The Pennsylvania State University

The Graduate School

Eberly College of Science

TOPICS ON SUPERVISED AND UNSUPERVISED DIMENSION REDUCTION

A Dissertation in

Statistics

by

Andreas A. Artemiou

Submitted in Partial Fullfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2010

The dissertation of Andreas A. Artemiou was reviewed and approved* by the following:

Bing Li
Professor of Statistics
Dissertation Adviser
Chair of Committee

Runze Li
Professor of Statistics
Graduate Studies Chair

Michael G. Akritas
Professor of Statistics

Francesca Chiaromonte
Associate Professor of Statistics and Public Health Sciences

Adam D. Smith
Assistant Professor of Computer Science and Engineering

*Signatures are on file with Graduate School

Abstract

In the first part of this work, we extend the results by Artemiou and Li (2009) and Ni (2010) in several interesting ways. First we extend them in the case of non random covariance matrix and in case there is a multivariate response in the linear regression setting. Second we try to explore if there is predictive potential of linear principal components in the case of non linear regression functions and especially in the context of sufficient dimension reduction. Third we propose an information criterion that in very limited number of cases can be used to check the predictive potential of linear principal components. Lastly, we explore the predictive potential of kernel principal component in the completely nonparametric regression function $Y = f(\boldsymbol{X}) + \epsilon$ where f is an arbitrary function. The most general form of our result, shows that the phenomenon goes far beyond the context of linear regression and classical principal components where it was originally noticed: if nature selects an arbitrary distribution for the predictor $\boldsymbol{X}$ and an arbitrary conditional distribution of the response Y given $\boldsymbol{X}$, then Y tends to have stronger correlation with higher-ranking kernel principal components than with lower-ranking kernel principal components. These two questions need the arbitrariness of function f and the arbitrariness of matrix $\boldsymbol{\Sigma}$ which are achieved by unitary invariance. A small data analysis, shows that this tendency holds in three different databases.

In the second part, SVMIR, a new method for sufficient dimension reduction using inverse regression and support vector machine algorithms is proposed. This method is known to have several advantages, in comparison, to previous inverse regression methods like SIR, SAVE and DR. First, since machine learning methods instead of sample moments are applied in estimating the directions in the Central Dimension Reduction Subspace this method is shown to be robust in the presence of outliers.

Second, through a modification of the objective function that we need to minimize, we can show that dimension reduction without matrix inversion can be achieved. Third, through simulations our method is shown to be robust in departures from ellipticity as well as in the presence of categorical predictors among our predictors. Finally, this method gives us a way of estimating nonlinear directions in the Central Dimension Reduction Subspace, and direction in the feature space using kernel functions. The above are shown in theory, through simulations and by application on real data examples; one to build a regression model for the relative performance of computer CPUs, the second for the classification of *E.coli* proteins on cellular localization sites.

Key Words and Phrases Kernel principal components; Regression; Unitary invariance; Sufficient Dimension Reduction; Support Vector Machines.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This work is the product of five years of research at the Department of Statistics at Pennsylvania State University.

I would like to thank my advisor, Professor of Statistics, Bing Li, for holding my hand since I was a "baby researcher". For helping me since I was just doing literature review for my Master Thesis until the final results of this work. Without his guidance and direction this work would not have bean a realization anywhere close to the time it took us to complete it.

Second, I would like to thank the members of my committee for their useful comments.

Third, I would like to thank everyone in the Statistics Department. All the Professors, the staff and the graduate students I met during my studies. Some memories will be unforgettable. Special thanks to Professor Emeritus William Harkness. Since the moment we stepped in the department, he showed his support to all graduate students, but more importantly for making available some of his own money to support graduate students for their travel to conferences giving them the opportunity to meet people, exchange ideas and get involved in the Statistics community.

I would like to thank the members of Holy Trinity Orthodox Church. For praying with me and for me. Especially, the Torbic family for being a second family to me since I moved in State College.

My family. My sisters. My parents for funding my trips back home twice a year. It was great to have sometime to relax back home.

Last but not least, my fiancee, Andria, for bearing my ups and downs. For listening to me, when all my nerves where in full tense and caring about me all this time. For getting out of her dream, to chase mine.

Dedication

TO THOSE,
WHOM THEIR PRESENCE IN MY LIFE,
WAS SOMETHING MORE THAN JUST A SHADOW

"TAKE THE RIGHT WAY MY SON,
AND A POWERFUL PEN"
IOANNIS CH. APOSTOLIDES

PART I

ON THE PREDICTIVE POTENTIAL OF LINEAR AND KERNEL PRINCIPAL COMPONENTS

Chapter 1

Introduction

1.1 History of Principal Components

The main idea of principal component analysis is to reduce the dimension of data sets that consist of many correlated variables. Usually, if we have n variables in the original data set, our objective is to find a set of $d(\ll n)$ new variables that are independent and at the same time describe as much as possible the variation in the original data set. These d new variables are linear combinations of the original variables and are called the principal components (PC). The procedure to find them is called principal component analysis (PCA).

Most statisticians agree that the earliest descriptions of PCA were given by Pearson (1901) and later by Hotelling (1933). Cook (2007) notes that there is an indication of principal components in the work by Adcock (1878) who wrote about the "principal axis" as the "most probable position of the straight line determined by the measured coordinates, ..., of n points". But Joliffe (2002) states, that "...

Preisendorfer and Mobley (1988) go even earlier and say that Beltrami (1873) and Jordan (1874) derived the singular value decomposition in a way that implies PCA." So, one can say that PCA was something people had been using, well before it was mathematically justified.

The absence of computing power set aside the development and further use of PCA for almost 30 years after Hotelling's work. Indeed, as Pearson (1901) noted, computation becomes difficult when the original data set consists of more than four variables. Scientists became interested in PCA again around mid 1960's when the obstacles of computation were overcome. Some works, such as Rao (1964), made important improvements in the PCA methods and motivated more researchers to study PCA, its theory and applications.

In recent years, researchers try to expand Principal Components beyond the well known applications that we have been using them since they were first introduced. For example, Jong and Kotz (1999) illustrate the relationship between the extra sum of squares in regression and the eigenvalues that are related with principal components. Tipping and Bishop (1999), present an EM algorithm that helps them find the principal axis. Their study can be considered an extension of the works by Lawley (1953) and Anderson and Rubin (1956) where principal component analysis is viewed as a maximum likelihood procedure on a probability density of the observed data.

Finally, there is also an extensive work on the idea of several nonlinear principal components methods, like kernel principal components which were introduced by Schölkopf, B., Smola, A., Müller (1997, 1998) and it is one of the most widely used method for nonlinear unsupervised dimension reduction. The idea is to map the observed predictor vectors into a higher dimensional space and then perform

linear principal component analysis in the higher dimensional space. Kernel principal components and their predictive potential is explored in Chapter 5. Also, other approaches to nonlinear feature extraction based on the principal component methodology, that are not explored further in this work, are principal curves (Hastie and Stuetzle, (1989)) and functional principal component analysis (Rice and Silverman, (1991) and Silverman, (1996)).

1.2 How Principal Component Analysis works

Principal Component Analysis is simple and easy to understand. Let $\boldsymbol{X}$ be a p-dimensional vector which denotes the original variables in a data set. Let also $\boldsymbol{\Sigma}$ to denote the covariance matrix of $\boldsymbol{X}$, that is $\boldsymbol{\Sigma} = \text{cov}(\boldsymbol{X})$.

To find the principal components of $\boldsymbol{X}$ one first finds the eigenvalues of $\boldsymbol{\Sigma}$. Denote those eigenvalues as $\lambda_i, i = 1, ..., p$ and for simplicity (and without loss of generality) assume $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$. Then using the equation $(\Sigma - \lambda_i I)\, \boldsymbol{v} = 0$ for each eigenvalue $\lambda_i, i = 1, ..., p$ separately, we can find the corresponding eigenvector $\boldsymbol{v}_i, i = 1, ..., p$.

The i^{th} principal component can be found by multiplying the eigenvector corresponding to λ_i (the i^{th} largest eigenvalue) with the variable vector $\boldsymbol{X}$. That is, the first principal component is $\boldsymbol{v}_1^T \boldsymbol{X}$, the second principal component is $\boldsymbol{v}_2^T \boldsymbol{X}$ and so on. Since the eigenvalue λ_i is proportional to the length of the i^{th} longest axis of the p-dimensional ellipsoid represented by $\boldsymbol{\Sigma}$, the first principal component explains most of the variation in the data, and so on.

The first principal component is sometimes called "***the principal component***".

As mentioned earlier, the main use of principal components is to reduce the dimension of $\boldsymbol{X}$. This can be done by selecting the first $d \ll p$ of the principal components. There are many ways to determine d. Usually, one can choose to keep only the principal components that account for a certain percentage (usually 80% to 90%) of the total variation, or to keep only the principal components that corresponds to the eigenvalues that are larger than a certain cutoff point (usually 1). There are many other subjective and inferential methods to determine d. The reader is referred to Joliffe (2002) Chapter 6 for details. Whatever way the principal components are selected, if d is not small enough, the reduction that is achieved may not be very useful.

1.3 Principal Components in Regression

Regression is the procedure we use in Statistics to find the relationship between a set of variables, called the predictors, and a variable, called response. Although there can be a multivariate response, in this introduction, the focus of the analysis is on univariate responses.

The use of principal components in regression is popular when we have a large number of predictors that make the regression analysis and statistical inference on the original predictors difficult. Moreover, if there is multicollinearity between the original predictors, we prefer to use the principal components, since they are uncorrelated, and we can avoid multicollinearity. (This causes other problems such as biased estimators for the coefficients of the regression, but this is minimal compared with the advantage we gain by avoiding multicollinearity).

Although not introducing his principal axis in terms of regression, Pearson

(1901) can be considered the first one who thought about principal components in a regression context. In his work he mentioned the following property

> "The best-fitting straight line to a system of points coincides in direction with the maximum axis of the correlation ellipsoid..."

Later, researchers discovered more properties of the principal components. The principal components as we are using them today were introduced by Hotelling (1933). In that work, he was interested in finding vectors $\boldsymbol{a}_1, ..., \boldsymbol{a}_p$ so that, $\boldsymbol{a}_i^T \boldsymbol{X}$ has maximum variance subject to the condition that cov $\left(\boldsymbol{a}_i^T \boldsymbol{X}, \boldsymbol{a}_j^T \boldsymbol{X}\right) = 0, j = 1, ...i-1$. Also, Kendall (1957), explained why doing regression using the principal components instead of the original predictors helps us for a better and easier interpretation of the effect of each principal component on the response, since they are mutually independent. It is clear that adding more principal components to our regression model the effect of each of the previous principal component will stay unaffected, while in the original predictors the effect can vary dramatically by adding a new variable, especially when there is multicollinearity among the predictors. On the other hand, one can argue that in case the principal components have no clear meaning the interpretation of the regression model can become difficult.

The fact that the interpretation doesn't change by adding principal components is very important, since in case of multicollinearities in the original predictor, by deleting the principal components that explain a small amount of the variance can give us better and more stable estimation for the coefficients. We can keep in our model only those predictors that have variance larger than a cutoff point. Another more sophisticated idea of doing this is by using variance inflation factors (VIF's) for the p predictor variables. If VIF's are close to 1 that means we have a good model, if VIF's are much larger than 1 then we delete the variables that have large

VIF. We subtract all those predictors that have VIF larger than a cutoff point.

1.4 Historic Debate

For the reader to completely understand the problem being attacked on this work, a presentation the debate that is actually still going on, between some scientists, is given. This debate was presented in Cook (2007) in greater details.

The debate seems to begin from the practice of regressing Y on the first few principal components of $\boldsymbol{X}$, as suggested and advocated in Kendall (1957). This idea is also supported by Mosteller and Tukey (1977) who while they identify the flaw of the procedure they believe that

> "A malicious person who knew our x's and the plan for them could always invent a y to make our choices look horrible. But we do not believe nature works that way..."

That is, they say that although there might be a problem on the way a malicious person can choose the response variable, nature is not a malicious person and it is more than fair in choosing the correct response for the predictors. Therefore, they believe that most of the cases, regressing on the principal components extracted from the principal components analysis will work fine. These ideas seems to be shared by others as well, like Hocking (1976) and Scott (1992).

On the other hand, there is Cox (1968) who clearly states that

> "A difficulty seems to be that there is no logical reason why the dependent variable should not be closely tied to the least important principal

component."

That is, he does not see why one can trust principal component analysis, if there is nothing to ensure that it will give us the best linear combination of the predictors in the end. The idea is shared by other scientists such as Hotelling (1957) and Hawkins and Fatti (1984). Moreover, Joliffe (1982) and Hadi and Ling (1998), showed by examples that deciding on the number of principal components solely based on the variance they explain, can actually be flawed. That is, sometimes the components with smaller variances can be the ones that are highly correlated with the response Y. In such case, dropping the principal components that have small variances will result in dropping a predictor that is highly correlated with the response. Although this has caused a growing debate over the years on the appropriateness of the method and there were plenty of discussions on what might be the phenomenon causing this to happen, it seems that there is not a satisfactory answer on how to solve this problem. Principal component analysis, though, is still being used as a dimension reduction technique in regression. It is really interesting that there is very little work done in identifying how often we get the wrong answer.

The reason this happens is clear to all scientists. The problem starts from the way principal components are calculated. Principal components are calculated, as explained earlier, using the covariance matrix of the predictors $\boldsymbol{X}$. We first order the eigenvalues and for each eigenvalue we calculate the respective eigenvector. Finally, multiplying the ordered eigenvectors (which are ordered beginning from the one corresponding to the largest eigenvalue) by the predictor vector $\boldsymbol{X}$ we get the principal components. As one can easily recognize, the predictor Y has nothing to do in any direct or indirect way in the calculation of principal components. That's why, as Cox (1968) said it, there is no logical reason why the first few principal

components should be highly correlated with the response variable and the least principal components should be less correlated with the predictor.

This question has received renewed interest recently due to the need for handling regression problems with very high dimensional predictors but relatively few observation units, as one encounters when analyzing microarray data, so that the sample covariance matrix of $\boldsymbol{X}$ is singular and the usual regression techniques cannot be directly applied. Under these circumstances regressing Y on the first few principal components is a practical solution and often gives reasonable results. For example, Chiaromonte and Martinelli (2002), are presenting a dimension reduction algorithm, which uses principal component analysis, to analyze gene expression and Bura and Pfeiffer (2003) are using another algorithm for class prediction of tumor status. Both works are dealing with microarray data and the algorithms find linear combinations of genes, in order to minimize the dimension and achieve the desired outcome.

In a comment of the paper by Cook (2007), Li (2007) gave a conjecture about the correlation between principal components and regression. This conjecture is presented in the next section.

1.5 Conjecture

Li (2007), in his comment on Cook (2007) made a conjecture in an attempt to explain probabilistically why the response should be related to the leading principal components of the predictors. It was stated roughly as follows:

> If nature arbitrarily selects a covariance matrix $\boldsymbol{\Sigma}$ for $\boldsymbol{X}$ and coefficients $\boldsymbol{\beta}$ for the regression of Y on $\boldsymbol{X}$, then the principal components of $\boldsymbol{X}$ of

higher ranks tend to have stronger correlations with Y than do those of lower ranks.

Intuitively, Li (2007) argued that if $\boldsymbol{X}$ is concentrated on a single direction, then the only way for Y to be correlated with $\boldsymbol{X}$ at all is to be correlated with its first principal component. Likewise if $\boldsymbol{X}$ has an elongated distribution the $\boldsymbol{X}$ components in the longer axes should on average bear stronger correlations with Y. Now if $\boldsymbol{\Sigma}$ is selected arbitrarily then $\boldsymbol{X}$ would have a large probability of having an elongated distribution, and would therefore affect the similar probabilistic ordering of correlations, even if the relation between Y and $\boldsymbol{X}$ is independent of the shape of the distribution of $\boldsymbol{X}$. He demonstrated this conjecture by several simulation studies, which invariably supported it.

