, the contour indicator becomes insensitive. In a result, additional assumptions need to be imposed to guarantee exhaustiveness. To resolve this drawback, the second method, named as general contour regression (GCR), is constructed using a more complex contour indicator. Compared to SCR, it is still efficient under the existence of a symmetric pattern between X and Y, hence it needs less assumption to be exhaustive. However, due to its complex form, properties of GCR such as asymptotic results become difficult to develop.

A common drawback of these two methods is that the computational cost is at least $O(n^2)$, which is higher than the other inverse regression methods. This becomes one motivation of directional regression proposed in Li and Wang (2007), in which the computational cost is reduced back to $O(n)$.

Following the idea of contour regression, Li and Wang (2007) proposed directional regression to exhaustively estimate the central subspace. Like the contour regression, directional regression also regress the direction $X - \tilde{X}$, however, it directly regresses the directions on the paired response $(Y, \tilde{Y})$ instead of the contour indicators. Compared to contour regression, this simplification makes it faster in

computation. Besides, it has a clearer connection to the original inverse regression - instead of regressing a single X on a single Y, it regresses a direction $X - \tilde{X}$ on the corresponding response pair $(Y, \tilde{Y})$. Again, this generalization is very natural in the sense that the goal of SDR is to find an appropriate set of directions in X that sufficiently summarizes the information about Y.

The candidate matrix of directional regression, as straightforward as its name, is

$$E\{2I_p - E[(X - \tilde{X})(X - \tilde{X})^\intercal | Y, \tilde{Y}]\}^2.$$

When the ellipticity condition (2.4) and mild additional conditions hold, its eigenvectors associated with nonzero eigenvalues span the central subspace. Hence like contour regression, directional regression is also exhaustive. Moreover, with some calculations it can be shown that the matrix above can be re-expressed in terms of $E(XX^\intercal|Y)$ and $E(X|Y)$. Especially, it can be expressed as a sum of the candidate matrices of SIR and SAVE, plus other semi-positive matrices. Thus the space derived from this method will include the spaces from SIR and SAVE as its subspaces. Compared to exhaustiveness, this is a more direct evidence showing that this method is more comprehensive than its ancestors. Another direct result from this expression is that the computational time of directional regression is in the same order as SIR and SAVE, which is considerably less than that of contour regressions.

Li and Wang (2007) also pointed out that under ellipticity condition (2.4), SAVE is also asymptotically consistent and exhaustive. In particular, asymptotically SAVE contains SIR as a part of its result. However, in practice this phenomenon is rarely observed due to the inefficiency of SAVE in finite sample cases.

Finally, if we look at the candidate matrix of directional regression carefully, it's straightforward to see how it represents the generalization from a single X to the direction $X - \tilde{X}$: If we replace $(X - \tilde{X})$ by $X - E(X|Y)$ and $(Y, \tilde{Y})$ by Y and $2I_p$ by I_p, we will get the candidate matrix of SAVE. More generally, if we treat the term $(X - \tilde{X})(X - \tilde{X})^\intercal$ as the sample covariance of two data points up to a multiplicative scalar, then we can generalize it to the m-tuple case. That is, we take m i.i.d data points of (X, Y) and compute the sample covariance matrix of X. In this way we generalize the directional regression. Interestingly, as $m \to \infty$, the new SDR method will approach SAVE. Hence conversely, SAVE is

also a generalization of directional regression.

2.3 MAVE-type SDR methods

In this section we will review several methods in the literature that use nonparametric techniques to estimate the central dimension reduction subspace. Compared to inverse regression methods, they don't invoke assumptions on marginal distribution of X and thus have a wider application. However, the nonparametric estimation usually slows down the convergence rate of estimator, and since numerical optimization is usually involved, these methods require more computation.

As mentioned multiple times before, Xia, Tong, Li and Zhu (2002) proposed MAVE, which is a sequence of three methods based on the regression of Y on X: outer product of gradients estimation (OPG), minimum average variance estimation (MAVE) and refined MAVE (RMAVE). Note that since the method is based on $E(Y|X)$, like pHd (Li 1992), MAVE also estimates the central mean subspace (See Cook's and Li's comments attached with the paper).

The idea of MAVE is that since $E(Y|X)$ is a function of β, the gradient of $E(Y|X)$ with respect to X must be a linear combination of β. To estimate this gradient, the authors apply local linear regression of Y on X. Then the estimation of β becomes straightforward. Based on different details in the estimation procedure, these three methods have their own advantages and disadvantages.

Since the gradient of $E(Y|X)$ lies in the central mean subspace, the most straightforward approach is to use the mean of its outer product as the candidate matrix, whose column space is also in the central mean subspace. OPG is constructed in this way. Here the gradient is estimated by minimizing the following objective function which estimates $E(\mathrm{Var}(Y|X))$ using local linear approximation:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_j - a_i - b_i^{\intercal}(X_j - X_i))^2 w_{ij} \tag{2.8}$$

over $\{(a_i, b_i^{\intercal}), i = 1, \ldots, n\}$ in which $a_i \in \mathbb{R}$, $b_i \in \mathbb{R}^p$, and w_{ij} is a standardized weight:

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

Here h is the bandwidth of the kernel function K_h and needs to be determined by data-driven methods such as cross-validation. Note that in (2.8), b_i is the gradient of $E(Y|X)$ at $X = X_i$ and thus the central mean subspace is estimated by the column space of the matrix $\sum_{i=1}^{n} b_i b_i^{\mathsf{T}}$. Since local linear approximation is involved with the kernel function of p-dimensional predictor, OPG suffers from the effect of its bias term, leading to a slow convergence rate. Nevertheless, since the b_i's can be found by weighted least squares estimation, OPG is very computationally efficient and thus can serve as a "good" initial value for other SDR methods.

To improve OPG, more sophisticated methods are constructed by reducing the bias in nonparametric estimation. MAVE is one of them with the following objective function:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_j - a_i - b_i^{\mathsf{T}}\beta^{\mathsf{T}}(X_j - X_i))^2 w_{ij} \tag{2.9}$$

in which $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^{p\times d}$. Since $E(Y|X)$ is treated explicitly as a function of $\beta^{\mathsf{T}}X$, the SDR assumption (2.2) is more efficiently used. As a result, the local linear approximation is on $\mathbb{R}^d$ instead of $\mathbb{R}^p$, and β becomes an explicit variable in the optimization, leading to a more accurate estimator. In addition, the authors also mention that if a higher-order local polynomial approximation is used, the resulting estimator will be further improved to be $\sqrt{n}$-consistent.

To minimize the objective function (2.9), an initial value of β is fixed and $a = (a_1, \ldots a_n)$ and $b = (b_1, \ldots, b_n)$ can be derived using weighted least squares estimation. Then fixing a and b, β can be updated using weighted least squares estimation again. This adaptive process will stop when the convergence threshold is reached. To make this process computationally efficient, a good initial value of β is needed. One such candidate is the estimate from OPG.

A further improvement of MAVE comes from using a more compact kernel function in the nonparametric estimation. Note that since $E(Y|X)$ is a function of $\beta^{\mathsf{T}}X$, so should be the kernel function in the nonparametric estimation. This results in RMAVE which minimizes the following objective function:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_j - a_i - b_i^{\mathsf{T}}\beta^{\mathsf{T}}(X_j - X_i))^2 w_{ij}^{\beta} \tag{2.10}$$

which is the same as (2.9) except for the weight function $w_{ij}^{\beta}(\cdot) = w_{ij}(\beta^{\mathsf{T}}\cdot)$. Since w_{ij}^{β} has a lower dimensional variable, the nonparametric estimation is less biased. In this way RMAVE improves MAVE and has a faster convergence rate. Similar to MAVE, here a good initial value of β is also needed to avoid the risk of divergence. In their paper, the authors suggest using the estimate derived from MAVE.

Therefore in practice, these three methods are usually applied sequentially, with one to be the initial value of another. Hereafter we will also call this sequence of methods as MAVE, if no ambiguity is caused.

From these two modifications of OPG, one can see that there are two ways of regularization: one is to reduce the dimension of the variable in local linear expansion, and the other is to reduce the dimension of the variable in the kernel function. MAVE adopts the first one, and RMAVE adopts both of them. Actually, if only the second way is adopted in the objective function, that is, only the kernel function is modified, OPG can still be improved. More details can be found in Hristache, Juditsky, Polzehl and Spokoiny (2001). In their method, the objective function is nearly the same as (2.8), with the modified kernel function that is slightly different from RMAVE by additional smoothness of bandwidth. Namely, in the kernel function of RMAVE, the bandwidth in the directions of β is always h and the bandwidth in the orthogonal directions is always ∞; whereas in their kernel function, the bandwidth is a continuous function of X and it approaches h and ∞ in the corresponding directions as the iteration number grows.

One advantage of MAVE, as easily seen from its objective function, is that the estimate will be exhaustive under fairly general conditions. An underlying reason is that $E(Y|X)$ is a function of $\beta^{\mathsf{T}}X$ if and only if the space spanned by its gradient is a subspace of span(β). This will not be true if the gradient is replaced by the columns of the Hessian matrix. That is another reason that pHd is not exhaustive, as discussed in the last section.

Besides, in practice MAVE has been shown effective in finite sample cases. In fact, our work (see Chapter 5) shows that it is semiparametrically efficient under the homoscedasticity assumption.

In addition, the simple setting in MAVE makes it easy to be generalized to other dimension reduction problems. A first example is groupwise dimension reduction

(Li, Li and Zhu 2010). A second example is discussed below in more details.

As mentioned above, MAVE only estimates the central mean subspace. To generalize it to estimate the central subspace, Xia (2007) proposed dMAVE, which is the same as MAVE except that the response Y is replaced by a symmetric density function. Consequently, the new method can recover the central subspace exhaustively, under fairly general conditions. dMAVE can be treated as an example of the theory in Yin and Li (2011), which states that if a dense subfamily in $L_2(\mu_Y)$ serves as the response variable, then MAVE will estimate the central subspace of $Y|X$. Candidates of such dense subfamilies include polynomials, power transforms, trigonometric functions and wavelets. The corresponding method in Yin and Li (2011) is called MAVE ensemble. Similar to MAVE, it is actually the sequence of OPG ensemble, MAVE ensemble and RMAVE ensemble. Under fairly general conditions, this method has been shown to exhaustively recover the central subspace.

The idea of this generalization is not unique to MAVE. As pointed out by Yin and Li (2011), the theory underlying this generalization is that the central subspace is spanned by the union of all the central mean subspaces of a dense family of functions $\{f(Y)\}$. By picking an adequate subset of this family and estimating the corresponding central mean subspaces, the central subspace will be fully recovered. Thus the crucial part is to find an exhaustive estimator of the central mean subspace. Besides MAVE, another estimator has been proposed by Zhu and Zeng (2006) using the Fourier method. Based on the same idea of generalization using $\{e^{itY}, t \in \mathbb{R}\}$ as the dense family, Zhu and Zeng (2006) has also extended this method to exhaustively estimate the central subspace.

In more detail, the Fourier method is also based on the fact that the gradient of $E(Y|X)$ is in the central mean subspace. However, unlike MAVE which nonparametrically estimates $E(Y|X)$, it adopts Fourier transform of the gradient, and takes a weighted average of the outer product of the Fourier transform. Then the objective function is a candidate matrix and consequently the optimal β is its principle eigenvectors. Since the Fourier transform can be treated as a weighted mean of the gradient, the candidate matrix can be treated as a weighted mean of the outer product of the gradient.

By carefully selecting the weight in the above method, the candidate matrix

has a closed form expression except for the gradient of the log-likelihood of X. If the distribution of X is known, the candidate matrix will be completely specified. Otherwise, it needs to be estimated nonparametrically. Since there is no iteration involved in this method, it will be computationally efficient.

To generalize this method for the central subspace, as mentioned above, Zhu and Zeng (2006) first replaces Y by $\{e^{itY}, t \in \mathbb{R}\}$, which results in a sequence of candidate matrices, and then carefully selects the weight so that the weighted sum of the candidate matrices has a closed form up to the gradient of the log-likelihood of X, which is accessible for estimation.

Since the gradient of $E(Y|X)$ is fully used in this method, it's not surprising that the Fourier method is exhaustive for estimating the central mean subspace and the central subspace. Although this method is based on moments of $Y|X$ while contour regression and directional regression are rather based on inverse moments, interestingly, all these methods are exhaustive and their candidate matrices contain $(X - \tilde{X})$ and $(Y - \tilde{Y})$. This phenomenon might lead to some common underlying reasoning that explain the exhaustiveness of these methods.

As mentioned above, the major drawback of inverse regression methods when compared to MAVE-type methods, is that they need to assume X to be elliptically distributed. However, there is no determinative relationship between the inverse regression methods and assumption (2.4) - as one can see, the only role (2.5) and (2.7) play is to specify $E(X|\beta^{\mathsf{T}}X)$ and $\mathrm{Var}(X|\beta^{\mathsf{T}}X)$ as functions of $\beta^{\mathsf{T}}X$. Apparently these functions can be estimated more generally, which makes these conditions unnecessary. For example, in SIR we have

$$\mathrm{Var}[E(X|Y)] - \mathrm{Var}[E(X|\beta^{\mathsf{T}}X)|Y] = 0.$$

If we impose a more general parametric model on $E(X|\beta^{\mathsf{T}}X)$, then β can still be estimated by solving the sample estimate of this estimating equation. Since the estimating equation is unbiased, the consistency of β is still achieved under mild assumptions. This generalization has been proposed in Li and Dong (2009) and Dong and Li (2010), for SIR and SAVE correspondingly. In these methods, $E(X|\beta^{\mathsf{T}}X)$ and $\mathrm{Var}(X|\beta^{\mathsf{T}}X)$ are assumed to lie in a finite dimensional functional space such as the space of cubic functions, which releases the ellipticity assump-

tion (2.4) significantly and make these inverse regression methods applicable to more real data problems.

As we have seen so far, in many cases people start from the same idea and end up with different SDR methods. So it is important to systematically explore the connection between these methods. One approach is to build a unifying framework in which we can locate each of these SDR methods. Recently, this framework has been discovered in Ma and Zhu (2012a) using the family of estimating equations. The idea is that since we are only interested in estimating β, we can treat β as the parameter of interest, and treat everything else unknown as nuisance parameters, namely, the marginal distribution of X and the conditional distribution of Y given X. Under different SDR assumptions, these nuisance parameters need to satisfy different constraints. Since they are all infinite dimensional, theories for semiparametric models are needed to derive the corresponding efficient scores.

In their paper, the families of influence functions have been derived for the central subspace and the central mean subspace, without extra distributional assumptions on X. These families enjoy a double-robustness property, which makes them extendable to larger families of unbiased estimating equations. The extended families are found to be inclusive in SDR, in the sense that all the popular SDR methods in the literature, including OLS, SIR, SAVE, iHt (Cook and Li 2002), MAVE and directional regression, are equivalent to solving the corresponding estimating equations, under different additional assumptions such as (2.5) which simplifies the estimation procedure. Besides, starting from these estimating equations, it's straightforward to relax the ellipticity condition (2.5) on X, so that the corresponding methods are generalized, with the price of nonparametric estimation on $E(X|\beta^{\intercal}X)$. More details of this paper will be discussed in Chapter 5. Motivated by their work, we have derived the semiparametrically efficient estimation for SDR with respect to a general statistical functional, which includes the central mean subspace as a special case.

2.4 Other types of SDR methods

Since there have been a large class of SDR methods in the literature, a natural question is how to combine them to get a more comprehensive method. One approach

is provided by Ye and Weiss (2003), in which linear combinations of candidate matrices are considered. When all the candidate matrices lie in the same central dimension reduction subspace, as does an arbitrary linear combination of them. Moreover, an appropriate combination of these candidate matrices will preserve all the information in each matrix and thus is more comprehensive. One example of such a linear combination, is seen in Li and Wang (2007) discussed above, in which directional regression is shown to be a specific linear combination of the matrices of SIR, SAVE and additional matrices, and is shown to be an improvement of both SIR and SAVE in practice. Another example is the Fourier method discussed in the last section. A third example, will be seen in the next paragraph.

Another natural question arises when the response Y is multi-dimensional - this phenomenon becomes popular recently with the "rise of Big Data". Theoretically, all the SDR methods can still be applied without additional adjustment. However, in practice many SDR methods fail to provide a consistent estimator when the dimension of Y is large relative to the sample size. For inverse regression methods, the failure happens because the number of slices of Y will increase in an exponential order of dimension of Y, so it can dominate the sample size.

To address this problem, a natural idea is that since we can estimate the central subspace for a univariate response, we can generate a set of one-dimensional transformations of Y and estimate the corresponding central subspace for each one in the set. When this set is "large" enough, the union of these spaces will span the central subspace of Y itself. To estimate each space, suitable candidates of SDR methods are inverse regression methods, by which the estimated spaces can be easily combined by taking the sum of the corresponding candidate matrices. Note that this idea is analogous to the definition of the characteristic function for multi-dimensional random variable via its linear combinations.

This idea is realized in Li, Wen and Zhu (2008), who proposed the projective re-sampling method for multi-dimensional response. As seen from its name, the transformations adopted in this method are normalized linear combinations. As the authors have shown, if the set of linear combinations is "dense" enough, the convergence rate of the new method will be the same as if the response were univariate. To generate this set, the authors suggested taking a random sample of sufficiently large size from the uniform distribution on the unit sphere in $\mathbb{R}^q$ in

which q is the dimension of Y. Other data-driven distributions of linear combinations can also be considered to further improve the efficiency of this method.

Another example based on this idea, though not implemented yet, can be applied to the Fourier method in Zhu and Zeng (2006) as mentioned before. Since the Fourier transform is a linear combination of e^{itY}, this method can be naturally generalized for multivariate Y, and the new method will still be consistent under mild assumptions.

Finally, in real data a variety of features of Y given X can become the center of the interest, such as the conditional mean, the conditional variance and the conditional median. However, in the literature only the central subspace and the central mean subspace have received great attention. Zhu and Zhu (2009) started to fill this gap, by considering assumption (2.3) and introducing the central variance subspace. In their paper the space is estimated using a MAVE-type method, in which however the central mean subspace needs to be estimated first. Our work (Chapter 5) gives the semiparametric efficient estimator for this problem, which cannot avoid the estimation of the central mean subspace either, though it is not of interest.

2.5 Order - determination methods

As introduced in the last chapter, an important issue in SDR is to determine the reduced dimension d, as pointed out by Schott (1995) and Cook (2004). In the literature, several order-determination methods have been commonly used, in combination with different SDR methods.

One common method that is used in combination with inverse regression SDR methods is sequential tests. This method has been discussed, for example in Li (1991), Li (1992) and Bura and Yang (2011). Since most of the inverse regression methods are associated with candidate matrices, the order d is equal to the rank of the matrix, so that the order-determination procedure can be embedded in the field of matrix algebra. In particular, in this method the test statistic is a function of eigenvalues of $\hat{M}$, which is the sample estimate of the candidate matrix M. Under the null hypothesis that the last $p-k$ eigenvalues are 0, the asymptotic distribution of this statistic is derived. The hypothesis of $d \leq k$ is then tested sequentially

as k increases from 1 until the failure of rejection, and d is determined to be the stopping value k. The method is usually easy to implement once the asymptotic null distribution is available. However, this distribution varies with the inverse-regression method and is usually nontrivial to derive. Besides, since the power of each test is not available, the heuristic 0.05 is chosen to be the probability of type-I error, which is clearly not desirable.

Another method that is associated with inverse regression SDR methods is BIC-type criterion, as first proposed in Zhu, Miao and Peng (2006). Similar to the sequential tests, it also estimates d based on sample eigenvalues. The basic idea is that since the first d sample eigenvalues are in order $\mathrm{O}_P(1)$ whereas the last $p-d$ of them are in order $\mathrm{o}_P(1)$, the sum of the first k sample eigenvalues will increase rapidly with k when $k < d$ and will increase negligibly when $k \geq d$. Hence it behaves similarly to the likelihood function in variable selection. By adding a penalty function of k with an appropriate order, it's plausible that the new function will be minimized at the true value d. In Zhu, Miao and Peng (2006), this penalty is chosen to be a linear function of k, which is analogous to the L_0 penalty in variable selection. Thus this order-determination method is called BIC-type criterion. Under fairly general conditions, it has been shown consistent. However, as a common issue in penalized optimization, this method can suffer from a careless selection of tuning parameter.

In Xia, Tong, Li and Zhu (2002), cross validation is used to determine the order d for MAVE-type methods. It has a natural parallel in variable selection that has been commonly used and shown effective in practice. Besides, under fairly general assumptions, it has been shown consistent asymptotically. However, in implementation of the method, one needs to compute the fitted value for each data point X_i under the semiparametric model, which is usually not available in other SDR methods due to the primary goal of SDR. This drawback restricts the generalization of the method.

Finally, Ye and Weiss (2003) proposed a bootstrap method that is used in combination with inverse regression SDR methods. In the literature it received great attention due to its unique advantages. First, in contrast to sequential tests, it doesn't rely on asymptotic distributions of sample candidate matrices. Second, in contrast to BIC-type criterion, it does not involve a tuning parameter. Third, in

contrast to cross-validation, it does not rely on fitted values under specific models. However, the asymptotic consistency of the method has not been proved yet, and its criterion to determine d is heuristic. More details about this method will be discussed in the next Chapter.

At the end of this chapter, I would like to clarify that due to limited space, what has been reviewed here is only a part of works in the literature that are most related to this dissertation, while many other works are skipped which however are important in the SDR literature and also related to our work. Some of these works include: likelihood approach (Yin and Cook 2005; Yin, Li and Cook 2008; Cook and Forzani 2009), non-linear SDR (Fukumizu, Bach and Jordan 2009; Lee, Li and Chiaromonte 2013), and single-index quantile regression (Kong and Xia 2012). Some of them will be mentioned again in the next three chapters.

Chapter 3

Order determination for dimension reduction using an alternating pattern of spectral variability

3.1 Introduction

In this chapter, we will present our first method to estimate the dimension d of the central dimension reduction subspace, with the aid of inverse regression SDR methods. As reviewed in the last chapter, in these SDR methods, the order-determination problem is translated into estimating the rank of the candidate matrix; meanwhile, all the existing order-determination methods rely either on the level of the eigenvalues or the variability of the eigenvectors of its sample estimator. In our work however, we exploit the alternating pattern of the spectral variability - the bootstrap variability of eigenvectors and the level of eigenvalues - in positive semi-definite matrix estimators. Using this pattern, we construct the ladle plot and the corresponding new estimator called the comprehensive spectral estimator (CSE). Therefore we regard this new estimator CSE as an intrinsic improvement over the existing methods in the literature. Under fairly general con-

ditions, we will show the asymptotic consistency of CSE. We will also illustrate its efficacy compared to other order-determination methods via simulation study. The validity of CSE will be further supported by real applications. As one can see, our method can also be applied beyond SDR to estimate the rank of other positive semi-definite matrix parameters, such as the covariance matrix of X in PCA.

In more detail, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an independent and identically distributed sample of (X, Y), a pair of random elements. Let $\hat{M}$ be a consistent estimator of a $p \times p$ - dimensional positive semi-definite candidate matrix M. Then the alternating pattern of the spectral variability can be summarized as follows:

> When the eigenvalues of $\hat{M}$ are far apart, the eigenvectors of $\hat{M}$ tend to have small variability in direction; when the eigenvalues of $\hat{M}$ are close together, the eigenvectors of $\hat{M}$ tend to have large variability in direction.

Ye and Weiss (2003) proposed a procedure to estimate d based on the bootstrap variability of eigenvectors. Given the full sample, this procedure first generates a large amount of bootstrap samples and the corresponding bootstrap estimates of M, and then for each $k < p$, computes the variability of the subspaces spanned by the first k eigenvectors of these bootstrap estimates. Ye and Weiss (2003) speculated that when $k = d$, due to the consistency of $\hat{M}$, each of the k-dimensional bootstrap subspaces converges to the column space of M, and hence has low bootstrap variability. When $k > d$, each of the k-dimensional bootstrap subspaces estimates the column space of M and an arbitrary part of the null space of M. This arbitrariness then leads to large bootstrap variability. When $k < d$, each of the k-dimensional bootstrap subspaces estimates a proper subspace of the column space of M, and the potential arbitrariness of this proper subspace leads to a more complex pattern of bootstrap variability. Based on this intuition, Ye and Weiss's bootstrap procedure estimates d as the largest k such that this bootstrap variability is small. The upper-left panel of Figure 3.1 shows a typical pattern of the bootstrap variability of eigenvectors, as computed from an example in Section 5 with $d = 2$. As one can see, the speculation in Ye and Weiss (2003) is valid in this case.

As reviewed in the last chapter, in contrast to the bootstrap procedure, both the sequential tests (Schott 1994; Cook and Li 2004; Bura and Yang 2011) and BIC-type criterion rely on the eigenvalues of $\hat{M}$ and they are asymptotically consistent

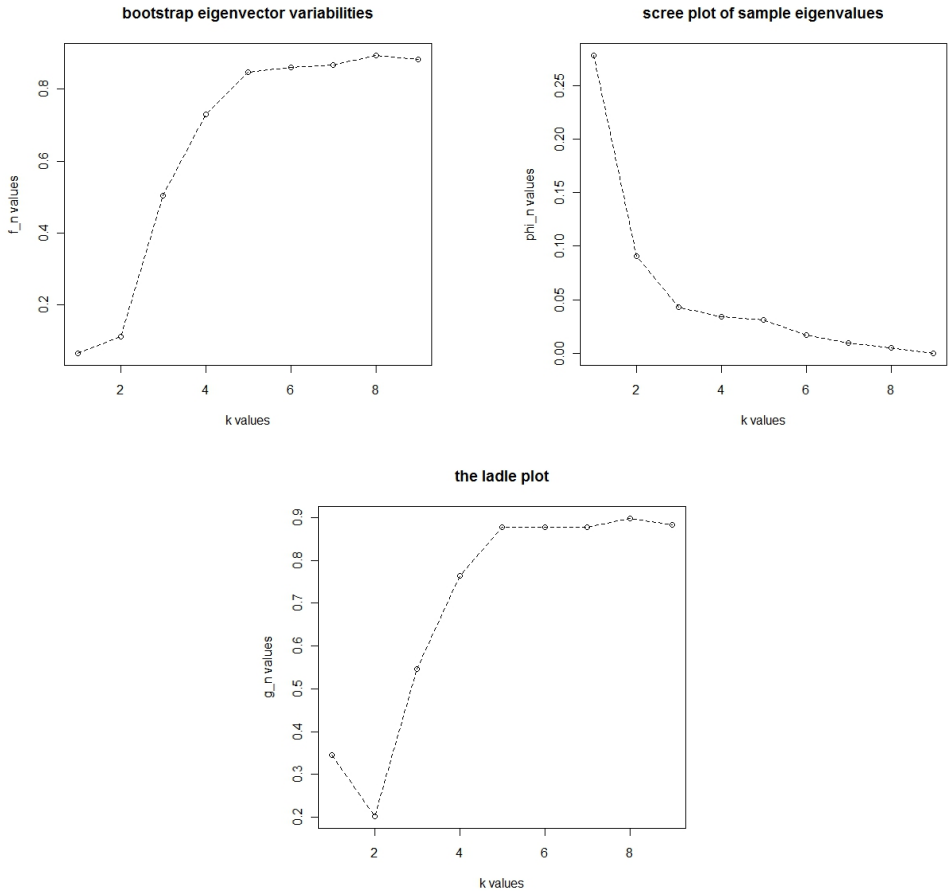

Figure 3.1. The benefit of combining eigenvalues and variability of eigenvectors: the upper-left panel is the plot of bootstrap variability of eigenvectors; the upper-right panel is the scree plot; the lower panel is the ladle plot based on a combination of eigenvalues and eigenvector variability.

under certain conditions. However, they both suffer from their inherent limitations. In particular, as our simulation studies indicate, BIC-type criterion is quite sensitive to the tuning parameter, and may perform differently under different models. The upper-right panel of Figure 3.1 is the scree plot (i.e. the plot of successive eigenvalues of $\hat{M}$) using the same example. Therefore both the sequential tests and the BIC-type criterion are based on the monotone transforms of this plot.

From the two plots in the upper panels of Figure 3.1, we see that, if we treat both the bootstrap variability of eigenvectors and the level of eigenvalues of $\hat{M}$ as functions of k, then for $k \leq d\,(d = 2)$, the bootstrap variability of eigenvectors is relatively flat but the level of eigenvalues sharply decreases; whereas for $k > d$, the

bootstrap variability of eigenvectors sharply increases whereas the level of eigenvalues becomes relatively flat. Clearly, both functions provide useful information about the value of d, and it is conceivable that, by combining them appropriately we could arrive at a sharper estimator of d. Indeed, the plot in the lower panel of Figure 3.1 is based on a scale-free combination of these two functions. We see that it is minimized at d. From our experience the ladle shape in the lower plot is common for many models. For this reason we call such a plot the ladle plot, as an alternative to the scree plot. The goal of this chapter is to introduce the ladle plot for order determination, study its consistency, and investigate its performance through simulations and applications.

The rest of the chapter is organized as follows. In Section 2, by inspecting the alternating pattern between the variability of eigenvectors and the level of eigenvalues, we propose the ladle plot, and a consequent dimension estimator. In Sections 3 and 4, we give a rigorous mathematical characterization of this alternating pattern, and use it to prove the consistency of the new estimator. In Section 5, we conduct simulation studies to compare the new estimator with the previous eigenvalue-based or eigenvector-based estimators. In Section 6 we apply the new estimator to a real-data example. Some discussions are made in Section 7. For continuity in exposition, we leave the proofs of all the lemmas and corollaries in Section 7.

3.2 The ladle plot

In this section we lay out the notation and the theoretical background, and explain the alternating pattern between the eigenvalue levels and the eigenvector variability, on which the ladle plot is based. We also propose the new estimator which we call the comprehensive spectral estimator, or CSE for short.

First, we set up the notation that will be used throughout the article. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathbb{N}$ be the set of natural numbers. Let Ω_S be the collection of all sequences $\{a_n : n \in \mathbb{N}\}$ where $a_n \in \mathbb{R}^{p+1}$, and $\mathcal{F}_S$ be the Borel σ-field in Ω_S. Let $S : \Omega \to \Omega_S$ be a random element that is measurable with respect to $\mathcal{F}/\mathcal{F}_S$. Thus, for each $\omega \in \Omega$, $S(\omega)$ is a sequence of $(p+1)$ - dimensional vectors $\{S_n(\omega) : n \in \mathbb{N}\}$ in which $S_n : \Omega \to \mathbb{R}^{p+1}$ is a random vector $\omega \mapsto S_n(\omega)$.

Let $P_S = P \circ S^{-1}$ be the probability on $(\Omega_S, \mathcal{F}_S)$ induced by S. Then $(\Omega_S, \mathcal{F}_S, P_S)$ is the probability space of the sequence of random variables $\{S_n : n \in \mathbb{N}\}$. Assume this sequence to be an independent and identically distributed sequence, and let $S_n = (X_n, Y_n)$ for each $n \in \mathbb{N}$, where X_n is a p-dimensional random vector representing the predictor, and Y_n is a random variable representing the response. In this way, we identify each random sample of size n with the set of the first n elements in a sequence $s \in \Omega_S$. Conversely, we identify each $s \in \Omega_S$ with a sequence of random samples with increasing sample size. Following the tradition of the bootstrap literature (Bickel and Freedman 1981), we will work directly with $(\Omega_S, \mathcal{F}_S, P_S)$ without referring to $(\Omega, \mathcal{F}, P)$.

Let F be the distribution of S_1 and F_n be the empirical distribution based on $\{S_1, \ldots, S_n\}$. This is a random measure defined on a finite support. For any fixed $s \in \Omega_S$, let $S^*_{1,n}, \ldots, S^*_{n,n}$ be an independent and identically distributed sample from F_n. Let F^*_n be the empirical distribution based on $\{S^*_{1,n}, \ldots, S^*_{n,n}\}$. Both F_n and F^*_n are random measures, but once $s \in \Omega_S$ is given, F_n is a fixed measure while F^*_n is still random. The additional layer of randomness comes from bootstrap re-sampling.

