MASTER'S THESIS

Eero Liski

On Sliced Inverse Regression

UNIVERSITY OF TAMPERE
Department of mathematics and statistics
Statistics

September 2009

University of Tampere

Department of mathematics and statistics

Liski, Eero: On Sliced Inverse Regression

Master's Thesis, 61 p., 7 appendix pages

Statistics

September 2009

Abstract

In statistics, dimension reduction is a method to reduce the number of variables, which will then be considered in the future analysis of the data. Often the new variables are just suitably chosen linear combinations of the original variables $X_1, ..., X_p$. Well known dimension reduction techniques are principal component analysis (PCA), factor analysis (FA) and independent component analysis (ICA), for example. Sliced inverse regression (SIR) is a dimension reduction method proposed by Li (1991). In sliced inverse regression it is assumed that the new variables are used to explain the variation of a response variable Y, and this is taken into account in the dimension reduction process. The inverse regression function is used to find an estimate of the so called central dimension reduction subspace (central DRS). This thesis presents main theoretical results behind SIR and reports the results of an extensive simulation study.

In our simulation study, the performance of three dimension reduction methods, sliced inverse regression, sliced average variance estimate (SAVE) and principal hessian directions (PHD), are compared under various experimental settings. We consider four different choices of dimensions of a vector-valued explanatory variable $\mathbf{X}$, four choices of distributions of $\mathbf{X}$, four different choices of sample sizes, seven different models for the dependence, and two different levels of noise.

Finally, a real data set from a study on coronary heart disease risk factors is analyzed using the three different dimension reduction techniques.

Keywords: inverse regression, dimension reduction, dimension reduction subspace, conditional independence

Contents

Notation and Symbols

$'$	transpose
Ω	sample space
$\in$	belongs to
$\subseteq$	subset of (or equal to)
$\|\|\cdot\|\|$	vector norm
$\perp$	orthogonal complement
$\oplus$	direct sum of subspaces
$\square$	end of proof
∂	partial derivative
$\sim$	distributed as
$rank(\cdot)$	rank of a matrix
i.e.	*id est*, that is
$\mathbf{I}_p$	$p \times p$ identity matrix
$\mathbf{P}$	projector matrix
$\mathscr{C}(\cdot)$	column space
$\mathscr{C}_{Y\|\mathbf{X}}$	central dimension reduction subspace (central DRS)
$\mathscr{C}_{Y\|\mathbf{Z}}$	standardized central dimension reduction subspace (standardized central DRS), where $E(\mathbf{Z}) = \mathbf{0}$ and $cov(\mathbf{Z}) = \mathbf{I}_p$
$\mathscr{C}_{E(\mathbf{X}\|Y)}$	inverse regression subspace
$\mathscr{C}_{E(\mathbf{Z}\|Y)}$	standardized inverse regression subspace

1 Introduction

Dimensionality is a major concern in analyzing large data sets. Dimension reduction for regression, as pioneered by such authors as Li & Duan (1989), Duan & Li (1991), Li (1991, 1992) and Cook & Weisberg (1991), is aimed at reducing the dimension of a vector-valued explanatory variable $\mathbf{X} = (X_1, \ldots, X_p)'$, while preserving its regression relation with a response Y.

Searching for dependencies between variables in the data often begins with looking at scatter plots of the data set. Computers are well equipped to handle one-, two- and three-dimensional graphics, but visualizing high dimensional data sets becomes difficult. If there is no specific prior information about the dependence between Y and $\mathbf{X}$, we could proceed, for example, with parametric regression, say linear regression. The problem of variable selection arises when we want to decide which variables to include in the model. If there are no persuasive models available, nonparametric regression might offer a solution. However, high dimensional data are difficult to work with due to `the curse of dimensionality'. Unless we have an extremely large sample size, the sparseness of the data points causes nonparametric regression methods to break down. (Li 1991; Li 2000; Wand & Jones 1995.) In this case, it would be very useful to reduce the dimension of $\mathbf{X}$ without loss of information about the dependence between $\mathbf{X}$ and Y. Even if parametric regression was able to handle the high dimensional case, reducing the dimension helps the calculations and the interpretation of the coefficients. Also, if the dimension could be reduced to one or two, then a two- or three-dimensional scatter plot consisting of Y and $\mathbf{X}$ would contain all the information about the dependencies within the data. Being able to visualize the data set might give clue to what kind of model or approach would be feasible.

Sliced inverse regression (SIR) is a dimension reduction method introduced by Li (1991). It is applicable in a regression situation, where we have a response variable Y and possibly a high dimensional vector $\mathbf{X}$ of explanatory variables. A traditional dimension reduction technique is to apply *principal component analysis* (PCA) on $\mathbf{X}$ first and then use the first few principal components to explain Y. This is known as *principal component regression* (PCR). However, the dimension reduction in PCR does not take Y into consideration at all. Therefore, if for two different data sets the distribution of the vector-valued explanatory variables $\mathbf{X}_1$ and $\mathbf{X}_2$ were the same, the data sets would reduce to the same linear combinations (see Appendix A). This would happen even if the dependence between Y and $\mathbf{X}_i$ $(i = 1, 2)$ was not the same for the two data

sets.

SIR is a useful dimension reduction technique in high-dimensional regression problems where the response variable Y depends on K ($\leq p$) unknown linear combinations of the explanatory variables $X_1, \ldots, X_p$, but K and the exact form of dependence are unknown. This thesis presents the core of the theory behind SIR and studies its applicability using a large simulation study and a real data example. In Chapter 2, the idea of dimension reduction in regression is presented. Chapter 3 discusses SIR and more advanced definitions of dimension reduction. In Chapter 4 we conduct a simulation study and present the essential results. Finally, in Chapter 5 we apply SIR along with two other dimension reduction methods to a real data.

2 Dimension reduction in regression

We start this chapter by discussing the concept of regression – one of the four words in the title *On Sliced Inverse Regression*.

2.1 Regression

Typically, the theory of regression is concerned with the prediction of one random variable Y, a response variable, using other random variables $X_1, \ldots, X_p$, that are called explanatory variables. It will be convenient to write p-variate random variables as random column vectors $\mathbf{X} = (X_1, \ldots, X_p)'$ of p components, where the prime denotes the operation of transposing a row to a column. Random variables are represented by capital letters, and their realizations by lowercase letters. For example, Y is a random variable and y denotes its observed value. The value of Y varies in the marginal sample space Ω_Y of Y, i.e. $y \in \Omega_Y \subseteq \mathbb{R}$, where the symbols $\in$ and $\subseteq$ denote `belongs to' and `subset of (or is included in)', respectively. Vectors are denoted by boldface letters $\mathbf{X}$ and $\mathbf{x}$. The value of $\mathbf{X}$ varies in the marginal sample space $\Omega_{\mathbf{X}}$ of $\mathbf{X}$, i.e. $\mathbf{x} \in \Omega_{\mathbf{X}} \subseteq \mathbb{R}^p$.

Since Y and $\mathbf{X}$ are random, they have a joint distribution. The conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ is possibly a different probability distribution for each value of $\mathbf{x}$. When we wish to describe this entire family, we will say `the distribution of $Y|\mathbf{X}$', and the phrase `the distribution of $Y|(\mathbf{X} = \mathbf{x})$' refers to a single conditional distribution of Y given $\mathbf{X} = \mathbf{x}$.

The goal of regression is to study the conditional distribution of the response variable Y given the p-dimensional random vector $\mathbf{X}$ of explanatory variables. In other words, how does the distribution of $Y|(\mathbf{X} = \mathbf{x})$ change as a function of $\mathbf{x}$. If we wish to determine a function $m(\mathbf{x})$ for predicting a future observation of a response Y at a given value of $\mathbf{x}$, the mean square error $E(Y - m(\mathbf{x}))^2$ is minimized by choosing $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, the mean of the conditional distribution of $Y|(\mathbf{X} = x)$:

$$E(Y|X_1 = x_1, \ldots, X_p = x_p) = m(x_1, \ldots, x_p).$$

Although traditionally attention is restricted to the mean function $E(Y|\mathbf{X} = \mathbf{x})$ and perhaps to the variance function $var(Y|\mathbf{X} = \mathbf{x})$ of the conditional distribution of $Y|(\mathbf{X} = \mathbf{x})$, in full generality the object of interest is the conditional distribution of $Y|(\mathbf{X} = \mathbf{x})$.

The study of conditional distributions is often made using the mean function. In the p-dimensional case, $E(Y|\mathbf{X} = \mathbf{x})$ denotes the mean function, which

is a function of $\mathbf{x}$. In the case of $p = 1$, the mean $E(Y|X = x)$ is a function of x. We shall write the definition of $E(Y|X = x)$ in the case of continuous random variables. Let $f(y, x), f(y|x)$ and $f_X(x)$ denote the joint density function of Y and X, the conditional density function of Y given $X = x$ and the marginal density function of X, respectively. Then the conditional expected value of Y given $X = x$ can be written as

$$E(Y|X = x) = \int y f(y|x)\, dy,$$

where the conditional density function $f(y|x)$ of Y given $X = x$ is the function of y defined by $f(y|x) = f(x, y)/f_X(x)$ for any x such that $f_X(x) > 0$. (Casella & Berger 1990.) Then $E(Y|X = x)$ is the mean of the conditional distribution of Y given $X = x$.

There is a very useful connection between the population mean $E(Y)$ and $E(Y|X)$. Note that $E(Y|X)$ is a random variable, whose observed value is $E(Y|X = x)$ when the observed value of X is x. Then we have

$$\begin{aligned}
E(E(Y|X)) &= \int E(Y|X = x) f_X(x)\, dx \\
&= \int \Big(\int y f(y|x)\, dy \Big) f_X(x)\, dx \\
&= \int \int y \frac{f(x, y)}{f_X(x)} f_X(x)\, dy\, dx = \int \int y f(x, y)\, dy\, dx \\
&= \int y \Big(\int f(x, y)\, dx \Big) dy = \int y f_Y(y)\, dy \\
&= E(Y),
\end{aligned}$$

provided that the expectations exist (Casella & Berger 1990, p. 154). Here $f_Y(y)$ denotes the marginal density function of Y. The identity $E(E(Y|X)) = E(Y)$ is often needed in theoretical derivations.

2.2 Simple linear regression

The simple linear regression model is usually written as

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{2.1}$$

where β_0 and β_1 are unknown fixed parameters, Y is a random variable, x is a known constant and ε is a random variable with $E(\varepsilon) = 0$. To emphasize the fact that our inferences about the relationship between Y and x assume knowledge of x, we could write (2.1) as

$$E(Y|x) = \beta_0 + \beta_1 x. \tag{2.2}$$

If the explanatory variable is random instead of fixed, as will be the case in this thesis, we write

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2.3}$$

where X and ε are assumed to be independent and $E(\varepsilon) = 0$. Then the conditional expected value

$$E(Y|X) = \beta_0 + \beta_1 X$$

is a random variable. Let X take any value $x \in \Omega_X$ and let us focus on $E(Y|X = x)$. One such model is the bivariate normal model, where (X, Y) follows the bivariate normal distribution with means $E(X)$ and $E(Y)$, variances σ_X^2 and σ_Y^2 and covariance σ_{XY}. For a bivariate normal model the conditional distribution of Y given $X = x$ is normal and

$$E(Y|X = x) = E(Y) + \frac{\sigma_{XY}}{\sigma_X^2}(x - E(X)) = (E(Y) - \frac{\sigma_{XY}}{\sigma_X^2}E(X)) + (\frac{\sigma_{XY}}{\sigma_X^2})x.$$

Consequently, the bivariate normal model implies a linear regression function.

On the other hand, suppose that random variables X and Y have a joint density function, but the joint density function is not necessarily normal. Then the linear hypothesis (2.2), that is, $E(Y|X = x) = \beta_0 + \beta_1 x$ for some $\beta_0, \beta_1 \in \mathbb{R}$ implies that

$$E(Y|X = x) = E(Y) + \frac{\sigma_{XY}}{\sigma_X^2}(x - E(X)). \tag{2.4}$$

The identity (2.4) follows straightforwardly, since by definition

$$E(Y|X = x) = \int y f(y|x)\, dy = \int y \frac{f(x, y)}{f_X(x)}\, dy = \beta_0 + \beta_1 x,$$

and therefore

$$\int y f(x, y)\, dy = \beta_0 f_X(x) + \beta_1 x f_X(x), \tag{2.5}$$

for all $x \in \Omega_X$. This gives

$$\int \int y f(x, y)\, dy\, dx = \beta_0 \int f_X(x)\, dx + \beta_1 \int x f_X(x)\, dx,$$

which is equivalent to

$$E(Y) = \beta_0 + \beta_1 E(X). \tag{2.6}$$

Now, multiplying (2.5) by x and integrating both sides, we obtain

$$\int \int xy\, f(x, y)\, dy\, dx = \beta_0 \int x f_X(x)\, dx + \beta_1 \int x^2 f_X(x)\, dx,$$

which is equivalent to

$$E(XY) = \beta_0 E(X) + \beta_1 E(X^2). \tag{2.7}$$

Solving (2.6) and (2.7) for β_0 and β_1, we get

$$\beta_1 = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E(X)^2} = \frac{\sigma_{XY}}{\sigma_X^2} \tag{2.8}$$

and

$$\beta_0 = E(Y) - \frac{\sigma_{XY}}{\sigma_X} E(X).$$

Consequently, the result (2.4) follows.

2.3 Multiple linear regression

The multiple linear regression model is similar to the previous model, except that it has multiple explanatory variables. If the explanatory variables are fixed, say $x_1, \ldots, x_p$, we write

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon,$$

which is an extension of the model (2.1). If the explanatory variable is a p-dimensional random vector, $\mathbf{X} = (X_1, \ldots, X_p)'$, we write

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

or simply

$$Y = \beta_0 + \boldsymbol{\beta}'\mathbf{X} + \varepsilon, \tag{2.9}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a fixed unknown parameter vector, $\mathbf{X}$ and ε are independent and $E(\varepsilon) = 0$. The model (2.9) is a generalization of the model (2.3). The conditional expectation of Y given $\mathbf{X}$ under (2.9) is

$$E(Y|\mathbf{X}) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}. \tag{2.10}$$

Suppose that the $p+1$ random vector $(Y, \mathbf{X}')'$ has a joint distribution, not necessarily normal, with mean vector $(E(Y), E(\mathbf{X})')'$ and covariance matrix

$$\begin{pmatrix} \sigma_Y^2 & \boldsymbol{\Sigma}_{Y\mathbf{X}} \\ \boldsymbol{\Sigma}_{\mathbf{X}Y} & \boldsymbol{\Sigma} \end{pmatrix},$$

where $\sigma_Y^2 = var(Y)$, $\boldsymbol{\Sigma}_{\mathbf{X}Y} = (\sigma_{X_1Y}, \ldots, \sigma_{X_pY})'$ and $\boldsymbol{\Sigma} = cov(\mathbf{X})$. Then the coefficients of the linear predictor (2.10) take the form (Johnson & Wichern 2002, Section 7.8)

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}Y} \text{ and } \beta_0 = E(Y) - \boldsymbol{\beta}'\boldsymbol{E}(\mathbf{X}). \tag{2.11}$$

This generalizes the identity (2.4) of the previous section to the case of p explanatory variables.

Furthermore, if the joint distribution of $(Y, \mathbf{X}')'$ is normal, then the conditional expectation of Y given $\mathbf{X}$ is also linear, given by (2.10), with coefficients (2.11) (Johnson & Wichern 2002, Section 4.2).

2.4 Dimension reduction

Dimension reduction is motivated by the hope that high dimensional data would be retrievable by observations in lower dimensions without losing any information about the dependence between $\mathbf{X}$ and Y. Li (1991) introduced the model

$$Y = g(\boldsymbol{\beta}_1'\mathbf{X}, \boldsymbol{\beta}_2'\mathbf{X}, \ldots, \boldsymbol{\beta}_K'\mathbf{X}, \varepsilon) \tag{2.12}$$

to describe such a situation. Here Y is a univariate random variable, $\mathbf{X} = (X_1, \ldots, X_p)'$ is a p-dimensional random column vector and ε is a random error independent of $\mathbf{X}$ and its probability distribution is unknown. The p-dimensional column vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, dimension K $(K \leq p)$ and the function g are all unknown.

Let us now compare the model (2.12) to the model written as

$$Y = f(X_1, \ldots, X_p, \varepsilon). \tag{2.13}$$

In model (2.13), function f has all the original explanatory variables as inputs. The key difference between the models (2.12) and (2.13) is that the function g uses variables $\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_K'\mathbf{X}$, which are linear combinations of the original explanatory variables. These K linear combinations are the new explanatory variables. The smaller the value of K, the greater the dimension reduction. It is assumed in model (2.12), that variables $\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_K'\mathbf{X}$ hold the same regression information as $\mathbf{X}_1, \ldots, \mathbf{X}_p$, so there is no loss of information switching from (2.13) to (2.12). (Li 2000.)

We take three examples to demonstrate what the model (2.12) might look like.

Example 2.1. The first model

$$Y = 5 + X_1 + X_2 + X_3 + 0X_4 + 0X_5 + \varepsilon$$

is linear. In this case, $p = 5$ and $\mathbf{X} = (X_1, \ldots, X_5)'$. Now one can choose $K = 1$, $\boldsymbol{\beta} = \boldsymbol{\beta}_1 = (1, 1, 1, 0, 0)'$ and $g(\boldsymbol{\beta}_1'\mathbf{X}, \varepsilon) = 5 + \boldsymbol{\beta}_1'\mathbf{X} + \varepsilon$.

Example 2.2. Assume that $\mathbf{X}$ is as in Example 2.1. Our model is

$$Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + \varepsilon.$$

Now one can choose $K = 2$, $\boldsymbol{\beta}_1 = (1,0,0,0,0)'$, $\boldsymbol{\beta}_2 = (0,1,0,0,0)'$ and $g(\boldsymbol{\beta}_1'\mathbf{X}, \boldsymbol{\beta}_2'\mathbf{X}, \varepsilon) = \boldsymbol{\beta}_1'\mathbf{X}/(0.5 + (\boldsymbol{\beta}_2'\mathbf{X} + 1.5)^2) + \varepsilon$. In this example Y can not be presented by only one linear combination of the X_i's ($i = 1, \ldots, 5$), but two linear combinations of X_i's are needed.