1.6 Formulation of the conjecture

In this section we present the main results as they appeared by Artemiou and Li (2009). More results can be found in Artemiou (2008). For completeness we also give the definitions that were presented in the aforementioned works. Those results formulate the conjecture of the previous section into a theorem as the last result in this section shows, that is Theorem 1.6.1.

Definition 1.6.1 *Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$, be p random elements. We say that they are exchangeable if, for any permutation $(i_1, \ldots, i_p)$ of $(1, \ldots, p)$, we have*

$$(\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_p}) \stackrel{\mathcal{D}}{=} (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p).$$

Definition 1.6.2 *We say that a $p \times p$ positive definite random matrix $\boldsymbol{\Sigma}$ has an*

orientationally uniform distribution if

$$\boldsymbol{\Sigma} = \sigma_1^2 \boldsymbol{v}_1 \boldsymbol{v}_1^T + \cdots + \sigma_p^2 \boldsymbol{v}_p \boldsymbol{v}_p^T,$$

where each $(\sigma_i^2, \boldsymbol{v}_i)$ *is a pair of random elements in which* σ_i^2 *is a positive random variable and* $\boldsymbol{v}_i$ *is a* p*-dimensional random vector, such that*

1. $(\sigma_1^2, \ldots, \sigma_p^2)$ *are exchangeable, and its distribution is dominated by the Lebesgue measure,*
2. $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ *are exchangeable, and* $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p\}$ *is an orthonormal set,*
3. $(\sigma_1^2, \ldots, \sigma_p^2)$ *and* $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ *are independent.*

Lemma 1.6.1 *Suppose* $\boldsymbol{\beta}$ *and* $\boldsymbol{v}_1, \boldsymbol{v}_2$ *are* p*-dimensional random vectors such that*

1. $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{v}_1, \boldsymbol{v}_2)$*;*
2. $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set* G*.*
3. $\boldsymbol{v}_1$ *and* $\boldsymbol{v}_2$ *are linearly independent and exchangeable.*

Then $(\boldsymbol{\beta}^T \boldsymbol{v}_2)^2 / (\boldsymbol{\beta}^T \boldsymbol{v}_1)^2$ *has a unique median, which equals 1.*

Theorem 1.6.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *orientationally uniform random matrix,*
2. $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$*,*

3. $Y = \boldsymbol{\beta}^T \boldsymbol{X} + \delta$, *where* $\boldsymbol{\beta}$ *is a p-dimensional random vector and* δ *is a random variable such that* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$, $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, $E(\delta) = 0$ *and* $\mathrm{var}(\delta) < \infty$.

4. $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set* $G \in \mathbb{R}^p$.

Let $w_1, \ldots, w_p$ *be the 1st, . . . , pth principal components of* $\boldsymbol{X}$, *and let* $\rho_i = \rho_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathrm{corr}^2(Y, w_i | \boldsymbol{\beta}, \boldsymbol{\Sigma})$. *Then, whenever* $i < j$, $P(\rho_i \geq \rho_j) > 1/2$.

1.7 Discussion of previous results

This theorem shows that Principal Component Analysis can be used to reduce the number of predictors of the regression model. Although, it doesn't prove that PCA is always effective on finding the most correlated principal components with the response, it proves that the probability that a principal component corresponding to a largest eigenvalue to be more correlated with the response variable, is greater than the probability of a principal component that corresponds to a smaller eigenvalue.

As Artemiou (2008) mentioned, the theorem is a very useful tool, that provides at least enough evidence why the principal components that can be found by principal component analysis are probabilistically more correlated with the response. But since this, is only based on probability it gives an answer, as to why (quoted by Mosteller and Tukey (1977))

> "A malicious person who knew our x's and our plan for them could always invent a y to make our choices look horrible"

and why Jolliffe (1982) and Hadi and Ling (1998) were able to find examples where the last few principal components are more correlated with the response. On the

other hand, the chances are still in favor of the fact that the nature is fair. So although risky, it has been proved that one can confidently use principal component analysis to find the principal components and being confident the principal components mostly correlated with the response will be given the first few ones (that is, the ones corresponding to the largest eigenvalues).

Since this is a procedure that people have been using for a long time, although this problem was well known and the reasons behind it were well understood, this is not something that will change how people think towards principal component analysis and its use in regression. The proof is probabilistic, so it doesn't say anything about the behavior of a single datasets. It just shows what's going to happen if you have a collection of dataset. For a single dataset it is obvious that you can still get a lower order principal component who is more correlated with the response than the higher order ones. People that were critical against the use of principal component analysis, will probably continue to be thinking critically against it as there is an unmeasurable risk it will not give you the correct results. On the other hand, those that are in favor of using principal component analysis in regression, they now have a rigorous proof that the probability they will get the desired results is higher than the probability to get the wrong result.

1.8 Extensions to the previous results

An unpublished manuscript, Ni (2010), extends the results presented above, in Artemiou (2008) and in Artemiou and Li (2009).

First of all, the author proved the following theorem:

Theorem 1.8.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *orientationally uniform random matrix,*

2. $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$*,*

3. $Y = \boldsymbol{\beta}^T\boldsymbol{X} + \delta$*, where* $\boldsymbol{\beta}$ *is a* p*-dimensional random vector and* δ *is a random variable such that* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$*,* $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$*,* $E(\delta) = 0$ *and* $\mathrm{var}(\delta) < \infty$*.*

4. $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set* $G \in \mathbb{R}^p$*.*

Let $\lambda_1, \ldots, \lambda_p$ *to be the ordered eigenvalues of the covariance matrix* $\boldsymbol{\Sigma}$ *and* $w_1, \ldots, w_p$ *be the 1st, . . . , pth principal components of* $\boldsymbol{X}$*, and let* $\rho_i = \rho_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathrm{corr}^2(Y, w_i|\boldsymbol{\beta}, \boldsymbol{\Sigma})$*. Then, whenever* $i < j$*,*

$$P(\rho_i \geq \rho_j) = \frac{2}{\pi}\mathrm{E}\left(\arctan\left[(\lambda_i/\lambda_j)^{\frac{1}{2}}\right]\right) \tag{1.1}$$

Mainly he showed that the left hand side of the inequality in Theorem 1.6.1 (Artemiou and Li (2009)) it is not just bounded to a number but there is an expression that it is exactly equal to; and this expression it is the right hand side of (1.1)in Theorem 1.8.1.

Ni (2010) showed also that the above result is true if we remove one level of randomness, that is, we do not need both, the covariance matrix $\boldsymbol{\Sigma}$ and the regression coefficients $\boldsymbol{\beta}$ to be random. The author showed that the result is true if the covariance matrix $\boldsymbol{\Sigma}$ is orientationally uniform and the regression coefficients are constant. The author, also proved that the result holds when the covariance matrix is constant and the regression coefficients $\boldsymbol{\beta}$ are spherically distributed. In the latter case the expectation on the right hand side of (1.1 is redundant as the quantity inside the expectation is constant.

In Chapter 2, we show the results for fixed covariance matrix and multivariate

response. In Chapter 3, we show two extensions of this result in the nonlinear case, for both the univariate and multivariate case. In Chapter 4, we talk about an information criterion that interestingly in a very special case match the results of Artemiou and Li (2009) and Ni (2010). The results of Chapters 3 and 4 describe the way we tried to attack the extension to the nonlinear case using linear principal components. The most general results and the most interesting ones are presented in Chapter 5. In Chapter 5, we use a completely different approach to extend the result by Artemiou and Li (2009) and Ni (2010) in the sufficient dimension reduction concept using linear principal components, and we use kernel principal components to extend the results in two other ways, the more general one doesn't assume any specific regression function for $Y|\boldsymbol{X}$. The only assumption that we have is that we randomly choose a conditional distribution for $Y|\boldsymbol{X}$ that makes Y to be dependent on $\boldsymbol{X}$ by excluding measures that imply Y and $\boldsymbol{X}$ are independent. In Chapter 6, there is a discussion of the results and possible future extensions.

Chapter 2

Extensions for fixed covariance matrix and multivariate response

In this Chapter we present how the results in Artemiou and Li (2009) can extend for fixed covariance matrix and for multivariate response.

The results for fixed covariance matrix were developed at the same time by Ni(2010) where he showed that his result holds for a fixed covariance matrix if the regression coefficients are spherically distributed. We show how we developed our Thorem, which results in the same condition; that is, the regression coefficients $\boldsymbol{\beta}$ need to be spherically distributed for the same result to hold. We first prove Lemma 2.1.1 about the exchangeability of two random variables (which is not shown in Ni (2010)) under the assumption of spherically distributed regression coefficients and then we show Lemma 2.1.2 that proves the median of a ratio of two specific

variables (which we need in the main result) is unique and equal to 1. A similar result as Lemma 2.1.2 was shown in Artemiou and Li (2010) when the covariance matrix has an orientationally uniform distribution. The result for multivariate response variable Y, is interesting but more investigation on the conditions for the uniqueness of median is needed.

2.1 Fixed covariance matrix

Lemma 2.1.1 *Let $\boldsymbol{\beta}$ be a p-dimensional random vector, with spherical distribution. Also let $\boldsymbol{a}, \boldsymbol{b}$ be p-dimensional fixed vectors with equal length, that is $\|\boldsymbol{a}\| = \|\boldsymbol{b}\| = k$. Then the random variables $\boldsymbol{\beta}^T\boldsymbol{a}$ and $\boldsymbol{\beta}^T\boldsymbol{b}$ are exchangeable.*

Proof. By definition, if $\boldsymbol{\beta}^T\boldsymbol{a}$ and $\boldsymbol{\beta}^T\boldsymbol{b}$ are exchangeable, it means

$$\begin{pmatrix} \boldsymbol{\beta}^T\boldsymbol{a} \\ \boldsymbol{\beta}^T\boldsymbol{b} \end{pmatrix} \overset{\mathcal{D}}{=} \begin{pmatrix} \boldsymbol{\beta}^T\boldsymbol{b} \\ \boldsymbol{\beta}^T\boldsymbol{a} \end{pmatrix}$$

Let $\mathbf{C}$ be a $p \times p$ matrix such that

1. $\mathbf{C}^T = \mathbf{C}^{-1}$ and
2. $\mathbf{C}\,(\boldsymbol{b}, \boldsymbol{a}) = (\boldsymbol{a}, \boldsymbol{b})$.

We then have $\boldsymbol{\beta} \overset{\mathcal{D}}{=} C\boldsymbol{\beta}$, since $\boldsymbol{\beta}$ is spherically distributed. Then

$$\begin{pmatrix} \boldsymbol{\beta}^T\boldsymbol{a} \\ \boldsymbol{\beta}^T\boldsymbol{b} \end{pmatrix} \overset{\mathcal{D}}{=} \boldsymbol{\beta}^T\,(\boldsymbol{a}, \boldsymbol{b}) \overset{\mathcal{D}}{=} \boldsymbol{\beta}^T\mathbf{C}^T\,(\boldsymbol{a}, \boldsymbol{b}) \overset{\mathcal{D}}{=} \boldsymbol{\beta}^T\,(\boldsymbol{b}, \boldsymbol{a}) \overset{\mathcal{D}}{=} \begin{pmatrix} \boldsymbol{\beta}^T\boldsymbol{b} \\ \boldsymbol{\beta}^T\boldsymbol{a} \end{pmatrix}$$

□

The above lemma, will help us prove the following lemma, which ensures uniqueness of the median for $\boldsymbol{\beta}^T\boldsymbol{b}/\boldsymbol{\beta}^T\boldsymbol{a}$. This result is important in proving the main result of this section, Theorem 2.1.1, that follows the following Lemma. The proof of the following Lemma is similar to the one for Lemma 1.6.1 as it appeared in Aremiou and Li (2009). The only difference is that we are using Lemma 2.1.1 to show exchangeability which in Lemma 1.6.1 was implied by the properties of orientationally uniform matrix.

Lemma 2.1.2 *Let $\boldsymbol{\beta}$ be a p-dimensional spherically distributed random vector, and let $\mathbf{u}, \mathbf{v}$ be p-dimensional fixed vectors. Suppose*

1. *$P(\boldsymbol{\beta} \in G) > 0$ for any nonempty open set G.*
2. *$\mathbf{u}, \mathbf{v}$ linearly independent with equal length*

Then the random variable $\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2 / \left(\boldsymbol{\beta}^T\mathbf{v}\right)^2$ has a unique median equal to 1.

PROOF. First we need to show that 1 is a median, that is 1 satisfies the following equation

$$P\left(\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2 / \left(\boldsymbol{\beta}^T\mathbf{v}\right)^2 < 1\right) \leq 1/2 \leq P\left(\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2 / \left(\boldsymbol{\beta}^T\mathbf{v}\right)^2 \leq 1\right). \tag{2.1}$$

Because of assumption 1 and Lemma 2.1.1 we have $\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2$ and $\left(\boldsymbol{\beta}^T\mathbf{v}\right)^2$ to be

exchangeable. That means

$$\begin{aligned} P\left(\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2/\left(\boldsymbol{\beta}^T\mathbf{v}\right)^2 \le 1\right) =& P\left(\left(\boldsymbol{\beta}^T\mathbf{v}\right)^2/\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2 \le 1\right) \\ =& 1 - P\left(\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2/\left(\boldsymbol{\beta}^T\mathbf{v}\right)^2 < 1\right) \end{aligned}$$

Hence

$$\begin{aligned} P\left(\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2/\left(\boldsymbol{\beta}^T\mathbf{v}\right)^2 < 1\right) &\le 1 - P\left(\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2/\left(\boldsymbol{\beta}^T\mathbf{v}\right)^2 < 1\right), \\ P\left(\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2/\left(\boldsymbol{\beta}^T\mathbf{v}\right)^2 \le 1\right) &\ge 1 - P\left(\left(\boldsymbol{\beta}^T\mathbf{u}\right)^2/\left(\boldsymbol{\beta}^T\mathbf{v}\right)^2 \le 1\right), \end{aligned}$$

which imply (2.1).

Now we need to show that 1 is the only number that satisfies (2.1). In other words, for any $0 < c_1 < 1$ and $c_2 > 1$ we have

$$P((\boldsymbol{\beta}^T\mathbf{u})^2/(\boldsymbol{\beta}^T\mathbf{v})^2 \le c_1) < 1/2 \;\text{ and }\; P((\boldsymbol{\beta}^T\mathbf{u})^2/(\boldsymbol{\beta}^T\mathbf{v})^2 < c_2) > 1/2.$$

We will show only the first one. Similarly we can prove the second one. Let $c_3 \in (c_1, 1)$. Since $(\mathbf{u}, \mathbf{v})$ has full column rank, the following system of equations

$$\begin{cases} \boldsymbol{\beta}^T\mathbf{u} = \sqrt{c}_3 \\ \boldsymbol{\beta}^T\mathbf{v} = 1 \end{cases}$$

has a solution, say $\boldsymbol{\beta}_0$. Note that $(\boldsymbol{\beta}_0^T\mathbf{u})^2/(\boldsymbol{\beta}_0^T\mathbf{v})^2 = c_3 \in (c_1, 1)$. Because $\boldsymbol{\beta} \mapsto (\boldsymbol{\beta}^T\mathbf{u})^2/(\boldsymbol{\beta}^T\mathbf{v})^2$ is continuous there is a neighborhood of $\boldsymbol{\beta}_0$, say G, such that

$$\boldsymbol{\beta} \in G \Rightarrow (\boldsymbol{\beta}^T\mathbf{u})^2/(\boldsymbol{\beta}^T\mathbf{v})^2 \in (c_1, 1).$$

By assumption 2, $P(\boldsymbol{\beta} \in G) > 0$. Therefore

$$P((\boldsymbol{\beta}^T \mathbf{u})^2/(\boldsymbol{\beta}^T \mathbf{v})^2 \in (c_1, 1)) > 0,$$

implies (2.1). □

Now that we have shown the previous two helpful results, we are ready to show the extension of our main theorem for fixed covariance matrix. The proof of Theorem 2.1.1 is similar to the proof of Theorem 1.6.1 as it appears in Artemiou and Li (2009). The only difference is that there is no conditioning on the covariance matrix $\boldsymbol{\Sigma}$ as it is considered non random and not random as Artemiou and Li (2009) had it.

