Sufficient Dimension Reduction

Jingyue Lu
Primary Supervisor: Professor Alan Welsh

July 2017

A thesis submitted for the degree of Master of Philosophy
of The Australian National University

Declaration of Authorship

I, Jingyue Lu, declare that this thesis titled, ‘Sufficient Dimension Reduction’ and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: ____

Date: ____

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Alan Welsh, for his valuable comments, remarks and engagement through the learning process of this master thesis. Alan has not only introduced me to the topic of the thesis, but also helped me gain a deeper understanding of statistics through many insightful conversations.

On a more personal level, I am thankful to my friend Xiaoyang Xu, for all the times he has lifted my spirit with his healthy cooking and shared laughter. Last but not least, I want to thank my parents for encouraging me in all my pursuits and inspiring me to follow my dreams.

This research is supported by Australian Government Research Training Program Fee-Offset Scholarship.

Abstract

In regression analysis, it is difficult to uncover the dependence relationship between a response variable and a covariate vector when the dimension of the covariate vector is high. To reduce the dimension of the covariate vector, one approach is sufficient dimension reduction. Sufficient dimension reduction is based on the assumption that the response variable relates to only a few linear combinations of the covariate vector. Thus, by replacing the covariate vector with these linear combinations, sufficient dimension reduction achieves dimension reduction. The goal of sufficient dimension reduction is to estimate the space spanned by these linear combinations of the covariate vector. We denote this space by S.

In this thesis, we give an introductory review on three important sufficient dimension reduction methods. They are Sliced Inverse Regression (SIR), Sliced Average Variance Estimate (SAVE) and Principle Hessian Directions (pHd). Li proposed SIR in 1991. SIR is a method that exploits the simplicity of the inverse regression. Given the univariate response variable and the high dimensional covariate, it is much easier to regress the covariate against the response variable than the other way around. Motivated by a theorem that connects forward regression and inverse regression, SIR estimates S using inverse regression lines. Since SIR uses first moments only, it fails when there exists symmetry dependence between the response variable and the covariate. To make up for this defect, Cook proposed SAVE in a comment on SIR in 1991. SAVE follows the general lines of SIR but uses second moments as well as first moments to estimate S. pHd is also a second moment method. Li developed pHd in 1992 based on the observation that the eigenvectors for the Hessian matrices of the regression function are closely related to the basis vectors of S. Therefore pHd provides an estimate of S by using these eigenvectors.

To compare these methods, a simulation study is presented at the end. From the simulation results, SIR is the most efficient method and SAVE is the most time consuming method. Since SIR fails when symmetry dependence exists, we recommend pHd when symmetry dependence presents and SIR in other cases.

Contents

Chapter 1

Introduction

With technological advances, datasets have grown in both size and complexity. One consequence of increasing amounts of data is that we often need to relate a response variable to a potentially large number of possible covariates. The high dimension of the covariate space makes it difficult to uncover this relationship. To reduce the dimension of the covariate space, two major approaches are developed based on different assumptions.

The first approach is variable selection. Variable selection is used when researchers believe that among all available predictors, only a few have explanatory effect. Thus, variable selection reduces the number of covariates by identifying and removing the covariates that have non explanatory effect. The second approach is sufficient dimension reduction. Sufficient dimension reduction, on the other hand, assumes that each covariate has explanatory effect, but the explantory effect is only represented through a few linear combinations of covariates. Therefore, sufficient dimension reduction aims to find these linear combinations. By replacing the collection of covariates with these linear combinations, sufficient dimension reduction achieve dimension reduction of the covariate space.

In this thesis, we focus on the second approach: sufficient dimension reduction.

1.1 Problem set up

Throughout the thesis, we denote the response variable as $y \in \mathbb{R}$ and the covariate vector as $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$.

Given the assumption of sufficient dimension reduction, the main problem of sufficient dimension reduction can be described by the model

$$y = f(\beta_1^T x, \beta_2^T x, \ldots, \beta_k^T x, \epsilon), \tag{1.1}$$

where β's are unknown column vectors of the matrix $\Phi := (\beta_1, \beta_2, \ldots, \beta_k)$, ϵ is independent of x, and f is an arbitrary unknown function on $\mathbb{R}^{k+1}$. If we can find Φ, we can replace p dimensional covariate x with $\beta_1^T x, \beta_2^T x, \ldots, \beta_k^T x$. Since k is typically much smaller than p, we hence achieve dimension reduction.

However, Φ is not identifiable. Let $S(A)$ be the space spanned by columns of an arbitrary matrix A. We observe that if (1.1) holds, then it also holds when we replace Φ with any matrix A such that $S(A) = S(\Phi)$. Therefore, it is appropriate to identify $S(\Phi)$ instead. We call a subspace $S(\Phi)$ satisfying (1.1) a *dimension reduction subspace* (DRS) (Li, 1991). Because $S(I_p)$ is by definition a DRS, DRS always exists and is not always unique.

To achieve maximum dimension reduction, we are interested in finding a minimum DRS. A minimum DRS S_{min} is a DRS such that $\dim(S_{min}) \leq \dim(S_{drs})$ for all DRSs S_{drs}. As we will see in Chapter 3, a minimum dimension reduction subspace may not be unique, leading to complications at later stages. In order to deal with this issue, we adopt Cook's idea (Cook, 2009) and introduce the concept of central dimension reduction subspaces (or central subspaces), denoted as $S_{y|x}$. A central subspace, when exists, is the unique minimum dimension reduction subspace. Since central subspaces exist under various reasonable conditions (Cook, 1994a, 1996), we restrict ourselves to the class of regressions for which the central subspace exist to ensure the uniqueness of the minimum dimension reduction subspace. Thus, we conclude the goal of sufficient dimension reduction is to find the central subspace of a problem of interest. More details are provided in Chapter 3.

1.2 Project outline

The purpose of this thesis is to provide readers with an introductory review on three sufficient dimension reduction (SDR) methods, which are Sliced Inverse Regression (SIR) (Li, 1991), Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991), and Principal Hessian Directions (pHd) (Li, 1992; Cook, 1998). In particular, we want to see how we can use these methods to at least partially recover the central subspace $S_{y|x}$ to achieve dimension reduction.

The rest of the thesis is organized in the following way. Chapter 2 is a preparation chapter. It gives a short introduction to elliptically contoured distributions and their properties. Since elliptically contoured distributions are closely related to the prerequisites of many SDR methods, studying them should help us gain a better understanding of SDR methods later.

Chapter 3-6 are about SDR methods. Chapter 3 sets up a theoretical framework for our studies of SDR methods. It studies central subspaces in detail by addressing the key questions: What are central subspaces? Why do we need central subspaces? And when do central subspaces exist? Discussions of the SDR methods SIR, SAVE and pHd are contained in Chapter 4 and Chapter 5. For each method, we not only examine the theoretical foundations, but also provide a step by step algorithm for estimating the central subspace $S_{y|x}$. Since each method has its advantages and disadvantages, a simulation study for testing and comparing the SDR methods SIR, SAVE and pHd is presented in the final chapter.

Chapter 2

Elliptically contoured distributions

Before we delve into specific sufficient dimension reduction (SDR) methods, we first introduce elliptically contoured distributions, which, as we will show in later chapters, are closely related to the key prerequisites required for most SDR methods to work. Elliptically contoured distributions are a natural generalization of Gaussian distributions. When the covariate has an elliptically contoured distribution, many SDR methods are able to exploit the nice properties of elliptically contoured distributions inherited from Gaussian distributions to attain neat and compact results. In this chapter, we examine the basic but essential properties of elliptically contoured distributions.

2.1 Definition and Characterisation

Despite being a generalization of Gaussian distributions, elliptically contoured distributions are generally treated as an extension of spherical distributions. In this section, we adopt this way of classifying them and start by introducing spherical distributions following the ideas of Kelker (1970) and Frahm (2004).

Definition 2.1 (Spherical distribution)**.** Let X be a p-dimensional random vector. X has a p-dimensional spherical distribution if and only if, for all $\mathbb{R}^{p\times p}$ orthonormal matrices Γ, X and ΓX have the same distribution such that $X =_d \Gamma X$.

Spherical distributions are also referred to as radial distributions. To better understand this definition, we first note that when a random vector X satisfies $X =_d \Gamma X$ for any p

by p orthonormal matrix Γ, X is rotationally symmetric. As a result, the above definition can be equally stated as follows.

Let X be a p-dimensional random vector. X has a p-dimensional spherical distribution if and only if it is rotationally symmetric.

Recall that if we let $U^{(p)}$ be a p-dimensional random vector that is uniformly distributed on the unit sphere

$$S^{p-1} := \{x \in \mathbb{R}^p : \|x\|_2 = 1\},$$

and assume $\mathcal{R}$ is a nonnegative random variable independent of $U^{(p)}$, then every p-dimensional random vector Y with the form of $Y := \mathcal{R}U$ is rotationally symmetric. Since spherical distributions and rotationally symmetric distributions are identical, Y is spherically distributed. Hence, we have found an explicit form that ensures a random vector follows a spherical distribution. A question arises naturally: can any spherically distributed random vector X be written in the form of $\mathcal{R}U$? If this is the case, the analysis of spherical distributions can be conducted in a straightforward manner, as we can work with U and R directly instead.

In order to answer this question, we consider a spherically distributed p-dimensional random vector X. Because X is, by definition, rotationally symmetric, for any $t \in \mathbb{R}^p$, the equality

$$\cos(\measuredangle(t, X)) =_d \cos(\measuredangle(v, U^{(p)})) =_d v^T U^{(p)} \tag{2.1}$$

holds for every $v \in S^{p-1}$ and random vector $U^{(p)}$ uniformly distributed on S^{p-1} (Frahm, 2004). Here, $\measuredangle(t, X)$ measures the angle between p-dimensional vector t and the random vector X and we have used the fact that $t^T X = \|t\|_2 \cdot \|X\|_2 \cdot \cos(\measuredangle(t, X))$. As a result of this equality, the characteristic function of $\cos(\measuredangle(t, X))$ satisfies

$$\begin{aligned} t \longmapsto \varphi_{\cos(\measuredangle(t,X))}(s) &= \varphi_{v^T U^{(p)}}(s) \\ &:= \mathrm{E}\{\exp(isv^T U^{(p)})\} = \mathrm{E}\{\exp(i(sv)^T U^{(p)})\} \\ &= \varphi_{U^{(p)}}(sv) \end{aligned} \tag{2.2}$$

where $v \in S^{p-1}$ is arbitrary and $\varphi_{U^{(p)}}$ is the characteristic function of $U^{(p)}$. This relationship between the characteristics functions of $\cos(\measuredangle(t, X))$ and $U^{(p)}$ will lead to our desired result. To see this, we first write the characteristic function of X in terms of $\varphi_{\cos(\measuredangle(t,X))}$

as follows

$$t \longmapsto \varphi_X(t) = \mathrm{E}\{\exp(it^TX)\} = \mathrm{E}\{\exp(i \cdot \|X\|_2 \cdot \|t\|_2 \cdot \cos(\measuredangle(t,X)))\}.$$

Then applying the law of total expectations to derive that

$$\begin{aligned} t \longmapsto \varphi_X(t) &= \mathrm{E}_X[\mathrm{E}\{\exp(i \cdot \|X\|_2 \cdot \|t\|_2 \cdot \cos(\measuredangle(t,X))) \mid \|X\|_2 = r\}] \\ &= \int_0^\infty \mathrm{E}\{\exp(i \cdot r\|t\|_2 \cos(\measuredangle(t,X)))\} dF_{\|X\|_2}(r) \\ &= \int_0^\infty \varphi_{\cos(\measuredangle(t,X))}(r\|t\|_2) dF_{\|X\|_2}(r), \end{aligned}$$

where $F_{\|X\|_2}$ is the cumulative distribution function (c.d.f) of $\|X\|_2$. Then by the relationship (2.2), we have

$$\begin{aligned} t \longmapsto \varphi_X(t) &= \int_0^\infty \varphi_{\cos(\measuredangle(t,X))}(r\|t\|_2) dF_{\|X\|_2}(r) \\ &= \int_0^\infty \varphi_{U^{(p)}}(r\|t\|_2 \cdot \frac{t}{\|t\|_2}) dF_{\|X\|_2}(r). \end{aligned}$$

Here, we have replaced the v in (2.2) with $t/\|t\|_2 \in S^{p-1}$. Finally, we obtain

$$\begin{aligned} t \longmapsto \varphi_X(t) &= \int_0^\infty \varphi_{U^{(p)}}(r\|t\|_2 \cdot \frac{t}{\|t\|_2}) dF_{\|X\|_2}(r) \\ &= \int_0^\infty \varphi_{U^{(p)}}(rt) dF_{\|X\|_2}(r) \\ &= \int_0^\infty \varphi_{rU^{(p)}}(t) dF_{\|X\|_2}(r) \end{aligned}$$

for any $r \geq 0$. We note that the last line of above equation can be viewed as the characteristic function of the random vector $\mathcal{R}U^{(p)}$, where $\mathcal{R}$ is a nonnegative random variable independent of $U^{(p)}$ and having the same distribution as $\|X\|_2$. We thus have successfully shown that any spherical distributed p-dimensional random vector X has the representation $X =_d \mathcal{R}U^{(p)}$. Once p is given, $U^{(p)}$ is fully determined and the spherical distribution of X is completely decided by the non-negative random variable $\mathcal{R}$. Therefore, $\mathcal{R}$ is often called "generating random variable" or "generating variate" of X (Frahm, 2004).

Remark 2.2. From the definition of spherical distributions, we see that a spherical distribution is invariant under rotation. This implies that spherical distributions are distributions that are centered about zero. Thus, when the expectation of a spherical distribution exists, the expectation has be to 0. We can prove this statement by using either the definition or the stochastic representation of spherical distributions.

Assume a p-dimensional random vector X is spherically distributed and its mean exists. We first show $\mathrm{E}(X) = 0$ by definition. Let $\Gamma_1 \neq \Gamma_2$ be orthonormal matrices. Since $X =_d \Gamma_1 X =_d \Gamma_2 X$, $\mathrm{E}(X) = \mathrm{E}(\Gamma_1 X) = \mathrm{E}(\Gamma_2 X)$. That is $\mathrm{E}((\Gamma_1 - \Gamma_2)X) = (\Gamma_1 - \Gamma_2)\mathrm{E}(X) = 0$. Because we know that $(\Gamma_1 - \Gamma_2) \neq 0$, we conclude that $\mathrm{E}(X) = 0$. On the other hand, we have shown that X has the representation $X =_d \mathcal{R}U^{(p)}$. Since $\mathcal{R}$ is independent of $U^{(p)}$ and $\mathrm{E}(U^{(p)}) = 0$, we also derive that $\mathrm{E}(X) = \mathrm{E}(\mathcal{R}) \cdot \mathrm{E}(U^{(p)}) = 0$.

Remark 2.3. Given the fact that every spherical distribution has the representation $X =_d \mathcal{R}U^{(p)}$, we can easily deduce the generating variable for standard normal distributions. Assume $X \sim N_p(0, I_p)$ and has the representation $X =_d \mathcal{R}U^{(p)}$ as defined above. Then we have

$$\chi_p^2 =_d X^T X = \mathcal{R}^2 U^{(p)T} U^{(p)} =_{a.s.} \mathcal{R}^2.$$

It follows that the generating variable of a standard normal distribution is $\sqrt{\chi_p^2}$, the square root of a random variable with a chi-squared distribution with p degrees of freedom.

In addition to the generating random variable $\mathcal{R}$, we can also find the characteristic generator function of a spherically distributed random vector X by closely examining the characteristic function φ_X and exploring the $\mathcal{R}U^{(p)}$ representation. The key observation to make is that, for the characteristic function $\varphi_{U^{(p)}}$ of $U^{(p)}$, we can always find a function $\phi_{U^{(p)}}$ such that $\varphi_{U^{(p)}}(sv) = \phi_{U^{(p)}}(s^2)$ for every $s \in \mathbb{R}$. As mentioned in (2.2), we note that, given that the point v is arbitrary, $\varphi_{U^{(p)}}(sv)$ only depends on s. In addition, since $\varphi_{U^{(p)}}((-s)v) = \varphi_{U^{(p)}}(s(-v))$, $\varphi_{U^{(p)}}(sv)$ is independent of the sign of s and hence can be treated as a function of s^2. We thus obtain

$$\begin{aligned} t \longmapsto \varphi_X(t) &= \int_0^\infty \varphi_{rU^{(p)}}(t) dF_{\mathcal{R}}(r) \\ &= \int_0^\infty \varphi_{U^{(p)}}(rt) dF_{\mathcal{R}}(r) \\ &= \int_0^\infty \varphi_{U^{(p)}}(r\|t\|_2 \cdot \frac{t}{\|t\|_2}) dF_{\mathcal{R}}(r) \\ &= \int_0^\infty \phi_{U^{(p)}}(r^2\|t\|_2^2) dF_{\mathcal{R}}(r). \end{aligned} \tag{2.3}$$

Consequently, φ_X can be equally represented through

$$s \longmapsto \phi_X(s) := \int_0^\infty \phi_{U^{(p)}}(r^2 s) dF_{\mathcal{R}}(r) \quad s \geq 0 \tag{2.4}$$

with

$$t \longmapsto \varphi_X(t) = \phi_X(\|t\|_2^2) = \phi_X(t^T t). \tag{2.5}$$

Moreover, we observe that if a p-dimensional random vector X has characteristic function $\varphi(t)$ satisfying $\varphi(t) = \phi(t^T t)$ for some function ϕ, then X is spherically distributed by definition, as the characteristic function implies that $X =_d \Gamma X$ for all orthonormal matrices $\Gamma \in \mathbb{R}^{p \times p}$.

Combining previous results, we conclude a random vector X belongs to the class of spherical distributions if and only if the equality (2.5) holds. As a result, ϕ_X is generally referred to as "characteristic generator" of X (Schmidt, 2002). We point out that the characteristic generator captures all the information contained in $\mathcal{R}$.

Remark 2.4. It is easy to deduce that the characteristic generator of a random vector X with standard normal distribution is $\phi_X(s) = \exp(-s/2)$ given that $\varphi_X(t) = \exp(-t^T t/2)$ and $\phi_X(t^T t) = \varphi_X(t)$.

We now extend our discussions to elliptically contoured distributions. As we mentioned at the beginning, elliptical contoured distributions are a generalization of spherical distributions. To be more specific, we will see shortly that every affine transformation of a spherically distributed random vector follows an elliptically contoured distribution and the converse is also true. Before we give proofs for these statements, we first give a formal definition of elliptically contoured distributions.

Definition 2.5 (Elliptically contoured distributions)**.** Let X be a p-dimensional random vector. X is said to be "elliptically distributed" or just "elliptical" if and only if there exits a constant vector $\mu \in \mathbb{R}^p$, a symmetric positive semidefinite matrix $\Sigma \in \mathbb{R}^{p \times p}$, and a function $\phi : \mathbb{R}^+ \to \mathbb{R}$ such that the characteristic function $\varphi_{X-\mu}(t)$ of $X - \mu$ satisfies $\varphi_{X-\mu}(t) = \phi(t^T \Sigma t)$. We write $X \sim EC_p(\mu, \Sigma, \phi)$, where EC is short for elliptically contoured.

Thus, to show that every affinely transformed spherical random vector is elliptically distributed, it is sufficient to find the characteristic function of the transformed random vector and check the existence of the function ϕ, satisfying $\varphi_{X-\mu}(t) = \phi(t^T \Sigma t)$. Based upon this idea, we have the following proposition.

Proposition 2.6. *Let X be a k-dimensional spherically distributed random vector with characteristic generator ϕ_X. Also assume $\Lambda \in \mathbb{R}^{p \times k}$ is an arbitrary matrix and $\mu \in \mathbb{R}^p$ is an arbitrary vector. Then $Y := \mu + \Lambda X$ has the characteristic function*

$$t \longmapsto \varphi_Y(t) = \exp(it^T \mu) \cdot \phi_X(t^T \Sigma t), \qquad t \in \mathbb{R}^p,$$

where $\Sigma := \Lambda\Lambda^T$. Consequently, Y is elliptically distributed.

Proof. We prove this proposition by directly computing the characteristic function of Y. We have

$$\begin{aligned} t \longmapsto \varphi_Y(t) &= \mathrm{E}(\exp(it^T(\mu + \Lambda X))) = \exp(it^T\mu) \cdot \varphi_X(\Lambda^T t) \\ &= \exp(it^T\mu) \cdot \phi_X((\Lambda^T t)^T(\Lambda^T t)) = \exp(it^T\mu) \cdot \phi_X(t^T \Sigma t). \end{aligned}$$

It follows

$$t \longmapsto \varphi_{Y-\mu}(t) = \varphi_X(\Lambda^T t) = \phi_X(t^T \Sigma t),$$

so Y has an elliptically contoured distribution by definition. □

Remark 2.7. In fact, this proposition partly motivates the definition of elliptically contoured distribution above and, to some extent, serves as a basis to define elliptically contoured distributions with a focus on the characteristic generators. Further, we emphasize that Y need not necessarily have the same dimension as X.

For the other direction, to show that every elliptically contoured distribution is an affinely transformed spherical distribution, we use the stochastic representation theorem. We recall that every spherically distributed random vector X has the representation $X =_d \mathcal{R}U^{(k)}$. The stochastic representation theorem proves the statement by showing that, under certain conditions, every $Y \sim EC_p(\mu, \Sigma, \phi)$ has the stochastic representation $Y =_d \mu + \mathcal{R}\Lambda U^{(k)}$, which is just the affine transformation of the spherically distributed random vector $\mathcal{R}U^{(k)}$.

Theorem 2.8 (Stochastic representation theorem)**.** *Let Y be p-dimensional random vector. Then $Y \sim EC_p(\mu, \Sigma, \phi)$ with $rank(\Sigma) = k$ if and only if*

$$Y =_d \mu + \mathcal{R}\Lambda U^{(k)}$$

where $\mathcal{R}$ is a nonnegative random variable, $U^{(k)}$ is k-dimensional random vector uniformly distributed on S^{k-1} that is independent of $\mathcal{R}$, $\mu \in \mathbb{R}^p$ and $\Lambda \in \mathbb{R}^{p \times k}$ with $rank(\Lambda) = k$.

Proof. We have proved the "if" direction in the proposition above. To show the "only if" direction, we note that every rank k symmetric positive semidefinite matrix Σ can be decomposed as $\Sigma = \Lambda\Lambda^T$ where $\Lambda \in \mathbb{R}^{p \times k}$. Then, define the random vector

$$X := \Lambda^{\dagger}(Y - \mu),$$

where $\Lambda^\dagger := (\Lambda^T\Lambda)^{-1}\Lambda^T$ is the Moore-Penrose pseudoinverse of Λ. Calculating the characteristic function of X, we obtain

$$\begin{aligned} t \longmapsto \varphi_X(t) &= \varphi_{Y-\mu}((\Lambda^\dagger)^T t) = \phi(t^T\Lambda^\dagger\Sigma(\Lambda^\dagger)^T t) \\ &= \phi(t^T(\Lambda^T\Lambda)^{-1}\Lambda^T(\Lambda\Lambda^T)\Lambda(\Lambda^T\Lambda)^{-1}t) = \phi(t^Tt), \qquad t \in \mathbb{R}^k. \end{aligned} \tag{2.6}$$

This implies that X is spherically distributed with the characteristic generator $\phi(t^Tt)$ and can be represented as $X =_d \mathcal{R}U^{(k)}$. Hence $Y = \mu + \Lambda X =_d \mu + \mathcal{R}\Lambda U^{(k)} \sim EC_p(\mu, \Sigma, \phi)$. □

We make some important comments about the stochastic representations of elliptically contoured distributions.

- Firstly, although each elliptically contoured distributed random vector can be formulated in stochastic representation, it should be emphasized that this representation is not unique. To be more specific, a stronger statement has been proved by Cambanis et al. (1981). It states that, given X is nondegenerate, if $X \sim EC_p(\mu, \Sigma, \phi)$ and $X \sim EC_p(\mu_0, \Sigma_0, \phi_0)$, then there exists a constant $c > 0$ such that $\Sigma_0 = c\Sigma$ and $\phi_0(\cdot) = \phi(c^{-1}\cdot)$ while $\mu = \mu_0$. It is possible for Σ and ϕ to be different from Σ_0 and ϕ_0 but the differences are up to a constant.

- Secondly, we note that an elliptically distributed random vector $X \sim EC_p(0, I_p, \phi)$ with $\mu = 0$ and $\Sigma = I_p$ is spherically distributed as $X = 0 + \mathcal{R}I_pU^{(p)} = \mathcal{R}U^{(p)}$. Using the same line of reasoning, we also find that affine transformations of elliptically distributed random vectors are also elliptically distributed. Consider $Y \sim EC_p(\mu, \Sigma, \phi)$ with stochastic representation $Y =_d \mu + \mathcal{R}\Lambda U^{(k)}$ where $\Lambda \in \mathbb{R}^{p\times k}$ and $\Lambda\Lambda^T = \Sigma$. Further, let $\alpha \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times p}$. Assume the random vector W is transformed from Y by

  $$W = \alpha + AY.$$

  Then we obtain

  $$W =_d \alpha + A(\mu + \mathcal{R}\Lambda U^{(k)}) = (\alpha + A\mu) + \mathcal{R}A\Lambda U^{(k)},$$

  which implies $W \sim EC_m(\alpha + A\mu, A\Sigma A^T, \phi)$. That is, W is elliptically distributed with $\varphi_{W-(\alpha+A\mu)}(t) = \phi_Y(t^TA\Sigma A^Tt)$. In conclusion, the class of elliptical contoured distributions is closed under affine transformations.

- Finally, the stochastic representation of an elliptically contoured distribution is generally preferred to its characteristic representation. Not only does the stochastic representation give a straightforward geometric interpretation of an elliptically distributed random vector X (μ determines the location of X, $\mathcal{R}$ specifies the shape, especially the tailedness of the distribution while Λ and $U^{(k)}$ together produces density surface), but also the explicit representations facilitate the simulation process of X (Frahm, 2004).

Remark 2.9. Multivariate normal distributions are a special case of elliptically contoured distributions. To see this, let $X \sim N_p(\mu, \Sigma)$ be a random vector with multivariate normal distribution, where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ is positive definite with the decomposition $\Sigma = \Lambda\Lambda^T$ with $\Lambda \in \mathbb{R}^{p\times k}$. Then from remark 2.3, we can derive that

$$X =_d \mu + \sqrt{\chi^2_k}\Lambda U^{(k)}$$

and hence X is elliptically distributed. In addition, from Remark 2.4, we have the characteristic function $\varphi_{X-\mu}$ of $X - \mu$ satisfies $t \longmapsto \varphi_{X-\mu}(t) = \exp(t^T \Sigma t)$.

So far, we have introduced elliptically contoured distributions as an extension of spherical distributions. We have also shown that elliptically contoured distributions have stochastic representations, that is they can be represented as an affine transformation of a spherical distribution $\mathcal{R}U^{(k)}$. With this explicit expression for elliptically contoured random vectors, we can easily develop the basic properties of this class of distributions, covered in the following section.

2.2 Basic properties

In this section, we study the density functions, marginal distribution and conditional distributions of elliptically contoured distributions. The main focus will be on the conditional distributions as their properties are the key for most sufficient dimension reduction methods to work.

2.2.1 Density functions

Adopting the same analysing procedure as that used above, to find the density function of the elliptically contoured distributions, we first derive the density function of the spherical distributions.

Theorem 2.10 (Spherical distributions). *Let X be a p-dimensional random vector with stochastic representation $X =_d \mathcal{R}U^{(p)}$ where the c.d.f of $\mathcal{R}$ is absolutely continuous. Then the c.d.f of X is given by*

$$x \longmapsto f_X(x) = \frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}} \cdot \|x\|_2^{-(p-1)} \cdot f_{\mathcal{R}}(\|x\|_2), \qquad x \in \mathbb{R}^p \setminus \{0\},$$

where $f_{\mathcal{R}}$ is the p.d.f of $\mathcal{R}$.

Proof. To start, we recall that the density function of a p-dimensional random vector uniformly distributed on the unit hypersphere S^{p-1} is $\frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}}$ and that $U^{(p)}$ and $\mathcal{R}$ are independent. Thus, given that the c.d.f of $\mathcal{R}$ is absolutely continuous, we have that the density function of the pair (r, u) is

$$(r, u) \longmapsto f_{(\mathcal{R},U^{(p)})}(r, u) = \frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}} \cdot f_{\mathcal{R}}(r), \qquad r > 0, u \in S^{p-1}.$$

In order to find the density function of $X =_d \mathcal{R}U^{(p)}$, we define the transformation $h : (0, \infty) \times S^{p-1} \to \mathbb{R}^p \setminus \{0\}$ by $h(r, u) = ru$. Clearly, h is injective. We thus have the p.d.f of X as

$$x \longmapsto f_X(x) = f_{\mathcal{R},U^{(p)}}(h^{-1}(x)) \cdot |J_h|^{-1}, \qquad x \in \mathbb{R}^p \setminus \{0\}, \tag{2.7}$$

where J_h is the Jacobian determinant of $\partial ru/\partial(r, u)^T$. Since for any $u \in S^{p-1}$, $\|u\| = 1$, it follows that $\|ru\| = r$. As a result, $h^{-1}(x) = (\|x\|_2, x/\|x\|_2)$. When the Jacobian J_h is considered, direct calculation gives us

$$|J_h| = det(\begin{pmatrix} 1 & 0 \\ 0 & rI_{p-1} \end{pmatrix}) = r^{p-1} = \|x\|_2^{p-1}.$$

Here we have used the fact that $\partial ru/\partial r$ has unit length and is orthogonal to $\partial ru/\partial u^T$ on $S_r^{p-1} := \{t \in \mathbb{R}^p : \|t\| = r\}$, the hypersphere with radius r.

Substituting the above results into (2.7), we have derived the p.d.f. of X as:

$$\begin{aligned} x \longmapsto f_X(x) &= f_{\mathcal{R},U^{(p)}}(\|x\|_2, x/\|x\|_2) \cdot \|x\|_2^{p-1} \\ &= \frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}} \cdot \|x\|_2^{-(p-1)} \cdot f_{\mathcal{R}}(\|x\|_2), \qquad x \in \mathbb{R}^p \setminus \{0\}. \end{aligned} \tag{2.8}$$

□

Remark 2.11. We can apply this theorem to derive the density function of a standard normally distributed random vector $X \sim N_p(0, I_d)$. Given that the p.d.f of χ_p^2 is

$$t \longmapsto f(t) = \frac{t^{\frac{p}{2}-1} \cdot \exp(-\frac{t}{2})}{2^{p/2} \cdot \Gamma(\frac{p}{2})}, \qquad t \geq 0,$$

and $\mathcal{R} = \sqrt{\chi_p^2}$, we get

$$t \longmapsto f_{\mathcal{R}}(t) = 2t \cdot f(t^2).$$

Then it follows from the above theorem, the density function of X is

$$\begin{aligned} x \longmapsto f_X(x) &= \frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}} \cdot \|x\|_2^{-(p-1)} \cdot 2\|x\|_2 \cdot f(x^T x) \\ &= \frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}} \cdot \|x\|_2^{-(p-1)} \cdot 2\|x\|_2 \cdot \frac{(x^T x)^{\frac{p}{2}-1} \cdot \exp(-\frac{x^T x}{2})}{2^{p/2} \cdot \Gamma(\frac{p}{2})} \\ &= \frac{1}{(2\pi)^{p/2}} \cdot \exp(-\frac{x^T x}{2}). \end{aligned}$$

The result for spherical distributions can be easily extended to elliptically contoured distributions with a positive definite Σ.

Theorem 2.12 (Elliptically contoured distributions)**.** *Let $X \sim EC(\mu, \Sigma, \phi)$ where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ is symmetric positive definite. Equivalently, we can write X in its stochastic representation $X =_d \mu + \mathcal{R}\Lambda U^{(p)}$ where $\Lambda\Lambda^T = \Sigma$ and $\Lambda \in \mathbb{R}^{p \times p}$. Assume the c.d.f of $\mathcal{R}$ is absolutely continuous, then the p.d.f of X is given by*

$$x \longmapsto f_X(x) = |det(\Sigma)|^{-1/2} \cdot g_{\mathcal{R}}((x-\mu)^T \Sigma^{-1}(x-\mu)), \quad x - \mu \in S_\Lambda \setminus \{0\}$$

where

$$t \longmapsto g_{\mathcal{R}}(t) := \frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}} \cdot \sqrt{t}^{-(p-1)} \cdot f_{\mathcal{R}}(\sqrt{t}), \quad t > 0,$$

S_Λ is the linear subspace of $\mathbb{R}^p$ spanned by Λ and $f_{\mathcal{R}}$ is the p.d.f of $\mathcal{R}$.