Example 2.3. Assume that $\mathbf{X}$ is as in Examples 2.1 and 2.2. Our model is

$$Y = X_1^2 + X_2^2\,\sigma\varepsilon.$$

Now one can choose $K = 2$, $\boldsymbol{\beta}_1 = (1,0,0,0,0)'$, $\boldsymbol{\beta}_2 = (0,1,0,0,0)'$ and $g(\boldsymbol{\beta}_1'\mathbf{X}, \boldsymbol{\beta}_2'\mathbf{X}, \varepsilon) = (\boldsymbol{\beta}_1'\mathbf{X})^2 + (\boldsymbol{\beta}_2'\mathbf{X})^2\,\sigma\varepsilon$.

In Examples 2.1, 2.2 and 2.3 dimension K, vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and function g are given by the fixed model. Example 2.1 presents a multiple linear regression model, where the parameter vector $\boldsymbol{\beta}_1$ is $(1,1,1,0,0)'$. In traditional multiple linear regression analysis it is assumed that $K = 1$, the function g is known and the aim is to estimate $\boldsymbol{\beta}_1$. As mentioned before, Li's model (2.12) differs completely from this situation, because in Li's model the function g, vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ and K are all *unknown*.

Model (2.12) is essential in sliced inverse regression – SIR is a dimension reduction method which seeks the unknown vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, or more specifically, the space spanned by these vectors. This is a key point in SIR, because the individual vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ in model (2.12) cannot be identified since the function g is unknown. Thus, we are interested in estimating the subspace $\mathscr{C}(\mathbf{B})$, the linear space spanned by the columns $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ of the matrix $\mathbf{B} = (\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_K)$. Example 2.4 illustrates this situation.

Note that the model (2.12) is not well defined in the sense that if $Y = g(\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_K'\mathbf{X}, \varepsilon)$, then for any nonsingular $K \times K$ matrix $\mathbf{H}$, there exists a function g^* and a matrix $\mathbf{G} = \mathbf{BH}$ such that $Y = g^*(\boldsymbol{\gamma}_1'\mathbf{X}, \ldots, \boldsymbol{\gamma}_K'\mathbf{X}, \varepsilon)$, where $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$ are the column vectors of $\mathbf{G}$ (Halmos 1958).

Example 2.4. Assume that we have the following model:

$$\begin{aligned} Y &= X_1 X_2 + \varepsilon \\ &= (\boldsymbol{\beta}_1'\mathbf{X})(\boldsymbol{\beta}_2'\mathbf{X}) + \varepsilon, \end{aligned}$$

where $K = 2$, $\boldsymbol{\beta}_1 = (1,0,0,0,0)'$, $\boldsymbol{\beta}_2 = (0,1,0,0,0)'$, $\mathbf{X} = (X_1, \ldots, X_5)'$ and ε is a random error independent of $\mathbf{X}$. Now $\mathbf{B} = (\boldsymbol{\beta}_1 : \boldsymbol{\beta}_2)$ is a 5×2 matrix, where the column vectors are $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Since $\operatorname{rank}(\mathbf{B}) = 2$, $\mathscr{C}(\mathbf{B})$ is a plane in $\mathbb{R}^5$.

Let us choose a nonsingular matrix

$$\mathbf{H} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and perform the matrix multiplication $\mathbf{G} = \mathbf{BH}$. This yields a 5×2 matrix $\mathbf{G}$ with two linearly independent column vectors, $\boldsymbol{\gamma}_1 = (1, 1, 0, 0, 0)'$ and $\boldsymbol{\gamma}_2 = (1, -1, 0, 0, 0)'$. Since $\mathscr{C}(\mathbf{B}) = \mathscr{C}(\mathbf{G})$ (Halmos 1958), the column vectors of $\mathbf{G}$ generate the same plane in $\mathbb{R}^5$ as the column vectors of $\mathbf{B}$.

Now we can write Y in two different ways:

$$\begin{aligned} Y &= X_1 X_2 + \varepsilon \\ &= (\boldsymbol{\beta}_1' \mathbf{X})(\boldsymbol{\beta}_2' \mathbf{X}) + \varepsilon \\ &= g(\boldsymbol{\beta}_1' \mathbf{X}, \boldsymbol{\beta}_2' \mathbf{X}, \varepsilon) \end{aligned}$$

and

$$\begin{aligned} Y &= [(X_1 + X_2)^2 - (X_1 - X_2)^2]/4 + \varepsilon \\ &= [(\boldsymbol{\gamma}_1' \mathbf{X})^2 - (\boldsymbol{\gamma}_2' \mathbf{X})^2]/4 + \varepsilon \\ &= g^*(\boldsymbol{\gamma}_1' \mathbf{X}, \boldsymbol{\gamma}_2' \mathbf{X}, \varepsilon). \end{aligned}$$

Example 2.1 shows that the model presented in the first row of the example can be reparameterized by a different pair of independent parameter vectors $(\boldsymbol{\gamma}_1 : \boldsymbol{\gamma}_2) \neq (\boldsymbol{\beta}_1 : \boldsymbol{\beta}_2)$, provided that the function g is changed from g to g^* correspondingly.

3 Sliced inverse regression

In Section 2.4 we introduced the idea of dimension reduction and tentatively the aim of SIR. In this chapter we will introduce the theoretical basics of SIR and illustrate the key results in a way that helps to understand the procedure. Then we will present the SIR algorithm to compute the SIR estimate for a practical data set.

3.1 Inverse regression

The usual forward regression $Y|\mathbf{X}$ was introduced in Section 2.1. In the inverse regression $\mathbf{X}|Y$ the roles of $\mathbf{X}$ and Y are reversed. Instead of using $E(Y|\mathbf{X})$, SIR uses $E(\mathbf{X}|Y)$, which naturally has the expected value $E(E(\mathbf{X}|Y)) = E(\mathbf{X})$. The benefit of the change of roles comes from the fact that $\mathbf{X}|Y$ is composed of p simple regressions, $X_j|Y,\ j = 1, \ldots, p$. Due to `the curse of dimensionality', the response surface $E(Y|\mathbf{X} = \mathbf{x})$ is very difficult to estimate directly. For realistic sample sizes, standard nonparametric regression methods such as kernel methods, nearest neighbor methods or smoothing splines break down quickly, when the dimension p is larger than two (Duan & Li 1991). The conditional expectation $E(\mathbf{X}|Y = y)$ on the other hand can be estimated taking one coordinate at a time. This way the estimation of $E(\mathbf{X}|Y = y)$ comes down to a one-dimensional case. That is, we can write $E(\mathbf{X}|Y = y)$ as $(E(X_1|Y = y), \ldots, E(X_p|Y = y))'$ and estimate each term separately. Thus we obtain p one-dimensional curve smoothing problems.

Whether or not the change of roles of Y and $\mathbf{X}$ feels natural, it offers a way to circumvent the dimensionality problem. In some cases, the inverse regression might be of interest on its own. For example, inverse regression was used in calibration problems in Krutchkoff (1967, 1969).

3.2 Dimension reduction subspace

Consider a regression problem consisting of a univariate response variable Y and a p-dimensional vector $\mathbf{X}$ of random explanatory variables with a joint cumulative distribution function (cdf) $F(y, \mathbf{x})$. The goal of a regression analysis is to infer how the cdf $F(y|\mathbf{x})$ of the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ changes as the value of $\mathbf{x}$ varies in the sample space.

The motivation of dimension reduction was presented at the beginning of

Section 2.4. Let

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}'\mathbf{X}, \tag{3.1}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_K)$ is a $p \times K$ matrix $(K \leq p)$. Statement (3.1) indicates that Y and $\mathbf{X}$ are conditionally independent given any value for the random vector $\mathbf{B}'\mathbf{X}$ (see Appendix C). Cook (1998) introduced the following important concept.

Definition 3.1. If (3.1) holds, then the subspace $\mathscr{C}(\mathbf{B})$ is a *dimension reduction subspace* (DRS) for the regression of Y on $\mathbf{X}$, where the columns $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ of the $p \times K$ matrix $\mathbf{B}$ form a basis for the DRS.

To be exact, $\mathscr{C}(\mathbf{B})$ is a subspace of $\mathbb{R}^p$, i.e. $\mathscr{C}(\mathbf{B}) \subseteq \mathbb{R}^p$, and it is defined as $\mathscr{C}(\mathbf{B}) = \{\mathbf{z} : \mathbf{z} = a_1\boldsymbol{\beta}_1 + \cdots + a_K\boldsymbol{\beta}_K, \text{ for some } a_1, \ldots, a_K \in \mathbb{R}\}$. The statement (3.1) is equivalent to $F_{Y|\mathbf{X}}(y|\mathbf{x}) = F_{Y|\mathbf{B}'\mathbf{X}}(y|\mathbf{B}'\mathbf{x})$ for all values of $\mathbf{x}$ in the sample space. This means that the p-dimensional vector $\mathbf{X}$ of explanatory variables can be replaced by the K-dimensional vector $\mathbf{B}'\mathbf{X}$ of explanatory variables without loss of information about the dependence between Y and $\mathbf{X}$. Such a $\mathbf{B}$ always exists, because (3.1) is trivially true when $\mathbf{B} = \mathbf{I}_p$, where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. (Cook 1998.)

A fundamental goal here is to reduce the dimension of $\mathbf{X}$. Hence the idea of the smallest dimension arises naturally.

Definition 3.2. A subspace of $\mathbb{R}^p$ is said to be a minimum DRS for the regression of Y on $\mathbf{X}$ if it is a DRS with the smallest dimension within all DRSs.

Minimum DRSs always exist, but they are not necessarily unique (Cook 1994). The following example presents a case of a non-unique minimum DRS.

Example 3.1. (Cook 1998, p.105) Let $p = 2$ and let $\mathbf{X} = (X_1, X_2)'$ be uniformly distributed on the unit circle, $||\mathbf{X}|| = 1$, where $||\ ||$ denotes a vector norm. Set $Y = X_1^2 + \varepsilon$, where ε is an independent error. Therefore we have $Y \perp\!\!\!\perp \mathbf{X} | (1,0)\mathbf{X}$. It follows from $X_1^2 = 1 - X_2^2$ that $Y = X_1^2 + \varepsilon = (1 - X_2^2) + \varepsilon$, and consequently also $Y \perp\!\!\!\perp \mathbf{X} | (0,1)\mathbf{X}$. Therefore $\mathscr{C}((1,0)')$ and $\mathscr{C}((0,1)')$ are both minimum DRSs.

Note that, in Example 3.1, $\mathbf{X}$ is not a genuinely bivariate random variable. If we change the Cartesian coordinate system to the polar coordinate system and write $\mathbf{X} = (cos\, V, sin\, V)'$, then $Y = cos^2 V + \varepsilon$. Therefore, the model for Y can be written simply as a function of the univariate random variable V and a bivariate random vector is not necessarily needed.

The intersection of all DRSs is a subspace of $\mathbb{R}^p$ (Halmos 1958, p.17), but it is not necessarily a DRS. In the Example 3.1, $\mathscr{C}((1,0)')$ and $\mathscr{C}((0,1)')$ are both DRSs, but their intersection {(0,0)'} is not a DRS. It turns out that the intersection of all DRSs is a DRS under various reasonable conditions (Cook 1994, 1996, 1998). The following definition introduces an essential concept of dimension reduction.

Definition 3.3. If the intersection of all DRSs is a DRS, it is the *central* DRS and denoted by $\mathscr{C}_{Y|\mathbf{X}}$.

Cook (1994, 1996, 1998) introduced the concept of the central DRS and proved that it is the *unique minimum* DRS (Cook 1998, Proposition 6.2).

Any vector in the central DRS is called an *effective dimension reduction (e.d.r.) direction*. In this thesis from now on we will assume that the central DRS exists.

3.3 Models for dimension reduction

Let $\mathscr{C}_{Y|\mathbf{X}}$ be the central DRS with basis $\mathbf{B} = (\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_K)$ and dimension K. Then the statement

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}'\mathbf{X} \tag{3.2}$$

can be thought of as a dimension reduction model. Here $\mathscr{C}_{Y|\mathbf{X}}$ is well-defined and unique. The parameter of a dimension reduction model is the central DRS $\mathscr{C}_{Y|\mathbf{X}}$.

Duan & Li (1991) and Li (1991, 1992) represented dimension reduction in the following way:

$$Y = g(\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_K'\mathbf{X}, \varepsilon), \tag{3.3}$$

where function g and vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are unknown and $\mathbf{X}$ and ε are independent. Model (3.3) was introduced without addressing existence or uniqueness issues. When the central DRS exists, the dimension reduction models (3.2) and (3.3) are technically equivalent and they can be connected by requiring that $\mathbf{B} = (\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_K)$ is a basis for $\mathscr{C}_{Y|\mathbf{X}}$ (Cook 1998, p.114).

Example 3.2. (Cook 1998, p.187) Suppose that $(Y, \mathbf{X}')'$ follows a multivariate normal distribution, where $cov(\mathbf{X}) = \boldsymbol{\Sigma}$, $cov(Y, \mathbf{X}) = \boldsymbol{\Sigma}_{Y\mathbf{X}}$ and $var(Y) = \sigma^2$. Assuming that the conditional distribution of Y conditioned on $\mathbf{X}$ depends on $\mathbf{X}$ only through the conditional expectation of $Y|\mathbf{X}$, then

$$E(Y|\mathbf{X}) = E(Y) + \boldsymbol{\Sigma}_{Y\mathbf{X}}\boldsymbol{\Sigma}^{-1}(\mathbf{X} - E(\mathbf{X}))$$

and $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}Y}$ spans $\mathscr{C}_{Y|\mathbf{X}}$, where $\boldsymbol{\Sigma}_{\mathbf{X}Y} = \boldsymbol{\Sigma}_{Y\mathbf{X}}'$. The inverse regression function is

$$\begin{aligned} E(\mathbf{X}|Y) &= E(\mathbf{X}) + \boldsymbol{\Sigma}_{\mathbf{X}Y}\sigma^{-2}(Y - E(Y)) \\ &= E(\mathbf{X}) + \boldsymbol{\Sigma}\boldsymbol{\beta}\sigma^{-2}(Y - E(Y)). \end{aligned}$$

Thus the values of the centered inverse regression function $E(\mathbf{X}|Y = y) - E(\mathbf{X})$ fall in the one-dimensional subspace of $\mathscr{C}(\boldsymbol{\Sigma}\boldsymbol{\beta})$ as the value of $Y = y$ ranges over Ω_Y. Note that $\mathscr{C}(\boldsymbol{\Sigma}\boldsymbol{\beta}) = \boldsymbol{\Sigma}\mathscr{C}_{Y|\mathbf{X}}$.

It turns out that important characteristics of this example can be preserved when multivariate normality does not hold, but the following condition is assumed.

Condition 3.1. *(L.D.C.) Linear Design Condition. For any* $\mathbf{b} \in \mathbb{R}^p$ *the conditional expectation* $E(\mathbf{b}'\mathbf{X}|\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_K'\mathbf{X}) = c_0 + c_1\boldsymbol{\beta}_1'\mathbf{X} + \cdots + c_K\boldsymbol{\beta}_K'\mathbf{X}$ *for some constants* $c_i \in \mathbb{R}, i = 0, 1, \ldots, K$*, where vectors* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ *are defined as in (3.2).*

The L.D.C. condition is fulfilled when the distribution of $\mathbf{X}$ is elliptic. For example, the L.D.C. condition holds for the multivariate normal distribution. However, the distribution of $\mathbf{X}$ does not have to be elliptic in order to fulfill the L.D.C. condition, as will be seen in the simulation study.

As mentioned before, the concept of inverse regression is essential in SIR. Let $\mathscr{C}_{E(\mathbf{X}|Y)}$ denote the subspace spanned by $\{E(\mathbf{X}|Y = y) - E(\mathbf{X})|y \in \Omega_Y\}$. The subspace $\mathscr{C}_{E(\mathbf{X}|Y)}$ is called the *inverse regression subspace.*

The following result is called the main theorem of SIR.

Theorem 3.2. *Under the model (3.2) and the Linear Design Condition 3.1,* $\mathscr{C}_{E(\mathbf{X}|Y)} \subseteq \mathscr{C}(\boldsymbol{\Sigma}\mathbf{B}) = \boldsymbol{\Sigma}\mathscr{C}_{Y|\mathbf{X}}$*, where* $\mathbf{B}$ *is defined as in (3.2) and* $\boldsymbol{\Sigma} = cov(\mathbf{X})$ *is positive definite.*

Let $\mathbf{Z}$ denote the vector of standardized explanatory variables.

$$\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - E(\mathbf{X})).$$

Clearly, $E(\mathbf{Z}) = \mathbf{0}$ and $cov(\mathbf{Z}) = \mathbf{I}_p$. Working in terms of the vector of standardized explanatory variables involves no loss of generality, because we can always back-transform to the original scale. Also, since $\mathbf{Z}$ is a 1 - 1 linear transformation of $\mathbf{X}$,

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}'\mathbf{X} \Longleftrightarrow Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{T}'\mathbf{Z}, \tag{3.4}$$

where $\mathbf{T} = (\boldsymbol{\theta}_1 : \cdots : \boldsymbol{\theta}_K) = \boldsymbol{\Sigma}^{1/2}\mathbf{B}$ and $\boldsymbol{\theta}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_i$, $i = 1, \ldots, K$. The vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ are called *standardized e.d.r. directions* and $\mathscr{C}_{Y|\mathbf{Z}}$ is called the *standardized central DRS.* Under the assumption that the central DRS exists, the models (3.2) and (3.3) are equivalent. Hence by (3.4) we can rewrite (3.3) as

$$Y = g(\boldsymbol{\theta}_1'\mathbf{Z}, \ldots, \boldsymbol{\theta}_K'\mathbf{Z}, \varepsilon), \tag{3.5}$$

where the function g is not the same as in (3.3), but it is simply written as g to denote some unknown function.