Let $\mathfrak{F}$ be the family of all distributions of S_1. Without causing ambiguity, we denote M both as a statistical functional $M : \mathfrak{F} \to \mathbb{R}^{p \times p}$ and the candidate matrix $M(F)$. We write $M(F_n)$ as $\hat{M}$ and call it the sample estimator of M; we write $M(F^*_n)$ as M^* and call it the bootstrap estimator of M. Note again that once $s \in \Omega_S$ is given, or equivalently a sequence of full samples is given, $\{\hat{M} : n \in \mathbb{N}\}$ is non-stochastic but $\{M^* : n \in \mathbb{N}\}$ is still random. Suppose $\mathrm{rank}(M) = d$ and $\lambda_1 \geq \ldots \geq \lambda_d > 0 = \lambda_{d+1} = \ldots = \lambda_p$ are the eigenvalues of M in the descending order. Let $v_1, \ldots, v_p$ be the eigenvectors of M such that for each $i = 1, \ldots p$, $Mv_i = \lambda_i v_i$ and the matrix $(v_1, \ldots, v_p)$ is orthogonal. Suppose $\mathcal{S}_1, \ldots, \mathcal{S}_l$ are the eigenspaces of M in which $\mathcal{S}_l$ is the null space. Note that $\{v_{d+1}, \ldots v_p\}$ spans $\mathcal{S}_l$. Similarly, we define $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_p, \hat{v}_1, \ldots, \hat{v}_p\}$ and $\{\lambda^*_1, \ldots, \lambda^*_p, v^*_1, \ldots, v^*_p\}$ for $\hat{M}$ and M^*. For each $k \leq p$, let

$$B_k = (v_1, \ldots, v_k), \ \hat{B}_k = (\hat{v}_1, \ldots, \hat{v}_k), \ B^*_k = (v^*_1, \ldots, v^*_k).$$

Given an independent and identically distributed sample of size n, we generate n independent copies of bootstrap samples, then for each $k < p$ we have n random

matrices $B^*_{k,1}, \ldots, B^*_{k,n}$. Define a function

$$f_n : \{1, \ldots, p-1\} \to \mathbb{R}, \quad f_n(k) = n^{-1} \sum_{i=1}^{n} [1 - |\det(\hat{B}_k^{\intercal} B^*_{k,i})|].$$

Then f_n approximates the bootstrap variability of eigenvectors as k varies. Note that the range of f_n is $[0, 1]$ - it reaches 0 if each $B^*_{k,i}$ spans the same column space as $\hat{B}_k$, and it reaches 1 if each $B^*_{k,i}$ spans a space orthogonal to $\hat{B}_k$.

For each $k < p$, similar to M^*, $\{f_n(k) : n \in \mathbb{N}\}$ is a sequence of bootstrap random variables. That is, given $s \in \Omega_S$, it is still a sequence of random variables. To formulate the appropriate asymptotic framework, for any sequence of bootstrap random variables $\{Q_n\}$ and any probabilistic property $\mathbb{P}$, we define

$$Q_n \text{ has property } \mathbb{P} \quad \text{ almost surely } \ P_S \tag{3.1}$$

if there exists $A \in \mathcal{F}_S$ such that $P_S(A) = 1$ and given any $s \in A$, the sequence of random variables $\{Q_n\}$ satisfies $\mathbb{P}$. This framework has been used in, for example, Bickel and Freedman (1981). It will be frequently used once we begin rigorous proofs in the next section.

As mentioned in the last section, Ye and Weiss (2003) speculated a large value of f_n at all $k > d$ and a small value of f_n at $k = d$. This intuition can be expanded into the following alternating pattern between $f_n(k)$ and λ_k. For each $k < p$, consider the two cases:

Case 1: $\lambda_k > \lambda_{k+1}$. Suppose $v_k \in \mathcal{S}_i$, then $v_{k+1} \in \mathcal{S}_{i+1}$. In this case, in each of the bootstrap samples we are always estimating the same k - dimensional space spanned by the union of $\mathcal{S}_1, \ldots, \mathcal{S}_i$. The consistency of $\hat{M}$ then leads to a small value of $f_n(k)$. In particular, this is always the case for $k = d$ at which $\lambda_d > 0 = \lambda_{d+1}$.

Case 2: $\lambda_k = \lambda_{k+1}$. Suppose $v_k \in \mathcal{S}_i$, then so is v_{k+1}. In this case, in each of the bootstrap samples we are estimating the k - dimensional space spanned by the union of $\mathcal{S}_1, \ldots, \mathcal{S}_{i-1}$ and a random proper subspace of $\mathcal{S}_i$. This randomness causes large variation in the estimates, and consequently a large value of $f_n(k)$. In particular, this is always the case for all $k > d$ at which $\lambda_k = \lambda_{k+1} = 0$.

When $k < d$, both of these cases may occur, depending on the equalities among $\lambda_1, \ldots, \lambda_d$. However, the eigenvalue λ_k is always large. Therefore, by carefully

combining the eigenvector variability $f_n(k)$ and the sample eigenvalue $\hat{\lambda}_k$, we can use the alternating pattern above to characterize the true rank d.

First, we define the following transformation of eigenvalues, in the sample level,

$$\phi_n : \{1, \ldots, p-1\} \to \mathbb{R}, \quad \phi_n(k) = \hat{\lambda}_{k+1} / (\sum_{i=1}^{p} \hat{\lambda}_i). \tag{3.2}$$

Here we normalize ϕ_n to make it invariant to scale change in the data, as well as f_n. Note that by shifting the eigenvalues 1 unit to the left, ϕ_n takes a small value at $k = d$, which will be seen crucial. Then we define the objective function

$$g_n : \{1, \ldots, p-1\} \to \mathbb{R}, \quad g_n(k) = f_n(k) + \phi_n(k). \tag{3.3}$$

Note that g_n collects the information from both eigenvectors and eigenvalues. In the spirit of the BIC-type criterion, f_n serves as the penalty term to lift the objective function on the right-hand side of d. However, compared to the BIC-type criterion, this new penalty term is an intrinsic property of matrices, and is combined with the "likelihood" term ϕ_n scale-freely, so that tuning parameter selection is avoided.

Intuitively, when the signal is distinguishable from noise, e.g. λ_d is not close to 0, then on the left-hand side of d the eigenvalue term ϕ_n is large; on the right-hand side of d the eigenvector term f_n is large; and they are both small at d. Therefore we expect g_n to have the "ladle" shape and reach its minimum likely at d. In other words, if we define a new estimator

$$\hat{d} = \arg\min \{g_n(k) : k = 1, \ldots, p-1\} \tag{3.4}$$

then we expect $\hat{d}$ to be consistent under fairly general conditions. By the definition of g_n, we call this new estimator the comprehensive spectral estimator, or CSE for short.

3.3 Theoretical analysis of alternating spectral variation

In this section, we give a rigorous characterization of the alternating pattern of spectral variability discussed in the last section. First, we introduce the regularity assumptions that will be adopted throughout the rest of the article.

Assumption 1. *(asymptotic linearity) There is a random matrix $H(X)$ with mean 0 and finite second moment such that*

$$\hat{M} = M + E_n H(X) + o_P(n^{-1/2}). \tag{3.5}$$

Assumption 2. *(self-similarity) The bootstrap estimator M^* satisfies*

$$n^{1/2}(\mathrm{vech} \circ M^* - \mathrm{vech} \circ \hat{M}) \xrightarrow{\mathcal{D}} N\left(0, \mathrm{var}_F(\mathrm{vech} \circ H(X))\right) \text{ almost surely } P_S \tag{3.6}$$

where vech$\circ$ *is the vectorization of the upper triangular part of a matrix and* $\mathrm{var}_F(\mathrm{vech} \circ H(X))$ *is positive definite.*

Assumption 3. *(Existence of mean) For any sequence of nonnegative random variables $\{Z_n : n \in \mathbb{N}\}$ involved hereafter, if $Z_n = O_P(c_n)$ for some sequence $\{c_n : n \in \mathbb{N}\}$ with $c_n > 0$, then $E(c_n^{-1} Z_n)$ exists for each $n \in \mathbb{N}$ and $E(c_n^{-1} Z_n) = O(1)$.*

Assumption 1 is quite mild: it is satisfied if the statistical functional M is Hadamard differentiable. See, for example, Bickel, Klaassen, Ritov and Wellner (1993), Page 19 and Fernholz (1983). Because $\{M^* = M(F_n^*) : n \in \mathbb{N}\}$ is a sequence of bootstrap random variables, Assumption 2 follows the asymptotic framework (3.1) and states that the asymptotic behavior of $n^{1/2}(M^* - \hat{M})$ mimics that of $n^{1/2}(\hat{M} - M)$. Instead of writing down the sufficient conditions for this assumption, we take it as a starting point of our asymptotic development. For more details about the validity of self-similarity, see, for example, Bickel, Klaassen, Ritov and Wellner (1981), Parr (1985), Liu, Singh and Lo (1989) and Gill (1989). Assumption 3 is a standard moment regulation that has been commonly used in the literature. All these assumptions are fairly general and will not significantly restrict the application of our estimator.

Assumption 3 regulates the behavior of the sample mean, as shown in the following lemma.

Lemma 1. *Suppose $X_{n,1}, \ldots, X_{n,n}$ are independent and identically distributed nonnegative random variables for each $n \in \mathbb{N}$, and there exists $\{c_n : n \in \mathbb{N}\} \subset \mathbb{R}$ such that $X_{n,1} = O_P(c_n)$. Then*

$$\bar{X}_n := \sum_{i=1}^{n} X_{n,i} \,/\, n = O_P(c_n)$$

The next lemma provides the first key to our result in this section, as it regulates the order between the variability of eigenvectors and the level of eigenvalues. In particular, it provides the essential argument to show the first part of the alternating pattern, that is, far apart eigenvalues are associated with small variability of eigenvectors.

Lemma 2. *Suppose $\{\Sigma_n : n \in \mathbb{N}\} \subset \mathbb{R}^{p\times p}$ is a sequence of non-stochastic positive semi-definite matrices, and $\{\hat{\Sigma}_n : n \in \mathbb{N}\} \subset \mathbb{R}^{p\times p}$ is a sequence of random positive semi-definite matrices. Moreover, suppose $\hat{\Sigma}_n - \Sigma_n = O_P(n^{-1/2})$. For $i = 1, \ldots, p$, define $\lambda_i(\Sigma_n)$, $v_i(\Sigma_n)$ and $v_i(\hat{\Sigma}_n)$ accordingly. Then*

$$|v_i(\Sigma_n)^\intercal v_j(\hat{\Sigma}_n)| \times [\lambda_i(\Sigma_n) - \lambda_j(\Sigma_n)] = O_P(n^{-1/2})$$

for any $i, j = 1, \ldots, p$ with $i \neq j$.

Let Σ_n be $\hat{M}$ and $\hat{\Sigma}_n$ be M^*, then a direct application of this lemma leads to:

Lemma 3. *Under Assumptions 1 and 2, for any positive semi-definite candidate matrix $M \in \mathbb{R}^{p\times p}$ and any $i, j \in \{1, \ldots, p\}$, if $\lambda_i > \lambda_j$, then*

$$\hat{v}_i^\intercal v_j^* = O_P(n^{-1/2}) \quad \textit{almost surely} \quad P_S. \tag{3.7}$$

For each $k < p$, recall that both $\{\hat{B}_k, \hat{v}_{k+1}, \ldots, \hat{v}_p\}$ and $\{B_k^*, v_{k+1}^*, \ldots, v_p^*\}$ are orthogonal, thus this lemma implies that if $\lambda_k > \lambda_{k+1}$, then the spaces spanned by B_k^* and $\hat{B}_k$ are nearly identical almost surely P_S. This leads to the next theorem, which states that $f_n(k)$ is small in this case.

Theorem 1. *Under Assumptions 1, 2 and 3, for any positive semi-definite candidate matrix $M \in \mathbb{R}^{p\times p}$, if $k < p$ and $\lambda_k > \lambda_{k+1}$, then*

$$f_n(k) = O_P(n^{-1}) \quad \textit{almost surely } P_S.$$

PROOF. For simplicity in notation we denote

$$M_{11} \equiv \hat{B}_k^\intercal B_k^* \quad \text{and} \quad M_{21} \equiv [\hat{v}_{k+1}, \ldots, \hat{v}_p]^\intercal B_k^*.$$

By Assumption 3 and Lemma 1, we only need to show that

$$1 - |\det(M_{11})| = O_P(n^{-1}) \quad \text{almost surely } P_S.$$

Denote $A_1, A_2 \in \mathcal{F}_S$ as in the proof of Lemma 3. Then by the Law of the Iterated Logarithm, Zhao, Krishnaiah and Bai (1986), Assumptions 1 and 2, we have $P_S(A_1 \cap A_2) = 1$. Conditioning on any $s \in A_1 \cap A_2$, by Lemma 3, we know $M_{21} = O_P(n^{-1/2})$. Because $\hat{B}_p$ is an orthogonal matrix and $(B_k^*)^\intercal B_k^* = I_k$, we have

$$M_{11}^\intercal M_{11} + M_{21}^\intercal M_{21} = (B_k^*)^\intercal \hat{B}_p \hat{B}_p^\intercal B_k^* = I_k.$$

Thus $M_{11}^\intercal M_{11} = I_k - M_{21}^\intercal M_{21} = I_k - O_P(n^{-1})$. Since $\det(\cdot)$ is a Lipschitz continuous function, we have

$$\det(M_{11})^2 = \det(M_{11}^\intercal M_{11}) = \det(I_k) - O_P(n^{-1}),$$

which implies that

$$1 - |\det(M_{11})| = (\det(I_k) - \det(M_{11})^2) \,/\, (1 + |\det(M_{11})|) = O_P(n^{-1}).$$

This completes the proof. □

From this theorem, when $\lambda_k > \lambda_{k+1}$, for almost any large-sized full sample, $f_n(k)$ takes a small value. This reconfirms our intuition about Case 1 in Section 2, which is the first part of the alternating pattern. As the second part of the alternating pattern, when $\lambda_k = \lambda_{k+1}$, we expect $f_n(k)$ to take a large value. Note that $O_P(1)$ includes $o_P(1)$ as a special case, thus it is not strong enough to iden-

tify a sequence of "large" random variables. Therefore, in the following we will first rigorously define a sequence of "large" random variables and then show that $\{f_n(k) : n \in \mathbb{N}\}$ is such a sequence in this case.

Definition 1. *(bounded below from 0 in probability) A sequence of random variables $\{Z_n : n \in \mathbb{N}\}$ is bounded below from 0 in probability and written as $Z_n = O_P^+(1)$, if for any positive non-stochastic sequence $\{\epsilon_n : n \in \mathbb{N}\}$ with $\epsilon_n \to 0$, $Z_n > \epsilon_n$ in probability, that is,*

$$\lim_{n\to\infty} P(Z_n > \epsilon_n) = 1. \tag{3.8}$$

Furthermore, for any positive non-stochastic sequence $\{c_n : n \in \mathbb{N}\}$, if $Z_n/c_n = O_P^+(1)$ then we write $Z_n = O_P^+(c_n)$.

In a rough sense, this concept is the asymptotic analogue of a random variable Z taking positive values with probability 1. Some of its properties are listed in the next lemma. In particular, they justify that if a sequence of random variables (Z_n) is $O_P^+(1)$, then any sequence of random variables greater than (Z_n) is also $O_P^+(1)$, and (Z_n) is greater than any sequence of "small" random variables, i.e. $o_P(1)$. Hence this concept indeed defines a sequence of "large" random variables.

Lemma 4. *For any sequences of random variables $\{X_n : n \in \mathbb{N}\}$, $\{Y_n : n \in \mathbb{N}\}$ and $\{Z_n : n \in \mathbb{N}\}$,*

(a) if $Z_n \geq X_n$ and $X_n = O_P^+(1)$, then $Z_n = O_P^+(1)$;

(b) if $X_n = O_P^+(1)$ and $Y_n = o_P(1)$, then $X_n - Y_n = O_P^+(1)$. In particular, $X_n > Y_n$ in probability.

(c) if X_n can be written as $X_n = c_n + R_n$ where $\{c_n \in \mathbb{R} : n \in \mathbb{N}\}$ is a non-stochastic sequence with $\liminf_{n\to\infty} c_n > 0$ and $R_n = o_P(1)$, then $X_n = O_P^+(1)$.

Although being $O_P^+(1)$ seems a restrictive condition, based on part (c) in this lemma, the following corollary provides a general condition for the sample mean to be $O_P^+(1)$. Since f_n is the bootstrap sample mean, this corollary simplifies the argument in the part two of the alternating pattern.

Corollary 1. *Suppose for each $n \in \mathbb{N}$, $X_{n,1}, \ldots, X_{n,n}$ are independent and identically distributed nonnegative random variables, and $\bar{X}_n = (\sum_{i=1}^n X_{n,i})/n$. Then $\bar{X}_n = O_P^+(1)$ if one of the following equivalent statements holds:*

(a) There exists $\epsilon > 0$ *such that* $\liminf_{n\to\infty} P(X_{n,1} > \epsilon) > 0$*;*

(b) For any sequence $\{\epsilon_n > 0 : n \in \mathbb{N}\}$ *with* $\epsilon_n \to 0$, $\liminf_{n\to\infty} P(X_{n,1} > \epsilon_n) > 0$.

The second issue in investigating f_n is that it involves the determinant of matrices, which is often difficult to handle. This issue is addressed in the next lemma.

Lemma 5. *Suppose* $A = (a_{i,j}) \in \mathbb{R}^{p\times p}$ *and each column of* A *is of unit length. For each* $k < p$*, let* A_k *be the* $k \times k$ *submatrix of* A *formed by the elements of* A *in the first* k *rows and first* k *columns. Then*

$$1 - |\det(A_k)| \geq a_{k+1,k}^2 \,/\, 2.$$

Let $A_k = \hat{B}_k^\intercal B_k^*$, then since both $\hat{B}_p$ and B_p^* are orthogonal matrices, it is easy to see that each column of $\hat{B}_p^\intercal B_p^*$ is of unit length. Hence this lemma indicates that

$$1 - |\det(\hat{B}_k^\intercal B_k^*)| \geq (\hat{v}_{k+1}^\intercal v_k^*)^2 \,/\, 2.$$

Based on these results, we are ready to show the next theorem, which demonstrates the part two of the alternating pattern that close enough eigenvalues are associated with large bootstrap variability of eigenvectors.

Theorem 2. *Let* $c_n = \{\log[\log(n)]\}^{-2}$*. Under Assumptions 1 and 2, for any positive semi-definite candidate matrix* $M \in \mathbb{R}^{p\times p}$*, if* $k < p$ *and* $\lambda_k = \lambda_{k+1}$*, then*

$$f_n(k) = O_P^+(c_n) \quad \textit{almost surely } P_S.$$

Although $\{c_n : n \in \mathbb{N}\}$ converges to 0, in practice the convergence rate is very slow. For example, $\{\log[\log(10^4)]\}^{-2} \approx 0.2$. Hence in practice the difference between $O_P^+(c_n)$ and $O_P^+(1)$ is negligible.

PROOF. For simplicity in explanation we assume $\lambda_{k-1} > \lambda_k = \lambda_{k+1} > \lambda_{k+2}$. The general case can be shown in the same way. Denote $A_1, A_2 \in \mathcal{F}_S$ as in the proof of Lemma 3, then again by the Law of the Iterated Logarithm, Zhao, Krishnaiah and Bai (1986), Assumptions 1 and 2, we have $P_S(A_1 \cap A_2) = 1$. Given any $s \in A_1 \cap A_2$, since $\lambda_k = \lambda_{k+1}$, we have

$$\{(nc_n)^{1/2}\,[\hat{\lambda}_k - \hat{\lambda}_{k+1}] : n \in \mathbb{N}\} = O(1). \tag{3.9}$$

By Corollary 1 and Lemma 5, we only need to show that for any sequence $\{\epsilon'_n > 0 : n \in \mathbb{N}\}$ with $\epsilon'_n \to 0$,

$$\liminf_{n\to\infty} P(|\hat{v}_{k+1}^{\intercal} v_k^*| > \epsilon'_n c_n^{1/2}) > 0 \tag{3.10}$$

If (3.10) is not true, then there exists a sequence $\{\epsilon'_n > 0 : n \in \mathbb{N}\} = o(1)$ and a subsequence $(m) \subset \mathbb{N}$ such that

$$\lim_{m\to\infty} P(|\hat{v}_{k+1}^{\intercal} v_k^*| \leq \epsilon'_m c_m^{1/2}) = 1 \tag{3.11}$$

For simplicity in notation, in (3.11) and the rest of the proof for each $i \leq p$ we denote $\hat{v}_i$ as the ith eigenvector of $M(F_m)$ without referring to m, and similarly we denote v_i^*. Since $\{(\hat{v}_k, \hat{v}_{k+1}) : m \in \mathbb{N}\}$ is bounded, by Bolzano-Weierstrass Theorem we can further suppose $(\hat{v}_k, \hat{v}_{k+1}) \to (u, v)$ for some $u, v \in \mathbb{R}^p$ as $m \to \infty$. Then by Slutsky's Theorem, Assumption 2 implies that

$$m^{1/2}\,\mathrm{vech}\circ[(\hat{v}_k, \hat{v}_{k+1})^{\intercal}(M^* - \hat{M})(\hat{v}_k, \hat{v}_{k+1})] \xrightarrow{\mathcal{D}} N(0, \Sigma). \tag{3.12}$$

in which $\Sigma = \mathrm{var}_F\{\mathrm{vech}\circ[(u, v)^{\intercal} H(X)(u, v)]\}$. Let $\sigma_{1,2}^2$ be the diagonal element of Σ corresponding to the $(1, 2)$th position of M^*. Then since $\mathrm{var}_F(\mathrm{vech}\circ H(X))$ is positive definite, we have $\sigma_{1,2}^2 > 0$. Then (3.12) implies

$$m^{1/2} \sum_{i=1}^{p} \lambda_i^* \,(\hat{v}_k^{\intercal} v_i^*)\,(\hat{v}_{k+1}^{\intercal} v_i^*) \xrightarrow{\mathcal{D}} N(0, \sigma_{1,2}^2). \tag{3.13}$$

By Lemma 3, for any $i \notin \{k, k+1\}$, $(\hat{v}_k^{\intercal} v_i^*)(\hat{v}_{k+1}^{\intercal} v_i^*) = O_P(m^{-1})$. Moreover, since both B_p^* and $(\hat{v}_k, \hat{v}_{k+1})$ are orthogonal, $\hat{v}_k^{\intercal} B_p^* B_p^{*\intercal} \hat{v}_{k+1} = 0$. Therefore,

$$(\hat{v}_k^{\intercal} v_k^*)(\hat{v}_{k+1}^{\intercal} v_k^*) + (\hat{v}_k^{\intercal} v_{k+1}^*)(\hat{v}_{k+1}^{\intercal} v_{k+1}^*) = -\sum_{i\notin\{k,k+1\}} (\hat{v}_k^{\intercal} v_i^*)(\hat{v}_{k+1}^{\intercal} v_i^*) = O_P(m^{-1}).$$

Thus (3.13) indicates that

$$m^{1/2}\,(\hat{v}_k^{\intercal} v_k^*)\,(\hat{v}_{k+1}^{\intercal} v_k^*)\,(\lambda_k^* - \lambda_{k+1}^*) \xrightarrow{\mathcal{D}} N(0, \sigma_{1,2}^2). \tag{3.14}$$

Consequently, fix any $\delta > 0$,

$$\lim_{m\to\infty} P(m\,(\hat{v}_k^{\mathsf{T}} v_k^*)^2\,(\hat{v}_{k+1}^{\mathsf{T}} v_k^*)^2\,(\lambda_k^* - \lambda_{k+1}^*)^2 > \delta) > 0.$$

Since $0 \le (\hat{v}_k^{\mathsf{T}} v_k^*)^2 \le 1$, this implies

$$\lim_{m\to\infty} P(m\,(\hat{v}_{k+1}^{\mathsf{T}} v_k^*)^2\,(\lambda_k^* - \lambda_{k+1}^*)^2 > \delta) > 0. \tag{3.15}$$

For each $m \in \mathbb{N}$, denote the event in (3.11) as C_m, and the event in (3.15) as D_m. Then (3.11) implies that $\lim_{m\to\infty} P(C_m \cap D_m) > 0$. On the other hand, $C_m \cap D_m$ implies that $m(\epsilon'_m)^2 c_m (\lambda_k^* - \lambda_{k+1}^*)^2 > \delta$. Hence

$$\liminf_{m\to\infty} P(m\,c_m\,(\lambda_k^* - \lambda_{k+1}^*)^2 > \delta\,/\,(\epsilon'_m)^2) > 0 \tag{3.16}$$

On the other hand, by Assumption 2, Zhao, Krishnaiah and Bai (1986) and (3.9), $\lambda_k^* - \lambda_{k+1}^* = O_P((mc_m)^{-1/2})$. Since $\epsilon'_m \to 0$, we have $\delta/(\epsilon'_m)^2 \to \infty$, which means

$$\lim_{m\to\infty} P(mc_m(\lambda_k^* - \lambda_{k+1}^*)^2 > \delta\,/\,(\epsilon'_m)^2) = 0.$$

This contradiction shows that (3.10) is true, which means that $f_n(k) = O_P^+(c_n)$ given any $s \in A_1 \cap A_2$. This completes the proof. □

Combine Theorem 1 and Theorem 2, we have the following result:

$$f_n(k) = \begin{cases} O_P^+(c_n) & \text{if } \lambda_k = \lambda_{k+1} \\ O_P(n^{-1}) & \text{if } \lambda_k > \lambda_{k+1} \end{cases} \quad \text{almost surely } P_S$$

which now completely characterizes the alternating pattern between the bootstrap variability of eigenvectors and level of eigenvalues.

3.4 Consistency of CSE

In this section we show the consistency of the new estimator CSE defined in Section 2, as a natural consequence of the alternating pattern proved in the last section.

First of all, in conjunction with the bootstrap variability of eigenvectors, the next lemma characterizes the "reverse" trend in the sample eigenvalues as k varies.

Lemma 6. *Let $c_n = \{\log[\log(n)]\}^{-2}$. Under Assumption 1, for any positive semi-definite candidate matrix $M \in \mathbb{R}^{p\times p}$ of rank d, and for each $k < p$,*

$$\phi_n(k) = \begin{cases} O_P^+(1) & \text{if } k < d \\ O_P(c_n^{-1/2} \cdot n^{-1/2}) & \text{if } k \geq d \end{cases} \quad \text{almost surely } P_S.$$

Given the full sample, although ϕ_n is non-stochastic, we treat it as a degenerate random element for convenience in asymptotic development. As this lemma points out, under the framework (3.1) the convergence rate is $(nc_n)^{1/2}$, which is slightly slower than the usual $n^{1/2}$ convergence rate.

Based on all the discussions above, the following theorem states the consistency of CSE.

Theorem 3. *Under Assumptions 1, 2 and 3, for any positive semi-definite candidate matrix $M \in \mathbb{R}^{p\times p}$ of rank d, the estimator $\hat{d}$ defined in (3.4) is a consistent estimator of d in the sense that*

$$\lim_{n\to\infty} P(\hat{d} = d) = 1 \quad \text{almost surely } P_S \tag{3.17}$$

where $P(\hat{d} = d)$ is the conditional probability given the full sample sequence $s \in \Omega_S$.

PROOF. Let $c_n = \{\log[\log(n)]\}^{-2}$, then it is easy to see that

$$O_P(n^{-1}) = o_P(c_n), \quad O_P^+(1) = O_P^+(c_n), \quad O_P(c_n^{-1/2} \cdot n^{-1/2}) = o_P(c_n).$$

Thus Theorem 1, Theorem 2 and Lemma 6 have shown that

$$\begin{cases} f_n(k) = O_P(1), \phi_n(k) = O_P^+(c_n) & \text{if } k < d \\ f_n(k) = o_P(c_n), \phi_n(k) = o_P(c_n) & \text{if } k = d \\ f_n(k) = O_P^+(c_n), \phi_n(k) = O_P(1) & \text{if } k > d \end{cases} \quad \text{almost surely } P_S.$$

Since both f_n and ϕ_n are nonnegative and $g_n = f_n + \phi_n$, $g_n \geq \max\{f_n, \phi_n\}$. Thus Lemma 4.1 implies that

$$g_n(k) = \begin{cases} O_P^+(c_n) & \text{if } k \neq d \\ o_P(c_n) & \text{if } k = d \end{cases} \quad \text{almost surely } P_S.$$

By Lemma 4.2 , g_n is minimized at d in probability almost surely P_S. This completes the proof. □

By this theorem, given almost any large-sized full sample, the curve of g_n tends to have the ladle shape and CSE tends to truly estimate d. Besides, the one-unit shift of eigenvalues makes ϕ_n small at $k = d$, which enables the combination of the two pieces in g_n to be scale-free and simultaneously generate consistent estimates.

In addition, since ϕ_n always takes a large value on the left-hand side of d, when f_n is also large on the left-hand side of d, g_n is substantially greater than both of them. In other words, the information from the eigenvectors and the eigenvalues, when both are valuable, will be amplified by the suitable combination g_n. Therefore in practice, we expect CSE to be stabler than the estimators solely based on the eigenvectors or the eigenvalues, when they are all theoretically consistent.

3.5 Comparison with other methods via simulation studies

In this section we conduct simulation studies to compare our estimator with the order-determination methods mentioned in Section 1, namely, the BIC-type criterion, the sequential tests and the bootstrap procedure.

For a fair comparison, we chose the best two types of sequential tests suggested by Bura and Yang (2011): the scaled version of weighted Chi-square test and the Wald-type Chi-square test. And we set the level of the tests to be 0.05 as usual. For the BIC-type criterion, we followed the suggestion in Zhu, Miao and Peng (2006) to fix the number of slices H in SDR procedures to be 25, and the tuning parameter to be $C_n = H\,(0.5\log(n) + 0.1n^{1/3})\,/\,(2\,n)$. For the bootstrap procedure, since we were not aware of any criterion in the literature to tune the cutting point that identifies small values of f_n, we picked the cutting point to be $\delta \cdot \max\{f_n(k) : 1 \le k < p\}$, and estimated d as the largest k such that $f_n(k)$ is below the cutting point. If no points of f_n are below the cutting point, we chose the estimate to be the minimum value of f_n. We let δ be 20%, 40% and 60% sequentially, so that the corresponding three estimators are able to reveal how the bootstrap procedure performs as the cutting point varies. Originally we had also used $\delta = 10\%$, but in an omitted simulation, the corresponding estimator

was found to perform universally worse than any of these three estimators. In the following we label all these estimators as "seq.-1", "seq.-2", "BIC", "YW-1", "YW-2", "YW-3" in the order above.

For a comprehensive comparison, we applied the order-determination methods to three commonly used dimension reduction procedures: SIR, SAVE and directional regression. For each of them we simulated three underlying models. We designed each model to ensure that the corresponding candidate matrix has the desired rank $d = 2$. Besides, in each model we let $p = 10$ and $X \sim N(0, I_{10})$ be independent of $\epsilon \sim N(0, 0.5^2)$. We generated 200 independent copies of random sample, each of size $n = 500$. Thus for each order-determination method, there were 200 independent estimates, among which the proportion of true estimation (i.e. the estimate equals to 2) was recorded to measure the performance of the method. Except for applying to the BIC-type criterion, we fixed the number of slices H in the dimension reduction procedures to be 10.

The 9 models are listed in the following. We apply SIR to Models 1–3, apply SAVE to Models 4–6, and apply directional regression to Models 7–9.

1. $Y = X_1 \,/\, [0.5 + (1.5 + X_2)^2] + \epsilon$;
2. $Y = \exp(X_1)\, I(\epsilon > 0) - \exp(X_2)\, I(\epsilon < 0)$;
3. $Y = X_1\,(X_1 + X_2 + 1) + \epsilon$;
4. $Y = \cos(X_1) + \cos(2\,X_2) + \epsilon$;
5. $Y = X_1^2 + X_2^2 + \epsilon$;
6. $Y = X_1^2 + (X_2 + X_3)^2 \cdot \epsilon$;
7. $Y = X_1^2 + X_2 + \epsilon$;
8. $Y = 0.4\,(X_1 + X_2 + X_3)^2 + 3\,\sin[0.25\,(X_1 + X_5 + 3\,X_6)] + 0.4\,\epsilon$;
9. $Y = 3\,\sin[0.25\,(X_1 + X_2 + X_3)] + 3\,\sin[0.25\,(X_1 + X_5 + 3\,X_6)] + 0.4\,\epsilon$.

For reference, Model 1 and Model 3 have been used in Li (1991); a similar version of Model 7 has been used in Bura and Yang (2011); and Model 8 and Model 9 have been used in Li and Wang (2007). In all the 9 models, the central subspace (Cook 1998) is 2-dimensional. In Models $1 - 3$, the monotone trend between

X and Y enables SIR to be exhaustive. In Models 4 - 6, the symmetric pattern between X and Y makes SAVE efficient and exhaustive. Since directional regression is exhaustive under fairly general conditions, in all the models the candidate matrices are of rank 2.

Besides, it is easy to see that in Model 2 and Model 5 the candidate matrices have equal nonzero eigenvalues: in Model 2, since $E(X|Y) = (\log(Y)\, I(Y > 0), \log(-Y)\, I(Y < 0), 0, \ldots, 0)^\intercal$, the nonzero eigenvalues of $M = \mathrm{var}(E(X|Y))$ are $(0.5, 0.5)$. In Model 5, for any 2-dimensional orthogonal matrix A, since $\|A\,(X_1, X_2)^\intercal\|^2 = X_1^2 + X_2^2$, and $((X_1, X_2)A^\intercal, X_3, \ldots, X_p)$ has the same marginal distribution as X, the joint distribution of $((X_1, X_2)A^\intercal, X_3, \ldots, X_p, Y)$ is invariant of A, and so is the candidate matrix M. Hence the first two eigenvectors of M lie in the same eigenspace. Equivalently, the two nonzero eigenvalues of M are equal.