Proof. From the theorem above, the density function of $Y \coloneqq \mathcal{R}U^{(p)}$ is

$$y \longmapsto f_Y(y) = \frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}} \cdot \|y\|_2^{-(p-1)} \cdot f_{\mathcal{R}}(\|y\|_2).$$

To derive the density function of X, we introduce the transformation $h : \mathbb{R}^p \backslash \{0\} \to S_\Lambda \backslash \{0\}$ with $h(y) = \Lambda y$. We note that $h(y) = x - \mu$ and h is injective as Λ is invertible , so we have

$$x \longmapsto f_X(x) = f_Y(h^{-1}(x-\mu)) \cdot |J_h|^{-1}.$$

Since $h^{-1}(x-\mu) = \Lambda^{-1}(x-\mu)$ and $\partial(\mu + \Lambda y)/\partial y^T = \Lambda$ implies that $|J_h| = |det(\Lambda)|$, we hence conclude the p.d.f of X is: for $x - \mu \in S_\Lambda \setminus \{0\}$

$$\begin{aligned} x \longmapsto f_X(x) &= f_Y(\Lambda^{-1}(x-\mu)) \cdot |det(\Lambda)|^{-1} \\ &= |det(\Lambda)|^{-1} \cdot \frac{\Gamma(\frac{p}{2})}{2\pi^{p/2}} \cdot \|\Lambda^{-1}(x-\mu)\|_2^{-(p-1)} \cdot f_{\mathcal{R}}(\|\Lambda^{-1}(x-\mu)\|_2). \end{aligned} \tag{2.9}$$

Finally as

$$|det(\Lambda)| = |det(\Sigma)|^{1/2}$$

$$(\Lambda^{-1})^T\Lambda^{-1} = (\Lambda^T)^{-1}\Lambda^{-1} = (\Lambda\Lambda^T)^{-1} = \Sigma^{-1},$$

we can replace $|det(\Lambda)|^{-1}$ with $|det(\Sigma)|^{-1/2}$ and $\|\Lambda^{-1}(x-\mu)\|_2$ with $\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$ respectively. The desired result is thus obtained. □

From the theorem above, we see that when elliptically contoured distributions have a positive definite Σ, their density functions can be expressed in terms of the density function of the generating random variable $\mathcal{R}$.

2.2.2 Moments

We can also use stochastic representations to find the mean and covariance of a p-dimensional elliptical random vector X. Assume the X has the stochastic representation $X =_d \mu + \mathcal{R}\Lambda U^{(k)}$ with $\mathcal{R}$, $U^{(k)}$ defined as above and $\Lambda \in \mathbb{R}^{p\times k}, \mu \in \mathbb{R}^k$. Then the mean of X is

$$\mathrm{E}(X) = \mathrm{E}(\mu + \mathcal{R}\Lambda U^{(k)}) = \mu + \Lambda\mathrm{E}(\mathcal{R}) \cdot \mathrm{E}(U^{(k)}),$$

where the last equality used the fact that $\mathcal{R}$ is independent of $U^{(k)}$. Provided $\mathrm{E}(\mathcal{R})$ is finite, applying the fact that $E(U^{(k)}) = 0$, we obtain that $\mathrm{E}(X) = \mu$.

When the covariance of X is considered, we compute

$$\mathrm{Cov}(X) = \mathrm{E}((\mathcal{R}\Lambda U^{(k)})(\mathcal{R}\Lambda U^{(k)})^T) = \mathrm{E}(\mathcal{R}^2) \cdot \Lambda \mathrm{E}(U^{(k)}U^{(k)T})\Lambda^T. \tag{2.10}$$

Here we require that $E(\mathcal{R}^2) < \infty$. To derive the explicit formula for $\mathrm{Cov}(X)$, we thus need to obtain the explicit expressions for $\mathrm{E}(\mathcal{R}^2)$ and $\mathrm{E}(U^{(k)}U^{(k)T})$ respectively.

We start with the simpler calculation: $\mathrm{E}(U^{(k)}U^{(k)T})$. Since the distribution of $U^{(k)}$ is known, $\mathrm{E}(U^{(k)}U^{(k)T})$ is a fixed number. Thus, by letting $\mathcal{R}$ be $\sqrt{\chi^2_k}$ distributed, we can derive its value by using familiar facts of normal and chi-square distributions. We recall that from previous remarks, we have concluded that $\sqrt{\chi^2_k}U^{(k)} \sim N_k(0, I_k)$. It follows that

$$I_k = \mathrm{E}((\sqrt{\chi^2_k}U^{(k)})(\sqrt{\chi^2_k}U^{(k)})^T) = \mathrm{E}(\chi^2_k) \cdot \mathrm{E}(U^{(k)}U^{(k)T}) = k \cdot \mathrm{E}(U^{(k)}U^{(k)T}),$$

which implies $E(U^{(k)}U^{(k)T}) = I_k/k$.

To obtain the value for $E(\mathcal{R}^2)$, we need the following theorem proved by Cambanis et al. (1981).

Theorem 2.13. *Let $X \sim EC_p(\mu, \Sigma, \phi)$. Further, assume X is nondegenerate and has stochastic representation $X =_d \mu + \mathcal{R}\Lambda U^{(k)}$ where $\Sigma = \Lambda\Lambda^T$. Then $\mathrm{E}(\mathcal{R}^2)$ exists if and only if the right hand-side derivative of $\phi(u)$ at $u = 0$, denoted as $\phi'(0)$, exists and is finite. Moreover,*

$$\mathrm{E}(\mathcal{R}^2) = -2k\phi'(0).$$

Proof. We observe that if we let $U^{(k)}_1$ be the first component of $U^{(k)}$, then it is direct to see that $\mathrm{E}(\mathcal{R}^2) < \infty$ if and only if $\mathrm{E}((RU^{(k)}_1)^2) < \infty$ as $\mathrm{E}(\mathcal{R}^2) = k \cdot \mathrm{E}((RU^{(k)}_1)^2)$. Thus to prove the existence part of the theorem, we only need to show that $\phi'(0)$ exists if and only if $E((RU^{(k)}_1)^2) < \infty$.

Since, from previous discussions, we know that $RU^{(k)}$ has the characteristic function $t \longmapsto \phi(t^Tt), t \in \mathbb{R}^k$, it follows that $RU^{(k)}_1$ has the characteristic function

$$\phi_{RU^{(k)}_1}(u) = \phi(u^2), \quad u \in \mathbb{R}. \tag{2.11}$$

We first assume $\mathrm{E}((RU_1^{(k)})^2)$ exists. It follows that $\phi_{RU_1^{(k)}}$ is twice differentiable. Using this result, we can derive the following key equality

$$\begin{aligned}\mathrm{E}((RU_1^{(k)})^2) &= -\phi''_{RU_1^{(k)}}(0) = -\lim_{h\to 0}\frac{\phi(h^2)-2\phi(0)+\phi((-h)^2)}{h^2}\\ &= -2\lim_{h\to 0}\frac{\phi(h^2)-\phi(0)}{h^2} = -2\phi'(0) < \infty.\end{aligned}$$

As a result, the existence of $\mathrm{E}((RU_1^{(k)})^2)$ guarantees the existence and finiteness of $\phi'(0)$ and $\mathrm{E}((RU_1^{(k)})^2) = -2\phi'(0)$.

For the other direction, let $\phi'(0)$ exist and be finite. We want to show that

$$\mathrm{E}((RU_1^{(k)})^2) = \int_{-\infty}^{\infty} x^2 dH(x) < \infty, \tag{2.12}$$

where H is the distribution function of $RU_1^{(k)}$. To show this inequality, we first note that

$$x^2 = 2\lim_{h\to 0}\frac{1-\cos hx}{h^2}. \tag{2.13}$$

In addition, due to the relationship (2.11), for $h \neq 0$, we have

$$\begin{aligned}\frac{1-\phi(h^2)}{h^2} &= \frac{-\phi_{RU_1^{(k)}}(h)+2\phi_{RU_1^{(k)}}(0)-\phi_{RU_1^{(k)}}(-h)}{2h^2}\\ &= \int_{-\infty}^{\infty}\frac{-(\cos hx + i\sin hx)+2-(\cos hx - i\sin hx)}{2h^2}dH(x)\\ &= \int_{-\infty}^{\infty}\frac{1-\cos hx}{h^2}dH(x).\end{aligned} \tag{2.14}$$

Substituting the results of (2.13) and (2.14) into (2.12) and applying Fatou's lemma, we obtain

$$\begin{aligned}\mathrm{E}((RU_1^{(k)})^2) &= 2\int_{-\infty}^{\infty}\lim_{h\to 0}\frac{1-\cos hx}{h^2}dH(x)\\ &\leq 2\lim_{h\to 0}\int_{-\infty}^{\infty}\frac{1-\cos hx}{h^2}dH(x)\\ &= 2\lim_{h\to 0}\frac{1-\phi(h^2)}{h^2} = -2\phi'(0) < \infty.\end{aligned} \tag{2.15}$$

As a result, the existence of $\phi'(0)$ implies the existence of $\mathrm{E}((RU_1^{(k)})^2)$. □

With the discussions above, we have successfully evaluated both $\mathrm{E}(U^{(k)}U^{(k)T})$ and $\mathrm{E}(\mathcal{R}^2)$ under the assumption that the covariance of X exists. As a result, the covariance of X is

$$\begin{aligned}\mathrm{Cov}(X) &= \mathrm{E}(\mathcal{R}^2)\cdot\Lambda\mathrm{E}(U^{(k)}U^{(k)T})\Lambda^T \\ &= -2k\phi'(0)\cdot\Lambda(I_k/k)\Lambda^T = -2\phi'(0)\Sigma.\end{aligned} \tag{2.16}$$

At last, we note that we can always find a representation such that $Cov(X) = \Sigma$ by multiplying $\mathcal{R}$ with $(-2\phi'(0))^{-1/2}$.

2.2.3 Marginal distributions

To study the marginal distributions of elliptically contoured distributions, we adopt Hult and Lindskog's idea (Hult and Lindskog, 2002) and introduce matrices $P_k \in \{0,1\}^{k\times p}(k \leq p)$, such that P_k only contains 0 or 1 entries and $P_kP_k^T = I_k$. The P_k matrices are also referred to as "permutation and deletion" by Frahm (2004), as P_k affects a p-dimensional random vector X by permuting k components of X and deleting the remaining $p-k$ components of X. In terms of stochastic representations, we observe that: given $X \sim EC_p(\mu,\Sigma,\phi)$ with $X =_d \mu + \mathcal{R}\Lambda U^{(k)}$ and $Y := P_kX$,

$$Y =_d P_k(\mu + \mathcal{R}\Lambda U^{(k)}) = P_k\mu + \mathcal{R}P_k\Lambda U^{(k)}. \tag{2.17}$$

This implies that, Y, as an affine transformation of X, is also elliptically distributed with $Y \sim EC_p(P_k\mu, P_k\Sigma P_k^T, \phi)$.

With this observation, a direct application of "permutation and deletion" matrices can give us the marginal distribution of a random variable. For example, consider again the p-dimensional random vector $X \sim EC_p(\mu,\Sigma,\phi)$ and partition the arrays as

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad \text{and} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{with dimensions} \quad \begin{pmatrix} k\times 1 \\ (p-k)\times 1 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \text{with dimensions} \begin{pmatrix} k\times k & k\times(p-k) \\ (p-k)\times k & (p-k)\times(p-k) \end{pmatrix}.$$

Then by setting

$$P_1 = \begin{pmatrix} I_k & 0_{k\times(p-k)} \end{pmatrix} \qquad P_2 = \begin{pmatrix} 0_{(p-k)\times k} & I_{p-k} \end{pmatrix},$$

we have $P_1X = X_1$ and $P_2X = X_2$. As a result, the distribution of X_1 is $EC_p(\mu_1, \Sigma_{11}, \phi)$ and the distribution of X_2 is $EC_p(\mu_2, \Sigma_{22}, \phi)$.

Moreover, from the analyses above, another important observation to make is that for elliptically contoured distributions, the characteristic function of the parent distribution always has the same functional form as the characteristic function of the marginal distribution. For example, if a marginal density of an elliptical random vector X is a normal density, then X is normally distributed. In fact, to show that an elliptically distributed random vector $X \sim EC_p(\mu, \Sigma, \phi)$ is normally distributed, Kelker (1970) showed that it is sufficient to check that the matrix Σ is diagonal and the components of X are independent.

Lemma 2.14. *Let $X \sim EC_p(\mu, \Sigma, \phi)$. If Σ is a diagonal matrix and the components of X are independent, then X is normally distributed.*

Proof. Without loss of generality, we assume $\mu = 0$. Since Σ is diagonal and the components of X are independent, we have

$$\phi(\sigma_{11}t_1^2 + \sigma_{22}t_2^2 + \cdots + \sigma_{pp}t_p^2) = \prod_{i=1}^{p} \phi(\sigma_{ii}t_i^2).$$

The above equation is also known as Hamel's equation and has the solution $\phi(x) = e^{cx}$ for some constant c, $c \leq 0$, as ϕ is a characteristic function. Since the characteristic function of X takes the form $\phi(t^T\Sigma t) = \exp(ct^T\Sigma t)$, X is normally distributed. □

For more results on the independence and correlation of components of random vectors, we refer interested readers to Johnson (1987).

2.2.4 Conditional distribution

Lastly and most importantly, we introduce some key results on conditional distributions, which are essential to the development of some sufficient dimension reduction methods.

In order to study marginal distributions of elliptically contoured distributions, we adopt the methodology developed by Cambanis et al. (1981): we start by analysing the marginal distributions of a random vector uniformly distributed on a hypersphere and then use the stochastic representations of random vectors to find the explicit form of the conditional distributions.

Theorem 2.15 (Cambanis et al. (1981)). *For any positive integer k, let $U^{(k)}$ be uniformly distributed on the unit hypersphere $S^{k-1} := \{x \in \mathbb{R}^k : \|x\|_2 = 1\}$. Then, given $U^{(k)}$ and any partition of $U^{(k)}$ with $m := dim(U_1^{(k)})$, we have $(U^{(k)})^T = \{(U_1^{(k)})^T, (U_2^{(k)})^T\} =_d \{\beta(U^{(m)})^T, (1-\beta^2)^{1/2}(U^{(k-m)})^T\}$, where $\beta, U^{(m)}, U^{(k-m)}$ are independent and*

$$\beta^2 \sim Beta(\frac{m}{2}, \frac{k-m}{2}).$$

Proof. To start, assume $X = (X_1^T, X_2^T)^T \sim N_k(0, I_k)$ with $\dim(X_1) = m$. Clearly, X_1 and X_2 are independent. We also observe that since the mapping $x \longmapsto (\|x\|_2, x/\|x\|_2)$ is Borel measurable on $\mathbb{R}^k - \{0\}$, we obtain that, given that $X =_d \mathcal{R}U^{(k)}$,

$$(\|X\|_2, X/\|X\|_2) =_d (R, U^{(k)}). \tag{2.18}$$

Because X_1 and X_2 are independent and the equality (2.18), it follows that $\frac{X_1}{\|X_1\|_2}, \frac{X_2}{\|X_2\|_2}$, $\|X_1\|_2$, $\|X_2\|_2$ are jointly independent and

$$\frac{X_1}{\|X_1\|_2} =_d U^{(m)}, \qquad \frac{X_2}{\|X_2\|_2} =_d U^{(k-m)}, \tag{2.19}$$

$$\|X_1\|_2 =_d \sqrt{\chi_m^2}, \qquad \|X_2\|_2 =_d \sqrt{\chi_{k-m}^2}. \tag{2.20}$$

Let

$$\{(U_1^{(k)})^T, (U_2^{(k)})^T\}^T = U^{(k)} =_d \frac{X}{\|X\|_2} = (\frac{X_1^T}{\|X\|_2}, \frac{X_2^T}{\|X\|_2})^T. \tag{2.21}$$

To derive the distribution of $\frac{X_1}{\|X\|_2}$ and $\frac{X_2}{\|X\|_2}$, we define

$$\beta := \frac{\|X_1\|_2}{\|X\|_2} = \frac{\|X_1\|_2}{(\|X_1\|_2^2 + \|X_2\|_2^2)^{1/2}}. \tag{2.22}$$

Given (2.19), (2.20) and the independence between $\|X_1\|_2, \|X_2\|_2$, β^2 has the Beta$(\frac{m}{2}, \frac{k-m}{2})$ distribution. In addition, since β can be seen as a function of $\|x_1\|_2$ and $\|x_2\|_2$ and we know that $\frac{X_1}{\|X_1\|_2}, \frac{X_2}{\|X_2\|_2}, \|X_1\|_2, \|X_2\|_2$ are jointly independent, β, $\frac{X_1}{\|X_1\|_2}$ and $\frac{X_2}{\|X_2\|_2}$ are independent as well. As a result, we derive that

$$\frac{X_1}{\|X\|} = \frac{X_1}{\|X_1\|} \cdot \frac{\|X_1\|}{\|X\|} =_d \beta U^{(m)}, \tag{2.23}$$

and consequently $\frac{X_2}{\|X\|} =_d (1-\beta^2)^{1/2} U^{k-m}$. □

We now follow the proofs given by Frahm (2004) to obtain explicit stochastic representations of conditional distributions with the help of the above theorem.

Theorem 2.16. *Let $X \sim EC_p(\mu, \Sigma, \phi)$, where $\Sigma \in \mathbb{R}^{p\times p}$ is positive definite with rank$(\Sigma) = r$. We partition X as $X = (X_1^T, X_2^T)^T$ with dim$(X_1) = k \leq r$ and $\mu = (\mu_1^T, \mu_2^T)^T$. Further assume that*

$$C = \begin{pmatrix} C_{11} & 0 \\ C_{21} & C_{22} \end{pmatrix} \text{ with dimensions } \begin{pmatrix} k \times k & k \times (r-k) \\ (p-k)\times k & (p-k)\times(r-k) \end{pmatrix}$$

is the generalized Cholesky root of Σ. Then a regular conditional distribution of X_2 given $X_1 = x_1$ is the elliptical distribution that has the stochastic representation:

$$(X_2|X_1 = x_1) =_d \mu^* + \mathcal{R}^* C_{22} U^{(r-k)}, \tag{2.24}$$

where

- $U^{(r-k)}$ *is uniformly distributed on* S^{r-k-1}
- $\mathcal{R}^* =_d (\mathcal{R}\sqrt{1-\beta}|\mathcal{R}\sqrt{\beta}U^{(k)} = C_{11}^{-1}(x_1 - \mu_1))$ *with* $\beta \sim Beta(\frac{k}{2}, \frac{r-k}{2})$
- $\mu^* = \mu_2 + C_{21}C_{11}^{-1}(x_1 - \mu_1)$.

Proof. By Theorem 2.15, we have

$$U^{(r)} = \begin{pmatrix} U_1^{(r)} \\ U_2^{(r)} \end{pmatrix} =_d \begin{pmatrix} \sqrt{\beta}\cdot U^{(k)} \\ \sqrt{1-\beta}\cdot U^{(r-k)} \end{pmatrix}. \tag{2.25}$$

Substituting this result into the stochastic representation of X, we get

$$X = (X_1^T, X_2^T)^T =_d \begin{pmatrix} \mu_1 + C_{11}\mathcal{R}\sqrt{\beta}U^{(k)} \\ \mu_2 + C_{21}\mathcal{R}\sqrt{\beta}U^{(k)} + C_{22}\mathcal{R}\sqrt{1-\beta}U^{(r-k)} \end{pmatrix}. \tag{2.26}$$

Since $X_1 = x_1$, we have $x_1 = \mu_1 + C_{11}\mathcal{R}\sqrt{\beta}U^{(k)}$ and consequently $\mathcal{R}\sqrt{\beta}U^{(k)} = C_{11}^{-1}(x_1 - \mu_1)$. As a result,

$$\mu^* = \mu_2 + C_{21}\mathcal{R}\sqrt{\beta}U^{(k)} = \mu_2 + C_{21}C_{11}^{-1}(x_1 - \mu_1)$$

and

$$\mathcal{R}^* =_d (\mathcal{R}\sqrt{1-\beta}|\mathcal{R}\sqrt{\beta}U^{(k)} = C_{11}^{-1}(x_1 - \mu_1)).$$

□

Remark 2.17. In fact, we do not need to calculate the Cholesky root of the matrix Σ to find the conditional distributions as they can be expressed directly through the components of Σ. We adopt the same notation used in the theorem above. Let

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad \text{with dimensions} \quad \begin{pmatrix} k \times k & k \times (p-k) \\ (p-k) \times k & (p-k) \times (p-k) \end{pmatrix}.$$

We observe that

$$C_{21}C_{11}^{-1} = (C_{21}C_{11}^T)(C_{11}^{T-1}C_{11}^{-1}) = \Sigma_{21}\Sigma_{11}^{-1} \tag{2.27}$$

and

$$\begin{aligned} C_{22}C_{22}^T &= C_{21}C_{21}^T + C_{22}C_{22}^T - C_{21}C_{21}^T \\ &= (C_{21}C_{21}^T + C_{22}C_{22}^T) - (C_{21}C_{11}^T)(C_{11}^{T-1}C_{11}^{-1})(C_{11}C_{21}^T) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \end{aligned} \tag{2.28}$$

Given these two equalities, we can replace components of the Cholesky root C with that of Σ. Hence, $(X_2|X_1 = x_1) \sim EC_{p-k}(\mu^*, \Sigma^*, \phi^*)$ with

$$\begin{aligned} \mu^* &= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \\ \Sigma^* &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \tag{2.29}$$

while ϕ^* corresponds to the characteristic generator of $\mathcal{R}^* U^{(r-k)}$.

To facilitate future discussions, we also summarise the mean and covariance results of $(X_2|X_1 = x_1)$ in the following corollary.

Corollary 2.18. *Beginning as in Theorem 2.16, we have*

$$\mathrm{E}(X_2|X_1 = x_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1),$$

and

$$\mathrm{Var}(X_2|X_1 = x_1) = w(x_1)(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}),$$

where $w(x_2)$ is a function of x_1 through the quadratic form $(x_1 - \mu_1)^T\Sigma_{11}^{-1}(x_1 - \mu_1)$.

Proof. Based on our previous discussions about moments of elliptically contoured distributions, this corollary is a direct result of Theorem 2.16 and Remark 2.17. □

Remark 2.19. For the formula of $\mathrm{Var}(X_2|X_1 = x_1)$, Kelker (1970) showed that w is constant if and only if $X = (X_1^T, X_2^T)^T$ is normally distributed.

The reason that we are interested in conditional distributions of elliptically distributed random vectors is that their conditional distributions enjoy several nice properties. To close this section, we introduce the most important result of this chapter, proved by Eaton (1986), about the expected value of conditional distributions of elliptically contoured distributions.

Theorem 2.20. *Assume the random vector X in $\mathbb{R}^p$ has a mean vector. Suppose $v \neq 0$ is an arbitrary p-dimensional vector. Then, for any vector u that is orthogonal to v,*

$$\mathrm{E}(u^T X | v^T X) = 0, \tag{2.30}$$

if and only if X is spherical.

Proof. Let $\varphi(t) = \mathrm{E}\{\exp(it^T X)\}$ be the characteristic function of X. We note that given that the mean vector of X exists, the gradient of φ exists and

$$\nabla \varphi(t) = i\mathrm{E}\{X \exp(it^T X)\}. \tag{2.31}$$

To prove the statement, we first assume that (2.30) holds. Then because

$$\mathrm{E}\mathrm{E}\{u^T X \exp(iv^T X) | vX\} = \mathrm{E}\{\exp(iv^T X)\mathrm{E}(u^T X|v^T X)\} = E[0] = 0 \tag{2.32}$$

for all u such that $u^T v = 0$ and (2.31),

$$u^T \nabla \varphi(v) = 0. \tag{2.33}$$

Now consider a smooth curve $c : (0, 1) \longmapsto \{x| \, \|x\| = r\}$ such that, for any Γ in the orthogonal group O_p, $c(z_1) = t$ and $c(z_2) = \Gamma t$ for some $z_1, z_2 \in (0, 1)$. As $\|c(z)\|^2 = r^2$ for all $z \in (0, 1)$, we have

$$(\dot{c}(z))^T c(z) = 0 \qquad \forall z \in (0, 1). \tag{2.34}$$

The vector of derivatives $\dot{c}$ is perpendicular to c at any $z \in (0,1)$. Combining the results of (2.33) and (2.34), we derive that

$$\frac{d}{dz}\varphi(c(z)) = (\dot{c}(z))^T \nabla \varphi(c(z)) = 0. \tag{2.35}$$

Hence, the characteristic function φ is constant over the whole curve c and consequently,

$$\varphi(t) = \varphi(\Gamma t), \qquad \forall \Gamma \in O_p,$$

which indicates that X is spherically distributed.

For the other direction, we consider the random vector $Y := (u,v)^T X = (u^T X, v^T X)^T$. Since X is spherically distributed with $\mathrm{E}(X) = 0$, Y, as a linear transformation of X, has an elliptical distribution $Y \sim EC(0, \Sigma, \phi)$, where

$$\Sigma = (u,v)^T(u,v) = \begin{pmatrix} u^T u & 0 \\ 0 & v^T v \end{pmatrix}.$$

Finally, a direct application of Corollary 2.18 gives the desired result. □

We note that the theorem above can be generalised to matrices. Let Φ be an arbitrary $k \times p$ matrix, with $k \leq p$. Define P_Φ to be the projection operator for the column space of Φ and $Q_\Phi = I_p - P_\Phi$. Then by the same line of reasoning, we can easily show that $\mathrm{E}(Q_\Phi x | \Phi^T x) = 0$ for all Φ if and only if the random vector X is spherically distributed. Furthermore, because the expected value operator E is linear, it can also be derived that, for all Φ,

$$\mathrm{E}(x|\Phi^T x) = \mathrm{E}(P_\Phi x + Q_\Phi x|\Phi^T x) = \mathrm{E}(P_\Phi x|\Phi^T x) + \mathrm{E}(Q_\Phi x|\Phi^T x) = P_\Phi x, \tag{2.36}$$

if and only if the random vector X is spherically distributed.

Finally, when elliptically contoured distributions are considered, we observe that since elliptically contoured distributions are simply affine transformations of spherically distributed random vectors, the equality (2.36) implies that

$E(x|\Phi^T x)$ is a linear function of $\Phi^T x$ for all conforming matrices Φ if and only if X is an elliptically contoured random vector.

This is an important property of elliptically distributed random vectors. We will frequently refer back to this property when we study sufficient dimension reduction methods in later chapters.

Chapter 3

Central subspaces

We start investigating sufficient dimension reduction (SDR) methods from this chapter onwards. In the introduction, we mentioned that we are interested in finding a minimum dimension reduction subspace. However, as we will see shortly, such a minimum dimension reduction subspace may not be unique. This will lead to complications and misleading results when we apply SDR methods. To facilitate our discussions of SDR methods, it is important that we deal with the issue of non-uniqueness first. One possible solution, proposed by Cook (2009), is to introduce the concept of central dimension reduction subspaces (or central subspaces). A central dimension reduction subspace is the unique minimum dimension reduction subspace when it exists. Cook suggested that we should restrict ourselves to the class of regressions for which the central subspace exists to ensure the uniqueness of the minimum dimension reduction subspace. In this chapter, we focus on studying central subspaces. To understand the need of central subspaces, we will carefully study the abtract mathematical problem of sufficient dimension reduction and dimension reduction subspaces. Then, we will closely examine the conditions that ensure the existence of the central dimension reduction subspace. We need to determine whether these conditions are weak enough for Cook's idea to be relevant in practice.

3.1 Conditional Independence

To facilitate our studies of sufficient dimension reduction, we first present some useful results on conditional independence, which will be needed in the following discussions.

Proposition 3.1. *Let U, V, W be random vectors. Then, $U \perp\!\!\!\perp V|W$ if and only if $U \perp\!\!\!\perp (V,W)|W$.*

Proposition 3.2. *Let U, V, W be random vectors and assume U^* is a function of U. Then, if $U \perp\!\!\!\perp V|W$,*

1. *$U^* \perp\!\!\!\perp V|W$,*
2. *$U \perp\!\!\!\perp V|(W,U^*)$.*

Proposition 3.3 (conditional independence)**.** *Assume U, V, W and Z are random vectors. Then the following two conditions are equivalent:*

- *$U \perp\!\!\!\perp W|(Z,V)$ and $U \perp\!\!\!\perp V|Z$,*
- *$U \perp\!\!\!\perp (V,W)|Z$.*

For the purpose of this chapter, we omit the proofs for these propositions. Conditional independence is an important but challenging area of statistics and its results often play essential roles in helping us understand large data sets. For detailed proofs of the propositions above and background knowledge on conditional independence in general, we refer interested readers to Basu and Pereira (1983), Dawid (1979a), Dawid (1979b).

3.2 Problem set up

We start by setting up a mathematical framework of sufficient dimension reduction. Suppose y is a univariate response and x is a p-dimensional vector of explanatory variables. We have briefly mentioned in the introduction that the key assumption of sufficient dimension reduction methods is that there exist p-dimensional vectors $\beta_1, \ldots, \beta_k$ such that there is no loss of information when we regress the response variable y on $\beta_1^T x, \ldots, \beta_k^T x$ instead of x. In other words, the relationship between y and x can be described by the following model:

$$y = f(\beta_1^T x, \beta_2^T x, \ldots, \beta_k^T x, \epsilon), \tag{3.1}$$

where f is an arbitrary unknown function on $\mathbb{R}^{k+1}$ and ϵ is independent of x. When k is smaller than p, the dimension of the predictor x, we achieve dimension reduction.

At first sight, it feels that we need to find $\Phi := (\beta_1, \dots, \beta_k)$ in order to reduce the dimension of x. However, the problem is that Φ is not identifiable. To see it, let $S(\Phi)$ denotes a subspace that is spanned by column vectors of Φ and let $B = (b_1, \dots, b_q)$ be a basis matrix of the subspace $S(\Phi)$. Because $b_1, \dots, b_q$ are basis vectors, we can write each $\beta_1, \dots, \beta_k$ as a linear combinations of $b_1, \dots, b_q$. As the result, we can equally state (3.1) as

$$y = g(b_1^T x, \dots, b_q^T x, \epsilon) \tag{3.2}$$

for some function g on $\mathbb{R}^{q+1}$ and we derive a solution for B instead. In fact, the argument holds for any matrix B such that $S(B) = S(\Phi)$. Since it is impossible to solve for a particular matrix Φ, in sufficient dimension reduction, we are interested in identifying the subspace $S(\Phi)$. The subspace $S(\Phi)$ is called a dimension reduction subspace (DRS).

Before we carefully study dimension reduction subspaces, we point out that models other than (3.1) have been used in the literature of SDR methods. For instance, Cook (1994a,b, 1996) suggested that we can summarise the relationship between y and x using conditional independence. That is,

$$y \perp\!\!\!\perp x | \Phi^T x, \tag{3.3}$$

where $\Phi := (\beta_1, \dots, \beta_k)$ and $\perp\!\!\!\perp$ means independent of. Since we have assumed that all regression information is contained within $\Phi^T x$, y should be independent of x once we are given $\Phi^T x$. Furthermore, we can represent the underlying assumption of sufficient dimension reduction with conditional distribution functions:

$$F_{y|x}(a) = F_{y|\Phi^T x}(a) \qquad \text{for all } a \in \mathbb{R}. \tag{3.4}$$

The conditional distribution function of y given $\Phi^T x$ is the same as the conditional distribution function of y given x (Ma and Zhu, 2013; Zeng and Zhu, 2010).

Although models (3.1), (3.3) and (3.4) are different in formulations, they are in fact equivalent to each other.

Lemma 3.4 (Zeng and Zhu (2010))**.** *Assume the response variable y is one dimensional and $x \in \mathbb{R}^p$ is a vector of explanatory variables. Then models (3.1), (3.3) and (3.4) are equivalent.*

Proof. Model (3.3) is equivalent to model (3.4) by the definition of conditional independence (Basu and Pereira, 1983). Therefore, it is sufficient to show that model (3.1) and model (3.3) are equal.

First, we assume (3.1) holds. We observe that, given $\Phi^T x$, y depends on ϵ only. Since x is independent of ϵ, x is independent of y given $\Phi^T x$. Thus, (3.3) holds. The other direction is more involved and to prove it, an appropriate measure needs to be introduced. We omit the proof of this direction and refer interested readers to (Zeng and Zhu, 2010). □

In the following discussions, we will mainly use the formulation (3.3). There are two main reasons for this choice. Firstly, we observe that, apart from vectors $\beta_1, \ldots, \beta_k$, formulation (3.1) requires an arbitrary link function f on $\mathbb{R}^{k+1}$ and an independent error ϵ. In some applications, conceiving a link function f or a meaningful independent random error can be an obstacle. For instance, Cox and Snell (1968) showed that it is not possible to construct an independent error based on just y and $\Phi^T x$ when y is a binary variable, taking values 0 and 1 with probability depending on $\Phi^T x$. Formulation (3.3) avoids this drawback by using conditional independence instead of introducing f and ϵ. Secondly, Basu and Pereira (1983) proved several useful properties of conditional independence (some are covered in section 3.1). Since these properties play an important role in the analysis of SDR methods, adopting formulation (3.3) will greatly facilitate our discussions on SDR methods in later chapters.