Corollary 3.1. *Assume that (L.D.C.) holds. Then for model (3.5), the standardized inverse regression subspace, denoted by* $\mathscr{C}_{E(\mathbf{Z}|Y)}$*, is a subspace of the standardized central DRS* $\mathscr{C}_{Y|\mathbf{Z}}$*.*

As a random vector, $E(\mathbf{Z}|Y)$ has a covariance matrix $cov[E(\mathbf{Z}|Y)]$. By Corollary 3.1, $\mathscr{C}_{E(\mathbf{Z}|Y)} \subseteq \mathscr{C}_{Y|\mathbf{Z}}$. This does not guarantee equality between $\mathscr{C}_{E(\mathbf{Z}|Y)}$ and $\mathscr{C}_{Y|\mathbf{Z}}$, and thus inference about $\mathscr{C}_{E(\mathbf{Z}|Y)}$ possibly covers only a part of $\mathscr{C}_{Y|\mathbf{Z}}$. The next theorem provides a rationale that the estimate of $cov[E(\mathbf{Z}|Y)]$ serves to estimate $\mathscr{C}_{E(\mathbf{Z}|Y)}$ (see Cook 1998, Proposition 11.1).

Theorem 3.3. *Let* $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$ *and* $\mathscr{C}_{E(\mathbf{Z}|Y)}$ *the inverse regression space. Then*

$$\mathscr{C}\{cov[E(\mathbf{Z}|Y)]\} = \mathscr{C}_{E(\mathbf{Z}|Y)}.$$

Proof. By the projection theorem (Halmos 1958, p.129), $\mathbb{R}^p = \mathscr{C}_{E(\mathbf{Z}|Y)} \oplus \mathscr{C}^{\perp}_{E(\mathbf{Z}|Y)}$, where $\perp$ denotes the orthogonal complement and $\oplus$ denotes the direct sum of subspaces.

1° Suppose that $\mathbf{a} \in \mathscr{C}^{\perp}_{E(\mathbf{Z}|Y)}$. Then we have $E(\mathbf{a}'\mathbf{Z}|Y) = \mathbf{a}'E(\mathbf{Z}|Y) = 0$. But

$$cov[E(\mathbf{Z}|Y)] = E[E(\mathbf{Z}|Y)E(\mathbf{Z}'|Y)]$$

and hence

$$cov[E(\mathbf{Z}|Y)]\mathbf{a} = E[E(\mathbf{Z}|Y)E(\mathbf{a}'\mathbf{Z}|Y)] = 0,$$

if $\mathbf{a} \in \mathscr{C}^{\perp}_{E(\mathbf{Z}|Y)}$ and consequently $\mathbf{a} \in \mathscr{C}\{cov[E(\mathbf{Z}|Y)]\}^{\perp}$.

2° If $\mathbf{a} \in \mathscr{C}\{cov[E(\mathbf{Z}|Y)]\}^{\perp}$, then $cov[E(\mathbf{Z}|Y)]\mathbf{a} = 0$ and

$$\mathbf{a}'cov[E(\mathbf{Z}|Y)]\mathbf{a} = E[E(\mathbf{a}'\mathbf{Z}|Y)E(\mathbf{a}'\mathbf{Z}|Y)] = var[E(\mathbf{a}'\mathbf{Z}|Y)] = 0. \tag{3.6}$$

Since $E[E(\mathbf{a}'\mathbf{Z}|Y)] = E(\mathbf{a}'\mathbf{Z}) = 0$, it follows from (3.6) that $\mathbf{a}'E(\mathbf{Z}|Y) = E(\mathbf{a}'\mathbf{Z}|Y) = 0$ (with probability 1), and hence $\mathbf{a} \in \mathscr{C}^{\perp}_{E(\mathbf{Z}|Y)}$.

We have proved that $\mathscr{C}^{\perp}_{E(\mathbf{Z}|Y)} = \mathscr{C}\{cov[E(\mathbf{Z}|Y)]\}^{\perp}$ and hence we may conclude that $\mathscr{C}_{E(\mathbf{Z}|Y)} = \mathscr{C}\{cov[E(\mathbf{Z}|Y)]\}$. □

We find a basis for $\mathscr{C}\{cov[E(\mathbf{Z}|Y)]\}$, and consequently also for $\mathscr{C}_{E(\mathbf{Z}|Y)}$, by constructing the eigenvalue decomposition of matrix $cov[E(\mathbf{Z}|Y)]$:

$$cov[E(\mathbf{Z}|Y)]\mathbf{t}_i = \lambda_i \mathbf{t}_i \quad (i = 1, \ldots, p, \quad \lambda_1 \geq \cdots \geq \lambda_p), \tag{3.7}$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $cov[E(\mathbf{Z}|Y)]$ and $\mathbf{t}_1, \ldots, \mathbf{t}_p$ are the corresponding eigenvectors. Since $\mathscr{C}_{E(\mathbf{Z}|Y)} \subseteq \mathscr{C}_{Y|\mathbf{Z}}$, (3.7) must give at most K nonzero eigenvalues. The eigenvectors $\mathbf{t}_1, \ldots, \mathbf{t}_{K-d}$, where d denotes the number of nonzero eigenvalues, corresponding to the nonzero eigenvalues form the basis of $\mathscr{C}_{E(\mathbf{Z}|Y)}$. The value of d determines whether $\mathscr{C}_{E(\mathbf{Z}|Y)}$ is a subspace of $\mathscr{C}_{Y|\mathbf{Z}}$ or if it is equal to it. The possible values of d are $0, \ldots, K$, and for $d = 0$ we have $\mathscr{C}_{E(\mathbf{Z}|Y)} = \mathscr{C}_{Y|\mathbf{Z}}$, and we would be able to determine the standardized central DRS. However, the possible equality between $\mathscr{C}_{E(\mathbf{Z}|Y)}$ and $\mathscr{C}_{Y|\mathbf{Z}}$ and the value of d are unknown.

Since we are originally interested in the central DRS rather than in the standardized central DRS, we must transform the eigenvectors $\mathbf{t}_1, \ldots, \mathbf{t}_{K-d}$ to obtain a basis of $\mathscr{C}_{E(\mathbf{X}|Y)}$. From (3.4) we obtained the equalities $\boldsymbol{\theta}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_i,\ i = 1, \ldots, K$. By this result the vectors $\mathbf{b}_i = \boldsymbol{\Sigma}^{-1/2}\,\mathbf{t}_i\,(j = 1, \ldots, K-d)$ form a basis of $\mathscr{C}_{E(\mathbf{X}|Y)}$ so that $\mathscr{C}_{E(\mathbf{X}|Y)} \subseteq \mathscr{C}_{Y|\mathbf{X}}$. Therefore we are able to obtain a subspace of the central DRS, but we cannot infer the possible equality of the subspaces.

3.4 Proof of the main theorem of SIR

We write the proof of the main theorem in SIR (Theorem 3.2) and clarify several steps from the proof given by Li (2000). The proof in this thesis is given in the case of $K = 1$, which is the core of the proof, but it can be extended to $K > 1$ using matrix notations (see Li 2000).

We may assume without loss of generality that $\mathbf{X}$ is centered, i.e. $E(\mathbf{X}) = \mathbf{0}$. Let us write first $E(\mathbf{X}|\boldsymbol{\beta}'\mathbf{X})$ componentwise as

$$E(\mathbf{X}|\boldsymbol{\beta}'\mathbf{X}) = (E(X_1|\boldsymbol{\beta}'\mathbf{X}), \ldots, E(X_p|\boldsymbol{\beta}'\mathbf{X}))'.$$

By the L.D.C. condition there exists constants c_{i0}, c_{i1}, such that

$$E(X_i|\boldsymbol{\beta}'\mathbf{X}) = c_{i0} + c_{i1}\boldsymbol{\beta}'\mathbf{X}, \quad i = 1, \ldots, p. \tag{3.8}$$

Since

$$E(E(X_i|\boldsymbol{\beta}'\mathbf{X})) = E(c_{i0} + c_{i1}\boldsymbol{\beta}'\mathbf{X}) = c_{i0} + c_{i1}\boldsymbol{\beta}'E(\mathbf{X}) = c_{i0},$$

and on the other hand (Casella & Berger 1990, Theorem 4.4.1)

$$E(E(X_i|\boldsymbol{\beta}'\mathbf{X})) = E(X_i) = 0$$

for all $i = 1, \ldots, p$. Thus $c_{10} = \cdots = c_{p0} = 0$.

The expectations $E(X_i|\boldsymbol{\beta}'\mathbf{X}), i = 1, \ldots, p$ are linear and hence by (2.8)

$$c_{i1} = \frac{cov(X_i, \boldsymbol{\beta}'\mathbf{X})}{var(\boldsymbol{\beta}'\mathbf{X})}, i = 1, \ldots, p.$$

This yields

$$(c_{11}, \ldots, c_{p1})' = \frac{cov(\mathbf{X}, \boldsymbol{\beta}'\mathbf{X})}{var(\boldsymbol{\beta}'\mathbf{X})} = \boldsymbol{\Sigma}\boldsymbol{\beta}/\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}, \tag{3.9}$$

since

$$cov(\mathbf{X}, \boldsymbol{\beta}'\mathbf{X}) = cov(\mathbf{X})\boldsymbol{\beta} = \boldsymbol{\Sigma}\boldsymbol{\beta} \text{ and } var(\boldsymbol{\beta}'\mathbf{X}) = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}.$$

Due to (3.8) and (3.9) we have

$$E(\mathbf{X}|\boldsymbol{\beta}'\mathbf{X}) = (\boldsymbol{\Sigma}\boldsymbol{\beta}/\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\beta}'\mathbf{X}.$$

Since $E(\mathbf{X}|Y = y) = E(E(\mathbf{X}|\boldsymbol{\beta}'\mathbf{X})|Y = y)$ by Proposition 1 in Appendix B, we have

$$E(\mathbf{X}|Y = y) = E((\boldsymbol{\Sigma}\boldsymbol{\beta}/\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\beta}'\mathbf{X}|Y = y). \tag{3.10}$$

Since $\boldsymbol{\Sigma}\boldsymbol{\beta}/\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}$ is constant, we may rewrite (3.10) as

$$E(\mathbf{X}|Y = y) = \boldsymbol{\Sigma}\boldsymbol{\beta}E(\boldsymbol{\beta}'\mathbf{X}|Y = y)/\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}.$$

If we denote $m(y) = E(\boldsymbol{\beta}'\mathbf{X}|Y = y)/\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}$, a scalar function of y, then

$$E(\mathbf{X}|Y = y) = m(y)\boldsymbol{\Sigma}\boldsymbol{\beta}.$$

Hence $E(\mathbf{X}|Y = y) \in \mathscr{C}(\boldsymbol{\Sigma}\boldsymbol{\beta})$, as was to be proved.

3.5 SIR algorithm

So far in this thesis we have explained two words of the title *On Sliced Inverse Regression*, that is, inverse regression. The meaning of the second word, *Sliced*, becomes evident in this section. We argued in Section 3.1 the ease of estimation of the inverse regression function $E(\mathbf{X}|Y = y)$ compared to the forward regression function $E(Y|\mathbf{X} = \mathbf{x})$. There are several nonparametric regression methods to estimate the inverse regression function, such as kernel, nearest neighbor or smoothing splines, but SIR uses slicing because of its simplicity (Li 1991).

When Y is continuous, Li (1991) suggested replacing Y with a discrete version $\tilde{Y}$ based on partitioning the observed range of Y into S fixed, non-overlapping slices H_s, $s = 1, \ldots, S$. Let us define $\tilde{Y} = s$, when $\tilde{Y} \in H_s$. Because $\tilde{Y}$ is a function of Y, it follows (Cook 1998, Proposition 4.5) that $\tilde{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}'\mathbf{X}$, where $\mathbf{B}$ is a basis for $\mathscr{C}_{Y|\mathbf{X}}$. Thus the central DRS $\mathscr{C}_{\tilde{Y}|\mathbf{X}}$ from the regression of $\tilde{Y}$ on $\mathbf{X}$ provides information about $\mathscr{C}_{Y|\mathbf{X}}$:

$$\mathscr{C}_{\tilde{Y}|\mathbf{X}} \subseteq \mathscr{C}_{Y|\mathbf{X}}. \tag{3.11}$$

We can hope that $\mathscr{C}_{\tilde{Y}|\mathbf{X}} = \mathscr{C}_{Y|\mathbf{X}}$. Applying the Theorem 3.3 to $\mathbf{Z}|\tilde{Y}$ yields the identity

$$\mathscr{C}\{cov[E(\mathbf{Z}|\tilde{Y})]\} = \mathscr{C}_{E(\mathbf{Z}|\tilde{Y})}. \tag{3.12}$$

On the other hand, by Corollary 3.1 $\mathscr{C}_{E(\mathbf{Z}|\tilde{Y})} \subseteq \mathscr{C}_{\tilde{Y}|\mathbf{Z}}$, and by (3.11) $\mathscr{C}_{\tilde{Y}|\mathbf{Z}} \subseteq \mathscr{C}_{Y|\mathbf{Z}}$. Then we have the string of relationships

$$\mathscr{C}(cov[E(\mathbf{Z}|\tilde{Y})]) = \mathscr{C}_{E(\mathbf{Z}|\tilde{Y})} \subseteq \mathscr{C}_{\tilde{Y}|\mathbf{Z}} \subseteq \mathscr{C}_{Y|\mathbf{Z}}$$

(Cook 1998, p.204).

The relationship (3.12) shows that the inverse regression subspace $\mathscr{C}_{E(\mathbf{Z}|\tilde{Y})}$ is spanned by the eigenvectors corresponding to the nonzero eigenvalues of $cov[E(\mathbf{Z}|\tilde{Y})]$. A central idea of SIR is to construct an estimate of $\mathscr{C}_{E(\mathbf{Z}|\tilde{Y})}$ from an estimate of $cov[E(\mathbf{Z}|\tilde{Y})]$. This is done by constructing the *SIR algorithm* (Li 1991):

Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be the original data set with p+1 variables ($\mathbf{x}$ and y) and n cases. The SIR algorithm goes as follows:

1. Center and standardize $\mathbf{x}$ to get $\mathbf{z}_i = \hat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})\,(i = 1, \ldots, n)$, where $\hat{\mathbf{\Sigma}}$ and $\bar{\mathbf{x}}$ are, respectively, the sample covariance matrix and the sample mean of $\mathbf{x}$.

2. Order and reindex the data using variable y so that $y_1 \leq \cdots \leq y_n$:

$$\begin{array}{ll} y_1 & \mathbf{z}_1 = (z_{11}, z_{12}, \ldots, z_{1p})' \\ y_2 & \mathbf{z}_2 = (z_{21}, z_{22}, \ldots, z_{2p})' \\ \vdots & \\ y_n & \mathbf{z}_n = (z_{n1}, z_{n2}, \ldots, z_{np})' \end{array}$$

3. Use the ordered data to divide the data into S non overlapping slices H_s so that the number of observations in each slice is as equal as possible.

4. Estimate $E(\mathbf{Z}|Y \in H_s)$ by computing the sample mean of $\mathbf{z}$ within each slice:

$$\bar{\mathbf{z}}_s = n_s^{-1} \sum_{i=1}^{n} \mathbf{z}_i I_{H_s}(y_i),$$

where $s = 1, \ldots, S$ is the index of slices, n_s is the number of observations in slice H_s and the indicator $I_{H_s} = 1$ if $y_s \in H_s$ and zero otherwise. The data are displayed by writing

$$\begin{array}{rcl} \bar{\mathbf{z}}_1 & = & (\bar{z}_{11}, \bar{z}_{12}, \ldots, \bar{z}_{1p})' \\ \bar{\mathbf{z}}_2 & = & (\bar{z}_{21}, \bar{z}_{22}, \ldots, \bar{z}_{2p})' \\ & \vdots & \\ \bar{\mathbf{z}}_S & = & (\bar{z}_{S1}, \bar{z}_{S2}, \ldots, \bar{z}_{Sp})'. \end{array}$$

5. Estimate $cov(E(\mathbf{Z}|\tilde{Y}))$ by calculating the covariance matrix

$$\hat{\mathbf{V}} = n^{-1} \sum_{s=1}^{S} n_s \bar{\mathbf{z}}_s \bar{\mathbf{z}}_s'.$$

6. Estimate $\mathscr{C}_{\tilde{Y}|\mathbf{X}}$ by conducting a principal component analysis by forming the following eigenvalue decomposition

$$\hat{\mathbf{\Gamma}}' \hat{\mathbf{\Lambda}} \hat{\mathbf{\Gamma}} = \hat{\mathbf{V}},$$

where $\hat{\mathbf{\Lambda}}$ contains the eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ and $\hat{\mathbf{\Gamma}}$ the corresponding eigenvectors $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_p$. The first few eigenvectors corresponding to the K $(K \leq p)$ largest eigenvalues can be used as estimators for the standardized e.d.r. directions.

7. Compute the estimators for the original e.d.r. directions:

$$\hat{\boldsymbol{\beta}}_k = \hat{\mathbf{\Sigma}}^{-1/2}\hat{\boldsymbol{\eta}}_k, \quad k = 1, \ldots, K$$

As we have mentioned earlier in this thesis, in practice K is unknown and we have to estimate it. As soon as the value of K is selected, $\mathscr{C}(\hat{\boldsymbol{\beta}}_1 : \cdots : \hat{\boldsymbol{\beta}}_K)$ provides an estimate of $\mathscr{C}_{Y|\mathbf{X}}$. In the next chapter (Subsections 4.4.2 and 4.4.3) we discuss two procedures to estimate K.

4 Simulations

In this chapter we report on an extensive simulation study of the performance of three dimension reduction methods. For purposes of comparison, we included the following three dimension reduction methods: sliced inverse regression (SIR), *sliced average variance estimate* (SAVE) and *principal hessian directions* (PHD). SAVE and PHD are introduced in the following section. The aim of the study is to compare SIR, SAVE and PHD in a large number of different settings.

In our simulation study, we consider four different choices of dimensions, $p = 5, 10, 20, 40$, of $\mathbf{X}$, four different choices of distributions of $\mathbf{X}$, four different choices of sample sizes, $N = 100, 200, 400, 800$, seven different models, and two different levels of noise, $\sigma = 0.5, 1$.

In the simulations we used the R 2.9.0 software (R Development Core Team 2009) and the R-packages *dr, lattice, vcd* and *mvtnorm* (see Weisberg 2008, Sarkar 2009, Meyer et al. 2009 and Genz et al. 2009).

4.1 SAVE and PHD

In this section we will briefly introduce two alternative dimension reduction methods to SIR: SAVE and PHD. The simulation study conducted in this thesis gives a very useful insight into the performance of each of the three methods in the various settings.