Theorem 2.1.1 *Suppose*

1. *a non random covariance matrix* $\boldsymbol{\Sigma}$
2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$,
3. $Y = \boldsymbol{\beta}^T \boldsymbol{X} + \delta$, *where* $\boldsymbol{\beta}$ *is a p-dimensional spherically distributed random vector and* δ *is a random variable such that* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X})$, $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta})$, $E(\delta) = 0$ *and* $\text{var}(\delta) < \infty$.
4. $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set* $G \in \mathbb{R}^p$.

Let $w_1, \ldots, w_p$ *be the* 1^{st}, $\ldots$, p^{th} *principal components of* $\boldsymbol{X}$, *and let* $\rho_i = \rho_i(\boldsymbol{\beta}) = \text{corr}^2(Y, w_i|\boldsymbol{\beta})$. *Then, whenever* $i < j$,

$$P(\rho_i \geq \rho_j) > 1/2.$$

PROOF. Let τ^2 denote $\text{var}(\delta)$. Let $(\sigma^2_{(1)}, \boldsymbol{v}_{(1)}), \ldots, (\sigma^2_{(p)}, \boldsymbol{v}_{(p)})$ be the ordered $(\sigma^2_1, \boldsymbol{v}_1), \ldots, (\sigma^2_p, \boldsymbol{v}_p)$ such that $\sigma^2_{(1)} \geq \cdots \geq \sigma^2_{(p)}$. First, we derive an explicit expression for ρ_i. Note that

$$\begin{aligned}\text{cov}(Y, \boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}) =& \text{cov}(\boldsymbol{\beta}^T\boldsymbol{X} + \delta, \boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}) \\ =& \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{v}_{(i)} + \text{cov}(\delta, \boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}). \end{aligned} \tag{2.2}$$

Because $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta})$, we have $\delta \perp\!\!\!\perp (\boldsymbol{v}^T_{(i)}\boldsymbol{X}, \boldsymbol{\beta})$. This implies $\delta \perp\!\!\!\perp \boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}$, and hence that the second term in (2.2) is zero. Because $(\sigma^2_{(i)}, \boldsymbol{v}_{(i)})$ is an eigen pair of $\boldsymbol{\Sigma}$, we have $\boldsymbol{\Sigma}\boldsymbol{v}_{(i)} = \sigma^2_{(i)}\boldsymbol{v}_{(i)}$. Hence

$$\text{cov}^2(Y, \boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}) = \sigma^4_{(i)}(\boldsymbol{\beta}^T\boldsymbol{v}_{(i)})^2. \tag{2.3}$$

In the meantime

$$\text{var}(Y|\boldsymbol{\beta}) = \text{var}(\boldsymbol{\beta}^T\boldsymbol{X}|\boldsymbol{\beta}) + 2\text{cov}(\boldsymbol{\beta}^T\boldsymbol{X}, \delta|\boldsymbol{\beta}) + \text{var}(\delta|\boldsymbol{\beta}).$$

Because $\delta \perp\!\!\!\perp \boldsymbol{\beta}$, the last term on the right is simply τ^2. Because $\delta \perp\!\!\!\perp (\boldsymbol{\beta}, \boldsymbol{X})$, we have $\delta \perp\!\!\!\perp \boldsymbol{\beta}^T\boldsymbol{X}|\boldsymbol{\beta}$. So the second term on the right is 0. Hence

$$\text{var}(Y|\boldsymbol{\beta}) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2. \tag{2.4}$$

Moreover, using the facts $\boldsymbol{\Sigma}\boldsymbol{v}_{(i)} = \sigma^2_{(i)}\boldsymbol{v}_{(i)}$ and $\boldsymbol{v}^{\mathsf{T}}_{(i)}\boldsymbol{v}_{(i)} = 1$

$$\text{var}(\boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}) = \boldsymbol{v}^T_{(i)}\boldsymbol{\Sigma}\boldsymbol{v}_{(i)} = \sigma^2_{(i)}. \tag{2.5}$$

Now combine (2.3), (2.4), and (2.5) to obtain

$$\rho_i = \text{corr}^2(Y, \boldsymbol{v}_{(i)}^T \boldsymbol{X}) = \frac{\sigma_{(i)}^2 (\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2}. \tag{2.6}$$

Let $i < j$. Then, using (2.6) we deduce

$$P(\rho_i \geq \rho_j) = P\left(\frac{\sigma_{(i)}^2 (\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2} \geq \frac{\sigma_{(j)}^2 (\boldsymbol{\beta}^T \boldsymbol{v}_{(j)})^2}{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2} \right) = P\left(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_{(j)})^2} \geq \frac{\sigma_{(j)}^2}{\sigma_{(i)}^2} \right).$$

To prove the theorem we need to show that

$$P\left(\frac{(\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{(\boldsymbol{\beta}^T \boldsymbol{v}_{(j)})^2} \geq \frac{\sigma_{(j)}^2}{\sigma_{(i)}^2} \right) > \frac{1}{2} \tag{2.7}$$

Then by Lemma 2.1.2 inequality (2.7) above holds. □

The following proposition shows that if the above inequality is true for non-random covariance matrix, then it is also true if we have a random covariance matrix.

Proposition 2.1.1 *Let assume that* $\boldsymbol{\Sigma}$ *a* $p \times p$ *matrix and* $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$*. Let also* $w_1, \ldots, w_p$ *be the 1st, . . . , pth principal components of* $\boldsymbol{\Sigma}$*, and let* $\rho_i = \rho_i(\boldsymbol{\beta}) = \text{corr}^2(Y, w_i|\boldsymbol{\beta})$*. Then, whenever* $i < j$*, if* $P(\rho_i \geq \rho_j|\boldsymbol{\Sigma}) > 1/2$ *holds for every* $\boldsymbol{\Sigma}$ *it is implied that* $P(\rho_i \geq \rho_j) > 1/2$ *holds.*

Proof. Since $P(\rho_i \geq \rho_j|\boldsymbol{\Sigma}) > 1/2$ holds for every $\boldsymbol{\Sigma}$ then

$$P(\rho_i \geq \rho_j|\boldsymbol{\Sigma}) = E\left(P(\rho_i \geq \rho_j|\boldsymbol{\Sigma})\right) > 1/2 \Rightarrow P(\rho_i \geq \rho_j) > 1/2$$

□

2.2 Multivariate response

Since we have a multivariate response in this section the correlation formula that we used in the previous section will not work. So we use the squared multiple correlation coefficient, which is defined below. For more details the interested reader is referred to Hall and Mathiason (1990).

Definition 2.2.1 *The square multiple correlation between U, a p-dimensional random vector, and V, a q-dimensional random vector, is defined as follows*

$$\mathrm{mcor}^2\,(U,V) = \mathrm{tr}\left(\Sigma_U^{-\frac{1}{2}}\Sigma_{UV}\Sigma_V^{-1}\Sigma_{VU}\Sigma_U^{-\frac{1}{2}}\right)$$

where Σ_U, Σ_V are the covariance matrices or U and V respectively, and Σ_{UV}, is the covariance matrix between U and V.

Since Proposition 2.1.1 does not depend on the dimension of the response variable, it is still true for the multivariate response variable as the following corollary shows.

Corollary 2.2.1 *Let assume that $\boldsymbol{\Sigma}$ a $p \times p$ matrix and $\boldsymbol{X}$ is a p-dimensional random vector with $E(\boldsymbol{X}) = 0$ and $\mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Let also $w_1, \ldots, w_p$ be the 1st, $\ldots$, pth principal components of $\boldsymbol{\Sigma}$, and let $\rho_i = \rho_i(\boldsymbol{\beta}) = \mathrm{mcor}^2(\boldsymbol{Y}, w_i|\boldsymbol{\beta})$. Then, whenever $i < j$, if $P(\rho_i \geq \rho_j|\boldsymbol{\Sigma}) > 1/2$ holds for every $\boldsymbol{\Sigma}$ it is implied that $P(\rho_i \geq \rho_j) > 1/2$ holds.*

Using this fact, for the majority of this work (Chapters 3 and 4) we prove the results for the case with fixed covariance matrix and the results for random covariance matrix are therefore implied and presented as corollaries without proofs.

Theorem 2.2.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *fixed* $p \times p$ *and* $\boldsymbol{\Gamma}$ *fixed* $q \times q$ *matrices,*
2. $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$*,*
3. $\boldsymbol{\beta}$ *is a* $p \times q$ *spherically distributed random matrix independent of* $\boldsymbol{X}$*,*
4. $\boldsymbol{\epsilon}$ *is a* q *dimensional random vector independent of* $(\boldsymbol{X}, \boldsymbol{\beta})$*, with* $E(\boldsymbol{\epsilon}) = 0, \mathrm{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Gamma}$*,*
5. $\boldsymbol{Y} = \boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\epsilon}$*,*
6. $\dfrac{\boldsymbol{v}_{(i)}^T \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{v}_{(i)}}{\boldsymbol{v}_{(j)}^T \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{v}_{(j)}}$ *has unique variance equal to 1, for* $\boldsymbol{A} = \boldsymbol{\beta}\left(\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \boldsymbol{\Gamma}\right)^{\frac{1}{2}}$ *and* $\boldsymbol{v}_{(i)}, i = 1, \ldots, p$ *is the* i^{th} *ordered eigenvector of matrix* $\boldsymbol{\Sigma}$ *in the sense that it corresponds to the* i^{th} *largest eigenvalue*

Let $w_1, \ldots, w_p$ *be the* $1^{\text{st}}, \ldots, p^{\text{th}}$ *ordered principal components of* $\boldsymbol{\Sigma}$*, and let* $\rho_i = \rho_i(\boldsymbol{\beta}) = \mathrm{mcor}^2(Y, w_i|\boldsymbol{\beta})$*. Then, whenever* $i < j$*,*

$$P(\rho_i \geq \rho_j) > 1/2.$$

Proof. By Definition 2.2.1,

$$\rho_i = \boldsymbol{\Sigma}_{w_i \boldsymbol{Y}} \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{Y} w_i} \sigma_{w_i}^{-2} \tag{2.8}$$

Now

$$\boldsymbol{\Sigma}_{w_i \boldsymbol{Y}} = \operatorname{cov}\left(w_i, \boldsymbol{Y} \,|\, \boldsymbol{\beta}\right) = \operatorname{cov}\left(\boldsymbol{v}_{(i)}^T \boldsymbol{X}, \boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\epsilon} \,\middle|\, \boldsymbol{\beta}\right)$$

where $\boldsymbol{v}_{(i)}$ is the i^{th} ordered eigenvector of $\boldsymbol{\Sigma}$, in the sense that it corresponds to the i^{th} largest eigenvalue, and $\left(\sigma_{(i)}^2, \boldsymbol{v}_{(i)}\right), i = 1, \ldots, p$ are ordered eigen pairs of $\boldsymbol{\Sigma}$, in the sense that $\sigma_{(1)}^2 \geq \ldots \geq \sigma_{(p)}^2$.

The above is equal to

$$\boldsymbol{\Sigma}_{w_i \boldsymbol{Y}} = \boldsymbol{v}_{(i)}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \operatorname{cov}\left(\boldsymbol{v}_{(i)}^T \boldsymbol{X}, \boldsymbol{\epsilon} \,\middle|\, \boldsymbol{\beta}\right) = \boldsymbol{v}_{(i)}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \tag{2.9}$$

because $\boldsymbol{\epsilon} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta})$. Also,

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \operatorname{var}\left(\boldsymbol{\beta}^T \boldsymbol{X} + \boldsymbol{\epsilon} \,\middle|\, \boldsymbol{\beta}\right) = \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \boldsymbol{\Gamma} \tag{2.10}$$

Finally,

$$\sigma_{w_i}^2 = \operatorname{var}\left(\boldsymbol{v}_{(i)}^T \boldsymbol{X} \,\middle|\, \boldsymbol{\beta}\right) = \sigma_{(i)}^2 \tag{2.11}$$

By combining (2.8), (2.9), (2.10), (2.11),

$$\rho_i = \boldsymbol{v}_{(i)}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \left(\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \boldsymbol{\Gamma}\right)^{-1} \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{v}_{(i)} \sigma_{(i)}^{-2}$$

The objective is to prove that for $i < j$, $P(\rho_i \geq \rho_j) > 1/2$. This is equivalent to

$$P\left(\boldsymbol{v}_{(i)}^T \boldsymbol{\Sigma}\boldsymbol{\beta}\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}+\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{v}_{(i)}\sigma_{(i)}^{-2} \geq \boldsymbol{v}_{(j)}^T \boldsymbol{\Sigma}\boldsymbol{\beta}\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}+\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{v}_{(j)}\sigma_{(j)}^{-2}\right) > 1/2. \tag{2.12}$$

Since $\left(\sigma_{(i)}^2, \boldsymbol{v}_{(i)}\right)$ is an eigen pair we have that $\boldsymbol{v}_{(i)}^T\boldsymbol{\Sigma} = \boldsymbol{v}_{(i)}^T\sigma_{(i)}^2$, so that the above is equivalent to:

$$P\left(\sigma_{(i)}^2\boldsymbol{v}_{(i)}^T\boldsymbol{\beta}\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}+\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\beta}^T\boldsymbol{v}_{(i)} \geq \sigma_{(j)}^2\boldsymbol{v}_{(j)}^T\boldsymbol{\beta}\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}+\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\beta}^T\boldsymbol{v}_{(j)}\right) > 1/2. \tag{2.13}$$

Let $\boldsymbol{A} = \boldsymbol{\beta}\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}+\boldsymbol{\Gamma}\right)^{-\frac{1}{2}}$. Then the above is equivalent to

$$\begin{aligned} P\left(\sigma_{(i)}^2\boldsymbol{v}_{(i)}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{v}_{(i)} \geq \sigma_{(j)}^2\boldsymbol{v}_{(j)}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{v}_{(j)}\right) > 1/2 \Rightarrow \\ P\left(\frac{\boldsymbol{v}_{(i)}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{v}_{(i)}}{\boldsymbol{v}_{(j)}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{v}_{(j)}} \geq \frac{\sigma_{(j)}^2}{\sigma_{(i)}^2}\right) > 1/2 \end{aligned} \tag{2.14}$$

Inequality (2.14) holds true from the fact that $\sigma_{(j)}^2 < \sigma_{(i)}^2$ by assumption 6 for unique median. □

Using Proposition 2.2.1 we can show that the above result is true in the case of random covariance matrix as the following corollary states.