Remark 3.5. We point out that there is an underlying limitation of all sufficient dimension reduction models. Since SDR approach assumes that the explanatory effect of x about y is manifested through a few linear combinations of covariates, SDR models restrict parsimonious characterizations of $y|x$ to linear manifolds. Therefore, even for simple nonlinear manifolds, we may need to take all of $\mathbb{R}^p$ to characterize them (Cook, 2009). For instance, the only way to describe $y \perp\!\!\!\perp x | \|x\|$ with SDR models is to let $\Phi = I_p$ and $S(\Phi) = \mathbb{R}^p$.

3.3 Dimension reduction subspaces

Given a univariate response variable y and x of p-dimensional covariates, we want to identify dimension reduction subspaces (DRS) for $y|x$. Recall that a subspace $S(\Phi)$ is called a dimension reduction subspace if

$$y \perp\!\!\!\perp x | \Phi^T x$$

holds.

We first note that a DRS always exists. Because $y \perp\!\!\!\perp x|x$ is always true, we can find a DRS by letting $\Phi = I_p$. For the same reason, a dimension reduction subspace need not be unique. For example, if there exists a matrix $B \neq I_p$ such that $y \perp\!\!\!\perp x|B^T x$ holds, then both $S(B)$ and $S(I_p)$ are valid dimension reduction subspaces. Because our goal is to maximally reduce the dimension of x, what we are really interested in is identifying a DRS with minimum dimension among all possible DRSs. A subspace S is said to be a minimum DRS for $y|x$ if S is a DRS and $\dim(S) \leq \dim(S_{drs})$ for all DRSs S_{drs} (Cook, 1994a,b, 2009). Hence, we have narrowed down the subspaces of interest to minimum DRSs. A minimum dimension reduction subspace always exists by definition. To better understand minimum DRSs, we look at the following property of minimum DRSs.

Proposition 3.6 (Cook (2009))**.** *Let $S(\Phi)$ be a minimum dimension reduction subspace for the regression of y on x and assume $A \in \mathbb{R}^{p\times p}$ is an arbitrary full rank matrix. Then if $z = A^T x$, $S(A^{-1}\Phi)$ is a minimum dimension reduction subspace for the regression of y on z.*

Proof. To prove that $S(A^{-1}\Phi)$ is a minimum DRS for $y|z$, we first show that $S(A^{-1}\Phi)$ is a DRS for $y|z$. By the Proposition 3.2, we have $y \perp\!\!\!\perp x|\Phi^T x$ if and only $y \perp\!\!\!\perp A^T x|\Phi^T x$. Because A is full rank, it follows that $y \perp\!\!\!\perp x|\Phi^T x$ if and only $y \perp\!\!\!\perp z|(A^{-1}\Phi^T)^T z$. Therefore, $S(A^{-1}\Phi)$ is a DRS for $y|z$ by definition. Next, suppose there exists a DRS $S(C)$ for $y|z$ such that $\dim\{S(C)\} \leq \dim\{S(A^{-1}\Phi)\}$. Since $y \perp\!\!\!\perp A^T x|C^T A^T x$ implies $y \perp\!\!\!\perp x|(AC)^T x$, $S(AC)$ is a DRS for $y|x$. Because A is full-rank and $\dim\{S(C)\} \leq \dim\{S(A^{-1}\Phi)\}$, it follows that $\dim\{S(AC)\} \leq \dim\{S(\Phi)\}$, which contradicts the fact that $S(\Phi)$ is a minimum DRS. Thus, $S(A^{-1}\Phi)$ is a minimum dimension reduction subspace for the regression of y on z. □

This property gives a clear formula for a minimum DRS when the predictors are linearly transformed with full-rank matrices. With the help of this property, we can derive a minimum DRS for $y|x$ by standardizing the predictors first. Then, we identify a minimum DRS for the regression of y on the standardised predictors z. Finally, a linear transformation of the minimum DRS for $y|z$ gives us the desired result. Because it is often easier to deal with standardized variables, we will use this strategy frequently in the following chapters when we develop sufficient dimension reduction methods.

3.4 Central subspaces

Although a minimum DRS exists for all regressions, minimum DRSs are not generally unique. To see this, we consider the following example provided by Cook (2009).

Example 3.1. *Let $p = 2$. Assume that $x = (x_1, x_2)^T$ distributed uniformly on the unit circle $\|x\| = 1$. The true model is*

$$y|x = x_1^2 + \epsilon,$$

where the random error ϵ is independent of $\boldsymbol{x}$.
We observe that, since $x_1^2 + x_2^2 = 1$,

$$y|x = x_1^2 + \epsilon = (1 - x_2^2) + \epsilon.$$

Thus, both $S((1,0)^T)$ and $S((0,1)^T)$ are dimension reduction subspaces. Because both of them are one dimensional subspaces, $S((1,0)^T)$ and $S((0,1)^T)$ are minimum DRSs.

The non-uniqueness of minimum dimension reduction subspaces could lead to erroneous conclusions at later stages when we attempt to recover such minimum dimension reduction subspaces. For instance, in the paper of Chiaromonte and Cook (2002), it is mentioned that when using sliced inverse regression (Li, 1991) to recover minimum dimension reduction subspaces for the example above, we often take the minimum dimension reduction subspace as the intersection of $S((1,0)^T)$ and $S((0,1)^T)$, which is $\{0\}$. As a result, x and y are wrongly concluded to be independent.

To deal with the issues caused by non-uniqueness of minimum dimension reduction subspaces, we adopt Cook's idea (Cook, 1994a,b, 1996, 2009). Cook introduced a new type of space called central dimension reduction subspaces. When a regression has a central dimension reduction subspace, the regression can only have a unique minimum DRS. Cook suggested that we can avoid the problem of non-unique minimum DRSs by restricting our attention to regressions for which the central dimension reduction subspace exists. We give the formal definition of central dimension reduction subspaces below.

Definition 3.7 (Central dimension reduction subspace)**.** A subspace S is a central dimension reduction subspace (or central subspace for short) for the regression of y on x if S is a dimension reduction subspace and $S \subseteq S_{drs}$ for all dimension reduction subspaces S_{drs}. We denote the central dimension-reduction subspace by $S_{y|x}$ or $S_{y|x}(\Phi)$ when a matrix Φ that spans the central subspace needs to be referred to explicitly.

When a central subspace exists, it is the unique minimum DRS by definition. We can formally prove this statement by contradiction. Assume S_1 is a second minimum dimension reduction subspace for an arbitrary regression with a central subspace $S_{y|x}$. Then, because $S_{y|x} \subseteq S_1$ and $\dim(S_{y|x}) = \dim(S_1)$, we much have $S_{y|x} = S_1$. Therefore, the central subspace is the unique minimum DRS when it exists.

Nevertheless, it should be noted that a central subspace does not necessarily exist even when there is a unique minimum dimension reduction subspace. To see this, we consider a similar example but with $p = 3$.

Example 3.2. *Let $x \in \mathbb{R}^3$ be uniformly distributed on a unit sphere so that $\|x\| = 1$. We assume that*

$$y|x = x_1^2 + \epsilon.$$

For this example, the unique minimum direction reduction subspace S_1 is spanned by the vector $(1, 0, 0)^T$. However, since

$$y|x = x_1^2 + \epsilon = 1 - x_2^2 - x_3^2 + \epsilon,$$

another possible dimension reduction subpsace S_2 is spanned by vectors $(0, 1, 0)^T$ and $(0, 0, 1)^T$. The intersection of these two dimension reduction spaces is the origin.

In this case, the central subspace does not exist and the unique minimum dimension reduction subspace is not a central subspace.

3.5 Existence of the central subspace

To follow Cook's idea, it is important to identify conditions that ensure the existence of the central subspace for regression problems. Apart from enabling us to decide whether the results based on central subspaces are applicable to the regression problems of interest, investigating these conditions also allow us to determine whether the class of regression problems for which the central subspace exists is large enough for Cook's idea to be of practical use.

In order to study the existence conditions of the central spaces, we start by looking at a similar example to the one above.

Example 3.3. *Let $x \in \mathbb{R}^3$ be uniformly distributed on a unit sphere so that $\|x\| = 1$. This time, we modify the Example 3.2 slightly by letting*

$$y|x = x_1^2 + x_1 + \epsilon.$$

In this case, the central subspace exists and it is spanned by the vector $(1, 0, 0)^T$. As the sign of x_1 cannot be determined by x_2 and x_3, all possible dimension reduction subspaces must include the vector $(1, 0, 0)^T$. Since the space spanned by $(1, 0, 0)^T$ is a dimension reduction subspace, it is by definition a central subspace.

This and Example 3.2 in the previous section show that the existence of a central subspace depends on the conditional distribution of $y|x$ and on the marginal distribution of x.

To further explore the conditions that affect the existence of a central subspace, we assume that a problem of interest has a minimum dimension reduction subspace $S_m(\Phi)$ and we also let $S_{drs}(B)$ be an arbitrary dimension reduction subspace. Then, by the definition of DRS, we have

$$y \perp\!\!\!\perp x|x, \quad y \perp\!\!\!\perp x|\Phi^T x, \quad y \perp\!\!\!\perp x|B^T x.$$

Since $B^T x$ can be seen as a function of x and we know that $y \perp\!\!\!\perp x|\Phi^T x$, Proposition 3.2 shows that

$$y \perp\!\!\!\perp x|(\Phi^T x, B^T x).$$

Due to the equivalence between formulations (3.3) and (3.4), we thus have

$$F_{y|x}(a) = F_{y|\Phi^T x, B^T x}(a) = F_{y|\Phi^T x}(a) = F_{y|B^T x}(a), \quad \forall a \in \mathbb{R}. \tag{3.5}$$

The above equality is important because it helps us uncover essential relationships for studying central subspaces. We observe that, given the equality(3.5), for all $a \in \mathbb{R}$,

$$\begin{aligned} F_{y|\Phi^T x}(a) &= F_{y|B^T x}(a) \\ &= E_{\Phi^T x|B^T x}[F_{y|\Phi^T x, B^T x}(a)] \quad \text{(by defn of conditional expectation)} \\ &= E_{\Phi^T x|B^T x}[F_{y|\Phi^T x}(a)]. \end{aligned} \tag{3.6}$$

Thus, the fact that $S_m(\Phi)$ and $S_{drs}(B)$ are dimension reduction subspaces implies that $F_{y|\Phi^T x}(a) = E_{\Phi^T x|B^T x}[F_{y|\Phi^T x}(a)]$. In other words, $S_m(\Phi)$ and $S_{drs}(B)$ being dimension reduction subspaces ensures that, with respect to the conditional distribution of $\Phi^T x|B^T x$,

$F_{y|\Phi^T x}(a)$ is constant with probability 1. So, no further information is supplied to $F_{y|\Phi^T x}(a)$ by $B^T x$ given $\Phi^T x$.

The equality clearly holds when $S_m(\Phi)$ is a central subspace. That is $S_m(\Phi) \subset S_{drs}(B)$. However the equality (3.6) may hold under other conditions. If we can identify these conditions, we may be able to force the existence of a central subspace by imposing restrictions so that the equality (3.6) holds only when $S_m(\Phi)$ is a central subspace.

To explore different conditions for the equality (3.6), we start by assuming that $S_m(\Phi)$ is not a central subspace. Without loss of generality, let $S(C) := S_m(\Phi) \cap S_{drs}(B)$ and also let $S(\Phi_1) = S(C)^{\perp S(\Phi)}$ and $S(B_1) = S(C)^{\perp S(B)}$. Here, $S(C)^{\perp S}$ means the orthogonal complement to $S(C)$ in S. Since $S_m(\Phi)$ is not central, $S(\Phi_1)$ and $S(B_1)$ are nontrivial subspaces. Then, intuitively, the equality (3.6) implies that the information provided by $S(\Phi_1)$ to the response variable is the same as that provided by $S(B_1)$. To be more specific, if the information about the response variable contained in $S(\Phi_1)$ is contributed via a function of $f_\Phi(\Phi_1^T x)$, then there exists a function $f_B(B_1^T x)$ such that $f_B(B_1^T x) = f_\Phi(\Phi_1^T x)$. $f_\Phi(\Phi_1^T x)$ can be replaced by $f_B(B_1^T x)$.

Example 3.4. *To better understand this statement, we recall Example 3.2. Let $x \in \mathbb{R}^3$ be uniformly distributed on a unit sphere so that $\|x\| = 1$. We assume that*

$$y|x = x_1^2 + \epsilon.$$

In this case, $S_m(\Phi) = S((1,0,0)^T)$, $S_{drs}(B) = S((0,1,0)^T, (0,0,1)^T)$ and $S(C) = \{0\}$. We also note that

$$f_\Phi(\Phi_1^T x) = f_\Phi(\Phi^T x) = (\Phi^T x)^2 = x_1^2.$$

Moreover, since x follows a spherical distribution, we easily observe that by defining $f_B(B_1^T x)$ as

$$f_B(B_1^T x) = f_B(B^T x) := 1 - x_2^2 - x_3^2,$$

we can replace $f_\Phi(\Phi_1^T x)$ with $f_B(B_1^T x)$. Here, we tie the regression function to the distribution of x and thereby achieve the equality (3.6) without forcing the centrality. The possibility of replacement hence precludes the existence of a central subspace.

As a result, to ensure the existence of a central subspace, we have to eliminate the possibility of such replacement. In other words, if there exist functions such that $f_\Phi(\Phi_1^T x) =$

$f_B(B_1^T x)$, both functions f_Φ and f_B should be trivial. In order to enforce this requirement, we follow Chiaromonte and Cook (2002)'s approach and introduce a lemma from real analysis first.

Lemma 3.8. *Let $\Omega \subseteq \mathbb{R}^p$ be an open set, and $g : \mathbb{R}^p \to \mathbb{R}^1$ an analytic function. Also, let P_S be the orthogonal projection operator on S with respect to the standard inner product. Assume S_1 and S_2 are any two subspaces of $\mathbb{R}^p$. Then if*

$$g(x) = g(P_{S_1}x) = g(P_{S_2}x), \quad \forall x \in \Omega, \tag{3.7}$$

we have

$$g(x) = g(P_{S_1 \cap S_2}x), \quad \forall x \in \Omega. \tag{3.8}$$

Proof. Let $T = S_1 \cap S_2$. In addition, let $T_1 = T^{\perp S_1}$ and $T_2 = T^{\perp S_2}$. T_1 is the orthogonal complement of T of the subspace S_1 and T_2 is the orthogonal complement of T of the subspace S_2. Then for any $x \in \Omega$, we can decompose $P_{S_1}x$ and $P_{S_2}x$ as follows:

$$P_{S_1}x = P_T x + P_{T_1}x,$$

$$P_{S_2}x = P_T x + P_{T_2}x.$$

Here, we note that $P_{T_1}x$ and $P_{T_2}x$ are linearly independent by the way they are defined.

Now, we recall the defining property for an analytic function g is that, for any $a \in \mathbb{R}^p$, one can write

$$g(z) = b_0 + \sum_{k_1,\ldots,k_p=1}^{\infty} b_{k_1,\ldots,k_p}(z_1 - a_1)^{k_1} \ldots (z_p - a_p)^{k_p} \tag{3.9}$$

where z is in the neighbourhood of a and $b_0, b_{k_1,\ldots,k_p}$ are constants. Let $a = P_T x$. Then by this property, we have

$$g(P_{S_1}x) = b_0 + \sum_{k_1,\ldots,k_p=1}^{\infty} b_{k_1,\ldots,k_p}(u_1)^{k_1} \ldots (u_p)^{k_p}$$

and

$$g(P_{S_2}x) = b_0 + \sum_{k_1,\ldots,k_p=1}^{\infty} b_{k_1,\ldots,k_p}(v_1)^{k_1} \ldots (v_p)^{k_p},$$

where $(u_1, \dots, u_p)^T = P_{T_1}x$ and $(v_1, \dots, v_p)^T = P_{T_2}x$. Since by assumption $g(P_{S_1}x) = g(P_{S_2}x)$, the two summations are equal as well:

$$b_0 + \sum_{k_1,\dots,k_p=1}^{\infty} b_{k_1,\dots,k_p}(u_1)^{k_1}\dots(u_p)^{k_p} = b_0 + \sum_{k_1,\dots,k_p=1}^{\infty} b_{k_1,\dots,k_p}(v_1)^{k_1}\dots(v_p)^{k_p}. \tag{3.10}$$

However, given that $(u_1, \dots, u_p)^T$ and $(v_1, \dots, v_p)^T$ are linearly independent, the above equality (3.10) holds if and only if $b_{1,\dots,1}, \dots = 0$. It follows that

$$g(x) = g(P_{S_1}x) = g(P_{S_2}x) = b_0 = g(P_T x).$$

Since x is arbitrary, the lemma is proved. □

This lemma says that given $g(x) = g(P_{S_1}x) = g(P_{S_2}x)$, the fact that g is analytic ensures the information for evaluating $g(x)$ is completely captured by the projection of x into the intersection of the subspaces S_1 and S_2. We thus can use this lemma to derive the following proposition to secure the existence of the central subspace. We let $\mathcal{L}_X$ and $Supp_X$ denote the probability law and the closed support of X respectively.

Proposition 3.9 (Chiaromonte and Cook (2002))**.** *Assume that $Supp_X$ contains an open set Ω with $\mathcal{L}_X(\Omega) = 1$. If we are given that $Y \perp\!\!\!\perp X|E(Y|X)$, where Y admits finite first order moments and $E(Y|X)$ can be expressed as an analytic function of X, almost surely, the central subspace exists.*

Proof. Let S_m be a minimum dimension reduction subspace and S_{drs} an arbitrary dimension reduction subspace. Then by definition, $Y \perp\!\!\!\perp X|P_{S_m}X$ and $Y \perp\!\!\!\perp X|P_{S_{drs}}X$.

We note that we are given $Y \perp\!\!\!\perp X|E(Y|X)$, so the regression problem of interest is characterized by its regression function. When $Y \perp\!\!\!\perp X|E(Y|X)$ holds, Cook (1996) showed that, for any arbitrary DRS S_{drs}, $Y \perp\!\!\!\perp X|P_{S_{drs}}X$ if and only if $Y \perp\!\!\!\perp X|E(Y|P_{S_{drs}}X)$ and, additionally, $E(Y|X) = E(Y|P_{S_{drs}}X)$. Therefore, we have

$$E(Y|X) = E(Y|P_{S_m}X) = E(Y|P_{S_{drs}}X).$$

Since $E(Y|X)$ can be rewritten as an analytic function g of X, we can express the above equality as $g(X) = g(P_{S_m}X) = g(P_{S_{drs}}X)$, almost surely. It follows that $g(x) = g(P_{S_m}x) = g(P_{S_{drs}}x)$, $\forall x \in \Omega$. Then a direct application of Lemma 3.8 gives us $g(x) = g(P_{S_m \cap S_{drs}}x)$ for all $x \in \Omega$, which in turn implies that $Y \perp\!\!\!\perp X|E(Y|P_{S_m \cap S_{drs}}X)$. Because $Y \perp\!\!\!\perp$

$X|E(Y|P_{S_m \cap S_{drs}}X)$ holds if and only if $Y \perp\!\!\!\perp X|P_{S_m \cap S_{drs}}X$, $S_m \cap S_{drs}$ is a dimension reduction subspace.

Finally, given that S_m is a minimum DRS, we must have $\dim(S_m \cap S_{drs}) = \dim(S_m)$. It follows that $S_m = S_m \cap S_{drs} \subset S_{drs}$. Since S_{drs} is arbitrary, S_m is contained in all possible dimension reduction subspaces and hence S_m is the central subspace. □

The proposition is applicable to many standard regression models. For instance, it can be used on additive-error models. If the true model is $y = g(x) + \epsilon$ with $x \perp\!\!\!\perp \epsilon$, $E(\epsilon) = 0$ and $g(x)$ analytic, we have $E(y|x) = g(x)$ and $y \perp\!\!\!\perp x|g(x)$. In fact, we can also apply the proposition to problems with heteroscedastic variance, as the conditions required by the proposition are relatively loose. Consider the model: $y = g(x) + \sigma(g(x))\epsilon$, where $\epsilon \perp\!\!\!\perp X$, $E(\epsilon) = 0$ and $g(x)$ is analytic. In this case, we still have $E(y|x) = g(x)$ and $y \perp\!\!\!\perp x|g(x)$.

However, to apply this proposition, we do require that $y \perp\!\!\!\perp x|E(y|x)$ and the conditional mean $E(y|x)$ can be expressed as an analytic function of the predictor. For the idea of central subspaces to be of more general use, we need to develop conditions that guarantee the existence of the central subspace without constraining $Y|X$ in any fashion. Fortunately, this is achieved by the following proposition of Chiaromonte and Cook (2002); Cook (1994a, 1996).

Proposition 3.10. *Assume that $Supp_X$ contains an open and convex set Ω with $\mathcal{L}_X(\Omega) = 1$. Then the central subspace exists for the regression of any response Y on X.*

Proof. Assume $S(A)$ and $S(B)$ are arbitrary dimension reduction subspaces. Also, let $S(C) = S(A) \cap S(B)$, $S(A_1) = S(C)^{\perp S(A)}$ and $S(B_1) = S(C)^{\perp S(B)}$. We first want to show that $S(C)$ is a dimension reduction subspace as well.

Because $S(A)$ and $S(B)$ are DRSs, by the equality (3.5), we have

$$F_{y|x}(a) = F_{y|A_1^T x, B_1^T x, C^T x}(a) = F_{y|A_1^T x, C^T x}(a) = F_{y|B_1^T x, C^T x}(a), \quad \forall a \in \mathbb{R}. \tag{3.11}$$

Since x has a density with convex open support and (A_1, B_1, C) is a full-rank operator, $(A_1^T x, B_1^T x, C^T x)^T$ has a density with convex open support, denoted by Ω_w. Let the conditional values for $A_1^T x, B_1^T x, C^T x$ be w_1, w_2, w_3 respectively. We observe that, by a similar argument, the distribution of $(A_1^T x, B_1^T x)|(C^T x = w_3)$ has a density with a convex open support as well. We denote this support as $\Omega_{12|3}(w_3)$. To prove $S(C)$ is a DRS, we need to show, for all $a \in \mathbb{R}$, the equality

$$F_{y|(A_1^T x = w_1, B_1^T x = w_2, C^T x = w_3)}(a) = F_{y|C^T x = w_3}(a), \quad \forall (w_1, w_2, w_3)^T \in \Omega_w, \tag{3.12}$$

holds. Because $\Omega_w = \cup_{\Omega_3} \Omega_{12|3}(w_3)$, where Ω_3 is the support of $C^T x$, we can rewrite the equality (3.12) as, for all $a \in \mathbb{R}$ and any arbitrary $w_3 \in \Omega_3$,

$$F_{y|(A_1^T x = w_1, B_1^T x = w_2, C^T x = w_3)}(a) = F_{y|C^T x = w_3}(a), \quad \forall (w_1, w_2)^T \in \Omega_{12|3}(w_3) \tag{3.13}$$

Consequently, we prove $S(C)$ is a DRS by showing the above equality instead.

Fix any $w_3 \in \Omega_3$ and let $u = (w_1, w_2)$ and $v = (w_1', w_2')$ be two arbitrary points in $\Omega_{12|3}(w_3)$. Since $\Omega_{12|3}(w_3)$ is convex and open, there exists a linked sequence $l^1 = (l_1^1, l_2^1), \ldots, l^N = (l_1^N, l_2^N) \in \Omega_{12|3}(w_3)$ such that

1. $l^1 = (l_1^1, l_2^1) = (w_1, w_2)$;
2. $l^N = (l_1^N, l_2^N) = (w_1', w_2')$;
3. for all $n = 2, \ldots, N$, either $l_1^n = l_1^{n-1}$ or $l_2^n = l_2^{n-1}$.

We claim that for all $n = 2, \ldots, N$, we have $F_{y|l^n}(a) = F_{y|l^{n-1}}(a)$ for all $a \in \mathbb{R}$. To see it, we note that l^n, l^{n-1} are linked by either $l_1^n = l_1^{n-1}$ or $l_2^n = l_2^{n-1}$. When $l_1^n = l_1^{n-1}$, by the equality (3.11), we have, for all $a \in \mathbb{R}$

$$\begin{aligned} F_{y|(A_1^T x = l_1^n, B_1^T x = l_2^n, C^T x = w_3)}(a) &= F_{y|(A_1^T x = l_1^n, C^T x = w_3)}(a) = F_{y|(A_1^T x = l_1^{n-1}, C^T x = w_3)}(a) \\ &= F_{y|(A_1^T x = l_1^{n-1}, B_1^T x = l_2^{n-1}, C^T x = w_3)}(a). \end{aligned} \tag{3.14}$$

Similarly, when $l_2^n = l_2^{n-1}$, the equality (3.11) implies that

$$\begin{aligned} F_{y|(A_1^T x = l_1^n, B_1^T x = l_2^n, C^T x = w_3)}(a) &= F_{y|(B_1^T x = l_2^n, C^T x = w_3)}(a) = F_{y|(B_1^T x = l_2^{n-1}, C^T x = w_3)}(a) \\ &= F_{y|(A_1^T x = l_1^{n-1}, B_1^T x = l_2^{n-1}, C^T x = w_3)}(a) \end{aligned} \tag{3.15}$$

for all $a \in \mathbb{R}$. Therefore, $F_{y|l^1}(a) = F_{y|l^2}(a) = \cdots = F_{y|l^N}(a)$ for all $a \in \mathbb{R}$. Since $l^1 = u$ and $l^N = v$, we obtain

$$F_{y|(A_1^T x = w_1, B_1^T x = w_2, C^T x = w_3)}(a) = F_{y|(A_1^T x = w_1', B_1^T x = w_2', C^T x = w_3)}(a) \qquad \forall a \in \mathbb{R}.$$

Finally, because u, v are arbitrary, for any $a \in \mathbb{R}$, $F_{y|C^Tx=w_3}(a)$ is constant over the set $\Omega_{12|3}(w_3)$ and the equality (3.13) follows.

Up to now, we have shown that for any DRSs $S(A)$ and $S(B)$, $S(C) = S(A) \cap S(B)$ is also a DRS. To prove the existence of the central subspace, let $S_m(\Phi)$ be a minimum DRS. Since, for any DRS $S(A)$, $S_m(\Phi) \cap S(A)$ is a DRS and $\dim(S_m(\Phi)) = \dim(S_m(\Phi) \cap S(A))$, we have $S_m(\Phi) \subset S(A)$. Consequently, $S_m(\Phi)$ is the central subspace. □

This proposition is important, because it eliminates the constraint on $Y|X$ and purely focuses on the distribution of X. Unlike $Y|X$, the object of study, the distribution of X is at least partially known and sometimes controllable. We thus can check whether the central subspace exists. In addition, we observe that the conditions required for the distribution of X are quite weak. The proposition always holds when $\mathcal{L}_X$ is absolutely continuous and $Supp_X$ is convex, conditions which are satisfied by many problems of interest. For example, the central subspace always exists for any predictor with positive density over $\mathbb{R}^p$. Even if the distribution of X does not satisfy these requirements, it is also possible to modify the distribution of X to ensure the existence of the central subspace in some cases.

So far, we have introduced possible conditions that force the existence of the central subspace. We see that these conditions are fairly weak, so the class of regressions for which the central subspace exists should be large enough to be relevant in practice. To facilitate our following discussions, we assume regression problems of interest have the central subspace thereafter.

Chapter 4

SIR and SAVE

In the following two chapters, we will study sufficient dimension reduction (SDR) methods under the assumption that central subspaces exist. Simulations of SDR methods are provided in Chapter 6.

In this chapter, we will carefully study the method Sliced Inverse Regression (SIR) and its extension Sliced Average Variance Estimation (SAVE). SIR and SAVE are two important and widely used methods. They tackle traditional challenges in a different yet efficient way, as they extract information about the central subspace via inverse regression lines. In the following discussions of these methods, we will focus on addressing two key questions. Firstly, how can we apply these methods to recover at least a portion of the central subspace? We aim to provide a step-by-step procedure for each method. Secondly, how effective are these methods? In order to make the best use of SIR and SAVE methods, we need to find their strengths and limitations respectively.

4.1 Sliced Inverse Regression

We first introduce the Sliced Inverse Regression methodology, which was proposed by Li (1991). Sliced inverse regression(SIR), as its name suggests, is a method based on the inverse regression $x|y$ instead of the forward regression $y|x$. Since the covariate x is generally of much higher dimensions than that of the response variable y, the inverse regression is significantly easier to obtain than the forward regression. In our case, y is one dimensional. The inverse regression is composed of p simple regressions $x_i|y$, $i = 1, \dots, p$, each of which can be easily computed and studied in a $2D$ plot. The key idea of SIR is to make use of the efficiency enjoyed by inverse regression to infer about central subspaces. To be more specific, we want to establish a connection between inverse regression lines and central subspaces, so we can take the advantage of inverse regression to efficiently derive at least a portion of the central subspace. The word "Sliced" is included in the name of the method because slicing techniques are used during the procedure.

4.1.1 Inverse Regression Subspace

In order to find the connection between inverse regression lines and the central subspace, we start with a simple example provided by Cook (2009).

Example 4.1. *Assume* (y, x^T) *follows a non-singular multivariate normal distribution, where* $x \in \mathbb{R}^p$ *and* $y \in \mathbb{R}$. *Also, assume that* $y \perp\!\!\!\perp x|E(y|x)$. *Let* $\Sigma_{yx} = \mathrm{Cov}(y, x), \Sigma_{xy} = \mathrm{Cov}(x, y)$, $\Sigma = \mathrm{Var}(x)$ *and* $\sigma^2 = \mathrm{Var}(y)$. *Applying Corollary 2.18 of Chapter 2, we derive the following equations for regressing* y *on* x *and regressing* x *on* y*:*

$$\mathrm{E}(y|x) = \mathrm{E}(y) + \Sigma_{yx}\Sigma^{-1}(x - \mathrm{E}(x)), \tag{4.1}$$

and

$$\mathrm{E}(x|y) = \mathrm{E}(x) + \Sigma_{xy}\sigma^{-2}(y - \mathrm{E}(y)). \tag{4.2}$$

From the first equation for $\mathrm{E}(y|x)$, *we observe that given* $y \perp\!\!\!\perp x|\mathrm{E}(y|x)$, *we have* $y \perp\!\!\!\perp x|\Sigma_{yx}\Sigma^{-1}x$ *or equivalently* $y \perp\!\!\!\perp x|\eta^T x$, *where* $\eta := (\Sigma_{yx}\Sigma^{-1})^T = \Sigma^{-1}\Sigma_{xy}$. *It follows that, the subspace* $S(\eta)$, *spanned by the columns of the matrix* η, *is a DRS. Moreover, because* η *is a* $p \times 1$ *vector, the subspace* $S(\eta)$ *is contained in any possible DRS and hence a central subspace* $S_{y|x}(\eta)$.

When the inverse regression $\mathrm{E}(x|y)$ *is considered, we note that if we define the inverse regression subspace as*

$$S_{\mathrm{E}(x|y)} = span\{\mathrm{E}(x|y) - \mathrm{E}(x) \,|\, y \in \mathbb{R}\},$$

the equality (4.2) indicates that the inverse regression subspace is spanned by $\Sigma_{xy} = \Sigma\eta$. *We omit* σ^{-2} *here, because it is a scalar and has no impact on a subspace.*

Therefore, in this simple example, the inverse regression subspace $S_{\mathrm{E}(x|y)}$ *is a one dimensional subspace spanned by the vector* $\Sigma\eta$. *We can equally write* $S_{\mathrm{E}(x|y)}$ *as* $S(\Sigma\eta)$. *Since the central space* $S_{y|x}(\eta)$ *is related to* $S(\Sigma\eta)$ *via a linear transformation* Σ*:* $\Sigma S_{y|x}(\eta) =$

$S(\Sigma\eta)$, we can easily derive the central subspace via the formula $S_{y|x}(\eta) = \Sigma^{-1}S(\Sigma\eta) = \Sigma^{-1}S_{\mathrm{E}(x|y)}$.

The above example gives us a brief idea of how inverse regression can be used to find central subspaces. In addition, we have introduced an important type of subspace: inverse regression subspace $S_{\mathrm{E}(x|y)}$. An inverse regression subspace is spanned by the centered inverse regression curve $\mathrm{E}(x|y) - \mathrm{E}(x)$ as y varies.

In essence, estimating $S_{y|x}$ with inverse regression is consisted of two steps. Firstly, we need to establish a connection between the central subspace $S_{y|x}$ and the inverse regression subspace $S_{\mathrm{E}(x|y)}$. Secondly, we approximate the $S_{\mathrm{E}(x|y)}$ of the regression of interest. As the name suggests, slicing techniques are used in estimating $S_{\mathrm{E}(x|y)}$. Once we have an estimate of $S_{\mathrm{E}(x|y)}$, we can find $S_{y|x}$ using the relationship between $S_{y|x}$ and $S_{\mathrm{E}(x|y)}$. We note that, in the example above, we have assumed (y, x^T) follows a multivariate normal distribution. In the following discussions, we will focus on applying SIR to general regression problems. We will investigate each step in detail.