4.1.1 SAVE

SAVE was introduced by Cook & Weisberg (1991) in the discussion initiated by Li (1991). As Li, Cook and Weisberg pointed out, SIR is not a valid method for every situation. In particular, SIR is known to fail when the response surface is symmetric about origin. One of the models used in this paper describes this situation (M3 in Subsection 4.2.2). Whereas SIR uses only the first moment, $E(\mathbf{X}|Y)$, SAVE uses the second moment, $cov(\mathbf{X}|Y)$, and can therefore detect, for example, symmetric dependence more efficiently than SIR. SAVE is not very efficient in estimating monotone trends for small to moderate sample sizes.

The objective of SAVE is to estimate the central DRS by estimating the *SAVE matrix*

$$E(\mathbf{I}_p - cov(\mathbf{Z}|Y))^2,$$

its eigenvalues and the corresponding eigenvectors. (Cook & Weisberg 1991.)

More information about SAVE can be found in literature (see for example Cook & Weisberg 1991 and Cook 1998).

4.1.2 PHD

PHD is a dimension reduction method introduced by Li (1992). Li was well aware of the deficiency of SIR in the case of symmetric dependence, but PHD is a method which can handle many symmetric cases for finding the central DRS. PHD uses the *Hessian matrix*

$$\mathbf{H}_{\mathbf{X}}(\mathbf{x}) = ((\partial^2/\partial x_i \partial x_j) f(\mathbf{x})), \quad i, j = 1, \ldots, p$$

of the regression function $f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, where ∂ denotes the partial derivative. The motivation is that $\mathbf{H}_{\mathbf{X}}(\mathbf{x})$ is degenerate along any directions that are orthogonal to $\mathscr{C}_{Y|\mathbf{X}}$, which is defined as in (3.2). The principal hessian directions with respect to the distribution of $\mathbf{X}$ are defined as the eigenvectors $\mathbf{b}_1, \ldots, \mathbf{b}_p$ of the matrix $E(\mathbf{H}_{\mathbf{X}}(\mathbf{X}))\mathbf{\Sigma}$:

$$E(\mathbf{H}_{\mathbf{X}}(\mathbf{X}))\mathbf{\Sigma}\mathbf{b}_j = \lambda_j \mathbf{b}_j, \quad j = 1, \ldots, p$$

where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$. (Li 1992.)

PHD can be conducted using two different methods: the response-based method or the residual-based method. In the simulation studies we have used the residual based method.

More information about PHD can be found in the literature (see for example Li 1992, 2000 and Cook 1998).

4.2 Models

In the simulation study conducted in this thesis, data are generated from seven different models, which are presented in this section. The models are selected from four papers (Cook & Weisberg 1991; Li 1992; Li 2000; Zhu & Zhu 2007) in such a way, that for SIR there are at least two models for which it should work, for PHD there are at least two models for which it should work and for SAVE there are at least three models for which it should work. That is, for SIR there are two models from Li (2000), for SAVE there are two models from Cook & Weisberg (1991) and one model from Zhu & Zhu (2007), and for PHD there are two models from Li (1992). In other words, each model has been selected from a paper in which the method in question was developed by the author or authors, except Zhu & Zhu (2007). In addition, each of these groups contain models of two different dimensions of the central DRS ($K = 1, 2$).

In the following subsections we introduce the models used in the simulation study. Each subsection is named according to the method suitable for the models introduced in the corresponding subsection, so that it is easier to compare the performance of different methods. Let us emphasize, however, that all three

methods are applied to every model. It is convenient to mention at this point that for every model $\varepsilon \sim N(0,1)$ and $\sigma = 0.5, 1$, where the symbol $\sim$ denotes `distributed as'.

4.2.1 Models suitable for SIR

We consider the following models, which we introduced in Section 2.4, suitable for SIR:

$$\begin{aligned} \text{M1:} \quad & Y = 5 + \boldsymbol{\beta}'\mathbf{X} + \sigma\varepsilon \\ \text{M2:} \quad & Y = \frac{\boldsymbol{\beta}_1'\mathbf{X}}{0.5 + (\boldsymbol{\beta}_2'\mathbf{X} + 1.5)^2} + \sigma\varepsilon \end{aligned}$$

For M1, $K = 1$ and $\boldsymbol{\beta} = (1, 1, 1, 0, \ldots, 0)'$. For M2, $K = 2$, $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)'$ and $\boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)'$.

4.2.2 Models suitable for SAVE

The models suitable for SAVE are:

$$\begin{aligned} \text{M3:} \quad & Y = (\boldsymbol{\beta}'\mathbf{X})^2 + \sigma\varepsilon \\ \text{M4:} \quad & Y = (\mu + \sqrt{2}\,\boldsymbol{\beta}'\mathbf{X})^2 + \sigma\varepsilon \\ \text{M5:} \quad & Y = (\boldsymbol{\beta}_1'\mathbf{X})^2 + (\boldsymbol{\beta}_2'\mathbf{X})^2\,\sigma\varepsilon \end{aligned}$$

For M3, $K = 1$ and $\boldsymbol{\beta} = (1, 0, \ldots, 0)'$. For M4, $K = 1$ and $\boldsymbol{\beta} = (1, 1, 0, \ldots, 0)'$. For M5, $K = 2$, $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)'$ and $\boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)'$.

4.2.3 Models suitable for PHD

The models suitable for PHD are:

$$\begin{aligned} \text{M6:} \quad & Y = \boldsymbol{\beta}'\mathbf{X}\, sin(2\boldsymbol{\beta}'\mathbf{X}) + \sigma\varepsilon \\ \text{M7:} \quad & Y = cos(2\boldsymbol{\beta}_1'\mathbf{X}) - cos(\boldsymbol{\beta}_2'\mathbf{X}) + \sigma\varepsilon \end{aligned}$$

For M6, $K = 1$ and $\boldsymbol{\beta} = (1, 0, \ldots, 0)'$. For M7, $K = 2$, $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)'$ and $\boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)'$.

4.3 Distributions of X

We use four different distributions of $\mathbf{X} = (X_1, \ldots, X_p)'$ in the simulation study. These distributions are presented in the following subsections.

4.3.1 Distributional setting D1

The first distribution is a multivariate normal distribution

$$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I}_p),$$

which is elliptic and the L.D.C. condition is fulfilled.

4.3.2 Distributional setting D2

The second distribution is a multivariate t-distribution

$$\mathbf{X} \sim T_p(\mathbf{0}, \mathbf{I}_p),$$

with 5 degrees of freedom. The distribution is elliptic and the L.D.C. condition is fulfilled. The multivariate t-distribution has heavier tails than the normal distribution and the components are uncorrelated, but not independent.

4.3.3 Distributional setting D3

The third distribution is constructed such that

$$\begin{aligned} X_1 &\sim (V-2)/8, \text{with } V \sim \Gamma(2,2) \\ X_2 &\sim \sqrt{3/5} \times t(5) \\ X_3 &\sim U(-\sqrt{3}, \sqrt{3}), \end{aligned}$$

where $\Gamma(2,2)$ denotes the gamma distribution with the shape parameter 2 and the scale parameter 2, $t(5)$ denotes the t-distribution with 5 degrees of freedom and $U(-\sqrt{3}, \sqrt{3})$ denotes the uniform distribution on $(-\sqrt{3}, \sqrt{3})$. The rest of the variables, denoted by $\mathbf{X}_{p-3} = (X_4, \ldots, X_p)'$, are distributed as

$$\mathbf{X}_{p-3} \sim N_{p-3}(\mathbf{0}, \mathbf{I}_{p-3}).$$

The distribution of $\mathbf{X}$ is not elliptic and it is related to an independent component model (IC model) (see Hyvärinen, Karhunen & Oja 2001). Notice that X_1, the first component of the distributional setting D3, is skew, whereas the other components are symmetric. All the components have $E(X_i) = 0$ and $var(X_i) = 1$.

4.3.4 Distributional setting D4

The fourth distribution is inspired by Velilla (1998). The paper introduces some examples of distributions of $\mathbf{X}$ which are not elliptic, but the L.D.C. condition is still fulfilled. We expand these special cases and conduct a general way to construct distributions of this type.

Let $\mathbf{B} = (\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_K)$ be a $p \times K$ matrix, where vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are defined as in (3.2). Since vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are linearly independent, $rank(B) =$

K, where $rank$ denotes the rank of a matrix. We want to find a $p \times (p-K)$ matrix $\mathbf{C}$ such that $rank(\mathbf{C}) = p-K$ and $\mathbf{C}'\mathbf{B} = \mathbf{0}$. Velilla (1998) showed, that when $\mathbf{X}$ is constructed as

$$\mathbf{X} = \mathbf{C}\mathbf{v} + \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{w}, \tag{4.1}$$

where $\mathbf{v}$ and $\mathbf{w}$ are two independent random vectors of appropriate dimensions, then the distribution of $\mathbf{X}$ is not elliptic but the L.D.C. condition is still fulfilled.

Let us choose a $p \times p$ matrix

$$\mathbf{P_{I-B}} = \mathbf{I} - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}',$$

which is the projector to the orthocomplement of $\mathbf{B}$. Because $rank(\mathbf{P_B}) = rank(\mathbf{B}) = K$, we have $rank(\mathbf{P_{I-B}}) = p-K$. Since $\mathbf{P_{I-B}}$ is the projector to the orthocomplement of $\mathbf{B}$, we have $\mathbf{P_{I-B}}\mathbf{B} = \mathbf{0}$. Because $rank(\mathbf{P_{I-B}}) = p-K$, matrix $\mathbf{P_{I-B}}$ has $p-K$ linearly independent column vectors. We find those linearly independent column vectors by constructing the eigenvalue decomposition of $\mathbf{P_{I-B}}$:

$$\mathbf{TDT}' = \mathbf{TD}^{1/2}\mathbf{D}^{1/2}\mathbf{T}' = \mathbf{P_{I-B}},$$

where the diagonal matrix $\mathbf{D}$ contains the eigenvalues and $\mathbf{T}$ contains the corresponding eigenvectors of matrix $\mathbf{P_{I-B}}$. Again, since $rank(\mathbf{P_{I-B}}) = p-K$ and $\mathbf{P_{I-B}}$ is a projector, $p-K$ diagonal elements of $\mathbf{D}$ are nonzero and K diagonal elements are zero. By calculating $\mathbf{TD}$ we obtain a $p \times p$ matrix, denoted by $\mathbf{C}^*$, which has $p-K$ nonzero column vectors and K zero column vectors. Lets choose the $p-K$ nonzero column vectors from $\mathbf{C}^*$ and construct a $p \times (p-K)$ matrix $\mathbf{C}$. Matrix $\mathbf{C}$ has the desired properties $rank(\mathbf{C}) = p-K$ and $\mathbf{C}'\mathbf{B} = \mathbf{0}$. Finally, we construct $\mathbf{X}$ as in (4.1).

In the simulations we have chosen $\mathbf{v} \sim U(-4,4)$ and depending on whether we have a model with $K=1$ or $K=2$ [$\mathbf{w} = w_1$ or $\mathbf{w} = (w_1, w_2)'$], we have chosen $w_1 \sim 0.5N(0,4) + 0.5N(0,16)$ and $w_2 \sim U(-4,4)$ (see Velilla 1998).

Figures 4.1- 4.4 show the graphs of the four distributional settings, when $p=5$ and $N=200$. The figures demonstrate the differences between the different settings. Both the normal and the T_5 data have the same spherical shape, but the T_5 data has much heavier tails. In the IC data in Figure 4.3 however, the different pairwise scatter plots show that some components are clearly not elliptic and that the first component is skew. The distribution D4 in Figure 4.4 looks quite different from the other distributions. All pairwise scatter plots are clearly not elliptical. For model D4, $\mathbf{B}$ is in this case $(1,1,1,0,0)'$.

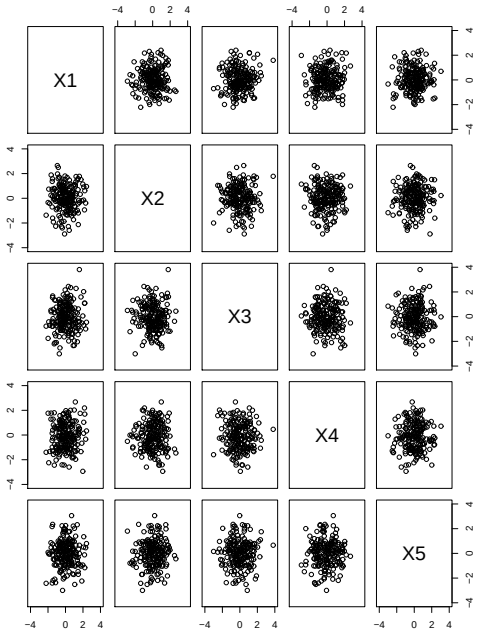

Figure 4.1. Scatterplot matrix for a data set of 200 observations, generated from the multivariate normal distibution $N_5(\mathbf{0}, \mathbf{I}_5)$ (the setting D1).

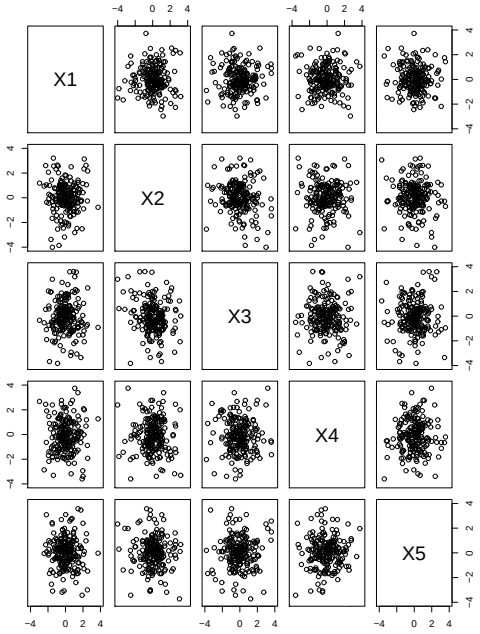

Figure 4.2. Scatterplot matrix for a data set of 200 observations, generated from the multivariate t-distibution $T_5(\mathbf{0}, \mathbf{I}_5)$ (the setting D2).

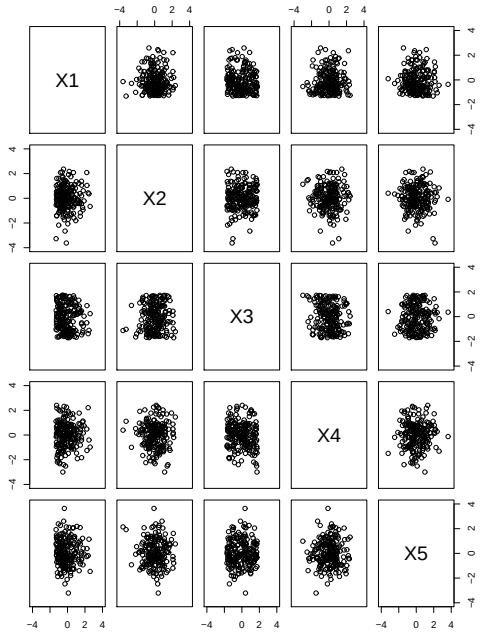

Figure 4.3. Scatterplot matrix for a data set of 200 observations, generated from a distribution under the distributional setting D3 with $p = 5$.

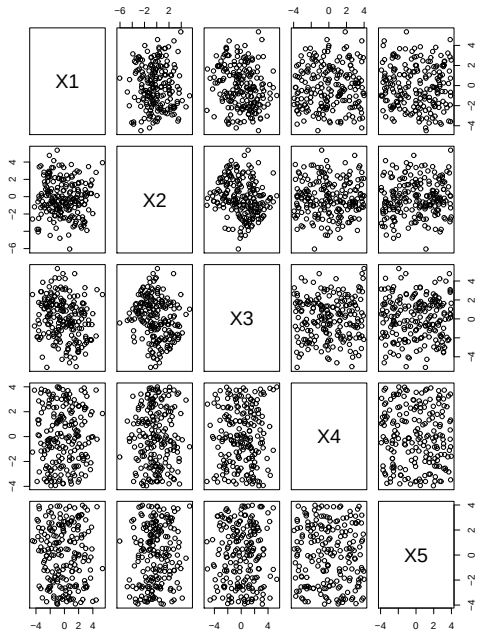

Figure 4.4. Scatterplot matrix for a data set of 200 observations, generated from a distribution under the distributional setting D4, where matrix $\mathbf{B} = (1, 1, 1, 0, 0)'$ is constructed under the model M1 with $p = 5$.

4.4 The objective of the simulation study

The objective of the simulation study is to compare the performance of the dimension reduction methods SIR, SAVE and PHD in the large number of different settings. Our investigations concentrate on two points: 1. How close the estimated e.d.r. directions are to the true e.d.r. directions when the dimension K of the central DRS is known and 2. How many times K is estimated correctly when K is unknown. Criteria to evaluate these objectives are discussed in the following subsections.

4.4.1 The R^2 criterion

The closeness of the estimated and true e.d.r. directions when K is known is measured by using the R^2 criterion introduced by Li (1991). For $K = 1$, the criterion calculates the squared multiple correlation coefficient between the projected variable $\mathbf{b}'\mathbf{X}$ and the ideally reduced variables $\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_K'\mathbf{X}$ and can be written as

$$R^2(\mathbf{b}) = \max \frac{(\mathbf{b}'\boldsymbol{\Sigma}\boldsymbol{\beta})^2}{\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b} \cdot \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}},$$

where $\boldsymbol{\beta} \in \mathscr{C}(\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_K)$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{X}$. When $K > 1$, we use the squared trace correlation, denoted by $R^2(\mathbf{B})$, i.e. the average of the squared canonical correlation coefficients between $\mathbf{b}_1'\mathbf{X}, \ldots, \mathbf{b}_K'\mathbf{X}$ and $\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_K'\mathbf{X}$.

The R^2 criterion takes values between [0,1], where 0 is total failure and 1 is the perfect estimate. The values of R^2 from the simulation results are shown using boxplots. The criterion is used for each of the methods SIR, SAVE and PHD.