In addition, the error structure are additive in all but Model 2 and Model 6. Hence these models represent a wide range of association types between X and Y.

The following table records the sample proportion of true estimation for the 7 order-determination methods, when applied with the dimension reduction procedures to these models.

Table 3.1. Comparison of order-determination methods

Models	seq.-1	seq.-2	BIC	YW-1	YW-2	YW-3	CSE
1	0.825	0.500	0.695	0.970	0.980	0.720	**1.000**
2	0.855	0.475	0.530	0.790	0.935	0.560	**1.000**
3	0.810	0.415	0.765	0.585	**0.930**	0.820	0.920
4	0.115	0.590	0.015	0.070	0.650	0.880	**0.920**
5	0.310	0.500	0.330	0.580	0.970	0.885	**1.000**
6	0.355	0.040	0.645	0.800	0.940	0.875	**0.980**
7	0.485	0.915	0.050	0.485	0.880	0.760	**1.000**
8	0.810	0.950	0.285	0.950	0.980	0.640	**1.000**
9	0.830	0.455	0.125	0.025	0.390	0.645	**0.990**

From this table, except in Model 8, the two sequential tests fail to reach their nominal level 0.05 at the current sample size, especially in Models 4–6. The BIC-type criterion provides consistent estimates in some models, but its performance

is model-dependent: in extreme cases such as Model 4 and Model 7, it constantly falsely estimates d. This indicates that the criterion is sensitive to the tuning parameter and the value suggested by Zhu, Miao and Peng (2006) is not universally appropriate. The bootstrap procedures with different cutting points tend to truly estimate d in most cases, among which the top candidate is "YW-2" with the cutting point being 40% of f_n's maximum value. However, as Model 4 and Model 9 suggest, this cutting point is not universally optimal, so a tuning parameter selection is inevitable. In contrast to all these methods, CSE consistently estimates d in all the models. In particular, compared to the second top competitor "YW-2", it performs substantially better in Model 4 and Model 9, and is comparable otherwise.

3.6 A real-data example

In this section we apply directional regression to the data set "wine" (Forina, Leardi, Armanino and Lanteri 1988), with the aid of CSE to determine the dimension of the reduced predictor. We see that directional regression visualizes the data set. From the graphics, CSE does provide a reasonable estimate of the reduced dimension.

The data set contains 178 wine samples grown in the same area but from three different cultivars. For each wine sample, its cultivar and its value in 13 chemical constituents are recorded. The goal of data analysis is to classify the cultivars based on these 13 chemical constituents, including "Alcohol", "Malic acid", "Ash", "Alkalinity of ash", "Magnesium", "Total phenols", "Flavanoids", "Nonflavanoid phenols", "Proanthocyanins", "Color intensity", "Hue", "OD280 / OD315 of diluted wines" and "Proline".

Intuitively, one would expect to find three disjoint regions in the predictor space such that each region identifies a cultivar. However, since the predictor X is 13-dimensional and it is generally difficult to tell whether three regions in $\mathbb{R}^{13}$ are disjoint, this task becomes impractical. An alternative is to first find a lower-dimensional sufficient statistic if it exists. And if the reduced dimension is less than 4, then we can visually identify these disjoint regions. A natural way to realize this is sufficient dimension reduction. Because of its favorable properties introduced in Li and Wang (2007), here we use directional regression.

To meet the linearity condition required in directional regression, we took log transformation on "Malic acid", "Color intensity" and "Proline", and took reciprocal on "Magnesium", and then standardized each component of the modified predictor. Since the response variable Y has 3 categories, the number of slices in directional regression was set to be 3. To illustrate the spectral variability in this data set, in Figure 4.1 we draw the plots as in Figure 3.1 of Section 1.

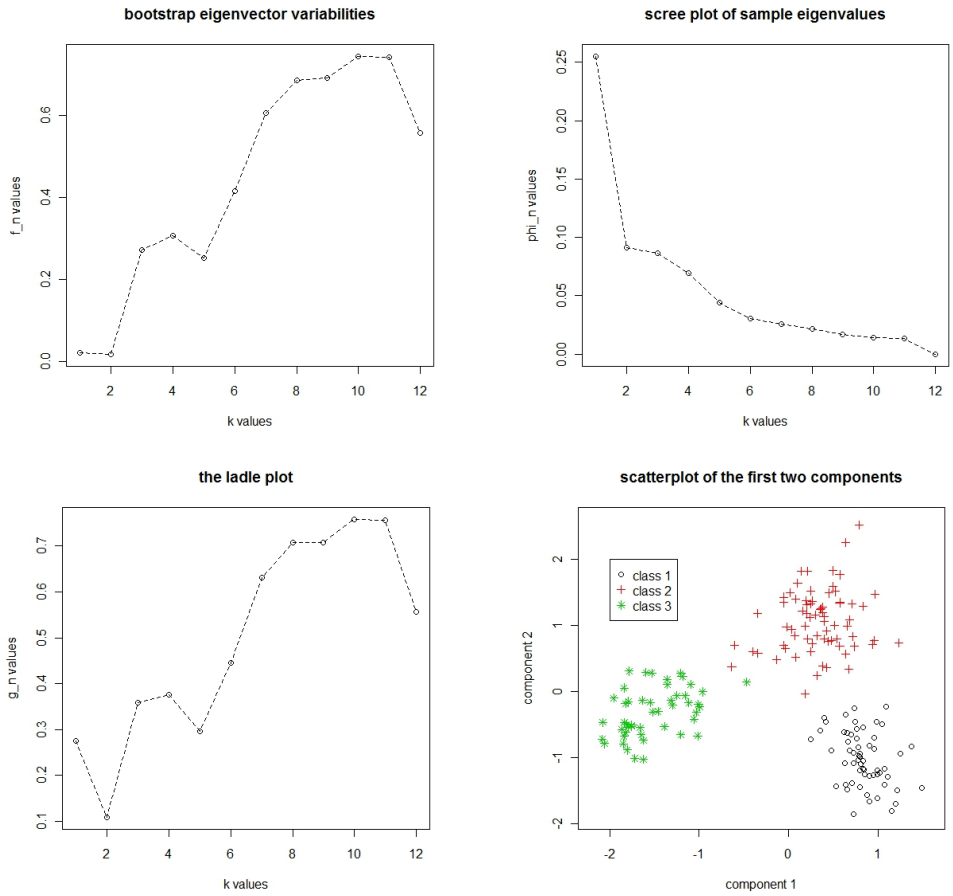

Figure 3.2. The upper-left panel shows the bootstrap eigenvector variability; the upper-right panel is the scree plot of sample eigenvalues; the lower-left panel is the ladle plot; the lower-right panel is the scatter plot of the reduced predictor indexed by cultivars.

From the plot in the upper-left panel of Figure 4.1, the bootstrap estimators "YW-1", "YW-2", "YW-3" choose 2, 5 and 6 as the estimates, while from the ladle plot in the lower-left panel, CSE chooses 2 as the estimate.

To see which estimate is reasonable, let $\hat{\beta}_1, \hat{\beta}_2$ be the first 2 directions from directional regression, and we standardized $(\hat{\beta}_1^{\intercal} X, \hat{\beta}_2^{\intercal} X)$ to have identity covariance

matrix. In the lower-right panel of Figure 4.1 is the scatter plot of $(\hat{\beta}_1^{\mathsf{T}} X, \hat{\beta}_2^{\mathsf{T}} X)$, marked with different colors and symbols according to the cultivars. In this plot, the three cultivar groups are distinguished from each other, except for one outlier. Hence $(\hat{\beta}_1^{\mathsf{T}} X, \hat{\beta}_2^{\mathsf{T}} X)$ can nearly perfectly classify the cultivars, which means that it is a linear sufficient statistic. Equivalently, the central subspace is at most 2-dimensional. On the other hand, neither $\hat{\beta}_1^{\mathsf{T}} X$ nor $\hat{\beta}_2^{\mathsf{T}} X$ are redundant: the first two cultivars are not distinct in the direction of $\hat{\beta}_1^{\mathsf{T}} X$, and none of the cultivars are distinct in the direction of $\hat{\beta}_2^{\mathsf{T}} X$. Consequently, the reasonable choice of the reduced dimension is 2, which is picked by "YW-1" and CSE.

Since the cultivars of wine are nearly perfectly identified by the reduced predictor, directional regression is a useful data-refining tool to preserve the information contained in data and enable data-visualization simultaneously.

3.7 Discussion

In this chapter we exploited the alternating pattern of eigenvalue levels and eigenvector variability, to propose a new order-determination method for sufficient dimension reduction. Previously developed order-determination methods for sufficient dimension reduction use either the level of eigenvalues or the variability of eigenvectors, but not both. The main point of this chapter is to show that much can be gained by employing the alternating pattern of these two types of variability. Based on this alternating pattern we introduced a new plot - the ladle plot - as an alternative to the scree plot, and a new estimator, the comprehensive spectral estimator.

We established the asymptotic consistency of the new estimator. Along with these results, we also gave a rigorous characterization of the alternating pattern through asymptotic analysis. Through simulation studies under a wide range of models, and through real data analysis, we demonstrated the superb performance of the new estimator.

Although in this chapter we have focused on order-determination of sufficient dimension reduction, determining the rank of a matrix-valued parameter is a fundamental problem lying at the core of many statistical procedures such as model selection, variable selection and sparse estimation. It is conceivable that the alternating pattern of eigenvalue levels and eigenvector variabilities can be used to

improve many of these methods.

3.8 Proofs of Lemmas and Corollaries

Proof of Lemma 1 By Assumption 3, we know $E(c_n^{-1}X_{n,1}) = O(1)$. Since $X_{n,i}$ is nonnegative, so is $c_n^{-1}\bar{X}_n$. By Markov's Inequality, for any $x > 0$,

$$P(c_n^{-1}\bar{X}_n > x) \leq E(c_n^{-1}\bar{X}_n)/x = E(c_n^{-1}X_{n,i})/x$$

Hence $\lim_{x\to\infty}\lim_{n\to\infty} P(c_n^{-1}\bar{X}_n > x) = 0$, which implies that $\bar{X}_n = O_P(c_n)$. □

Proof of Lemma 2 For any $j = 1, \ldots, p$, suppose $v_j(\hat{\Sigma}_n) = \sum_{i=1}^p c_{ij} v_i(\Sigma_n)$. Since $v_j(\hat{\Sigma}_n)$ is of unit length and $(v_1(\Sigma_n)), \ldots, v_p(\Sigma_n))$ is orthogonal, $c_{ij} = v_i(\Sigma_n)^\intercal v_j(\hat{\Sigma}_n)$ and $\sum_{i=1}^p c_{ij}^2 = 1$. Then

$$\Sigma_n\, v_j(\hat{\Sigma}_n) = \Sigma_n(\textstyle\sum_{i=1}^p c_{ij} v_i(\Sigma_n)) = \sum_{i=1}^p c_{ij}\lambda_i(\Sigma_n)v_i(\Sigma_n). \tag{3.18}$$

On the other hand,

$$\begin{aligned}\Sigma_n v_j(\hat{\Sigma}_n) &= \hat{\Sigma}_n v_j(\hat{\Sigma}_n) + (\Sigma_n - \hat{\Sigma}_n)v_j(\hat{\Sigma}_n) \\ &= \lambda_j(\hat{\Sigma}_n)v_j(\hat{\Sigma}_n) + (\Sigma_n - \hat{\Sigma}_n)v_j(\hat{\Sigma}_n) \end{aligned} \tag{3.19}$$

Since $\hat{\Sigma}_n - \Sigma_n = O_P(n^{-1/2})$, by Zhao, Krishnaiah and Bai (1986), $\lambda_j(\hat{\Sigma}_n) = \lambda_j(\Sigma_n) + O_P(n^{-1/2})$. Combining (3.18) and (3.19), we have:

$$\begin{aligned}\textstyle\sum_{i=1}^p c_{ij}\lambda_i(\Sigma_n)v_i(\Sigma_n) &= \lambda_j(\Sigma_n)v_j(\hat{\Sigma}_n) + O_P(n^{-1/2}) \\ &= \textstyle\sum_{i=1}^p c_{ij}\lambda_j(\Sigma_n)v_i(\Sigma_n) + O_P(n^{-1/2})\end{aligned}$$

Hence

$$\textstyle\sum_{i=1}^p c_{ij}[\lambda_i(\Sigma_n) - \lambda_j(\Sigma_n)]v_i(\Sigma_n) = O_P(n^{-1/2})$$

Since $(v_1(\Sigma_n), \ldots, v_p(\Sigma_n))$ is orthogonal, for each $i \neq j$,

$$|c_{ij}|[\lambda_i(\Sigma_n) - \lambda_j(\Sigma_n)] = O_P(n^{-1/2})$$

Since $c_{ij} = v_i(\Sigma_n)^\intercal v_j(\hat{\Sigma}_n)$, this lemma has been proved. □

Proof of Lemma 3 Let $A_1 \in \mathcal{F}_S$ be the event that

$$\{n^{1/2}(\hat{\lambda}_i - \lambda_i)/\log[\log(n)] : n \in \mathbb{N}\}$$

is a bounded sequence for each $i = 1, \ldots, p$. By the Law of the Iterated Logarithm and Zhao, Krishnaiah and Bai (1986), $P_S(A_1) = 1$. Note that for any $s \in A_1$,

$$|\hat{\lambda}_i - \hat{\lambda}_j| = |\lambda_i - \lambda_j| + o(1) \quad \text{as} \quad n \to \infty \tag{3.20}$$

for any $i, j = 1, \ldots, p$. Let $A_2 \in \mathcal{F}_S$ be the event in (3.6). then $P_S(A_2) = 1$. Hence $P_S(A_1 \cap A_2) = 1$. For any fixed $s \in A_1 \cap A_2$, let

$$\Sigma_n = \hat{M}, \quad \hat{\Sigma}_n = M^*.$$

By Lemma 2,

$$(\hat{v}_i^\intercal v_j^*)\,(\hat{\lambda}_i - \hat{\lambda}_j) = O_P(n^{-1/2}).$$

By (3.20), then, $\hat{v}_i^\intercal v_j^* = O_P(n^{-1/2})$. □

Proof of Lemma 4 Statement (a) is straightforward by the definition of $O_P^+(1)$, so it's omitted here. Note that to show (b), we only need to show that in this case $X_n > Y_n$ in probability: since for any $\{\epsilon_n : n \in \mathbb{N}\} = o(1)$, $Y_n + \epsilon_n = o_P(1)$, it indicates that $X_n > Y_n + \epsilon_n$ in probability. Then by definition $X_n - Y_n = O_P^+(1)$.

To show that $X_n > Y_n$ in probability, fix any $\omega \in (0, 1)$. For each n, define

$$\delta_{\omega,n} := \inf\{x : P(Y_n > x) \le \omega\}.$$

Since $Y_n = o_P(1)$, for any $\epsilon > 0$, $\lim P(Y_n > \epsilon) = 0$. Hence for all large n, $P(Y_n > \epsilon) \le \omega$, which means $\delta_{\omega,n} \le \epsilon$. Therefore $\delta_{\omega,n} = o(1)$. Since $X_n = O_P^+(1)$, by definition $\lim P(X_n > \delta_{\omega,n}) = 1$. Thus we have

$$\begin{aligned}
\liminf P(X_n > Y_n) &\ge \liminf P(X_n > \delta_{\omega,n}, Y_n \le \delta_{\omega,n}) \\
&= \liminf[P(X_n > \delta_{\omega,n}) - P(X_n > \delta_{\omega,n}, Y_n > \delta_{\omega,n})] \\
&\ge \liminf[P(X_n > \delta_{\omega,n}) - P(Y_n > \delta_{\omega,n})]
\end{aligned}$$

$$\begin{aligned} &= \lim P(X_n > \delta_{\omega,n}) - \limsup P(Y_n > \delta_{\omega,n}) \\ &\geq 1 - \omega. \end{aligned}$$

Let $\omega \to 0$, we have $\lim P(X_n > Y_n) = 1$, which means $X_n > Y_n$ in probability.

To show (c), by (b) we can assume $R_n = 0$ for each $n \in \mathbb{N}$. Then for any sequence $\{\epsilon_n : n \in \mathbb{N}\} = o(1)$, since $\liminf c_n > 0$, let $\epsilon = \liminf c_n/2$. Then for all large n, $c_n > \epsilon$ and $\epsilon_n < \epsilon$, therefore $\liminf P(c_n > \epsilon_n) = 1$, which implies that $X_n = c_n = O_P^+(1)$. □

Proof of Corollary 1 We will first show that condition (a) implies the result. Then we will show the equivalence between the two conditions.

Fix any $B > \epsilon$, and let $Z_{n,i} = \min\{X_{n,i}, B\}$ be the truncated version of $X_{n,i}$. Since $X_{n,i}$ is nonnegative, we have $0 \leq Z_{n,i} \leq B$. Since $B > \epsilon$, $X_{n,i} > \epsilon$ implies $Z_{n,i} > \epsilon$. Thus we still have $\liminf_{n\to\infty} P(Z_{n,1} > \epsilon) > 0$, and consequently $\liminf_{n\to\infty} E(Z_{n,1}) \geq \epsilon \liminf_{n\to\infty} P(Z_{n,1} > \epsilon) > 0$. Since $\{Z_{n,i}\}$ is uniformly bounded by B, $\{\mathrm{var}(Z_{n,i}) : n \in \mathbb{N}\}$ exists and is uniformly bounded by B^2. Therefore it is easy to see that $\bar{Z}_n = E(Z_{n,1}) + o_P(1)$. By Lemma 4 (c), $\bar{Z}_n = O_P^+(1)$. Since $\bar{X}_n \geq \bar{Z}_n$, Lemma 4 (a) indicates that $\bar{X}_n = O_P^+(1)$.

To show that (a) $\Rightarrow$ (b), denote ϵ as in (a). Note that for any $\epsilon_n \to 0$, $\epsilon_n < \epsilon$ for all large n, thus $P(X_{n,1} > \epsilon_n) \geq P(X_{n,1} > \epsilon)$ for all large n. Then (b) follows straightforwardly.

To show that (b) $\Rightarrow$ (a), assume (a) is not true. Then for each $n \in \mathbb{N}$, there exists $f(n) \in \mathbb{N}$ such that $P(X_{f(n),1} > 2^{-n}) < 2^{-n}$. Without loss of generality we assume $f(n)$ to be a monotone increasing function of n and $f(n) \to \infty$ as $n \to \infty$. Let $\{\epsilon_n : n \in \mathbb{N}\}$ be a sequence such that $\epsilon_n \to 0$ and $\epsilon_{f(n)} = 2^{-n}$. Then we have $\liminf_{n\to\infty} P(X_{n,1} > \epsilon_n) \leq \lim P(X_{f(n),1} > \epsilon_{f(n)}) = 0$, which contradicts (b). Hence (a) is true. □

Proof of Lemma 5 The proof is based on the following general proposition:

For any $k \in \mathbb{N}$ and $U = (u_1, \ldots, u_k)$ in which each $u_i \in \mathbb{R}^k$, we have

$$|\det(U)| \leq \prod_{i=1}^{k} \|u_i\|_2.$$

To see why it holds, since $\det(U)$ is multiplied by c if one of the u_i's is multiplied by c, we assume $\|u_i\|_2 = 1$ without loss of generality. Let $\lambda_1, \ldots, \lambda_k$ be the

eigenvalues of $U^\intercal U$. Then we know $\det(U^\intercal U) = \prod_i \lambda_i = \det(U)^2$. On the other hand, because $tr(U^\intercal U) = \sum_i \|u_i\|^2 = k$, we have $\sum_i \lambda_i = k$. Then $|\det(U)| \leq 1$ is implied by the inequality of arithmetic and geometric means.

Now for each $j \leq k$, since $\sum_{i=1}^p a_{i,j}^2 = 1$, we have $\sum_{i=1}^k a_{i,j}^2 \leq 1$. Especially, when $j = k$, we have $\sum_{i=1}^k a_{i,k}^2 \leq 1 - a_{k+1,k}^2$. A direct application of the proposition above implies that $|\det(A_k)| \leq (1 - a_{k+1,k}^2)^{1/2}$. Hence

$$1 - |\det(A_k)| \geq 1 - (1 - a_{k+1,k}^2)^{1/2} \geq a_{k+1,k}^2 \,/\, (1 + (1 - a_{k+1,k}^2)^{1/2}) \geq a_{k+1,k}^2 \,/\, 2$$

□

Proof of Lemma 6 Let A_1 be the same in the proof of Lemma 3. Then by the Law of Iterated Logarithm and Zhao, Krishnaiah and Bai (1986), $P_S(A_1) = 1$. Conditional on any $s \in A_1$, for each $k < p$, $\hat{\lambda}_k = \lambda_k + O(c_n^{-1/2} \cdot n^{-1/2})$, which implies that

$$\phi_n(k) = \frac{\lambda_{k+1} + O(c_n^{-1/2} \cdot n^{-1/2})}{\sum_{i=1}^p \lambda_i + O(c_n^{-1/2} \cdot n^{-1/2})} = \frac{\lambda_{k+1}}{\sum_{i=1}^p \lambda_i} + O(c_n^{-1/2} \cdot n^{-1/2}).$$

Denote the first term on the right hand side by $\phi_0(k)$, then $\lim_{n\to\infty} \phi_n(k) = \phi_0(k)$. If $k < d$, then $\lambda_{k+1} > 0$, which means $\phi_0(k) > 0$. Then $\phi_n(k) = O_P^+(1)$ by Lemma 4 (c).
If $k \geq d$, then $\lambda_{k+1} = 0$, which means $\phi_0(k) = 0$. Thus

$$\phi_n(k) = O(c_n^{-1/2} \cdot n^{-1/2}) = O_P(c_n^{-1/2} \cdot n^{-1/2}).$$

□

Chapter 4

Order-determination for dimension reduction using augmentation predictors

4.1 Introduction

In this chapter, we will continue discussing the order-determination problem, and we will address it with a new approach. The resulting estimator is designed for inverse regression SDR methods, and it shares the same framework as CSE combining the information from both eigenvalues and eigenvectors of the sample candidate matrices. Therefore it enjoys the same advantages of CSE when compared to other order-determination methods such as the sequential tests and BIC-type criterion. Rather than measuring the bootstrap variability in CSE, in this approach we extract information from eigenvectors with the aid of an augmentation predictor. We call the new estimator the augmented spectral estimator (ASE). With the details introduced later, one can clearly see that ASE is computationally much more efficient than CSE. On the other hand, compared to the literature of using the augmentation predictor in variable selection, ASE employs augmentation in a novel way, as it naturally incorporates it into the SDR assumptions. Under fairly general conditions, we will show the asymptotic consistency of ASE. We will further show its efficacy in finite sample cases by simulation. At the end of this chapter, we will

discuss its potential generalization in multiple directions.

To begin with, suppose X is the p-dimensional predictor and Y is the response variable in the data. An augmentation predictor is a random variable that is generated independently of (X, Y) and then merged with X to form a new predictor. In the literature, its usage has been noticed by multiple authors in stochastic modeling (Gammaitoni, Hanggi, Jung and Marchesoni 1998) and variable selection (Miller 2002; Wu, Boos and Stefanski 2007). It has also been used to construct the permutation tests in SDR (Cook and Weisberg 1991; Cook and Yin 2001). Typically in these methods, the key assumption is that the augmentation predictor behaves in the same way as the "inactive" components of X in the corresponding procedures, thus it provides relative information about these components of X. For example, in Wu, Boos and Stefanski (2007), the number of falsely selected components of the augmentation predictor is used to estimate the False Selection Rate; in other methods such as the permutation tests, the augmentation predictor is used to simulate the baseline distribution that distinguishes significant signals from noise. Since the augmentation predictor is always recognizable as the noise, in real data analysis it provides oracle information as if in simulation. Therefore these methods are efficient in practice.

Nonetheless, in general there is a lack of theory to support the aforementioned key assumption, which makes these methods suffer from being heuristic. In contrast, in this chapter we will verify this assumption in ASE under slightly extra conditions. Furthermore, due to the unique way the augmentation predictor is employed, we will also show the consistency of ASE under fairly general conditions where this assumption doesn't hold.

The motivation for ASE can be explained as follows. Suppose the columns of the candidate matrix M lie in the central dimension reduction subspace, and $\text{rank}(M) = d$. To estimate d, a simple case will arise if some components of X are known to be independent of Y conditional on all the other components. In that case, it is easy to see that all the vectors in the central subspace take 0 on these components, and so do all the vectors in the central dimension reduction subspace, and all the d eigenvectors of M with nonzero eigenvalues. Suppose $\hat{M}$ is a consistent estimator of M, then if $k \leq d$, the k^{th} eigenvector of $\hat{M}$ will take negligible values on these components; if $k > d$, plausibly the k^{th} eigenvector of $\hat{M}$ will take non-negligible values on these components. Thus an appropriate

measure of these components will be a function of k that is small when $k \leq d$ and large when $k > d$.

In practice, these components of X are not available due to our limited knowledge about the data. However, by replacing X with the merged predictor of X and the augmentation predictor, these components become available as the augmentation predictor plays this role. Therefore from the discussion above, we have the following conjecture:

> An appropriate measure of the components of sample eigenvectors, corresponding to the augmentation predictor, is small when $k \leq d$ and is large when $k > d$.

Note that this pattern is exactly alternating with the sample eigenvalue of M which is large when $k \leq d$ and small when $k > d$. Therefore, similar to CSE, based on these alternating patterns we can construct a new objective function that is likely to be minimized at d. The corresponding minimizer is thus called the augmented spectral estimator (ASE).

In the rest of the chapter, we will discuss the consistency of ASE for the central subspace and normally distributed X in Section 2, and we will generalize it to any central dimension reduction subspace and elliptically distributed X in Section 3. A simulation study will be presented in Section 4. We will discuss potential generalizations in Section 5.

4.2 Consistency of ASE in a special case

We start to discuss the consistency of ASE with the special case where the SDR parameter is the central subspace and X is multivariate normally distributed. As one will see, in this case it requires the least for ASE to be consistent, due to the validity of the aforementioned assumption, which again states the identical behavior of the augmentation predictor and the "inactive" linear combinations of X. This assumption is revealed in an invariance property that naturally leads to the conjecture stated above. In the next section, when X is more generally elliptically distributed and the SDR parameter is any central dimension reduction subspace,

this invariance property no longer holds, which requires us to invoke additional assumptions.

To set up the notation, we write $Z \overset{\mathcal{D}}{=} W$ when two random variables Z and W have the same distribution. As in the previous chapter, we denote $X_n = O_P^+(1)$ if a sequence of random variables $\{X_n : n \in \mathbb{N}\}$ is "bounded above from 0", that is,

$$\lim_{n\to\infty} P(X_n > c_n) = 1$$

for any constant sequence $c_n = \mathrm{o}(1)$. Two important properties of $O_P^+(1)$ are restated in the following lemma. The proof can be found in the previous chapter.

Lemma 7. *1. For any two sequences of random variables $\{X_n : n \in \mathbb{N}\}$ and $\{Z_n : n \in \mathbb{N}\}$, if $X_n = O_P^+(1)$ and $Z_n \geq X_n$ for each $n \in \mathbb{N}$, then $Z_n = O_P^+(1)$.*

2. For any two sequences of random variables $X_n = O_P^+(1)$ and $Y_n = o_P(1)$, we have

$$\lim_{n\to\infty} P(X_n > Y_n) = 1.$$

Again these properties justify $O_P^+(1)$ as defining a sequence of "large" random variables.

Without loss of generality we suppose that X has been standardized so that $X \sim N(0_p, I_p)$. Since the support of X is $\mathbb{R}^p$, the central subspace of $Y|X$ does exist (Cook 1998). Suppose this space is d-dimensional with $d < p$. Suppose $X_a \sim N(0_q, I_q)$ is a q-dimensional augmentation predictor and $X_a \perp\!\!\!\perp (X, Y)$. Denote X^* as $(X^\intercal, X_a^\intercal)^\intercal$. Then we have $X^* \sim N(0_{p+q}, I_{p+q})$. Note that the central subspace of $Y|X^*$ exists in $\mathbb{R}^{p+q}$ and is also d-dimensional. Denote β as a $(p+q) \times d$ matrix whose columns form an orthonormal basis of this space, and γ as a $(p+q) \times (p+q-d)$ matrix whose columns form an orthonormal basis of span$(\beta)^\perp$. Note that neither β or γ are observable from the data. However, since $X_a \perp\!\!\!\perp (X, Y)$, the last q rows of β are 0. On the other hand, by definition $\beta^\intercal X^*$ contains all the information in X^* about Y, which indicates an invariance property presented in the following lemma.

Lemma 8. *Invariance Property: Suppose* $X^* \sim N(0_{p+q}, I_{p+q})$ *and* A *is a* $(p+q-d)$*-dimensional orthogonal matrix. We have*

$$(\beta^\intercal X^*, \gamma^\intercal X^*, Y) \overset{\mathcal{D}}{=} (\beta^\intercal X^*, A\gamma^\intercal X^*, Y). \tag{4.1}$$

PROOF. By definition, (β, γ) is an orthogonal matrix. Since $X^* \sim N(0_{p+q}, I_{p+q})$, we have $\gamma^\intercal X^* \sim N(0_{p+q-d}, I_{p+q-d})$ and $\gamma^\intercal X^* \perp\!\!\!\perp \beta^\intercal X^*$. Also by definition, we have $\gamma^\intercal X^* \perp\!\!\!\perp Y \,|\, \beta^\intercal X^*$. Hence $\gamma^\intercal X^* \perp\!\!\!\perp (\beta^\intercal X^*, Y)$. Since A is orthogonal, we have $A\gamma^\intercal X^* \perp\!\!\!\perp (\beta^\intercal X^*, Y)$ and $A\gamma^\intercal X^* \sim N(0_{p+q-d}, I_{p+q-d})$. Hence

$$(\beta^\intercal X^*, \gamma^\intercal X^*, Y) \overset{\mathcal{D}}{=} (I, II)$$

in which $I \overset{\mathcal{D}}{=} (\beta^\intercal X^*, Y)$ is independent of $II \overset{\mathcal{D}}{=} \gamma^\intercal X^*$. Meanwhile,

$$(\beta^\intercal X^*, A\gamma^\intercal X^*, Y) \overset{\mathcal{D}}{=} (I, III)$$

in which $III \overset{\mathcal{D}}{=} A\gamma^\intercal X^*$ is independent of I. Since $II \overset{\mathcal{D}}{=} III$, we have

$$(I, II) \overset{\mathcal{D}}{=} (I, III).$$

Hence (4.1) is true. □

Note that there are two nontrivial constraints in this property: the SDR parameter is the central subspace and X^* is multivariate normally distributed. It is easy to see that the first constraint cannot be removed, that is, this invariance property no longer holds for any other central dimension reduction subspace. However, theoretically the second constraint can be relaxed, as this property still holds if X^* is more generally elliptically distributed. Nonetheless, unless the original predictor X is normally distributed, it is generally impractical to find an appropriate X_a that makes X^* elliptically distributed. On the other hand, when X is normally distributed, any normal X_a will make X^* normal. Therefore we still restrict X^* to be normally distributed here as a practical issue.

A direct application of the invariance property implies the following theorem:

Theorem 4. *Suppose* $X^* \sim N(0_{p+q}, I_{p+q})$ *and the columns of the candidate matrix* M *span the central subspace* $\mathcal{S}_{Y|X^*}$*. Suppose rank*$(M) = d$ *and* $\hat{M}$ *is a consistent estimator of* M*. For* $k = 1, \ldots, p$*, denote* $\hat{\beta}_k = (\hat{\beta}_{k,1}, \ldots, \hat{\beta}_{k,p+q})^\intercal$ *as*

the eigenvector of $\hat{M}$ corresponding to its k^{th} largest eigenvalue. Then

$$\|(\hat{\beta}_{k,p+1},\ldots,\hat{\beta}_{k,p+q})\|_2 = \begin{cases} o_P(1) & \text{if } k \le d \\ O_P^+(1) & \text{if } k > d \end{cases}$$

The choice of L_2 norm here is arbitrary - this theorem still holds if other equivalent norm such as L_1 norm or L_∞ norm is used. This flexibility exists in all the results in the chapter. Therefore we set it as default and will not point it out again.