4.1.2 Finding a connection between $S_{y|x}$ and $S_{\mathrm{E}(x|y)}$

We need to find a connection between the central subspace $S_{y|x}$ and the inverse regression subspace $S_{\mathrm{E}(x|y)}$ of an arbitrary regression problem. To do so, we start by introducing Proposition 4.1 below.

Proposition 4.1 (Cook (2009)). *Let x be a $p\times 1$ random vector with $\mathrm{E}(x) = 0$ and positive definite covariance matrix Σ. Let $\Phi \in \mathbb{R}^{p\times q}$, where $q \leq p$, be an arbitrary full-rank matrix. Assume that $\mathrm{E}(x|\Phi^T x = u)$ is linear function of u: $\mathrm{E}(x|\Phi^T x = u) = Mu$ for some fixed matrix $M \in \mathbb{R}^{p\times q}$. Then*

- *$M = \Sigma\Phi(\Phi^T\Sigma\Phi)^{-1}$.*
- *M^T is a generalized inverse of Φ.*
- *ΦM^T is the orthogonal projection operator for $S(\Phi)$ relative to the inner product $(v_1, v_2)_\Sigma = v_1^T\Sigma v_2$.*

Proof. We prove each dot point in order.

Result One: $M = \Sigma\Phi(\Phi^T\Sigma\Phi)^{-1}$.

Because x has a positive definite covariance matrix Σ, we first derive that

$$\mathrm{Cov}(x, \Phi^T x) = \mathrm{Cov}(x, x)\Phi = \Sigma\Phi$$

and

$$\mathrm{Cov}(\Phi^T x, \Phi^T x) = \Phi^T \mathrm{Cov}(x, x)\Phi = \Phi^T \Sigma\Phi$$

Recall that for any two random variables W, U with mean 0, we have $\mathrm{Cov}(W, U) = \mathrm{E}(WU^T) - 0 = \mathrm{E}(\mathrm{E}(WU^T|U)) = \mathrm{E}(\mathrm{E}(W|U)U^T)$. By letting $U = \Phi^T x$ and $W = x$ and the fact that $\mathrm{E}(x|\Phi^T x = u) = Mu$, we obtain

$$\begin{aligned}\Sigma\Phi = \mathrm{Cov}(x, \Phi^T x) &= \mathrm{E}(\mathrm{E}(W|U)U^T)\\ &= \mathrm{E}(E(x|\Phi^T x)x^T\Phi) = M\mathrm{E}(\Phi^T x x^T \Phi) \qquad (4.3)\\ &= M\mathrm{Cov}(\Phi^T x, \Phi^T x) = M\Phi^T\Sigma\Phi.\end{aligned}$$

It directly follows that $M = \Sigma\Phi(\Phi^T\Sigma\Phi)^{-1}$, as $\Phi^T\Sigma\Phi$ is invertible.

Result two: M^T is a generalized inverse of Φ.
To prove that M^T is a generalized inverse of Φ, it is sufficient to show that $\Phi M^T\Phi = \Phi$. Using result one and the fact that both Σ and $\Phi^T\Sigma\Phi$ are symmetric, we easily derive that

$$\Phi M^T \Phi = \Phi(\Phi^T\Sigma\Phi)^{-T}\Phi^T\Sigma^T\Phi = \Phi(\Phi^T\Sigma\Phi)^{-1}(\Phi^T\Sigma\Phi) = \Phi.$$

Result three: ΦM^T is the orthogonal projection operator for $S(\Phi)$ relative to the inner product $(v_1, v_2)_\Sigma = v_1^T\Sigma v_2$.
Since Φ is at the front of the operator ΦM, ΦM is clearly an operator for the space $S(\Phi)$. We have to show that ΦM is a projection operator and is orthogonal. Because, by result two,

$$\Phi M^T\Phi M^T = (\Phi M^T\Phi)M^T = \Phi M^T,$$

ΦM^T is a projection operator. In addition, we observe that

$$(\Phi M^T x, y)_\Sigma = x^T M\Phi^T\Sigma y = x^T\Sigma\Phi(\Phi^T\Sigma\Phi)^{-1}\Phi^T\Sigma y,$$

$$(x, \Phi M^T y)_\Sigma = x^T\Sigma\Phi M^T y = x^T\Sigma\Phi(\Phi^T\Sigma\Phi)^{-1}\Phi^T\Sigma y.$$

The above two equalities imply $(\Phi M^T x, y)_\Sigma = (x, \Phi M^T y)_\Sigma$, so ΦM^T is orthogonal. Result three is proved. □

This proposition reveals how the conditional expectation $\mathrm{E}(x|\Phi^T x = u)$ is related to the orthogonal projection operator of the space spanned by columns of Φ, when the conditional expectation $E(x|\Phi^T x = u)$ is linear in u. Therefore, if a regression problem has the central subspace $S(\Phi)$ and satisfies that $E(x|\Phi^T x = u)$ is linear in u, we can use the results of Proposition 4.1 to establish a connection between $S_{y|x}$ and $S_{\mathrm{E}(x|y)}$. Based on this idea, Cook (2009), Li (1991) introduced the following proposition.

Remark 4.2. We point out that, by using this idea, we have loosened the condition that x is normally distributed to that $E(x|\Phi^T x = u)$ is linear in u. From Chapter 2, we know that this condition is satisfied when x follows an elliptically contoured distribution, but being elliptically distributed is not a necessary condition for linear conditional expectations. In addition, Carroll and Li (1992) showed that the assumption that $E(x|\Phi^T x = u)$ is linear in u is realistic for many high-dimensional data sets. It can be proved that, if Φ is a random matrix with a vague distribution, the probability that this assumption holds approaches to 1 when the dimensionality of x tends to infinity.

Proposition 4.3. *Let Φ be a basis for $S_{y|x}$, and let $\Sigma = \mathrm{Var}(x)$. Assume that $\mathrm{E}(x|\Phi^T x = u)$ is a linear function of u. Then*

$$\mathrm{E}(x|y) - \mathrm{E}(x) = P^T_{\Phi(\Sigma)}(\mathrm{E}(x|y) - \mathrm{E}(x))$$

and

$$S_{\mathrm{E}(x|y)} \subseteq S(\Sigma\Phi) = \Sigma S_{y|x}$$

where $P_{\Phi(\Sigma)}$ is the projection operator for $S_{y|x}$ relative to the inner product induced by Σ.

Proof. Since Φ is a basis for $S_{y|x}$, we have $y \perp\!\!\!\perp x|\Phi^T x$. We first relate $\mathrm{E}(x|y)$ and $\mathrm{E}(x|\Phi^T x)$. Because $y \perp\!\!\!\perp x|\Phi^T x$, we have

$$\begin{aligned} \mathrm{E}(x|y) &= \mathrm{E}_{\Phi^T x|y}\{\mathrm{E}(x|\Phi^T x, y)\}, \\ &= \mathrm{E}_{\Phi^T x|y}\{\mathrm{E}(x|\Phi^T x)\}. \end{aligned} \tag{4.4}$$

It follows

$$\mathrm{E}(x|y) - \mathrm{E}(x) = \mathrm{E}_{\Phi^T x|y}\{\mathrm{E}(x|\Phi^T x) - \mathrm{E}(x)\}. \tag{4.5}$$

Then, since $\mathrm{E}(x|\Phi^T x) - \mathrm{E}(x)$ is linear in $\Phi^T x$, by the result three of the Proposition 4.1, we obtain

$$\mathrm{E}(x|\Phi^T x) - \mathrm{E}(x) = P^T_{\Phi(\Sigma)}(x - \mathrm{E}(x)).$$

Substituting this result back into the equation (4.5), we have

$$\begin{aligned}\mathrm{E}(x|y) - \mathrm{E}(x) &= \mathrm{E}_{\Phi^T x|y}(\mathrm{E}(x|\Phi^T x) - \mathrm{E}(x)) \\ &= \mathrm{E}_{\Phi^T x|y}(P^T_{\Phi(\Sigma)}(x - \mathrm{E}(x))) \\ &= P^T_{\Phi(\Sigma)}(\mathrm{E}(x|y) - \mathrm{E}(x)),\end{aligned} \tag{4.6}$$

and conclusions follow. □

By a similar argument of Proposition 3.6, we can also adapt above results to standardized covariates.

Corollary 4.4. *Under the same assumptions of Proposition 4.3, we have*

$$S_{\mathrm{E}(z|y)} \subseteq S(\Sigma^{1/2}\Phi) = \Sigma^{-1/2} S_{y|x} = S_{y|z}, \tag{4.7}$$

where $z = \Sigma^{-1/2}(x - \mathrm{E}(x))$.

We see that when the covariate vector is standardised to z, the relationship between $S_{\mathrm{E}(y|z)}$ and $S_{y|z}$ is more straightforward than that between $S_{\mathrm{E}(y|x)}$ and $S_{y|x}$. $S_{\mathrm{E}(y|z)}$ is a subset of $S_{y|z}$ while $S_{\mathrm{E}(y|x)}$ is a subset of a linear transformation of $S_{y|x}$, $\Sigma S_{y|x}$. Since there is no loss of generality given that $\Sigma^{-1/2} S_{y|z} = S_{y|x}$, we can work in the scale of $z = \Sigma^{-1/2}(x - \mathrm{E}(x))$ to facilitate the discussion.

Remark 4.5. We have shown that $S_{\mathrm{E}(x|y)} \subseteq \Sigma S_{y|x}$, when $E(x|\Phi^T x = u)$ is linear in u. In most situations, $S_{\mathrm{E}(x|y)}$ is a strict proper subset of $\Sigma S_{y|x}$, but there are situations in which this may not be so. In some situations, it is possible for $S_{\mathrm{E}(x|y)}$ to contain no information about $\Sigma S_{y|x}$ by being trivial, or to contain all the information of $\Sigma S_{y|x}$ by satisfying the equality $S_{\mathrm{E}(x|y)} = \Sigma S_{y|x}$. The same statement holds true when $S_{\mathrm{E}(x|y)}$ and $\Sigma S_{y|x}$ are replaced by $S_{\mathrm{E}(z|y)}$ and $S_{y|z}$ respectively, where $z = \Sigma^{-1/2}(x - \mathrm{E}(x))$. To better illustrate this point, we give examples for each case and work in the scale of z.

Case 1: $S_{\mathrm{E}(z|y)}$ is trivial

Assume z follows a standard normal distribution. Suppose the true model is

$$y|z = (\gamma^T z)^2 + \epsilon,$$

where γ is a $p \times 1$ vector and ϵ is an independent normal error. In this case, the central subspace $S_{y|z}$ is one dimensional and spanned by the vector γ.

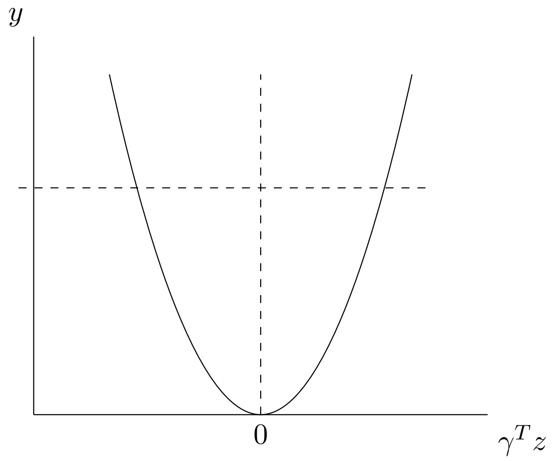

FIGURE 4.1: Stylised graph for $y|z = (\gamma^T z)^2 + \epsilon$

We also provide a stylised plot for the model. We observe from Figure 4.1 that the model has a upward parabola shape and is symmetric about $z = 0$ axis. Thus, for all values of y, we have $\mathrm{E}(\gamma^T z|y) = 0$. Moreover, since z follows a standard normal distribution, we could further deduce $\mathrm{E}(z|y) = 0$ for all values of y. Hence, the inverse regression subspace $S_{\mathrm{E}(z|y)} = \mathrm{span}\{\mathrm{E}(z|y) - \mathrm{E}(z) = 0 - 0 = 0 \,|\, y \in \mathbb{R}^+\}$ is a trivial subspace and $S_{\mathrm{E}(z|y)}$ contains no information about the central subspace. In fact, this reasoning holds for any symmetric dependence. When symmetry structure is present, the portion of $\mathrm{E}(z|y) - \mathrm{E}(z)$ contributed by the symmetry part is always 0. Since $S_{\mathrm{E}(z|y)}$ is a subspace of the first moment of the inverse regression only, $S_{\mathrm{E}(z|y)}$ reveals no information about the symmetry structure. It should be noted that even when symmetry dependence exists, $S_{\mathrm{E}(z|y)}$ is still capable of revealing other non-symmetric structures of the regression of interest and, in this case, $S_{\mathrm{E}(z|y)}$ is not necessarily trivial.

Case 2: $S_{\mathrm{E}(z|y)} = S_{y|z}$

We make the same assumption as above except the true model now is

$$y|z = c_0 + c_1(\gamma^T z) + (\gamma^T z)^2 + \sigma\epsilon$$

with $c_0, c_1 \in \mathbb{R}$. In this case, although a symmetry structure $(\gamma^T z)^2$ exists, $S_{\mathrm{E}(z|y)}$ is capable of recovering the linear part $\gamma^T z$ and $S_{\mathrm{E}(z|y)} = S(\gamma)$ (Cook, 2009). We observe that given the true model, the central subspace is $S_{y|z} = S(\gamma)$. Thus, $S_{\mathrm{E}(z|y)} = S_{y|z}$.

So far, we have successfully connected the inverse regression subspace $S_{\mathrm{E}(x|y)}$ and $S_{y|x}(\Phi)$; when $\mathrm{E}(x|\Phi^T x)$ is linear, $S_{\mathrm{E}(x|y)}$ is a subset of the transformed central subspace $\Sigma S_{y|x}$. Thus, by studying $S_{\mathrm{E}(x|y)}$, we could obtain at least a partial estimate of $S_{y|x}$, which is $\Sigma^{-1}S_{\mathrm{E}(x|y)}$. Alternatively, we can work with standardized covariate vector z. Since $S_{\mathrm{E}(z|y)} \subseteq S_{y|z}$, we first estimate $S_{y|z}$ by approximating $S_{\mathrm{E}(z|y)}$. Then we derive an estimate of $S_{y|x}$ using the linear transformation $S_{y|x} = \Sigma^{-1/2}S_{y|z}$.

Because it is more direct to work with standardised covariate vector z, we follow the second procedure and try to find an approximation to $S_{\mathrm{E}(z|y)}$ in the next section.

4.1.3 Estimating $S_{\mathrm{E}(z|y)}$

We want to find an efficient way to approximate the inverse regression subspace $S_{\mathrm{E}(z|y)}$. To start, we introduce a useful result, which is modified from Proposition 2.7 of (Eaton, 1983).

Proposition 4.6. *Suppose x is a random vector in an inner product space V with* $\mathrm{Cov}(x) = \Sigma$ *and* $\mathrm{E}(x) = \mu$. *Let $S(\Sigma)$ be the range space of Σ. Then,*

$$P\{x - \mu \in S(\Sigma)\} = 1.$$

Proof. To simplify the notation, denote $y = x - \mu$. Since y is just a horizontal shift of x, $\mathrm{Cov}(y) = \mathrm{Cov}(x) = \Sigma$. Thus, it is equivalent to show that $P\{y \in S(\Sigma)\} = 1$. If Σ is a full-rank matrix, then y has to be within the space $S(\Sigma)$ as $S(\Sigma) = V$. The interesting case is when Σ is singular.

Assume the null space of Σ is $\mathcal{N}(\Sigma)$ with dimension $k > 0$ and orthogonal basis $\{u_1, \ldots, u_k\}$. Since $\mathcal{N}(\Sigma) \oplus S(\Sigma) = V$, a vector $v \notin S(\Sigma)$ if and only if $(v, u_i) \neq 0$ for some index $i = 1, \ldots, k$. Thus,

$$\begin{aligned} P\{y \notin S(\Sigma)\} &= P\{(y, u_i) \neq 0 \text{ for some } i = 1, \ldots, k\} \\ &\leq \sum_1^k P\{(y, u_i) \neq 0\}. \end{aligned} \tag{4.8}$$

Because $\mathrm{E}(y) = 0$, (y, u_i) has mean 0. Because $u_i \in \mathcal{N}(\Sigma)$, $\mathrm{Var}\{(y, u_i)\} = (u_i, \Sigma u_i) = 0$. As a result, $P\{(y, u_i) = 0\} = 1$ for $i = 1, \ldots, k$. It follows that

$$0 \leq P\{y \notin S(\Sigma)\} \leq \sum_1^k P\{(y, u_i) \neq 0\} = 0 \tag{4.9}$$

and consequently $P\{y \in S\{\Sigma\}\} = 1$. □

We apply Proposition 4.6 to the random vector $\mathrm{E}(z|y)$. Since $\mathrm{E}(\mathrm{E}(z|y)) = 0$, we have $P\{E(z|y) \in S(\mathrm{Var}[\mathrm{E}(z|y)])\} = 1$, which implies $S_{\mathrm{E}(z|y)} \subset S\{\mathrm{Var}[\mathrm{E}(z|y)]\}$. In fact, it can also be shown that $S\{\mathrm{Var}[\mathrm{E}(z|y)]\} \subset S_{\mathrm{E}(z|y)}$. Assume the dimension of $S_{\mathrm{E}(z|y)}$ is d and let $S^{\perp}_{\mathrm{E}(z|y)}$ be an orthogonal complement of the subspace $S_{\mathrm{E}(z|y)}$ with an orthonormal basis $\{v_1, \dots, v_{p-d}\}$. It follows that $\mathrm{E}(z|y)^T v_i = 0$ for $i = 1, \dots, k$. Because $\mathrm{Var}[\mathrm{E}(z|y)] = \mathrm{E}[\mathrm{E}(z|y)\mathrm{E}(z|y)^T]$, we have $\mathrm{Var}[\mathrm{E}(z|y)]v_i = \mathrm{E}[\mathrm{E}(z|y)(\mathrm{E}(z|y)^T v_i)] = 0$. Therefore $S\{\mathrm{Var}[\mathrm{E}(z|y)]\}$ is contained in $S_{\mathrm{E}(z|y)}$ as well. Combining these results, we derive that

$$S_{\mathrm{E}(z|y)} = S\{\mathrm{Var}[\mathrm{E}(z|y)]\}. \tag{4.10}$$

The inverse regression subspace $S_{\mathrm{E}(z|y)}$ is equivalent to the range space of $\mathrm{Var}[\mathrm{E}(z|y)]$. We can thus construct an estimate of $S_{\mathrm{E}(z|y)}$ by finding an approximation to the subspace $S\{\mathrm{Var}[\mathrm{E}(z|y)]\}$.

To approximate $S\{\mathrm{Var}[E(z|y)]\}$, Li (1991) suggested replacing the response variable y with a discrete version $\tilde{y}$. We first partition the range of y into h (pre-determined) fixed, nonoverlapping slices J_s, $s = 1, \dots, h$. Then within each slice, we represent the range of y of that slice by a fixed number $\tilde{y}_s$ within the range of the slice. The vector $\tilde{y}$ consists of these fixed values $\tilde{y}_s$. Finally, we derive an estimate of $S\{\mathrm{Var}[\mathrm{E}(z|y)]\}$ by calculating the eigenvectors corresponding to the nonzero eigenvalues of $\mathrm{Var}[\mathrm{E}(z|\tilde{y})]$, which estimate the basis for $S\{\mathrm{Var}[\mathrm{E}(z|y)]\}$.

Remark 4.7. For the replacement of y by $\tilde{y}$ to be valid, we require $S_{\tilde{y}|z} \subset S_{y|z}$. This can be simply proved by a direct application of Proposition 3.2 of conditional independence. Let Φ be a basis of $S_{y|x}$. Then $\Sigma^{1/2}\Phi$ is a basis for $S_{y|z}$ and $y \perp\!\!\!\perp z|(\Sigma^{1/2}\Phi)^T z$. Since $\tilde{y}$ can be seen as a function of y, $y \perp\!\!\!\perp z|(\Sigma^{1/2}\Phi)^T z$ implies that $\tilde{y} \perp\!\!\!\perp z|(\Sigma^{1/2}\Phi)^T z$. $S_{\tilde{y}|z} \subset S_{y|z}$ clearly holds. Therefore, under the assumption that $\mathrm{E}(x|\Phi^T x = u)$ is a linear function of u,

$$S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\} = S_{\mathrm{E}(z|\tilde{y})} \subset S_{\tilde{y}|z} \subset S_{y|z}.$$

4.1.4 SIR Algorithm

So far, we have outlined the idea behind the sliced inverse regression and have carefully discussed all theoretical foundations required for this method to work. Since there is no loss of generality, we have worked in the scale of

$$z = \Sigma^{-1/2}(x - \mathrm{E}(x))$$

to facilitate the discussion.

Overall, we want to take advantage of the low dimension of the response variable y by trying to establish a relationship between the inverse regression subspace $S_{\mathrm{E}(z|y)}$ and the central subspace $S_{y|z}$. Fortunately, given $S_{y|x} = S(\Phi)$, if $\mathrm{E}(x|\Phi^T x)$ is linear, the relationship $S_{\mathrm{E}(z|y)} \subseteq S_{y|z}$ can be established. We then slice the range of y and use $\mathrm{Var}[\mathrm{E}(z|\tilde{y})]$ to obtain an estimation of $S_{\mathrm{E}(z|y)}$ to uncover information of the central subspace $S_{y|z}$. As the name sliced inverse regression suggests, this method provides information about $S_{y|x}$ via two key factors: the inverse regression space and an estimation obtained via the slicing technique. We summarise and list the step-by-step algorithm for SIR below.

SIR Algorithm

Assume we have a sample $\{(y_i, x_i), i = 1, \ldots, n\}$ and we divide the range of y into h slices so that each slice J_s contains n_s number of observations, $s = 1, \ldots, h$.

1. Standardize sample covariates. Denote the sample variance as $\hat{\Sigma}$ and the sample mean as $\bar{x}$. Compute the standardized covariate as
$$\hat{z}_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x}).$$

2. Slice the range of y in to h slices and replace each y with $\tilde{y}_s$ for $y \in J_s$. Estimate $\mathrm{E}(z|\tilde{y}_s)$, $s = 1, \ldots, h$ by
$$\bar{z}_s = \frac{\sum_{y_i \in J_s} \hat{z}_i}{n_s}.$$

3. Estimate the population matrix $\mathrm{Var}[\mathrm{E}(z|\tilde{y})] = \sum_{s=1}^{h} \Pr(y \in J_s)\mathrm{E}(z|y \in J_s)\mathrm{E}(z|y \in J_s)^T$ by the weighted sample covariance matrix
$$\hat{V} = \frac{1}{n}\sum_{s=1}^{h} n_s \bar{z}_s \bar{z}_s^T.$$

4. Perform the eigenvalue decomposition on $\hat{V}$. Denote the eigenvalues as $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ and their associated eigenvectors $\hat{l}_1, \ldots, \hat{l}_p$.

5. Let the dimension of $S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\}$ be d. Find the SIR estimate of $S_{\mathrm{E}(z|\tilde{y})}$ with
$$\hat{S}_{\mathrm{E}(z|\tilde{y})} = S(\hat{l}_1, \ldots, \hat{l}_d).$$

6. Linear transform $\hat{S}_{\mathrm{E}(z|\tilde{y})}$ by $\hat{\Sigma}^{1/2}$. The SIR estimate of the central subspace $S_{y|x}$ is

$$\hat{\Sigma}^{-1/2}\hat{S}_{\mathrm{E}(z|\tilde{y})} = S(\hat{\Sigma}^{-1/2}\hat{l}_1, \ldots, \hat{\Sigma}^{-1/2}\hat{l}_d).$$

We point out that the sample variance $\hat{V}$ converges to the population covariance matrix $\mathrm{Var}[E(z|\tilde{y})]$ at the rate of $\sqrt{n}$. Here we recall that we have used the sample covariance of x, $\hat{\Sigma}$ to standardise x. In asymptotic analyses of SIR, this case is referred to as the ignorant case. When the population covariance Σ is known, it is called the non-ignorant case. For both cases, the asymptotic behaviour of $\hat{V}$ can be derived by applying the Central Limit Theorem and the Delta method and the same convergence result will be obtained. Detailed proofs for both ignorant case and non-ignorant case can be found in Saracco (1997).

There have been many other approaches available for studying the asymptotic distribution of $\hat{V}$ and different results can be derived for specific settings. For example, by assuming $\mathrm{rank}(V) = 1$, Duan and Li (1991) used Taylor expansion of a related eigen-problem, whose solution is the largest eigenvector of $\hat{V}$, to study the asymptotic distribution of $\hat{V}$. Carroll and Li (1992) studied the asymptotic behaviour of the eigenvectors of $\hat{V}$ when the sample covariate x cannot be directly computed and surrogates of the covariate have to be introduced. Finally, Hsing and Carroll (1992) discussed the asymptotic properties of $\hat{V}$ when each slice J_s has two observations only; these results were later extended by Zhu and Ng (1995) for any fixed number of observations in each slice J_s.

Finally, it is important to note that the convergence results of $\hat{V}$ imply that the eigenvalues and eigenvectors of $\hat{V}$ converge at the rate of $\sqrt{n}$ to the eigenvalues and engenvectors of $\mathrm{Var}[\mathrm{E}(x|\tilde{y})]$ as well (Saracco, 1997).

4.1.5 A method for choosing the dimension $S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\}$

We recall that, in the SIR algorithm, we have assumed the dimension $d = \dim[S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\}]$ of the inverse regression space is known. Theoretically, if the covariance matrix $Var[E(x|\tilde{y})]$ is known, d is simply the number of its non-zero eigenvalues and the sum of the smallest $p - d$ eigenvalues is zero. However, in practice, we need to determine the value of d.

A widely used method is proposed by Bura and Cook (2001). Bura and Cook suggested that we should choose d by studying the asymptotic behaviour of the statistic:

$$\hat{\Delta}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j, \tag{4.11}$$

where $\hat{\lambda}_j$ are eigenvalues of $\hat{V}$. Once we know the asymptotic distribution of the statistic $\hat{\Delta}$, we can determine d by a series of hypotheses tests. Let m be an integer. We start by assuming $m = 0$ and then compute the statistic $\hat{\Delta}_m$ to test the hypotheses H_0: $d = m$ against H_1: $d > m$, using its asymptotic distribution. If the test concludes $d > m$, we increase m by 1 and repeat the test until either we accept H_0: $d = m$ or H_1: $d > m$ when $m = p - 1$ (In this case, we conclude $d = p$, as the possible maximum value of d is p).

Remark 4.8. Bura and Cook (2001)'s method was developed on Li (1991)'s original dimension test. Bura and Cook used the same test statistic as Li. However, because Li's dimension test requires normally distributed covariates, Busa and Cook extended Li's test for general situations. There are other methods for choosing the value of d. For example, by investigating the eigenvectors of $\hat{V}$, Schott (1994) proposed a test for choosing d under the assumption that the covariates are elliptically distributed. Under the same assumption, we can also choose d using permutation procures developed by Cook and Yin (2001), which can be computationally expensive.

We now investigate the asymptotic distribution of $\hat{\Delta}_d$. Since $\hat{\Delta}_d$ is the sum of the smallest $\min(p-d, h-d)$ eigenvalues of the covariance matrix $\hat{V}$, one possible approach to study $\hat{\Delta}_d$ is through the singular values of the Cholesky decomposition of $\hat{V}$, as the square of these singular values are the positive eigenvalues of $\hat{V}$. Denote the matrices

$$\hat{Z} = (\sqrt{\frac{n_1}{n}}\bar{z}_1, \ldots, \sqrt{\frac{n_h}{n}}\bar{z}_h)$$

and

$$Z = (\sqrt{\Pr(y \in J_1)}\mathrm{E}(z|y \in J_1), \ldots, \sqrt{\Pr(y \in J_h)}\mathrm{E}(z|y \in J_s)).$$

We have $\hat{V} = \hat{Z}\hat{Z}^T$. We first need to characterize the asymptotic distribution of the singular values of $\hat{Z}$. To do so, we borrow the general asymptotic result for singular values from Eaton and Tyler (1994). Assume the singular value decomposition of Z gives

$$\begin{aligned} Z &= U^T \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} V \\ &= \begin{pmatrix} U_I & U_0 \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_I & V_0 \end{pmatrix}^T \end{aligned} \tag{4.12}$$

where $U \in \mathbb{R}^{p\times p}$, $V \in \mathbb{R}^{h\times h}$ are orthonormal matrices with $U_0 \in \mathbb{R}^{p\times(p-d)}$, $V_0 \in \mathbb{R}^{h\times(h-d)}$ and $D \in \mathbb{R}^{d\times d}$ is a diagonal matrix of singular values. Eaton and Tyler (1994) showed that the asymptotic distribution of the smallest $\min(p-d, h-d)$ singular values of $\sqrt{n}(\hat{Z} - Z)$ is the same as the asymptotic distribution of the singular values of the $(p-d) \times (h-d)$ matrix

$$\sqrt{n}U_0^T(\hat{Z} - Z)V_0 = \sqrt{n}U_0^T\hat{Z}V_0. \tag{4.13}$$

This means the asymptotic distribution of the smallest $\min(p-d, h-d)$ singular values of $\sqrt{n}\hat{Z}$ is the same as the asymptotic distribution of $\sqrt{n}U_0^T\hat{Z}V_0$. As a result, we can focus on studying the asymptotic behaviour of $\sqrt{n}\text{vec}(U_0^T\hat{Z}V_0)$ and then derive the asymptotic distribution of the statistic $\hat{\Delta}_d$, which is equal to the asymptotic distribution of

$$\Delta_d = n\text{tr}[U_0^T\hat{Z}V_0(U_0^T\hat{Z}V_0)^T] = n\text{vec}(U_0^T\hat{Z}V_0)^T\text{vec}(U_0^T\hat{Z}V_0).$$

Remark 4.9. The above analysis was based on the assumption that h is large enough so that $d < \min(p, h-1)$ (Cook, 2009). Because $\text{E}(\text{E}(z|\tilde{y})) = 0$, we note that there is at least one linear dependency among columns of Z. Therefore, Z has possible maximum rank $\min(p, h-1)$. Consequently, the possible maximum rank of $\text{Var}[\text{E}(z|\tilde{y})] = ZZ^T$ is $\min(p, h-1)$. Since we are testing hypotheses about d, the number of non-zero eigenvalues of $\text{Var}[\text{E}(z|\tilde{y})]$, we require $d < \min(p, h-1)$ for our approach to be feasible. The constraint $d < \min(p, h-1)$ is always satisfied if we choose $h > p + 1$.

Proposition 4.10 (Cook (2009))**.** *Let U_0, $\hat{Z}$, V_0 be defined as above. Then*

$$\sqrt{n}vec(U_0^T\hat{Z}V_0) \longrightarrow_d N(0, \Sigma_Z). \tag{4.14}$$

Here,

$$\Sigma_Z = (V_0^TQ \otimes I_{p-d})\Upsilon_0(QV_0 \otimes I_{p-d}), \tag{4.15}$$

where Q is the orthogonal projection for $S^{\perp}((\sqrt{\Pr(y \in J_1)}, \ldots, \sqrt{\Pr(y \in J_h)})^T)$ and Υ_0 is a $(p-d)h \times (p-d)h$ block diagonal matrix with diagonal blocks $U_0^T \mathrm{Var}(z|\tilde{y}_s) U_0$, $s = 1, \ldots, h$.

Proof. To start, let us define

$$M_n \coloneqq (\bar{x}_1, \ldots, \bar{x}_h) \qquad \in \mathbb{R}^{p \times h}$$

$$C \coloneqq (\mathrm{E}(x|y \in J_1), \ldots, \mathrm{E}(x|y \in J_h)) \qquad \in \mathbb{R}^{p \times h}$$

The proof can be broken into three steps. During the first step, we follow Cook's idea and find an approximation to $\sqrt{n} U_0^T \hat{Z} V_0$. The approximation should be a function of the matrix $M_n - C$ and should allow us to transform the problem from studying the asymptotic distribution of $\sqrt{n} U_0^T \hat{Z} V_0$ to studying the asymptotic behaviour of its approximation. The reason that we want the approximation to be a function of $M_n - C$ is that we can apply the Central Limit theorem to find the asymptotic distribution of $M_n - C$. Finally, we apply the Delta method to derive the desired distribution.

Step One: Approximation to $\sqrt{n} U_0^T \hat{Z} V_0$
In order to find a function of the matrix $M_n - C$ that approximates $\sqrt{n}\mathrm{vec}(U_0^T \hat{Z} V_0)$, we try to find an equivalent expression for $\sqrt{n}\mathrm{vec}(U_0^T \hat{Z} V_0)$ that incorporates $M_n - C$.