Corollary 2.2.2 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *orientationally unifromly distributed random matrix and* $\boldsymbol{\Gamma}$ *is a fixed* $q \times q$ *matrix,*

2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\text{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$

3. $\boldsymbol{\beta}$ *is a* $p \times q$ *spherically distributed random matrix independent of* $(\boldsymbol{X}, \boldsymbol{\Sigma})$

4. $\boldsymbol{\epsilon}$ *is a q dimensional random vector independent of* $(\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$*, with* $E(\boldsymbol{\epsilon}) = 0, \text{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Gamma}$

5. $\boldsymbol{Y} = \boldsymbol{\beta}^T\boldsymbol{X} + \boldsymbol{\epsilon}$,

6. $\dfrac{\boldsymbol{v}_{(i)}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{v}_{(i)}}{\boldsymbol{v}_{(j)}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{v}_{(j)}}$ *has unique median equal to 1, for* $\boldsymbol{A} = \boldsymbol{\beta}\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + \boldsymbol{\Gamma}\right)^{\frac{1}{2}}$ *and* $\boldsymbol{v}_{(i)}, i = 1, \ldots, p$ *is the* i^{th} *ordered eigenvector of matrix* $\boldsymbol{\Sigma}$ *in the sense that it corresponds to the* i^{th} *largest eigenvalue.*

Let $w_1, \ldots, w_p$ *be the* $1^{\text{st}}, \ldots, p^{\text{th}}$ *ordered principal components of* $\boldsymbol{\Sigma}$*, and let* $\rho_i = \rho_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \text{mcor}^2(Y, w_i|\boldsymbol{\beta}, \boldsymbol{\Sigma})$*. Then, whenever* $i < j$*,*

$$P(\rho_i \geq \rho_j) > 1/2.$$

As in the univariate case in the multivariate case the spherically distributed assumption for the regression coefficients can be removed. This result needs further improvement by exploring the conditions under which Assumption 6 is valid. This is something we leave for future work. Also there is an open question whether an equivalent result as the one by Ni (2010) holds in this case.

In the univariate case the assumption of spherically distributed regression coefficient was necessary to achieve exchangeability of the ratio of two random variables in the case of non random covariance matrix. In the multivariate case, this assumption is similarly needed to achieve exchangeability of a ratio of two different random variables in the case of fixed covariance matrix. For the case of random covariance

matrix this assumption can be dropped. This is because in the case of fixed covariance matrix the spherical distribution is needed to show exchangeability which is needed to show that the median of the ratio of the two random variables is equal to 1. In the random covariance matrix, exchangeability is shown without the need of spherically distributed regression coefficients. We acknowledge the need of further research and development in this case to identify the conditions needed for Assumption 6 in Theorem 2.2.1 and Corollary 2.2.2 to be true and especially the conditions under which the median is unique.

Finally, it is interesting for one to investigate if an equivalent result as the one in Ni (2010) is true for the multivariate response case. This is left for future work as well.

Chapter 3

Beyond the linear regression model

In this chapter we extend the results in Artemiou and Li (2009) and Ni (2010) under different model assumption, that is, we remove the assumption of a linear model. We show first the results in the univariate case, for both fixed and random covariance matrices and then we show that one can extend them in the multivariate case. As in the previous Chapter (see Corollary 2.2.2 and the discussion that follows it) for the multivariate response case one needs to explore the conditions so that specific ratios have unique median equal to 1. Also, the conditions under which the results in Ni (2010) can be extended to the multivariate response case need to be investigated further, which we leave for future work.

For the univariate cases we will give only the main results, as the supporting lemmas were proved in the previous Chapter. Those are Lemmas 2.1.1 and 2.1.2, for the exchangeability of two random variables and the uniqueness of median which

is equal to 1 for the ratio of those two random variables, respectively.

3.1 Conditional independence model

3.1.1 Univariate response

Theorem 3.1.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *non random matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;
2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ *is a spherically distributed random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$;
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta})$ *is linear in* $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$.

If $Y \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta})$, *then, for any square-integrable function* $f(Y)$ *such that* $E[\boldsymbol{X}f(Y)|\boldsymbol{\beta}] \neq \boldsymbol{0}$, $\mathrm{var}[f(Y)|\boldsymbol{\beta}] > 0$ *and* $i < j$ *we have*

$$P\left(\mathrm{corr}^2(f(Y), \boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}) > \mathrm{corr}^2(f(Y), \boldsymbol{v}_j^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta})\right) > 1/2.$$

Proof. We denote with $\mathbf{w}_1, \ldots, \mathbf{w}_p$ the ordered principal components of the predictor vector $\boldsymbol{X}$. Let $i < j$. Since $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$, we have $E(\boldsymbol{X}|\boldsymbol{\beta}) = E(\boldsymbol{X}) = 0$. Hence

$$\mathrm{cov}(f(Y), \mathbf{w}_i|\boldsymbol{\beta}) = E[f(Y)\boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}] = E[f(Y)\boldsymbol{v}_i^\mathsf{T}E(\boldsymbol{X}|Y, \boldsymbol{\beta})|\boldsymbol{\beta}]. \tag{3.1}$$

The inner conditional expectation on the right hand side can be rewritten as

$$E(\boldsymbol{X}|Y,\boldsymbol{\beta}) = E[E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},Y,\boldsymbol{\beta})|Y,\boldsymbol{\beta}] = E[E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})|Y,\boldsymbol{\beta}], \tag{3.2}$$

where the second equality follows from the conditional independence $Y \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})$.

Now let us compute $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})$ in (3.2). Because, by assumption 3, $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$, we have

$$E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta}) = P^{\mathsf{T}}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X} = \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}.$$

Substitute this into the right hand side of (3.2) to obtain

$$E(\boldsymbol{X}|Y,\boldsymbol{\beta}) = E[E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})|Y,\boldsymbol{\beta}] = \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}E(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}|Y,\boldsymbol{\beta}).$$

Hence, by (3.1),

$$\begin{aligned}\operatorname{cov}(f(Y),\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}) =& \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}]E[f(Y)E(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}|Y,\boldsymbol{\beta})|\boldsymbol{\beta}] \\ =& \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}E(\boldsymbol{X}f(Y)|\boldsymbol{\beta}).\end{aligned}$$

In the meantime we note that $\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\Sigma} = \boldsymbol{v}_i^{\mathsf{T}}\sigma_i^2$ and

$$\operatorname{var}(\boldsymbol{v}_i\boldsymbol{X}|\beta) = \operatorname{var}(\boldsymbol{v}_i\boldsymbol{X}) = \sigma_i^2.$$

Using these and the assumption that $\operatorname{var}(f(Y)|\boldsymbol{\beta}) > 0$ we obtain

$$\operatorname{corr}^2(f(Y),\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}) = \frac{\sigma_i^4(\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\beta})^2(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-2}[\boldsymbol{\beta}^{\mathsf{T}}E(\boldsymbol{X}f(Y)|\boldsymbol{\beta})]^2}{\operatorname{var}(f(Y)|\boldsymbol{\beta})\sigma_i^2}$$

Using the assumption that $E(\boldsymbol{X}f(Y)) \neq \boldsymbol{0}$, we obtain

$$\frac{\text{corr}^2(f(Y), \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta})}{\text{corr}^2(f(Y), \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta})} = \frac{\sigma_i^2(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2}{\sigma_j^2(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2}$$

Hence

$$P\left(\text{corr}^2(f(Y), \boldsymbol{v}_i\boldsymbol{X}|\boldsymbol{\beta}) > \text{corr}^2(f(Y), \boldsymbol{v}_j\boldsymbol{X}|\boldsymbol{\beta})\right) = P((\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2/(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2 > \sigma_j^2/\sigma_i^2) > 1/2,$$

where the last inequality follows from $\sigma_j^2/\sigma_i^2 < 1$ and Lemma 2.1.2. □

This result is pretty straightforward to extend it to a similar one as the one in Ni (2010). For this we have the following corollary.

Corollary 3.1.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *non random matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$*;*
2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$*;*
3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ *is a spherically distributed random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$*;*
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal\boldsymbol{X}, \boldsymbol{\beta})$ *is linear in* $\boldsymbol{\beta}^\intercal\boldsymbol{X}$*.*

If $Y \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^\intercal\boldsymbol{X}, \boldsymbol{\beta})$*, then, for any square-integrable function* $f(Y)$ *such that* $E[\boldsymbol{X}f(Y)|\boldsymbol{\beta}] \neq \boldsymbol{0}$*,* $\text{var}[f(Y)|\boldsymbol{\beta}] > 0$ *and* $i < j$ *we have*

$$P\left(\text{corr}^2(f(Y), \boldsymbol{v}_i^\intercal\boldsymbol{X}|\boldsymbol{\beta}) > \text{corr}^2(f(Y), \boldsymbol{v}_j^\intercal\boldsymbol{X}|\boldsymbol{\beta})\right) = \frac{2}{\pi}\arctan\left[(\lambda_i/\lambda_j)^{\frac{1}{2}}\right].$$

Using Proposition 2.1.1 one can show that the above results are also true when we have random covariance matrix $\boldsymbol{\Sigma}$, as the following corollaries state.

Corollary 3.1.2 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *matrix from an orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;
2. $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ *is a random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$;
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is linear in* $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$.

If $Y \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, *then, for any square-integrable function* $f(Y)$ *such that* $E[\boldsymbol{X}f(Y)|\boldsymbol{\beta}, \boldsymbol{\Sigma}] \neq \boldsymbol{0}$, $\mathrm{var}[f(Y)|\boldsymbol{\beta}, \boldsymbol{\Sigma}] > 0$ *and* $i < j$ *we have*

$$P\left(\mathrm{corr}^2(f(Y), \boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(f(Y), \boldsymbol{v}_j^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) > 1/2.$$

Corollary 3.1.3 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *matrix from an orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;
2. $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ *is a spherically distributed random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$;
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is linear in* $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$.

If $Y \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$*, then, for any square-integrable function* $f(Y)$ *such that* $E[\boldsymbol{X}f(Y)|\boldsymbol{\beta}, \boldsymbol{\Sigma}] \neq \boldsymbol{0}$*,* $\mathrm{var}[f(Y)|\boldsymbol{\beta}, \boldsymbol{\Sigma}] > 0$ *and* $i < j$ *we have*

$$P\left(\mathrm{corr}^2(f(Y), \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(f(Y), \boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) \frac{2}{\pi}\mathrm{E}\left(\arctan\left[(\lambda_i/\lambda_j)^{\frac{1}{2}}\right]\right).$$

3.1.2 Multivariate response

Theorem 3.1.2 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *fixed matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1*,* $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$*;*

2. $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$*;*

3. $\boldsymbol{\beta}$ *is a* $p \times q$ *spherically distributed random matrix and* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$*;*

4. $\boldsymbol{v}_i^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_i / \boldsymbol{v}_j^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_j$ *having unique median equal to 1; where*

$$\mathbf{A} = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}E(f(\boldsymbol{Y})\boldsymbol{X}|\boldsymbol{\beta})\left(\mathrm{var}(f(\boldsymbol{Y})|\boldsymbol{\beta})\right)^{-1/2}.$$

5. $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta})$ *is linear in* $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$*.*

If $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta})$*, then, for any square-integrable function* $f(\boldsymbol{Y})$ *such that* $E[\boldsymbol{X}f(\boldsymbol{Y})|\boldsymbol{\beta}] \neq \boldsymbol{0}$*,* $\mathrm{var}[f(\boldsymbol{Y})|\boldsymbol{\beta}]$ *positive definite matrix and* $i < j$ *we have*

$$P\left(\rho_i > \rho_j\right) > 1/2, \tag{3.3}$$

where $\rho_i = \mathrm{mcor}^2(f(\boldsymbol{Y}), \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta})$*.*

PROOF. By Definition 2.2.1,

$$\rho_i = \boldsymbol{\Sigma}_{w_i f(\boldsymbol{Y})} \boldsymbol{\Sigma}^{-1}_{f(\boldsymbol{Y})} \boldsymbol{\Sigma}_{f(\boldsymbol{Y}) w_i} \sigma^{-2}_{w_i} \tag{3.4}$$

where $w_i = \boldsymbol{v}_i^{\mathsf{T}} \boldsymbol{X}$ the i^{th} principal component.

We have that

$$\sigma^2_{w_i} = \sigma^2_i, \tag{3.5}$$

$$\boldsymbol{\Sigma}_{f(\boldsymbol{Y})} = \text{var}(f(\boldsymbol{Y})|\boldsymbol{\beta}) \tag{3.6}$$

and

$$\boldsymbol{\Sigma}_{w_i f(\boldsymbol{Y})} = \text{cov}\left(w_i, f(\boldsymbol{Y})|\,\boldsymbol{\beta}\right) = \text{cov}\left(\boldsymbol{v}^{\mathsf{T}}_{(i)}\boldsymbol{X}, f(\boldsymbol{Y})\Big|\,\boldsymbol{\beta}\right) = E\left(\boldsymbol{v}^{\mathsf{T}}_{(i)}\boldsymbol{X}(f(\boldsymbol{Y}))^{\mathsf{T}}\Big|\,\boldsymbol{\beta}\right) \tag{3.7}$$

where the last equality holds since $E\left(\boldsymbol{X}|\boldsymbol{\beta}\right) = E(\boldsymbol{X}) = 0$ since $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{\beta}$. Now:

$$E\left(\boldsymbol{v}^{\mathsf{T}}_{(i)}\boldsymbol{X}(f(\boldsymbol{Y}))^{\mathsf{T}}\Big|\,\boldsymbol{\beta}\right) = E\left(\boldsymbol{v}^{\mathsf{T}}_{(i)}E(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\beta})f(\boldsymbol{Y})^{\mathsf{T}}\Big|\,\boldsymbol{\beta}\right) \tag{3.8}$$

where

$$E(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\beta}) = E[E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\beta})|\boldsymbol{Y},\boldsymbol{\beta}] = E[E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})|\boldsymbol{Y},\boldsymbol{\beta}] \tag{3.9}$$

where the second equality follows from the conditional independence $Y \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})$.

Now, since $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal\boldsymbol{X},\boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}^\intercal\boldsymbol{X}$, we have

$$E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal\boldsymbol{X},\boldsymbol{\beta}) = P_\beta^\intercal(\boldsymbol{\Sigma})\boldsymbol{X} = \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal\boldsymbol{X}.$$

By equation (3.9) we have

$$E(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\beta}) = \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}E(\boldsymbol{\beta}^\intercal\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\beta})$$

Combining with equation (3.8) we have

$$\begin{aligned} E\left(\boldsymbol{v}_{(i)}^\intercal\boldsymbol{X}(f(\boldsymbol{Y}))^\intercal\middle|\boldsymbol{\beta}\right) =& E\left(\boldsymbol{v}_{(i)}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}E(\boldsymbol{\beta}^\intercal\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\beta})(f(\boldsymbol{Y}))^\intercal\middle|\boldsymbol{\beta}\right) \\ =& \boldsymbol{v}_{(i)}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal E(\boldsymbol{X}(f(\boldsymbol{Y}))^\intercal|\boldsymbol{\beta}). \end{aligned}$$

Since $\boldsymbol{v}_i^\intercal\boldsymbol{\Sigma} = \boldsymbol{v}_i^\intercal\sigma_i^2$ the above becomes

$$E\left(\boldsymbol{v}_i^\intercal\boldsymbol{X}(f(\boldsymbol{Y}))^\intercal|\,\boldsymbol{\beta}\right) = \sigma_i^2\boldsymbol{v}_i^\intercal\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal E(\boldsymbol{X}(f(\boldsymbol{Y}))^\intercal|\boldsymbol{\beta}).$$

which by equation (3.18) means

$$\boldsymbol{\Sigma}_{w_i f(\boldsymbol{Y})} = \sigma_i^2\boldsymbol{v}_i^\intercal\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal E(f(\boldsymbol{Y})\boldsymbol{X}|\boldsymbol{\beta}) \tag{3.10}$$

Now using (3.25), (3.26), (3.27) and (3.10) we have

$$\rho_i = \sigma_i^2\boldsymbol{v}_i^\intercal\mathbf{A}\mathbf{A}^\intercal\boldsymbol{v}_i \tag{3.11}$$

where $\mathbf{A} = \boldsymbol{\beta}(\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal E(\boldsymbol{X}(f(\boldsymbol{Y}))^\intercal|\boldsymbol{\beta})\,(\mathrm{var}(f(\boldsymbol{Y})|\boldsymbol{\beta}))^{-1/2}$.