4.4.2 χ^2-test

In practice we do not know K, but we must estimate it. In this study we distinguish the three cases: underestimation, correct result and overestimation. The estimation is done by using two procedures: a χ^2-statistic and a BIC (Bayesian information criterion) criterion to be introduced in the next subsection. Li (1991) introduced a χ^2-statistic for SIR to estimate the dimension K. Li (1992) also introduced a different χ^2-statistic for PHD. Shao, Cook & Weisberg (2007) introduced a χ^2-statistic for SAVE. There are many versions of χ^2-statistic in the literature for estimating the dimension K, of which some assume normality of $\mathbf{X}$ and some do not. In this thesis we use procedures which assume normality of $\mathbf{X}$, since these procedures are implemented in the *dr* package for SIR, SAVE and PHD. Next, we will introduce the χ^2-statistic for SIR and leave the χ^2-statistics for SAVE and PHD for the reader to be found from the literature.

Li proposes the next theorem, where $\bar{\lambda}_{p-K}$ is the average of the $p-K$ smallest eigenvalues and S is the number of slices:

Theorem 4.1. *If* $\mathbf{X}$ *is normally distributed, then* $n(p-K)\bar{\lambda}_{(p-K)}$ *follows asymptotically a* χ^2*-distribution with* $(p-K)(S-K-1)$ *degrees of freedom.*

Using Theorem 4.1, we can assess how many linear combinations should be chosen. That is, we estimate the value of K. This can be done by using the criterion

$$p\text{-value}_j = P(\chi^2_{(p-j)(S-j-1)} \geq n(n-p)\bar{\lambda}_{(p-j)}),$$

presented by Li (2000).

The idea is to construct a sequence of p-values starting from $j = 0$. If p-value$_j$ is less than 0.05 for example, we can conclude that the dimension of the central DRS is at least $j+1$. Then increase the value of j until the first `too large' p-value indicates that we should choose $j-1$ as the estimate of the dimension of the central DRS ($K = j-1$). In the simulation study we tested only up to $K = 4$ for computational reasons and since this is enough for our purposes.

In summary, three different χ^2-tests are applied in the simulation study - one for each dimension reduction method. The results of the χ^2-estimates for SIR, SAVE and PHD are visualized using barplots.

4.4.3 BIC

We selected K also using the BIC type criterion of Zhu, Miao & Peng (2006). It is defined as follows: let

$$log\, L_k = \frac{n}{2} \sum_{i=1+min(\tau,k)}^{p} (log\, \hat{\theta}_i + 1 - \hat{\theta}_i),$$

where $\hat{\theta}_i$ is defined as $\hat{\theta}_i = \hat{\lambda}_i + 1$, where $\hat{\lambda}_i$ are the eigenvalues of the estimate for $cov(E(\mathbf{X}|Y))$, τ is the number of $\hat{\theta}_i$'s that are greater than 1 and k is the number of e.d.r. directions. Let

$$G(k) = logL_k - \frac{C_n k(2p-k+1)}{2},$$

where p is the dimension of $\mathbf{X}$. In our simulations we have used the penalty constant $C_n = c^{-1}W_n$, where c is the number of data points in each slice. According to results in Zhu et al. (2006), W_n can be selected in a fairly wide range. We have chosen $W_n = (0.5\, log(n) + 0.1n^{1/3})/2$, which performs best over all W_n's used in the simulation studies in Zhu et al. (2006).

The estimator of K is defined as the maximizer $\hat{K}$ of $G(k)$ over $k \in (0, \ldots, p-1)$, that is,

$$G(\hat{K}) = \max_{0 \geq k \geq p-1} G(k).$$

The BIC procedure is applied to SIR only and the results are visualized in the same barplots as the results from the χ^2-procedures. Therefore, the accuracy of the estimation of the e.d.r. directions of SIR is being measured by using both the χ^2 procedure suitable for SIR and the BIC procedure. For methods SAVE and PHD, only the χ^2-technique suitable for each method is applied.

4.5 Results

In this section we summarize the essential simulation results. We report the method that works the best for each model and observe graphs that include interesting information. The graphs do not necessarily present the general tendency in the results for a specific model, but may show some surprising results. The values of the R^2 criterion are visualized using boxplots and the barplots display the three possible estimation results: underestimation, correct estimation and overestimation, which are presented by three shades of grey. Underestimation is shown in light grey, correct estimation is shown in grey and overestimation is shown in dark grey. Note that only a small fraction of the results are displayed in figures. All extensive simulation summaries and figures are available upon request from the author.

4.5.1 Accuracy of subspace estimation assuming K known

Model M1

Figure 4.5 shows the values of the R^2 criterion under the model M1 for the distributional setting D1, when the standard deviation of ε is $\sigma = 0.5$. SIR works well throughout every value of p and N, except when $p = 40$ and $N = 100$. SAVE works well when p is 5 and 10, but it needs a large number of data points. When p is 5 or 10, PHD takes values from the whole range [0,1], which indicates its inability to detect a proper DRS. For $p = 20, 40$, PHD does not work at all. In the case of D2, SIR deteriorates and SAVE and PHD do not work at all. For D3, SIR is the only method that works, but for D4 also SAVE works when p is small and/or N is large.

When σ is increased to 1, the methods become unable to detect a proper DRS, that is, the range of values of R^2 becomes wider. It is interesting to observe that for D1 SAVE collapses totally when $\sigma = 1$. All in all, SIR is superior to SAVE and PHD in this setting.

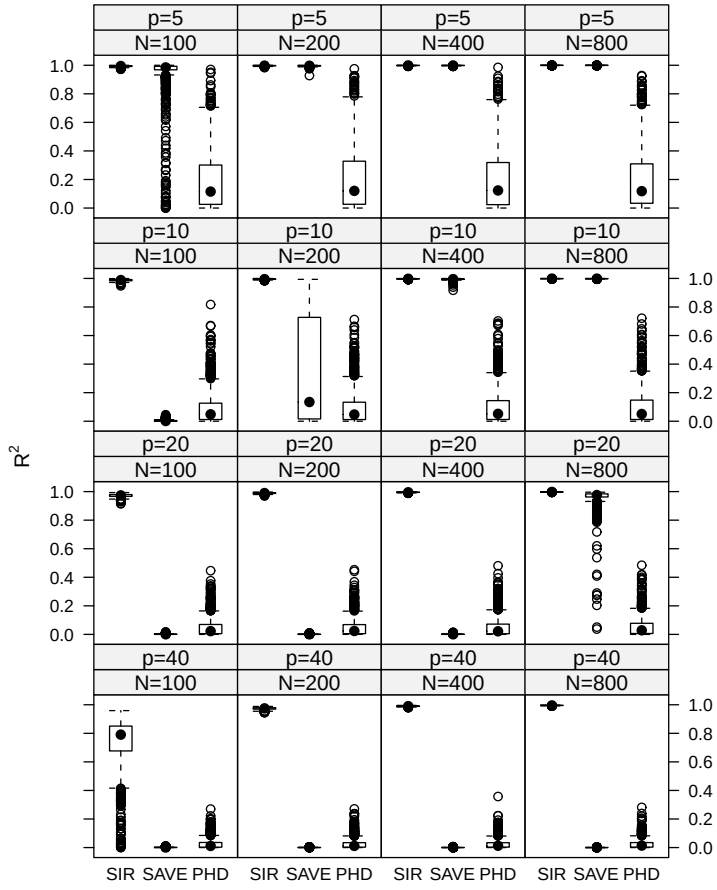

Figure 4.5. Boxplot for the R^2 criterion under the setting M1, D1, $\sigma = 0.5$.

Model M2

For M2, SIR does not work as well as it did for M1, but it works better than SAVE or PHD. It is best for D1 and D4. Figure 4.6 shows that for large dimensions SIR needs a large value of N in order to work. SAVE works only when p is small and N is large. However, the most interesting finding is that PHD does almost as well as SIR for every distributional setting.

When $\sigma = 1$, SIR and PHD have a wider range of values, and SAVE collapses totally.

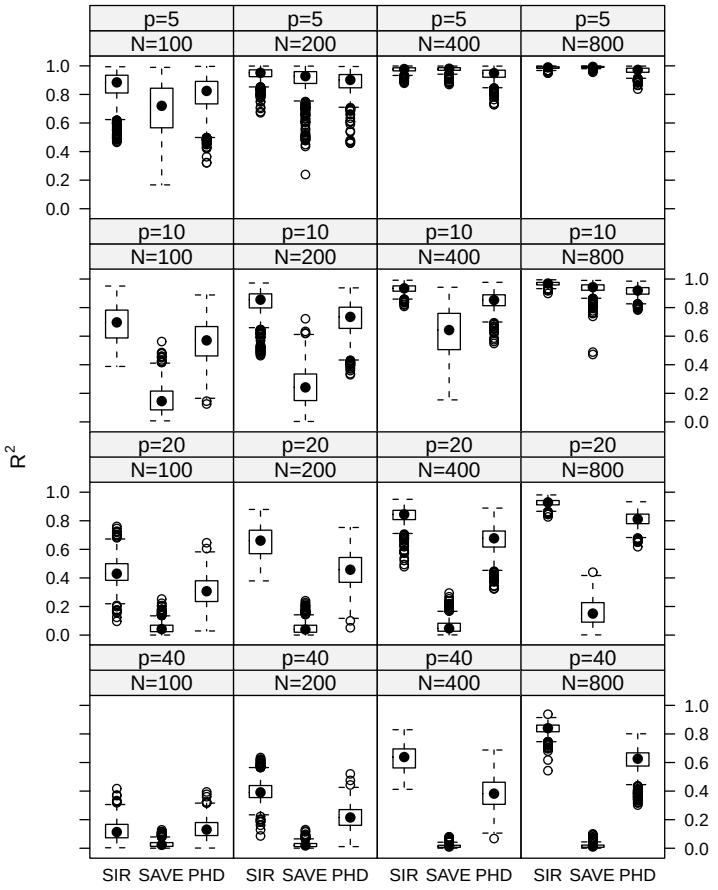

Figure 4.6. Boxplot for the R^2 criterion under the setting M2, D4, $\sigma = 0.5$.

Model M3

Figure 4.7 shows that under M3 SIR is not good for D1, because it takes values from the whole range of R^2. SAVE works when p/N is small. The overall performance of PHD is superior to other methods under the setting in Figure 4.7. For D2, no method work well except SAVE, when $p = 5$ and $N = 800$. For D3, both SIR and PHD perform well for all values of p when N is at least 200. The results for D4 are very close to the results for D1, except that SAVE works better for D4 than it does for D1.

In the case of $\sigma = 1$, the results deteriorate very little if at all.

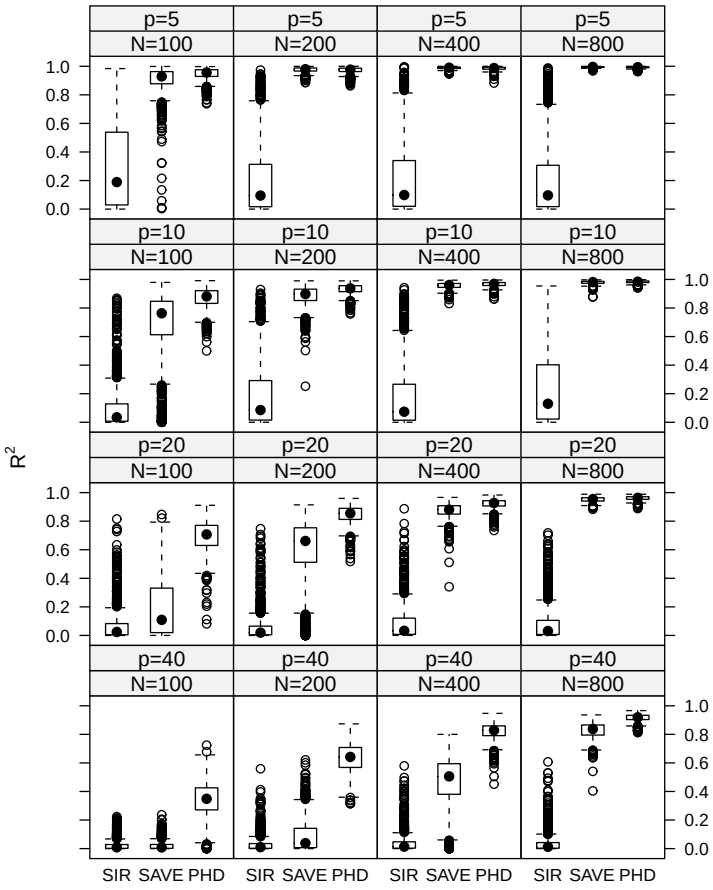

Figure 4.7. Boxplot for the R^2 criterion under the setting M3, D1, $\sigma = 0.5$.

Model M4

The results for M4 are quite similar to M3. In general, SIR does not work, but for D1 with $p = 5$ and $N = 800$ it performs decently, and for D3 it collapses totally. For SAVE and PHD the results are very similar to M3, that is, PHD performs a little bit better than SAVE. An interesting result is that none of the methods perform well for D2, although the distribution is elliptic. This can be seen in Figure 4.8.

The results for $\sigma = 1$ are very close to those for $\sigma = 0.5$.

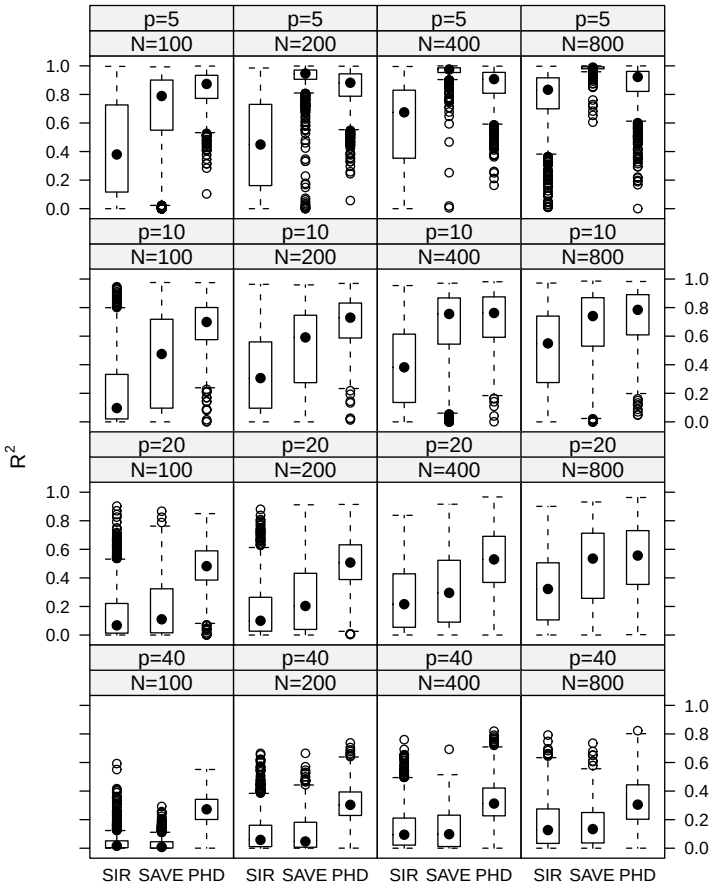

Figure 4.8. Boxplot for the R^2 criterion under the setting M4, D2, $\sigma = 1$.

Model M5

The overall performance of SAVE is superior to other methods under M5, but only when p is small and N is large. In general, SIR does not work at all, but for D3 the results are interesting: when N is large, SIR gets the majority of its values above 0.5 (see Figure 4.9). The results for PHD are quite similar to SIR.

When $\sigma = 1$, the results do not change significantly.

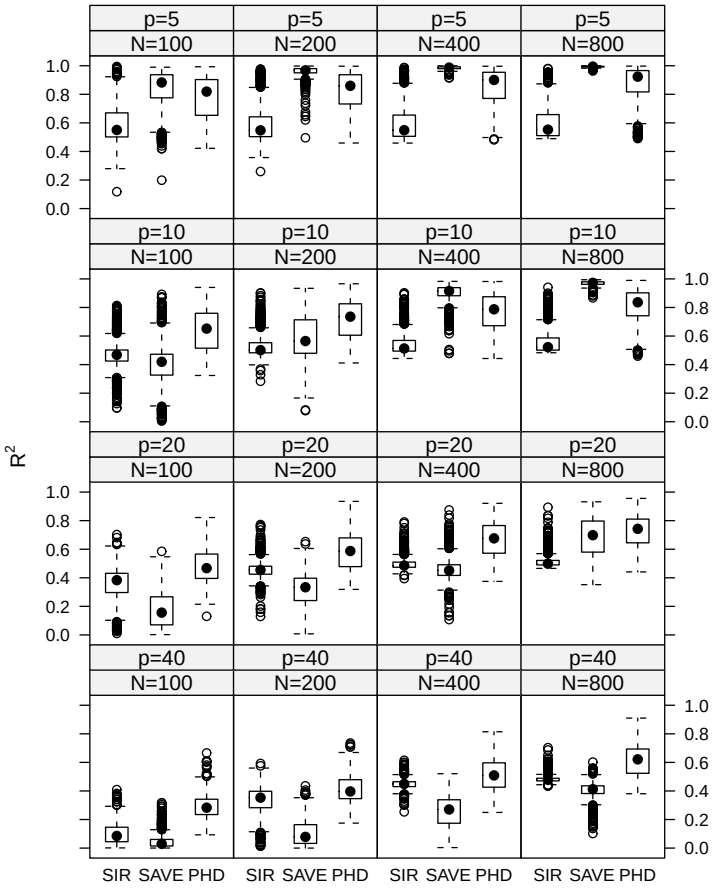

Figure 4.9. Boxplot for the R^2 criterion under the setting M5, D3, $\sigma = 0.5$.

Model M6

Under M6 no method is superior to the other methods. However, the performance of SAVE seems to be the best. For D1, SAVE and PHD work well only if $N = 800$ and also for $p = 5, N = 400$. For D2, none of the methods work. For D3, neither SAVE or PHD work and for D4 SAVE works only when p/N is small. PHD collapses totally for D4. SIR does not work for the majority of the cases under M6. It is unable to detect a proper DRS, as it takes values from the whole range of R^2. However, for D3 when $N = 800$, SIR works very well and is superior to SAVE and PHD (see Figure 4.10).

When $\sigma = 1$, SAVE collapses for D1. For D3, SIR and PHD take values from the whole range of R^2.