PROOF. Since $X_a \perp\!\!\!\perp (X,Y)$, each vector in $\mathcal{S}_{Y|X^*}$ has 0 on its last q components. Since the columns of M span $\mathcal{S}_{Y|X^*}$, its eigenvectors that correspond to nonzero eigenvalues also have 0 on the last q components. Since $\hat{M}$ is a consistent estimator of M, for any $k \le d$ and $i = 1,\ldots,q$, we have $\hat{\beta}_{k,p+i} = o_P(1)$. Consequently,

$$\|(\hat{\beta}_{k,p+1},\ldots,\hat{\beta}_{k,p+q})\|_2 = o_P(1).$$

For any matrix B, denote $P_B = B(B^\intercal B)^{-1}B^\intercal$ as the projection matrix of the space spanned by columns of B. Then $P_{(\hat{\beta}_1,\ldots,\hat{\beta}_d)} \xrightarrow{\mathcal{P}} P_\beta$. Since for any $k > d$, we have $(\hat{\beta}_1,\ldots,\hat{\beta}_d)^\intercal\hat{\beta}_k = 0$, it indicates that $P_\beta\hat{\beta}_k = o_P(1)$. For any $(p+q-d)$-dimensional orthogonal matrix A, denote $\hat{M}_A$ as the estimator of the candidate matrix with the predictor $(\beta^\intercal X^*, A\gamma^\intercal X^*)$ (though it is unobservable). Then $\hat{M}_A = (\beta,\gamma A^\intercal)^\intercal\hat{M}(\beta,\gamma A^\intercal)$. Since $\hat{M}_A$ is a measurable function of $(\beta^\intercal X, A\gamma^\intercal X, Y)$ and the data are i.i.d, by Lemma 8 we have

$$\hat{M}_A \overset{\mathcal{D}}{=} \hat{M}_{I_{p+q-d}} \tag{4.2}$$

Denote $\hat{\beta}_{k,A}$ as the eigenvector of $\hat{M}_A$ corresponding to its k^{th} largest eigenvalue. Since $\hat{M}_A = (\beta,\gamma A^\intercal)^\intercal\hat{M}(\beta,\gamma A^\intercal)$, we have $A\gamma^\intercal\hat{\beta}_k = (\hat{\beta}_{k,A,d+1},\ldots,\hat{\beta}_{k,A,p+q})$. By (4.2), the distribution of $\hat{\beta}_{k,A}$ is invariant of orthogonal matrix A, as does the distribution of $A\gamma^\intercal\hat{\beta}_k$. Hence $\gamma^\intercal\hat{\beta}_k$ has an elliptical distribution. Since $P_\beta\hat{\beta}_k = o_P(1)$ and $\|\hat{\beta}_k\|_2 = 1$, $\|\gamma^\intercal\hat{\beta}_k\|_2 \xrightarrow{\mathcal{P}} 1$. Thus the asymptotic distribution of $\gamma^\intercal\hat{\beta}_k$ is the uniform distribution on the unit sphere in $\mathbb{R}^{p+q-d}$. Since each vector in $\mathcal{S}_{Y|X^*}$ has 0 on the last q components, $(e_{p+1},\ldots,e_{p+q})$ is orthogonal to $\mathcal{S}_{Y|X^*}$, which means $(e_{p+1},\ldots,e_{p+q}) \in \text{span}(\gamma)$. Suppose $\gamma = (e_{p+1},\ldots,e_{p+q},v)$ where v is an appropriate matrix. Then asymptotically $(\hat{\beta}_{k,p+1},\ldots,\hat{\beta}_{k,p+q})$ is distributed as the first q components of the uniform distribution on the unit sphere in $\mathbb{R}^{p+q-d}$. It is

then straightforward to see that $\|(\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})\|_2 = O_P^+(1)$. □

Based on this theorem, if we plot $\|(\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})\|_2$ as k varies, then for large samples it approaches 0 when $k \leq d$, whereas it deviates from 0 when $k > d$. Consequently, d can be characterized as the largest k at which this curve stays flat and close to 0. The following Figure illustrates this curve under Model IVin the simulation study, where X is normally distributed and directional regression is used in the SDR procedure, and the central subspace is 2-dimensional.

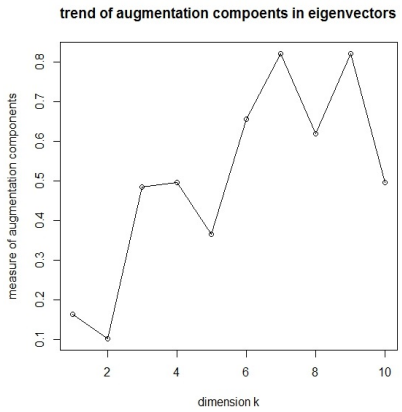

Figure 4.1. The trend of augmentation components in the sample eigenvectors

Starting from this point there are various approaches to estimating d, among which the most common strategy is to construct an objective function that tends to be minimized at d. In this spirit, it is desirable to find a conjugate function of $\|(\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})\|_2$ with the alternating pattern - it takes large values when $k < d$ and small values when $k \geq d$ - so that the objective function can be easily constructed as the sum of this pair of functions. Fortunately, ϕ_n defined in the last chapter can serve as this conjugate. Recall that ϕ_n is defined as

$$\phi_n : \{1, \ldots, p-1\} \to \mathbb{R}, \quad \phi_n(k) = \hat{\lambda}_{k+1} / (\sum_{i=1}^{p} \hat{\lambda}_i). \tag{4.3}$$

in which $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ are the eigenvalues of $\hat{M}$ in descending order. Then as Lemma 6 in the last chapter points out, the consistency of $\hat{M}$ implies that $\phi_n(k) = O_P^+(1)$ when $k < d$ and $\phi_n(k) = o_P(1)$ when $k \geq d$, which is exactly desirable. For ease of reading, we present this lemma again, though the proof is omitted.

Lemma 9. *Suppose M is a $p \times p$ positive semi-definite matrix with rank $(M) = d$, and $\hat{M}$ is a consistent estimator of M. Then ϕ_n defined in (4.3) satisfies,*

$$\phi_n(k) = \begin{cases} O_P^+(1) & \text{if } k < d \\ o_P(1) & \text{if } k \geq d \end{cases}$$

Again, this lemma holds as long as $\hat{M}$ is a consistent estimator of M. In particular, it doesn't require X to be normally distributed or the SDR parameter to be the central subspace. Therefore it still holds in the next section where these assumptions are relaxed.

Based on all the discussions above, it is now straightforward to see that the sum of the two functions introduced above tends to be minimized at d, which implies the consistency of ASE. This result is formally presented in the following theorem.

Theorem 5. *Suppose $X^* \sim N(0_{p+q}, I_{p+q})$ and the columns of the candidate matrix M span the central subspace $\mathcal{S}_{Y|X^*}$. Suppose rank$(M) = d$ and $\hat{M}$ is a consistent estimator of M. Define $h_n : \{1, \ldots, p-1\} \to \mathbb{R}$ as:*

$$h_n(k) = \|(\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})\|_2 + \phi_n(k). \tag{4.4}$$

Then h_n is minimized at d with probability going to 1.

PROOF. For simplicity in notation we let $l_n : \{1, \ldots, p-1\} \to \mathbb{R}$ be

$$l_n(k) = \|(\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})\|_2.$$

Then $h_n = l_n + \phi_n$. Theorem 5 and Lemma 9 imply that

$$\begin{cases} l_n(k) = o_P(1),\ \phi_n(k) = O_P^+(1) & \text{if } k < d, \\ l_n(k) = o_P(1),\ \phi_n(k) = o_P(1) & \text{if } k = d, \\ l_n(k) = O_P^+(1),\ \phi_n(k) = o_P(1) & \text{if } k > d. \end{cases}$$

Since both l_n and ϕ_n are nonnegative and $h_n = l_n + \phi_n$, $h_n \geq \max\{l_n, \phi_n\}$. By Lemma 7.1, we have $h_n(k) = O_P^+(1)$ when $k \neq d$ and $h_n(d) = o_P(1)$. By Lemma 7.2, for each $k \neq d$, $h_n(k) > h_n(d)$ with probability going to 1. Hence

$$\lim_{n\to\infty} P(\arg\min_{k=1,\ldots,p-1} h_n(k) = d) = 1.$$

□

As this theorem shows, ASE is a consistent order-determination method. Compared to the sequential tests, it doesn't involve asymptotic derivation, so it is more convenient to use when applied to new inverse regression SDR methods; compared to BIC-type criterion, it doesn't involve tuning parameters, so it is easier to use as no tuning parameter selection is needed; compared to CSE, it doesn't involve bootstrap re-sampling, so it is much more efficient in computation, especially when the sample size is large. Like CSE, ASE fully uses the intrinsic information from the matrices. Thus we expect it to be efficient in finite sample cases.

So far, we have shown the consistency of ASE and discussed its theoretical advantages compared to other order-determination methods, under the constraints that X is normally distributed and the SDR parameter is the central subspace. These constraints substantially limit the application of ASE. Hence in the next section, we will discuss the consistency of ASE in more general settings. Fortunately, it holds with few extra assumptions.

4.3 Consistency of ASE under the general setting

In this section we relax the SDR parameter to be a general central dimension reduction subspace, and we assume X to be elliptically distributed. Note that this setting includes the last section as a special case. Since the ellipticity assumption is commonly adopted to ensure the consistency of inverse regression SDR methods, this setting is most general so that it literally doesn't invoke any constraint on the application of ASE.

As a generalization of multivariate normal distribution, elliptical distribution doesn't preserve all the desired properties. One example mentioned before is that the family of multivariate normal distributions is closed under the operation of merging two independent random variables. Namely, if both of the two independent random variables are normally distributed, then so is their joint distribution. However, this property no longer holds for the family of elliptical distribution. That is, for any two random variables that are independent and elliptically distributed, their joint distribution may not be elliptically distributed. This failure does affect the generalization of ASE here, in the way that when the original predictor X is elliptically distributed, it is impractical to find an appropriate X_a without

extra assumptions such that the merged predictor X^* is still elliptically distributed. Fortunately, it will not affect the consistency of inverse regression SDR methods when applied to the merged data (X^*, Y). Furthermore, under negligible extra assumptions, ASE is still consistent. More details are discussed in the following.

As in the last chapter, we still generate X_a from a q-dimensional standard normal distribution $N(0_q, I_q)$ independently of (X, Y), and merge it with X to get X^*. Again note that X^* may not have an elliptical distribution. However, since the support of X^* is $\mathbb{R}^{p+q}$, the central dimension reduction subspace for (X^*, Y) still exists (Cook 1998, Cook and Li 2002, Zhu and Zhu 2009). Let (β, γ) be an orthogonal matrix in which β forms an orthonormal basis of the central dimension reduction subspace. Since $X_a \perp\!\!\!\perp (X, Y)$, β is $p \times d$-dimensional and the last q rows of β are 0. Let β_1 be the first p rows of β, then $\beta^\intercal X^* = \beta_1^\intercal X$ and consequently $E(X^* \,|\, \beta^\intercal X^*) = (E(X^\intercal \,|\, \beta_1^\intercal X), \mathbf{0})^\intercal = P_\beta^\intercal X^*$. Therefore although the ellipticity assumption (2.5) no longer holds on X^*, the weaker assumption (2.4) does, which ensures the consistency of inverse regression SDR methods when applied to (X^*, Y).

As seen in the last section, the pattern of sample eigenvalues is a direct result of the consistency of $\hat{M}$. However, the pattern of sample eigenvectors is implied by invariance property, which doesn't hold in the general setting here. This property plays a key role in Theorem 5, as it guarantees that the augmentation predictor X_a is well "mixed" with the irrelevant components $\gamma^\intercal X$ in the original predictor. Fortunately, it is not a necessary condition for the consistency of ASE, since the "mixture" between X_a and $\gamma^\intercal X$ can also be induced by the extra assumption of asymptotic normality of $\hat{M}$:

Assumption 4. *For any symmetric matrix B, let vech $\circ$ B be the vectorization of its upper triangular matrix. Then*

$$\sqrt{n}\,\textit{vech} \circ (\hat{M} - M) \xrightarrow{\mathcal{D}} N(0, \Sigma), \tag{4.5}$$

in which Σ is a $(p+q)\cdot(p+q+1)/2$-dimensional invertible matrix.

Usually $\hat{M}$ is constructed by sample moments, thus this assumption holds under general regularity conditions. In fact, it has been commonly adopted in the literature, for example in the sequential tests and CSE.

Under this assumption, the following theorem reveals the same behavior of sample eigenvectors as in Theorem 4 in the general setting.

Theorem 6. *Suppose $X \in \mathbb{R}^p$ is elliptically distributed and $X_a \sim N(0_q, I_q)$ is independent of (X, Y). Suppose M is a positive semi-definite candidate matrix for (X^*, Y) with rank $(M) = d$, and $\hat{M}$ is an estimator of M satisfying Assumption 4, then*

$$\|(\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})\|_2 = \begin{cases} o_P(1) & \text{if } k \leq d \\ O_P^+(1) & \text{if } k > d \end{cases}$$

PROOF. First, we transform X^* linearly to simplify the notation. Suppose e_i is the $(p+q)$-dimensional unit vector that takes 1 on the i^{th} component, then we know that $e_i \in \text{span}(\gamma)$ for any $i > p$. Let v be an appropriate matrix such that $\gamma = (v, e_{p+1}, \ldots, e_{p+q})$. Denote A as $(\beta, v, e_{p+1}, \ldots, e_{p+q})^\intercal$, and we transform X^* into AX^*. Consequently, M is transformed to $AMA^\intercal$ that is equal to

$$\begin{pmatrix} \Lambda_d & 0 & 0 \\ 0 & 0_{(p-d)\times(p-d)} & 0 \\ 0 & 0 & 0_{q\times q} \end{pmatrix}. \tag{4.6}$$

in which Λ_d is a $d \times d$ positive definite diagonal matrix. Since A is orthogonal, $A^\intercal A = I_{p+q}$. For $k = 1, \ldots, p-1$, we have $A\hat{M}A^\intercal A\hat{\beta}_k = \hat{\lambda}_k A\hat{\beta}_k$. Thus $A\hat{\beta}_k$ is the eigenvector of $A\hat{M}A^\intercal$ corresponding to the k^{th} largest eigenvalue. Since $((A\hat{\beta}_k)_{p+1}, \ldots, (A\hat{\beta}_k)_{p+q}) = (\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})$, the last q components of $\hat{\beta}_k$ are invariant under this linear transformation of X^*. Hence without loss of generality, we can assume that M is of the form (4.6). Accordingly, we write $\hat{M}$ and $\hat{\beta}_k$ as

$$\begin{pmatrix} \hat{M}_{11} & \hat{M}_{12} & \hat{M}_{13} \\ \hat{M}_{21} & \hat{M}_{22} & \hat{M}_{23} \\ \hat{M}_{31} & \hat{M}_{32} & \hat{M}_{33} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \hat{\beta}_{k1} \\ \hat{\beta}_{k2} \\ \hat{\beta}_{k3} \end{pmatrix}.$$

Note that $\hat{\beta}_{k3} = (\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})^\intercal$. To simplify the notation, we denote $\|\cdot\|_2$ by $\|\cdot\|$.

When $k \leq d$, since $e_{p+i} \in \text{span}(\beta)^\perp$ for $i = 1, \ldots, q$, the consistency of $\hat{M}$ implies $\|\hat{\beta}_{k3}\| = o_P(1)$.

When $d < k < p$, since $\hat{\beta}_k$ is the eigenvector of $\sqrt{n}\hat{M}$ corresponding to $\hat{\lambda}_k$, we have

$$\sqrt{n}\hat{M}_{21}\hat{\beta}_{k1} + \sqrt{n}\hat{M}_{22}\hat{\beta}_{k2} + \sqrt{n}\hat{M}_{23}\hat{\beta}_{k3} = \sqrt{n}\hat{\lambda}_k\hat{\beta}_{k2} \tag{4.7}$$

and

$$\sqrt{n}\hat{M}_{31}\hat{\beta}_{k1} + \sqrt{n}\hat{M}_{32}\hat{\beta}_{k2} + \sqrt{n}\hat{M}_{33}\hat{\beta}_{k3} = \sqrt{n}\hat{\lambda}_k\hat{\beta}_{k3} \tag{4.8}$$

By Assumption 4 and Zhao, Krishnaiah and Bai (1986), all the terms in the above two equations are $O_P(1)$. Suppose $E_\delta = (\|\hat{\beta}_{k1}\| < \delta) \cap (\|\hat{\beta}_{k3}\| < \delta)$ with $\delta > 0$, then as $\delta \to 0$, $E_\delta \to E_0 := (\|\hat{\beta}_{k1}\| = 0) \cap (\|\hat{\beta}_{k3}\| = 0)$. Denote P_n as the probability measure of $\sqrt{n}\,\text{vech} \circ (\hat{M} - M)$ when the sample size is n, and P_0 as the limiting probability measure of $\sqrt{n}\,\text{vech} \circ (\hat{M} - M)$ in Assumption 4. When E_0 occurs, equations (4.7) and (4.8) reduce to

$$\sqrt{n}\hat{M}_{22}\hat{\beta}_{k2} = \sqrt{n}\hat{\lambda}_k\hat{\beta}_{k2} \tag{4.9}$$

and

$$\sqrt{n}\hat{M}_{32}\hat{\beta}_{k2} = 0. \tag{4.10}$$

Note that E_0 implies that $\|\hat{\beta}_{k2}\| = 1$. If we denote $\hat{\alpha}_1, \ldots, \hat{\alpha}_{p-d}$ as the columns of the V-matrix in the singular value decomposition of $\sqrt{n}\hat{M}_{32}$, then (4.9) and (4.10) imply $(\cup_{i=1}^{p-d} A_{n,i})$ in which $A_{n,i}$ is $\{\sqrt{n}\hat{M}_{22}\hat{\alpha}_i = \sqrt{n}(\hat{\alpha}_i^{\mathsf{T}}\hat{M}_{22}\hat{\alpha}_i)\hat{\alpha}_i\}$. By Assumption 4, we know that almost surely $\sqrt{n}\hat{M}_{32}$,

$$\sqrt{n}\,\text{vech} \circ \hat{M}_{22} \,|\, \sqrt{n}\hat{M}_{32} \xrightarrow{\mathcal{D}} N(\sqrt{n}\,V \cdot \text{vec} \circ \hat{M}_{32},\, \Omega)$$

in which vec$\circ$ is the vectorization of a matrix and Ω is positive definite. Thus the asymptotic distribution of $\sqrt{n}\,\text{vech} \circ \hat{M}_{22} \,|\, \sqrt{n}\hat{M}_{32}$ is absolutely continuous with respect to Lebesque measure μ on $\mathbb{R}^{(p-d+1)\times(p-d)/2}$. Since $\hat{\alpha}_1, \ldots, \hat{\alpha}_{p-d}$ are fixed given $\sqrt{n}\hat{M}_{32}$, we have $\mu(\cup_{i=1}^{p-d} A_{n,i}) = 0$, which indicates that

$$P_0(\cup_{i=1}^{p-d} A_{n,i} \,|\, \sqrt{n}\hat{M}_{32}) = 0$$

almost surely $\sqrt{n}\hat{M}_{32}$. By integrating this function of $\sqrt{n}\hat{M}_{32}$, we have

$$P_0(\cup_{i=1}^{p-d} A_{n,i}) = 0.$$

Hence $P_0(E_0) = 0$, which is equivalent to that

$$\lim_{\delta \to 0} P_0(E_\delta) = 0. \tag{4.11}$$

Note that for each $\delta > 0$, P_0 is continuous on the boundary of E_δ. Thus

$$\lim_{n \to \infty} P_n(E_\delta) = P_0(E_\delta). \tag{4.12}$$

Also, since M is of the form (4.6), $(e_1, \ldots, e_d)$ span the central subspace. Thus the consistency of $\hat{M}$ implies that $\hat{\beta}_{k1} = o_P(1)$, which means that $\lim_{n\to\infty} P(\|\hat{\beta}_{k1}\| < \delta) = 1$ for any $\delta > 0$. Hence we have

$$\lim_{n \to \infty} P_n(E_\delta) = \lim_{n \to \infty} P_n(\|\hat{\beta}_{k3}\| < \delta),$$

which combined with (4.11) and (4.12) indicates that

$$\lim_{\delta \to 0} \lim_{n \to \infty} P(\|\hat{\beta}_{k3}\| < \delta) = 0. \tag{4.13}$$

Fix any constant sequence $\delta_n \to 0$, for any $\delta > 0$ there exists N such that $\delta_n < \delta$ for any $n \geq N$. Thus the event $\|\hat{\beta}_{k3}\| < \delta_n$ implies the event $\|\hat{\beta}_{k3}\| < \delta$, which means

$$\limsup_{n \to \infty} P(\|\hat{\beta}_{k3}\| < \delta_n) \leq \lim_{n \to \infty} P(\|\hat{\beta}_{k3}\| < \delta).$$

Let $\delta \to 0$, we have $\lim_{n\to\infty} P(\|\hat{\beta}_{k3}\| < \delta_n) = 0$ implied by (4.13). Hence $\|\hat{\beta}_{k3}\| = O_P^+(1)$. □

From this theorem and the discussion in the last section, it is straightforward to see the consistency of ASE for any central dimension reduction subspace and elliptically distributed X, as presented in the next theorem. The proof is exactly the same as in Theorem 5, and so is omitted.

Theorem 7. *Suppose $X \in \mathbb{R}^p$ is elliptically distributed and $X_a \sim N(0_q, I_q)$ is independent of (X, Y). Suppose M is a positive semi-definite candidate matrix for*

(X^*, Y) with rank $(M) = d$, and $\hat{M}$ is an estimator of M satisfying Assumption 4. Then h_n defined in (4.4) is minimized at d with probability going to 1.

As this theorem suggests, ASE is consistent for estimating the dimension of any central dimension reduction subspace when X is elliptically distributed. Again in this general setting, the invariance property in Lemma 8 doesn't hold. Therefore, in ASE it is not necessary to assume the identical behavior of the augmentation predictor and the "inactive" components of the original predictor. The underlying reason is that we employ the augmentation predictor to indicate special pattern of eigenvectors rather than simulating the baseline distribution. This is the fundamental difference between ASE and other statistical methods in the literature that employ augmentation predictors.

Since asymptotic normality is the only assumption needed to ensure the consistency of ASE, ASE has the same general application as the sequential tests and CSE. We have discussed the theoretical advantage of ASE in the end of the last section. Next we will explore its behavior in finite sample cases.

4.4 Simulation study

In this section, we conduct a simulation study to investigate the behavior of ASE in finite sample cases, compared to other order-determination methods. To be concise, we will only present part of the results in this dissertation.

Note that there are two directions in which ASE can vary: first, as mentioned after Theorem 4, different norms can be employed to measure the magnitude of $(\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})$; second, the augmentation predictor X_a can have different dimensions q. Intuitively, q should vary with n and p, and should be reasonably large so that the pattern of $\|(\hat{\beta}_{k,p+1}, \ldots, \hat{\beta}_{k,p+q})\|$ is stable while the sample size n still dominates $p + q$ so that the inverse regression SDR methods remain trustworthy. For the purpose of illustration, in this section we arbitrarily choose L_2 norm and $q = p - 1$. Further investigation on the optimal norm and q may be deserved, and other potential improvements of ASE, such as adding augmentation predictors repeatedly and taking the average, may also be of interest.

As in the last chapter, we will compare ASE with the most popular order-determination methods in the literature: the sequential tests, BIC-type criterion and

the bootstrap methods. We will also incorporate CSE in this comparison. Again for the sequential tests, we choose the best two versions as suggested in Bura and Yang (2011), and we denote them by "seq.-1" and "seq.-2"; for BIC-type criterion, we fix the value of the tuning parameter to be $C_n = H\left(0.5\log(n) + 0.1n^{1/3}\right)/\left(2\,n\right)$ as suggested in Zhu, Miao and Peng (2006); for the bootstrap methods, we set the threshold δ to be 20%, 40% and 60% sequentially, and denote them by YW-1, YW-2 and YW-3.

To illustrate a variety of cases in practice, we choose the following 4 models:

Model I: $Y = X_1^2 + (X_2 + X_3)^2 + \epsilon$;

Model II: $Y = X_1^2 + X_2^2 + \epsilon$;

Model III: $Y = |X_1| + X_2 + X_3\,\epsilon$;

Model IV: $Y = X_1 + X_2^2 + \epsilon$.

In all these models $\epsilon \sim N(0, 0.5^2)$ is generated independently of X. In Model I, the pattern between X and Y is symmetric, and the two nonzero eigenvalues are clearly nonequal in all the inverse regression SDR methods; in Model II, the pattern is also symmetric and the two nonzero eigenvalues are equal when X has an identity covariance matrix; in Model III and Model IV, the pattern is symmetric only in certain directions of X, while Model III is heteroscedastic. In each model we will estimate the dimension of the central mean subspace, which is the same as the central subspace in Model I, II and IV. Note that each space is 2-dimensional.

As well, since the theories behind ASE vary with distributions of X, here we generate X from the following 3 distributions:

1. $X \sim N(0, I_p)$;
2. $X = Z \cdot I(\|Z\|_{\max} < 3)$ where the components of Z are mutually independent and each has distribution t_3;
3. $X = U \cdot R$ in which U is uniformly distributed on the unit sphere in $\mathbb{R}^p$ and is independent of $R \sim N(2, 0.5^2)$.

Here Distribution 1 is commonly used in the simulation studies in the literature. Distribution 2 generates heavily-tailed X, while Distribution 3 generates "ring-shaped" X. The last two distributions represent deviations of elliptical distribution from normality to a certain extent.

Table 4.1. Comparison of order-determination methods

Models	P_X	seq.-1	seq.-2	BIC	YW-1	YW-2	YW-3	CSE	ASE
I	1	0.51	0.31	0.07	0.97	1.00	0.92	0.99	1
I	2	0	1	0	0.94	0.97	0.93	0.97	0.99
I	3	0.10	0.88	0.03	0.91	0.99	0.90	1	1
II	1	0.99	0.99	0	0.52	0.97	0.91	1	1
II	2	0	0	0	0.45	0.94	0.80	0.99	0.97
II	3	0.18	1	0	0.57	0.97	0.90	0.98	0.99
III	1	0.99	0.99	0	0.74	0.97	0.87	0.89	1
III	2	0	0	0	0.83	0.96	0.91	0.95	0.97
III	3	0.86	1	0	0.78	0.95	0.93	0.97	1
IV	1	0.40	0.93	0	0.52	0.93	0.90	1	1
IV	2	0	0.90	0	0.93	0.95	0.91	1	0.91
IV	3	0.32	0	0	0.75	0.91	0.86	0.99	0.99

We fix p to be 10, q to be 9, and n to be 500, and apply each order-determination method to each model and each distribution of X. In each case, 200 samples are generated independently which results in 200 estimates of d. The performance of each order-determination method is then measured by the sample proportion of true estimation. To ensure that the candidate matrix has the desired rank, we choose direction regression as the inverse regression SDR method. The results are summarized in the following table, in which each cell records the sample proportion of true estimation.

From this table, while the performance of other estimators vary along the models, clearly, YW-2, CSE and ASE are consistent in all the cases. Compared to YW-2 and CSE, ASE is slightly more accurate in Model II and Model III, while it performs equally well in other models. In general, ASE performs the best when X is normally distributed. This is possibly due to the validity of invariance property, which removes the assumption of asymptotic normality on the inverse regression SDR methods.

4.5 Discussion

In this section we will discuss potential generalization of ASE in multiple directions.

The first potential generalization is to high-dimensional problems, which has been increasingly popular in recent years. In the current setting of "small p, large n", the consistency and asymptotic normality of inverse regression SDR methods is quite general. However, it is no longer the case when we allow p to increase with n. On one hand, it has been commonly realized that the inverse regression SDR methods lose their consistency when applied to high-dimensional data, except for SIR (Zhu and Zeng 2006; Cook, Forzani and Rothman 2012). On the other hand, the definition of asymptotic normality itself is tricky enough to be modified in the high-dimensional setting.

Nevertheless, when X is normally distributed, as pointed out in Section 2, asymptotic normality is not a necessary condition, so that ASE can be possibly generalized to high-dimensional data when SIR is used in the SDR procedure. For other inverse regression methods, this generalization is also possible when additional sparsity structure is invoked in the SDR assumptions (Li 2007; Chen, Zou and Cook 2010).

A second potential generalization is to modify ASE to be applicable to MAVE-type methods. In this modification, it is natural to replace the sample eigenvectors by the estimates of the central dimension reduction subspace, and replace the sample eigenvalues by the decrement of the objective function as k increases. Note that a fundamental motivation of ASE is that for any vector in the central dimension reduction subspace, its components corresponding to the augmentation predictor are 0. Besides, as seen in the proof of Theorem 6, the asymptotic normality can be relaxed to the condition that the asymptotic distribution of $f(n)\,\mathrm{vech} \circ (\hat{M} - M)$ is absolutely continuous with respect to Lebesque measure for some function f. Therefore we expect ASE to be consistent when applied to MAVE-type methods, after appropriate modification.

A third potential generalization is to modify ASE to be applicable to principle component analysis (PCA). This is worthwhile since PCA is still widely used in applications. Note that in the candidate matrix of SDR procedures, the rows and columns corresponding to the augmentation predictor are 0. However, this is no

longer true in PCA which is unsupervised learning. Therefore, the current version of ASE needs to be modified substantially in this generalization.

Chapter 5

On efficient dimension reduction with respect to a statistical functional of interest

5.1 Introduction

As mentioned in Chapter 1, in this chapter we will introduce a new sufficient dimension reduction framework that targets a statistical functional of interest, and propose an efficient estimator for the semiparametric estimation problems of this type. The statistical functional covers a wide range of applications, such as conditional mean, conditional variance, and conditional quantile. We will derive the general forms of the efficient score function and efficient information as well as their specific forms for three important statistical functionals: the linear functional, the composite linear functional, and the implicit functional. In conjunction with our theoretical analysis we will also propose a class of one-step Newton-Raphson algorithms, and show by simulation study and real applications that they substantially outperform existing methods.

The purpose of this chapter is twofold: to introduce a new framework for suffi-

cient dimension reduction that targets a statistical functional, and to develop semi-parametrically efficient estimators for problems of this type.

As reviewed in Chapter 2, SDR provides us a mechanism to reduce the dimension of the predictor while preserving the conditional distribution of Y given X. However, in many applications our interests are only in some specific aspects of the conditional distribution $P_{Y|X}$. For example, in nonparametric regression we are interested in the conditional mean $E(Y|X)$; in median regression we are interested in the conditional median $M(Y|X)$, and in volatility analysis we are interested in the conditional variance $\mathrm{Var}(Y|X)$, and in supervised classification we are interested in the class label of Y given its covariates X. To illuminate this point further let us consider the model

$$Y = \mu(X_1) + \sigma(X_1 + X_2)\varepsilon, \tag{5.1}$$

where μ and σ are unknown functions. In this case the central subspace, which is the 2-dimensional subspace of $\mathbb{R}^p$ spanned by $(1, 0 \ldots, 0)$ and $(0, 1, 0, \ldots, 0)$, only tells us that the sufficient predictors can be any linear combination of X_1 and X_2, but it does not tell us that the conditional mean is a function of X_1, and the conditional variance is a function of $X_1 + X_2$. Thus, the information provided by the central subspace is clearly inadequate if we want to build a model like (5.1).

Under these circumstances it makes sense to reformulate sufficient dimension reduction to target a specific functional, so as to provide a more nuanced picture of the relation between X and Y than offered by the central subspace. Several such efforts have been made over the past decade or so. For example, Cook and Li (2002) introduced the central mean subspace, which is defined by the relation $E(Y|X) = E(Y|\alpha^\intercal X)$, where α is minimal in the same sense as it is in the central subspace. Yin and Cook (2002) introduced the kth central moment subspace through the relation $E(Y^k|X) = E(Y^k|\alpha^\intercal X)$. Zhu and Zhu (2009) introduced the central variance subspace by requiring that $\mathrm{Var}(Y|X)$ is a function of $\alpha^\intercal X$. Zhu, Dong, and Li (2012) introduced a general class of estimating equations for single-index conditional variance. The Minimum Average Variance Estimator (MAVE) by Xia, Tong, Li, and Zhu (2002) also targets the central mean subspace. Kong and Xia (2012) introduced an adaptive quantile estimator for single-index quantile regression, which targets the conditional quantile. It turns out the space spanned by

the columns of α in the above relations are all subspaces of the central subspace. They provide refined structures for the central subspace. As a consequence, $\alpha^{\mathsf{T}}X$ can be written as $\beta^{\mathsf{T}}\zeta^{\mathsf{T}}X$, and we can refine central subspace based on the sufficient predictor $\zeta^{\mathsf{T}}X$. For convenience, we reset $\zeta^{\mathsf{T}}X$ as X, and use $\tilde{X}$ to represent the original predictor throughout the rest of the chapter.

The first goal of this project is to unify these problems by introducing a general dimension reduction paradigm with respect to statistical functional T of the conditional density of Y given X, say $\eta(x, y)$, through the following statement

$$T(\eta(x, \cdot)) \text{ is a function of } \beta^{\mathsf{T}}x. \tag{5.2}$$

Note that sufficient dimension reduction for conditional mean, conditional moments, and conditional variance discussed in the last paragraph are all special cases of relation (5.2). The minimal subspace $\mathcal{S}(\beta)$ of $\mathbb{R}^p$ that satisfies this relation is called the T-central subspace.