From this we have that (3.3) becomes

$$P(\sigma_i^2 \boldsymbol{v}_i^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{v}_i > \sigma_j^2 \boldsymbol{v}_j^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{v}_j) = P\left(\frac{\boldsymbol{v}_i^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{v}_i}{\boldsymbol{v}_j^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{v}_j} > \frac{\sigma_j^2}{\sigma_i^2}\right) > \frac{1}{2}$$

from the assumption that the median is unique and equal to 1. □

Using Proposition 2.2.1 the above result can be extended for random covariance matrix as the following corollary shows.

Corollary 3.1.4 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *random matrix from an orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;
2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta}$ *is a* $p \times q$ *random matrix and* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$;
4. $\boldsymbol{v}_i^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{v}_i / \boldsymbol{v}_j^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{v}_j$ *having unique median equal to 1; where*

$$\mathbf{A} = \boldsymbol{\beta}(\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}^\top E(f(\boldsymbol{Y})\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \left(\mathrm{var}(f(\boldsymbol{Y})|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right)^{-1/2}.$$

5. $E(\boldsymbol{X}|\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is linear in* $\boldsymbol{\beta}^\top \boldsymbol{X}$.

If $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^\top \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, *then, for any square-integrable function* $f(\boldsymbol{Y})$ *such that* $E[\boldsymbol{X} f(\boldsymbol{Y})|\boldsymbol{\beta}, \boldsymbol{\Sigma}] \neq \boldsymbol{0}$, $\mathrm{var}[f(\boldsymbol{Y})|\boldsymbol{\beta}, \boldsymbol{\Sigma}]$ *positive definite matrix and* $i < j$ *we have*

$$P(\rho_i > \rho_j) > 1/2, \tag{3.12}$$

where $\rho_i = \text{mcor}^2(f(\boldsymbol{Y}), \boldsymbol{v}_i^\intercal \boldsymbol{X} | \boldsymbol{\beta}, \boldsymbol{\Sigma})$.

3.2 Assuming a model where $E(Y|X,\beta) = E(Y|\beta^T X, \beta)$

3.2.1 Univariate response

Theorem 3.2.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *non random matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;

2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$;

3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ *is a spherically distributed random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$;

4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal \boldsymbol{X}, \boldsymbol{\beta})$ *is linear in* $\boldsymbol{\beta}^\intercal \boldsymbol{X}$.

5. Y *is a random variable such that* $E(Y|\boldsymbol{\beta}) = 0$, $P(\text{var}(Y|\boldsymbol{\beta}) < \infty) = 1$, *and*

$$P(\text{cov}(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X} | \boldsymbol{\beta}) \neq 0) = 1.$$

If $E(Y|\boldsymbol{X}, \boldsymbol{\beta}) = E(Y|\boldsymbol{\beta}^T \boldsymbol{X}, \boldsymbol{\beta})$, *then for* $i < j$,

$$P\left(\text{corr}^2(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X} | \boldsymbol{\beta}) > \text{corr}^2(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X} | \boldsymbol{\beta})\right) > 1/2.$$

Proof. We denote with $\mathbf{w}_1, \ldots, \mathbf{w}_p$ the ordered principal components of the

predictor vector $\boldsymbol{X}$. Let $i < j$. Since $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$, we have $E(\boldsymbol{X}|\boldsymbol{\beta}) = E(\boldsymbol{X}) = 0$. Hence

$$\begin{aligned}
\operatorname{cov}(Y, \mathbf{w}_i|\boldsymbol{\beta}) =& \operatorname{cov}(Y, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}) \\
=& E[E(Y|\boldsymbol{X}, \boldsymbol{\beta})\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}] \\
=& E[E(Y|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta})\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}] \\
=& E[Y\boldsymbol{v}_i^{\mathsf{T}}E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta})|\boldsymbol{\beta}]. \qquad (3.13)
\end{aligned}$$

Now let us compute $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta})$ in (3.13). Because, by assumption 3, $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$, we have

$$E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta}) = P_{\boldsymbol{\beta}}^{\mathsf{T}}(\boldsymbol{\Sigma})\boldsymbol{X} = \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}.$$

Substitute this into the right hand side of (3.13) to obtain

$$\operatorname{cov}(Y, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}) = E[Y\boldsymbol{v}_i^{\mathsf{T}}E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta})|\boldsymbol{\beta}] = \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}E(Y\boldsymbol{X}|\boldsymbol{\beta}).$$

In the meantime we note that $\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\Sigma} = \boldsymbol{v}_i^{\mathsf{T}}\sigma_i^2$ and

$$\operatorname{var}(\boldsymbol{v}_i\boldsymbol{X}|\beta) = \operatorname{var}(\boldsymbol{v}_i\boldsymbol{X}) = \sigma_i^2.$$

Combining the above we have that:

$$\operatorname{corr}^2(Y, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}) = \frac{\sigma_i^4(\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\beta})^2(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-2}[\boldsymbol{\beta}^{\mathsf{T}}E(Y\boldsymbol{X}|\boldsymbol{\beta})]^2}{\operatorname{var}(Y|\boldsymbol{\beta})\sigma_i^2}$$

Using these and the assumptions that $\operatorname{var}(Y|\boldsymbol{\beta}) > 0$ and $P(\operatorname{cov}(Y, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta}) \neq 0) = 1$ we obtain

$$\frac{\operatorname{corr}^2(Y, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta})}{\operatorname{corr}^2(Y, \boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta})} = \frac{\sigma_i^2(\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\beta})^2}{\sigma_j^2(\boldsymbol{v}_j^{\mathsf{T}}\boldsymbol{\beta})^2}$$

Hence

$$P\left(\text{corr}^2(Y, \boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}) > \text{corr}^2(Y, \boldsymbol{v}_j^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta})\right) = P((\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^2/(\boldsymbol{v}_j^\mathsf{T}\boldsymbol{\beta})^2 > \sigma_j^2/\sigma_i^2) > 1/2,$$

where the last inequality follows from $\sigma_j^2/\sigma_i^2 < 1$ and from Lemma 2.1.2. □

The following corollary extends the above result to the equality as the one showed by Ni (2010).

Corollary 3.2.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *non random matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;
2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ *is a spherically distributed random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$;
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta})$ *is linear in* $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$.
5. Y *is a random variable such that* $E(Y|\boldsymbol{\beta}) = 0$, $P(\text{var}(Y|\boldsymbol{\beta}) < \infty) = 1$, *and*

$$P(\text{cov}(Y, \boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}) \neq 0) = 1.$$

If $E(Y|\boldsymbol{X}, \boldsymbol{\beta}) = E(Y|\boldsymbol{\beta}^T\boldsymbol{X}, \boldsymbol{\beta})$, *then for* $i < j$,

$$P\left(\text{corr}^2(Y, \boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}) > \text{corr}^2(Y, \boldsymbol{v}_j^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta})\right) = \frac{2}{\pi}\arctan\left[(\lambda_i/\lambda_j)^{\frac{1}{2}}\right].$$

Using Proposition 2.2.1, the above two results, Theorem 3.2.1 and Corollary 3.2.1, extend to the case of random covariance matrices as the following two corollaries show.

Corollary 3.2.2 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *random matrix from an orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;
2. $\boldsymbol{X}$ *is a* p*-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ *is a spherically distributed random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$;
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is linear in* $\boldsymbol{\beta}^\intercal \boldsymbol{X}$.
5. Y *is a random variable such that* $E(Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = 0$, $P(\mathrm{var}(Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) < \infty) = 1$, *and*

$$P(\mathrm{cov}(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \neq 0) = 1.$$

If $E(Y|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = E(Y|\boldsymbol{\beta}^T \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, *then for* $i < j$,

$$P\left(\mathrm{corr}^2(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \mathrm{corr}^2(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) > 1/2.$$

Corollary 3.2.3 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *random matrix from an orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;

2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\text{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$*;*

3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ *is a spherically distributed random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$*;*

4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is linear in* $\boldsymbol{\beta}^\intercal\boldsymbol{X}$*.*

5. Y *is a random variable such that* $E(Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = 0$*,* $P(\text{var}(Y|\boldsymbol{\beta}, \boldsymbol{\Sigma}) < \infty) = 1$*, and*

$$P(\text{cov}(Y, \boldsymbol{v}_i^\intercal\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \neq 0) = 1.$$

If $E(Y|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = E(Y|\boldsymbol{\beta}^T\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$*, then for* $i < j$*,*

$$P\left(\text{corr}^2(Y, \boldsymbol{v}_i^\intercal\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \text{corr}^2(Y, \boldsymbol{v}_j^\intercal\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) = \frac{2}{\pi}\text{E}\left(\arctan\left[(\lambda_i/\lambda_j)^{\frac{1}{2}}\right]\right).$$

3.2.2 Multivariate response

Theorem 3.2.2 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *non random matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1*,* $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$*;*

2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$*;*

3. $\boldsymbol{\beta}$ *is a* $p \times q$ *is a spherically distributed random matrix and* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$*;*

4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal\boldsymbol{X}, \boldsymbol{\beta})$ *is linear in* $\boldsymbol{\beta}^\intercal\boldsymbol{X}$*.*

5. $\boldsymbol{Y}$ *is a random vector such that* $E(\boldsymbol{Y}|\boldsymbol{\beta}) = 0$*,* $\text{var}(\boldsymbol{Y}|\boldsymbol{\beta})$ *positive definite matrix.*

6. $\boldsymbol{v}_i^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_i/\boldsymbol{v}_j^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_j$ *has unique median equal to 1, where*

$$\mathbf{A} = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}E(\boldsymbol{X}\boldsymbol{Y}^{\mathsf{T}}|\boldsymbol{\beta})\left(\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta})\right)^{-1/2}$$

If $E(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta}) = E(\boldsymbol{Y}|\boldsymbol{\beta}^T\boldsymbol{X},\boldsymbol{\beta})$*, then for* $i < j$*,*

$$P\left(\rho_i > \rho_j\right) > 1/2, \tag{3.14}$$

where $\rho_i = \mathrm{mcor}^2(\boldsymbol{Y}, \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}|\boldsymbol{\beta})$*.*

Proof. By Definition 2.2.1,

$$\rho_i = \boldsymbol{\Sigma}_{w_i\boldsymbol{Y}}\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}w_i}\sigma_{w_i}^{-2} \tag{3.15}$$

where $w_i = \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}$ the i^{th} principal component.

We have that

$$\sigma_{w_i}^2 = \sigma_i^2, \tag{3.16}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta}) \tag{3.17}$$

and

$$\boldsymbol{\Sigma}_{w_i\boldsymbol{Y}} = \mathrm{cov}\left(w_i, \boldsymbol{Y}\,|\,\boldsymbol{\beta}\right) = \mathrm{cov}\left(\boldsymbol{v}_{(i)}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{Y}\,\middle|\,\boldsymbol{\beta}\right) = E\left(\boldsymbol{v}_{(i)}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{Y}^{\mathsf{T}}\,\middle|\,\boldsymbol{\beta}\right) \tag{3.18}$$

where the last equality holds since $E\left(\boldsymbol{X}|\boldsymbol{\beta}\right) = E(\boldsymbol{X}) = 0$ since $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{\beta}$. Now we

get that:

$$\begin{aligned}\boldsymbol{\Sigma}_{w_i\boldsymbol{Y}} = E\left(\boldsymbol{v}_{(i)}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{Y}^{\mathsf{T}}\middle|\boldsymbol{\beta}\right) =& E[\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}(E(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta}))^{\mathsf{T}}|\boldsymbol{\beta}] \\ =& E[\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}(E(\boldsymbol{Y}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta}))^{\mathsf{T}}|\boldsymbol{\beta}] \\ =& E[\boldsymbol{v}_i^{\mathsf{T}}E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})\boldsymbol{Y}^{\mathsf{T}}|\boldsymbol{\beta}]. \qquad (3.19)\end{aligned}$$

Now let us compute $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})$ in (3.19). Because, by assumption 3, $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$, we have

$$E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta}) = P_{\boldsymbol{\beta}}^{\mathsf{T}}(\boldsymbol{\Sigma})\boldsymbol{X} = \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}.$$

Substitute this into the right hand side of (3.19) to obtain

$$\boldsymbol{\Sigma}_{w_i\boldsymbol{Y}} = E[\boldsymbol{v}_i^{\mathsf{T}}E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X},\boldsymbol{\beta})\boldsymbol{Y}^{\mathsf{T}}|\boldsymbol{\beta}] = \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}E(\boldsymbol{X}\boldsymbol{Y}^{\mathsf{T}}|\boldsymbol{\beta}).$$

In the meantime we note that $\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{\Sigma} = \boldsymbol{v}_i^{\mathsf{T}}\sigma_i^2$ and so

$$\rho_i = \sigma_i^2\boldsymbol{v}_i^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_i \qquad (3.20)$$

where $\mathbf{A} = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathsf{T}}E(\boldsymbol{X}\boldsymbol{Y}^{\mathsf{T}}|\boldsymbol{\beta})\left(\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta})\right)^{-1/2}$. From this we have that (3.14) becomes

$$P(\sigma_i^2\boldsymbol{v}_i^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_i > \sigma_j^2\boldsymbol{v}_j^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_j) = P\left(\frac{\boldsymbol{v}_i^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_i}{\boldsymbol{v}_j^{\mathsf{T}}\mathbf{A}\mathbf{A}^{\mathsf{T}}\boldsymbol{v}_j} > \frac{\sigma_j^2}{\sigma_i^2}\right) > \frac{1}{2}$$

from the assumption for unique median. □

As before, this result can be extended to the case that we have a random covariance matrix.

Corollary 3.2.4 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *random matrix from an orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$
2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$*;*
3. $\boldsymbol{\beta}$ *is a* $p \times q$ *is a random matrix and* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$*;*
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is linear in* $\boldsymbol{\beta}^\intercal \boldsymbol{X}$*.*
5. $\boldsymbol{Y}$ *is a random vector such that* $E(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = 0$, $\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\Sigma})$ *positive definite matrix.*
6. $\boldsymbol{v}_i^\intercal \mathbf{A}\mathbf{A}^\intercal \boldsymbol{v}_i / \boldsymbol{v}_j^\intercal \mathbf{A}\mathbf{A}^\intercal \boldsymbol{v}_j$ *has unique median equal to 1, where*

$$\mathbf{A} = \boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta})^{-1} \boldsymbol{\beta}^\intercal E(\boldsymbol{X}\boldsymbol{Y}^\intercal | \boldsymbol{\beta}, \boldsymbol{\Sigma}) \left(\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right)^{-1/2}$$

If $E(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = E(\boldsymbol{Y}|\boldsymbol{\beta}^T \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$*, then for* $i < j$*,*

$$P\left(\rho_i > \rho_j\right) > 1/2, \tag{3.21}$$

where $\rho_i = \mathrm{mcor}^2(\boldsymbol{Y}, \boldsymbol{v}_i^\intercal \boldsymbol{X} | \boldsymbol{\beta}, \boldsymbol{\Sigma})$.

3.3 More results under the assumption $E(Y|X, \beta) = E(Y|\beta^T X, \beta)$

In this section we present an effort to generalize previous results to any type of nonlinear model. One of the extensions we tried was to check if the relationship

between principal components and the response variable in a regression model holds also for any polynomial function of principal components. In this section we present the result for the correlation of the squared principal component with the response. We didn't develop this results to their full potential to achieve our objectives, the reason being that the more general the polynomial is, the messier the calculations gets, and we were also able to develop results in a much clearer and better way. Those results are presented in a later Chapter, and they are much more general, than the results we present here.

3.3.1 Univariate model

Theorem 3.3.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *non random covariance matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$.