Let

$$1_h \coloneqq (1, \ldots, 1)^T \qquad \in \mathbb{R}^{h \times 1},$$

$$\hat{\rho} \coloneqq (\frac{n_1}{n}, \ldots, \frac{n_h}{n})^T \qquad \in \mathbb{R}^{h \times 1},$$

and

$$\rho \coloneqq (\Pr(y \in J_1), \ldots, \Pr(y \in J_h))^T \qquad \in \mathbb{R}^{h \times 1}.$$

Also, assume that $\bar{x}$, $\bar{x}_s$ and $\hat{\Sigma}$ are the sample estimates of $\mu = \mathrm{E}(x)$, $\mu_{x|s} = \mathrm{E}(x|y \in J_s)$ and $\Sigma = \mathrm{Var}(x)$ respectively.

We observe that

$$(\bar{x}_1 - \bar{x}, \ldots, \bar{x}_h - \bar{x}) = M_n(I_h - \hat{\rho} 1_h^T)$$

and

$$(\mu_{x|1} - \mu, \ldots, \mu_{x|h} - \mu) = C(I_h - \rho 1_h^T).$$

It follows that

$$\hat{Z} = \hat{\Sigma}^{-1/2} M_n (I_h - \hat{\rho} 1_h^T) \hat{G}$$

and

$$Z = \Sigma^{-1/2} C (I_h - \hat{\rho} 1_h^T) G,$$

where $G \in \mathbb{R}^{h\times h}$ is a diagonal matrix with diagonal elements $\sqrt{\Pr(y \in J_s)}$ and $\hat{G} \in \mathbb{R}^{h\times h}$ is also a diagonal matrix with diagonal entries $\sqrt{n_s/n}$. Also, we note that

$$G^{-1}(I_h - \rho 1_h^T) G = I_h - \sqrt{\rho}\sqrt{\rho}^T = Q,$$

which implies $GQ = (I_h - \rho 1_h^T)G$. Here, $\sqrt{\rho} \coloneqq (\sqrt{\Pr(y \in J_1)}, \ldots, \sqrt{\Pr(y \in J_h)})^T \in \mathbb{R}^{h\times 1}$.

Denote $\hat{A} = \hat{\Sigma}^{-1/2}\Sigma^{1/2}$, $F = GQ$ and $\hat{F} = \hat{G} Q_{\sqrt{\hat{\rho}}}$. $Q_{\sqrt{\hat{\rho}}}$ is the orthogonal projection for $S^{\perp}(\sqrt{\hat{\rho}})$. It follows that

$$\hat{F} = \hat{G} Q_{\sqrt{\hat{\rho}}} = (I_h - \hat{\rho} 1_h^T)\hat{G}. \tag{4.16}$$

In addition, we can express $\sqrt{n} U_0^T \hat{Z} V_0$ as

$$\sqrt{n} U_0^T \hat{Z} V_0 = \sqrt{n} U_0^T (\hat{A} - I_p + I_p) \Sigma^{-1/2} (M_n - C + C)(\hat{F} - F + F) V_0. \tag{4.17}$$

We now expand the equation in terms $(\hat{A} - I_p)$, $(M_n - C)$, and $(\hat{F} - F)$ with the error term $o_p(n^{-1/2})$, which gives:

$$\begin{aligned} \sqrt{n} U_0^T \hat{Z} V_0 =& \sqrt{n} U_0^T (\hat{A} - I_p) C F V_0 \\ &+ \sqrt{n} U_0^T \Sigma^{-1/2} (M_n - C) F V_0 \\ &+ \sqrt{n} U_0^T \Sigma^{-1/2} C (\hat{F} - F) V_0 \\ &+ \sqrt{n} U_0^T \Sigma^{-1/2} C F V_0 + o_p(n^{-1/2}). \end{aligned} \tag{4.18}$$

Since $Z = \Sigma^{-1/2} C F$ and $Z V_0 = 0_{p\times(h-d)}$, the first and the fourth terms are zero. For the third term, we notice that

$$\begin{aligned} \sqrt{n} U_0^T \Sigma^{-1/2} C (\hat{F} - F) V_0 &= \sqrt{n} U_0^T \Sigma^{-1/2} C \hat{F} V_0 - \sqrt{n} U_0^T \Sigma^{-1/2} C F V_0 \\ &= \sqrt{n} U_0^T \Sigma^{-1/2} C \hat{F} V_0 \\ &= \sqrt{n} (C^T \Sigma^{-1/2} U_0)^T (I_h - \hat{\rho} 1_h^T) \hat{G} V_0. \end{aligned} \tag{4.19}$$

The last equality uses the equation (4.16). Because G is invertible, $Z = \Sigma^{-1/2} C (I_h - \rho 1_h^T) G$ and $U_0^T Z = 0_{(p-d)\times h}$, we have $(C^T \Sigma^{-1/2} U_0)^T (I_h - \rho 1_h^T) = 0_{(p-d)\times h}$. Also, we know that if $v^T (I_h - \rho 1_h^T) = 0_{1\times h}$, where $v \in \mathbb{R}^{h\times 1}$, v has to be in the space $S(1_h)$. It follows that $C^T \Sigma^{-1/2} U_0 \in S(1_h)$.

Now, direct algebra shows that $1_h^T(I_{h\times h} - \hat{\rho}1_h^T) = 0_{1\times h}$. Using the fact that $C^T\Sigma^{-1/2}U_0 \in S(1_h)$, we conclude $(C^T\Sigma^{-1/2}U_0)^T(I_{h\times h} - \hat{\rho}1_h^T) = 0$. As a result, the equation (4.19) is zero.

Combining above results, we have derived that

$$\begin{aligned}\sqrt{n}U_0^T\hat{Z}V_0 &= \sqrt{n}U_0^T\Sigma^{-1/2}(M_n - C)FV_0 + o_p(n^{-1/2}) \\ &= \sqrt{n}U_0^T\Sigma^{-1/2}(M_n - C)GQV_0 + o_p(n^{-1/2}).\end{aligned} \tag{4.20}$$

Step 2: the Central Limit theorem

Direct application of the Central limit theorem to $\sqrt{n}\text{vec}(M_n - C)$ gives that

$$\sqrt{n}\text{vec}(M_n - C) \longrightarrow_d N(0, (G^{-1}\otimes I_p)\Upsilon_x(G^{-1}\otimes I_p)), \tag{4.21}$$

where Υ_x is a block diagonal matrix with diagonal blocks $\text{Var}(x|\tilde{y}_s)$, $s = 1, \ldots, h$.

Step 3: the Delta method

Finally, we introduce the function:

$$\begin{aligned}f_1 :&\mathbb{R}^{ph\times 1} \mapsto \mathbb{R}^{(p-d)(h-d)\times 1} \\ &\text{vec}(X) \mapsto \text{vec}(U_0^T\Sigma^{-1/2}XGQV_0).\end{aligned}$$

Applying the Delta method gives us the desire result:

$$\sqrt{n}\text{vec}(U_0^T\hat{Z}V_0) \longrightarrow_d N(0, (V_0^TQ\otimes I_{p-d})\Upsilon_0(QV_0\otimes I_{p-d})),$$

with Υ_0 being a $(p-d)h\times(p-d)h$ block diagonal matrix with diagonal blocks $U_0^T\text{Var}(z|\tilde{y}_s)U_0$, $s = 1, \ldots, h$. □

Given the above proposition, we know that $\sqrt{n}\text{vec}(U_0^T\hat{Z}V_0)$ converges to a normal distribution asymptotically. Since $\hat{\Delta}_d$ is the square of $\sqrt{n}\text{vec}(U_0^T\hat{Z}V_0)$, we conclude that $\hat{\Delta}_d$ is distributed as a linear combination of independent chi-square random variables. Formally, we summarise the result in Proposition 4.11 below.

Proposition 4.11 (Cook (2009))**.** *Let $d = dim(S_{\text{E}(z|y)})$, where $d < h-1$ and $d < p$. Also, define the statistic $\hat{\Delta}_d$ and Δ_d as above. Then the asymptotic distribution of Δ_d, as well* *as $\hat{\Delta}_d$, is the same as the distribution of*

$$C = \sum_{k=1}^{(p-d)(h-d)} \omega_k \chi^2(1), \tag{4.22}$$

where the $\chi^2(1)$ are independent chi-square random variables with one degree of freedom, and $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_{(p-d)(h-d)}$ are the eigenvalues of the covariance matrix Σ_Z defined in Proposition 4.10.

Proof. Based on the result of Proposition 4.10, this proposition directly follows from p.112 of Eaton (1983). □

Since both Δ_d and $\hat{\Delta}_d$ converge in distribution to C in equation (4.22), we are now able to determine d through the sample estimate $\hat{\Sigma}_Z$ of Σ_z. In other words, the asymptotic distribution of $\hat{\Delta}_d$ is approximated by

$$\hat{C} = \sum_{k=1}^{(p-d)(h-d)} \hat{\omega}_k \chi^2(1), \tag{4.23}$$

where $\{\hat{\omega}_1, \ldots, \hat{\omega}_{(p-d)(h-d)}\}$ are eigenvalues of $\hat{\Sigma}_Z$, computed using sample versions of the various quantities required to compute Σ_Z.

To conclude, we outline the algorithm for determining d. This algorithm is also referred to as the marginal dimension test or dimension test (Weisberg, 2015; Cook, 2004).

An algorithm for choosing the dimension of $S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\}$

1. Compute the singular value decomposition of $\hat{Z}$ to estimate U and V by their sample versions. In addition, compute the sample version of $\mathrm{Var}(z|\tilde{y}_s)$, $s = 1, \ldots, h$.

2. Set $m = 0$.

3. Set $d = m$. Use formula (4.15) and the sample estimates of U_0, V_0 and $\mathrm{Var}(z|\tilde{y}_s)$, $s = 1, \ldots, h$ to calculate $\hat{\Sigma}_Z$.

4. Compute the eigenvalues of $\hat{\Sigma}_Z$ and denote them as $\hat{\omega}_1 \geq \cdots \geq \hat{\omega}_{(p-d)(p-d)}$.

5. Calculate $\hat{\Delta}_d$ using equation (4.11). Then compute the p-value as $Pr(\hat{C} > \hat{\Delta}_\kappa)$, where
$$\hat{C} = \sum_{k=1}^{(p-d)(h-d)} \hat{\omega}_k \chi^2(1).$$

6. Compare the calculated p-value with the pre-determined cutoff value. If the p-value is larger than the pre-determined cutoff value, then $d = m$ is the final estimate. If not, proceed as if $d > m$ holds. Let $m = m + 1$ and return to step three.

4.1.6 Comments on SIR

To close our discussions of SIR, we give some further comments on SIR.

4.1.6.1 Comment One: e.d.r. directions

Li first introduced the slice inverse regression method in 1991. In his paper, Li worked with the formulation
$$y = f(\beta_1^T x, \ldots, \beta_k^T x, \epsilon), \tag{4.24}$$
where $\epsilon \perp\!\!\!\perp x$, the β's are an unknown vectors, and f is an unknown arbitrary function on $\mathbb{R}^{k+1}$. Li called the vectors $\beta_1, \ldots, \beta_k$ *effective dimension reduction directions* (e.d.r. directions) and, correspondingly, the space spanned by these vectors an *effective dimension reduction subspace*. In our discussion above, we adopted Cook's idea instead, which uses the formulation
$$y \perp\!\!\!\perp x | \Phi^T x. \tag{4.25}$$

It is mainly because Li did not address the issues about existence and uniqueness of the effective dimension reduction subspace. Hence, careless use of effective dimension reduction subspaces could lead to misleading conclusions. On the other hand, the existence and uniqueness conditions for the dimension reduction space based on (4.25) have been established by Cook. However, it should be noted that when the central subspace exists, the models (4.24) and (4.25) are technically equivalent. We can connect them by requiring that the central subspace $S_{y|x}$ is spanned by $\{\beta_1, \ldots, \beta_k\}$ or, equally, columns of Φ.

4.1.6.2 Comment Two: slices

The choice of the number of slices h is often treated as a less critical issue in the analysis of SIR. Although h may affect the asymptotic variance of the output estimate, the difference is often considered as unimportant in practice (Li, 1991). This is probably the reason that there is no method available to select an optimal h and choose an optimal bandwidth in the literature (Ma and Zhu, 2013). However, different opinions have been voiced recently. For instance, Becker and Gather (2007) showed through simulations that when h is much larger than $0.1n$, SIR results will be strongly influenced by the choice of h. Therefore, further investigations on the choice of h might be worthwhile.

Despite there being no general rules for choosing h, some caution should be taken when deciding on a value for h. Firstly, as we discussed in remark 4.9, h should be large enough to satisfy $\min(p, h-1) > d$ (Cook, 2009). Generally, we should choose h to be sufficiently large to avoid any loss of population structure after the replacement. Secondly, in terms of the range of each slice, it is often preferred to allow it to vary so that the number of observations within each slice is as similar as possible (Li, 1991). Finally, in the situation where each slice contains a fixed number L observations, Li (2000) mentioned that if the estimated eigenvalue is smaller than $\frac{1}{L}$, then the true eigenvalue is probably zero.

4.1.6.3 Comment Three: limitations of SIR

There two main limitations to SIR: the requirement of linear conditional expectation and the failure of SIR under symmetry dependence. In terms of linear conditional expectation, we require $E(x|\Phi^T x)$ to be linear in $\Phi^T x$ given that $S_{y|x} = S(\Phi)$. We have pointed out that this assumption is realistic in many high dimensional data problems. Still, we should always check whether linear conditional expectation is met before applying SIR, as serious violation of the assumption will lead to wrong results. Because we do not have information about Φ beforehand (we want to use SIR to derive Φ), a stronger condition is tested in practice. That is, whether $E(x|B^T x)$ is linear in $B^T x$ for any arbitrary matrix B. Or equivalently, whether x is elliptically distributed.

On the other hand, we recall that the failure of SIR under symmetry dependence is mainly caused by the fact that SIR uses first moment only to recover the relationship between the covariates and the response variable. To tackle this issue, methods using higher moments have been developed. We will introduce one of such methods below.

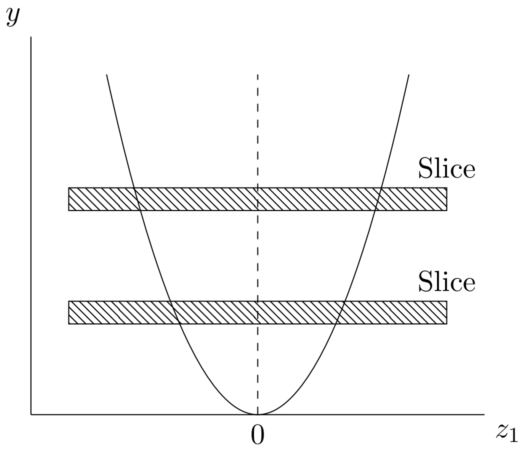

FIGURE 4.2: Stylised graph of $y|z = ((1, 0, \dots, 0)z)^2 + \epsilon$

We see that the average of z_1 within each slice is 0, but the variance of z_1 changes over slices. It follows that $\mathrm{E}(z|\tilde{y}_s) = 0$ for each arbitrary slice J_s and $\mathrm{Var}(z|\tilde{y}_s)$ is different for different slices.

4.2 SAVE

We introduce Sliced Average Variance Estimation(SAVE) in this section. SAVE, proposed by Cook and Weisberg (1991), was specifically designed to overcome the inability of SIR to detect symmetry dependence. The idea behind SAVE was that, although $\mathrm{E}(z|\tilde{y}_s) = 0$ for each slice J_s, the variance $\mathrm{Var}(z|\tilde{y}_s)$ does change from slice to slice (For example, see Figure 4.2). Therefore, SAVE extracts information about the central subspace that is missed by SIR by using the second moment as well as the first moment.

In order to understand how the information about the central subspace $S_{y|x}$ is contained in the second moment, we first assume that $S_{y|x} = S(\Phi)$ so that $y \perp\!\!\!\perp x|\Phi^T x$. We also assume that x follows an elliptically contoured distribution to simplify the discussion. Since x is elliptically distributed, a direct application of Corollary 2.18 shows that

$$\mathrm{E}(x|\Phi^T x) = \mu + \Sigma\Phi(\Phi^T\Sigma\Phi)^{-1}\Phi^T(x - \mu), \tag{4.26}$$

$$\mathrm{Var}(x|\Phi^T x) = w(\Phi^T x)[\Sigma - \Sigma\Phi(\Phi^T\Sigma\Phi)^{-1}\Phi^T\Sigma] \tag{4.27}$$

where $\mu = \mathrm{E}(x)$, $\Sigma = \mathrm{Var}(x)$ and $w(\Phi^T x)$ is function about $\Phi^T x$ through the quadratic form $(x - \mu)^T\Phi[\mathrm{Var}(\Phi^T x)]^{-1}\Phi^T(x - \mu)$. Since we have shown that standardising x and working on the z-scale involves no loss of generality, we standardise x to be consistent with our analysis of SIR method. Also, because we have shown that $x \perp\!\!\!\perp y|\Phi^T x$ is equivalent to $z \perp\!\!\!\perp y|\Psi^T z$ where $\Psi = \Sigma^{1/2}\Phi$, it follows from equations (4.26), (4.27) that

$$\mathrm{E}(z|\Psi^T z) = \Psi(\Psi^T\Psi)^{-1}\Psi^T z, \tag{4.28}$$

$$\mathrm{Var}(z|\Psi^T z) = w(\Psi^T z)[I - \Psi(\Psi^T\Psi)^{-1}\Psi^T]. \tag{4.29}$$

We observe that $\Psi(\Psi^T\Psi)^{-1}\Psi^T$ by its form is an orthogonal projection operator onto the space $S(\Psi)$ with the inner product $(x, y) = x^T y$. By letting $P_\Psi = \Psi(\Psi^T\Psi)^{-1}\Psi^T$ and $Q_\Psi = I - P_\Psi$, we can equally write

$$\mathrm{E}(z|\Psi^T z) = P_\Psi z \tag{4.30}$$

$$\mathrm{Var}(z|\Psi^T z) = w(\Psi^T z)Q_\Psi. \tag{4.31}$$

With these results, we are now able to derive an alternative formula for $\mathrm{Var}(z|y)$ via the law of total variance:

$$\begin{aligned}\mathrm{Var}(z|y) &= \mathrm{E}[\mathrm{Var}(z|\Psi^T)|y] + \mathrm{Var}[\mathrm{E}(z|\Psi^T z)|y]\\ &= \mathrm{E}[w(\Psi^T z)Q_\Psi|y] + \mathrm{Var}(P_\Psi z|y)\\ &= \mathrm{E}[w(\Psi^T z)|y]Q_\Psi + P_\Psi\mathrm{Var}(z|y)P_\Psi\\ &= w_y Q_\Psi + P_\Psi\mathrm{Var}(z|y)P_\Psi,\end{aligned} \tag{4.32}$$

where $w_y \coloneqq E[w(\Psi^T z)|y]$ is a function of y.

We make some important comments on the above equation. Firstly, we note that w_y is a scalar function. Therefore, assuming Ψ has rank d, w_y is an eigenvalue of $\mathrm{Var}(z|y)$ with multiplicity $p-d$ and its associated eigenvectors span the space $S(Q_\Psi)$. The remaining eigenvectors of $\mathrm{Var}(z|y)$ span the central subspace $S_{y|z} = S(\Psi)$. Secondly and more importantly, by rearranging, we observe that

$$w_y I_p - \mathrm{Var}(z|y) = w_y P_\Psi - P_\Psi\mathrm{Var}(z|y)P_\Psi = P_\Psi[w_y I_p - \mathrm{Var}(z|y)]P_\Psi. \tag{4.33}$$

The eigenvectors of $w_y I_p - \mathrm{Var}(z|y)$ are in the space $S(\Psi)$. Thus, if we can estimate $w_y I_p - \mathrm{Var}(z|y)$ and find its eigenvectors that correspond to nonzero eigenvalues, we can estimate $S(\Psi)$ by the space spanned by these eigenvectors.

In order to estimate $w_y I_p - \mathrm{Var}(z|y)$, we recall from Chapter 2 (remark 2.19) that $w(\Psi^T z)$ is a constant function if and only if z is normally distributed. Thus, by assuming x

follows a normal distribution, we make z normally distributed and $w(\Psi^T z)$ a constant. Consequently, w_y is a constant. In fact, we can further conclude that $w_y = 1$ for all y when x has a normal distribution (Cook and Weisberg, 1991). In this case, we do not need to worry about the value of w_y any more. We directly estimate $I_p - \mathrm{Var}(z|y)$ and then find its eigenvectors. In general, to avoid negative eigenvalues, we calculate the eigenvectors via $[I_p - \mathrm{Var}(z|y)]^2$ instead. Estimating $[I_p - \mathrm{Var}(z|y)]^2$ can be challenging due to its relatively complicated form. To deal with this, we adopt the same approach Li used in developing SIR. We slice the range of y into h fixed slices $J_1, \ldots, J_h$ with $n_1, \ldots, n_h$ elements and approximate $[I_p - \mathrm{Var}(z|y)]^2$ by

$$\hat{\Sigma}_{save} = \sum_{s=1}^{h} \frac{n_s}{n} (I_p - \widehat{\mathrm{Var}}(z|y \in J_s))^2, \tag{4.34}$$

the sample version of the population quantity $\Sigma_{save} = \sum_{s=1}^{h} \mathrm{Pr}(y \in J_s)(I_p - \mathrm{Var}(z|y \in J_s))^2$.

So far, we have outlined the key ideas of SAVE. During our discussion, we have required the assumption that x is normally distributed. In fact, this condition can be loosened, as shown in (Cook and Lee, 1999). For the above reasoning to hold and hence for $S(\Sigma_{save}) \subseteq S_{y|z}$, it is sufficient to require the following two conditions:

Condition 1: The conditional expectation $\mathrm{E}(x|\Phi^T x)$ is a linear function of $\Phi^T x$.
Condition 2: The matrix $\mathrm{Var}(x|\Phi^T x)$ is constant.

The first condition is automatically satisfied when x follows an elliptically contoured distribution and when x is normally distributed, the second condition is automatically satisfied ($w(\Phi^T x)$ is constant in equation 4.27).

In summary, SAVE was developed using the similar methodology as SIR. SAVE is also an extension of SIR; SIR only relies on the first moment but SAVE employs the second moment as well. However, it should be noted that SIR has wider applicability than SAVE, as SIR only requires the conditional expectation $E(x|\Phi^T x)$ to be linear in $\Phi^T x$ while, besides the linear conditional expectation, SAVE also requires the matrix $\mathrm{Var}(x|\Phi^T x)$ to be constant. In terms of algorithms, due to the similar ideas adopted by SIR and SAVE, the algorithm of SAVE is exactly the same as that of SIR except that we need to replace $\hat{V}$ with the new matrix $\hat{\Sigma}_{save}$ and let d be the dimension of $S(\Sigma_{save})$ instead.

4.2.1 A method for choosing the dimension of $S(\Sigma_{save})$

To apply SAVE in practice, we need a method for choosing the dimension d of $S(\Sigma_{save})$. Although SAVE has been considered a useful complement to SIR, the development of suitable tests for d has lagged. It is technically difficult to find the asymptotic distribution of the eigenvalues of a quadratic function of the variance, but progress has been made on asymptotic analysis.

Following the idea behind the test proposed by Li for SIR, Cook and Ni successfully derived the asymptotic distribution of a similar test statistic for $\hat{\Sigma}_{save}$ in 2005. Denote the eigenvalues of $\hat{\Sigma}_{save}$ by $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ with $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_p$. For the hypothesis $\dim(S(\Sigma_{save})) = m$, Cook and Ni suggested the SAVE test statistic

$$\Delta_{save} = n \sum_{i=m+1}^{p} \hat{\lambda}_i.$$

When n goes to infinity, Δ_{save} approaches to a weighted linear combination of p^2h independent chi-square random variables with one degree of freedom. The weights are computed as the eigenvalues of a symmetric matrix of size $p^2h \times p^2h$. We note that when using this test statistic, for a moderate number of slices h and dimension p, it is computationally expensive to compute all the weights and we need a large sample for the test to be reliable. For instance, if $p = 10$ and $h = 20$, we need to find the eigenvalues of a matrix of order 2000×2000.

Due to this drawback, we will choose the dimension of Σ_{save} using a computationally feasible test for d, developed by Shao et al. (2007). Instead of working with eigenvalues, Shao et al. (2007) proposed a different test statistic using a set of eigenvectors of $\hat{\Sigma}_{save}$. Again, suppose the hypothesis is $\dim(S(\Sigma_{save})) = m$. Let Θ, $\hat{\Theta} \in \mathbb{R}^{p\times(p-m)}$ be matrices with columns being orthnormalized eigenvectors that correspond to the smallest $(p-m)$ eigenvalues of Σ_{save} and $\hat{\Sigma}_{save}$ respectively. Also, define the population quantity $A_s = \Pr(y \in J_s)^{1/2}(I_p - \mathrm{Var}(z|y \in J_s))$ and its sample estimator $\hat{A}_s = (\frac{n_s}{n})^{1/2}(I_p - \widehat{\mathrm{Var}}(z|y \in J_s))$, so $\Sigma_{save} = \sum_{s=1}^{h} A_s^2$ and $\hat{\Sigma}_{save} = \sum_{s=1}^{h} \hat{A}_s^2$. Shao, Cook and Weisberg uses the following test statistic

$$T_m(\hat{\Theta}) = \frac{n}{2} \sum_{s=1}^{h} \mathrm{tr}\{(\hat{\Theta}^T \hat{A}_s \hat{\Theta})^2\}. \tag{4.35}$$

This test statistic was first proposed by Cook (2004) for SIR and then extended to SAVE by Shao, Cook and Weisberg. For more information and the intuition behind this test statistic, see (Cook, 2004).

Before we study the asymptotic distribution of $T(\hat{\Theta})$, we introduce one more condition that is required for the following asymptotic analysis to hold.

Condition three: For any non-zero $\beta \in S_{y|x}$ and all J_s, either $\mathrm{Var}\{\mathrm{E}(\beta^T x|y \in J_s)\} > 0$ or $\mathrm{Var}\{\mathrm{Var}(\beta^T x|y \in J_s)\} > 0$ holds.

Condition three is also referred as the coverage condition. Recall that, in the above discussion, we have shown that $S(\Sigma_{save}) \subseteq S_{y|z}$ under condition one and two. However, if condition three is also satisfied, $S(\Sigma_{save}) = S_{y|z}$. This equality can be proved by contradiction. Assume $S(\Sigma_{save})$ is a strict subset of $S_{y|z}$. Then there exists a $\beta \neq 0$ and $\beta \in S_{y|z}$ such that $\beta \in S(\Sigma_{save})^{\perp}$. It follows that $(I_p - \mathrm{Var}(z|y \in J_s))\beta = 0$ for all J_s. Thus, $\mathrm{Var}(\beta^T z|y \in J_s) = \beta^T \mathrm{Var}(z|y \in J_s)\beta = \beta^T\beta$. Consequently, $\mathrm{Var}\{\mathrm{Var}(\beta^T z|y \in J_s)\} = 0$ and

$$\mathrm{Var}\{\mathrm{E}(\beta^T z|y \in J_s)\} = \mathrm{Var}(\beta^T z) - \mathrm{E}\{\mathrm{Var}(\beta^T z|y \in J_s)\} = \beta^T\beta - \beta^T\beta = 0.$$

Since these results contradict condition three, we have $S(\Sigma_{save}) = S_{y|z}$.

Theorem 4.12 (Shao et al. (2007)). *Assume Conditions 1-3 hold and* $\mathrm{Var}(\Theta^T z \otimes \Theta^T z|\Psi^T z)$ *is constant. Then, under the hypothesis $d = m$, when n goes to infinity,*

$$2T_m(\hat{\Theta}) \longrightarrow_d \sum_i \omega_i \chi_i^2(h-1), \tag{4.36}$$

where ω_i, $i = 1, \ldots, (p-m)(p-m+1)/2$ are the largest $(p-m)(p-m+1)/2$ eigenvalues of $\mathrm{Var}(\Theta^T z \otimes \Theta^T z)$ *and $\chi_i^2(h-1)$ are independent χ^2 random variables with $h-1$ degrees of freedom.*

If, in addition, x is normally distributed, then

$$T_m(\hat{\Theta}) \longrightarrow_d \chi^2\{(h-1)(p-m)(p-m+1)/2\}, \tag{4.37}$$

where $\chi^2\{(h-1)(p-m)(p-m+1)/2\}$ is a χ^2 random variable with $(h-1)(p-m)(p-m+1)/2$ degrees of freedom.

Proof. Here, we only provide an outline of the proof due to its length. For the detailed proof, see Shao et al. (2007).

Let P_Θ be the projection operator for the space $S(\Theta)$. To start, we apply Lemma 2.1 of Tyler (1981) to show that $P_{\hat{\Theta}} = P_\Theta + O(n^{-1/2})$. Substituting this result into $P_{\hat{\Theta}} \hat{A}_s P_{\hat{\Theta}}$ and using the fact that $\hat{\Theta}^T \hat{\Theta} = I_{p-m}$, we can derive that $T_m(\hat{\Theta}) = T_m(\Theta) + o_p(1)$. Therefore, it is sufficient to derive the distribution of $T_m(\Theta)$.

To study the distribution of $T_m(\Theta)$, we first apply results from perturbation theory (see Appendix B of Li (1992) and Kato (1976)) to find an approximation of $\hat{A}$, the formula of which is ommited due to its length. Then, using this approximation, we obtain that

$$\begin{aligned} B &:= (\Theta^T \hat{A}_1 \Theta, \Theta^T \hat{A}_2 \Theta, \ldots, \Theta^T \hat{A}_h \Theta)^T \\ &= \frac{1}{n} \sum_{i=1}^{n} \{g_i \otimes (V_i V_i^T - I_{p-m})\} - \mathrm{E}\{G \otimes (VV^T - I_{p-m})\} + o_p(n^{-1/2}), \end{aligned}$$

where $V_i = \Theta^T z_i$, $g_i = ((\mathbb{1}_{y_i \in J_1} - \Pr(y \in J_1)) \Pr(y \in J_1)^{-1/2}, \ldots, (\mathbb{1}_{y_i \in J_h} - \Pr(y \in J_h)) Pr(y \in J_h)^{-1/2})^T$, and $G = ((\mathbb{1}_{y \in J_1} - \Pr(y \in J_1)) \Pr(y \in J_1)^{-1/2}, \ldots, (\mathbb{1}_{y \in J_h} - \Pr(y \in J_h)) \Pr(y \in J_h)^{-1/2})^T$. $\mathbb{1}_E$ is an indicator function indicating whether or not the event E is true.

Because $g_i \otimes (V_i V_i^T - I_{p-m})$ are independent and identically distributed with mean $\mathrm{E}\{G \otimes (VV^T - I_{p-m})\}$ and finite variance, the Central Limit Theorem implies that

$$\sqrt{n}\mathrm{vec}(B) \longrightarrow_d N(0, \mathrm{Var}\{G \otimes (VV^T - I_{p-m})\}). \tag{4.38}$$

We could further break down $\mathrm{Var}\{G \otimes (VV^T - I_{p-m})\}$. To simplify the notation, let $W = VV^T - I_{p-m}$. We know that $\mathrm{Var}(V) = \mathrm{Var}\{\mathrm{E}(V|\Psi^T z)\} + \mathrm{E}\{\mathrm{Var}(V|\Psi^T z)\}$. Since $\mathrm{Var}(V|\Psi^T z)$ is constant by Condition two and $\mathrm{E}(V|\Psi^T z) = 0$ by Condition one, we have $\mathrm{Var}(V|\Psi^T z) = \mathrm{Var}(V) = I_{p-m}$. Consequently, $\mathrm{E}(W|\Psi^T z) = 0$.