The second, and the main goal of this project is to develop semiparametrically efficient estimators for the T-central subspace. In a series of recent papers, Ma and Zhu (2012, 2013a, 2013b) use semiparametric theory to study sufficient dimension reduction and develop semiparametrically efficient estimators of the central subspace. These are related to the earlier developments by Li and Dong (2009) and Dong and Li (2010), which use estimating equations to relax the elliptical distribution assumption for sufficient dimension reduction. We extend Ma and Zhu's approach to find semiparametrically efficient estimator for the T-central subspace. We derive the general formulas for the efficient score and efficient information for the semiparametric family specified by the relation (5.2), and further deduce their specific forms for three important statistical functionals: the linear functionals (L-functionals), the composite linear functionals (C-functionals), and implicit functionals (I-functionals). These functionals cover a wide range of applications. For example, all conditional moments are L-functionals, all conditional cumulants (see, for example, McCullagh, 1987) are C-functionals, and quantities such as conditional median, conditional quantile, and conditional support vector machine (Li, Artemiou, and Li, 2011) are I-functionals.

Using the formulas for efficient score and efficient information, we propose a one-step Newton-Raphson alorithm to implement semiparametrically efficient

estimation. Compared with the semiparametric estimators of Ma and Zhu (2012, 2013a), our algorithm has two distinct and attractive features. First, since our algorithm relies on the MAVE-type procedure for minimization, it can be implemented by iterations of a least squares problem without resorting to high-dimensional search-based optimization. Second, unlike Ma and Zhu (2012, 2013a), our method does not require any specific parameterization of β that potentially restricts the generality of their method.

The rest of this chapter is organized as follows. In Section 5.2 we give a general formulation of sufficient dimension reduction with respect to a statistical functional of interest. To set the stage for further development, we lay out the semiparametric structure of our problem in Section 5.3. In Section 5.4 we derive the efficient score and efficient information for a general statistical functional. In Sections 5.5, 5.6, and 5.7 we further deduce the specific forms of the efficient score and efficient information for the L-, C-, and I-functionals. In section 5.8 we discuss the effect of estimating the central subspace on the efficient score. In section 5.9 we develop the one-step Newton-Raphson estimation procedure for semiparametrically efficient estimation. In section 5.10 we conduct simulation studies to compare our method with other methods, and in Section 5.11 we apply our method to a data set. Some concluding remarks are made in Section 5.12. The proofs of some technical results are given in Section 5.13 at the end.

The following notation will be consistently used throughout the rest of the chapter. The symbol I_k denotes the $k \times k$ dimensional identity matrix; e_k denotes a vector whose kth entry is 1 and other entries are 0; $\perp\!\!\!\perp$ indicates independence or conditional independence between two random elements — that is, $A \perp\!\!\!\perp B$ means A and B are independent, and $A \perp\!\!\!\perp B|C$ means A and B are independent given C. For integers s and t, $\mathbb{R}^s$ denotes the s dimensional Euclidean space, and $\mathbb{R}^{s\times t}$ denotes the set of $s \times t$ dimensional matrices. For a function with multiple arguments, say $f(x, y, z)$, we use the dot notation to represent mappings of a subset of the arguments. For example, $f(\cdot, y, z)$ represents the mapping $x \mapsto f(x, y, z)$ where y and z are fixed, and $f(\cdot, \cdot, z)$ represents the mapping $(x, y) \mapsto f(x, y, z)$ where z is fixed. We use superscripts of X to index components and subscripts of X index subjects. Thus X_i^j means the jth component of the ith observation in a sample $X_1, \ldots, X_n$. However, a^i represents power when a is not X.

5.2 Dimension reduction for conditional statistical functional

Let $(\Omega_X, \mathcal{F}_X, \mu_X)$ and $(\Omega_Y, \mathcal{F}_Y, \mu_Y)$ be σ-finite measure spaces, where $\Omega_X \subseteq \mathbb{R}^d$ and $\Omega_Y \subseteq \mathbb{R}$ and $\mathcal{F}_X$ and $\mathcal{F}_Y$ are σ-fields of Borel sets in Ω_X and Ω_Y. Let (X, Y) be a pair of random elements that takes values in $(\Omega_X \times \Omega_Y, \mathcal{F}_X \times \mathcal{F}_Y)$. Let $\mathscr{M}$ be a family of densities of (X, Y) with respect to $\mu = \mu_X \times \mu_Y$. We assume that $\mathscr{M}$ is a semiparametric family; that is, there exist $\Theta \subseteq \mathbb{R}^r$ and a family $\mathscr{F}$ of functions $\phi : \Omega_X \times \Omega_Y \times \Theta \to \mathbb{R}$ such that

$$\mathscr{M} = \cup\{\mathscr{M}_\theta : \theta \in \Theta\}, \text{ where } \mathscr{M}_\theta = \{\phi(\cdot, \cdot, \theta) : \phi \in \mathscr{F}\}.$$

Furthermore, we assume that each $\phi \in \mathscr{F}$ can be factorized into $\lambda(x)\eta(x, y, \theta)$ where λ is the marginal density of X, $\eta(x, \cdot, \theta)$ is the conditional density of Y given X. The real assumption in this factorization is that θ appears only in the conditional density.

As an illustration, consider the single index model where

$$Y = m(\beta^\intercal X) + \varepsilon, \quad \beta \in \mathbb{R}^d, \quad X \perp\!\!\!\perp \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \quad X \sim N(0, \Sigma),$$

and m is an unknown function. Since m is unknown, β is identified up to a proportional constant. To avoid the trivial case let us assume it has at least one nonzero component, and further assume this is the first component for convenience. We can then assume without loss of generality $\beta^\intercal = (1, \theta^\intercal)$ where $\theta \in \mathbb{R}^{d-1}$. Then

$$\beta = e_1 + \Gamma\theta, \quad m(\beta^\intercal X) = m(X^1 + \theta^\intercal \Gamma^\intercal X),$$

where $\Gamma^\intercal$ is the $d \times (d-1)$ matrix $(0, I_{d-1})$. In this case, $\lambda(x)$ is the p.d.f. of $N(0, \Sigma)$ and $\eta(x, \cdot, \theta)$ is the p.d.f. of $N(m(x^1 + \theta^\intercal \Gamma^\intercal x), \sigma^2)$ for a given x.

Now let $\mathscr{F}_1$ denote the family $\{\eta : \phi \in \mathscr{F}\}$, and

$$\mathscr{L} = \{\lambda : \phi \in \mathscr{F}\}, \quad \mathscr{H}_\theta = \{\eta(\cdot, \cdot, \theta) : \eta \in \mathscr{F}_1\}, \quad \mathscr{H} = \cup\{\mathscr{H}_\theta : \theta \in \Theta\}.$$

We assume that $\mathscr{M}$ contains the true density of (X, Y). That is, there exist $\theta_0 \in \Theta$, $\lambda_0 \in \mathscr{L}$, and $\eta_0 \in \mathscr{F}_1$ such that $\phi_0(x, y, \theta_0) = \lambda_0(x)\eta_0(x, y, \theta_0)$ is the true density

of (X, Y). For convenience, we abbreviate $\phi_0(x, y, \theta_0)$, $\eta_0(x, y, \theta_0)$, and $\mathscr{M}_{\theta_0}$ as $\phi_0(x, y)$, $\eta_0(x, y)$, and $\mathscr{M}_0$.

Let $\mathscr{G}$ be a class of densities of Y with respect to μ_Y that contains all $\eta(x, \cdot, \theta)$ for $\eta \in \mathscr{F}_1$, $\theta \in \Theta$, and $x \in \Omega_X$. Let $T : \mathscr{G} \to \mathbb{R}$ be a mapping from $\mathscr{G}$ to $\mathbb{R}$. Such mappings are called statistical functionals. The functional T induces the random variable

$$x \mapsto T(\eta(x, \cdot, \theta)),$$

on Ω_X, which we write as $T(\eta(X, \cdot, \theta))$. Following the convention of the conditional expectation, we write $T(\eta_0(X, \cdot, \theta_0))$ as $T(Y|X)$. This random variable can be used to characterize a wide variety of features of a conditional density $\eta(x, \cdot, \theta)$ that might interest us, as detailed by the following example.

Example 1 Let $T : \mathscr{G} \to \mathbb{R}$ be the functional $g \mapsto \int_{\Omega_Y} yg(y)d\mu_Y$. Then, each $\eta(\cdot, \cdot, \theta) \in \mathscr{H}$ uniquely defines the mapping

$$x \mapsto T(\eta(x, \cdot, \theta)) = \int_{\Omega_Y} y\eta(x, y, \theta)d\mu_Y(y).$$

That is, $T(\eta(X, \cdot, \theta))$ is the conditional expectation $E(Y|X)$ under $\eta(\cdot, \cdot, \theta)$.

Let $T : \mathscr{G} \to \mathbb{R}$ be the mapping

$$T(g) = \int y^2 g(y)d\mu_Y(y) - (\int yg(y)d\mu_Y(y))^2.$$

Then, $T(\eta(X, \cdot, \theta))$ is the conditional variance $\mathrm{Var}(Y|X)$ under the conditional density $x \mapsto \eta(x, \cdot, \theta)$.

Let $T : \mathscr{G} \to \mathbb{R}$ be the functional defined by the equation in m

$$\int \mathrm{sgn}(y - m)g(y)d\mu_Y(y) = 0, \tag{5.3}$$

where $\mathrm{sgn}(a)$ is the sign function that takes the value 1 if $a \geq 0$ and takes the value -1 if $a < 0$. The solution to (5.3) is the median of Y. Each $\eta(\cdot, \cdot, \theta) \in \mathscr{H}$ uniquely defines the mapping $T(\eta(X, \cdot, \theta))$, which is the conditional median of Y given X under the conditional density $x \mapsto \eta(x, \cdot, \theta)$. □

We now give a rigorous definition of the T-central subspace.

Definition 2. *If there is a matrix $\gamma \in \mathbb{R}^{d\times u}$, with $u < d$, such that $T(\eta_0(X,\cdot,\theta_0))$ is measurable with respect to $\sigma(\gamma^\intercal X)$, then we call $\mathcal{S}(\gamma)$ a sufficient dimension reduction subspace for T. The intersection of all such spaces is called the central subspace for conditional functional T, or the T-central subspace.*

We denote the T-central subspace by $\mathcal{S}_{T(Y|X)}$. For example, if T is the conditional mean functional, then $\mathcal{S}_{T(Y|X)}$ becomes the central mean subspace, which we write as $\mathcal{S}_{E(Y|X)}$; if T is the conditional median functional, then $\mathcal{S}_{T(Y|X)}$ becomes the central median subspace, which we write as $\mathcal{S}_{M(Y|X)}$; if T is the conditional variance functional, then $\mathcal{S}_{T(Y|X)}$ becomes the central variance subspace, which we write as $\mathcal{S}_{V(Y|X)}$. It is easy to see that $\mathcal{S}_{T(Y|X)} \subseteq \mathcal{S}_{Y|X}$: this is because $Y \perp\!\!\!\perp X|\beta^\intercal X$ implies

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

In the following, we denote the dimension of the $\mathcal{S}_{T(Y|X)}$ by s and any basis matrix of $\mathcal{S}_{T(Y|X)}$ (of dimension $d \times s$) as β.

5.3 Formulation of the semiparametric problem

To set the stage for further development we first outline the basic semiparametric structure of our problem. Let $L_2(\phi_0) = \{r : \int r^2\phi_0 d\mu < \infty\}$. Let $\langle\cdot,\cdot\rangle_{\phi_0}$ and $\|\cdot\|_{\phi_0}$ denote the inner product and norm in $L_2(\phi_0)$. For a technical reason, it is easier to work with an embedding of $\mathscr{M}$ into $L_2(\phi_0)$, defined as

$$R : \phi \mapsto 2(\phi^{1/2} - \phi_0^{1/2})/\phi_0^{1/2} \equiv r.$$

Let $R(\mathscr{M}) = \{R(\phi) : \phi \in \mathscr{M}\}$. This transformation ensures that $R(\mathscr{M}) \subseteq L_2(\phi_0)$; whereas additional assumptions are needed to ensure $\mathscr{M} \subseteq L_2(\phi_0)$. Also note that $R(\phi_0)$ is the 0 element in $L_2(\phi_0)$.

A curve in $R(\mathscr{M}_0)$ that passes through $r_0 = 0$ is any mapping $\alpha \mapsto r_\alpha(\cdot)$ from $[0,1) \to R(\mathscr{M}_0)$ that is Fréchet differentiable at $\alpha = 0$. That is, there is a member $\dot{r}_0$ of $L_2(\phi_0)$ such that

$$\|r_\alpha - r_0 - \dot{r}_0\alpha\|_{\phi_0} = o(|\alpha|).$$

The tangent space $\mathscr{T}_\phi$ of $R(\mathscr{M}_0)$ at r_0 is the closure of the subspace of $L_2(\phi_0)$ spanned by $\dot{r}_0$ along all curves.

Let $\mathring{r}_0 \in [L_2(\phi_0)]^r$ be the score with respect to θ; that is,

$$\|R(\phi_0(\cdot,\theta)) - R(\phi_0(\cdot,\theta_0)) - \mathring{r}_0^{\mathsf{T}}(\theta-\theta_0)\|_{\phi_0} = o(\|\theta-\theta_0\|).$$

Let $\Pi(\mathring{r}_0|\mathscr{T}_\phi^\perp)$ be the componentwise projection of the random vector $\mathring{r}_0$ on to the orthogonal complement of the tangent space $\mathscr{T}_\phi$. This projection is called the efficient score, and we denote it by $S_{\text{eff}}(X,Y,\theta_0)$. The matrix

$$J_{\text{eff}}(\theta_0) = E[S_{\text{eff}}(X,Y,\theta_0)S_{\text{eff}}^{\mathsf{T}}(X,Y,\theta_0)]$$

is called the efficient information. Now let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be an i.i.d. sample of (X,Y). For a function h of (x,y), let $E_nh(X,Y)$ denote the sample average of $h(X_1,Y_1),\ldots,h(X_n,Y_n)$. Under some conditions, if $\hat{\theta}$ is the solution to the estimating equation

$$E_nS_{\text{eff}}(X,Y,\theta) = 0, \tag{5.4}$$

then $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{\mathcal{D}} N(0, J_{\text{eff}}^{-1}(\theta_0))$. Moreover, for any estimate $\tilde{\theta}$ of θ_0 that is regular with respect to $\mathscr{T}_\phi$, $\sqrt{n}(\tilde{\theta}-\theta_0)$ can be decomposed as $\sqrt{n}(\hat{\theta}-\theta_0)+\Delta_n$ where two terms are asymptotically independent. This result, well known in the semiparametric literature as the Hájek-LeCam convolution theorem, implies $\hat{\theta}$ has the smallest asymptotic variance among all regular estimators with respect to $\mathscr{T}_\phi$. That is, $\hat{\theta}$ is semiparametrically efficient. For a comprehensive exposition of this theory, see Bickel, Klaassen, Ritov, and Wellner (1993, Chapter 3) or van der Vaart (1998, Chapter 25).

We now investigate how the sufficient dimension reduction in Definition 2 specifies the semiparametric family $\mathscr{M}$, and what is the meaning of the parameter θ in this context. Since our goal is to estimate $\mathcal{S}(\beta)$, we need fewer parameters than ds. In fact, the set $\{\mathcal{S}(\beta) : \beta \in \mathbb{R}^{d\times s},\ \text{rank}(\beta) = s\}$ is a Grassmann manifold, which has dimension $s(d-s)$ (see, for example, Edelman, Arias, and Smith, 1998). There always exists a smooth parameterization $\beta = \beta(\theta)$, where $\theta \in \mathbb{R}^{s(d-s)}$, because $\mathcal{S}(\beta)$ is determined if a certain $s\times s$ submatrix of β is fixed as I_s and the complementary $(d-s)\times s$ block has free varying entries. The specific

form of the parameterization is not important to us.

Let $\sigma_\theta(X)$ be the σ-field generated by $\beta^\intercal(\theta)X$. Because $T(\eta(X,\cdot,\theta))$ is measurable with respect to $\sigma_\theta(X)$ if and only if

$$T(\eta(X,\cdot,\theta)) = E[T(\eta(X,\cdot,\theta))|\sigma_\theta(X)],$$

the semiparametric family for our purpose is $\mathscr{M} = \cup\{\mathscr{M}_\theta : \theta \in \mathbb{R}^{s(d-s)}\}$ where

$$\mathscr{M}_\theta = \{\phi(\cdot,\cdot,\theta) : \phi \in \mathscr{F},\ T(\eta(x,\cdot,\theta)) = E_\lambda[T(\eta(X,\cdot,\theta))|\sigma_\theta(X)]_x\ \ \forall\, x \in \Omega_X\}.$$

Here, for a sub-σ-field $\mathscr{A}$ of $\mathcal{F}_X$ and a function $f(X,Y)$, $E[f(X,Y)|\mathscr{A}]_x$ denotes the evaluation of the conditional expectation $E[f(X,Y)|\mathscr{A}]$ at x.

In this chapter we will focus on the development of the efficient score, the efficient information, and an accompanying estimation procedure, but will not give a rigorous proof of the asymptotic results (including the asymptotic distribution of $\sqrt{n}(\hat{\theta}-\theta_0)$ and the convolution theorem), because it would far exceed the scope of this chapter and because we do not expect the proof will fundamental deviate from that given in Ma and Zhu (2013a). In addition, as mentioned earlier, by design our method is applied to the sufficient predictor corresponding to the central subspace, whose dimension is relatively low. Consequently, we expect no surprises as regards the validity of $\sqrt{n}$-rate of convergence of our estimator. In the meantime, our simulation studies provide strong evidence that the our efficient estimator does approach the theoretical semiparametric variance bound for modestly large sample sizes.

5.4 Efficient score and efficient information

In this section we derive the efficient score and efficient information for the semiparametric problem set up in Section 5.3. To this end we first derive the tangent space $\mathscr{T}_\phi$ for a fixed $\theta \in \Theta$. Let $\mathscr{T}_\eta$ be the tangent space of $R(\mathscr{H}_\theta)$ at $R(\eta_0(\cdot,\cdot,\theta)) = 0$, and $\mathscr{T}_\lambda$ be the tangent space of $R(\mathscr{L})$ at $R(\lambda_0) = 0$.

Proposition 1. *The following relations hold:*

1. $\mathscr{T}_\phi = \mathscr{T}_\eta + \mathscr{T}_\lambda$;
2. $\mathscr{T}_\eta \perp \mathscr{T}_\lambda$ *in terms of the inner product in* $L_2(\phi_0)$;

3. $\mathscr{T}_\phi^\perp = \mathscr{T}_\lambda^\perp \cap \mathscr{T}_\eta^\perp$.

This proposition was verified and used in Ma and Zhu (2012, 2013a). Since the family $\mathscr{L}$ has no constraint, its tangent space is straightforward, as given in the next proposition, which is taken from Bickel et al (1993, page 52).

Proposition 2. *$\mathscr{T}_\lambda$ consists of all functions h in $L_2(\lambda_0)$ with $E_{\lambda_0} h(X) = 0$.*

To compute $\mathscr{T}_\eta$, we introduce a new functional for each fixed $x \in \Omega_X$. Let $\mathscr{H}_{\theta,x}$ be the class of densities $\{\eta(x,\cdot,\theta) : \eta \in \mathscr{H}_\theta\}$. Let R_x denote the mapping

$$\begin{aligned} R_x : \mathscr{H}_{\theta,x} &\to L_2(\eta_0(x,\cdot,\theta)), \\ \eta(x,\cdot,\theta) &\mapsto 2[\eta^{1/2}(x,\cdot) - \eta_0^{1/2}(x,\cdot,\theta)]/\eta_0^{1/2}(x,\cdot,\theta). \end{aligned} \tag{5.5}$$

This mapping is different from R, which is from $\mathscr{M}$ to $L_2(\phi_0)$. Nevertheless, note that $R(\eta(\cdot,\cdot,\theta))(x,y) = R_x(\eta(x,\cdot,\theta))(y)$. Let

$$T_x :\ L_2(\eta_0(x,\cdot,\theta)) \to \mathbb{R}, \quad r(x,\cdot,\theta) \mapsto T \circ R_x^{-1}(r(x,\cdot,\theta)). \tag{5.6}$$

The Fréchet derivative of T_x at $r(x,\cdot,\theta)$ is denoted by $\dot{T}_x(r(x,\cdot,\theta))$. This is a bounded linear functional from $L_2((\eta_0(x,\cdot,\theta))$ to $\mathbb{R}$.

Theorem 8. *Suppose, for each $x \in \Omega_X$, the functional T_x is Fréchet differentiable at 0. Let $\tau(x,\cdot,\theta)$ be the Riesz representation of $\dot{T}_x(0)$ and assume $\tau(\cdot,\cdot,\theta) \in L_2(\phi_0)$. Then*

$$\mathscr{T}_\eta \subseteq \{[h(x) - E(h(X)|\sigma_\theta(X))_x]\tau(x,y,\theta) + g(x) :\ h,\ g \in L_2(\lambda_0)\}^\perp \equiv \mathscr{U}^\perp \tag{5.7}$$

Moreover, if, for each $x \in \Omega_X$, the function $r(x,\cdot,\theta) \mapsto T_x(r(x,\cdot,\theta))$ is continuously Fréchet differentiable in a neighborhood of $0 \in L_2(\eta_0(x,\cdot,\theta))$, then $\mathscr{T}_\eta \supseteq \mathscr{U}^\perp$.

The proof is this theorem is technical and is presented in the Appendix (Section 5.13). From Theorem 8 and Propositions 1 and 2 we can easily derive the form of $\mathscr{T}_\phi^\perp$, as follows.

Corollary 2. *Under the assumptions of Theorem 8,*

$$\mathscr{T}_\phi^\perp = \{[h(x) - E(h(X)|\sigma_\theta(X))_x][\tau(x,y,\theta) - E(\tau(X,Y,\theta)|X)_x] :\ h \in L_2(\lambda_0)\}.$$

We now compute the efficient score, which is the projection of the true score with respect to θ on to $\mathscr{T}_\phi^\perp$. Let

$$r_0(x,y,\theta) = 2[\phi_0^{1/2}(x,y,\theta) - \phi_0^{1/2}(x,y,\theta_0)]/\phi_0^{1/2}(x,y,\theta_0).$$

The true score for the parameter of interest is the Fréchet derivative

$$\partial r_0(x,y,\theta)/\partial\theta|_{\theta=\theta_0}.$$

To differentiate from $\dot{r}_0(x,y,\theta_0)$, we denote the above derivative by $\mathring{r}_0(x,y,\theta_0)$. This is an $s(d-s)$-dimensional vector. Since the mapping $T_x : L_2(\eta_0(x,\cdot,\theta)) \to \mathbb{R}$ and R_x also depend on θ, we now write them as $T_{x,\theta}$ and $R_{x,\theta}$. We use T_x to denote the mapping T_{x,θ_0}. Following Bickel et al (1993, Chapter 3), we use $\Pi(f|\mathscr{A})$ to represent the projection of a function f on to a subspace $\mathscr{A}$ of $L_2(\phi_0)$.

Theorem 9. *Suppose the following conditions hold.*

1. *For each $x \in \Omega_X$ and θ in a neighborhood of θ_0, the mapping*

$$T_{x,\theta} : L_2(\eta_0(x,\cdot,\theta)) \to \mathbb{R}$$

is continuously Fréchet differentiable in a neighborhood of 0. Let $\tau(x,y,\theta)$ be the Riesz representation of $\dot{T}_{x,\theta}(0)$ and $\tau_c(x,y,\theta)$ be its centered version $\tau(x,y,\theta) - E_\theta[\tau(X,Y,\theta)|X]_x$.

2. *The function $\theta \mapsto r_0(x,\cdot,\theta)$ is Fréchet differentiable at $\theta = \theta_0$ with Fréchet derivative $\mathring{r}_0(x,\cdot,\theta_0)$.*

3. *If*

$$\begin{aligned}
q_1(x,\theta_0) &= E[\mathring{r}_0(X,Y,\theta_0)\tau_c(X,Y,\theta_0)|X]_x \\
q_2(x,\theta_0) &= E[\tau_c^2(X,Y,\theta_0)|X]_x \\
q_2^*(x,\theta_0) &= E_{\theta_0}[q_1(X,\theta_0)q_2^{-1}(X,\theta_0)|\sigma_{\theta_0}(X)]_x / E[q_2^{-1}(X,\theta_0)|\sigma_{\theta_0}(X)]_x \\
q_3(x,\theta_0) &= q_1(x,\theta_0) - q_2^*(x,\theta_0) \\
q_4(x,\theta_0) &= q_2^{-1}(x,\theta_0)q_3(x,\theta_0),
\end{aligned}$$

where $q_2^{-1}(x,\theta_0)$ is the reciprocal of $q_2(x,\theta_0)$, then $q_4(x,\theta_0) \in L_2(\lambda_0)$.

Then

$$S_{\text{eff}}(x, y, \theta_0) = \Pi(\mathring{r}_0(x, y, \theta_0)|\mathscr{T}_\phi^\perp) = q_4(x, \theta_0)\tau_c(x, y, \theta_0). \tag{5.8}$$

PROOF. Let $u^*(x, y, \theta_0) = q_4(x, \theta_0)\tau_c(x, y, \theta_0)$. By the projection theorem, it suffices to show

(a) $u^*(\cdot, \cdot, \theta_0) \in \mathscr{T}_\phi^\perp$;

(b) for any $u \in \mathscr{T}_\phi^\perp$,

$$\langle \mathring{r}_0(\cdot, \cdot, \theta_0), u\rangle_{\phi_0} = \langle u^*(\cdot, \cdot, \theta_0), u\rangle_{\phi_0}. \tag{5.9}$$

By Condition 3, $q_4(\cdot, \theta_0) \in L_2(\lambda_0)$. Moreover, by the definition of q_4 in Condition 3 it is easy to verify that $E(q_4(X, \theta)|\sigma_{\theta_0}(X)) = 0$. Hence, assertion (a) holds.

Because $u \in \mathscr{T}_\phi^\perp$, it has the form $h(x, \theta_0)\tau_c(x, y, \theta_0)$ for some $h(\cdot, \theta_0) \in L_2(\lambda_0)$ satisfying $E[h(X, \theta_0)|\sigma_{\theta_0}(X)] = 0$. Hence the right hand side of (5.9) is

$$\begin{aligned} E_\theta[h(X, \theta_0)q_4(X, \theta_0)\tau_c^2(X, Y, \theta_0)] &= E[h(X, \theta_0)q_4(X, \theta_0)q_2(x, \theta_0)] \\ &= E[h(X, \theta_0)q_3(X, \theta_0)]. \end{aligned}$$

Substitute the definition of $q_3(x, \theta_0)$ into the right hand side, and it becomes

$$E_\theta\{h(X, \theta)\{q_1(x, \theta_0) - E[q_1(x, \theta_0)q_2^{-1}(x, \theta_0)|\sigma_{\theta_0}(X)]/E[q_2^{-1}(x, \theta_0)|\sigma_{\theta_0}(X)]\}.$$

However, because $E(h(X, \theta_0)|\sigma_{\theta_0}(X)) = 0$, the equation above reduces to

$$E[h(X, \theta_0)q_1(X, \theta_0)] = E[h(X, \theta_0)\mathring{r}_0(X, Y, \theta_0)\tau_c(X, Y, \theta_0)],$$

which is the left hand side of (5.9). □

The next corollary, which follows directly from Theorem 9, gives the general form for the efficient information estimating $\mathcal{S}_{T(Y|X)}$.

Corollary 3. *Under the assumptions of Theorem 9, the efficient information for estimating $\mathcal{S}_{T(Y|X)}$ is given by*

$$J_{\text{eff}}(\theta_0) = E[q_3(X, \theta_0)q_3^{\mathsf{T}}(X, \theta_0)q_2^{-1}(X, \theta_0)]. \tag{5.10}$$

In the next three sections we apply the general result in Theorem 9 to derive the explicit forms of the efficient scores for three types of commonly used statistical functionals: the linear functionals, the composite linear functionals, and the implicit functionals. The common thread that runs through these developments is the calculation of the Riesz representation $\tau(x, y, \theta_0)$ of the Fréchet derivative $\dot{T}_x(0)$.

5.5 Linear statistical functionals

Dimension reduction of this type is the direct generalizations of the central mean subspace (Cook and Li, 2002) and the central moment subspace (Yin and Cook, 2002). It can also be viewed as a generalization of the single- and multiple-index models (see, for example, Härdle, Hall, and Ichimura, 1993). Let $f : \Omega_Y \to \mathbb{R}$ be a square-integrable function. Let L be the functional

$$L : \mathscr{G} \to \mathbb{R}, \quad g \mapsto \int_{\Omega_Y} f(y)g(y)d\mu_Y(y).$$

The corresponding conditional statistical functional is

$$L_{x,\theta}(r(x, \cdot, \theta)) \equiv L \circ R^{-1}_{x,\theta}(r(x, \cdot, \theta)) = \int_{\Omega_Y} f(y)(1 + r(x, y, \theta)/2)^2 \eta_0(x, y, \theta)dy.$$

The L-central subspace is defined by the relation

$$E[f(Y)|X] = E[f(Y)|\sigma_{\theta_0}(X)]. \tag{5.11}$$

Theorem 10. *Suppose the conditions 1, 2, 3 in Theorem 9 are satisfied for $L_{x,\theta}$. Then the efficient score for θ in problem (5.11) is given by (5.8) in which*

$$\begin{aligned} \tau_c(x, y, \theta_0) &= f(y) - E_{\theta_0}[f(Y)|\sigma_{\theta_0}(X)], \\ q_1(x, \theta_0) &= \partial E_\theta[f(Y)|\sigma_\theta(X)]_x/\partial\theta|_{\theta=\theta_0}, \\ q_2(x, \theta_0) &= E_{\theta_0}[f^2(Y)|X]_x - E^2_{\theta_0}[f(Y)|\sigma_{\theta_0}(X)]_x. \end{aligned}$$

PROOF. Because L_x is Fréchet differentiable at 0, its Fréchet derivative is the

same as the Gâteaux derivative (Bickel et al, 1993, page 455), which is defined by

$$r(x,\cdot,\theta_0) \mapsto \partial L_x(\epsilon r(x,\cdot,\theta_0))/\partial\epsilon|_{\epsilon=0}.$$

However, because

$$\begin{aligned}\partial L_x(\epsilon r(x,\cdot,\theta_0))/\partial\epsilon|_{\epsilon=0} &= \partial[\textstyle\int_{\Omega_Y} f(y)(1+\epsilon r(x,y,\theta_0)/2)^2\eta_0(x,y,\theta_0)dy]/\partial\epsilon|_{\epsilon=0}\\ &= \textstyle\int_{\Omega_Y} f(y)r(x,y,\theta_0)\eta_0(x,y,\theta_0)dy = \langle f, r(x,\cdot,\theta_0)\rangle_{\eta_0(x,\cdot,\theta_0)},\end{aligned}$$

the Riesz representation of $\dot{T}_x(0)$ is f. Hence, by Theorem 9,

$$q_2(x,\theta_0) = E[\tau_c^2(X,Y,\theta_0)|X]_x = E[f^2(Y)|X]_x - E^2[f(Y)|\sigma_{\theta_0}(X)]_x.$$

Also, for each θ,

$$\begin{aligned}\textstyle\int f(y)(1+r_0(x,y,\theta))^2\eta_0(x,y,\theta_0)d\mu_Y(y) &= \textstyle\int f(y)\eta_0(x,y,\theta)d\mu_Y(y)\\ &= E_\theta[f(Y)|X]_x = E_\theta[f(Y)|\sigma_\theta(X)]_x.\end{aligned}$$

Take Fréchet derivative with respect to θ on both sides to obtain

$$q_1(x,\theta_0) = E[f(Y)\mathring{r}_0(X,Y,\theta_0)|X]_x = \partial E_\theta[f(Y)|\sigma_\theta(X)]_x/\partial\theta|_{\theta=\theta_0},$$

as desired. □

Example 2. The central mean subspace introduced by Cook and Li (2002) is a special case of the L-central subspace with $f(y) = y$. The efficient score and efficient information are given by (5.8) and (5.10) where

$$\begin{aligned}\tau_c(x,y,\theta_0) &= y - E_{\theta_0}[Y|\sigma_{\theta_0}(X)]_x,\\ q_1(x,\theta_0) &= \partial E_\theta[Y|\sigma_\theta(X)]_x/\partial\theta|_{\theta=\theta_0},\\ q_2(x,\theta_0) &= \mathrm{Var}_{\theta_0}(Y|X)_x = E_{\theta_0}(Y^2|X) - E^2_{\theta_0}[Y|\sigma_{\theta_0}(X)].\end{aligned} \tag{5.12}$$

For example, if the central mean subspace has dimension 1 and is spanned by $c + \Gamma\theta_0$ for some $c \in \mathbb{R}^p$ and $\Gamma \in \mathbb{R}^{p\times(p-1)}$, as described in the second paragraph of Section 5.2, then

$$\tau_c(x,y,\theta_0) \quad = \quad y - E_{\theta_0}(Y|X^1 + \theta_0^{\mathsf{T}}\Gamma^{\mathsf{T}}X)_x,$$

$$q_1(x,\theta_0) \quad = \quad \Gamma^\intercal x[\partial E_\theta(Y|X^1+\theta^\intercal\Gamma^\intercal X)_x/\partial(\theta^\intercal\Gamma^\intercal x)|_{\theta=\theta_0}].$$

Therefore the efficient score is

$$S_{\text{eff}}(x,y,\theta) = \frac{1}{\text{Var}_\theta(Y|X)_x}\frac{\partial E_\theta(Y|X^1+\theta^\intercal\Gamma^\intercal X)_x}{\partial(\theta^\intercal\Gamma x)} \left\{x - \frac{E_\theta[X/\text{Var}_\theta(Y|X)|X^1+\theta^\intercal\Gamma^\intercal X]_x}{E_\theta[1/\text{Var}_\theta(Y|X)|X^1+\theta^\intercal\Gamma^\intercal X]_x}\right\}[y - E_\theta(Y|X^1+\theta^\intercal\Gamma^\intercal X)_x].$$

The efficient information is

$$J_{\text{eff}}(\theta) = E_\theta\left[\frac{\partial E_\theta(Y|X^1+\theta^\intercal\Gamma^\intercal X)}{\partial(\theta^\intercal\Gamma^\intercal X)}\left(X - \frac{E_\theta[X/\text{Var}(Y|X)|X^1+\theta^\intercal\Gamma^\intercal X]}{E_\theta[1/\text{Var}_\theta(Y|X)|X^1+\theta^\intercal\Gamma^\intercal X]}\right)\right]^{\otimes 2}$$

where $A^{\otimes 2}$ denotes $AA^\intercal$ by for a matrix A.