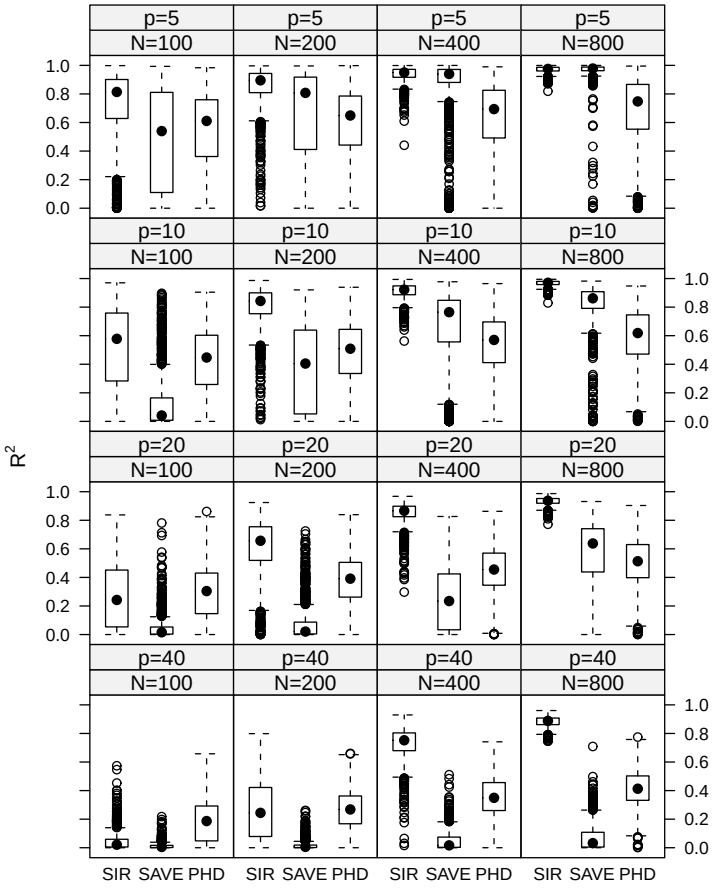

Figure 4.10. Boxplot for the R^2 criterion under the setting M6, D3, $\sigma = 0.5$.

Model M7

For M7, SIR works for no settings and SAVE works only for D1 and D3, when N is large and p is small. The performance of PHD is superior to the other methods for D1 when p/N is small, as can be seen in Figure 4.11. A similar result holds for D3.

When $\sigma = 1$, the results for SAVE and PHD deteriorate so that they perform well only for D1, when $N = 800$ and $p = 5$. PHD performs well also when $N = 800$ and $p = 10$.

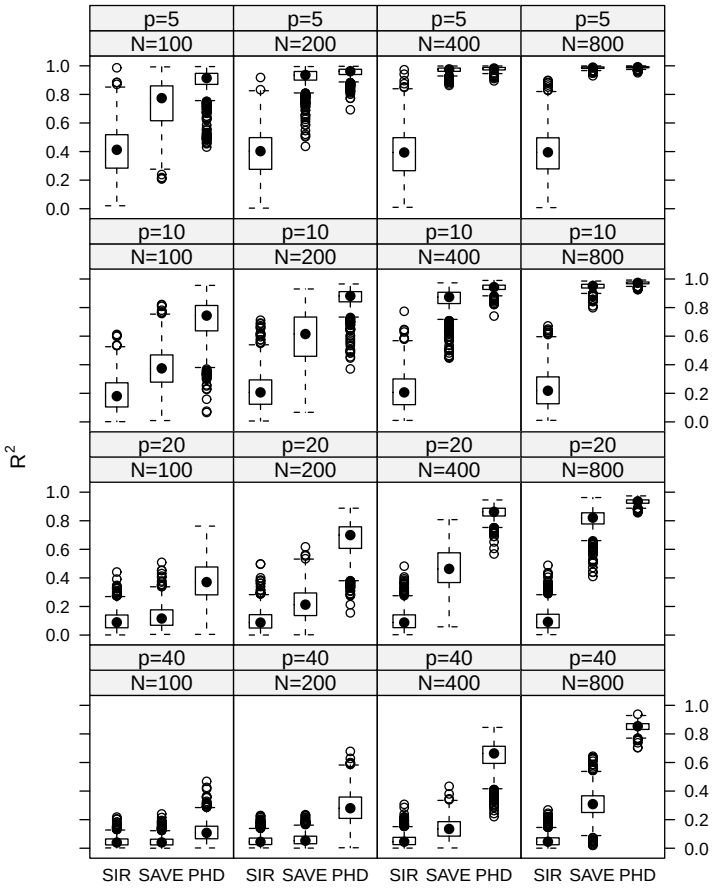

Figure 4.11. Boxplot for the R^2 criterion under the setting M7, D1, $\sigma = 0.5$.

4.5.2 Accuracy of subspace estimation assuming K unknown

Model M1

Figure 4.12 shows that for D1, χ^2_{SIR} estimates the dimension K correctly in the majority of the cases. BIC_{SIR} tends to overestimate, whereas χ^2_{PHD} underestimates. χ^2_{SAVE} estimates correctly only when p is small and N is large. In other cases it underestimates.

For D4 the results are quite similar to D1, but for D2 and D3 the results move towards overestimation. The results change very little when $\sigma = 1$.

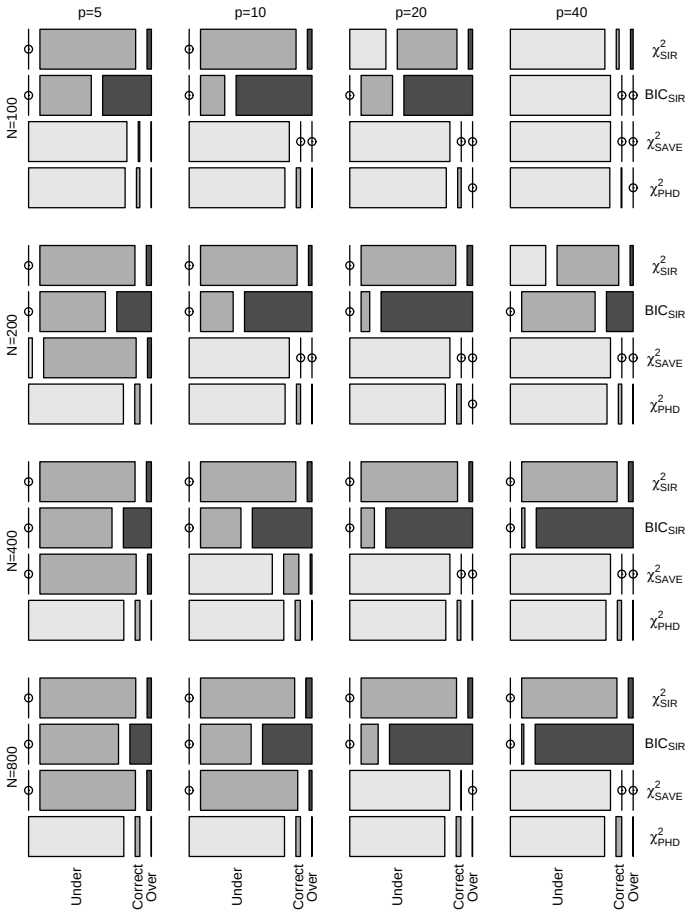

Figure 4.12. Barplot for the estimate of the dimension of the central DRS under the setting M1, D1, $\sigma = 0.5$.

Model M2

Compared to M1, the results for M2 are poorer. Figure 4.13 shows the difference from D2 quite clearly. χ^2_{SIR} estimates correctly in just a few cases and otherwise overestimates. All of the other methods are worse than χ^2_{SIR} and overestimate radically. An interesting tendency is that for a fixed N, increasing p makes the results move towards underestimation, whereas for a fixed p, increasing N makes the results move towards overestimation.

It is also interesting that for $\sigma = 1$ the results are better than for $\sigma = 0.5$, because there is a lot less overestimation.

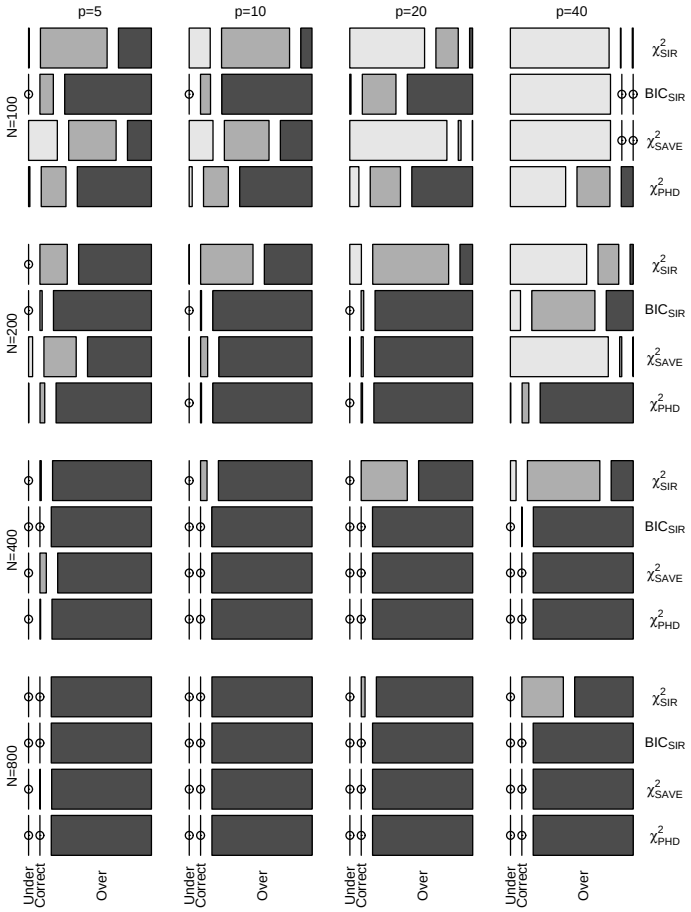

Figure 4.13. Barplot for the estimate of the dimension of the central DRS under the setting M2, D2, $\sigma = 0.5$.

Figure 4.14 shows that χ^2_{SIR} works very well for D3. BIC_{SIR} tends to overestimate, whereas χ^2_{SAVE} underestimates. However, χ^2_{PHD} is the best. For D1, D2 and D4, χ^2_{SIR} underestimates. χ^2_{SAVE} works for D1 and D4 when p/N. BIC_{SIR} does not work for any distributional setting. χ^2_{PHD} works the best for D1 and D4 as well as for D3, but none of the methods work for D2.

The results for $\sigma = 1$ differ very little from those for $\sigma = 0.5$.

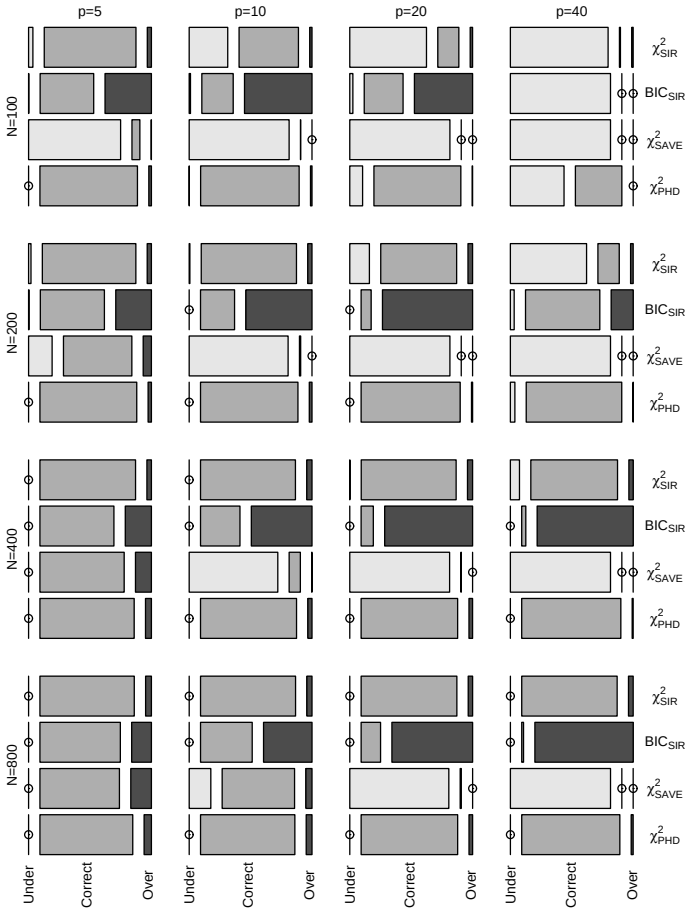

Figure 4.14. Barplot for the estimate of the dimension of the central DRS under the setting M3, D3, $\sigma = 0.5$.

Model M4

The results for M4 are quite similar to those for M3. The largest deviance between the results is that χ^2_{SIR} works for D1 as well as for D3, as long as N is large (see Figure 4.15). Also, χ^2_{PHD} does not work for D3, and consequently χ^2_{SIR} is the best for D3.

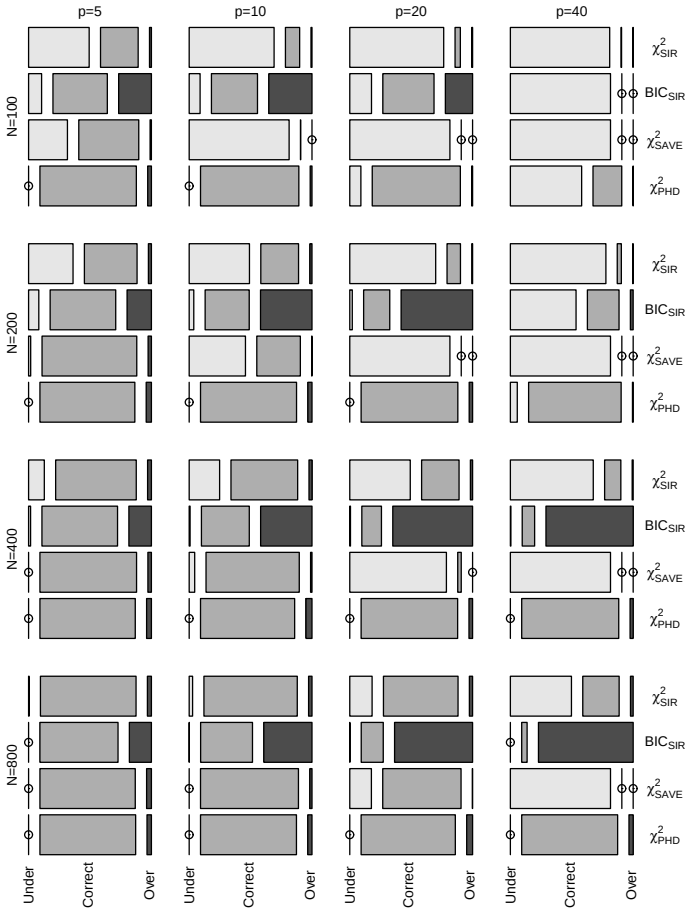

Figure 4.15. Barplot for the estimate of the dimension of the central DRS under the setting M4, D1, $\sigma = 0.5$.

Model M5

Figure 4.16 shows the following tendency of the results for M5: χ^2_{SIR} underestimates dramatically, BIC_{SIR} estimates fall rather uniformly into the three possible estimation outcomes, χ^2_{SAVE} underestimates and χ^2_{PHD} does very well.

The results for $\sigma = 1$ differ very little from the results for $\sigma = 0.5$.

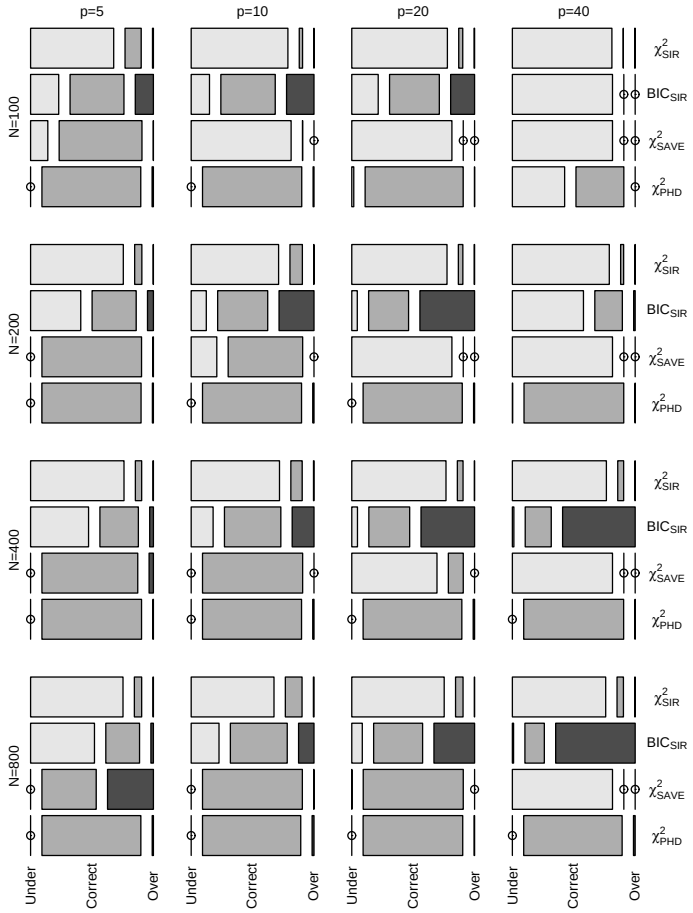

Figure 4.16. Barplot for the estimate of the dimension of the central DRS under the setting M5, D4, $\sigma = 0.5$.

Model M6

The results for M6 are quite similar to those for M5, Figure 4.17 showing the general pattern of the results. χ^2_{SIR} underestimates dramatically, BIC_{SIR} estimates distribute equally between the three outcomes, χ^2_{SAVE} underestimates and χ^2_{PHD} is the best-working method.