Using $\mathrm{E}(W|\Psi^T z) = 0$ and the fact that G and W are conditional independent given $\Psi^T z$, we derive

$$\begin{aligned}
\mathrm{Var}(G \otimes W) &= \mathrm{Var}\{G \otimes \mathrm{vec}(W)\} \\
&= \mathrm{E}[\mathrm{Var}\{G \otimes \mathrm{vec}(W)|\Psi^T z\}] \\
&= \mathrm{E}[\mathrm{E}\{GG^T \otimes \mathrm{vec}(W)\mathrm{vec}(W)^T|\Psi^T z\}] \qquad (4.39)\\
&= \mathrm{E}[\mathrm{E}(GG^T|\Psi^T z) \otimes \mathrm{E}\{\mathrm{vec}(W)\mathrm{vec}(W)^T|\Psi^T z\}] \\
&= \mathrm{E}[\mathrm{E}(GG^T|\Psi^T z) \otimes \mathrm{Var}(W|\Psi^T z)].
\end{aligned}$$

Because we have assumed that $\mathrm{Var}(W)$ is nonrandom and we know that $\mathrm{E}(W|\Psi^T z) = 0$, $\mathrm{Var}(W) = \mathrm{Var}(W|\Psi^T z)$. In addition, by the definition of G, we have $\mathrm{Var}(G) = \mathrm{E}(GG^T|\Psi^T z)$. Substituting these results in to the equality (4.39) gives

$$\mathrm{Var}(G \otimes W) = \mathrm{Var}(G) \otimes \mathrm{Var}(W) = \mathrm{Var}(G) \otimes \mathrm{Var}(VV^T). \qquad (4.40)$$

Finally, by direct computation, we find that $\mathrm{Var}(G)$ is a projection matrix with rank $h-1$ and $\mathrm{Var}(VV^T)$ has at most $(p-m)(p-m+1)/2$ nonzero eigenvalues due to the symmetry of VV^T. The eigenvalues of $\mathrm{Var}(G) \otimes \mathrm{Var}(VV^T)$ are the eigenvalues of $\mathrm{Var}(VV^T)$, each with multiplicity $h-1$. Using these facts combining with results (4.38) and (4.40), we obtain the desired result

$$2T_m(\hat{\Theta}) \longrightarrow_d \sum_i \omega_i \chi_i^2(h-1),$$

with ω_i, $i = 1, \ldots, (p-m)(p-m+1)/2$, being the largest $(p-m)(p-m+1)/2$ eigenvalues of $\mathrm{Var}(VV^T)$.

Finally, when x is normally distributed, it can be shown that $\Theta^T z \sim N(0, I_{p-m})$ and $\mathrm{Var}(\Theta^T z \otimes \Theta^T z)/2$ is a projection matrix with only eigenvalue 1 of multiplicity $(p-m)(p-m+1)/2$. It follows that,

$$T_m(\hat{\Theta}) \longrightarrow_d \chi^2\{(h-1)(p-m)(p-m+1)/2\}.$$

□

In summary, we provide the algorithm for determining d using the test statistic $T_m(\hat{\Theta})$.

An algorithm for choosing the dimension of $S(\Sigma_{save})$: d

1. Given the standardised sample (z_i, y_i) for $i = 1, \dots, n$ and slices $J_1, \dots, J_h$, compute and store
$$\hat{A}_s = \left(\frac{n_s}{n}\right)^{1/2} (I_p - \widehat{\text{Var}}(z|y \in J_s))$$
for $s = 1, \dots, h$. Then compute $\hat{\Sigma}_{save} = \sum_{s=1}^{h} \hat{A}_s^2$.

2. Perform eigenvalue decomposition on $\hat{\Sigma}_{save}$. Denote computed eigenvalues as $\hat{\lambda}_1 > \cdots > \hat{\lambda}_p$, and their corresponding orthonormalised eigenvectors as $\hat{l}_1, \dots, \hat{l}_p$.

3. Set $m = 0$.

4. Let $\hat{\Theta} = (\hat{l}_{m+1}, \dots, \hat{l}_p)$. Compute the test statistic $T_m(\hat{\Theta}) = \frac{n}{2} \sum_{s=1}^{h} \text{tr}\{(\hat{\Theta}^T \hat{A}_s \hat{\Theta})^2\}$.

5. - When x is normally distributed:

     Compute the p-value as $\Pr(\hat{C} > 2T_m(\hat{\Theta}))$, where $\hat{C}$ has asymptotic distribution $\chi^2\{(h-1)(p-m)(p-m+1)/2\}$

   - Otherwise:

     Calculate the eigenvalues of $\text{Var}(\hat{\Theta}^T z \otimes \hat{\Theta}^T z)$ and denote them as $\hat{\omega}_1 \geq \cdots \geq \hat{\omega}_{(p-m)(p-m+1)}$. Then compute the p-value as $\Pr(\hat{C} > 2T_m(\hat{\Theta}))$, where $\hat{C}$ has asymptotic distribution $\sum_{i=1}^{(p-m)(p-m+1)/2} \omega_i \chi_i^2(h-1)$

6. Compare the calculated p-value with the pre-determined cutoff value. If the p-value is larger than the pre-determined cutoff value, then $d = m$ is the final estimate. If not, proceed as if $d > m$ holds. Let $m = m + 1$ and return to step four.

4.3 Conclusion

SIR and SAVE are methods that use inverse regression lines and slicing techniques to recover central subspaces. Because SIR uses the first moment only, it fails when symmetry dependence presents. SAVE tackles this issue by employing both first and second moments. Overall, SAVE is more comprehensive than SIR, but SIR is more efficient (Cook and Lee, 1999). We will see more concrete results in Chapter 6 when we conduct a simulation study on SIR and SAVE. Since SIR and SAVE have their own advantages and disadvantages, hybrid methods have been proposed. For the purpose of this thesis, we will not discuss these hybrid methods. We refer interested readers to Zhu et al. (2007) Li and Wang (2007).

Chapter 5

pHd

In this chapter, we will continue our discussion of sufficient dimension reduction methods. We will introduce a new type of second moment based methods, namely the Principal Hessian Directions(pHd) methods. As their name suggests, pHd methods recover information about the central subspace using the Hessian matrix of the regression function.

5.1 Principal Hessian Directions

We have mentioned during our study of the SIR that its effectiveness in reducing the dimension of covariates can be greatly impaired when the forward regression function has little linear trend. The non-linearity can lead to zero average within each slice, rendering SIR ineffective. To deal with such cases, higher moments are introduced to recover the information missed by SIR. We have studied one such method, SAVE, in the previous chapter. SAVE adopts similar ideas and the slicing technique used by SIR so can be seen as an extension of SIR. In this chapter, we introduce completely different second moment based methods, the methods of principal Hessian directions(pHd). Li (1992) first introduced the idea of using Hessian matrices to estimate central subspaces. Based on this idea, Li then developed response based pHd (pHdy). However, there are several limitations to response based pHd and because of these limitations, Cook (1998) suggested a modified version: residual based pHd (pHdres).

In the following discussions, we will first briefly introduce the response based pHd method. In particular, we want to understand where its major deficiencies come from. Then, we will carefully examine residual based pHd to see how the drawbacks of pHdy are avoided in this modified version. To be consistent with our discussions of SIR and SAVE, we provide a step by step algorithm for the residual based pHd method and study related dimension tests in detail.

We also make some basic assumptions to facilitate our discussions. We assume the central subspace exists and is spanned by the matrix $\Phi \in \mathbb{R}^{p\times d}$, so that $y \perp\!\!\!\perp x|\Phi^T x$. Since we have shown that there is no information lost by standardizing the covariates, we work with the standardized predictor z hereafter. The central subspace $S_{y|z}$ exists and is spanned by the columns of $\Psi := \Sigma^{1/2}\Phi$. Therefore, it is sufficient for us to derive an estimate of Ψ, as we can obtain Φ and consequently the desired space $S_{y|x} = S(\Phi)$ by a simple linear transformation $\Sigma^{-1/2}\Psi$. We compute the sample version $\hat{z}$ by

$$\hat{z} = \hat{\Sigma}^{-1/2}(x - \bar{x}),$$

where $\hat{\Sigma}$ and $\bar{x}$ are sample estimates of $\Sigma = \mathrm{Var}(x)$ and $\mathrm{E}(x)$.

5.1.1 Response based pHd

Li (1992) proposed response based pHd shortly after he introduced SIR. To begin our discussion of response based pHd, we introduce the key idea that motivated the development of pHd in the first place. Consider a set of independent and identically distributed data (y_i, x_i), $i = 1, \ldots, n$, with each x_i standardized to $\hat{z}_i$. Also, denote the Hessian matrix of the forward regression as $H(z) \in \mathbb{R}^{p\times p}$, which is of the form:

$$H(z) = \frac{\partial^2 E(y|z)}{\partial z \partial z^T}. \tag{5.1}$$

Since we have assumed that the central subspace $S_{y|z}$ has a basis Ψ, we can replace the conditional mean $\mathrm{E}(y|z)$ with $\mathrm{E}(y|\Psi^T z)$, which results in

$$\begin{aligned} H(z) &= \frac{\partial^2 \mathrm{E}(y|\Psi^T z)}{\partial z \partial z^T} \\ &= \Psi \frac{\partial^2 \mathrm{E}(y|\Psi^T z)}{\partial(\Psi^T z)\partial(zT\Psi)}\Psi^T. \end{aligned} \tag{5.2}$$

This representation of the hessian matrix $H(z)$ shows that $H(z)$ is degenerate in any direction that is orthogonal to $S_{y|z}$. Furthermore, we observe all the eigenvectors corresponding to nonzero eigenvalues of $\mathrm{E}[H(x)]$ are in $S_{y|z}$. Hence, by finding a way to estimate the average hessian matrix $\mathrm{E}[H(z)]$, we should be able to find at least a subspace, spanned by the eigenvectors of $\mathrm{E}[H(z)]$ that associated with nonzero eigenvalues, of $S_{y|z}$. Based on this idea, the response based Principal Hessian directions (pHdy) method extracts information about $S_{y|z}$ by providing us with estimates of these eigenvectors.

In order to construct estimates of $\mathrm{E}[H(z)]$ and consequently its eigenvectors, Li (1992) applied Stein's lemma, introduced below.

Lemma 5.1 (Stein's Lemma). *(Stein (1981)) Let Y be a normally distributed random variable with mean ξ and variance 1. Also, we assume g, g' are indefinite integrals of the Lebesgue measurable function g' and g'' and all g, g', g'' have finite expectations. Then*

$$\mathrm{E}\{(Y-\xi)g(Y)\} = \mathrm{E}g'(Y), \tag{5.3}$$

$$\mathrm{E}\{(Y-\xi)^2 g(Y)\} = \mathrm{E}\{g(Y)+g''(Y)\}. \tag{5.4}$$

Proof. Because this lemma is covered in many textbooks, we only provide a sketch of the proof. For a detailed proof, we refer interested readers to Stein (1981).

Let $\phi(y)$ be the density of Y. We prove equation (5.3) mainly by applying integration by parts to $\mathrm{E}g'(Y) = \int_{-\infty}^{\infty} g'(y)\phi(y)dy$. During the process, we also need the equality that $\phi'(y) = -y\phi(y)$ to substitute $\phi'(y)$ with $-y\phi(y)$ and Fubini's theorem to change order of integration. Then the result follows.

Equation (5.4) is a consequence of equation (5.3). Without loss of generality, we assume $\xi = 0$. We prove equation (5.4) as follows:

$$\mathrm{E}\{Y^2 g(Y)\} = \mathrm{E}[Y\{Yg(Y)\}] = \mathrm{E}\{Yg(Y)\}' = \mathrm{E}\{g(Y)+Yg'(Y)\} = \mathrm{E}\{g(Y)+g''(Y)\}. \tag{5.5}$$

Here, equation (5.3) is used in the second and the last steps. □

Remark 5.2. Landsman and Neslehova (2008) showed that Stein's Lemma can be extended to multivariate normal vectors. Suppose $Y \in \mathbb{R}^p$ is a multivariate normal vector with mean ξ and variance matrix I. Also, let $g : \mathbb{R}^p \mapsto \mathbb{R}$ be a differentiable function such that

$$\int_{\mathbb{R}^p} \| \frac{\partial g(Y)}{\partial Y_i} \| dv(Y) < \infty, \quad i = 1, \ldots, p$$

$$\int_{\mathbb{R}^p} \| \frac{\partial^2 g(Y)}{\partial Y_i \partial Y_j} \| dv(Y) < \infty, \quad i, j = 1, \ldots, p$$

where v is the measure of Y. Then

$$\mathrm{E}\{(Y-\xi)g(Y)\} = \mathrm{E}\nabla g(Y), \tag{5.6}$$

$$\mathrm{E}\{g(Y)(Y-\xi)(Y-\xi)^T\} = \mathrm{E}\{g(Y)I + \frac{\partial^2 g(Y)}{\partial Y \partial Y^T}\}. \tag{5.7}$$

The proof for the above equations is very similar to the proof of Lemma 5.1. We refer interested readers to Landsman and Neslehova (2008) for details.

Since equation (5.7) can be rearranged as

$$\mathrm{E}\{\frac{\partial^2 g(Y)}{\partial Y \partial Y^T}\} = \mathrm{E}\{g(Y)(Y-\xi)(Y-\xi)^T\} - \mathrm{E}\{g(Y)I\}, \tag{5.8}$$

Stein's lemma provides an alternative way to compute the expectation of the second derivative of a function when Y is distributed normally with variance I. Hence, if we further assume that x follows a normal distribution, we can use this formula to compute the expectation of the Hessian matrix $\mathrm{E}[H(z)]$. We replace the $g(x)$ function with $E(y|z)$. Then equation (5.8) gives

$$\begin{aligned}\mathrm{E}\{\frac{\partial^2 E(y|z)}{\partial z \partial z^T}\} &= \mathrm{E}\{\mathrm{E}(y|z)zz^T\} - \mathrm{E}\{\mathrm{E}(y|z)I\} \\ &= \mathrm{E}\{\mathrm{E}(yzz^T|z)\} - \mathrm{E}(y)I \\ &= \mathrm{E}(yzz^T) - \mathrm{E}(y)\mathrm{E}(zz^T) \\ &= \mathrm{E}((y-\mathrm{E}(y))zz^T).\end{aligned} \tag{5.9}$$

By denoting $\Sigma_{yzz} \coloneqq \mathrm{E}((y-\mathrm{E}(y))zz^T)$, we conclude that

$$\Sigma_{yzz} = \Phi\mathrm{E}(\frac{\partial^2 \mathrm{E}(y|\Phi^T z)}{\partial(\Phi^T z)\partial(z^T\Phi)})\Phi^T$$

and consequently $\Sigma_{yzz} \in S_{y|z}$. Therefore, estimating $S_{y|z}$ with the response based Hessian matrix is, in essence, finding the eigenvectors corresponding to the non-zero eigenvalues of the population moment matrix Σ_{yzz}. We denote the ordered eigenvalues of Σ_{yzz} as $\delta_1, \ldots, \delta_p$ with $|\delta_1| \geq |\delta_2| \cdots \geq |\delta_p|$ and their associated eigenvectors as $l_1, \ldots, l_p$. If the rank of Σ_{yzz} is d, $l_1, \ldots, l_d$ are then called the principal Hessian directions (Li, 1992). Our pHdy estimate of $S_{y|z}$ is the space spanned by $l_1, \ldots, l_d$, denoted as S_{yzz}.

Finally, we need to develop tests for determining d to apply pHdy in practice. Similar to what we have did for SIR, we introduce the test statistic

$$\hat{\Delta}_{pHdy}(m) = \frac{n}{2\widehat{\mathrm{Var}}(y)} \sum_{j=m+1}^{p} \hat{\delta}_j^2.$$

Li (1992) proved that

$$\hat{\Delta}_{pHdy}(m) \sim \chi^2\{(p-m)(p-m+1)/2\},$$

We use this asymptotic result to estimate d by testing hypotheses $d = m$ vs $d > m$, starting from $m = 0$.

So far, we have outlined the idea behind pHdy and how we proceed with this method to derive an estimate of the central subspace. Although it is straightforward and easy to apply, pHdy has several drawbacks that greatly hinder its use in applications. To start, the pre-requirement for pHdy to work is fairly strict. pHdy requires x to be normally distributed, as it relies on Stein's lemma to estimate $\mathrm{E}[H(z)]$. However, given that $y \perp\!\!\!\perp x|\Phi^T x$, SIR simply requires the conditional expectation for the predictor to be linear for Φ, a much looser condition that is generally met in most high dimensional data problems. SAVE additionally requires constant variance, but still has wider applicability than pHdy.

More importantly, Cook (1998) pointed out that pHdy is not effective in finding linear trends. Since the Hessian matrix $H(z)$ is a second order differential operator, it does not change when a linear term of the predictor is added to the regression function. When the true regression is a linear function of the covariate, for instance,

$$y|z = \alpha + \eta^T z + \epsilon, \tag{5.10}$$

where z normally distributed, $\epsilon \perp\!\!\!\perp z$ and $\mathrm{E}(\epsilon) = 0$, it straightforward to see that $H(z) = 0$ and consequently $\mathrm{E}[H(z)] = \Sigma_{yzz} = \mathbf{0}$. Because of these properties of $H(z)$, it is likely that pHdy cannot produce satisfactory estimate when linear trends present. An example showing pHdy's lack of ability in detecting linear trends was given in (Cook, 1998). In his example, the plot of all data points and the fitted regression using the ordinary least squares (OLS) exhibited a clear linear relationship. However, applying the pHdy method suggested one important direction, which lead to an inappropriate curved relationship between the response and the predictor. The sample correlation is low at 0.11.

However, there are still cases when pHdy does have some power in revealing linear trends (For instance, see section 4.1 of Cook (1998)). The reason behind this surprising result is also the cause for the third and the last drawback of pHdy we will cover. An extra condition was implicitly assumed when Li proved the asymptotic result for the test statistic $\hat{\Delta}_{save}(m)$ (Cook, 2009, 1998; Weisberg, 2015). Let $\beta = \mathrm{Cov}(z, y)$ and $\Theta_0 = (l_{d+1}, \ldots, l_p)$, the eigenvectors of Σ_{yzz} corresponding to the eigenvalue 0. In Li's proof, Li used the condition $\Theta_0^T \beta = 0$ to derive the final asymptotic result. This condition, however, is not generally true. To illustrate, we consider the model (5.10). In this case, $d = \dim(S_{yzz}) = \dim(\mathbf{0}) = 0$, Θ_0 is the identity matrix and $\beta = \eta$, so $\Theta_0^T \beta$ is clearly nonzero. Since the condition $\Theta_0^T \beta = 0$ does not always hold, the method used by pHdy for choosing d could be unreliable. Furthermore, in the proof, Li showed that the asymptotic distribution of $\hat{\Delta}_{save}(m)$ is dependent on β via $\Theta_0^T \beta$. Because $\Theta_0^T \beta$ is not necessarily zero, the distribution of $\hat{\Delta}_{save}(m)$ can depend on β, contrary to Li's claim. This dependence relationship also accounts for pHdy's success in detecting linear trends in some cases. In summary, depending on whether $\Theta_0^T \beta = 0$, pHdy's performance in estimating the central subspace may fluctuate drastically, causing unnecessary complexities. We will provide more information about the assumption $\Theta_0^T \beta = 0$ in our later discussions of choosing d for pHdres.

Given the above discussions, pHdy has stricter requirements, compared to other available methods. To apply pHdy, we need x to be normally distributed and $\Theta_0^T \beta = 0$. Assuming x follows a normal distribution, a possible scenario for pHdy to work consistently is when $S_{yzz} = S_{y|z}$, as this condition forces $\Theta_0^T \beta = 0$. Still, with $\Theta_0^T \beta = 0$, pHdy is highly unlikely to detect any linear trend. Due to all these complexities and restrictions of pHdy, an improved and modified version of pHdy is needed.

5.1.2 Residual based pHd

Development of Residual based pHd is mainly motivated by the fact that pHdy is, in general, not effective in revealing linear trends of forward regressions. Thus, to maximise the use of Hessian matrices in extracting information about the central subspace, Cook (2009) suggested that we start by removing the linear relationship between the response variable and the predictor variable from the response variable. Then we can apply the pHdy method on the residual to obtain an estimate of the central subspace for the residual. Hopefully, the union of the linear coefficient vector and the pHdy estimate based on the residual can provide us with a satisfactory estimate of the central subspace $S_{y|z}$. Because this method estimates $S_{y|z}$ mainly relying on pHdy except, in this case, pHdy is applied to the residuals instead of the response variable, we call this method residual based pHd (pHdres).

We know that a useful tool for estimating linear relationships is ordinary least square(OLS) regression of y on z. Thus, we can study the population OLS residual e, calculated as

$$e = y - \mathrm{E}(y) - \beta^T z, \tag{5.11}$$

where $\beta \coloneqq \mathrm{Cov}(z, y)$. In terms of sample residuals $\hat{e}_i$, we let $\hat{z}$ be the sample version of the standardized predictor. Then, given the standardized data set $(y_i, \hat{z}_i)$, $i = 1, \ldots, n$, we compute sample residuals $\hat{e}_i$ similarly by

$$\hat{e}_i = y_i - \bar{y} - \hat{\beta}^T \hat{z}_i. \tag{5.12}$$

Here, we apply OLS regression of y on $\hat{z}$ to obtain an estimate $\hat{\beta}$ of the linear coefficient β.

Since the linear trend has been removed, pHdy should be effective in recovering the central subspace for the regression of e on z. Let $S_{e|z}$ denote this central subspace. Both β and $S_{e|z}$ can be easily described using the formula $\mathrm{Cov}(z, y)$ and pHdy. Thus, if we can verify that the union of β and $S_{e|z}$ is at least a subspace of $S_{y|z}$, pHdres should be an effective and efficient method for finding an approximation to the central subspace $S_{y|z}$. To unravel the relationship between $S_{e|z} \cup S(\beta)$ and $S_{y|z}$, we start by investigating the connection between β and the central subspace $S_{y|z}$.

We recall that β is the solution that minimizes the objective function $R(a, b) \coloneqq \mathrm{E}(L(a + b^T z, y))$, where $L(a + b^T z, y) = (y - a - b^T z)^2$ and the expectation is with respect to the joint distribution of y and z. That is

$$(\mathrm{E}(y), \beta) = \arg\min_{a,b} R(a, b).$$

Here, we point out that the loss function L takes input variables $(a + b^T z, y)$ instead of (z, y) due to its underlying assumption that the objective function has a linear kernel $a + b^T z$. Since, in this setting, the explicit form of the loss function L shows it is a strictly convex function about $a + b^T z$, a theorem from Li and Duan (1989) can shed some light on the connection between β and $S_{y|z}$.

Theorem 5.3 (Li and Duan (1989)). *Let $S_{drs}(\Phi)$ be a dimension-reduction subspace for the regression of y on x. Also assume*

$$(\alpha, \beta_x) = \arg\min_{a,b} R(a, b) := \arg\min_{a,b} \mathrm{E}[L(a + b^T x, y)].$$

Then $\beta_x \in S_{drs}(\Phi)$, if

1. *β_x is unique.*

2. *$L(u, v)$ is convex in u.*

3. *The conditional expectation $\mathrm{E}(x|\Phi^T x)$ is a linear function of $\Phi^T x$ and $\Sigma = \mathrm{Var}(x)$ is positive definite.*

Proof. The key to the proof is to use Jensen's inequality. To do so, we first write $R(a, b)$ as a conditional expectation incorporating the fact that $S_{drs}(\Phi)$ is a dimension reduction subspace, that is $y \perp\!\!\!\perp x|\Phi^T x$:

$$\begin{aligned} R(a, b) &= \mathrm{E}[L(a + b^T x, y)] = \mathrm{E}_{y,\Phi^T x}\mathrm{E}_{x|y,\Phi^T x}[L(a + b^T x, y)] \\ &= \mathrm{E}_{y,\Phi^T x}\mathrm{E}_{x|\Phi^T x}[L(a + b^T x, y)]. \end{aligned} \tag{5.13}$$

Given that L is convex in its first argument, Jensen's inequality gives that

$$R(a, b) \geq \mathrm{E}_{y,\Phi^T x}[L\{a + b^T E(x|\Phi^T x), y\}].$$

Without loss of generality, we assume that $\mathrm{E}(x) = 0$. Since $\mathrm{E}(x|\Phi^T x)$ is linear in $\Phi^T x$, by Proposition 4.1, we derive

$$R(a, b) \geq \mathrm{E}_{y,\Phi^T x}[L(a + (P_{\Phi(\Sigma)} b)^T x, y)].$$

It follows that

$$R(a, b) \geq R(a, P_{\Phi(\Sigma)} b).$$

We know that $P_{\Phi(\Sigma)} b \in S_{drs}(\Phi)$. Because a, b are arbitrary, β_x is a minimiser and β_x is unique, we must have $\beta_x \in S_{drs}(\Phi)$. □

In our setting, condition two of the theorem is automatically satisfied, as $L(a+b^Tz,y) = (y-a-b^Tz)^2$ is convex by definition. Moreover, we note that $L(a+b^Tz,y)$ is actually strictly convex, ensuring the uniqueness of β. Therefore, given this theorem, we can force $\beta \in S_{y|z}(\Psi)$ by requiring $\mathrm{E}(z|\Psi^Tz) = P_\Psi z$.

So far, we have shown that when $\mathrm{E}(z|\Psi^Tz) = P_\Psi z$, $\beta \in S_{y|z}(\Psi)$. We are now interested in the relationship between $S_{e|z}$ and $S_{y|z}$ under the assumption $E(z|\Psi^Tz) = P_\Psi z$. In fact, by adding the additional requirement that $\mathrm{E}(z|\Psi^Tz) = P_\Psi z$, the combination of $S_{e|z}$ and $S(\beta)$ recovers the whole central subspace.

Proposition 5.4 (Cook (2009))**.** *Let (y_i, x_i), $i = 1, \dots, n$ be a set of i.i.d data and z_i's be standardised predictor variables. Also let e_i be defined as in the equation (5.11). Assume that the central subspaces $S_{e|z}$ and $S_{y|z}$ are spanned by the columns of the matrices Υ and Ψ respectively. Then if $\mathrm{E}(z|\Psi^Tz) = P_\Psi z$, we have*

$$S_{y|z} = S_{e|z} + S(\beta). \tag{5.14}$$

Proof. To prove the proposition, we first observe that, by the definition of e and Υ, we have

$$y - \beta^Tz \perp\!\!\!\perp z|\Upsilon^Tz.$$

Then, direction applications of Propositions 3.1 and 3.2 on conditional independence give

$$y - \beta^Tz \perp\!\!\!\perp z|(\Upsilon^Tz, \beta^Tz)$$

and

$$(y - \beta^Tz, \beta^Tz) \perp\!\!\!\perp z|(\Upsilon^Tz, \beta^Tz).$$

Applying Proposition 3.2 again shows that

$$y \perp\!\!\!\perp z|(\Upsilon^Tz, \beta^Tz),$$

which indicates that $S(\Upsilon, \beta)$ is also a dimension reduction subspace for the regression of y on z. Since the central subspace $S_{y|z}$ is contained in any dimension reduction subspace,

$$S_{y|z} \subset S(\Upsilon, \beta) = S_{e|z} + S(\beta). \tag{5.15}$$

Because we have assumed that $\mathrm{E}(z|\Phi^T z) = P_\Phi z$, by Theorem 5.3, we know $\beta \in S_{y|z}$. In addition, as $\beta \in S_{y|z}$, the formula

$$e = y - \mathrm{E}(y) - \beta^T z \tag{5.16}$$

implies that $S_{e|z} \subset S_{y|z}$. Combining these results with equation (5.15) , we have the desired conclusion

$$S_{y|z} \subset S_{\Upsilon,\beta} \subset S_{\Psi,\beta} = S_{y|z}. \tag{5.17}$$

□

Remark 5.5. We point out that the key assumption for the above proposition to hold is $\mathrm{E}(z|\Psi^T z) = P_\Psi z$. The reasons for its importance are twofold. Firstly, this assumption is required to apply the Theorem from Li and Duan in order to force $\beta \in S_{y|z}$. Secondly, it is the fact $\beta \in S_{y|z}$ that leads us to conclude that $S_{e|z} \subset S_{y|z}$. If $\beta \notin S_{y|z}$, the regression of the residual on z can be more complicated than the regression of y on z. To be more specific, when $\beta \notin S_{y|z}$, we will have $\dim[S_{e|z}] > \dim[S_{y|z}]$, as the formula (5.16) indicates the central subspace $S_{e|z}$ has to contain the dimension determined by β. As a side note, we also remind the reader that the central subspace can be trivial. For example, if the regression of y over z follows a linear model

$$y|z = \beta_0 + \beta_1 z + \epsilon,$$

we have $S_{e|z} = 0$.

In general, given the existence of $S_{y|z}$ and $S_{e|z}$, we have $S_{y|z} \subset S(\Upsilon, \beta) = S_{e|z} + S(\beta)$. However, with the additional assumption of $E(z|\Phi^T z) = P_\Phi z$, $S_{e|z} + S(\beta)$ is restricted to be a subset of $S_{y|z}$, establishing the equality.

This Proposition establishes a nice equivalence relationship between the desired result $S_{y|z}$ and the union of $S_{e|z}$ and $S(\beta)$ under the key assumption that $\mathrm{E}(z|\Psi^T z) = P_\Psi z$. We observe that this assumption is similar to the defining condition of elliptically contoured distributions. A random variable x is elliptically distributed if and only if $\mathrm{E}(z|B^T z)$ is linear function in $B^T z$ for all conforming matrix B. Because our requirement $\mathrm{E}(z|\Psi^T z) = P_\Psi z$ only requires it to be true at a fixed matrix Φ, it is more specific and less strict than that of elliptically distributed variables. The assumption $\mathrm{E}(z|\Psi^T z) = P_\Psi z$ is therefore automatically satisfied when z has a elliptically contoured distribution.

Since we are working with linear regression, $S(\beta)$ can be easily computed. We now focus on recovering the central subspace of the regression of e on z using the key idea behind pHdy.

5.1.3 Estimating $S_{e|z}$

We now estimate the central subspace $S_{e|z}$. Based on the key idea of pHdy, we want to estimate $S_{e|z}$ with the eigenvectors corresponding to the non-zero eigenvalues of the expected Hessian matrix for the regression function on the residual:

$$\mathrm{E}[H_e(z)] = \mathrm{E}(\frac{\partial^2 \mathrm{E}(e|z)}{\partial z \partial z^T}) = \Upsilon \mathrm{E}(\frac{\partial^2 \mathrm{E}(e|\Upsilon^T z)}{\partial(\Upsilon^T z)\partial z^T \Upsilon})\Upsilon^T. \tag{5.18}$$

We know that $S(\mathrm{E}[H_e(z)]) \subseteq S_{e|z}$ due to the formula of $\mathrm{E}[H_e(z)]$. The only question remains is that how do we derive an explicit form for the expected Hessian matrix in order to compute its eigenvectors and eigenvalues?

We recall in pHdy, Li (1992) used Stein's Lemma to estimate Σ_{yzz} under the assumption z is normally distributed. Since this pre-requirement is fairly strict, Cook (1998) extended and refined Li's idea to estimate $S_{e|z}$ under relatively loose assumptions. In the following, we will quickly go through the procedure for estimating $S_{e|z}$ using Li's proposal. After that, we will carefully discuss Cook's approach.

We assume that z is normally distributed. With the help of Stein's Lemma, equation (5.8) gives

$$\begin{aligned} \mathrm{E}\{\frac{\partial^2 \mathrm{E}(e|z)}{\partial z \partial z^T}\} &= \mathrm{E}\{zz^T \mathrm{E}(e|z)\} - \mathrm{E}\{\mathrm{E}(e|z)\} \\ &= \mathrm{E}\{\mathrm{E}(ezz^T|z)\} - \mathrm{E}(e) \\ &= \mathrm{E}(ezz^T). \end{aligned} \tag{5.19}$$

Denoting $\mathrm{E}(ezz^T)$ as Σ_{ezz}, we have

$$\Sigma_{ezz} = \Upsilon \mathrm{E}(\frac{\partial^2 E(e|\Upsilon^T z)}{\partial(\Upsilon^T z)\partial z^T \Upsilon})\Upsilon^T.$$

Let $S_{ezz} := S(\Sigma_{ezz})$. Then $S_{ezz} \subseteq S_{e|z}$. Thus, we estimate $S_{e|z}$ by finding S_{ezz}, which is spanned by the eigenvectors that correspond to nonzero eigenvalues of Σ_{ezz}.

5.1.3.1 Cook's approach

In this subsection, we follow Cook (2009)'s idea to estimate $S_{e|z}$. Instead of requiring z to be normally distributed, Cook loosened the condition to require $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$ only. Although Cook did not rely on Stein's Lemma, he adopted Li's idea and estimated $S_{e|z}$ by establishing a connection between $S_{e|z}$ and Σ_{ezz} as well.

To start, we want to find the properties of Σ_{ezz} when $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$.