Alternatively, the efficient score and information can be written in the original parameterization β. See the Appendix (Section 5.13) for their explicit forms in the β-parameterization.

The central kth moment space (Yin and Cook, 2002) is a special case of the L-functional with $f(y) = y^k$. The efficient score and efficient information where (5.8) q_1, q_2, and τ_c given by formulas similar to (5.12) with Y replaced by Y^k.

Zhu and Zeng (2006) considered a SDR problem defined through the characteristic function $E_{\theta_0}(e^{itY}|X) = E_{\theta_0}[e^{itY}|\sigma_{\theta_0}(X)]$. They used this relation to recover the central space, but if our goal is to estimate θ defined through this relation, then

$$\begin{aligned}\tau_c(x,y,\theta_0) &= e^{ity} - E_{\theta_0}[e^{itY}|\sigma_{\theta_0}(X)]_x,\\ q_2(x,\theta_0) &= \text{Var}_{\theta_0}(e^{itY}|X)_x,\\ q_1(x,\theta_0) &= \partial E_\theta[e^{itY}|\sigma_\theta(X)]_x/\partial\theta|_{\theta=\theta_0}.\end{aligned}$$

The efficient score can be obtained by substituting the above into (5.8). □

5.6 Composite linear statistical functionals

We now consider a nonlinear function of several linear functionals, which is motivated by dimension reduction for conditional variance considered in Zhu and Zhu (2009) and the single-index conditional heteroscedasticity model in Zhu, Dong,

and Li (2012). See also Xia, Tong, and Li (2002). In fact, all cumulants are functionals of this type. Let $T_1, \ldots, T_k$ be bounded linear functionals from $\mathscr{G}$ to $\mathbb{R}$. That is,

$$T_\ell(g) = \int f_\ell(y) g(y) d\mu_Y(y), \quad \ell = 1, \ldots, k,$$

where $f_1, \ldots, f_k$ are square-integrable with respect to any density $g \in \mathscr{G}$. Let $\rho : \mathbb{R}^k \to \mathbb{R}$ be a differentiable function. Then $C : \ g \mapsto \rho(T_1(g), \ldots, T_k(g))$ defines a statistical functional on $\mathscr{G}$ to $\mathbb{R}$. We call such functionals composite linear functionals, or C-functionals. For example, if

$$T_1(g) = \int y g(y) d\mu_Y(y), \quad T_2(g) = \int y^2 g(y) d\mu_Y(y), \quad \rho(s_1, s_2) = s_2 - s_1^2,$$

then $C(g) = \mathrm{Var}_g(Y)$ is the variance functional. The corresponding conditional statistical functional is defined by

$$C_{x,\theta}(r(x,\cdot,\theta)) \ = C \circ \mathbb{R}_{x,\theta}^{-1}(r(x,\cdot,\theta_0)) = \rho[T_{1,x,\theta}(r(x,\cdot,\theta)), \ldots, T_{k,x,\theta}(r(x,\cdot,\theta))]$$

where $T_{\ell,x,\theta}$ denotes $T_\ell \circ \mathbb{R}_{x,\theta}^{-1}$. We will use the following notation:

$$\begin{aligned} \rho_\ell(X,\theta) &= \partial\rho(s)/\partial s_\ell|_{s=(T_{1,x,\theta}(0),\ldots,T_{\ell,x,\theta}(0))}, \\ G(X,\theta) &= (\rho_1(X,\theta), \ldots, \rho_k(X,\theta))^\intercal, \\ F(Y) &= (f_1(Y), \ldots, f_k(Y))^\intercal. \end{aligned} \tag{5.13}$$

Also note that, in our case, $T_{\ell,x,\theta}(0) = E_\theta(f_\ell(Y)|X)_x$. Again, we use symbols such as $T_{\ell,x}$ and C_x to indicate T_{ℓ,x,θ_0} and C_{x,θ_0}.

Theorem 11. *Suppose the conditions 1, 2, 3 in Theorem 9 hold for $C_{x,\theta}$. Then the efficient score for $\mathcal{S}_{C(Y|X)}$ is given by (5.8), in which*

$$\begin{aligned} \tau_c(x,y,\theta_0) &= G^\intercal(x,\theta_0)F(y) - G^\intercal(x,\theta_0)E_{\theta_0}(F(Y)|X)_x, \\ q_1(x,\theta_0) &= \partial E_\theta(F^\intercal(Y)|X)_x/\partial\theta|_{\theta=\theta_0} G(x,\theta_0), \\ q_2(x,\theta_0) &= G^\intercal(x,\theta_0)\mathrm{Var}_{\theta_0}(F(Y)|X)_x G(x,\theta_0). \end{aligned} \tag{5.14}$$

PROOF. As shown in Section 5.5, the Riesz representation of $\dot{T}_{\ell,x}(0)$ is simply f_ℓ.

By the chain rule of Fréchet differentiation and definition (5.13), we have

$$\dot{C}_x(0) = \sum_{\ell=1}^k \rho_\ell(X, \theta_0)\, \dot{T}_{\ell,x}(0).$$

Hence the Riesz representation of $\dot{C}_x(0)$ is

$$\tau(x, y, \theta_0) = \sum_{\ell=1}^k \rho_\ell(x, \theta_0) f_\ell(y) = G^\intercal(x, \theta_0) F(y).$$

In the meantime for each θ we have

$$\int \tau_c(x, y, \theta) \eta_0(x, y, \theta) d\mu_Y(y) = 0.$$

Differentiate both sides of this equation with respect to θ, to obtain

$$\int \partial \tau_c(x, y, \theta)/\partial\theta\, \eta_0(x, y, \theta) d\mu_Y(y) + \int \tau_c(x, y, \theta) \mathring{r}_0\,(x, y, \theta) d\mu_Y(y) = 0.$$

Hence

$$\begin{aligned}\int \tau_c(x, y, \theta) \mathring{r}_0\,(x, y, \theta) d\mu_Y(y) &= -E_\theta[\partial \tau_c(X, Y, \theta)/\partial\theta | X] \\ &= -\partial G^\intercal(X, \theta)/\partial\theta\, [F(Y) - E_\theta(F(Y)|X)] + \partial E_\theta(F^\intercal(Y)|X)/\partial\theta\, G(X, \theta).\end{aligned}$$

Now take conditional expectation $E_\theta(\cdots|X)$ on both sides to prove the second relation in (5.14). □

It is easy to see that an alternative expression of $q_1(\theta, X)$ in Theorem 11 is

$$q_1(x, \theta_0) = \partial \rho(E_\theta[f_1(Y)|X]_x, \ldots, E_\theta[f_k(Y)|X]_x)/\partial\theta|_{\theta=\theta_0}.$$

This expression is useful because $\rho(E_\theta[f_1(Y)|X]_x, \ldots, E_\theta[f_k(Y)|X]_x)$ is a function of $\sigma_\theta(X)$, and its derivative with respect to θ can be estimated by local linear regression, as we will see in Section 5.9.

Example 3. For the central variance subspace, we have

$$k = 2,\ f_1(y) = y,\ f_2(y) = y^2,\ \rho(s_1, s_2) = s_2 - s_1^2.$$

Hence $F(y) = (y, y^2)^\intercal$, and

$$\begin{aligned}\rho_1(X, \theta_0) &= \partial(s_2 - s_1^2)/\partial s_1|_{s_1 = E(F(Y)|X)} = -2E(Y|X),\\ \rho_2(X, \theta_0) &= \partial(s_2 - s_1^2)/\partial s_2|_{s_2 = E(F(Y)|X)} = 1.\end{aligned}$$

The Riesz representation of $\dot{C}_x(0)$ and its centered version are

$$\begin{aligned}\tau(X, Y, \theta_0) &= -2E(Y|X)Y + Y^2,\\ \tau_c(X, Y, \theta_0) &= -2E(Y|X)Y + Y^2 - E[-2E(Y|X)Y + Y^2|X]\\ &= [Y - E(Y|X)]^2 - E[\mathrm{Var}(Y|X)|\sigma_{\theta_0}(X)].\end{aligned}$$

Hence,

$$\begin{aligned}q_1(X, \theta_0) &= -\partial E_\theta[\tau_c(X, Y, \theta)]/\partial\theta|_{\theta=\theta_0} = \partial E_\theta[\mathrm{Var}_\theta(Y|X)|\sigma_\theta(X)]/\partial\theta|_{\theta=\theta_0},\\ q_2(X, \theta_0) &= \mathrm{Var}[\tau_c(X, Y, \theta_0)|X] = \mathrm{Var}\{[Y - E(Y|X)]^2|X\}.\end{aligned}$$

In this case the subspace $\mathscr{T}_\phi^\perp$ consists of functions of the form

$$[h(x) - E(h(X)|\sigma_{\theta_0}(X))][(Y - E(Y|X))^2 - E(\mathrm{Var}(Y|X)|\sigma_{\theta_0}(X))],$$

where h is an arbitrary member of $L_2(\lambda_0)$. Interestingly, the estimating equation proposed by Zhu, Dong, and Li (2012) is a special member of $\mathscr{T}_\phi^\perp$ with $h(x)$ taken to be the components of x. □

5.7 Implicit statistical functionals

In this section we study statistical functionals defined implicitly through estimating equations. Many robust estimators, such as conditional medians and quantiles, are of this type. Let $\Xi \subseteq \mathbb{R}$, and $e : \Xi \times \Omega_Y \to \mathbb{R}$ be a function of the parameter ξ and the variable y. Such functions are called estimating functions (see, for example, Godambe, 1960; Li and McCullagh, 1994). If the equation

$$\int_{\Omega_Y} e(\xi, y)g(y)d\mu_Y(y) = 0 \tag{5.15}$$

has a unique solution for each $g \in \mathscr{G}$, then it defines a functional $I : \mathscr{G} \to \mathbb{R}$ that assigns each g the solution to (5.15). We call such functionals implicit functionals, or I-functionals. If we replace $g \in \mathscr{G}$ with a conditional density $\eta(\cdot, \cdot, \theta) \in \mathscr{H}$, then (5.15) becomes

$$\int_{\Omega_Y} e(\xi, y)\eta(x, y, \theta)d\mu_Y(y) = 0.$$

The corresponding statistical functional is $I_{x,\theta}(r(x, \cdot, \theta)) = I \circ \mathbb{R}^{-1}_{x,\theta}(r(x, \cdot, \theta))$. The I-central subspace is defined by the statement

$$I_X(0) \text{ is measurable with respect to } \sigma_{\theta_0}(X).$$

Naturally, we write the function $(\theta, x) \mapsto I_{x,\theta}(0)$ as $\xi(\theta, x)$.

To simplify the presentation we use the notion of generalized functions. Let $\mathscr{K}$ be the class of functions defined on a bounded set B in $\mathbb{R}$ that have derivatives of all orders, whose topology is defined by the uniform convergence of all derivatives. Any continuous linear functional $U : \mathscr{K} \to \mathbb{R}$ with respect to this topology is called a generalized function. For example, let $a \in B$. Then it can be shown that the linear functional

$$\delta_a : \mathscr{K} \to \mathbb{R}, \quad \phi \mapsto \phi(a)$$

is continuous with respect to this topology. This continuous linear functional is called the Dirac delta function. We identify the functional δ_a with an imagined function $x \mapsto \delta_a(x)$ on B and write $\delta_a(\phi)$ formally as the integral

$$\delta_a(\phi) = \int \phi(x)\delta_a(x)d\lambda(x).$$

A consequence of this convention is that if we pretend $\delta_a(x)$ to be the derivative $\partial I(x \le a)/\partial a$ of the indicator function $I(x \le a)$ then we get correct answers at the integral level. For example, for any constant a and small number ϵ, we have

$$\int [I(y \le a + \epsilon) - I(y \le a)]g(y)dy = \int \delta_a(y)\epsilon g(y)dy + o(\epsilon) = \epsilon g(a) + o(\epsilon).$$

Thus, the pretended linearization $I(y \le a + \epsilon) - I(y \le a) = \delta_a(y)\epsilon + o(\epsilon)$ has caused no inconsistency. We use this device to simplify our presentation of

quantiles.

Theorem 12. *Suppose the conditions 1, 2, 3 in Theorem 9 hold for $I_{x,\theta}$. Moreover, suppose for each $\xi \in \Xi$, $g \in \mathscr{G}$, there is a (generalized) function $\dot{e}(\xi, y)$, which plays the role of $\partial e(\xi, y)/\partial\xi$, such that*

$$\left| \int_{\Omega_Y} [e(\xi + a, y) - e(\xi, y) - \dot{e}(\xi, y)a] g(y) d\mu_Y(y) \right| = o(a). \tag{5.16}$$

Then the efficient score for the I-central subspace is (5.8) in which

$$\begin{aligned} \tau_c(x, y, \theta_0) &= -e(\xi(\theta_0, x), y)/E_{\theta_0}[\dot{e}(\xi(\theta_0, x), Y)|X]_x, \\ q_1(x, \theta_0) &= \partial\xi(\theta, x)/\partial\theta|_{\theta=\theta_0}, \\ q_2(x, \theta_0) &= E_{\theta_0}[e^2(\xi(\theta, X), Y)|X]_x / E^2_{\theta_0}[\dot{e}(\xi(\theta_0, X), Y)|X]_x. \end{aligned} \tag{5.17}$$

PROOF. Differentiating both sides of the equation

$$\int_{\Omega_Y} e(I_{x,\theta}(\epsilon r(x, \cdot, \theta)), y)(1 + \epsilon r(x, y, \theta)/2)^2 \eta_0(x, y, \theta) d\mu_Y(y) = 0$$

with respect to ϵ at $\epsilon = 0$, and using the relation

$$\partial(1 + \epsilon r(x, y, \theta)/2)^2/\partial\epsilon|_{\epsilon=0} = 2(1 + 0r(x, y, \theta)/2) r(x, y, \theta)/2 = r(x, y, \theta),$$

we find

$$\partial I_{x,\theta}(\epsilon r(x, \cdot, \theta))/\partial\epsilon|_{\epsilon=0} = -E\left[\frac{e(I_{x,\theta}(0), Y)}{E(\dot{e}(I_{x,\theta}(0), Y)|X)_x} r(x, Y, \theta)|X \right]_x.$$

Hence the Riesz representation of the Fréchet derivative $\dot{I}_{x,\theta}(0)$ is

$$\begin{aligned} \tau(x, y, \theta) &= -e(I_{x,\theta}(0), y)/E[\dot{e}(I_{x,\theta}(0), Y)|X]_x \\ &= -e(\xi(\theta, x), y)/E[\dot{e}(\xi(\theta, x), Y)|X]_x = \tau_c(x, y, \theta), \end{aligned} \tag{5.18}$$

where the last equality holds because, by definition, $E[e(\xi(\theta, X), Y)|X] = 0$. By (5.18) and the definition of q_1 in Theorem 9,

$$q_1(x, \theta_0) = -E[e(\xi(\theta_0, x), y)\mathring{r}_0(x, y, \theta_0)|X]_x / E[\dot{e}(\xi(\theta_0, x), Y)|X]_x. \tag{5.19}$$

To further simplify the numerator of the right hand side, differentiate both sides of the equation $\int e(\xi(\theta,x),y)\kappa_0(x,y,\theta)d\mu_Y(y) = 0$ to obtain

$$\begin{aligned}\left[\int \dot{e}(\xi(\theta_0,x),y)\,\kappa_0(x,y,\theta_0)d\mu_Y(y)\right]\dot{\xi}(\theta_0,x)\\ + \int e(\xi(\theta_0,x),y)\mathring{r}_0(x,y,\theta_0)\kappa_0(x,y,\theta_0)d\mu_Y(y) = 0,\end{aligned}$$

where $\dot{\xi}(\theta_0,x)$ denotes $\partial\xi(\theta,x)/\partial\theta|_{\theta=\theta_0}$. Hence,

$$E[e(\xi(\theta_0,X),Y)\mathring{r}_0(X,Y,\theta_0)|X]_x = -E[\dot{e}(\xi(\theta_0,X),Y)|X]_x\,\dot{\xi}(\theta_0,x).$$

Substitute this into (5.19) to prove the first relation in (5.17). Substitute (5.18) into the definition of q_2 in Theorem 9 to prove the second relation in (5.17). □

In the next example we derive the efficient score for a particular type of I-functional — the quantile functional.

Example 4. If we assume all densities in $\mathscr{G}$ are continuous, then the pth quantile is the solution to the equation $EI(Y \le \xi) = p$. Equivalently, ξ is the solution to the equation

$$E[e(\xi,y)] = E[-\mathrm{sgn}(y-\xi) + 1 - 2p] = 0.$$

Because the generalized derivative of $\mathrm{sgn}(t)$ is $2\delta_0(t)$, we have $\dot{e}(\xi,y) = 2\delta_\xi(y)$. Hence

$$E[\dot{e}(\xi(\theta,X),Y)|X]_x \;=\; \textstyle\int_{\Omega_Y} 2\delta_{\xi(\theta,x)}(y)\eta_0(x,y,\theta)dy = 2\eta_0(x,\xi(\theta,x),\theta).$$

Because $e(\xi(\theta,x),Y)$ is a binary random variable that takes the value $2(1-p)$ with probability p and $-2p$ with probability $(1-p)$, we have

$$E[e^2(\xi(\theta,x),Y)|X]_x \;=\; [2(1-p)]^2p + (-2p)^2(1-p) = 4(1-p)p.$$

Hence, in the efficient score,

$$q_2(x,\theta_0) = \eta_0^2(x,\xi(\theta_0,x),\theta_0)/[(1-p)p],$$

and $q_1(x,\theta_0)$ is as given in (5.17) with $\xi(\theta,x)$ being the conditional pth quantile of Y given X. □

5.8 Effect of estimating the central subspace

Throughout the previous sections we have treated X as the true predictor from the central subspace; that is, $X = \zeta^\intercal \tilde{X}$ where $\tilde{X}$ is the original predictor and ζ is a basis matrix of the central subspace based for $Y|\tilde{X}$. However, in practice, ζ itself needs to be estimated and, in theory at least, should affect the form of the efficient score about β. While our simulation studies in Section 5.10 indicate that this effect is very small, for theoretical rigor we present here the efficient score treating the central subspace as an additional (finite-dimensional) nuisance parameter. For convenience, we use ζ to denote both a $p \times d$ matrix and the corresponding $(p-d)d$-dimensional Grassmann manifold.

Let $S_{\text{eff}}(\zeta^\intercal \tilde{X}, Y, \theta)$ denote the efficient score in Theorem 9 with X replaced by $\zeta^\intercal \tilde{X}$. Let $S^*_{\text{eff}}(\tilde{X}, Y, \zeta, \theta)$ denote the efficient score for θ with ζ treated as an additional nuisance parameter. Let $s_\zeta(\tilde{x}, y, \zeta, \theta)$ denote the score with respect to ζ. In the Appendix (Section 5.13) it is shown that

$$S^*_{\text{eff}} = S_{\text{eff}} - \Pi(S_{\text{eff}} \,|\, \mathcal{S}(\Pi(s_\zeta|\mathscr{T}_\phi^\perp))) \equiv S_{\text{eff}} - g, \tag{5.20}$$

where g is the function

$$\begin{aligned} g = \quad & q_2^{-1}(\zeta_0^\intercal \tilde{x}, \theta_0)\tau_c(\zeta_0^\intercal \tilde{x}, y, \theta_0) E\{q_3^*(\zeta_0^\intercal \tilde{X}, \theta_0) q_3^{*\intercal}(\zeta_0^\intercal \tilde{X}, \theta_0) q_2^{-1}(\zeta_0^\intercal \tilde{x}, \theta_0)\} \\ & E\{q_3^*(\zeta_0^\intercal \tilde{X}, \theta_0) q_3^{*\intercal}(\zeta_0^\intercal \tilde{X}, \theta_0) q_2^{-1}(\zeta_0^\intercal \tilde{x}, \theta_0)\}^\dagger \, q_3^*(\zeta_0^\intercal \tilde{x}, \theta_0), \end{aligned}$$

in which

$$\begin{aligned} q_1^*(\zeta_0^\intercal \tilde{x}, \theta_0) &= E[s_\zeta(\tilde{X}, Y, \zeta_0, \theta_0)\tau_c(\zeta_0^\intercal \tilde{X}, Y, \theta_0)|\zeta_0^\intercal \tilde{X}]_{\tilde{x}} \\ q_3^*(\zeta_0^\intercal \tilde{x}, \theta_0) &= q_1^*(\zeta_0^\intercal \tilde{x}, \theta_0) \\ & \quad - E_{\theta_0}[q_1^*(\tilde{X}, \zeta_0, \theta_0) q_2^{-1}(\zeta_0^\intercal \tilde{x}, \theta_0)|\sigma_{\theta_0}(\zeta_0^\intercal \tilde{X})]_{\tilde{x}} / E[q_2^{-1}(\zeta_0^\intercal \tilde{x}, \theta_0)|\sigma_{\theta_0}(\zeta_0^\intercal \tilde{X})]_{\tilde{x}}. \end{aligned}$$

In theory, the asymptotic variance bound based on S_{eff} is lower than or equal to that based on S^*_{eff}. However, in the simulation studies (Table 5.2 in Section 5.10) we see that the theoretical lower bound based on S_{eff} is nearly reached, which indicates that effect of estimating the central subspace on the efficient score for β is very small.

5.9 Estimation

In this section we introduce semiparametrically efficient estimators using the theory developed in the previous sections. For the L-functionals we develop the estimator in full generality, but for the C- and I-functionals we focus on the conditional variance functional and the conditional quantile functional. Procedures for other C- or I-functionals can be developed by analogy.

We first clarify two points related to the algorithm we will propose. First, since we will rely heavily on the MAVE-type algorithms, it is more convenient to use the β-parameterization rather than the θ-parameterization, and avoid redundancy in β by taking the generalized matrix inverse. We will justify the β-parameterization after introducing the algorithm. Second, the MAVE algorithm actually has two variants: the outer product gradient (OPG) and a refined version of MAVE (RMAVE). Typically, OPG, MAVE, and RMAVE are progressively more accurate and require more computation. In the following, the MAVE-type algorithm can be replaced either by RMAVE for greater accuracy or by OPG for less computation.

Our estimation procedure is divided into four steps. In step 1, we estimate the central subspace and project $\tilde{X}$ on to this subspace to obtain X. In step 2, we estimate $T(\eta_0(X_1,\cdot)),\ldots,T(\eta_0(X_n,\cdot))$ using a d-dimensional kernel estimate. These estimates are used as the proxy response, and we denote them by $\hat{Y}_1,\ldots,\hat{Y}_n$. In step 3, we apply MAVE to $(X_1,\hat{Y}_1),\ldots,(X_n,\hat{Y}_n)$ to estimate an initial value for β. In step 4, we use one-step Newton-Raphson algorithm based on the efficient score and efficient information to approximate the semiparametrically efficient estimate. We call our estimator SEE, which stands for semiparametrically efficient estimator.

Preparation step: a MAVE code. Let $(X_1,U_1),\ldots,(X_n,U_n)$ be a random sample from (X,U), and $K_h(\cdot)$ be a kernel with bandwith h. That is, $K_h(t)=K(t/h)/h$ for some symmetric function K that integrates to 1 and $h>0$. Compute the objective function

$$\Gamma(a,b,A)=\sum_{i=1}^n\sum_{j=1}^n\left[(U_j-a_i-b_i^\tau A^\tau(X_j-X_j)\right]^2K_h(X_j-X_i) \quad (5.21)$$

where $a_1,\ldots,a_n\in\mathbb{R}$, $b_1,\ldots,b_n\in\mathbb{R}^d$, a and b on the left denote $(a_1,\ldots,a_n)^\tau$ and $(b_1^\tau,\ldots,b_n^\tau)^\tau$ respectively, and $A\in\mathbb{R}^{p\times d}$. We will use this objective function in several ways. It is well known that minimization of (5.21) can be solved

by iterations of least squares, and in each iteration there is an explicit solution. Thus purely search-based numerical optimization (such as the simplex method) is avoided. See, for example, Li, Li, and Zhu (2010) and Yin and Li (2011).

Step 1: estimation of central subspace. We use the MAVE-ensemble in Yin and Li (2011) to estimate the central subspace. In this procedure, U in (5.21) is taken to be a set of functions $\{f_1(Y), \ldots, f_m(Y)\}$ randomly sampled from a dense family in $L_2(\mu_Y)$. In this chapter we take this set to be $\{(\sin(t_i y), \cos(t_i y)) : i = 1, \ldots, 10\}$, where $t_1, \ldots, t_{10}$ are i.i.d. unif$(0, 4)$. The sample of responses $Y_1, \ldots, Y_n$ are standardized so that the range $(0, 4)$ of the uniform distribution represents a reasonably rich class of functions relative to the range of Y. The basis matrix ζ of $\mathcal{S}_{Y|X}$ is then estimated by the MAVE-ensemble. The projected predictor $X = \hat{\zeta}^{\mathsf{T}}\tilde{X}$ is taken as the predictor in steps 2–4. Since our goal is to estimate $\mathcal{S}_{T(Y|X)}$, the choice of the working dimension $\hat{d}$ of $\mathcal{S}_{Y|X}$ is not crucial.

As was shown in Yin and Li (2011), at the population level, the MAVE ensemble is guaranteed to recover central subspace exhaustively as long as the functions of Y form a characterizing family. In practice, it is true any information lost in the initial step will be inherited by SEE. However, our experiences indicate that this problem can be mitigated by using a sufficiently rich ensemble family — for example, by increasing the range of the uniform distribution and the number of t_i's.

Several other methods are available for exhaustive estimation of the central subspace, such as the semiparametrically efficient estimator of Ma and Zhu (2013a), DMAVE by Xia (2007), and Sliced Regression by Wang and Xia (2008). Here, we have chosen the MAVE-ensemble for its computational simplicity.

Step 2: estimation of proxy response. This step is trivial for the L-functionals: because $T(\eta(X, \cdot)) = E[f(Y)|X]$ for some function f, we can use $f(Y)$ itself as the proxy response $\hat{Y}$. For the conditional variance functionals, this step needs not be fully performed: we can use $Y^2 - \hat{E}^2(Y|X)$ as the proxy response $\hat{Y}$, where $\hat{E}(Y|X)$ is the kernel estimator of $E(Y|X)$. If simplification of this type is not applicable, then we need to perform nonparametric estimation of $T(\eta(X, \cdot))$. For example, for the I-functionals, we estimate the proxy response $\hat{Y}_i$ by the minimizer ξ_i^* of the function $E_n[e(Y, \xi)K_h(X - X_i)]$.

Step 3: initial estimate of β. Apply MAVE to $(X_1, \hat{Y}_1), \ldots, (X_n, \hat{Y}_n)$ to obtain an initial estimate of β by minimizing $\Gamma(a, b, \beta)$ over all a, b, β. Denote this initial

estimate as $\tilde{\beta}$.

Step 4: the one-step Newton-Raphson algorithm. Rather than attempting to solve the score equation (5.4), we propose a one-step Newton-Raphson procedure. Let $S_{\text{eff}}(X, Y, \beta)$ be the efficient score in β, which is obtained by replacing $\sigma_\theta(X)$ by $\beta^\intercal X$ and $\partial/\partial\theta$ by $\partial/\partial\text{vec} \circ (\beta)$ whenever applicable. Let $J_{n,\text{eff}}(\beta) = E_n[S_{\text{eff}}(X, Y, \beta)S_{\text{eff}}^\intercal(X, Y, \beta)]$. We estimate β by

$$\hat{\beta} = \tilde{\beta} + [J_{n,\text{eff}}(\tilde{\beta})]^\dagger E_n[S_{\text{eff}}(X, Y, \tilde{\beta})] \tag{5.22}$$

where $\dagger$ is the Moore-Penrose inverse, and $\tilde{\beta}$ is the initial value from Step 3.

We now describe in detail how to compute $E_n[S_{\text{eff}}(X, Y, \tilde{\beta})]$ for different functionals. For the L-functional, the efficient score involves the following functions

$$(a) \quad E[f(Y)|\beta^\intercal X], \quad (b) \quad E[f^2(Y)|X], \quad (c) \quad \partial E[f(Y)|\beta^\intercal X]/\partial\text{vec} \circ (\beta).$$

For the C-functional, it involves the functions:

$$(d) \quad E_{\theta_0}[f_\ell(Y)|X], \quad (e) \quad \partial\rho(E_\beta[f_1(Y)|X], \ldots, E_\beta[f_\ell(Y)|X])/\partial\text{vec} \circ (\beta),$$
$$(f) \quad \text{cov}[f_i(Y), f_j(Y)|X].$$

For the conditional quantile functional, it involves the functions

$$(g)\, E[\xi(X)|\beta^\intercal X], \quad (h)\, \partial E[\xi(X)|\beta^\intercal X]/\partial\text{vec} \circ (\beta), \quad (i)\, \eta_0(X, E[\xi(X)|\beta^\intercal X]).$$

These functions can be categorized into three types: (a), (b), (d), (f), (g) are conditional means of random variables; (c), (e), (h) are derivatives of functions of $\beta^\intercal X$ with respect to vec $\circ\, (\beta)$; (i) is the conditional density evaluated at a quantile. The first two types can be solved by minimizing $\Gamma(a, b, A)$ in (5.21) with specific U_i and A: the following table gives the details of these random variables and matrices, as well as which parts of the MAVE output are needed in SEE.

Finally, we estimate $\eta_0(x, y_0)$ for any y_0 by the kernel conditional density estimator:

$$E_n[K_{h_1}(Y - y_0)K_h(X - x)]/E_n[K_h(X - x)],$$

where h_1 is a different bandwidth for the response variable.

Table 5.1. Using MAVE to estimate quantities (a) through (h) in efficient score

quantities	A	U_i	MAVE output
(a)	β	$f(Y)$	a^*
(b)	I_p	$f^2(Y)$	a^*
(c)	β	$f(Y)$	b^*
(d)	I_p	$f_\ell(Y)$	a^*
(e)	β	$\rho(E_n[f_1(Y)\|X], \ldots, E_n[f_k(Y)\|X])$	b^*
(f)	I_p	$f_\ell(Y)$, $f_\ell(Y)f_{\ell'}(Y)$	a^*
(g)	β	$\xi_n(X)$	a^*
(h)	β	$\xi_n(X)$	b^*

Tuning of bandwidths. We need to determine the kernel bandwidths h at various stages in the above algorithm. We use the Gaussian kernel with optimal bandwidth $h = cn^{-1/(p+4)}$ for nonparametric regression (see, for example, Xia, Tong, Li, and Zhu, 2002), where c is determined by five-fold cross validation. That is, we randomly divide the data into 5 subsets of roughly equal sizes. For each $i = 1, \ldots, 5$, we use the ith subset as the testing set and the rest as the training set. For a given c, we conduct dimension reduction on the training set, and use the sufficient predictor to evaluate a certain prediction criterion at each point on the testing set and average these evaluations over the testing set, and finally average the five averages of the criterion to obtain a single number. We then minimize the resulting criterion over c by a grid search. Naturally, the prediction criterion depends on the object to be estimated using that kernel. Below is a list of the prediction criteria we propose for the four steps in the estimation procedure.

In Step 1: We use the distance correlation introduced by Székely and Rizzo (2009, Theorem 1, expression (2.11); p and q therein are taken to be 2).

In Step 2: For the L-functionals, no tuning is needed. For the conditional variance functional, we need to estimate $E(Y|X)$, and we use the prediction criterion $[Y - \hat{E}(Y|X)]^2$, where $\hat{E}(Y|X)$ is the kernel estimate of $E(Y|X)$ based on the training set using a tuning constant c. For conditional median, we use the prediction criterion $|Y - \hat{M}(Y|X)|$, where $\hat{M}(Y|X)$ is the kernel estimate of the conditional median based on the training set using a specific tuning constant.