2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(X) = 0$ *and* $\mathrm{var}(X) = \boldsymbol{\Sigma}$;

3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ *is a a spherically distributed random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$;

4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal X, \boldsymbol{\beta})$ *is linear in* $\boldsymbol{\beta}^\intercal X$; $\mathrm{var}(X|\beta^\intercal X)$ *is nonrandom.*

If $E(Y|\boldsymbol{X}, \boldsymbol{\beta}) = E(Y|\boldsymbol{\beta}^\intercal \boldsymbol{X}, \boldsymbol{\beta})$ *and* $i < j$ *then*

$$P\left(\mathrm{corr}^2(Y, (\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}) > \mathrm{corr}^2(Y, (\boldsymbol{v}_j^\intercal \boldsymbol{X})^2|\boldsymbol{\beta})\right) > 1/2.$$

PROOF. We note that

$$\operatorname{cov}(Y,(\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}) = E(Y(\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}) - E(Y)E((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}). \tag{3.22}$$

First let us compute $E((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta})$:

$$E((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta}) = \operatorname{var}(\boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta}) + E^2(\boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta}).$$

By assumption 3,

$$\begin{aligned}
E(\boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta}) =& \boldsymbol{v}_i^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal \boldsymbol{X},\\
\operatorname{var}(\boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta}) =& \boldsymbol{v}_i^\intercal \left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\right)\boldsymbol{v}_i\\
=& \boldsymbol{v}_i^\intercal \left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\right)\boldsymbol{v}_i
\end{aligned}$$

So from the above and using the fact that $\boldsymbol{\Sigma}\boldsymbol{v}_i = \sigma_i^2\boldsymbol{v}_i$ we have that:

$$E((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta}) = \sigma_i^4(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-2}(\boldsymbol{\beta}^\intercal \boldsymbol{X})^2 + \sigma_i^2\boldsymbol{v}_i^\intercal \boldsymbol{v}_i - \sigma_i^4\boldsymbol{v}_i^\intercal \boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal \boldsymbol{v}_i \tag{3.23}$$

The first term on the right-hand side in (3.22) is rewritten as:

$$\begin{aligned}
E(Y(\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}) =& E(E(Y(\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{X},\boldsymbol{\beta})|\boldsymbol{\beta})\\
=& E(E(Y|\boldsymbol{X},\boldsymbol{\beta})(\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta})\\
=& E(E(Y|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta})(\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta})\\
=& E(YE((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}^\intercal \boldsymbol{X},\boldsymbol{\beta})|\boldsymbol{\beta})
\end{aligned}$$

Substitute (3.23) into the above to obtain:

$$E(Y(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta}) = \left(\sigma_i^2\boldsymbol{v}_i^\mathsf{T}\boldsymbol{v}_i - \sigma_i^4\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{v}_i\right)E(Y|\boldsymbol{\beta}) + \sigma_i^4(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^2(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-2}E\left(Y(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta}\right)$$

So substituting the above into (3.22) and simplifying we get that

$$\begin{aligned}\operatorname{cov}(Y,(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta}) =& -\sigma_i^4\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{v}_iE(Y|\boldsymbol{\beta}) + \sigma_i^4(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^2(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-2}E\left(Y(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta}\right)\\ =&\sigma_i^4(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^2c\end{aligned}$$

where $c = c(\boldsymbol{\beta})$ does not depend on i. Similarly

$$\operatorname{cov}(Y,(\boldsymbol{v}_j^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta}) = \sigma_j^4(\boldsymbol{v}_j^\mathsf{T}\boldsymbol{\beta})^2c$$

Now, we have that:

$$\operatorname{var}\left((\boldsymbol{v}_i^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta}\right) = E\left((\boldsymbol{v}_i^\mathsf{T}\boldsymbol{X})^4|\boldsymbol{\beta}\right) - \left(E\left((\boldsymbol{v}_i^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta}\right)\right)^2 = d\sigma_1^4\left(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{v}_i\right)^2 = d\sigma_1^4$$

because $\boldsymbol{v}_i^\mathsf{T}\boldsymbol{v}_i = 1$ since $\boldsymbol{v}_i, i = 1,\ldots,p$ form an orthonormal basis. Hence:

$$\operatorname{corr}^2(Y,\boldsymbol{v}_i^\mathsf{T}\boldsymbol{X}|\boldsymbol{\beta}) = \frac{\sigma_i^8(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^4c^2}{\operatorname{var}(Y|\boldsymbol{\beta})d\sigma_1^4} = \frac{\sigma_i^4(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^4c^2}{\operatorname{var}(Y|\boldsymbol{\beta})d}$$

Thus,

$$\begin{aligned}P\left(\operatorname{corr}^2(Y,(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta}) > \operatorname{corr}^2(Y,(\boldsymbol{v}_j^\mathsf{T}\boldsymbol{X})^2|\boldsymbol{\beta})\right) =& P\left(\frac{\sigma_i^4(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^4c^2}{\operatorname{var}(Y|\boldsymbol{\beta})d} > \frac{\sigma_j^4(\boldsymbol{v}_j^\mathsf{T}\boldsymbol{\beta})^4c^2}{\operatorname{var}(Y|\boldsymbol{\beta})d}\right)\\ =& P\left(\frac{(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\beta})^4}{(\boldsymbol{v}_j^\mathsf{T}\boldsymbol{\beta})^4} > \frac{\sigma_j^4}{\sigma_i^4}\right) > 1/2.\end{aligned}$$

where the last inequality is derived from Lemma 2.1.2. □

The above result can be extended as before in the case of a random covariance matrix as the following corollary shows.

Corollary 3.3.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *random matrix from an orientationally uniform matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;
2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\text{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ *is a random vector and* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$;
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^\intercal X, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is linear in* $\boldsymbol{\beta}^\intercal \boldsymbol{X}$; $\text{var}(\boldsymbol{X}|\beta^\intercal X)$ *is nonrandom.*

If $E(Y|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = E(Y|\boldsymbol{\beta}^\intercal \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *and* $i < j$ *then*

$$P\left(\text{corr}^2(Y, (\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > \text{corr}^2(Y, (\boldsymbol{v}_j^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}, \boldsymbol{\Sigma})\right) > 1/2.$$

Extensions to a similar one as the one in Ni (2010) needs further development and is not as straightforward as before. Further investigation is needed in this case to see exactly the type of equality that will holds.

3.3.2 Multivariate response

Theorem 3.3.2 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *non random covariance matrix and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$

corresponds to λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$*;*

2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}) = 0$ *and* $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$*;*

3. $\boldsymbol{\beta}$ *is a* $p \times q$ *is a spherically distributed random matrix and* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$*;*

4. $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta})$ *is linear in* $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$ *and* $\text{var}(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X})$ *is nonrandom.*

5. $\boldsymbol{Y}$ *is a random vector such that* $E(\boldsymbol{Y}|\boldsymbol{\beta}) = 0$*,* $\text{var}(\boldsymbol{Y}|\boldsymbol{\beta})$ *positive definite matrix.*

6. $(\boldsymbol{v}_i^{\mathsf{T}}\mathbf{C}\mathbf{C}^{\mathsf{T}}\boldsymbol{v}_i k + \mathbf{A}_i\mathbf{A}_i^{\mathsf{T}}) \,/\, \left(\boldsymbol{v}_j^{\mathsf{T}}\mathbf{C}\mathbf{C}^{\mathsf{T}}\boldsymbol{v}_j k + \mathbf{A}_j\mathbf{A}_j^{\mathsf{T}}\right)$ *has unique median equal to 1, where*

$$\begin{aligned} k =& (\mathrm{E}(\boldsymbol{Y}|\boldsymbol{\beta}))^{\mathsf{T}}(\text{var}(\boldsymbol{Y}|\boldsymbol{\beta}))^{-1}\mathrm{E}(\boldsymbol{Y}|\boldsymbol{\beta}) \\ \mathbf{C} =& \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1/2} \\ \mathbf{A}_i =& (\mathrm{E}((\boldsymbol{v}_i^{\mathsf{T}}\mathbf{C}\mathbf{C}^{\mathsf{T}}\boldsymbol{X})^2\boldsymbol{Y}|\boldsymbol{\beta}))^{\mathsf{T}}(\text{var}(\boldsymbol{Y}|\boldsymbol{\beta}))^{-1/2} \end{aligned}$$

If $E(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}) = E(\boldsymbol{Y}|\boldsymbol{\beta}^T\boldsymbol{X}, \boldsymbol{\beta})$*, then for* $i < j$*,*

$$P\left(\rho_i > \rho_j\right) > 1/2, \tag{3.24}$$

where $\rho_i = \text{mcor}^2(\boldsymbol{Y}, (\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X})^2\,|\boldsymbol{\beta})$*.*

Proof. By Definition 2.2.1,

$$\rho_i = \boldsymbol{\Sigma}_{w_i^2\boldsymbol{Y}|\boldsymbol{\beta}}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{Y}|\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\boldsymbol{Y}w_i^2|\boldsymbol{\beta}}\sigma^{-2}_{w_i^2|\boldsymbol{\beta}} \tag{3.25}$$

where $w_i = \boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{X}$ the i^{th} principal component.

We have that

$$\sigma^2_{w_i^2|\boldsymbol{\beta}} = \text{var}\left((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2 \,|\boldsymbol{\beta}\right) = E\left((\boldsymbol{v}_i^\intercal \boldsymbol{X})^4|\boldsymbol{\beta}\right) - \left(E\left((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}\right)\right)^2 = d\sigma_1^4 \left(\boldsymbol{v}_i^\intercal \boldsymbol{v}_i\right)^2 = d\sigma_1^4, \tag{3.26}$$

because $\boldsymbol{v}_i^\intercal \boldsymbol{v}_i = 1$ since $\boldsymbol{v}_i, i = 1, \ldots, p$ form an orthonormal basis. Also:

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}|\boldsymbol{\beta}} = \text{var}(\boldsymbol{Y}|\boldsymbol{\beta}) \tag{3.27}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_{w_i^2\boldsymbol{Y}|\boldsymbol{\beta}} =& \text{cov}\left(w_i^2, \boldsymbol{Y}\,\middle|\,\boldsymbol{\beta}\right) = \text{cov}\left((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2, \boldsymbol{Y}\,\middle|\,\boldsymbol{\beta}\right) \\ =& E\left((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2 \boldsymbol{Y}^\intercal\,\middle|\,\boldsymbol{\beta}\right) - E(\boldsymbol{Y}|\boldsymbol{\beta})E((\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta})\end{aligned}$$

which is similar to equation (3.22) in the proof of Theorem 3.3.1. So following a similar derivation (the only difference is that now $\boldsymbol{\beta}$ is a matrix and $\boldsymbol{Y}$ is a vector) we get that:

$$\begin{aligned}\text{cov}(\boldsymbol{Y}, (\boldsymbol{v}_i^\intercal \boldsymbol{X})^2|\boldsymbol{\beta}) =& -\sigma_i^4 \boldsymbol{v}_i^\intercal \boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal \boldsymbol{v}_i E(\boldsymbol{Y}|\boldsymbol{\beta}) + \sigma_i^4 E\left((\boldsymbol{v}_i^\intercal \boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\intercal \boldsymbol{X})^2 \boldsymbol{Y}|\boldsymbol{\beta}\right) \\ =& \sigma_i^4 \left(\boldsymbol{v}_i^\intercal \mathbf{C}\mathbf{C}^\intercal \boldsymbol{v}_i E(\boldsymbol{Y}|\boldsymbol{\beta}) + E((\boldsymbol{v}_i^\intercal \mathbf{C}\mathbf{C}^\intercal \boldsymbol{X})^2 \boldsymbol{Y}|\boldsymbol{\beta})\right)\end{aligned}$$

where $\mathbf{C} = \boldsymbol{\beta}(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta})^{-1/2}$. Combining everything we get that:

$$\begin{aligned}\rho_i =& d^{-2}\sigma_i^4 \left(\boldsymbol{v}_i^\intercal \mathbf{C}\mathbf{C}^\intercal \boldsymbol{v}_i E(\boldsymbol{Y}|\boldsymbol{\beta}) + E((\boldsymbol{v}_i^\intercal \mathbf{C}\mathbf{C}^\intercal \boldsymbol{X})^2 \boldsymbol{Y}|\boldsymbol{\beta})\right)^\intercal (\text{var}(\boldsymbol{Y}|\boldsymbol{\beta}))^{-1} \\ & \left(\boldsymbol{v}_i^\intercal \mathbf{C}\mathbf{C}^\intercal \boldsymbol{v}_i E(\boldsymbol{Y}|\boldsymbol{\beta}) + E((\boldsymbol{v}_i^\intercal \mathbf{C}\mathbf{C}^\intercal \boldsymbol{X})^2 \boldsymbol{Y}|\boldsymbol{\beta})\right) \\ =& d^{-2}\sigma_i^4 \left(\boldsymbol{v}_i^\intercal \mathbf{C}\mathbf{C}^\intercal \boldsymbol{v}_i k + \mathbf{A}_i \mathbf{A}_i^\intercal\right)\end{aligned}$$

where $k = (\mathrm{E}(\boldsymbol{Y}|\boldsymbol{\beta}))^{\mathsf{T}}(\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta}))^{-1}\mathrm{E}(\boldsymbol{Y}|\boldsymbol{\beta})$ and

$$\mathbf{A}_i = (\mathrm{E}((\boldsymbol{v}_i^{\mathsf{T}}\mathbf{C}\mathbf{C}^{\mathsf{T}}\boldsymbol{X})^2\boldsymbol{Y}|\boldsymbol{\beta}))^{\mathsf{T}}(\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta}))^{-1/2}$$

The above means that what we are trying to prove is:

$$P\left(\rho_i > \rho_j\right) > 1/2 \Rightarrow$$
$$P\left(\frac{(\boldsymbol{v}_i^{\mathsf{T}}\mathbf{C}\mathbf{C}^{\mathsf{T}}\boldsymbol{v}_i k + \mathbf{A}_i\mathbf{A}_i^{\mathsf{T}})}{\left(\boldsymbol{v}_j^{\mathsf{T}}\mathbf{C}\mathbf{C}^{\mathsf{T}}\boldsymbol{v}_j k + \mathbf{A}_j\mathbf{A}_j^{\mathsf{T}}\right)} > \frac{\sigma_j^4}{\sigma_i^4}\right) > 1/2$$

□

The following corollary shows the extension when the covariance matrix is randomly distributed from an orientationally uniform distribution.

Corollary 3.3.2 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *random matrix from an orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$;
2. $\boldsymbol{X}$ *is a p-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$;
3. $\boldsymbol{\beta}$ *is a* $p \times q$ *is a random matrix and* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$;
4. $E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ *is linear in* $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X})$ *is nonrandom.*
5. $\boldsymbol{Y}$ *is a random vector such that* $E(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = 0$, $\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\Sigma})$ *positive definite matrix.*

6. $(\boldsymbol{v}_i^{\intercal}\mathbf{C}\mathbf{C}^{\intercal}\boldsymbol{v}_i k + \mathbf{A}_i\mathbf{A}_i^{\intercal}) \,/\, \left(\boldsymbol{v}_j^{\intercal}\mathbf{C}\mathbf{C}^{\intercal}\boldsymbol{v}_j k + \mathbf{A}_j\mathbf{A}_j^{\intercal}\right)$ *has unique median equal to 1, where*

$$\begin{aligned} k =& (\mathrm{E}(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\Sigma}))^{\intercal}(\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\Sigma}))^{-1}\mathrm{E}(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\Sigma}) \\ \mathbf{C} =& \boldsymbol{\beta}(\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1/2} \\ \mathbf{A}_i =& (\mathrm{E}((\boldsymbol{v}_i^{\intercal}\mathbf{C}\mathbf{C}^{\intercal}\boldsymbol{X})^2\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\Sigma}))^{\intercal}(\mathrm{var}(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\Sigma}))^{-1/2}. \end{aligned}$$

If $E(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\Sigma}) = E(\boldsymbol{Y}|\boldsymbol{\beta}^T\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\Sigma})$*, then for* $i < j$*,*

$$P\left(\rho_i > \rho_j\right) > 1/2, \tag{3.28}$$

where $\rho_i = \mathrm{mcor}^2(\boldsymbol{Y}, (\boldsymbol{v}_i^{\intercal}\boldsymbol{X})^2\,|\boldsymbol{\beta},\boldsymbol{\Sigma})$.