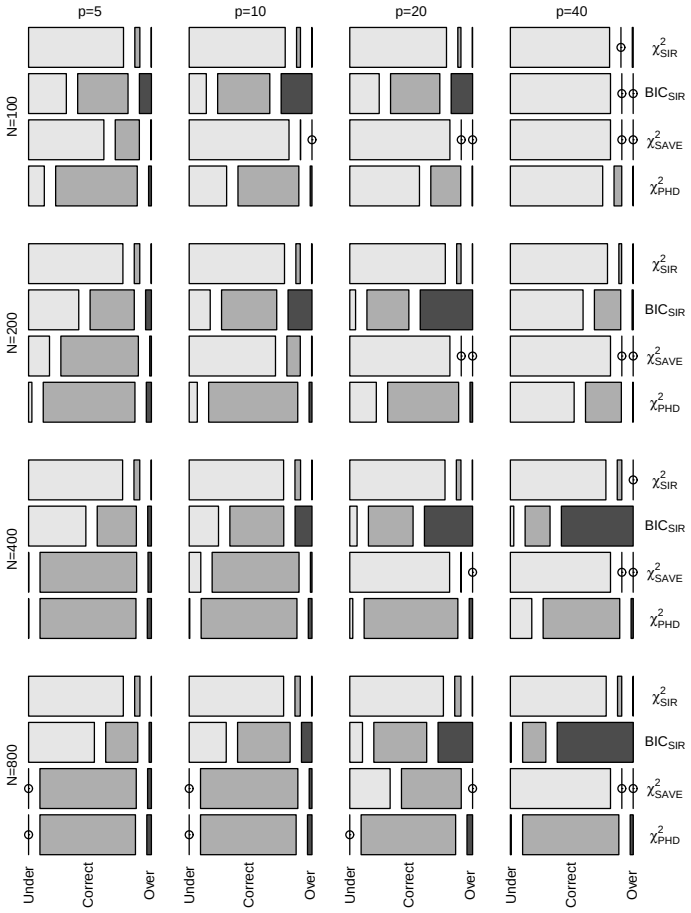

Figure 4.17. Barplot for the estimate of the dimension of the central DRS under the setting M6, D1, σ = 0.5.

Model M7

No methods seem to work for M7. The results are best for D4 – χ^2_{PHD} performs well when p/N is small, χ^2_{SIR} and χ^2_{SAVE} underestimate, and BIC_{SIR} gives estimates equally into each outcome category (see Figure 4.18).

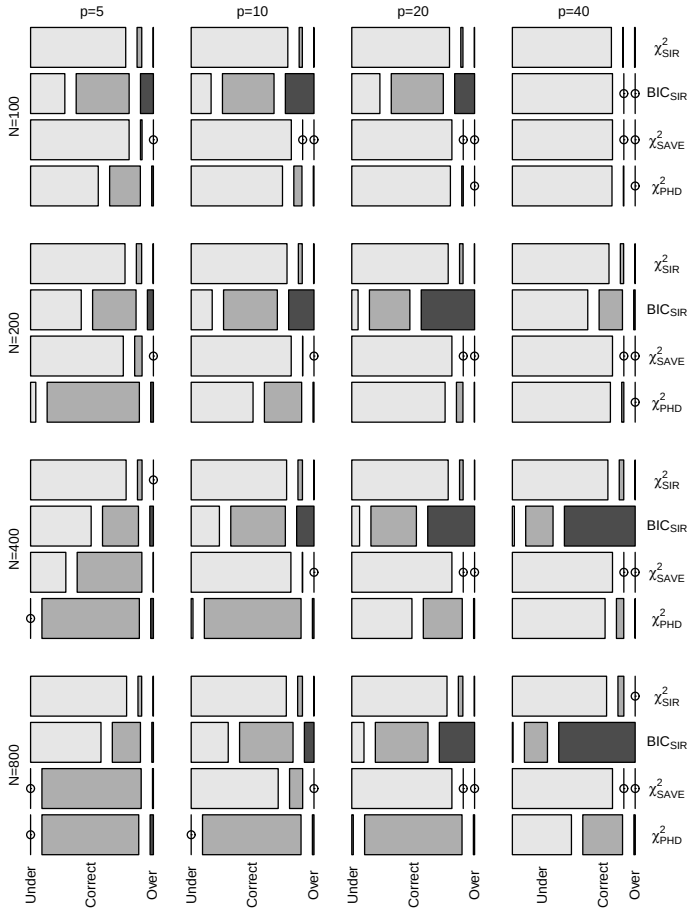

Figure 4.18. Barplot for the estimate of the dimension of the central DRS under the setting M7, D4, $\sigma = 0.5$.

4.6 Summary of the simulation study

When K is known, SIR is superior to SAVE and PHD for models M1 and M2, but for the rest of the models SIR performs worse than SAVE and PHD. PHD is superior to SIR and SAVE for models M3, M4, M5 and M7. SAVE is not really superior to both SIR and PHD for any model. For model M6 none of the methods seem to perform well.

In the more realistic case when K is unknown, SIR still performs best for models M1 and M2, but for the rest of the models PHD is quite superior to SIR and SAVE.

For both cases, when K is known and the more realistic case of K being unknown, the ratio of p/N seems important. Overall, the results are best for distributions D1 and D4, and worse for distribution D2.

5 Example

In this chapter we apply SIR, SAVE and PHD (residual-based) to a hemodynamic data set[1], which was analyzed for example in Tahvanainen et al. (2009). In short, the data is from the LASERI (Lasten Sepelvaltimotaudin Riskitekijät) study, where each of the 179 subjects was first 5 minutes in a supine position and then moved upwords for 5 minutes, and again tilted down for 5 minutes. Our data consists of the averages of the second and fifth minute for each phase. A more detailed description of the data can be found in Tahvanainen et al. (2009).

Our objective is to model the average difference to the tilt for the augmentation index AIx, which is defined as

$$Y = (AIxT3 + AIxT4)/2 - (AIxT1 + AIxT2)/2,$$

where $T1, T2, T3$ and $T4$ are the first four time points.

The number of quantitative explanatory variables is 22 and the number of categorical explanatory variables is 1 (SEX). All of the variables are introduced in Appendix D. Since the number of explanatory variables is large, we wish to reduce the dimension of the vector of explanatory variables. We use SIR, SAVE and PHD to study if they can find the central DRS by applying χ^2-tests suitable for each dimension reduction method.

It is worth mentioning that in Tahvanainen et al. (2009), a linear robust MM-regression model was fitted. This suggests that the dimension K of the central DRS was believed to be 1.

In a practical situation it is often of interest to keep only the variables really needed in the model to have an easier interpretation of the effects of different variables on the response. In this example also the categorical variable SEX should be considered when trying to reduce the dimension. These two issues will be addressed in the context of SIR in the following two sections.

5.1 Marginal coordinate tests

Marginal coordinate tests were proposed by Cook (2004). Suppose that $\mathscr{C}_{Y|\mathbf{X}}$ represents the central DRS with basis given by the columns of a $p \times K$ matrix $\mathbf{B}$, as defined in (3.2). Coordinate tests allow testing hypotheses that $\mathscr{C}_{Y|\mathbf{X}}$ is

[1]The data was kindly provided by Professor Ilkka Pörsti, Department of Internal Medicine, Tampere University Hospital, Tampere, Finland.

orthogonal to one (or more linear combinations) of the explanatory variables; were this so, we could achieve dimension reduction simply by dropping these explanatory variables.

Let us partition $\mathbf{X}' = (\mathbf{X}_1', \mathbf{X}_2')$ and $\mathbf{B}' = (\mathbf{B}_1', \mathbf{B}_2')$ according to the partition of $\mathbf{X}$. Consider a typical application to test the hypothesis that only r selected explanatory variables $\mathbf{X}_1$ contribute to the regression. By definition of $\mathscr{C}_{Y|\mathbf{X}}$ we have $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}'\mathbf{X}$. We wish to test the hypothesis $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}_1'\mathbf{X}_1$.

Define a hypothesis matrix $\mathbf{H}$. The marginal coordinate hypotheses are

$$(\mathbf{I} - \mathbf{P_H})\mathbf{B} = \mathbf{0} \text{ versus } (\mathbf{I} - \mathbf{P_H})\mathbf{B} \neq \mathbf{0},$$

where $\mathbf{P_H}$ is the projection on $\mathbf{H}$. In short, we are studying if all vectors in $\mathscr{C}_{Y|\mathbf{X}}$ can be represented as linear combinations of the columns of $\mathbf{H}$.

Marginal coordinate tests are currently defined in the *dr* package for SIR and SAVE, but not for PHD. Therefore, for PHD we do not apply marginal coordinate tests.

5.2 Categorical explanatory variables

Suppose we have also a categorical explanatory variable or grouping variable G with g levels, that might influence Y. Chiaromonte, Cook & Li (2002) described how such a categorical explanatory variable could be included in a dimension reduction regression problem. The basic idea is to divide the problem into g regression problems, and define the central DRS as the union of the subspaces for the regression problems $(Y|\mathbf{X}, G = 1), (Y|\mathbf{X}, G = 2), \ldots, (Y|\mathbf{X}, G = g)$. Chiaromonte et al. (2002) called this *partial sir*.

Categorical explanatory variables in a dimension reduction regression problem are currently defined in the *dr* package for SIR and SAVE, but not for PHD. Therefore, for PHD we do not use the categorical variable *SEX*.

5.3 Results for the hemodynamic data example

5.3.1 SIR

Ignoring at the beginning the variable SEX and using only the quantitative variables, the results suggest 0 as the dimension of the central DRS. Therefore, if the central DRS exists, SIR cannot find an estimate for it.

We continue by using marginal coordinate tests – dropping explanatory variables that have the largest p-value one by one. After each time the variable with the largest p-value is dropped, a new SIR is performed. Proceeding in this manner we end up with five variables, from which all but one have p-values less than 0.05. Performing SIR using these five variables produce results, which suggest 1 as the dimension of the central DRS.

Next, we consider the starting point of our dimension reduction procedure with all of the 22 variables, but we add the categorical variable *SEX* as an

explanatory variable. The results suggest 2 as the dimension of the central DRS.

In Tahvanainen et al. (2009), doctors with subject knowledge in the hemodynamic data presumed that there are only 12 relevant quantitative explanatory variables (*SAPr, DAPr, Waist, Hip, Cornell, CRP, Hkr, HDL, LDL, Trigly, AIx, Age*) out of the 22 variables, along with variable *GFR*, which is defined as follows

$$GFR = \begin{cases} ((140 - Age) * Weight)/Krea * 1.23, & SEX = \text{male}, \\ ((140 - Age) * Weight)/Krea * 1.04, & SEX = \text{female}. \end{cases}$$

We will continue by applying SIR to these 13 variables. SIR is performed to the whole data and to both sexes separately. SAVE and PHD will not be applied to these 13 variables, but only to the original 22 variables. Furthermore, SAVE will be applied to the categorical variable *SEX*. Note that performing a dimension reduction method to a subset of different sexes deviates from the situation where we have a categorical explanatory variable *SEX* in the model.

SIR strongly suggests 0 as the dimension of the central DRS for the whole data and for both subsets of sexes.

Previous studies have shown that SIR is sensitive to outliers (Gather, Hilker & Becker 2002). This is of relevance, because Tahvanainen et al. (2009) point out that this data might contain many outliers, since the measurements are for example sensitive to small movements of the subjects. To explore this, we show in Figures 5.1 and 5.2 Chi-square QQ-Plots for the different sexes, where the robust distances are based on the MCD-scatter matrix. As can be seen from these figures, both groups seem to have very heavy tails and seem not normally distributed. If the groups were in fact normally distributed, then all points would lie on the dashed line.

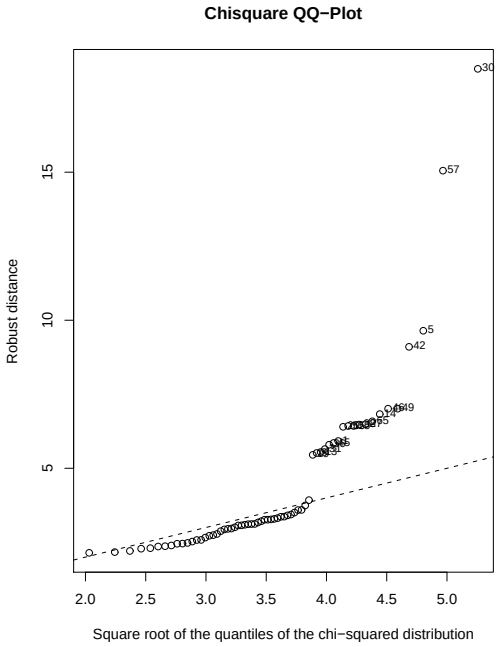

Figure 5.1. Chi-square Quantile-Quantile (QQ) Plot for evaluating the normality of the male subset of the data.

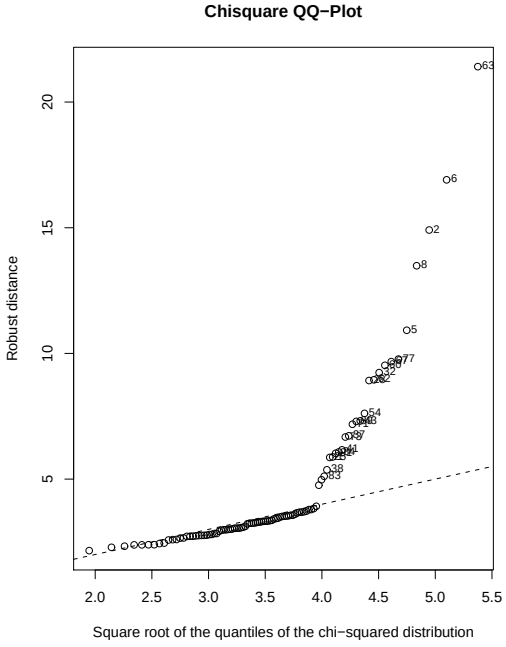

Figure 5.2. Chi-square Quantile-Quantile (QQ) Plot for evaluating the normality of the female subset of the data.

To get an idea of SIR in the clean data, we proceed by removing observations that have a robust distance larger than 5 and also observations that have missing values for any of the 13 variables. After this procedure we end up with 127 subjects. The results still suggest 0 as the dimension of the central DRS for the whole data and for both subsets of sexes, but the results especially for the whole data and for the subset of females are very close to suggesting 1 as the dimension of the central DRS.

We mentioned earlier that in Tahvanainen et al. (2009), a linear robust MM-regression model was fitted and therefore K was believed to be 1. By cleaning the data from outliers, we pursued normality. If these objectives were actually reached, we could make an interesting comparison to our simulation results for model M1, distribution D1 (see Figure 4.12). For the setting $p = 10$ and $N = 100$, namely, which is the setting closest to our situation, χ^2_{SIR} estimates K very efficiently.

5.3.2 SAVE and PHD

Applying SAVE and PHD to the 22 variables, the results suggest 0 as the dimension of the central DRS. Therefore, if the central DRS exists, then SAVE cannot find an estimate for it.

By using the marginal coordinate tests we end up with five variables, from which all but one have p-values larger than 0.05. Therefore, it seems that all variables except one seem to be orthogonal to the central DRS. After adding *SEX* as an explanatory variable, the results suggest 0 as the dimension of the central DRS.

For PHD, the estimate of the dimension of the central DRS is 0.

6 Conclusions

This thesis discusses sliced inverse regression (SIR). We begin by studying the basics of regression and dimension reduction, which are needed to build a foundation for SIR. Chapter 3 introduces the important concept of inverse regression. The other sections of Chapter 3 discuss the dimension reduction subspace (DRS), the key concept of central dimension reduction subspace (central DRS) and the main theorem of SIR, followed by the SIR algorithm based on the theory of SIR. The SIR algorithm provides a tool to obtain an estimate of the central DRS of a finite sample. We continue by comparing SIR to two other dimension reduction methods, sliced average variance estimate (SAVE) and principal hessian directions (PHD), in an extensive simulation study. Finally, we apply SIR, SAVE and PHD to a real data set.

One strength of this thesis is the extensive simulation study, which widens the view on the applicative scope of dimension reduction. Our study considers four different dimensions of the vector-valued explanatory variable $\mathbf{X}$, from 5 up to 40, four different distributions of $\mathbf{X}$, four different sample sizes, seven different models, and two different levels of noise. Moreover, the way we compare the methods is very extensive – we use the R^2 criterion, where the dimension K of the central DRS is assumed to be known, and three different χ^2-tests, one for each method, and a BIC type criterion for SIR to estimate K when it is unknown.

The results show that none of the three dimension reduction methods is superior to the others in all settings. However, SIR is superior to SAVE and PHD, when the dependence between the response Y and $\mathbf{X}$ is linear. When the dependence is symmetric, then SIR often underestimates the dimension of the central DRS and therefore performs poorly, SAVE does not work when p exceeds 10 and PHD works better for larger values of p as well. Although SIR usually performs poorly when the dependence is symmetric, for some distributional settings SIR is actually superior to SAVE and PHD. All in all, in the case of symmetric dependence, PHD is superior to SIR and SAVE. We programmed the BIC criterion to estimate the dimension K, and it turns out that this method tends to overestimate the dimension. Adding noise increases the range of values of the R^2 criterion or it does not affect much at all. In one case it strangely improves the estimation of K.

As a consequence from this simulation study we can point out that in a practical situation, the choice of dimension reduction technique will matter and its performance will depend on the underlying dependence structure as

well as on the distribution of $\mathbf{X}$. If one has some prior information about the dependence between the response and the vector of explanatory variables, the results in this thesis might help to select the right dimension reduction method. The ratio p/N also seems to be relevant. On the basis of the simulation results as a whole, PHD is superior to SIR and SAVE. However, for linear dependence between the response and the explanatory variables we suggest SIR. For symmetric dependence we suggest PHD.

Applying SIR, SAVE and PHD to a real data example does not yield an estimate of the central DRS. This is because the central DRS does not exist or these methods cannot find it.

Future studies might consist of expanding the simulation study by increasing dimension p, sample size N and the level of noise. For the estimation of K, it would be interesting to see how much each of the methods overestimate or underestimate, instead of only studying if a method underestimates or overestimates. Also, more accurate methods for estimating K would be of interest, since in practise it is usually unknown. Furthermore, as the analyses of the example data show, also robust dimension reduction methods are needed.

Acknowledgements

I wish to thank my supervisors, Professor Hannu Oja and Klaus Nordhausen, for great guidance. I would like to thank Professor Hannu Oja for giving me the opportunity to work in a truly inspiring group of people and for introducing me to a topic not only interesting, but also very challenging. I would like to thank Klaus Nordhausen for his patience in answering my many questions and especially for his expert R software guidance.