Proposition 5.6. *Assume the central subspace $S_{e|z}$ is spanned by columns of Υ. Let P_Υ be an orthogonal projection operator for $S_{y|z}$ and $Q_\Upsilon = I - P_\Upsilon$. Then if $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$, we have*

$$\Sigma_{ezz} = Q_\Upsilon \mathrm{E}[e \times \mathrm{Var}(z|\Upsilon^T z)] Q_\Upsilon + P_\Upsilon \Sigma_{ezz} P_\Upsilon. \tag{5.20}$$

Proof. Firstly, we recall that when $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$, a direct application of Proposition 4.1 gives

$$\mathrm{E}(z|e) = P_\Upsilon E(z|e) \in S_{e|z}(\Upsilon). \tag{5.21}$$

Secondly, from the law of total variance, we can write

$$\begin{aligned}\Sigma_{z|e} &:= \mathrm{Var}(z|e)\\ &= \mathrm{E}[\mathrm{Var}(z|\Upsilon^T z, e)|e] + \mathrm{Var}[\mathrm{E}(z|\Upsilon^T z, e)|e]\\ &= \mathrm{E}[\mathrm{Var}(z|\Upsilon^T z)|e] + \mathrm{Var}[\mathrm{E}(z|\Upsilon^T z)|e] \quad (\text{because } e \perp\!\!\!\perp z|\Upsilon^T z)\\ &= \mathrm{E}[\mathrm{Var}(z|\Upsilon^T z)|e] + P_\Upsilon \Sigma_{z|e} P_\Upsilon \quad (\text{using the equation } \mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z).\end{aligned} \tag{5.22}$$

Combining the above results, we can obtain

$$\begin{aligned}\mathrm{E}(zz^T|e) &= \Sigma_{z|e} + \mathrm{E}(z|e)\mathrm{E}(z^T|e)\\ &= \mathrm{E}[\mathrm{Var}(z|\Upsilon^T z)|e] + P_\Upsilon \Sigma_{z|e} P_\Upsilon + P_\Upsilon \mathrm{E}(z|e)\mathrm{E}(z^T|e) P_\Upsilon \quad (\text{using the equation (5.21)})\\ &= \mathrm{E}[\mathrm{Var}(z|\Upsilon^T z)|e] + P_\Upsilon \mathrm{E}(zz^T|e) P_\Upsilon\\ &= Q_\Upsilon \mathrm{E}[\mathrm{Var}(z|\Upsilon^T z)|e] Q_\Upsilon + P_\Upsilon \mathrm{E}(zz^T|e) P_\Upsilon.\end{aligned} \tag{5.23}$$

The last equality uses the facts that $\mathrm{Var}(z|\Upsilon^T z) = \mathrm{Var}(P_\Upsilon z + Q_\Upsilon z|\Upsilon^T z) = \mathrm{Var}(Q_\Upsilon z|\Upsilon^T z)$ and Q_Υ is an orthogonal projector.

Since $\Sigma_{ezz} = \mathrm{E}[e \times \mathrm{E}[zz^T|e]]$, we hence conclude:

$$\Sigma_{ezz} = Q_\Upsilon \mathrm{E}[e \times \mathrm{Var}(z|\Upsilon^T z)]Q_\Upsilon + P_\Upsilon \Sigma_{ezz} P_\Upsilon. \tag{5.24}$$

□

This proposition tells us that when $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$, the eigenvectors, corresponding to nonzero eigenvalues, of Σ_{ezz} must be in either $S_{e|z}(\Upsilon)$ or its orthogonal complement. However, there is no clear cut way to distinguish which eigenvectors belong to $S_{e|z}(\Upsilon)$ and which eigenvectors contain no information about $S_{e|z}(\Upsilon)$. However, when $\dim(S_{ezz})$ is small, which is often the case in practice, we can use graphical methods to make a decision. For example, we may need to plot the response variable against the direction of each eigenvector and then rule out eigenvectors for which the graphs show independence relationships. Once we can develop a method for determining the rank of S_{ezz}, it is feasible to identify $S_{e|z}$ related eigenvectors with graphs in this way.

Remark 5.7. We observe that if we impose a further restriction by making $\mathrm{Var}(z|\Upsilon^T z)$ constant, we will have

$$\Sigma_{ezz} = Q_\Upsilon \mathrm{E}[e \times \mathrm{Var}(z|\Upsilon^T z)]Q_\Upsilon + P_\Upsilon \Sigma_{ezz} P_\Upsilon = P_\Upsilon \Sigma_{ezz} P_\Upsilon.$$

As a result, S_{ezz} is a subspace of $S_{e|z}$. We then proceed in the same way as Li (1992) suggested.

In fact, we recall from Chapter two that when z is normally distributed, $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$ and $\mathrm{Var}(z|\Upsilon^T z)$ is constant. Hence, z being normally distributed can be seen as a special case of the conditions required by Cook. It follows that, when z is normally distributed, we can estimate $S_{e|z}$ by either Cook's method or Li's method; the derivation of these methods is different but the implementation is the same. That is, we estimate $S_{e|z}$ with S_{ezz}.

To end our short discussion of Cook's methodology, we emphasise that for Cook's methodology to work, we require $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$ in addition to $\mathrm{E}(z|\Phi^T z) = P_\Phi z$, where Υ and Φ span $S_{e|z}$ and $S_{y|z}$ respectively. Although it seems that $S(\Upsilon)$ and $S(\Phi)$ only differ by the vector β, these two conditions do not necessarily imply to each other. Nevertheless, if z has an elliptically contoured distribution, these two conditions are automatically satisfied.

5.1.4 pHdres algorithm

We summarise the step-by-step algorithm for pHdres. Assume we are given a set of independent and identically distributed samples (x_i, y_i), $i = 1, \ldots, n$.

1. Standardizing covariate variables. Denote the sample variance of x as $\hat{\Sigma}$ and the sample mean as $\bar{x}$. Then compute the standardized covariates as
$$\hat{z}_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x}).$$
2. Compute $\hat{\beta} = \text{Cov}(\hat{z}, y)$ and $\bar{y}$.
3. Given $\hat{\beta}$ and $\bar{y}$, calculate the residual $\hat{e}_i \coloneqq y_i - \bar{y} - \hat{\beta}^T \hat{z}_i$ for $i = 1, \ldots, n$.
4. Calculate the sample estimate of the population moment matrix Σ_{ezz}, using the formula
$$\hat{\Sigma}_{ezz} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i \hat{z}_i \hat{z}_i^T.$$
5. Perform the eigenvalue decomposition of $\hat{\Sigma}_{ezz}$. Denote the eigenvalues as $\hat{\delta}_1, \ldots, \hat{\delta}_p$, with $|\hat{\delta}_1| \geq \ldots |\hat{\delta}_p|$, and their associated eigenvectors as $\hat{l}_1, \ldots, \hat{l}_p$.
6. Let $d = \dim(S_{ezz})$. The span of $\hat{l}_1, \ldots, \hat{l}_d$ gives an estimate $\hat{S}_{ezz}$.
    - Assume Υ spans $S_{e|z}$ and d is small. If $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$, use graphs to determine which eigenvecoters of $\hat{l}_1, \ldots, \hat{l}_d$ estimate $S_{e|z}$. Denote these vectors by $\hat{l}_1^e, \ldots, \hat{l}_d^e$. The union of $\text{span}(\hat{l}_1^e, \ldots, \hat{l}_d^e)$ and $S(\hat{\beta})$ is the pHdres estimate of the central subspace $S_{y|z}$.
    - If $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$ and $\text{Var}(z|\Upsilon^T z)$ is a constant, $S_{ezz} \subset S_{e|z}$. The union of $\hat{S}_{ezz}$ and $S(\hat{\beta})$ is the pHdres estimate of the central subspace $S_{y|z}$.
7. Finally, since $S_{y|x} = \Sigma^{-1/2} S_{y|z}$, back transform the pHdres estimate by left multiplying $\hat{\Sigma}^{-1/2}$ to obtain the pHdres estimate of $S_{y|x}$.

We make some comments about the pHd algorithm listed above. Firstly, it is important to note that for the above procedure to work, we have implicitly assumed that:

$$E(z|\Psi^T z) = P_\Psi z,$$

given $S_{y|z} = S(\Psi)$. Secondly, the idea of using Hessian principal directions is used to estimate the space S_{ezz}. If we back-transform the eigenvectors $\hat{l}_1, \dots, \hat{l}_d$ to the original scale and denote them as $\hat{u}_1 := \hat{\Sigma}^{-1/2}\hat{l}_1, \dots, \hat{u}_p := \hat{\Sigma}^{-1/2}\hat{l}_p$, we can refer to the linear combinations $\hat{u}_1^T x, \dots, \hat{u}_d^T x$ as pHd predictors. Finally, we point out that although we introduced two different methods (one from Li and one from Cook) for estimating $S_{e|z}$, they both result in estimating $S_{e|z}$ with S_{ezz}. Li requires x to be normally distributed and concludes that $S_{ezz} \subset S_{e|z}$. Cook's method is more general and only requires $\mathrm{E}(z|\Upsilon^T z) = P_\Upsilon z$. Depending on whether $\mathrm{Var}(z|\Upsilon^T z)$ is a constant or not, we can either conclude $S_{ezz} \subset S_{e|z}$ or use graphical methods to estimate $S_{e|z}$. It is important to note that graphical methods are feasible only when the dimension of S_{ezz} is small.

5.1.5 A method for choosing the dimension of S_{ezz}

In this section, we introduce a method to choose the dimension $d = \dim(S_{ezz}) = \dim(S(\Sigma_{ezz}))$ so we can use pHdres in practice. Similarly to the method proposed for SIR, a widely used method is to formulate a test statistic for d, find the asymptotic distribution of such test statistic, and then test hypotheses about d to choose a value of d. Adopting this idea, we introduce the following test statistic, proposed by Li (1992):

$$\hat{\Delta}_{pHdres}(m) = \frac{n \sum_{j=m+1}^{p} \hat{\delta}_j^2}{2\mathrm{Var}(\hat{e})}. \tag{5.25}$$

The asymptotic distribution of $\hat{\Delta}_{pHdres}(m)$ has been carefully studied by both Li and Cook. We combine their results in the following proposition.

Proposition 5.8. *Let d be the dimension of the space S_{ezz} and define $\hat{\Delta}_d$ as in the equation (5.25). The asymptotic distribution of $\hat{\Delta}_{pHdres}(d)$ is the same as*

$$C = \frac{1}{2\mathrm{Var}(e)} \sum_{j=1}^{(p-d)(p-d+1)/2} \omega_j \chi(1), \tag{5.26}$$

where the $\chi(1)$'s are independent Chi-square variables, each with one degree of freedom and $\omega_1 \geq \dots \geq \omega_{(p-d)(p-d+1)/2}$ are eigenvalues of the matrix $\mathrm{Var}(eW)$. Here, W is defined *as*

$$W := \begin{pmatrix} \begin{pmatrix} v_1^2 - 1 \\ \sqrt{2} v_1 v_2 \\ \sqrt{2} v_1 v_3 \\ \vdots \\ \sqrt{2} v_1 v_{p-d} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} v_j^2 - 1 \\ \sqrt{2} v_j v_{j+1} \\ \vdots \\ \sqrt{2} v_j v_{p-d} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} v_{p-d-1}^2 - 1 \\ \sqrt{2} v_{p-d-1} v_{p-d} \\ (v_{p-d}^2 - 1) \end{pmatrix} \end{pmatrix} \in \mathbb{R}^{(p-d)(p-d+1)/2 \times 1}, \tag{5.27}$$

where $v_1 = l_{d+1}^T z, \ldots, v_{p-d} = l_p^T z$ *and* $l_{d+1}, \ldots, l_p$ *are eigenvectors of* Σ_{ezz} *corresponding to the zero eigenvalues of* Σ_{ezz}.

Before we prove the Proposition, we introduce the following classical results from perturbation theory (see, for example Kato (1976), Eaton and Tyler (1991)), because they play an important role in the proof of the Proposition.

Lemma 5.9. *Consider the second-order expansion*

$$T(w) = T + wT^{(1)} + w^2 T^{(2)} + o(w^2)$$

where $T(w), T, T^{(1)}, T^{(2)} \in \mathbb{R}^{p \times p}$ *are symmetric matrices and the rank of* T *is* k. *Let* $\lambda(w)$ *be the sum of the* $p - k$ *eigenvalues of* $T(w)$ *that are closest to* 0, *and let* $\Pi(w)$ *be the projection matrix of the space spanned by the* $p - k$ *associated eigenvectors. Also denote the projection matrix of the null space of* T *to be* Π, *so that* $\Pi T = T\Pi = 0$. *Then,*

$$\Pi(w) = \Pi - w\Pi T^{(1)} T^{\dagger} T^{\dagger} T^{(1)} \Pi + o(w), \tag{5.28}$$

and

$$\lambda(w) = w\lambda^{(1)} + w^2 \lambda^{(2)} + o(w^2), \tag{5.29}$$

with $\lambda^{(1)} = tr(T'\Pi)$, $\lambda^{(2)} = tr[T^{(2)}\Pi - T^{(1)}T^{\dagger}T^{(1)}\Pi]$. *Here the superscript* $\dagger$ *denotes the Moore-Penrose generalised inverse of a matrix.*

We are now ready to prove the proposition.

Proof. Since the problem of interest is invariant under affine transformation, we assume that $E(y) = 0$. We also define the following terms to simplify our discussions:

$$\beta_e = \mathrm{Cov}(z, e),$$

$$\Xi_i = z_i z_i^T - I_p, \qquad \bar{\Xi} = \frac{1}{n}\sum_i^n \Xi_i,$$

and

$$C_i = e_i \Xi_i - \Sigma_{ezz}, \qquad \bar{C} = \frac{1}{n}\sum_i^n C_i.$$

We note that we are interested in the distribution of the sum of squared eigenvalues while the Lemma 5.9 concerns the sum of eigenvalues. To overcome this limitation and apply Lemma 5.9 in our context, we use the fact that the eigenvalues of $\hat{\Sigma}_{ezz}\hat{\Sigma}_{ezz}^T$ are exactly the square of the eigenvalues of $\hat{\Sigma}_{ezz}$. Hence, to start, we need to find the second-order expansion of $\hat{\Sigma}_{ezz}\hat{\Sigma}_{ezz}^T$. Given the formula for $\hat{\Sigma}_{ezz}$ and Σ_{ezz}, direct calculation gives

$$\hat{\Sigma}_{ezz}\hat{\Sigma}_{ezz}^T = \Sigma_{ezz}\Sigma_{ezz} + (B_n\Sigma_{ezz} + \Sigma_{ezz}B_n) + \{B_nB_n + o_p(n^{-1/2})\Sigma_{ezz} + \Sigma_{ezz}o_p(n^{-1/2})\} + o_p(n^{-1}), \tag{5.30}$$

where

$$B_n = \bar{C} - \bar{z}\beta_e^T - \beta_e^T\bar{z}^T - (1/2)\bar{\Xi}\Sigma_{ezz} - (1/2)\Sigma_{ezz}\bar{\Xi}.$$

Let $\Theta_0 \in \mathbb{R}^{p\times(p-d)}$ be a matrix with columns $l_{d+1}, \ldots, l_p$. Since Σ_{ezz} is symmetric (the eigenvectors form an orthonormal basis), $P_0 := \Theta_0\Theta_0^T$ is a projection matrix associated with the null space of Σ_{ezz}. Then, with equation (5.30) and the fact that $P_0\Sigma_{ezz} = \Sigma_{ezz}P_0 = 0$, Lemma 5.9 shows that

$$\begin{aligned}\sum_{j=d+1}^{p} \hat{\delta}_j^2 =& \mathrm{tr}\{(B_n\Sigma_{ezz} + \Sigma_{ezz}B_n)P_0\} + \mathrm{tr}[\{B_nB_n + o_p(n^{-1/2})\Sigma_{ezz} + \Sigma_{ezz}o_p(n^{-1/2})\}P_0] \\ &- \mathrm{tr}\{(B_n\Sigma_{ezz} + \Sigma_{ezz}B_n)(\Sigma_{ezz}\Sigma_{ezz})^{\dagger}(B_n\Sigma_{ezz} + \Sigma_{ezz}B_n)P_0\} + o_p(n^{-1}).\end{aligned} \tag{5.31}$$

Since $P_0\Sigma_{ezz} = \Sigma_{ezz}P_0 = 0$, $P_0P_0 = P_0$ and trace operator is invariant under cyclic permutations, we derive that

$$\begin{aligned}\operatorname{tr}\{(B_n\Sigma_{ezz} + \Sigma_{ezz}B_n)P_0\} &= \operatorname{tr}(\Sigma_{ezz}P_0B_n) + \operatorname{tr}(P_0\Sigma_{ezz}B_n)\\ &= 0,\end{aligned}$$

and

$$\begin{aligned}&\operatorname{tr}[\{B_nB_n + o_p(n^{-1/2})\Sigma_{ezz} + \Sigma_{ezz}o_p(n^{-1/2})\}P_0]\\ =&\operatorname{tr}(B_nB_nP_0P_0) + \operatorname{tr}(P_0\Sigma_{ezz}o_p(n^{-1/2})) + \operatorname{tr}(P_0\Sigma_{ezz}o_p(n^{-1/2}))\\ =&\operatorname{tr}(P_0B_nB_nP_0).\end{aligned}$$

Similarly,

$$\begin{aligned}&\operatorname{tr}\{(B_n\Sigma_{ezz} + \Sigma_{ezz}B_n)(\Sigma_{ezz}\Sigma_{ezz})^{\dagger}(B_n\Sigma_{ezz} + \Sigma_{ezz}B_n)P_0\}\\ =&0 + \operatorname{tr}\{B_n\Sigma_{ezz}(\Sigma_{ezz}\Sigma_{ezz})^{\dagger}\Sigma_{ezz}B_nP_0\} + 0\\ =&\operatorname{tr}\{P_0B_n\Sigma_{ezz}(\Sigma_{ezz}\Sigma_{ezz})^{\dagger}\Sigma_{ezz}B_nP_0\}.\end{aligned}$$

Substituting the above results into equation (5.31), we obtain that

$$\begin{aligned}\sum_{j=d+1}^{p}\hat{\delta}_j^2 &= \operatorname{tr}(P_0B_nB_nP_0) - \operatorname{tr}(P_0B_n\Sigma_{ezz}(\Sigma_{ezz}\Sigma_{ezz})^{\dagger}\Sigma_{ezz}B_nP_0) + o_p(n^{-1})\\ &= \operatorname{tr}(P_0B_nP_0B_nP_0) + o_p(n^{-1})\\ &= \operatorname{tr}[(\Theta_0^TB_n\Theta_0)^2] + o_p(n^{-1}).\end{aligned} \tag{5.32}$$

As a result, the asymptotic distribution is the same as the asymptotic distribution of

$$\Delta_d^* = \frac{n}{2\operatorname{Var}(e)}\operatorname{tr}[(\Theta_0^TB_n\Theta_0)^2] = \frac{n}{2\operatorname{Var}(e)}\sum_{i,j}^{p-d}(\Theta_0^TB_n\Theta_0)_{i,j}^2, \tag{5.33}$$

where $(\Theta_0^TB_n\Theta_0)_{i,j}^2$ is the ijth elements of $\Theta_0^TB_n\Theta_0$.

Next, we evaluate $\Theta_0^T B_n \Theta_0$ to find the asymptotic distribution of Δ_d^*. Because, by definition, $\Sigma_{ezz}\Theta_0 = 0$ and $\beta_e = 0$, we have

$$\begin{aligned}
\Theta_0^T B_n \Theta_0 &= \Theta_0^T [\bar{C} - \bar{z}\beta_e^T - \beta_e^T \bar{z}^T - \frac{1}{2}\Xi\Sigma_{ezz} - \frac{1}{2}\Sigma_{ezz}\bar{C}]\Theta_0 \\
&= \Theta_0^T \bar{C}\Theta_0 \\
&= \frac{1}{n}\sum_i^n [e_i(\Theta_0^T z_i z_i^T \Theta_0 - \Theta_0^T\Theta_0)] - \Theta_0^T \Sigma_{ezz}\Theta_0 \\
&= \frac{1}{n}\sum_i^n [e_i(\Theta_0^T z_i z_i^T \Theta_0 - I_{p-d})].
\end{aligned} \tag{5.34}$$

We observe that $\Theta_0^T B_n \Theta_0$ is in fact an average of independent and identically distributed matrices with mean $\mathrm{E}[e_i(\Theta_0^T z_i z_i^T \Theta_0 - I_{p-d})] = \Theta_0^T \Sigma_{ezz}\Theta_0 = 0$. Thus, given the definition of W, we conclude that $\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i W_i$ is asymptotically normally distributed with mean 0 and variance $\mathrm{Var}(eW)$ by the multivariate Central Limit Theorem.

Finally, because

$$\Delta_d^* = \frac{n}{2\mathrm{Var}(e)}\sum_{i,j}^{p-d}(\Theta_0^T B_n \Theta_0)_{i,j}^2 = \frac{1}{2\mathrm{Var}(e)}\|\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i W_i\|^2$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i W_i \longrightarrow_d N(0, \mathrm{Var}(eW)),$$

following a similar argument to that of Proposition 4.11 gives the desired final result. □

Remark 5.10. We point out that the proof of the asymptotic behaviour of $\hat{\Delta}_{pHdres}(m)$ is in fact almost identical to the proof used by Li in proving the asymptotic distribution of the test statistic $\hat{\Delta}_{pHdy}(m)$ for pHdy. In Li's original proof for the asymptotic distribution of $\hat{\Delta}_{pHdy}(m)$, he implicitly assumed that $\Theta_0^T\beta = 0$. Since this assumption is in general not true, the proof is not always valid. However, in the pHdres case, because the regression is no longer of the response y but of e on z, the coefficient $\beta_e = \mathrm{Cov}(z, e)$ has to be zero. Consequently, $\Theta_0^T\beta_{ols} = 0$ will always hold and the asymptotic result for $\hat{\Delta}_{pHdres}(m)$ will always be true.

With the above proposition, we are now able to choose d using the statistic $\hat{\Delta}_d$. We give the detailed procedure below.

An algorithm for choosing the dimension d of S_{ezz}

1. Set $m = 0$.

2. Set $d = m$. Given d, form an estimate $\hat{W}_i$ of the matrix W_i as defined in equation (5.27) using $\hat{l}_{d+1}, \ldots, \hat{l}_p$ and $\hat{z_1}, \ldots, \hat{z_n}$.

3. Estimate $\mathrm{Var}(eW)$ by computing the variance matrix $\hat{\Sigma}_{eW}$ of vectors $\hat{e}_i\hat{W}_i$, $i = 1, \ldots, n$.

4. Compute the eigenvalues of $\hat{\Sigma}_{eW}$ and denote them as $\hat{\omega}_1 \geq \cdots \geq \hat{w}_{(p-d)(p-d+1)/2}$.

5. Calculate $\hat{\Delta}_{pHdres}(d)$ using equation (5.25). Then compute the p-value as $\Pr(\hat{C} > \hat{\Delta}_{pHdres}(d))$, using the result that the asymptotic distribution is the same as that of

$$\hat{C} = \frac{1}{2\mathrm{Var}(\hat{e})} \sum_{j=1}^{(p-d)(p-d+1)/2} \hat{\omega}_j \chi(1),$$

   where the $\chi(1)$'s are independent Chi-square distribution variables, each with one degree of freedom.

6. Compare the calculated p-value with the pre-determined cutoff value. If the p-value is larger than the pre-determined cutoff value, then $d = m$ is the final estimate. If not, proceed as if $d > m$ holds. Let $m = m + 1$ and return to step two.

5.2 Conclusion

In this chapter, we have introduced two methods based on principal Hessian directions: pHdy and pHdres. pHdy was developed first and is more straightforward, as we directly use the average Hesssion matrix of the regression function to estimate the central subspace. However, there are several serious drawbacks of pHdy: pHdy requires x to be normally distributed; it is not effective in detecting linear trends, and its asymptotic analysis may not always hold. Due to these drawbacks of pHdy, pHdres, a modified version of pHdy, was proposed. pHdres first removes the linear trend from the response variable using OLS. Then it uses the average Hessian matrix of the regression of residual on z to estimate $S_{e|z}$. Finally, it combines the OLS estimate with the estimate of $S_{e|z}$ to approximate the central subspace $S_{y|z}$. The application of OLS in the first step makes pHdres more effective in detecting linear trends and because it applies pHdy to the residual, the issue causing the invalidity of the asymptotic analysis for pHdy is avoided in pHdres's asymptotic analysis. We will test the effectiveness of pHdres in the next chapter.

Chapter 6

Simulations

In this chapter, we conduct a simulation study to compare the methods SIR, SAVE and pHdres. We do not include the pHdy method, because it is not always reliable. We use the **dr** package in R(R Core Team, 2015), first documented in Weisberg (2002) and revised in Weisberg (2015). The **dr** package was specifically developed for dimension reduction regression and it has implemented SIR, SAVE, pHdres, and IRE (not included in the thesis). In terms of choosing the dimension d of $S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\}$ (SIR), $S(\Sigma_{save})$ (SAVE) and $S(\Sigma_{ezz})$(pHdres), **dr** package used the same methods as we introduced in previous Chapters. To evaluate and compare SIR, SAVE and pHdres' performance in recovering the central subspace, three different examples will be studied.

6.1 Example One:

We start with the simplest case: a linear model. Assume the true model is

$$y_1 = x_1 + x_2 + x_3 + \epsilon, \tag{6.1}$$

where $x = (x_1, \ldots, x_5)^T$ follows a multivariate standard normal distribution, ϵ is normally distributed and the x_i's and ϵ are independent. In this case, the central subspace exists and is spanned by the vector $l_1 = (1, 1, 1, 0, 0)^T$.

We simulated 400 data points from the model (6.1). Since x is multivariate normally distributed, we can apply SIR, SAVE and pHdres to recover the dependence relationship between the covariates and the response variable. We ran the **dr** function with "method" equal to “sir”, “save”, “phdres” respectively and we repeated this on 1000 samples. For SIR and SAVE, the number of slices h was set to 20. For the tests choosing the dimension of $S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\}$ (SIR), $S(\Sigma_{save})$ (SAVE) and $S(\Sigma_{ezz})$ (pHdres), we chose the significance level to be 0.01 for better comparison. To facilitate the discussion, let k be the dimension of the final estimate of the central subspace for each method.

We first look at the values of k chosen by tests for all simulations. Before we study the results, it is important to note that the value of k for pHdres may be overestimated. We recall that pHdres estimates the central subspace using the formula $S_{y|z} = S(\beta) + S_{e|z}$ where $\beta = \mathrm{Cov}(z, y)$. In the algorithm of pHdres, we first compute β, which we record as the first possible direction. Then we separately estimate the basis vectors that span $S_{e|z}$. Assume we estimate the basis vectors of $S_{e|z}$ to be $\{\hat{\iota}_1, \ldots, \hat{\iota}_d\}$. We compute k as $k = 1 + d$, the sum of $\dim(S(\beta))$ and $\dim(S_{e|z})$. Since it is possible that $\beta \in S_{e|z}$, k is likely to be overestimated by one. As a side note, we mention that the reason we do not directly orthogonalise the set $\{\hat{\beta}, \hat{\iota}_1, \ldots, \hat{\iota}_r\}$ to find a basis for $S_{y|z}$ and then determine the value of k is that this orthogonalisation introduces additional errors and consequently produces misleading results in most simulations. For SIR, $k = d = \dim[S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\}]$ and for SAVE, $k = d = \dim(S(\Sigma_{save}))$.

Methods	$k = 0$	$k = 1$	$k = 2$
SIR	0	991	9
SAVE	29	969	2
pHdres	0	987	13

TABLE 6.1: Value of k over 1000 simulations($\alpha = 0.01$)

We now study the table above. In this example, the desired value of k is one, because the central subspace is one dimensional. From the table, we see that all three methods performed satisfactorily, choosing $k = 1$ in at least 950 out of 1000 simulations. SIR performed the best with the highest success rate (991/1000) in choosing $k = 1$ while SAVE has the lowest success rate (987/1000). Because $S_{e|z}$ is a trivial space in this example, pHdres did not overestimate k in this case.

Since the dimension of the central subspace is one, we computed the mean and standard deviation (sd) of the components of the first computed direction (standardised), over 1000 simulations for each method. We denote the first computed direction as $\hat{l}_1 = (\hat{l}_{11}, \ldots, \hat{l}_{15})$, because it estimates l_1.

Methods		$\hat{l}_{11}$	$\hat{l}_{12}$	$\hat{l}_{13}$	$\hat{l}_{14}$	$\hat{l}_{15}$
SIR	mean	0.577	0.577	0.577	0.000	0.000
	sd	(5.81e-03)	(5.57e-03)	(5.56e-03)	(6.85e-03)	(6.73e-03)
SAVE	mean	0.577	0.577	0.577	0.000	0.000
	sd	(5.96e-03)	(5.98e-03)	(5.96e-03)	(7.45e-03)	(7.26e-03)
pHdres	mean	0.577	0.577	0.577	0.000	0.000
	sd	(2.40e-04)	(2.40e-04)	(2.27e-04)	(2.88e-04)	(2.77e-04)

Table 6.2: Means and standard deviations of $\hat{l}_1 = (\hat{l}_{11}, \dots, \hat{l}_{15})$

The above table provides us with an general idea of the results of each method. From the table, it seems all three methods have been successful in estimating the true direction $l_1 = (1, 1, 1, 0, 0)^T$, as the mean of $\hat{l}_1$ for each method is in the same direction as l_1. When the standard deviation is considered, pHdres performed the best with the smallest sd for each component while SIR and SAVE performed about the same. To better understand the performances of the three methods, we looked at cosine of the angle between an estimated direction and the true direction. Denote the angle between an estimated direction and the true direction as θ $(-180 \leq \theta \leq 180)$. Since two estimates perform the same in estimating the true direction when their associated angles are the opposite of each other, we computed $|\cos(\theta)|$ for each simulation. We summarize the results below.

Methods	mean($\|\cos(\theta)\|$)	sd($\|\cos(\theta)\|$)
SIR	0.9999061	7.670389e-05
SAVE	0.9998925	8.543894e-05
pHdres	0.9999998	1.184175e-07

Table 6.3: Means and standard deviations of $|\cos(\theta)|$

We know that the smaller the absolute value of the angle θ, the better its related estimated direction is. Thus, a method performs well when $|\cos(\theta)|$ of its estimates are close to $\cos(0) = 1$. From Table(6.3), we observe that all three methods performed satisfactorily. All three means of $|\cos(\theta)|$ are very close to 1 and variances of $|\cos(\theta)|$ are close to 0. pHdres performed the best with the highest mean and the smallest standard deviation. The boxplots (6.1) show similar results. The performance of SIR and SAVE is about the same.

Finally, we evaluate the efficiency of all three methods by comparing the time each method took to run 1000 simulations. It is clear that SIR is a much more efficient method than SAVE and pHdres. SIR only took around one fifth of the time required by pHdres and one ninth of the time required by SAVE.

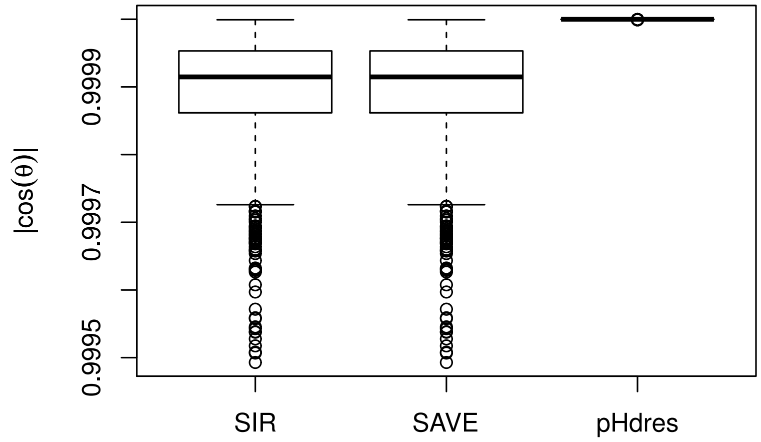

FIGURE 6.1: Boxplots of $|\cos(\theta)|$ for SIR, SAVE and pHdres

	SIR	SAVE	pHdres
Time(seconds)	20.17	189.76	99.23

TABLE 6.4: Time required for 1000 simulations

Overall, all three methods are effective in detecting the central subspace. pHdres provided the best estimates. This is not surprising, as pHdres uses ordinary least square regression to identify linear trends. SIR and SAVE performed similarly but SIR is the most efficient method.

6.2 Example Two:

We next consider a relatively complicated example. Consider the true model

$$y_2 = \frac{x_1}{(1 + (x_2 + 2)^2)} + \epsilon. \tag{6.2}$$

Again assume $x = (x_1, \ldots, x_5)^T$ follows a multivariate standard normal distribution, ϵ is normally distributed and the x_i's and ϵ are independent. In this case, the central subspace exists and is spanned by the vectors $l_1 = (1, 0, 0, 0, 0)^T$ and $l_2 = (0, 1, 0, 0, 0)^T$.

To be consistent, we simulated 400 data points from model (6.2). For SIR and SAVE, we set the number of slices h to 20. For the tests choosing the dimension of $S\{\text{Var}[\text{E}(z|\tilde{y})]\}$

(SIR), $S(\Sigma_{save})$ (SAVE) and $S(\Sigma_{ezz})$ (pHdres), we used $\alpha = 0.01$ to be consistent. For each method, we repeated the simulation 1000 times. Results are summarized below.

Methods	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
SIR	0	0	987	13	0
SAVE	59	932	9	0	0
pHdres	0	0	0	989	11

Table 6.5: Value of k over 1000 simulations($\alpha = 0.01$)

Since the central subspace is spanned by the vectors l_1 and l_2, the desired value of k is 2. We observe that SIR is the only method that chose $k = 2$ in most (987 out of 1000) of its simulations. In terms of the method SAVE, the vast majority (991/1000) of simulations chose the value $k = 1$; only 9 simulations chose the value $k = 2$. We will explore this below when studying all the p-values for SAVE. pHdres overestimated the value of k in all simulation. As we mentioned in the previous example, this overestimation is likely to be caused by the fact that $\beta = \text{Cov}(z, y) \in S_{e|z}$. To find out, we will study all first three directions computed in each simulation by pHdres. If it is true that $\beta = \text{Cov}(z, y) \in S_{e|z}$, the first direction (the vector β) should be contained in the space spanned by the second and the third directions (the basis vectors of $S_{e|z}$).