3.4 Discussion

The results in this Chapter, is the first effort we made to extend the results by Artemiou and Li (2009) and Ni (2010) in a more general framework than the linear model. What we did was to extend the results under two different assumptions, the one is the conditional independence model and the other is the equality of two conditional expectations, $\mathrm{E}(Y|X,\boldsymbol{\beta}) = \mathrm{E}(Y|\boldsymbol{\beta}^{\intercal}\boldsymbol{X},\boldsymbol{\beta})$. Those extensions, study the relationship of the linear principal components with the response.

There is ground for several other developments to be made and more importantly to improve the developments presented here especially in the following 3 areas:

1. The investigation of the conditions that for the multivariate response cases,

the desired ratios of variables will have unique median equal to 1.

2. The investigation of an exact equality as the one in Ni (2010) in the multivariate cases.

3. The investigation of an exact equality as the one in Ni (2010) for the univariate case in the last section of the three, which deals with the relationship of the squared principal components with the response.

Some of the results presented are special cases of the more general results developed in the next Chapters, but they are presented for completeness of the work developed, and for smoother presentation of the material.

Chapter 4

Information Criterion

As in the previous Chapter, the objective in this one is to extend the results of Artemiou and Li (2009) and Ni (2010) in a more general framework than the linear regression model. In this Chapter, we propose an information criterion which we believe can be used as a measure of relation between linear principal components and the response variable. Correlation is a measure associated with the linear relationship between variables. This criterion can be used in a more general framework than the linear regression model and can serve as a measure of general association (and not just linear association) between variables. This might be helpful in extending the previous results in a more general framework than the ones presented in the previous Chapters.

The results presented in this Chapter are very limited. We were not able to extend it beyond the linear case and beyond the normality assumption for the predictors. Basically this is because of the fact that we found the results in the next Chapter more interesting so that we found it more worthy to develop those results as best as we could.

Definition 4.0.1 *We define the information criterion between two random variables X and Y conditional on random variable W to be:*

$$I(X,Y|W) = E\left(\log\left(\frac{f(X,Y|W)}{f(Y|W)f(X|W)}\right)|W\right).$$

In the following sections we will first give the proof in the simple 2 dimensional predictor case and then we will prove the more general results. The simple case is presented for completeness and for the smoother introduction of the reader in the more general result.

4.1 Simple case - 2 dimensional predictor

First, we prove a helpful lemma about the normal distribution and then we prove the main theorem for the inequality (as the one in Artemiou (2008)) which is a special case of the more general inequality.

Lemma 4.1.1 *Suppose $Z \sim N_r(0,\Lambda)$. Then:*

$$\mathrm{E}(\log f(Z)) = -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log\det(\Lambda) - \frac{r}{2}.$$

Proof. From the distribution function of multivariate normal we have that:

$$\log f(z) = -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log\det(\Lambda) - \frac{1}{2}(z^{\top}\Lambda^{-1}z).$$

So:

$$\begin{aligned}\mathrm{E}(\log f(z)) =& \mathrm{E}\left(-\frac{r}{2}\log(2\pi) - \frac{1}{2}\log\det(\Lambda) - \frac{1}{2}(z^{\intercal}\Lambda^{-1}z)\right)\\ =& -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log\det(\Lambda) - \frac{1}{2}\mathrm{E}\left((z^{\intercal}\Lambda^{-1}z)\right)\\ =& -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log\det(\Lambda) - \frac{r}{2}.\end{aligned}$$

□

Theorem 4.1.1 *Let*

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix},$$

where σ_1^2 *and* σ_2^2 *are iid* G*. Let* $\theta \sim U(0,\pi)$*. Let* Γ *be the random matrix*

$$\Gamma = \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$$

Let $\boldsymbol{\Sigma} = \Gamma\boldsymbol{\Sigma}_0\Gamma^T$*. Suppose*

1. $\boldsymbol{X}$ *is a 2-dimensional random vector with* $E(\boldsymbol{X}|\boldsymbol{\Sigma}) = 0$ *and* $\mathrm{var}(\boldsymbol{X}|\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}$*,*
2. $Y = \boldsymbol{\beta}^T\boldsymbol{X} + \delta$*, where* $\boldsymbol{\beta}$ *is a p-dimensional random vector and* δ *is a random variable such that* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$*,* $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$*,* $E(\delta) = 0$ *and* $\mathrm{var}(\delta) < \infty$*.*
3. $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set* $G \in \mathbb{R}^2$*.*

Let λ_1, λ_2 *to be the ordered eigenvalues of the covariance matrix* $\boldsymbol{\Sigma}$ *and* w_1, w_2 *be the 1st and 2nd principal components of* $\boldsymbol{X}$*, and let* $\rho_i = \rho_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathrm{corr}^2(Y, w_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}), i =$

1, 2. *Then,*

$$P(\rho_1 \geq \rho_2) > \frac{1}{2}. \tag{4.1}$$

PROOF. We have, for $i = 1, 2$,

$$\begin{aligned} I(Y, X_i|\beta) &= E[\log p(Y, X_i|\beta)|\beta] - E[\log p(Y|\beta)|\beta] - E[\log p(X_i|\beta)|\beta] \\ &= E[\log p(Y, X_i|\beta)|\beta] - E[\log p(Y|\beta)|\beta] - E[\log p(X_i)], \end{aligned}$$

where the second equality follows from $\beta \perp\!\!\!\perp X$. So $I(Y, X_1|\beta) > I(Y, X_2|\beta)$ is equivalent to

$$E[\log p(Y, X_1|\beta)|\beta] - E[\log p(X_1)] > E[\log p(Y, X_2|\beta)|\beta] - E[\log p(X_2)]$$

This is equivalent to

$$E[\log p(Y, X_1|\beta)|\beta] - E[\log p(Y, X_2|\beta)|\beta] > E[\log p(X_1)] - E[\log p(X_2)]$$

Let us see what will happen when X is normal and Y is linear as described in the above. We have

$$\log p(x_1) = -(1/2)\log(2\pi) - \log(\sigma_1) - (1/2)(x_1^2/\sigma_1^2)$$

So

$$E \log p(X_1) = -(1/2)\log(2\pi) - \log(\sigma_1) - (1/2)$$

So

$$E\log(X_1) - E\log(X_2) = -\log(\sigma_1) + \log(\sigma_2) = \log(\sigma_2/\sigma_1).$$

Now from Lemma 4.1.1 for $r = 2$ we have that

$$\mathrm{var}(Y) = \beta^T\Sigma\beta + \tau^2, \quad \mathrm{var}(X_1) = \sigma_1^2, \quad \mathrm{cov}(Y, X_1) = \beta_1\sigma_1^2$$

So

$$\det\Lambda_1 = \det\mathrm{var}[(Y, X_1)^\intercal] = \sigma_1^2(\beta^T\Sigma\beta + \tau^2) - \beta_1^2\sigma_1^4.$$

So

$$\begin{aligned}
&E[\log p(Y, X_1|\beta)|\beta] - E[\log p(Y, X_2|\beta)|\beta] \\
&= (-1/2)\log[\sigma_1^2(\beta^T\Sigma\beta + \tau^2) - \beta_1^2\sigma_1^4] + (1/2)\log[\sigma_2^2(\beta^T\Sigma\beta + \tau^2) - \beta_2^2\sigma_2^4]
\end{aligned}$$

In the above

$$\begin{aligned}
\sigma_1^2(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + \tau^2) - \beta_1^2\sigma_1^4 &= \sigma_1^2(\beta_2^2\sigma_2^2 + \tau^2) \\
\sigma_2^2(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + \tau^2) - \beta_2^2\sigma_2^4 &= \sigma_2^2(\beta_1^2\sigma_1^2 + \tau^2)
\end{aligned}$$

So

$$\begin{aligned}
&E[\log p(Y, X_1|\beta)|\beta] - E[\log p(Y, X_2|\beta)|\beta] \\
&= (-1/2)\log[\sigma_1^2(\beta_2^2\sigma_2^2 + \tau^2)] + (1/2)\log[\sigma_2^2(\beta_1^2\sigma_1^2 + \tau^2)]
\end{aligned}$$

So we want

$$(-1/2)\log[\sigma_1^2(\beta_2^2\sigma_2^2+\tau^2)]+(1/2)\log[\sigma_2^2(\beta_1^2\sigma_1^2+\tau^2)]>(1/2)\log\sigma_2^2-(1/2)\log\sigma_1^2.$$

This is equivalent to

$$\begin{aligned}
&\log[\sigma_2^2(\beta_1^2\sigma_1^2+\tau^2)]-\log[\sigma_1^2(\beta_2^2\sigma_2^2+\tau^2)]>\log\sigma_2^2-\log\sigma_1^2\\
\Leftrightarrow&\log[(\beta_1^2\sigma_1^2+\tau^2)]-\log[(\beta_2^2\sigma_2^2+\tau^2)]>0\\
\Leftrightarrow&\beta_1^2\sigma_1^2+\tau^2>\beta_2^2\sigma_2^2+\tau^2\\
\Leftrightarrow&\beta_1^2\sigma_1^2>\beta_2^2\sigma_2^2
\end{aligned}$$

This is the same inequality that Artemiou (2008) shows for the 2-dimensional case which is a special case for the more general result which is presented in Artemiou and Li (2009)

4.2 General case - p dimensional predictor

In this section we extend the main result of the previous section for a p-dimensional predictor vector $\boldsymbol{X}$.

Theorem 4.2.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p\times p$ *non random matrix and* $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1\geq\ldots\geq\lambda_p$ *are the eigenvalues of* $\boldsymbol{\Sigma}$*;*

2. $\boldsymbol{X}$ *is a p-dimensional random vector such that* $\boldsymbol{X}\sim N(0,\boldsymbol{\Sigma})$*;*

3. *$\boldsymbol{\beta}$ is a p-dimensional spherically distributed random vector, $\boldsymbol{\beta} \perp\!\!\!\perp X$ and $P(\boldsymbol{\beta} \in G) > 0$ for any nonempty open set G;*

4. *ϵ is a random variable with $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \tau^2$, $\epsilon \perp\!\!\!\perp \boldsymbol{X}$.*

5. *$Y = \boldsymbol{\beta}^\intercal \boldsymbol{X} + \epsilon$*

Then, for $i < j$,

$$P(I(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}) > I(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta})) > 1/2. \tag{4.2}$$

PROOF. First, we use the definition of information, Definition 4.0.1, to expand the two information involved on the left hand side of inequality(4.6).

$$\begin{aligned} I(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}) =& E\left(\log\left(\frac{f(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta})}{f(Y|\boldsymbol{\beta}) f(\boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta})}\right)|\boldsymbol{\beta}\right) \\ =& E(\log f(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta})) - E(\log(f(Y|\boldsymbol{\beta}))) - E(\log(f(\boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}))) \end{aligned} \tag{4.3}$$

Now we know that $\boldsymbol{v}_i^\intercal \boldsymbol{X} \sim N(0, \boldsymbol{v}_i^\intercal \boldsymbol{\Sigma} \boldsymbol{v}_i)$ then from Lemma 4.1.1 we have that:

$$E(\log(f(\boldsymbol{v}_i^\intercal \boldsymbol{X}))) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \boldsymbol{v}_i^\intercal \boldsymbol{\Sigma} \boldsymbol{v}_i - \frac{1}{2} \tag{4.4}$$

Now we have that the joint distribution of Y and $\boldsymbol{v}_i^\intercal \boldsymbol{X}$ is the following:

$$\begin{pmatrix} Y \\ \boldsymbol{v}_i^\intercal \boldsymbol{X} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathrm{var}(Y) & \mathrm{cov}(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}) \\ \mathrm{cov}(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}) & \mathrm{var}(\boldsymbol{v}_i^\intercal \boldsymbol{X}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2 & \sigma_i^2 \boldsymbol{v}_i^\intercal \boldsymbol{\beta} \\ \sigma_i^2 \boldsymbol{v}_i^\intercal \boldsymbol{\beta} & \sigma_i^2 \end{pmatrix}\right).$$

Using Lemma 4.1.1 and the fact that $\det \operatorname{var}(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}) = \sigma_i^2(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2) - \sigma_i^4(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2$ we get the following result:

$$E(\log(f(\boldsymbol{v}_i^\intercal \boldsymbol{X}))) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log(\sigma_i^2(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2) - \sigma_i^4(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2) - \frac{2}{2} \tag{4.5}$$

Combining equations (4.3), (4.4) and (4.5) we have the following equation

$$\begin{aligned} I(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}) = &-\frac{1}{2}\log 2\pi - \frac{1}{2}\log(\sigma_i^2(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2) - \sigma_i^4(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2) - \frac{2}{2} \\ &- E(\log(f(Y|\boldsymbol{\beta}))) + \frac{1}{2}\log 2\pi + \frac{1}{2}\log \boldsymbol{v}_i^\intercal \boldsymbol{\Sigma}\boldsymbol{v}_i + \frac{1}{2} \end{aligned}$$

Similarly,

$$\begin{aligned} I(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta}) = &-\frac{1}{2}\log 2\pi - \frac{1}{2}\log(\sigma_j^2(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2) - \sigma_j^4(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2) - \frac{2}{2} \\ &- E(\log(f(Y|\boldsymbol{\beta}))) + \frac{1}{2}\log 2\pi + \frac{1}{2}\log \boldsymbol{v}_j^\intercal \boldsymbol{\Sigma}\boldsymbol{v}_j + \frac{1}{2} \end{aligned}$$

So using the facts that $\boldsymbol{v}_i^\intercal \boldsymbol{\Sigma} = \sigma_i^2 \boldsymbol{v}_i^\intercal$, $\boldsymbol{v}_i^\intercal \boldsymbol{v}_i = 1$ and by canceling similar terms, the left hand side of inequality (4.6) reduces to:

$$\begin{aligned} &P(I(Y, \boldsymbol{v}_i^\intercal \boldsymbol{X}|\boldsymbol{\beta}) > I(Y, \boldsymbol{v}_j^\intercal \boldsymbol{X}|\boldsymbol{\beta})) \\ =&P(-\log(\sigma_i^2(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2) - \sigma_i^4(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2) + \log \sigma_i^2 > -\log(\sigma_j^2(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2) - \sigma_j^4(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2) + \log \sigma_j^2) \\ =&P(-\log(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2 - \sigma_i^2(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2) > -\log(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2 - \sigma_j^2(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2)) \\ =&P(\log(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2 - \sigma_j^2(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2) > \log(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2 - \sigma_i^2(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2)) \\ =&P(\boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2 - \sigma_j^2(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2 > \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}\boldsymbol{\beta} + \tau^2 - \sigma_i^2(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2) \\ =&P(-\sigma_j^2(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2 > -\sigma_i^2(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2) \\ =&P(\sigma_i^2(\boldsymbol{v}_i^\intercal \boldsymbol{\beta})^2 > \sigma_j^2(\boldsymbol{v}_j^\intercal \boldsymbol{\beta})^2) \end{aligned}$$

which is simplified to:

$$P\left(\frac{(\boldsymbol{v}_i^\intercal\boldsymbol{\beta})^2}{(\boldsymbol{v}_j^\intercal\boldsymbol{\beta})^2} > \frac{\sigma_j^2}{\sigma_i^2}\right)$$

which is greater than 1/2 because of Lemma 2.1.2 and the fact that $\sigma_j^2/\sigma_i^2 < 1$. □

The following corollary shows that we can extend the result for random covariance matrix using Proposition 2.2.1.