Finally, I wish to thank my father, Professor Erkki Liski, for his constant support during my studies.

Bibliography

Casella, G. & Berger, R. L. (1998), Statistical Inference, *Duxbury Press, Belmont*.

Chiaromonte, F., Cook, R. D. & Li, B. (2002), Sufficient Dimension Reduction in Regressions with Categorical Predictors, *Annals of Statistics*, 30, 475-497.

Cook, R. D. (1994), Using Dimension-Reduction Subspaces to Identify Important Inputs in Models of Physical Systems, *Proceedings of the Section on Physical and Engineering Sciences*, pp. 18-25. Alexandria, VA: American Statistical Association.

Cook, R. D. (1996), Graphics for Regressions with a Binary Response, *Journal of the American Statistical Association*, 91, 983-992.

Cook, R. D. (1998), Regression Graphics: Ideas for Studying Regression through Graphics, *John Wiley & Sons, Inc., New York*.

Cook, R. D. (2004), Testing Predictor Contributions in Sufficient Dimension Reduction, *Annals of Statistics*, 32, 1062-1092.

Cook, R. D. & Weisberg, S. (1991), Sliced Inverse Regression for Dimension Reduction: Comment, *Journal of the American Statistical Association*, 86, 328-332.

Dawid, A. P. (1979), Conditional Independence in Statistical Theory, *Journal of the Royal Statistical Society*. Series B (Methodological), 41, 1-31.

Duan, N. & Li, K. C. (1991), Slicing Regression: a Link-Free Regression Method, *The Annals of Statistics*, 19, 505-530.

Gather, U., Hilker, T. & Becker, C. (2002), A Note On Outlier Sensitivity of Sliced Inverse Regression, *Statistics*, 36, 271-281.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. & Hothorn, T. (2009). mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-7. URL http://CRAN.R-project.org/package=mvtnorm

Halmos, P. R. (1958), Finite-Dimensional Vector Spaces, *Springer, Princeton*.

Hyvärinen, A., Karhunen, J. & Oja, E. (2001), Independent Component Analysis, *John Wiley & Sons, Inc., New York*.

Johnson, R. A. & Wichern, D. W. (2002), Applied Multivariate Statistical Analysis, *Prentice-Hall, Inc., Upper Saddle River*.

Krutchkoff, R. G. (1967), Classical and Inverse Regression Methods of Calibration, *Technometrics*, 9, 425-439.

Krutchkoff, R. G. (1969), Classical and Inverse Regression Methods of Calibration in Extrapolation, *Technometrics*, 11, 605-608.

Li, K. C. (1991), Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, 86, 316-327.

Li, K. C. (1992), On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma, *Journal of the American Statistical Association*, 87, 1025-1039.

Li, K. C. (2000), High Dimensional Data Analysis via the SIR/PHD Approach, UCLA Department of Statistics. Available from Internet: http://www.stat.ucla.edu/ kcli/sir-PHD.pdf.

Li, K. C. & Duan, N. (1989), Regression Analysis Under Link Violation, *Annals of Statistics*, 17, 1009-1052.

Meyer, D., Zeileis, A. & Hornik, K. (2009). vcd: Visualizing Categorical Data. R package version 1.2-4.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sarkar. D. (2009). lattice: Lattice Graphics. R package version 0.17-22. http://CRAN.R-project.org/package=lattice

Schott, J. R. (2005), Matrix Analysis for Statistics, *John Wiley & Sons, Inc., Hoboken*.

Shao, Y., Cook, R. D. & Weisberg, S. (2007), Marginal Tests with Sliced Average Variance Estimation, *Biometrica*, 94, 285-296.

Tahvanainen, A., Leskinen, M., Koskela, J., Ilveskoski, E., Nordhausen, K., Oja, H., Kähönen, M., Kööbi, T., Mustonen, J. & Pörsti, I. (2009), Ageing and Cardiovascular Responses to Head-Up Tilt in Healthy Subjects. *Atherosclerosis (2009)*, doi:10.1016/j.atherosclerosis.2009.06.001

Velilla, S. (1998), Assessing the Number of Linear Components in a General Regression Problem, *Journal of the American Statistical Association*, 93, 1088-1098.

Wand, M. P. & Jones, M. C. (1995), Kernel Smoothing, *Chapman & Hall, London*.

Weisberg, S. (2008). dr: Methods for Dimension Reduction for Regression. R package version 3.0.3. http://www.r-project.org

Zhu, L., Miao, B. & Peng, H. (2006), On Sliced Inverse Regression With High-Dimensional Covariates, *Journal of the American Statistical Association*, 101, 630-643.

Zhu, L. P. & Zhu, L. X. (2007), On Kernel Method for Sliced Variance Estimation, *Journal of Multivariate analysis*, 98, 970-991.

Appendix A: Principal component analysis

We present principal component analysis by following the results from Schott (2005, p.107).

Let $\mathbf{X}$ be an $p \times 1$ random vector having the covariance matrix $\mathbf{\Sigma}$. Suppose that we wish to find the $p \times 1$ vector $\mathbf{a}_1$ so as to make the variance of $\mathbf{a}_1'\mathbf{X}$ as large as possible. The variance of $\mathbf{a}_1'\mathbf{X}$ is

$$var(\mathbf{a}_1'\mathbf{X}) = \mathbf{a}_1'\{cov(\mathbf{X})\}\mathbf{a}_1 = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1.$$

Clearly, we can make this variance arbitrarily large by taking $\mathbf{a}_1 = \alpha\mathbf{c}$ for some scalar α and some vector $\mathbf{c} \neq \mathbf{0}$ and then let $\alpha \longrightarrow \infty$. We will remove this effect of the scale of $\mathbf{a}_1$ by imposing the constraint $\mathbf{a}_1'\mathbf{a}_1 = 1$. In this case, we are searching for the one direction in $\mathbb{R}^p$ that is the line for which the variability of observations of $\mathbf{X}$ projected onto that line is maximized. It can be proved (see Schott 2005, Theorem 3.17) that this direction is given by the normalized eigenvector of $\mathbf{\Sigma}$ corresponding to its largest eigenvalue. Suppose we also wish to find a second direction, given by $\mathbf{a}_2$ and orthogonal to $\mathbf{a}_1$, where $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $var(\mathbf{a}_2'\mathbf{X})$ is maximized. It can be shown (see Schott 2005, Theorem 3.17), that this second direction is given by the normalized eigenvector of $\mathbf{\Sigma}$ corresponding to its second largest eigenvalue. Continuing in this fashion, we would obtain p directions identified by the set $\mathbf{a}_1, \ldots, \mathbf{a}_p$ of orthonormal eigenvectors of $\mathbf{\Sigma}$. Effectively, what we have done is to find a rotation of the original axes to a new set of orthogonal axes, where each successive axis is selected so as to maximize the dispersion among the observations of $\mathbf{X}$ along that axis. Note that the components of the transformed vector $(\mathbf{a}_1'\mathbf{X}, \ldots, \mathbf{a}_p'\mathbf{X})'$, which are called the principal components of $\mathbf{\Sigma}$, are uncorrelated because for $i \neq j$,

$$cov(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_j = \mathbf{a}_i'(\lambda_j\mathbf{a}_j) = \lambda_j\mathbf{a}_i'\mathbf{a}_j = 0,$$

where λ_j's $j = 1, \ldots, p$ are the eigenvalues of $\mathbf{\Sigma}$.

Appendix B: Conditional expectation

Let X, Y and Z be univariate random variables. We write $E(X|Y = y, Z = z), f(x|y,z), f(x,y,z), f(z|y), f(y,z), f(x,y), f(x|y), f_Y(y)$ and $E(X|Y = y)$, respectively, to denote the expected value of X given $Y = y$ and $Z = z$, the conditional density function of X given $Y = y$ and $Z = z$, the joint density function of X, Y and Z, the conditional density function of Z given $Y = y$, the joint density function of Y and Z, the joint density function of X and Y, the conditional density function of X given $Y = y$, the marginal density function of Y and the expected value of X given $Y = y$. Now the following useful identity holds, providing that the expectations exist:

$$E(E(X|Y,Z)|Y) = E(X|Y). \tag{1}$$

To prove (1) we need to compute $E(E(X|Y,Z)|Y = y)$. $E(X|Y,Z)$ is a random variable, and for $Y = y$ its possible values are $E(X|Y = y, Z = z)$, where z varies over the range of Z. Now

$$\begin{aligned}
E(E(X|Y,Z)|Y = y) &= \int E(X|Y = y, Z = z) f(z|y)\, dx \\
&= \int\int x f(x|y,z) f(z|y)\, dx\, dz \\
&= \int\int x \frac{f(x,y,z)}{f(y,z)} \frac{f(z,y)}{f_Y(y)}\, dx\, dz \\
&= \int\int x \frac{f(x,y,z)}{f_Y(y)}\, dx\, dz \\
&= \int x \frac{f(x,y)}{f_Y(y)}\, dx \\
&= \int x f(x|y)\, dx \\
&= E(X|Y = y).
\end{aligned}$$

This proves (1).

Proposition 1. If the conditional independence $Y \perp\!\!\!\perp \mathbf{X}|Z$ holds, then

$$E(\mathbf{X}|Y) = E(E(\mathbf{X}|Z)|Y). \tag{2}$$

We prove Proposition 1 in the case of $p = 1$.

Proof. It is sufficient to prove the identity $E(E(X|Y,Z)|Y) = E(E(X|Z)|Y)$, since then our claim (2) follows by (1). We have by direct calculation

$$\begin{aligned} E(E(X|Y,Z)|Y=y) &= \int E(X|Y=y,Z=z)f(z|y)\,dz \\ &= \int\int xf(x|y,z)f(z|y)\,dx\,dz \\ &= \int\int x\frac{f(x,y,z)}{f(y,z)}\frac{f(z,y)}{f_Y(y)}\,dx\,dz \\ &= \int\int x\frac{f(x,y,z)}{f_Y(y)}\,dx\,dz. \end{aligned}$$

Since $f(x,y,z) = f_Z(z)f(x,y|z)$ and $f(x,y|z) = f(x|z)f(y|z)$ by conditional independence $Y \perp\!\!\!\perp X|Z$ (see identity (5) in Appendix C), we obtain

$$\begin{aligned} \int\int x\frac{f(x,y,z)}{f_Y(y)}\,dx\,dz &= \int\int x\frac{f_Z(z)f(x|z)f(y|z)}{f_Y(y)}\,dx\,dz \\ &= \int\int x\frac{f_Z(z)f(x|z)}{f_Y(y)}\frac{f(y,z)}{f_Z(z)}\,dx\,dz \\ &= \int\int xf(x|z)\frac{f(y,z)}{f_Y(y)}\,dx\,dz \\ &= \int\int xf(x|z)f(z|y)\,dx\,dz \\ &= \int E(X|Z=z)f(z|y)\,dz \\ &= \int E(E(X|Z)|Y=y), \end{aligned}$$

which yields the required result. Here $E(X|Y=y,Z=z), f(x|y,z), f_Z(z)$, $f(x|z), f(y|z), E(X|Z=z)$ and $E(E(X|Z)|Y=y)$ denote, respectively, the expected value of X given $Y=y$ and $Z=z$, the conditional density function of X given $Y=y$ and $Z=z$, the marginal density function of Z, the conditional density function of X given $Z=z$, the conditional density function of Y given $Z=z$, the expected value of X given $Z=z$ and the expected value of $E(X|Z)$ given $Y=y$.

□

Appendix C: Independence and conditional independence

The bulk of the following results are from Dawid (1979).

Introduction

Consider three continuous random variables A, B and C, and their probability functions $f(a), f(b)$ and $f(c)$, respectively, for any values $A = a$, $B = b$ and $C = c$. The joint probability function of the three variables can be written as

$$f(a, b, c) = f(a)f(b, c|a) = f(a)f(c|a)f(b|a, c).$$

As we can see from the latter form, C is conditional on A and B is conditional on A and C. This situation is presented in Figure 1.

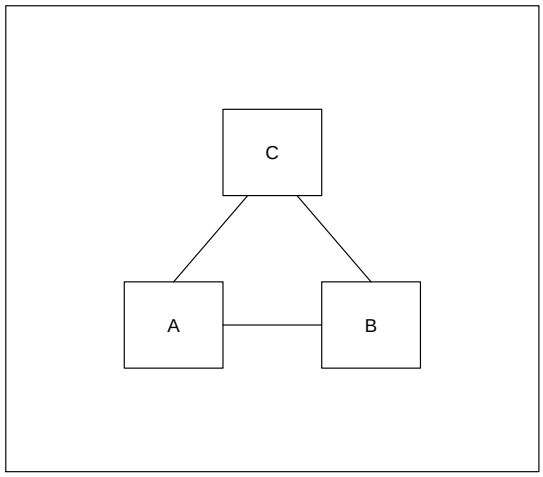

Figure 1. B is conditional on A and C.

If however B is conditionally independent of C given A, we could write $f(b|a, c) = f(b|a)$. In this case we could remove the arrow going from C to B from Figure 1:

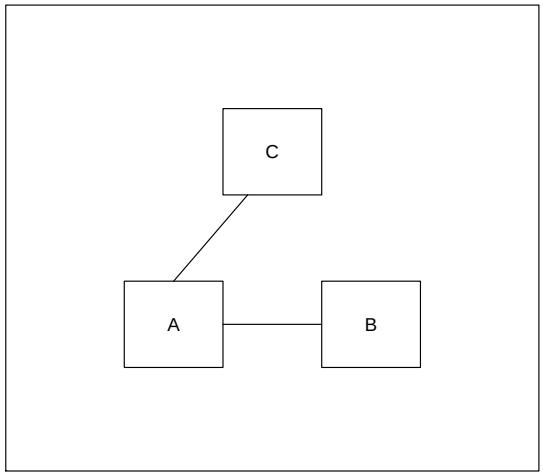

Figure 2. B is conditionally independent of C given A.

Independence

Let X and Y be univariate and continuous variables and denote $f(x, y)$, $f(x|y)$ and $f(x)$, respectively, the joint density function of X and Y, conditional density function of X given Y and the marginal density function of X. We write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent. That is,

$$f(x, y) = f_X(x) f_Y(y), \text{ for all values of x and y.}$$

Let's examine the conditional density function $f(x|y)$. If X and Y are independent, that is, if any information about Y does not change the uncertainty about X, we have for $f_Y(y) > 0$ that

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_X(x) f_Y(y)}{f_Y(y)} = f_X(x),$$

which is the marginal density function of X.

Conditional independence

Now we introduce a third variable Z, which is also univariate and continuous. We write $X \perp\!\!\!\perp Y|Z$ to denote that X and Y are *conditionally independent* given Z. Let us first write the joint density function of X, Y and Z as

$$f(x, y, z) = f(z) f(x, y|z).$$

Let us study more closely the expression $f(x, y|z)$. We can write this in another way by first introducing expressions

$$f(x|y,z) = \frac{f(x,y,z)}{f(y,z)} \tag{3}$$

and

$$f(y|z) = \frac{f(y,z)}{f_Z(z)}. \tag{4}$$

for $f(y,z) > 0$ and $f_Z(z) > 0$, respectively.

Multiplying (3) by (4) yields

$$f(x|y,z)f(y|z) = \frac{f(x,y,z)}{f(y,z)}\frac{f(y,z)}{f(z)} = \frac{f(x,y,z)}{f(z)} = f(x,y|z).$$

Now we have

$$f(x,y,z) = f_Z(z)f(x,y|z) = f_Z(z)f(y|z)f(x|y,z).$$

Lets study the expression $f(x|y,z)$. By definition, X is conditionally independent of Y given Z if the distribution of X is completely determined by Z alone, Y being superfluous. Then we can write

$$f(x,y|z) = f(y|z)f(x|y,z) = f(y|z)f(x|z), \tag{5}$$

which is a also a notation for $X \perp\!\!\!\perp Y|Z$ along with solely the expression

$$f(x|y,z) = f(x|z).$$

And finally, in the case of conditional independence, the joint density function of X, Y and Z can be written as

$$f(x,y,z) = f_Z(z)f(x,y|z) = f_Z(z)f(y|z)f(x|y,z) = f_Z(z)f(y|z)f(x|z).$$

Appendix D: Variables in the hemodynamic data

Table 1. Categorical variables in the data.

Variable	Total number of subjects	Female	Male
SEX	179	109	70

Table 2. Quantitative variables in the data.

Variable	Median (Quartiles)
Age (years)	43.00 (33.50, 51.00)
Weight (kg)	73.7 (65.0, 86.2)
Height (cm)	170.8 (165.0, 178.0)
Waist (cm)	86.00 (77.25, 98.75)
Hip (cm)	100 (96, 107)
Cornell (Cornell voltage product (ms × mm))	108.0 (100.0, 116.8)
Hkr (Hematocrit (% of blood volume formed by blood cells))	0.4100 (0.3900, 0.4400)
Krea (Creatinine (μmol/l))	73.00 (65.00, 82.00)
CRP (C-reactive protein (mmol/l))	0.800 (0.500, 1.500)
Trigly (Triglycerides (mmol/l))	0.960 (0.650, 1.420)
HDL (HDL-cholesterol (mmol/l))	1.59 (1.30, 1.90)
LDL (LDL-cholesterol (mmol/l))	2.700 (2.100, 3.300)
HR (Heart rate: baseline)	61.75 (56.17, 69.82)
CI (Cardiac index: baseline)	2.839 (2.450, 3.196)
SVRI (Systemic vascular resistance index: baseline)	2168 (1837, 2537)
SI (Stroke index: baseline)	45.04 (40.02, 50.81)
SAPr (Systolic blood pressure)	132.5 (122.3, 138.7)
DAPr (Diastolic blood pressure)	80.67 (74.42, 86.58)
PPa (Pulse pressure at the aortic level)	36.84 (32.17, 42.43)
TRa (Time of pressure wave reflection in the aorta)	148.5 (140.4, 162.6)
AIx (Augmentation index: baseline)	20.79 (11.27, 29.00)
PWV (Pulse wave velocity: baseline)	9.577 (8.560, 11.640)
GFR (Glomerular filtration rate)	110.70 (96.74, 130.30)