To get a general idea of the outputs for each method, we list the means and standard deviations (sd) of the components of the estimated directions (standardised) for the 1000 simulations.

SIR		$\hat{l}_{i1}$	$\hat{l}_{i2}$	$\hat{l}_{i3}$	$\hat{l}_{i4}$	$\hat{l}_{i5}$
$i = 1$	mean	0.998	-0.002	-0.001	0.002	-0.000
	sd	(2.12e-03)	(5.30e-02)	(2.51e-02)	(2.40e-02)	(2.40e-02)
$i = 2$	mean	-0.002	0.990	0.000	0.006	0.001
	sd	(7.29e-02)	(6.67e-03)	(6.73e-02)	(6.53e-02)	(7.00e-02)

Table 6.6: Means and standard deviations of computed directions (standardised) by SIR

SAVE		$\hat{l}_{i1}$	$\hat{l}_{i2}$	$\hat{l}_{i3}$	$\hat{l}_{i4}$	$\hat{l}_{i5}$
$i = 1$	mean	0.998	-0.002	-0.001	0.000	-0.001
	sd	(1.85e-03)	(4.94e-02)	(2.13e-02)	(2.25e-02)	(2.18e-02)
$i = 2$	mean	-0.002	0.968	0.004	-0.003	0.001
	sd	(6.78e-02)	(6.03e-02)	(1.26e-01)	(1.41e-01)	(1.37e-01)

Table 6.7: Means and standard deviations of computed directions (standardised) by SAVE

From above tables, we see that the two directions estimated by SIR are basically in the same directions as that of $l_1 = (1, 0, 0, 0, 0)^T$ and $l_2 = (0, 1, 0, 0, 0)^T$. SAVE provides

pHdres		$\hat{l}_{i1}$	$\hat{l}_{i2}$	$\hat{l}_{i3}$	$\hat{l}_{i4}$	$\hat{l}_{i5}$
$i = 1$	mean	0.995	-0.001	-0.003	0.002	0.000
	sd	(3.57e-03)	(6.51e-02)	(4.05e-02)	(3.98e-02)	(4.06e-02)
$i = 2$	mean	0.701	0.698	-0.003	0.005	0.000
	sd	(3.41e-02)	(3.41e-02)	(7.94e-02)	(8.09e-02)	(7.85e-02)
$i = 3$	mean	0.698	-0.701	-0.002	-0.002	0.004
	sd	(3.43e-02)	(3.44e-02)	(8.04e-02)	(8.10e-02)	(7.94e-02)

Table 6.8: Means and standard deviations of computed directions (standardised) by pHdres

similar results. Although the inference tests indicate that it is highly likely only the first direction computed is in the central subspace, Table (6.7) shows that the second direction computed by SAVE is in the central subspace as well. The first and second directions computed by SAVE look like good estimates of l_1 and l_2 respectively. Fairly different results are provided by pHdres. From Table(6.10), the three directions estimated by pHdres are basically in the directions of $(1, 0, 0, 0, 0)^T$, $(1, 1, 0, 0, 0)^T$ and $(1, -1, 0, 0, 0)^T$. These results suggest that $S(\beta)$ is spanned by $(1, 0, 0, 0, 0)^T$ and $S_{e|z}$ is spanned by $(1, 1, 0, 0, 0)^T$ and $(1, -1, 0, 0, 0)^T$. If this is true, we observe that $S(\beta) \subset S_{e|x}$ and $S_{y|x} = S_{e|x} = S((1, 0, 0, 0, 0)^T, (0, 1, 0, 0, 0)^T)$. Then, $S(\beta) \subset S_{e|x}$ explains the overestimation of k and the first three directions estimated by pHdres provide a good estimate of the central subspace.

To better understand the estimated directions, we also computed the absolute value of the cosine of angles between the estimated directions and the true directions l_1, l_2. Denote the angle between l_1 and its estimate as θ_1 and the angle between l_2 and its estimate as θ_2. Also let $l_3 = (1, 1, 0, 0, 0)^T$ and $l_4 = (1, -1, 0, 0, 0)^T$. For pHdres, to test our conjecture, we computed the cosines of angles between the first estimated direction and l_1, the second estimated direction and l_3, and the third estimated direction and l_4. We refer to these angles as θ_1, θ_3 and θ_4.

Methods		$(\lvert\cos(\theta_1)\rvert)$	$(\lvert\cos(\theta_2)\rvert)$	$(\lvert\cos(\theta_3)\rvert)$	$(\lvert\cos(\theta_4)\rvert)$
SIR	mean	0.998	0.990		
	sd	(2.18e-03)	(6.66e-03)		
SAVE	mean	0.998	0.968		
	sd	(1.85e-03)	(6.04e-02)		
pHdres	mean	0.995		0.989	0.989
	sd	(3.57e-03)		(8.34e-03)	(9.26e-03)

Table 6.9: Means and standard deviations of $|\cos(\theta_i)|$, $i = 1, 2, 3, 4$

Due to the similar results produced by SIR and SAVE, we compare their estimates for l_1 and l_2. We look at the distributions of $|\cos(\theta_1)|$, $|\cos(\theta_3)|$ and $|\cos(\theta_4)|$ for pHdres

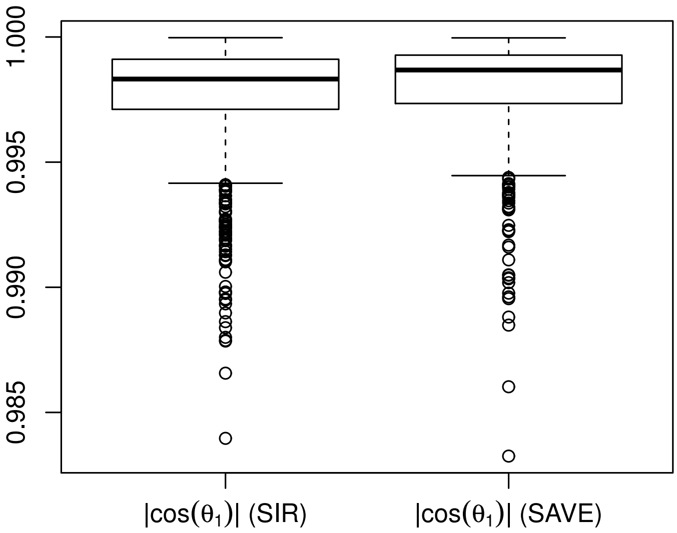

FIGURE 6.2: Boxplots of $|\cos(\theta_1)|$ for SIR and SAVE

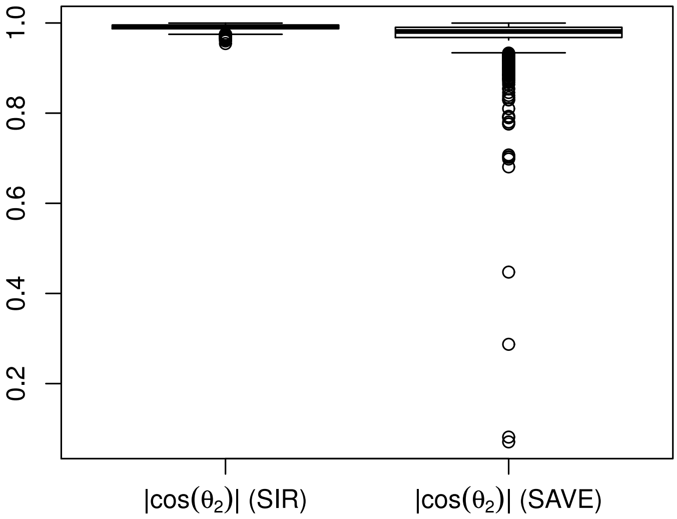

FIGURE 6.3: Boxplots of $|\cos(\theta_2)|$ for SIR and SAVE

separately.

From Table 6.9, Figure 6.2 and Figure 6.3, we see that both SIR and SAVE estimated l_1 and l_2 well with means close to 1 and standard deviations close to 0. To be more specific,

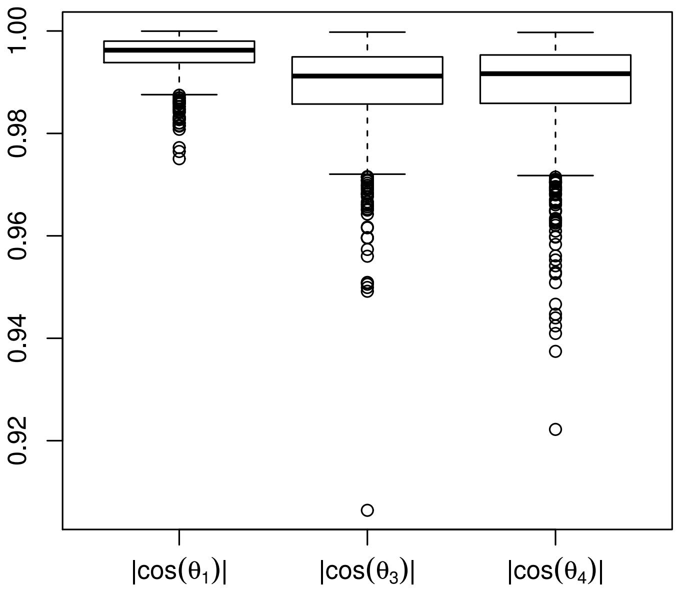

FIGURE 6.4: Boxplots of $|\cos(\theta_1)|$, $|\cos(\theta_3)|$,$|\cos(\theta_4)|$ for pHdres

SIR performed about the same as SAVE in estimating l_1, but much better than SAVE in estimating l_2. We observe from Figure 6.3 that $|\cos(\theta_2)|$ for SIR has less spread than SAVE. For SAVE, although the majority of $|\cos(\theta_2)|$ are close to 1, there are many cases with the value of $|\cos(\theta_2)|$ close to 0 instead. We hence conclude that SIR performed better in this example than SAVE. Nevertheless, the first two directions computed by SAVE still gave satisfactory estimates of l_1 and l_2.

For pHdres, we see from Figure 6.4 that all three distributions of $|\cos(\theta_1)|$, $|\cos(\theta_3)|$,$|\cos(\theta_4)|$ are left skewed with variance close to 0. For each distribution, even the smallest value is above 0.90. Therefore, pHdres estimates l_1, l_3, l_4 well. Consequently, pHdres gives good estimates of the central subspace.

Finally, we plot the values of $|\cos(\theta_1)|$ and $|\cos(\theta_2)|$ against their corresponding p-values for SAVE to examine the power of its test for choosing k.

Figure 6.5 shows the test of SAVE is effective in rejecting the null hypothesis $H_0 : k = 0$. However, Figure (6.6) indicates a serious problem. We observe that although nearly all $|\cos(\theta_2)|$ are close to 1, their corresponding p-values are nearly uniformly distributed. SAVE seems to be ineffective in testing the hypothesis $k = 1$ against $k > 1$.

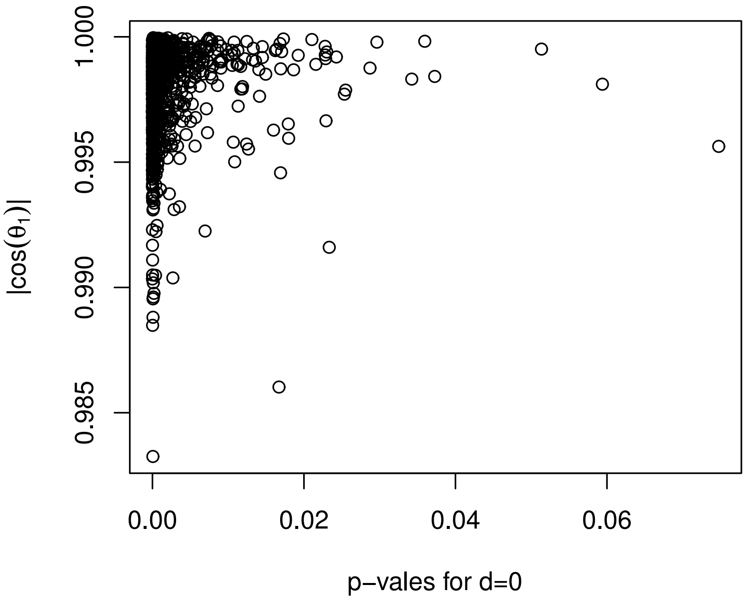

FIGURE 6.5: $|\cos(\theta_1)|$ vs p-values for SAVE

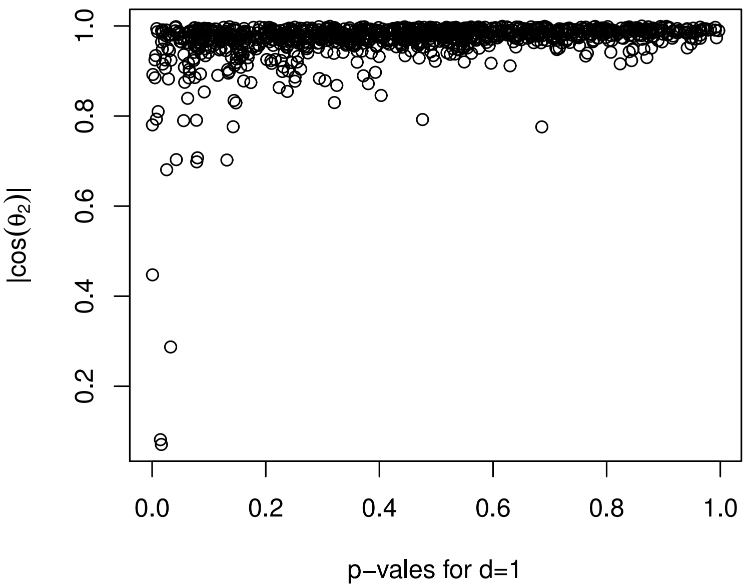

FIGURE 6.6: $|\cos(\theta_2)|$ vs p-values for SAVE

Finally, we list the time required by each method to run 1000 simulations. The results are similar to those of example one. SIR is the most efficient method while SAVE took the longest time.

	SIR	SAVE	pHdres
Time(seconds)	21.94	189.21	95.09

Table 6.10: Time required for 1000 simulations

Overall, we see that all three methods gave satisfactory estimates of the central subspace. Apart being the most time-efficient, SIR gave the best and most straightforward results. pHdres is also very effective but its results require more interpretation. SAVE performed well in estimating l_1, but less well when estimating l_2. Still, the average estimates from 1000 simulations of SAVE are satisfactory. Finally, we point out that SAVE does not seem to be effective in choosing the correct value of k.

6.3 Example Three:

In our final example, we consider a model that includes both a linear part and a quadratic part to test each method's power in detecting linear and nonlinear trends. Let the true model be

$$y_3 = (x_1 + x_2) + (x_3 + x_4)^2 + \epsilon. \tag{6.3}$$

Here, $x = (x_1, \ldots, x_5)^T$ has a multivariate standard normal distribution with mean 0, ϵ is normally distributed and is independent of x. Again, we simulated 400 data points using this model and repeated the simulation 1000 times. For SIR and SAVE, we set the number of slices h to 20. For the tests choosing the dimension of $S\{\mathrm{Var}[\mathrm{E}(z|\tilde{y})]\}$ (SIR), $S(\Sigma_{save})$ (SAVE) and $S(\Sigma_{ezz})$ (pHdres), we used $\alpha = 0.01$. In this case, the central subspace exists and is spanned by the vectors $l_1 = (1, 1, 0, 0, 0)^T$ and $l_2 = (0, 0, 1, 1, 0)^T$.

Methods	$k = 1$	$k = 2$	$k = 3$
SIR	978	21	1
SAVE	906	93	1
pHdres	0	984	16

Table 6.11: Value of k over 1000 simulations($\alpha = 0.01$)

We start by looking at the values of k chosen by the methods. In this case, the desired value of k is 2. Both SIR and SAVE estimated k to be smaller than 2 in the majority (over 900 out of 1000) of simulations. This result is expected from SIR, as we showed in Chapter 4 that SIR is ineffective in detecting the quadratic part due to the symmetry pattern. We expect that SIR is only able to provide estimates for l_1. The small k value provided by SAVE is unexpected, as SAVE was specifically developed to make up for the defect of

SIR in diagnosing symmetry dependence. SAVE should have been able to estimate both l_1 and l_2. From previous examples, we may suspect this unsatisfactory performance of SAVE is caused by its problems choosing k. We will explore this more below. Among all three methods, pHdres is the only method that gave the desired value of k. It estimated $k = 2$ in 984 out of 1000 simulations. However, because it is possible that pHdres may overestimate the value of k, we need to examine the results to evaluate the performance of pHdres.

We next look at the means and standard deviations (sd) of the components of standardised estimates computed in 1000 simulations for each method.

Methods		$\hat{l}_{11}$	$\hat{l}_{12}$	$\hat{l}_{13}$	$\hat{l}_{14}$	$\hat{l}_{15}$
SIR	mean	0.702	0.701	0.003	-0.001	-0.002
	sd	(1.56e-03)	(1.60e-03)	(3.81e-03)	(4.18e-03)	(3.80e-03)
SAVE*	mean	0.692	0.689	0.007	0.004	-0.002
	sd	(8.02e-02)	(8.01e-02)	(1.10e-01)	(1.05e-01)	(1.07e-01)
pHdres	mean	0.684	0.678	-0.008	-0.007	0.000
	sd	(6.93e-02)	(7.15e-02)	(1.66e-01)	(1.60e-01)	(9.80e-02)

TABLE 6.12: Means and standard deviations of $\hat{l}_1 = (\hat{l}_{11}, \dots, \hat{l}_{15})$

Methods		$\hat{l}_{21}$	$\hat{l}_{22}$	$\hat{l}_{23}$	$\hat{l}_{24}$	$\hat{l}_{25}$
SIR	mean	-0.015	0.013	0.011	-0.011	-0.007
	sd	(3.59e-01)	(3.57e-01)	(4.93e-01)	(5.08e-01)	(4.92e-01)
SAVE*	mean	-0.003	-0.005	0.698	0.701	-0.002
	sd	(7.98e-02)	(7.81e-02)	(4.92e-02)	(4.97e-02)	(6.70e-02)
pHdres	mean	-0.002	-0.002	0.701	0.703	0.000
	sd	(5.88e-02)	(5.94e-02)	(4.42e-02)	(4.42e-02)	(5.59e-02)

TABLE 6.13: Means and standard deviations of $\hat{l}_2 = (\hat{l}_{21}, \dots, \hat{l}_{25})$

We put a star over the method SAVE because, unlike SIR and pHdres, SAVE computed an estimate for l_2 first and then an estimate for l_1. In other words, SAVE concluded that the estimate for l_2 was associated with a larger eigenvalue that the estimate for l_1. Since l_2 corresponds to the quadratic part, SAVE might be more sensitive to the nonlinear trend than the linear trend.

From Table 6.12, we see that all three methods estimated l_1 satisfactorily with all three means basically in the same direction as l_1. However, we observe that SIR failed in estimating l_2. The mean of its estimations for l_2 is close to a zero vector. This is consistent with the results of SIR for choosing k. On the other hand, both SAVE and pHdres performed well in estimating l_2. Therefore, pHdres did not overestimate the value of k but the test of SAVE is not effective in choosing the correct value for k.

We next computed the cosine of the angles between the estimates and the true directions. Again, denote the angle between l_1 and its estimate as θ_1 and the angle between l_2 and its estimate as θ_2.

Methods		$(\lvert\cos(\theta_1)\rvert)$	$(\lvert\cos(\theta_2)\rvert)$
SIR	mean	0.992	0.417
	sd	(5.85e-03)	(2.84e-01)
SAVE	mean	0.976	0.989
	sd	(2.15e-02)	(8.47e-03)
pHdres	mean	0.963	0.993
	sd	(3.22e-02)	(5.32e-03)

TABLE 6.14: Means and standard deviations of $|\cos(\theta_1)|$, $|\cos(\theta_2)|$

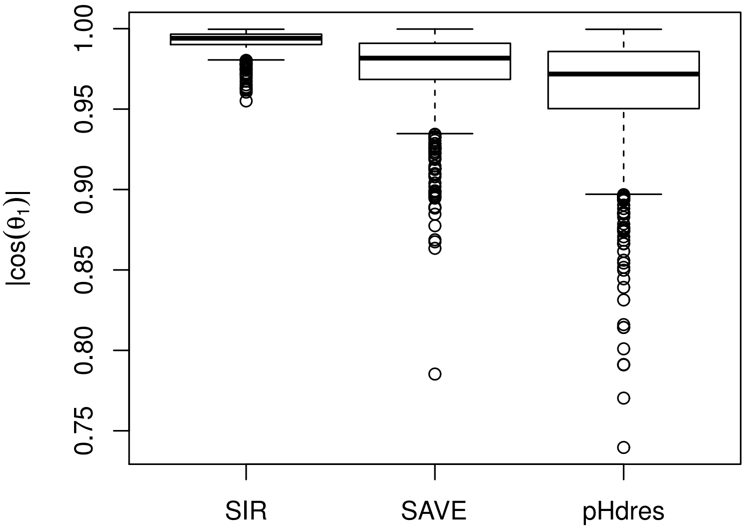

FIGURE 6.7: Boxplots of $|\cos(\theta_1)|$ for SIR, SAVE and pHdres

We now examine the distributions of $|\cos(\theta_1)|$, $|\cos(\theta_2)|$. We first look at the distribution of $|\cos(\theta_1)|$. From Table 6.14, all three methods gave satisfactory estimates for l_1 with all means above 0.95 and variances smaller than 3.3e-02. Among the three methods, SIR gave the best estimates and SAVE gave the second best estimates. The boxplot 6.7 indicates similar results. The $|\cos(\theta_1)|$ for SIR has the smallest box width and shortest whisker lengths. Although the estimates from pHdres are satisfactory in general, they are more widely spread and contain more small values.

In terms of the distributions of $|\cos(\theta_2)|$, we see from the left sub-figure of Figure 6.8 that SIR failed in estimating l_2. SAVE and pHdres gave satisfactory estimates for l_2.

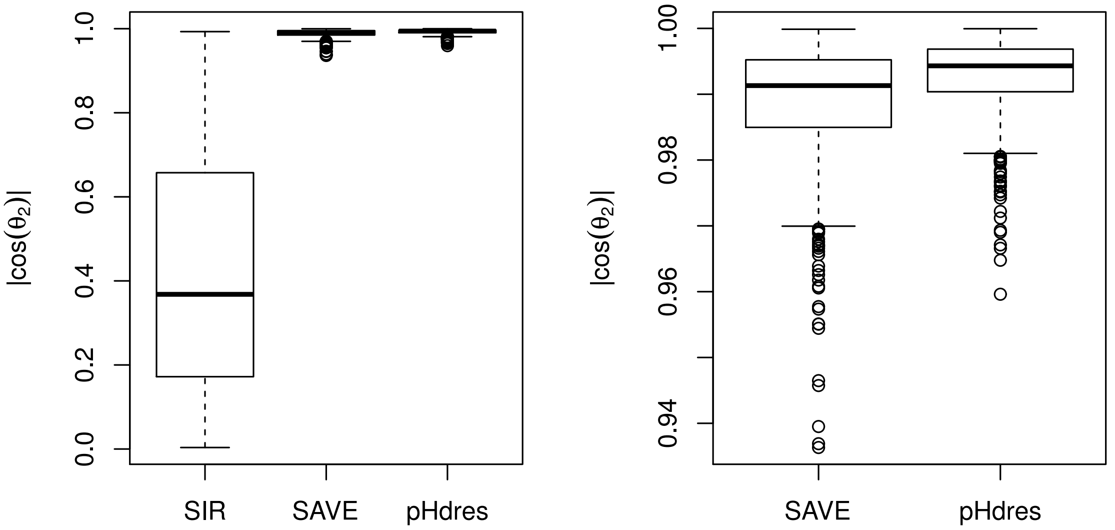

Figure 6.8: Boxplots of $|\cos(\theta_2)|$ for SIR, SAVE and pHdres

When compared with each other (see the right sub-figure 6.8), we conclude that pHdres provided better estimates for l_2 than SAVE. The distribution of $|\cos(\theta_2)|$ for pHdres has higher mean and smaller variance than SAVE. Also, the $|\cos(\theta_2)|$ for estimates of pHdres are more tightly distributed than for SAVE.

Similar to the previous example, although SAVE is able to provide good estimates for both l_1 and l_2, the test suggested that SAVE should only estimate a one dimensional subspace of the central subspace. We plotted $|\cos(\theta_i)|$ against the p-values for $i = 1, 2$ to explore this situation further.

We see that the tests have been useful in suggesting that the first directions computed (estimates for l_2) are in the central subspace. However, the test of SAVE failed to reject the null hypothesis $H_0 : k = 1$ in most simulations.

Finally, the relative times required for each method are consistent with the previous examples. SIR is the most efficient method while SAVE is the most time-consuming method.

	SIR	SAVE	pHdres
Time(seconds)	20.33	186.98	95.02

Table 6.15: Time required for 1000 simulations

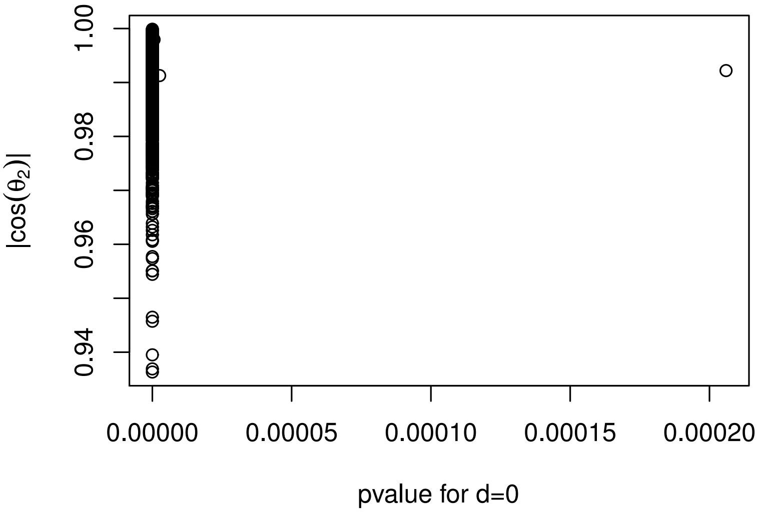

FIGURE 6.9: $|\cos(\theta_2)|$ vs p-values for SAVE

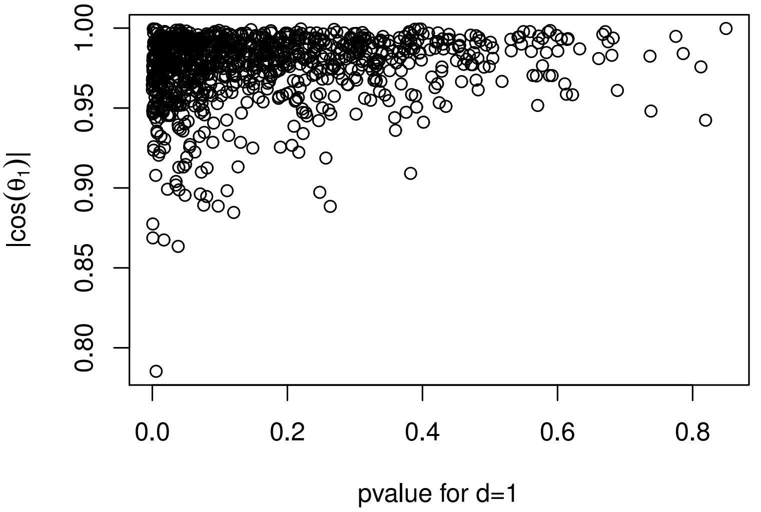

FIGURE 6.10: $|\cos(\theta_1)|$ vs p-values for SAVE

The example supports our previous claim that SIR is unable to diagnose symmetry dependence. Since $(x3+x4)^2$ is symmetrically distributed, SIR failed to detect the direction $l_2 = (0, 0, 1, 1, 0)$. Both SAVE and pHdres successfully estimated both l_1 and l_2. We also find that SAVE is more effective in detecting the nonlinear trend than the linear trend in this example; SAVE gave estimates for l_2 first.

6.4 Conclusion

Based on all three examples, we recommend SIR when we are sure there is no symmetry dependence and pHdres in other cases. When there is no symmetry dependence between the covariates and the response variable, SIR gives the best estimates and is the most efficient method. When it is not clear whether symmetry dependence exits, pHdres is a good option. pHdres provides satisfactory estimates within a moderate period of time in all three examples. However, since pHdres estimates the central subspace by estimating two subspace separately, we may need to be careful when interpreting results from pHdres. Finally, although SAVE also provides good estimates for all three examples, SAVE has some serious drawbacks. Firstly, it is a very time-consuming method. In all three examples, it takes approximately nine times and twice as much time as SIR and pHdres respectively. Secondly and more importantly, the method currently used by SAVE for choosing k is not reliable.

Bibliography

Basu, D. and Pereira, C. (1983). Conditional independence in statistics. *Sankhya: The Indian journal of Statistics*, 45:324–337.

Becker, C. and Gather, U. (2007). A note on the choice of the number of slices in sliced inverse regression. *Technical Report, Komplexitatsreduktion in Multivariaten Datenstrukturen, Universitat Dortmund*, 475.

Bura, E. and Cook, R. (2001). Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association*, 96:996–1003.

Cambanis, S., Huang, S., and Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 7:368–385.

Carroll, R. and Li, K. (1992). Measurement error regression with unknown link: dimension reduction and data visualization. *Journal of the American Statistical Association*, 87:1040–1050.

Chiaromonte, F. and Cook, R. (2002). Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, 54(4):768–795.

Cook, R. (1994a). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189.

Cook, R. (1994b). Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the Section on Physical and Engineering Sciences*, pages 18–25.

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

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

Cook, R. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, 32(3):1062–1092.

Cook, R. (2009). *Regression graphics: Ideas for studying regressions through graphics*, volume 482. John Wiley and Sons.

Cook, R. and Lee, H. (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association*, 94:1187–1200.

Cook, R. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332.

Cook, R. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis. *Australian and New Zealand Journal of Statistics*, 43:147–199.

Cox, D. and Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–275.

Dawid, A. (1979a). Conditional independence in statistical theory(with discussions). *Journal of the Royal Statistical Society*, 41:1–31.

Dawid, A. (1979b). Some misleading arguments involving conditional independence. *Journal of the Royal Statistical Society*, 41:249–252.

Duan, N. and Li, K. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, 19(2):505–530.

Eaton, M. (1983). *Multivariate statistics: A vector space approach*, volume 198. John Wiley and Sons.

Eaton, M. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis*, 20(2):272–276.

Eaton, M. and Tyler, D. (1991). On Wieland's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Annals of Statistics*, 19:260–271.

Eaton, M. and Tyler, D. (1994). The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis. *Journal of Multivaraite Analysis*, 50:238–264.

Frahm, G. (2004). *Generalized elliptical distributions: theory and applications.* Universitat zu Koln.

Hsing, T. and Carroll, R. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 20(2):1041–1061.

Hult, H. and Lindskog, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34(3):587–608.

Johnson, M. (1987). *Multivariate Statistical Simulation.* New York: Wiley.

Kato, T. (1976). *Perturbation Theory for Linear Operators (2nd ed.).* Berlin: Springer.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 32(4):419–430.

Landsman, Z. and Neslehova, J. (2008). Stein's Lemma for elliptical random vectors. *Journal of Multivariate Analysis*, 99(5):912–927.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of American Statistician Association*, 102:997–1008.

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

Li, K. (1992). On the principal Hessian directions for data visualization and dimension reduction: Another application of Stein's Lemma. *Journal of the American Statistical Association*, 87:1025–1040.

Li, K. (2000). Sampling properties of SIR. pages 28–39.

Li, K. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052.

Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, 81:134–150.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Saracco, J. (1997). An asymptotic theory for sliced inverse regression. *Communications in Statistics Theory and Methods*, 26(9):2141–2171.

Schmidt, R. (2002). Tail dependence for elliptically contoured distributions. *Mathematical Methods of Operations Research*, 55:301–327.

Schott, J. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89:141–148.

Shao, Y., Cook, R., and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika*, 94:285–296.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151.

Tyler, D. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, 9:725–736.

Weisberg, S. (2002). Dimension reduction regression in R. *Journal of Statistical Software*, 7(2).

Weisberg, S. (2015). The dr package. pages 1–28.

Zeng, P. and Zhu, Y. (2010). An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis*, 101(1):271–290.

Zhu, L. and Ng, K. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, 5:727–736.

Zhu, L., Ohtaki, M., and Li, Y. (2007). On hybrid methods of inverse regression based algorithms. *Computational Statistics and Data Analysis*, 51(5):2621–2635.