In Step 3: We use the prediction criterion $[\hat{Y} - \hat{E}(\hat{Y}|\tilde{\beta}^\tau X)]^2$, where $\hat{Y}$ is the proxy

response obtained from Step 2, and $\hat{E}(\hat{Y}|\tilde{\beta}^\tau X)$ is the MAVE output a^*.

In Step 4: There are three types of kernels in this step: the kernel for X, the kernel for $\tilde{\beta}^\tau X$, and the kernel for Y (the last one is needed only for the conditional median functional). Corresponding to these types we use bandwidths

$$h = c\, n^{-1/(d+4)}, \quad h = c\, n^{-1/(s+4)}, \quad h = c\, n^{-1/(1+4)}.$$

We use cross validation to determine the common c, using the estimate of β from the one-step Newton-Raphson algorithm. Once again, we use different prediction criteria for different functionals. For the mean functional, we use the criterion $[Y - \hat{E}(Y|\hat{\beta}^\tau X)]^2$. For the conditional median functional, we use the criterion $|Y - \hat{E}[\hat{M}(Y|X)|\hat{\beta}^\tau X]|$. For conditional variance functional we use the criterion $\{(Y - \hat{E}(Y|X))^2 - \hat{E}[Y - \hat{E}(Y|X)|\hat{\beta}^\tau X]^2\}^2$.

Justification of parameterization. We now justify the one-step iteration formula (5.22) as an equivalent form of

$$\hat{\theta} = \tilde{\theta} + J_{n,\text{eff}}(\tilde{\theta})^{-1} E_n[S_{\text{eff}}(X, Y, \tilde{\theta})]. \tag{5.23}$$

Since β is a function of a $s(d-s)$ dimensional parameter θ, the efficient information $J_{n,\text{eff}}(\beta)$ has rank $s(d-s)$. Let Γ denote the $sd \times [s(d-s)]$ matrix whose columns are eigenvectors of $J_{n,\text{eff}}(\beta)$ corresponding to its nonzero eigenvalues, and let $\beta = \Gamma\theta$, where θ is a free parameter in $\mathbb{R}^{s(d-s)}$. In this parameterization,

$$S_{\text{eff}}(X, Y, \theta) = \Gamma^\tau S_{\text{eff}}(X, Y, \beta), \quad J_{n,\text{eff}}(\theta) = \Gamma^\tau J_{n,\text{eff}}(\beta)\Gamma.$$

Hence (5.23) is equivalent to

$$\hat{\theta} = \tilde{\theta} + [\Gamma^\tau J_{n,\text{eff}}(\tilde{\beta})\Gamma]^{-1}\Gamma^\tau E_n[S_{\text{eff}}(X, Y, \tilde{\beta})].$$

Multiply both sides by Γ from the left, we find

$$\Gamma\hat{\theta} = \Gamma\tilde{\theta} + \Gamma[\Gamma^\tau J_{n,\text{eff}}(\tilde{\beta})\Gamma]^{-1}\Gamma^\tau E_n[S_{\text{eff}}(X, Y, \tilde{\beta})].$$

By construction, $\Gamma\hat{\theta} = \hat{\beta}$, $\Gamma\tilde{\theta} = \tilde{\beta}$, and $\Gamma[\Gamma^\tau J_{n,\text{eff}}(\tilde{\beta})\Gamma]^{-1}\Gamma^\tau$ is the Moore-Penrose inverse of $J_{n,\text{eff}}(\tilde{\beta})$. Thus the above iterative formula is the same as (5.22).

In concluding this section we point out two attractive features of our algorithm, which were briefly touched on in the Introduction. First, since our algorithm is implemented by repeated applications of variations of MAVE, it essentially consists of a sequence of least squares algorithms, thus avoiding any search-based numerical optimization, which can be infeasible when the dimension of θ is high. The second advantage is that, since our algorithm is based on the β-parameterization, we do not need any subjectively chosen parameterization. In comparison, Ma and Zhu (2013a) used the parameterization $\beta = (I_s, \theta)^\intercal$, where $\theta \in \mathbb{R}^{(d-s)\times s}$ is a matrix with free-varying entries. Note that this is not without loss of generality, because in reality the first d rows of β can be linearly dependent.

5.10 Simulation comparisons

In this section we conduct simulation comparisons between SEE and other methods for estimating three types of T-central subspaces: $\mathcal{S}_{E(Y|X)}$, $\mathcal{S}_{V(Y|X)}$, and $\mathcal{S}_{M(Y|X)}$. We use the distance between two subspaces proposed by Li, Zha, and Chiaromonte (2005) to measure estimation errors, which is defined as

$$\text{dist}(\mathcal{S}_1, \mathcal{S}_2) = \|\Pi_{\mathcal{S}_1} - \Pi_{\mathcal{S}_2}\|_2, \tag{5.24}$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ are subspaces of $\mathbb{R}^p$, and $\|\cdot\|_2$ is the L_2 norm in $\mathbb{R}^{p\times p}$. For each of the following models, the sample size is taken to be $n = 200$ or 500, or both; each sample is repeatedly drawn for $n_{\text{sim}} = 100$ times in the simulation. In all simulations we fix the working dimension in Step 1 at $\hat{d} = 3$, even though d is no greater than 2 in all examples.

The explicit forms of the efficient scores efficient information for Models I $\sim$ VII used in the following comparisons are derived in the Appendix (Section 5.13).
(a) Comparison for the central mean subspace. In this case the functional $T(\eta(X,\cdot))$ is the conditional mean $E(Y|X)$. We compare SEE with RMAVE under the following models

$$\begin{aligned}
&\text{Model I}: \; Y = X_1 + (1 + |X_2|)\,\varepsilon,\\
&\text{Model II}: \; Y = X_1(X_1 + X_2 + 1) + 0.5\varepsilon,\\
&\text{Model III}: \; Y = X_1 + (1 + |X_1|)\,\varepsilon,
\end{aligned}$$

$$\text{Model IV}: \; Y|X \sim \text{Poisson}(|X_1 + X_2|),$$

where $X \sim N(0, I_{10})$, $X \perp\!\!\!\perp \varepsilon$, and $\epsilon \sim N(0,1)$. These models represent a variety of scenarios one might encounter in practice. Specifically, the central mean subspace is a proper subspace of the central subspace in Model I, but coincides with the latter in the other models. The conditional variance $\text{Var}(Y|X)$ is a constant in Model II, but depends on X in the other models. Because of its additive error structure Model II is favorable to MAVE. Finally, model IV has a discrete response and the error only enters implicitly. Model I and Model III will be used again for Comparison 2, where conditional variance is the target; Model II was also used in Li (1991) and Xia, Tong, Li, and Zhu (2002).

The results with sample sizes $n = 200$ and $n = 500$ are presented in Table 5.2, in the blocks indicated by $E(Y|X)$. The entries are in the form $a(b)$, where a is the mean, and b the standard error, of the distance (5.24) between the true and estimated $\mathcal{S}_{E(Y|X)}$, based on $n_{\text{sim}} = 100$ simulated samples.

(b) Comparison for the central variance subspace. Let $T(\eta(X,\cdot))$ be the conditional variance $\text{Var}(Y|X)$. We compare SEE with the estimator proposed in Zhu and Zhu (2009) and Zhu, Dong, and Li (2012). In Model I, the central variance subspace is different from either the central mean subspace or the central subspace, while in Model III, the three spaces coincide.The results with $n = 100$ and $n = 500$ are reported in Table 5.2, in the blocks indicated by $\text{Var}(Y|X)$.

(c) Comparison for the central median subspace. Let $T(\eta(X,\cdot))$ be the conditional median $M(Y|X)$. We compare SEE with the adaptive quantile estimator (AQE) introduced by Kong and Xia (2012), which can also be used to estimate the central median subspace. We use the following models:

$$\text{Model V}: \; Y = X_1^2 + X_2\varepsilon, \quad \text{Model VI}: \; Y = 3X_1 + X_2 + \varepsilon,$$

where $X \sim N(0, I_{10})$ and $X \perp\!\!\!\perp \varepsilon$. For model V, ε has a skewed-Laplace distribution with p.d.f.

$$f(\varepsilon) = \begin{cases} (5/4)\, e^{-5\varepsilon/2} & \varepsilon \geq -(2/5)\log(4/3) \\ (80/27)\, e^{5\varepsilon} & \varepsilon < -(2/5)\log(4/3) \end{cases}.$$

In this case,

$$E(Y|X) = X_1^2 + X_2[1/5 - 2/5\log(4/3)], \quad M(Y|X) = X_1^2.$$

It follows that

$$\mathcal{S}_{E(Y|X)} = \mathcal{S}\{(1,0,\ldots,0)^\intercal, (0,1,0,\ldots,0)^\intercal\}, \quad \mathcal{S}_{M(Y|X)} = \mathcal{S}\{(1,0,\ldots,0)^\intercal\}.$$

For model VI, $\varepsilon \sim t_{(3)}$. Although for this model the central mean subspace coincides with the central median subspace, due to the heavy-tailed error distribution the conditional median is preferred to the conditional mean. Similar models can be found, for example, in Zou and Yuan (2008). The results for sample sizes $n = 200, 500$ are presented in Table 5.2, in the blocks indicated by $M(Y|X)$.

(d) Comparison with theoretical lower bound To see how closely the theoretical asymptotic lower bound (ALB) is approached by SEE for finite samples, we now compute the limit

$$\lim_{n\to\infty} \sqrt{n}\, E(\|\Pi_{\text{span}(\hat{\beta})} - \Pi_{\text{span}(\beta_0)}\|_2), \tag{5.25}$$

where $\hat{\beta}$ is the semiparametrically efficient estimate. This is the best we can do to estimate the T-central subspace. The explicit form and the derivation of (5.25) is given in the Appendix (Section 5.13). We present the numerical values of this limit in the last column (under the heading ALB) of Table 5.2 for different models and T functionals.

(e) Conclusions for comparisons in (a) $\sim$ (d) From Table 5.2 we see that SEE achieves substantially improved accuracy across all models and functionals considered. Stability of the estimates is also improved as can be seen from the decrease in standard errors. Our simulation studies (not presented here) indicate that the results are not significantly affected by the working dimension of the central subspace. For example, we repeated the analysis with $d = 2, 4$ and the patterns of the comparisons are not significantly altered.

Comparing the results for $n = 200$ and $n = 500$, we see that the proportion of improvement is smaller for the large sample size, as to be expected.

We see that the actual errors of the SEE computed from simulations are very close to the theoretical lower bounds both sample sized $n = 200, 500$ and the

Table 5.2. Comparison of SEE with other estimators for three statistical functionals

n	Functionals		Competing Estimators		
200	$E(Y\|X)$		RMAVE	SEE	ALB
200	$E(Y\|X)$	I	0.519 (0.127)	0.153 (0.067)	0.175
200	$E(Y\|X)$	II	0.164 (0.063)	0.124 (0.054)	0.115
200	$E(Y\|X)$	III	0.490 (0.156)	0.165 (0.058)	0.147
200	$E(Y\|X)$	IV	0.206 (0.072)	0.078 (0.036)	0.071
200	$\mathrm{Var}(Y\|X)$		Zhu, Zhu; Zhu, Dong, Li	SEE	ALB
200	$\mathrm{Var}(Y\|X)$	I	0.656 (0.231); 0.283 (0.126)	0.125 (0.051)	0.116
200	$\mathrm{Var}(Y\|X)$	III	0.843 (0.197); 0.408 (0.193)	0.116 (0.055)	0.113
200	$M(Y\|X)$		AQE	SEE	ALB
200	$M(Y\|X)$	V	0.029 (0.012)	0.019 (0.013)	0.016
200	$M(Y\|X)$	VI	0.087 (0.022)	0.049 (0.019)	0.043
500	$E(Y\|X)$		RMAVE	SEE	ALB
500	$E(Y\|X)$	I	0.100 (0.030)	0.081 (0.021)	0.083
500	$E(Y\|X)$	II	0.081 (0.018)	0.073 (0.013)	0.073
500	$E(Y\|X)$	III	0.095 (0.033)	0.047 (0.015)	0.046
500	$E(Y\|X)$	IV	0.079 (0.015)	0.047 (0.010)	0.045
500	$\mathrm{Var}(Y\|X)$		Zhu, Zhu; Zhu, Dong, Li	SEE	ALB
500	$\mathrm{Var}(Y\|X)$	I	0.315 (0.066); 0.219 (0.034)	0.109 (0.020)	0.104
500	$\mathrm{Var}(Y\|X)$	III	0.236 (0.035); 0.183 (0.037)	0.071 (0.025)	0.066
500	$M(Y\|X)$		AQE	SEE	ALB
500	$M(Y\|X)$	V	0.017 (0.005)	0.009 (0.003)	0.010
500	$M(Y\|X)$	VI	0.042 (0.015)	0.031 (0.009)	0.027

differences become negligible for $n = 500$. Since in the estimator the central subspace is estimated from the sample and in the lower bounds the central subspace is treated as known, the closeness of these errors to their corresponding ALB also indicates that the lower bound based on S^*_{eff} in Section 5.8 is close to the lower bound based on S_{eff}. In other words the effect of estimating the central subspace on the efficient score is small.

(f) Comparison under dependent components of X We now repeat comparisons in (a) through (c) using an X with dependent components. Rather than taking $\mathrm{Var}(X) = I_{10}$ we now take

$$\mathrm{cov}(X^i, X^j) = 0.5^{|i-j|}, \quad i, j = 1, \ldots 10. \tag{5.26}$$

Table 5.3. Comparison of SEE with other estimators with correlated predictors

Functionals	Model	Competing Estimators		
$E(Y\|X)$		RMAVE	SEE	ALB
$E(Y\|X)$	I	$0.520\,(0.155)$	$0.164\,(0.066)$	0.168
$E(Y\|X)$	II	$0.404\,(0.165)$	$0.160\,(0.080)$	0.149
$E(Y\|X)$	III	$0.571\,(0.134)$	$0.304\,(0.075)$	0.289
$E(Y\|X)$	IV	$0.283\,(0.090)$	$0.075\,(0.023)$	0.085
$\mathrm{Var}(Y\|X)$		Zhu, Zhu; Zhu, Dong, Li	SEE	ALB
$\mathrm{Var}(Y\|X)$	I	$0.539\ (0.174)$; $0.431\ (0.230)$	$0.131\ (0.077)$	0.108
$\mathrm{Var}(Y\|X)$	III	$0.617\ (0.222)$; $0.303\ (0.169)$	$0.204\ (0.066)$	0.227
$M(Y\|X)$		AQE	SEE	ALB
$M(Y\|X)$	V	$0.074\ (0.023)$	$0.015\ (0.012)$	0.012
$M(Y\|X)$	VI	$0.076\ (0.061)$	$0.032\ (0.015)$	0.041

The same covariance matrix was used in Ma and Zhu (2012). The results parallel to those in Table 5.2 are presented in Table 5.3. We see that the errors are larger than those for X with independent components, but the degree by which SEE improves upon the other estimators, and to which it approaches theoretical asymptotic lower bound, are similar to those for the independent-component case.

(g) Comparison for conditional upper quartile We now apply SEE to estimating the central upper-quartile subspace in which the functional of interest is solution to the equation $P(Y \leq c|X) = 0.75$. We generate X from $N(0, \Sigma)$ with Σ given by (5.26). We compare SEE with AQE for models V and VI, and the additional model

$$\text{Model VII}: \; Y = 1 + X_1 + (1 + 0.4\, X_2)\, \varepsilon,$$

where $\varepsilon \sim N(0, 1)$. In Model V, the central upper-quartile subspace has dimension 2, spanned by $(1, 0, \ldots, 0)^\intercal$ and $(0, 1, 0, \ldots, 0)^\intercal$; in Models VI and VII, the central upper-quartile subspaces have dimension 1 and are spanned by $(3, 1, 0, \ldots, 0)^\intercal$ and $(1, 0.4\Phi^{-1}(0.25), 0, \ldots, 0)^\intercal$, respectively, where Φ is the c.d.f. of the standard normal distribution. The performance of the estimators is summarized in Table 5.4. We see that SEE outperforms AQE both in average accuracy and estimation stability. It is also interesting to note that, for Model VI, the central median subspace coincides with the central upper-quartile subspace, and the SEE based on the conditional median (Table 5.3) performs better than the SEE based on the conditional

Table 5.4. Comparison of SEE with AQE at $\tau = 0.75$

Models	AQE	SEE	ALB
V	$0.176\ (0.083)$	$0.043\ (0.017)$	0.042
VI	$0.160\ (0.037)$	$0.069\ (0.021)$	0.053
VII	$0.245\ (0.117)$	$0.109\ (0.072)$	0.111

upper quartile (Table 5.4).

5.11 Application: age of abalones

In this section we evaluate the performance of SEE in an application, which is concerned with predicting the age of abalones using their physical measurements. The data can be found at the website http://archive.ics.uci.edu/ml/datasets.html, and consist of observations from 4177 abalones, of which 1528 are male, 1307 are female, and 1342 are infant. The observations on each subject contain 7 physical measurements and the age of the subject, as measured by the number of rings in its shell. We only use the subset of male abalones. The 532th subject in this subset is an outlier, and is deleted. Thus we have a sample of size 1527 with 7 predictors and 1 response. For objective evaluation of the estimators we further split the data into two subsets: the first 764 subjects are used as the training set to estimate the sufficient predictors and the rest 763 subjects are used as the testing set to plot the derived sufficient predictors versus the response.

We estimate both the central mean subspace (CMS) and the central variance subspace (CVS) of this data set. The CMS is estimated by RMAVE, SEE, and the method implicitly contained in Zhu, Dong, and Li (2012). The CVS is estimated by the methods proposed by Zhu and Zhu (2009) and Zhu, Dong, and Li (2012), and the SEE. The results are presented in Figure 5.1. The three upper panels are scatter plots of Y versus the sufficient predictor in the CMS as estimated by RMAVE, Zhu-Dong-Li, and SEE in that order. The three lower panels are the scatter plots of the absolute residuals $|Y - \hat{E}(Y|X)|$ versus the sufficient predictor in the CVS as estimated by Zhu-Zhu, Zhu-Dong-Li, and SEE.

To give an objective numerical comparison, we use a bootstrapped error measurement akin to that introduced by Ye and Weiss (2003), which is reasonable

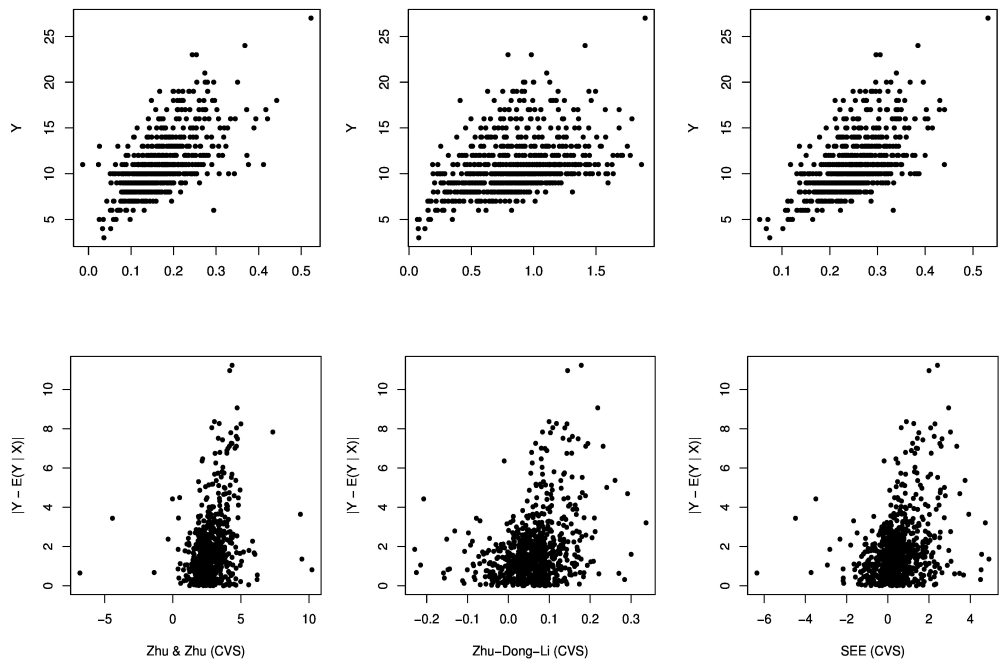

Figure 5.1. Comparison of SEE with other estimators of CMS and CVS for the abalone data. In the upper panels, the x-axes are the predictors obtained by RMAVE, Zhu-Dong-Li, and SEE estimates for the CMS; the y-axes are the abalones' ages. In the lower panels, the x-axes are the predictors derived from Zhu-Zhu, Zhu-Dong-Li, and SEE estimates of the CVS; the y-axes are the estimated absolute residuals.

because all estimators involved are consistent. Since the predictors in the abalone data set are highly correlated, two estimates of β that span substantially different linear spaces can correspond to nearly identical $\beta^{\mathsf{T}}X$. For this reason, rather than measuring the error in $\hat{\beta}$, as we did in the simulations, here we directly measure the error in $\hat{\beta}^{\mathsf{T}}X$. Specifically, we generate 500 bootstrap samples, and for each sample we compute the estimate $\tilde{\beta}$. We also compute the full-sample estimate $\hat{\beta}$. For each bootstrap sample, we evaluate the sample correlation between

$$\{\tilde{\beta}^{\mathsf{T}}X_1, \ldots, \tilde{\beta}^{\mathsf{T}}X_n\}, \quad \{\hat{\beta}^{\mathsf{T}}X_1, \ldots, \hat{\beta}^{\mathsf{T}}X_n\}.$$

We denote sample correlations for the 500 bootstrap samples as $\rho_1, \ldots, \rho_{500}$. We then compute $1 - \sum_{i=1}^{500} |\rho_i|/500$ and call it the bootstrap error of the estimator. The result is summarized in the following table.

We see that SEE is the top performer for estimating both CMS and CVS, fol-

Table 5.5. Bootstrap error of the estimators

Functionals	RMAVE	Zhu-Dong-Li	SEE
$E(Y\|X)$	0.145	0.009	0.003
$\mathrm{Var}(Y\|X)$	0.213	0.131	0.105

lowed by the estimator of Zhu, Dong and Li (2012), and then by RMAVE. We also observe that the estimation of central variance subspace is in general less accurate than that of central mean subspace, as has been observed in many other cases, for example in Zhu, Dong and Li (2012).

5.12 Discussion

In this chapter we introduce a general paradigm for sufficient dimension reduction with respect to a conditional statistical functional, along with semiparametrically efficient procedures to estimate the sufficient predictors of that functional. This method is particularly useful when we want to select sufficient predictors with some specific purposes in mind, such as estimating the conditional quantiles in a population. This work is a continuation, synthesis, and refinement of previous work on nonparametric mean regression, nonparametric quantile regression, and nonparametric estimation of heteroscedasticity, under the unifying framework of SDR. It provides us with tools to explore the detailed structures of the central subspace, making SDR more specific to our goals. Our work has also substantially broadened the scope of the semiparametric approach recently introduced to SDR by Ma and Zhu (2012, 2013a, and 2013b).

In a wide range of simulation studies the SEE is shown to outperform several previously proposed estimators for conditional mean, conditional quantile, and conditional variance. Moreover, the theoretical semiparametric lower bound is approximately achieved by the actual error based on simulation. Finally, the algorithm we developed for SEE has a special advantage over that proposed in Ma and Zhu (1012, 2013a): it does not rely on any specific parameterization of the central subspace, which means we do not need to subjectively assign any element of β to be nonzero from the outset.

5.13 Appendix

In this appendix we give the proofs of some of the technical results in this chapter. In this process we generated some new definitions, lemmas, and displayed formulas.

I. Proof of Theorem 1

Because θ will be fixed throughout this proof, we abbreviate symbols such as $\phi(x, y, \theta), \eta(x, y, \theta)$ as $\phi(x, y), \eta(x, y)$.

Part 1 (before "Moreover"). The defining relation of a member in $\mathscr{H}_\theta$ is

$$T(\eta(x,\cdot)) = E[T(\eta(X,\cdot)|\sigma_\theta(X)]_x \iff T_x(r(x,\cdot)) = E[T_X(r(X,\cdot))|\sigma_\theta(X)]_x.$$

Here, the conditional expectation $E(\cdots|\sigma_\theta(X))$ is under the true marginal density λ_0 and is therefore fixed as η varies over $\mathscr{H}_\theta$. Let $\alpha \mapsto r_\alpha$ be a curve in $\mathscr{H}_\theta$. Then

$$T_x(r_\alpha(x,\cdot)) = E[T_X(r_\alpha(X,\cdot))|\sigma_\theta(X)]_x$$

for every $\alpha \in [0, 1)$, which implies

$$\partial T_x(r_\alpha(x,\cdot))/\partial\alpha|_{\alpha=0} = E[\partial T_x(r_\alpha(X,\cdot))/\partial\alpha|_{\alpha=0}|\sigma_\theta(X)]_x. \tag{5.27}$$

That T_x is Fréchet differentiable at $r_0(x,\cdot)$ means

$$|T_x(r(x,\cdot)) - T_x(r_0(x,\cdot)) - \langle\tau(x,\cdot), r(x,\cdot)\rangle_{\eta_0(x,\cdot)}| = o(\|r(x,\cdot) - r_0(x,\cdot)\|_{\eta_0(x,\cdot)}).$$

By the chain rule of Fréchet differentiation we have

$$\partial T_x(r_\alpha(x,\cdot))/\partial\alpha|_{\alpha=0} = \langle\tau(x,\cdot), \dot{r}_0(x,\cdot)\rangle_{\eta_0(x,\cdot)} = E[\tau(X,Y)\dot{r}_0(X,Y)|X]_x \tag{5.28}$$

Substitute (5.28) into (5.27) to obtain

$$E[\tau(X,Y)\dot{r}_0(X,Y)|X] = E[\tau(X,Y)\dot{r}_0(X,Y)|\sigma_\theta(X)].$$

By the definition of conditional expectation, this is equivalent to

$$E\{h(X)[\tau(X,Y)\dot{r}_0(X,Y) - E(\tau(X,Y)\dot{r}_0(X,Y)|\sigma_\theta(X))]\} = 0,$$

for any $h \in L_2(\lambda_0)$, which can be rewritten as

$$E\{[h(X) - E(h(X)|\sigma_\theta(X))]\tau(X,Y)\dot{r}_0(X,Y)\} = 0. \tag{5.29}$$

As a curve in $\mathscr{H}_\theta$, r_α must also satisfy $E[(r_\alpha(X,\cdot)/2+1)^2|X] = 1$, which implies

$$E[(r_\alpha(X,\cdot)/2+1)^2 g(X)] = Eg(X)$$

for any $g \in L_2(\lambda_0)$. Differentiating both sides with respect to α at $\alpha = 0$ leads to

$$E[g(X)\dot{r}_0(X,Y)] = 0. \tag{5.30}$$

Combining (5.29) and (5.30) we see that $\mathscr{T}_\eta \perp \mathscr{U}$, which is equivalent to $\mathscr{T}_\eta \subseteq \mathscr{U}^\perp$.

Part 2 (after "Moreover"). Following Bickel et al (1993, section 3.2), our proof hinges on the construction of a curve whose tangent is in $\mathscr{U}^\perp$. To this end we need to construct some special functions.

Definition 3. *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a function that satisfies the following properties:*

1. *ψ is bounded and continuously differentiable;*
2. *the functions $\dot{\psi}$ and $\dot{\psi}/\psi$ are bounded;*
3. *$\psi(0) = \dot{\psi}(0) = 1$.*

Let u be a member of $L_2(\psi_0)$ that satisfies

$$E[u(X,Y)|X] = 0, \quad E[u(X,Y)\tau(X,Y)|X] = 1.$$

The functions ψ, u in the above definition do exist: for example, if

$$\psi(x) = 2(1+e^{-2x})^{-1},\ u(x,y) = \{\tau(x,y) - E[\tau(X,Y)|X]_x\}/\mathrm{Var}[\tau(X,Y)|X]_x,$$

then the conditions in Definition 3 are satisfied.

We need to show that, for any $t \in \mathscr{U}^\perp$, there is a curve $\alpha \mapsto r_\alpha$ such that

$$r_\alpha \in \mathscr{H}_\theta, \quad \partial r_\alpha(x,y)/\partial\alpha|_{\alpha=0} = t(x,y). \tag{5.31}$$

Since $t \perp \mathscr{U}$, we have $t \perp L_2(\lambda_0)$, which implies $E[t(X,Y)|X] = 0$. Let ψ and u be functions satisfying Definition A3. For each $(\alpha, \epsilon) \in \mathbb{R} \times \mathbb{R}$, let

$$\eta(x, y, \alpha, \epsilon) = \frac{\psi(\alpha t(x,y) + \epsilon u(x,y))\eta_0(x,y)}{c(x, \alpha, \epsilon)}, \quad r(x, \cdot, \alpha, \epsilon) = R_x(\eta(x, \cdot, \alpha, \epsilon)),$$

where

$$c(x, \alpha, \epsilon) = \textstyle\int_{\Omega_Y} \psi(\alpha t(x,y) + \epsilon u(x,y))\eta_0(x,y) d\mu_Y(y).$$

By construction, $\eta(x, \cdot, \alpha, \epsilon)$ is a probability density function.

By the assumption of continuous Fréchet differentiability and the assumptions in Definition A3 it can be shown that, in a neighborhood of $(\alpha, \epsilon) = (0, 0)$, the function $(\alpha, \epsilon) \mapsto T_x(r(x, \cdot, \alpha, \epsilon))$ is continuously differentiable. Moreover, it is easy to check that

$$\begin{aligned} \partial r(x, y, \alpha, 0)/\partial\alpha|_{\alpha=0} &= \partial \log \eta(x, y, \alpha, 0)/\partial\alpha|_{\alpha=0}, \\ \partial r(x, y, 0, \epsilon)/\partial\epsilon|_{\epsilon=0} &= \partial \log \eta(x, y, 0, \epsilon)/\partial\epsilon|_{\epsilon=0}. \end{aligned}$$

By construction we have

$$\begin{aligned} \partial \log \eta(x, y, \alpha, 0)/\partial\alpha|_{\alpha=0} &= [\dot{\psi}(0)/\psi(0)]t(x,y) - \partial \log c(x, \alpha, 0)/\partial\alpha|_{\alpha=0} \\ &= t(x,y), \end{aligned} \tag{5.32}$$

where the second equality holds because $[\dot{\psi}(0)/\psi(0)]t(x,y) = t(x,y)$ and

$$\partial c(x, \alpha, 0)/\partial\alpha|_{\alpha=0} = \textstyle\int_{\Omega_Y} \dot{\psi}(0)t(x,y)\eta_0(x,y)d\mu_Y(y) = E[t(X,Y)|X]_x = 0.$$

By the similar argument (using $E(u(X,Y)|X) = 0$) we deduce that

$$\partial \log \eta(x, y, 0, \epsilon)/\partial\epsilon|_{\epsilon=0} = u(x,y). \tag{5.33}$$

Hence, by the chain rule for Fréchet differentiation,

$$\begin{aligned} &\partial T_x(r(x, \cdot, \alpha, 0))/\partial\alpha|_{\alpha=0} = \langle \tau(x, \cdot), t(x, \cdot) \rangle_{\eta_0(x,\cdot)} \\ &\partial T_x(r(x, \cdot, 0, \alpha))/\partial\epsilon|_{\epsilon=0} = \langle \tau(x, \cdot), u(x, \cdot) \rangle_{\eta_0(x,\cdot)} = E[\tau(X,Y)u(X,Y)|X]_x = 1. \end{aligned}$$

Thus the differentiability of the function $(\alpha, \epsilon) \mapsto T_x(r(x, \alpha, \epsilon))$ implies

$$T_x(r(x, \cdot, \alpha, \epsilon)) = T_x(0) + \alpha\langle\tau(x, \cdot), t(x, \cdot)\rangle_{\eta_0(x,\cdot)} + \epsilon + \mathrm{rem}(x, \alpha, \epsilon), \quad (5.34)$$

where the remainder $\mathrm{rem}(x, \alpha, \epsilon)$ is of the order $o((\alpha^2 + \epsilon^2)^{1/2})$. It follows that

$$\begin{aligned} \partial\mathrm{rem}(x, \alpha, 0)/\partial\alpha|_{\alpha=0} &= \lim_{\alpha\to 0} \mathrm{rem}(x, \alpha, 0)/\alpha = 0, \\ \partial\mathrm{rem}(x, 0, \epsilon)/\partial\epsilon|_{\epsilon=0} &= \lim_{\epsilon\to 0} \mathrm{rem}(x, 0, \epsilon)/\epsilon = 0. \end{aligned} \quad (5.35)$$

The continuous differentiability of $(\alpha, \epsilon) \mapsto T_x(r(x, \cdot, \alpha, \epsilon))$ also implies the continuous differentiability of $\epsilon + \mathrm{rem}(x, \alpha, \epsilon)$ over the same neighbood. Hence

$$\partial[\epsilon + \mathrm{rem}(x, \alpha, \epsilon)]/\partial\epsilon = 1 + \partial\mathrm{rem}(x, \alpha, \epsilon)/\partial\epsilon$$

is positive in a neighborhood of $(0, 0)$. By the implicit function theorem, in a neighborhood of $\alpha = 0$, the equation

$$\epsilon + \mathrm{rem}(x, \alpha, \epsilon) = 0 \quad (5.36)$$

has a unique solution, which defines a continuously differentiable function $\alpha \mapsto \epsilon(x, \alpha)$. Consequently, if we let $r_\alpha(x, y) = r(x, y, \alpha, \epsilon(x, \alpha))$, then

$$T_x(r_\alpha(x, \cdot)) = T_x(0) + \alpha\langle\tau(x, \cdot), t(x, \cdot)\rangle_{\eta_0(x,\cdot)}. \quad (5.37)$$

Because the function $x \mapsto T_x(0)$ is measurable with respect to $\sigma_\theta(X)$,

$$E\{[h(X) - E(h(X)|\sigma_\theta(X))]T_X(0)\} = 0 \quad (5.38)$$

for any $h \in L_2(\lambda_0)$. Also, because $t \in \mathscr{U}^\perp$, we have

$$E\{[h(X) - E(h(X)|\sigma_\theta(X))]\tau(X, Y)t(X, Y)\} = 0. \quad (5.39)$$

Combining (5.37), (5.38), and (5.39), we find

$$E\{[h - E(h|\sigma_\theta(X))]T_X(r_\alpha(X, \cdot))\} = 0.$$

This implies, for any $h \in L_2(\lambda_0)$,

$$E\{h(X)[T_X(r_\alpha(X,\cdot)) - E(T_X(r_\alpha(X,\cdot))|\sigma_\theta(X))]\} = 0.$$

Hence $T_X(r_\alpha(X,\cdot)) = E(T_X(r_\alpha(X,\cdot))|\sigma_\theta(X))$, and the first relation in (5.31) holds.

Now differentiate both sides of the equation (5.36) with respect to α to obtain

$$\begin{aligned}&\partial\epsilon(x,\alpha)/\partial\alpha|_{\alpha=0} + \partial\mathrm{rem}(x,\alpha,0)/\partial\alpha|_{\alpha=0} \\ &\quad +[\partial\mathrm{rem}(x,0,\epsilon)/\partial\epsilon|_{\epsilon=0}]\,[\partial\epsilon(x,\alpha)/\partial\alpha|_{\alpha=0}] = 0.\end{aligned}$$

Thus, by (5.35), we have

$$\partial\epsilon(x,\alpha)/\partial\alpha|_{\alpha=0} = 0. \tag{5.40}$$

From (5.32), (5.33) and (5.40) we see that

$$\begin{aligned}\partial r_\alpha(x,y)/\partial\alpha|_{\alpha=0} &= \partial r(x,y,\alpha,0)/\partial\alpha|_{\alpha=0} + [\partial r(x,y,0,\epsilon)/\partial\epsilon|_{\epsilon=0}][\partial\epsilon(x,\alpha)/\partial\alpha|_{\alpha=0}] \\ &= t(x,y) + u(x,y)\times 0 = t(x,y),\end{aligned}$$

which is the second relation in (5.31).

II. Explicit formulas for the efficient scores in Section 10

For easy understanding we now use the β-parameterization rather than the θ-parameterization in sections 5, 6 and 7.

For central mean subspaces in Models I $\sim$ IV. By Theorem 3, for these models we have

$$\begin{aligned}\tau_c(x,y,\beta) &= y - E(Y|\beta^\tau X)_x, \\ q_1(x,\beta) &= \partial E(Y|\beta^\tau X)_x/\partial\mathrm{vec}\circ(\beta^\tau) = \partial E(Y|\beta^\tau X)_x/\partial(\beta^\tau x)\otimes x, \\ q_2(x,\beta) &= \mathrm{Var}(Y|X)_x,\end{aligned}$$

where $\otimes$ is the Kronecker product. Hence

$$q_3(x,\beta) = \frac{\partial E(Y|\beta^\tau X)_x}{\partial(\beta^\tau x)} \otimes \left\{x - \frac{E[X/\mathrm{Var}(Y|X)|\beta^\tau X]_x}{E[1/\mathrm{Var}(Y|X)|\beta^\tau X]_x}\right\}[y - E(Y|\beta^\tau X)_x].$$

The efficient score is then

$$S_{\text{eff}}(x,y,\beta) = \frac{1}{\text{Var}(Y|X)_x}\frac{\partial E(Y|\beta^\tau X)_x}{\partial(\beta^\tau x)} \otimes \left\{x - \frac{E[X/\text{Var}(Y|X)|\beta^\tau X]_x}{E[1/\text{Var}(Y|X)|\beta^\tau X]_x}\right\}[y - E(Y|\beta^\tau X)_x], \tag{5.41}$$

which is a ds-dimensional random vector. The efficient information is the $ds \times ds$ matrix

$$E[S_{\text{eff}}(X,Y,\beta)S_{\text{eff}}^\tau(X,Y,\beta)]. \tag{5.42}$$

For central variance subspaces in Models I $\sim$ III. Following the development in Section 6, Example 3, all we need to do is to replace Y in (5.41) $\tilde{Y} = [Y - E(Y|X)]^2$, which gives

$$\begin{aligned} S_{\text{eff}}(x,y,\beta) \;=\; & \frac{1}{\text{Var}[(Y-E(Y|X))^2|X]_x}\frac{\partial E[(Y-E(Y|X))^2|\beta^\tau X]_x}{\partial(\beta^\tau x)} \otimes \\ & \left\{x - \frac{E[X/\text{Var}((Y-E(Y|X))^2|X)|\beta^\tau X]_x}{E[1/\text{Var}((Y-E(Y|X))^2|X)|\beta^\tau X]_x}\right\}\cdot \\ & \{(Y-E(Y|X))^2 - E[(Y-E(Y|X))^2|\beta^\tau X]_x\}. \end{aligned}$$

The efficient information is given by (5.42).

For central quantile subspaces in Models V and VII. We first note that, if $\xi(X,\beta)$ is the conditional quantile that is measurable with respect to $\beta^\tau X$, then $\xi(Y,X)$ is the conditional quantile of Y given $\beta^\tau X$. That is, if we let $Q(U|V)$ denote the conditional quantile of a random variable U given a random vector V, then

$$\xi(X,\beta) = E[\xi(X,\beta)|\beta^\tau X] \;\Rightarrow \xi(X,\beta) = Q(Y|\beta^\tau X).$$

This is because $\xi(X,\beta)$ is defined to satisfy

$$P(Y \le \xi(X,\beta)|X) = p,$$

which implies

$$E[P(Y \le \xi(X,\beta))|X)|\beta^{\mathsf{T}}X] = P[Y \le \xi(X,\beta)|\beta^{\mathsf{T}}X] = p.$$

Since $\xi(X,\beta) = E[\xi(X,\beta)|\beta^{\mathsf{T}}X]$, we have

$$P[Y \le E(\xi(X,\beta)|\beta^{\mathsf{T}}X)|\beta^{\mathsf{T}}X] = p.$$

This means $E(\xi(X,\beta)|\beta^{\mathsf{T}}X)$ is the pth conditional quantile of Y given $\beta^{\mathsf{T}}X$. This justifies denoting $\xi(X,\beta)$ by $Q(Y|\beta^{\mathsf{T}}X)$.

Following the development in Example 4, we have

$$\begin{aligned}
\tau_c(x,y,\beta) &= [\text{sgn}(y - Q(Y|\beta^{\mathsf{T}}X)_x) - 1 + 2p] \,/\, [2\eta_0(x, Q(Y|\beta^{\mathsf{T}}X)_x, \beta)], \\
q_1(x,\beta) &= \partial Q(Y|\beta^{\mathsf{T}}X)_x/\partial\beta = [\partial Q(Y|\beta^{\mathsf{T}}X)_x \,/\, \partial\beta^{\mathsf{T}}x] \otimes x, \\
q_2(x,\beta) &= p(1-p) \,/\, [\eta_0^2(x, Q(Y|\beta^{\mathsf{T}}X)_x, \beta)].
\end{aligned}$$

Hence the efficient score is

$$\begin{aligned}
S_{\text{eff}}(x,y,\beta) \;=\; & \frac{\eta_0(x, Q(Y|\beta^{\mathsf{T}}X)_x, \beta)}{2p(1-p)} \frac{\partial Q(Y|\beta^{\mathsf{T}}X)_x}{\partial(\beta^{\mathsf{T}}x)} \;\otimes \\
& \left\{ x - \frac{E[X\eta_0^2(x, Q(Y|\beta^{\mathsf{T}}X)_x, \beta)|\beta^{\mathsf{T}}X]_x}{E[\eta_0^2(x, Q(Y|\beta^{\mathsf{T}}X)_x, \beta)|\beta^{\mathsf{T}}X]_x} \right\} \text{sgn}[y - Q(Y|\beta^{\mathsf{T}}X)_x - 1 + 2p].
\end{aligned}$$

The efficient information is given by (5.42).

III. Efficient score when the central subspace is unknown.

As in this chapter, suppose the central subspace is of dimension d and the T-central subspace is of dimension s. To avoid the trivial case we assume $s < d$; that is, the T-central subspace is a proper subspace of the central subspace. Recall that $\tilde{X} \in \mathbb{R}^p$ represents the original predictor, ζ a basis matrix of the central subspace for $Y|\tilde{X}$, and β a basis matrix of the T-central subspace of $Y|\zeta^{\mathsf{T}}X$. Note that ζ is a $p \times d$ matrix and β is a $d \times s$ matrix. Also recall that β has parameterization $\beta(\theta)$ where θ is a $(d-s)s$-dimensional free parameter. To keep notation simple, we also use ζ to denote an element in the $d \times (p-d)$ Grassmann manifold.

In the efficient score function developed in this chapter up to Section 8, ζ is treated a known constant rather than an unknown parameter. Here we derive the efficient score for β when ζ is unknown and treated as a (finite-dimensional)

nuisance parameter, along with the unknown (infinite-dimensional) nuisance functions η and λ. As in this chapter, we use $S_{\text{eff}}(\zeta^\intercal \tilde{x}, y, \theta)$ to denote the efficient score function with known ζ, and $S^*_{\text{eff}}(\tilde{x}, y, \zeta, \beta)$ to denote the efficient score function with ζ as the nuisance parameter. Besides, we use $s_\zeta(\tilde{x}, y, \zeta, \theta)$ to denote the score function with respect to ζ, which is

$$s_\zeta(x, y, \zeta, \theta) = \partial \log \eta(\zeta^\intercal \tilde{x}, y, \theta) \,/\, \partial \zeta.$$

Similarly, we use s_θ to denote the score function with respect to θ.

Let $\mathscr{T}_\zeta$ denote the finite-dimensional linear space spanned by components of s_ζ. Then S^*_{eff} is the projection of s_θ onto $\mathscr{T}_\phi^\perp \cap \mathscr{T}_\zeta^\perp$. Because $\mathscr{T}_\phi^\perp \cap \mathscr{T}_\zeta^\perp$ is a subspace of $\mathscr{T}_\phi^\perp$ and S_{eff} is the projection of s_θ onto $\mathscr{T}_\phi^\perp$, S^*_{eff} is the projection of S_{eff} on to $\mathscr{T}_\phi^\perp \cap \mathscr{T}_\zeta^\perp$. We have the following result which shows how to derive S^*_{eff}.

Lemma 10. *The efficient scores S^*_{eff} and S_{eff} are related by the following formula*

$$S^*_{\text{eff}} = S_{\text{eff}} - \Pi(S_{\text{eff}} \,|\, \mathcal{S}(\Pi(s_\zeta | \mathscr{T}_\phi^\perp))). \tag{5.43}$$

Thus, to find S^*_{eff}, we can first find $\Pi(s_\zeta | \mathscr{T}_\phi^\perp)$, and then $\Pi(S_{\text{eff}} \,|\, \mathcal{S}(\Pi(s_\zeta | \mathscr{T}_\phi^\perp))$, then finally subtract the result from S_{eff}.

PROOF OF LEMMA 10. Let $g = \Pi(S_{\text{eff}} \,|\, \mathcal{S}(\Pi(s_\zeta | \mathscr{T}_\phi^\perp)))$. We need to show that g is the projection of S_{eff} onto $\mathscr{T}_\phi^\perp \cap \mathscr{T}_\zeta^\perp$, or equivalently,

(a) $g \in \mathscr{T}_\phi^\perp \cap \mathscr{T}_\zeta^\perp$;

(b) $(S_{\text{eff}} - g) \perp \mathscr{T}_\phi^\perp \cap \mathscr{T}_\zeta^\perp$

To show (a), note that S_{eff} is the projection of s_θ onto $\mathscr{T}_\phi^\perp$, therefore $S_{\text{eff}} \in \mathscr{T}_\phi^\perp$. Similarly, we know

$$\Pi(S_{\text{eff}} \,|\, \mathcal{S}(\Pi(s_\zeta | \mathscr{T}_\phi^\perp))) \in \mathcal{S}(\Pi(s_\zeta | \mathscr{T}_\phi^\perp)) \subseteq \mathscr{T}_\phi^\perp$$

Hence $g \in \mathscr{T}_\phi^\perp$. On the other hand, by definition of projection, we know that $g \perp \mathcal{S}(\Pi(s_\zeta | \mathscr{T}_\phi^\perp))$. Besides, $g \in \mathscr{T}_\phi^\perp$ implies that $g \perp \mathcal{S}(\Pi(s_\zeta | \mathscr{T}_\phi))$. Because

$$s_\zeta = \Pi(s_\zeta | \mathscr{T}_\phi) + \Pi(s_\zeta | \mathscr{T}_\phi^\perp), \tag{5.44}$$

we have $g \perp \mathcal{S}(s_\zeta)$, which means that $g \in \mathscr{T}_\zeta^\perp$. Thus (a) does hold.

To show (b), again by definition of projection, $S_{\text{eff}} - g \in \mathcal{S}(\Pi(s_\zeta|\mathscr{T}_\phi^\perp))$. On the other hand, any $v \in \mathscr{T}_\phi^\perp \cap \mathscr{T}_\zeta^\perp$ satisfies $v \perp s_\zeta$ and $v \perp \mathscr{T}_\phi$. Because $\Pi(s_\zeta|\mathscr{T}_\phi) \in \mathscr{T}_\phi$, we have $v \perp \mathcal{S}(\Pi(s_\zeta|\mathscr{T}_\phi))$. By (5.44), $v \perp \Pi(s_\zeta|\mathscr{T}_\phi^\perp)$. Hence $v \perp (S_{\text{eff}} - g)$, which proves (b). □

Note that the form of S_{eff} given in Theorem 3 can be directly generalized to projections of any function $g \in L_2(\phi_0)$ onto $\mathscr{T}_\phi^\perp$. For example, when $g = s_\zeta$, let

$$\begin{aligned}
q_1^*(\zeta_0^\intercal \tilde{x}, \theta_0) &= E[s_\zeta(\tilde{X}, Y, \zeta_0, \theta_0)\tau_c(\zeta_0^\intercal \tilde{X}, Y, \theta_0)|\zeta_0^\intercal \tilde{X}]_{\tilde{x}} \\
q_2(\zeta_0^\intercal \tilde{x}, \theta_0) &= E[\tau_c^2(\zeta_0^\intercal \tilde{X}, Y, \theta_0)|\zeta_0^\intercal \tilde{X}]_{\tilde{x}} \\
q_3^*(\zeta_0^\intercal \tilde{x}, \theta_0) &= q_1^*(\zeta_0^\intercal \tilde{x}, \theta_0) - E_{\theta_0}[q_1^*(\zeta_0^\intercal \tilde{X}, \theta_0) q_2^{-1}(\zeta_0^\intercal \tilde{X}, \theta_0)|\sigma_{\theta_0}(\zeta_0^\intercal \tilde{X})]_{\tilde{x}} / \\
&\qquad E[q_2^{-1}(\zeta_0^\intercal \tilde{X}, \theta_0)|\sigma_{\theta_0}(\zeta_0^\intercal X)]_{\tilde{x}} \\
q_4^*(\zeta_0^\intercal \tilde{x}, \theta_0) &= q_2^{-1}(\zeta_0^\intercal \tilde{x}, \theta_0) q_3^*(\zeta_0^\intercal \tilde{x}, \theta_0).
\end{aligned}$$

Then, $\Pi(s_\zeta|\mathscr{T}_\phi^\perp) = q_4^*(\zeta_0^\intercal \tilde{x}, \theta_0)\tau_c(\zeta_0^\intercal \tilde{x}, y, \theta_0)$. Since the subspace $\mathcal{S}(\Pi(s_\zeta|\mathscr{T}_\phi^\perp))$ has finite dimension (no more than $s\,(d-s)$), the projection of S_{eff} on to it can be written down explicitly in terms of the Gram matrix. That is, if $A^\dagger$ denotes the Moore-Penrose inverse of a matrix A, then

$$\begin{aligned}
&\Pi(S_{\text{eff}}|\mathcal{S}(\Pi(s_\zeta|\mathscr{T}_\phi^\perp))) \\
&\quad = E\{S_{\text{eff}}\Pi(s_\zeta|\mathscr{T}_\phi^\perp)^\intercal\}\, E\{\Pi(s_\zeta|\mathscr{T}_\phi^\perp)\Pi(s_\zeta|\mathscr{T}_\phi^\perp)^\intercal\}^\dagger\, \Pi(s_\zeta|\mathscr{T}_\phi^\perp) \\
&\quad = q_2^{-1}(\zeta_0^\intercal \tilde{x}, \theta_0)\tau_c(\zeta_0^\intercal \tilde{x}, y, \theta_0) E\{q_3^*(\zeta_0^\intercal \tilde{X}, \theta_0) q_3^{*\intercal}(\zeta_0^\intercal \tilde{X}, \theta_0) q_2^{-1}(\zeta_0^\intercal \tilde{X}, \theta_0)\} \\
&\qquad E\{q_3^*(\zeta_0^\intercal \tilde{X}, \theta_0) q_3^{*\intercal}(\zeta_0^\intercal \tilde{X}, \theta_0) q_2^{-1}(\zeta_0^\intercal \tilde{X}, \theta_0)\}^\dagger\, q_3^*(\zeta_0^\intercal \tilde{x}, \theta_0).
\end{aligned} \tag{5.45}$$

Now $S_{\text{eff}}^*(\tilde{x}, y, \zeta_0, \theta_0)$ can be obtained by substituting (5.45) into Lemma 10.

IV. Asymptotic value of $E(\|\Pi_{\text{span}(\hat{\beta})} - \Pi_{\text{span}(\beta_0)}\|_2)$

In this section we derive the theoretical asymptotic lower bound (5.25), which is used to compute the ALB columns of Tables $5.2 \sim 5.4$ in Section 10. We only derive it for the case where β is a vector; the case where β is a matrix can be derived by obvious analogy. By Taylor's theorem,

$$\begin{aligned}
0 &= E_n[S_{\text{eff}}(X, Y, \hat{\theta})] \\
&= E_n[S_{\text{eff}}(X, Y, \theta_0)] + \partial E_n[S_{\text{eff}}(X, Y, \theta)]/\partial\theta|_{\theta=\theta_0}(\hat{\theta} - \theta_0) + O_P(n^{-1}) \\
&= E_n[S_{\text{eff}}(X, Y, \theta_0)] - J_{\text{eff}}(\theta_0)(\hat{\theta} - \theta_0) + O_P(n^{-1}).
\end{aligned}$$

Therefore,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n} J_{\text{eff}}^{-1}(\theta_0) E_n[S_{\text{eff}}(X, Y, \theta_0)] + o_P(1).$$

Let $\Pi_\beta = \beta(\beta^\intercal\beta)^{-1}\beta^\intercal$ be the projection matrix on to $\text{span}(\beta)$. Note that Π_β is in fact a function of θ. If $\text{span}(\beta)$ is has dimension 1, then the gradient of the function $\beta \mapsto \text{vec} \circ (\Pi_\beta)$ is $\Delta(\beta) \in \mathbb{R}^{p^2 \times p}$ in which the $((p-1)i + j, k)$th entry is

$$[\beta_i I(j = k) + \beta_j I(i = k)]/\|\beta\|_2^2 - 2\beta_i\beta_j\beta_k/\|\beta\|_2^4.$$

for each $i, j, k = 1, \ldots, p$. As described in Section 5.2, we can assume without loss of generality that $\beta = c + \Gamma\theta$ for some constant vector c and constant matrix Γ. Hence the gradient of $\text{vec} \circ (\Pi_\beta)$ with respect to θ is $\Delta(\beta)\Gamma$. Since

$$\sqrt{n}\, E_n S_{\text{eff}}(X, Y, \theta_0) \xrightarrow{\mathcal{D}} N(0, J_{\text{eff}}(\theta_0)),$$

by the Delta method we have

$$\sqrt{n}(\text{vec} \circ (\Pi_{\hat{\beta}}) - \text{vec} \circ (\Pi_{\beta_0})) \xrightarrow{\mathcal{D}} N(0, \Delta(\beta_0)\Gamma J_{\text{eff}}^{-1}(\theta_0)\Gamma^\intercal\Delta^\intercal(\beta_0)).$$

Denote $J_{\text{eff}}^\dagger(\beta_0)$ as the Moore-Penrose generalized inverse of $J_{\text{eff}}(\beta_0)$. Then we have $J_{\text{eff}}^\dagger(\beta_0) = \Gamma J_{\text{eff}}^{-1}(\theta_0)\Gamma^\intercal$. Therefore

$$\sqrt{n}(\text{vec} \circ (\Pi_{\hat{\beta}}) - \text{vec} \circ (\Pi_{\beta_0})) \xrightarrow{\mathcal{D}} N(0, \Delta(\beta_0) J_{\text{eff}}^\dagger(\beta_0)\Delta^\intercal(\beta_0)).$$

Note that in simulation we can estimate $J_{\text{eff}}^\dagger(\beta_0)$ by the Moore-Penrose inverse of $E_n[S_{\text{eff}}(X, Y, \beta_0)S_{\text{eff}}(X, Y, \beta_0)^\intercal]$.

Let $\omega_1, \ldots, \omega_{p^2}$ be the eigenvalues of $\Delta(\beta_0)J_{\text{eff}}^\dagger(\beta_0)\Delta^\intercal(\beta_0)$, and $Z_1, \ldots, Z_{p^2}$ be i.i.d random variables with χ_1^2 distribution. Then we have

$$n\|\Pi_{\hat{\beta}} - \Pi_{\beta_0}\|_2^2 \to_d \sum_{i=1}^{p^2} \omega_i Z_i.$$

Let Z be a random variable that has the distribution on the right-hand side above. Then $\sqrt{n}\, E\|\Pi_{\hat{\beta}} - \Pi_{\beta_0}\|_2 \to E\sqrt{Z}$. Therefore we can estimate the asymptotic semi-parametric lower bound by first estimating $\delta_1, \ldots, \delta_{p^2}$ via $J_{\text{eff}}^\dagger(\beta_0)$, and then

simulating the distribution of Z.

Bibliography

[1] Bai, Z.D., Krishnaiah, P.R. and Liang W.Q. (1986) On asymptotic joint distribution of the eigenvalues of noncentral MANOVA matrix for nonnormal populations. *the Indian Journal of Statistics* **48**, 153 - 162.

[2] Bickel, P. and Freedman, D. (1981) Some asymptotic theory for the bootstrap *Annals of Statistics* **9**, 1196-1217.

[3] Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press. Baltimore and London.

[4] Billingsley, P. (1968) *Convergence of probability measures*. Wiley Series in Probability and Mathematical Statistics, John Wiley, New York.

[5] Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society. Series B*. **63** 393 – 410.

[6] Bura, E. and Yang, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *Journal of Multivariate Analysis* **102**, 130-142

[7] Chikuse, Y. (2003). *Statistics on Special Manifolds*. Lecture Notes in Statistics, **74**, Springer, New York.

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

[9] Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proceedings of the Section on Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, 18–25.

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

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

[12] Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics*. **32**, 1062 – 1092.

[13] Cook, R. D. (2007). Fisher Lecture: Dimension Reduction for Regression (with discussion). *Statistical Science*. **22**, 1-152.

[14] Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*. **104**, 197-208.

[15] Cook, R. D., Forzani, L. and Rothman, A. (2012). Estimating sufficient reductions of the predictors in abundant high dimensional regressions. *Annals of Statistics*. **40**, 353-384.

[16] Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Annals of Statistics*, **30**, 455–474.

[17] Cook, R. D. and Li, B. (2004). Determining the dimension of Iterative Hessian Transformation. *Annals of Statistics* **32**, 2501 – 2531.

[18] Cook, R. D. and Li, L. (2009). Dimension reduction in regressions with exponential family predictors. *Journal of Computational and Graphical Statistics*, **18**, 774-791.

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

[20] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 316 – 342.

[21] Cook, R. D. and Yin, X. (2001). Dimension-reduction and visualization in discriminant analysis (Invited with discussion; subsequent award article). *Australia & New Zealand Journal of Statistics*, **43**, 147-200.

[22] Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika*, **97**, 279–294.

[23] Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *Annals of Statistics.* **19**, 505 – 530.

[24] Edelman, A, Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications*. **20**, 303–353.

[25] Evett, I. W. and Spiehler, E. J. (1988). *Rule Induction in Forensic Science.*. Halsted Press.

[26] Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Springer. New York.

[27] Forina M., Leardi, R., Arimanino, C. and Lanteri, S. (1988). *PARVUS - An extendable package of programs for data exploration, classification and correlation*. Springer.

[28] Fukumizu, K., Bach, F. and Jordan, M. (2009). Kernel dimension reduction in regression. *Annals of Statistics*. **37**, 1591-2082.

[29] Gammaitoni, L., Hanggi, P., Jung, P., and Marchesoni, F. (1998). Stochastic Resonance. *Reviews of Modern Physics.* **70**, 223-288.

[30] Gill, R. (1989). Non- and semi-parametric maximum likelihood estimators and the Von Mises method (Part 1). *Scandinavian Journal of Statistics* **16**, 97-128.

[31] Godambe, V. P. (1960). An optimum property of maximum likelihood estimation. *Annals of Mathmematical Statistics*. **31**, 1208–1211.

[32] Hall, P. and Li, K. -C. (1993). On almost Linearity of Low Dimensional Projections from High Dimensional Data. *Annals of Statistics*. **21**, 867-889.

[33] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *Annals of Statistics*, **21(1)**,157-178.

[34] Horton, P. and Nakai, K. (1996). A Probablistic Classification System for Predicting the Cellular Localization Sites of Proteins. *Intelligent Systems in Molecular Biology* **4**, 109-115.

[35] Hristache, M., Juditsky, A., Polzehl, J., Spokoiny, V. (2001) Structure Adaptive Approach for Dimension Reduction. *Annals of Statistics*, **29**, 1537 – 1566.

[36] Kong, E. and Xia, Y. (2012). A single-index quantile regression model and its estimation. *Econometric Theory*, **28**, 730–768.

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

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

[39] Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Annals of Statistics*, **37**, 1272–1298.

[40] Li, B. and McCullagh, P. (1994). Potential Functions and Conservative Estimating Functions *Annals of Statistics*. **22**, 340–356.

[41] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **35**, 2143 – 2172.

[42] Li, B., Wen, S., and Zhu, L. (2008) On a Projective Resampling Method for Dimension Reduction With Multivariate Responses *Journal of American Statistical Association*. **103**, 1177 – 1186.

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

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

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

[46] Li, K. -C. and Duan, N. (1989). Regression analysis under link violation. *Annals of Statistics.* **17**, 1009-1052.

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

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

[49] Liu, R. Y., Singh, K. and Lo, S. -H. (1989) On a representation related to the bootstrap. *Sankhyā: The Indian Journal of Statistics* **51**, 168-177.

[50] Luo, W. and Li, B. (2014). Order determination for dimension reduction using an alternating pattern of spectral variability. *Biometrika*. Under revision.

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

[52] Luo, W., Li, B. and Yin, X. (2014b). Supplement to "On efficient dimension reduction with respect to a statistical functional of interest." DOI:10.1214/13-AOS1195SUPP.

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

[54] Ma, Y. and Zhu, L. (2013a). Efficient estimation in sufficient dimension reduction. *Annals of Statistics*. In press.

[55] Ma, Y. and Zhu, L. (2013b). Efficiency loss caused by linearity condition in dimension reduction. *Biometrika*. In press.

[56] Mallows, C. L. (1972) A note on asymptotic joint normality. *Annals of Mathematical Statistics* **43**, 508-515

[57] McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall/CRC.

[58] Miller, A. J. (2002). *Subset Selection in Regression*. Chapman & Hall. London.

[59] Parr, W. (1985) The bootstrap: some large sample theory and connections with robustness. *Statistics & Probability Letters* **3**, 97-100.

[60] Schott, J. R. (1994). Determining the demensionality in sliced inverse regression. *Journal of the American Statistical Association* **89**, 141-148.

[61] Street, W. N., Wolberg, W. H. and Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis. *International Symposium on Electronic Imaging: Science and Technology* **1905**, 861-870.

[62] Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *Annals of Applied Statistics*, **3**, 1236–1265.

[63] Tanaka, H. (1973) An inequality for a functional of probability distributions and its application to Kac's one-dimensional model of a Maxwellian gas. *Z. Wahrscheinlichkeitstheorie verw.* **27**, 47-52.

[64] van der Vaart (2000). *Asymptotic Statistics*. Cambridge Press.

[65] Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, **103**, 811-821, 2008.

[66] Wu, Y., Boos, D. D., and Stefanski, L. A. (2007). Controlling Variable Selection by the Addition of Pseudo Variables. *Journal of the American Statistical Association*, **477**, 235-243.

[67] Wu, Q., Mukherjee, S. and Liang, F. (2009). Localized sliced inverse regression. *NIPS*. **21**, 1785-1792.

[68] Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Annals of Statistics*, **35**, 2654-2690.

[69] Xia, Y., Tong, H. and Li, W. K. (2002). Single-index volatility models and estimation. *Statistica Sinica*. **12**, 785C799.

[70] Xia, Y., Tong, H., Li, W. K., and Zhu, L-X. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 363–3410.

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

[72] Yin, X. and Cook, R. D. (2002). Dimension reduction for conditional kth moment in regression. *Journal of the Royal Statistical Society: Series B*, **64**, 159–175.

[73] Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika*. **92**, 371-384.

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

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

[76] Zhao, L. C., Krishnaiah, P. R. and Bai, Z. D. (1986) On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis* **20**, 1-25.

[77] Zhu, L., Dong, Y., and Li, R. (2012). Semiparametric estimation of conditional heteroscedasticity via single-index modeling. *Statistica Sinica*. In press.

[78] Zhu, L.-X. and Fang, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Annals of Statistics* **3**, 1053 – 1068.

[79] Zhu, L., Miao, B. and Peng, H. (2006) On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630-642.

[80] Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, **101**, 1638–1651.

[81] Zhu, L. and Zhu, L.-X. (2009). Dimension reduction for conditional variance in regression. *Statistica Sinica*, **19**, 869–883.

[82] Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, **36**, 1108–1126.

Wei Luo

Department of Statistics
the Pennsylvania State University
325 Thomas Building
University Park, PA 16802

Phone: (814) 206-4985
Email: wzl118@psu.edu

Education

Ph.D candidate, Department of Statistics, the Pennsylvania State University. August 2009 - August 2014 (expected).

- ◇ Adviser: Dr. Bing Li, Professor of Statistics.
- ◇ GPA: 3.98/4.00.

B. S. in Mathematics and Applied Mathematics, Department of Mathematics, Zhejiang University, China. September 2004 - July 2008.

Honors / Awards

- ◇ **Pass with distinction** in Ph.D candidacy exam, 2011, Department of Statistics, PSU.
- ◇ Student Travel Grant for Joint Statistical Meetings 2012, Department of Statistics, PSU
- ◇ Student Travel Grant for Joint Statistical Meetings 2013, Department of Statistics, PSU.

Research Interest

dimension reduction; high-dimensional data analysis; semi-parametric methods; extension of principal component analysis.

Publications / Manuscripts

1. **Luo, W.** and Altman, N.S. (2013) A characterization of conjugate priors in linear exponential families with application to inverse regression. *Statistical and Probability Letters.* **83**, 650 - 654.
2. **Luo, W.**, Li, B. and Yin, X. (2014) On efficient dimension reduction with respect to a statistical functional of interest. *Annals of Statistics.* **42**, 382 - 412.
3. **Luo, W.** and Li, B. (2014) Order determination for dimension reduction using an alternating pattern of spectral variability. *Biometrika.* Under revision.
4. **Luo, W.** and Li, B. Order-determination for dimension reduction by augmentation predictors. *Manuscript.*
5. **Luo, W.** and Li, B. A fast algorithm for semiparametrically efficient estimator in sufficient dimension reduction. *Manuscript.*
6. Altman, N.S., Raskutti, G. and **Luo, W.** Exploratory and inferential methods for massive data - the many roles of Principal Component Analysis. *Manuscript.*