Corollary 4.2.1 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *random matrix from a orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$ *are the eigenvalues of* $\boldsymbol{\Sigma}$*;*

2. $\boldsymbol{X}$ *is a p-dimensional random vector such that* $\boldsymbol{X}|\boldsymbol{\Sigma} \sim N(0, \boldsymbol{\Sigma})$*;*

3. $\boldsymbol{\beta}$ *is a p-dimensional random vector,* $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$ *and* $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set* G*;*

4. ϵ *is a random variable with* $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \tau^2$, $\epsilon \perp\!\!\!\perp \boldsymbol{X}$.

5. $Y = \boldsymbol{\beta}^\intercal\boldsymbol{X} + \epsilon$

Then, for $i < j$,

$$P(I(Y, \boldsymbol{v}_i^\intercal\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) > I(Y, \boldsymbol{v}_j^\intercal\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma})) > 1/2. \tag{4.6}$$

The following two corollaries show that the results above can be extended to the result that was proved by Ni (2010).

Corollary 4.2.2 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *non random matrix from a orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$ *are the eigenvalues of* $\boldsymbol{\Sigma}$*;*

2. $\boldsymbol{X}$ *is a p-dimensional random vector such that* $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$*;*

3. $\boldsymbol{\beta}$ *is a p-dimensional spherically distributed random vector,* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$ *and* $P(\boldsymbol{\beta} \in G) > 0$ *for any nonempty open set G;*

4. ϵ *is a random variable with* $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \tau^2$, $\epsilon \perp\!\!\!\perp \boldsymbol{X}$.

5. $Y = \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X} + \epsilon$

Then, for $i < j$,

$$P(I(Y, \boldsymbol{v}_i^{\mathsf{T}} \boldsymbol{X} | \boldsymbol{\beta}) > I(Y, \boldsymbol{v}_j^{\mathsf{T}} \boldsymbol{X} | \boldsymbol{\beta})) = \frac{2}{\pi} \arctan \left[(\lambda_i / \lambda_j)^{\frac{1}{2}} \right] \tag{4.7}$$

Corollary 4.2.3 *Suppose*

1. $\boldsymbol{\Sigma}$ *is a* $p \times p$ *random matrix from a orientationally uniform distribution and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ *are the eigenvectors of* $\boldsymbol{\Sigma}$ *in the sense that* $\boldsymbol{v}_1$ *corresponds to the largest eigenvalue* λ_1, $\boldsymbol{v}_2$ *corresponds to* λ_2 *and so on, were* $\lambda_1 \geq \ldots \geq \lambda_p$ *are the eigenvalues of* $\boldsymbol{\Sigma}$*;*

2. $\boldsymbol{X}$ *is a p-dimensional random vector such that* $\boldsymbol{X} | \boldsymbol{\Sigma} \sim N(0, \boldsymbol{\Sigma})$*;*

3. *$\boldsymbol{\beta}$ is a p-dimensional random vector, $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$ and $P(\boldsymbol{\beta} \in G) > 0$ for any nonempty open set G;*

4. *ϵ is a random variable with $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \tau^2$, $\epsilon \perp\!\!\!\perp \boldsymbol{X}$.*

5. *$Y = \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X} + \epsilon$*

Then, for $i < j$,

$$P(I(Y, \boldsymbol{v}_i^{\mathsf{T}} \boldsymbol{X} | \boldsymbol{\beta}, \boldsymbol{\Sigma}) > I(Y, \boldsymbol{v}_j^{\mathsf{T}} \boldsymbol{X} | \boldsymbol{\beta}, \boldsymbol{\Sigma})) = \frac{2}{\pi} \mathrm{E}\left(\arctan\left[(\lambda_i / \lambda_j)^{\frac{1}{2}} \right] \right). \tag{4.8}$$

The proof is straightforward as Ni (2010) showed that the last probability statement of the proof for Theorem 4.2.1 and Corollary 4.2.1 is equal to the right hand side of equation(4.7) and equation (4.8) respectively.

As we said at the beginning of this Chapter, this is an interesting result which has potential but is not yet developed as completely as it should be. It is limited in a very special case and is obviously not serving the purpose to extend the results in the non linear regression model. Further development of this result is left for future work.

Chapter 5

Generalizations using Kernel Principal Components

In this Chapter we will present the most important results for the first part of this work. We will extend the results by Artemiou and Li (2009) and Ni (2010) to the nonlinear regression setting, by assuming first an arbitrary nonparametric regression setting, then an arbitrary relation between $\boldsymbol{X}$ and Y and finally we will connect those results to the sufficient dimension reduction setting.

The extension to a more general setting than the linear regression setting, has been the objective in the two previous Chapters as well. The main difference in this Chapter, that enables us to generalize this result even in the cases that it essentially has no restrictive assumptions on the relationship between $\boldsymbol{X}$ and Y, is the use of kernel principal components instead of the linear principal components we previously used. A short review on kernel principal components followed by some motivating examples, are given in the next few sections before we present our

results.

5.1 Introducing Kernel Principal Components

Kernel principal component analysis was introduced by Schölkopf, B., Smola, A., Müller (1997, 1998) and it is one of the most widely used method for nonlinear unsupervised dimension reduction. The idea is described as follows; one can map the observed vectors into a higher dimensional space, called the feature space, using a kernel function $\phi(\cdot)$ and then try to perform linear principal component analysis in the feature space. When the linear functions in the feature space are mapped back into the original input space where the observed vectors lie, they are considered nonlinear functions of the predictors. Thus, the first kernel principal component will be the direction that captures the most variation among all possible functions (linear and nonlinear) in the input space. This, expands the idea of linear principal components that we were using in previous Chapters, to extract nonlinear features and thus to achieve nonlinear dimension reduction. In the literature there are other approaches to nonlinear feature extraction based on the principal component methodology, like principal curves (Hastie and Stuetzle, (1989)) and functional principal component analysis (Rice and Silverman, (1991) and Silverman, (1996)).

In the linear principal component analysis, if we assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are the observed vectors one performs an eigenvalue decomposition of the sample covariance matrix, which assuming the observed vectors are centered at $\mathbf{0}$, is equal to

$$\mathbf{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\top}$$

For the kernel principal component, we have the mapped observed vectors into the feature space $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$ and our objective is to perform an eigenvalue decomposition of the sample covariance matrix of those mappings, which assuming they are centered at $\mathbf{0}$ it is equal to

$$\mathbf{\Sigma}^* = \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\intercal$$

where

$$\phi : \mathbb{R}^p \to F$$

and F is our arbitrary feature space which might be infinite dimensional. Schölkopf, B., Smola, A., Müller (1997) showed that for certain choices of ϕ, it is easy to perform linear PCA using kernel functions as they were presented in the support vector machine literature (Boser, B. E., Guyon, I. M. and Vapnik V. (1992) and Cortes, C. and Vapnik, V. (1995)). This is known in the literature as the "kernel trick"; that is, if an operation depends only on inner products one can extract lower dimensional projections without dealing with the projection coefficients (in our case $\phi(\mathbf{x}_i), i = 1, \ldots, n$) which reside in the higher dimensional space (feature space). This idea is widely used in the machine learning literature (Vapnik V. (1998)) and recently in the sufficient (or supervised) dimension reduction literature (Fukumizu, Bach, and Jordan (2004, 2009), Yeh, Huang, and Lee (2009), Hsing and Ren (2009), Shi, Belkin, and Yu (2009)).

5.2 Kernel PCA and its predictive potential

A small clarification for the notation that is being used is that the notation $\stackrel{\mathcal{D}}{=}$ stands for "equal in distribution". Thus $U \stackrel{\mathcal{D}}{=} V$ means that U and V have the same distribution, and $U|W \stackrel{\mathcal{D}}{=} V|W$ means that the conditional distribution of U given W is the same as that of V given W.

Suppose that $\boldsymbol{X}$ and Y are defined on a probability space $(\Omega, \mathfrak{F}, P)$, and let $\Omega_{\boldsymbol{X}} = \{\boldsymbol{X}(\omega) : \omega \in \Omega\}$ denote the range of $\boldsymbol{X}$. Let $\mathcal{H}$ be a separable Hilbert space whose members are real-valued functions defined on $\Omega_{\boldsymbol{X}}$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denote the inner product in $\mathcal{H}$, and $\| \cdot \|_{\mathcal{H}}$ denote the induced norm. We assume throughout this Chapter that $\mathcal{H}$ is relative to the scalar field of real numbers $\mathbb{R}$.

At the population level, kernel principal component analysis (Schölkopf, Smola, and Müller, 1997), or kernel PCA, can be described as follows. The first kernel principal component is the function u_1 in $\mathcal{H}$ that maximizes

$$\text{var}[f(\boldsymbol{X})] \tag{5.1}$$

among all $f \in \mathcal{H}$ satisfying $\|f\|_{\mathcal{H}} = 1$. For $k = 2, 3, \ldots$, the kth kernel principal is the member of u_k of $\mathcal{H}$ that maximizes (5.1) subject to the constraints

$$\text{cov}[u_k(\boldsymbol{X}), u_i(\boldsymbol{X})] = 0, \quad i = 1, \ldots, k-1, \ \|u_k\|_{\mathcal{H}} = 1.$$

This is much more general than the classical (linear) PCA because the maximization is carried out among all functions in $\mathcal{H}$ — not just linear functions of the form $a^{\intercal}\boldsymbol{X}$. The term "kernel" comes from the fact that $\mathcal{H}$ may be taken to be a reproducing kernel Hilbert space derived from a positive definite mapping, or

kernel function, $K : \Omega_{\boldsymbol{X}} \times \Omega_{\boldsymbol{X}} \to \mathbb{R}$. In this context, $\mathcal{H}$ is the closed linear span of functions of the form

$$a_1 K(\cdot, x_1) + \cdots + a_m K(\cdot, x_m), \quad x_1, \ldots, x_m \in \Omega_{\boldsymbol{X}},\ a_1, \ldots, a_m \in \mathbb{R}, \tag{5.2}$$

and the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is specified by $\langle K(\cdot, x_1), K(\cdot, x_2) \rangle_{\mathcal{H}} = K(x_1, x_2)$. The reader is referred to Aronszajn (1950) for more details on reproducing kernel Hilbert spaces. However, it is important to note that this particular form of $\mathcal{H}$ although very useful in several kernel methods in the kernel literature has no bearing on our problem, and in the rest of the paper we only assume $\mathcal{H}$ to be a separable Hilbert space.

Similar to the classical PCA, the kernel PCA can be represented as an eigen-decomposition problem of a covariance matrix; in this context a covariance operator. To define a covariance matrix in this context one needs to first consider the bilinear form $b : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ defined by

$$b(f, g) = \operatorname{cov}[f(\boldsymbol{X}), g(\boldsymbol{X})].$$

Suppose that b is bounded. Then there is a bounded and self-adjoint linear operator $\boldsymbol{\Sigma} : \mathcal{H} \to \mathcal{H}$ such that

$$b(f, g) = \langle f, \boldsymbol{\Sigma} g \rangle_{\mathcal{H}} = \langle \boldsymbol{\Sigma} f, g \rangle_{\mathcal{H}}.$$

See, for example, Conway (1990, page 31). This operator $\boldsymbol{\Sigma}$ is called the covariance operator of $\boldsymbol{X}$ (Baker, 1973; Fukumizu, Bach, and Jordan, 2004, 2009). Under the

assumption that Σ is a compact operator, it has a discrete spectral decomposition

$$\sum_{i=1}^{\infty} \lambda_i P_i,$$

where $\lambda_1 > \lambda_2 > \cdots$ are real numbers, and P_i is the projection on to the linear subspace

$$\ker(\boldsymbol{\Sigma} - \lambda_i) = \{f \in \mathcal{H} : \Sigma f = \lambda_i f\}.$$

These projections are orthogonal to each other; that is, $P_i P_j = 0$ whenever $i \neq j$. It can be shown that any function in $\ker(\boldsymbol{\Sigma} - \lambda_i)$ is the ith kernel principal component defined in the last paragraph.

The central question pursued in this Chapter can be formulated at three different levels. The first level is the fully nonparametric mean regression model

$$Y = f(\boldsymbol{X}) + \varepsilon, \tag{5.3}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is arbitrary and $\varepsilon \perp\!\!\!\perp \boldsymbol{X}$. Given a randomly selected regression function f and a randomly selected covariance operator $\boldsymbol{\Sigma}$ for $\boldsymbol{X}$, would the kernel PCA enjoy the similar predictive tendency as possessed by the classical PCA in the context of linear regression?

The second level is the most general. Suppose Y and $\boldsymbol{X}$ are dependent but the dependence is not restricted by any model, parametric or nonparametric. Then, given a randomly selected conditional distribution of $Y|\boldsymbol{X}$, and a randomly selected covariance operator $\boldsymbol{\Sigma}$ for $\boldsymbol{X}$, would the kernel PCA possess the similar predictive power? This question, imposes virtually no assumptions on the form of the relationship between the response variable Y and the predictors $\boldsymbol{X}$

The third level is an extension of Artemiou and Li (2009) and Ni (2009) into sufficient dimension reduction context. Suppose that Y and $\boldsymbol{X}$ are conditionally independent given $\beta^{\mathsf{T}}X$; that is,

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}.$$

Then, given randomly selected $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, would the linear PCA possess the similar predictive power?

Although the conclusion at the second level is the most general, the results at the other two levels are not technically the special cases of those at the second level. This is because each of the three questions expressed above require its own conditions and assumptions which are different in each case. the non-parametric case is the one that requires the most of the techniques that we need to develop all the results so we solve it first and then we solve the rest of the problems.

The predictive potential of the conventional PCA in the context of linear regression (Artemiou and Li, 2009) and the results by Ni (2010) needs the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ to be exchangeable and independent and for the eigenvectors to be also orthogonal. Similar developments are needed in this case as well and they are described in the sections following the next one, where motivation examples are presented.

5.3 Motivating examples

Before introducing the population version of kernel principal component and explaining exactly what we are trying to show in more theoretical detail we present some real example analysis that show that there is some predictive potential in the

first few kernel principal components, which tend to be more correlated with the response than the last few kernel principal components tend to.

Remember, that what we try to do is a probabilistic phenomenon, that exists in nature in collections of datasets and is not necessarily true for each individual dataset separately. That is, in a single dataset this phenomenon might not be true, but if there is a collection of datasets, then there is a tendency that he first few kernel principal components are more correlated with the response. Based on this, we have three databases of datasets where we explore if this phenomenon tends to be true. From each database, we select data sets according to three pre-specified criteria which are the same as the ones used in Artemiou and Li (2009); (i) They have univariate responses; when a data set has multivariate responses we randomly select one of them. (ii) They have no categorical predictors. (iii) They are not artificially constructed. The first database is that provided in the *Arc* software (http://www.stat.umn.edu/arc/software.html), from which we select 33 data sets according to the above criteria. This database is also used in Artemiou and Li (2009), but here kernel principal component analysis rather than linear principal component analysis is applied. The second database consists of data sets from a multivariate analysis textbook by Johnson and Wichern, (2007), from which we select 53 data sets. The third database is the CMU StatLib database (http://lib.stat.cmu.edu/index.php), from which we select 54 data sets.

For each data set in a database, we compute the first 5 kernel principal components and their sample correlations with the response. To compute the kernel principal components, we use the centered Gram matrix described in Fukumizu, Bach, and Jordan (2009) with the Gaussian, the exponential, the laplacian, the sigmoid and the second degree polynomial kernel. The parameter σ for the Gaussian, the exponential, the laplacian kernel is determined adaptively for each data set, as

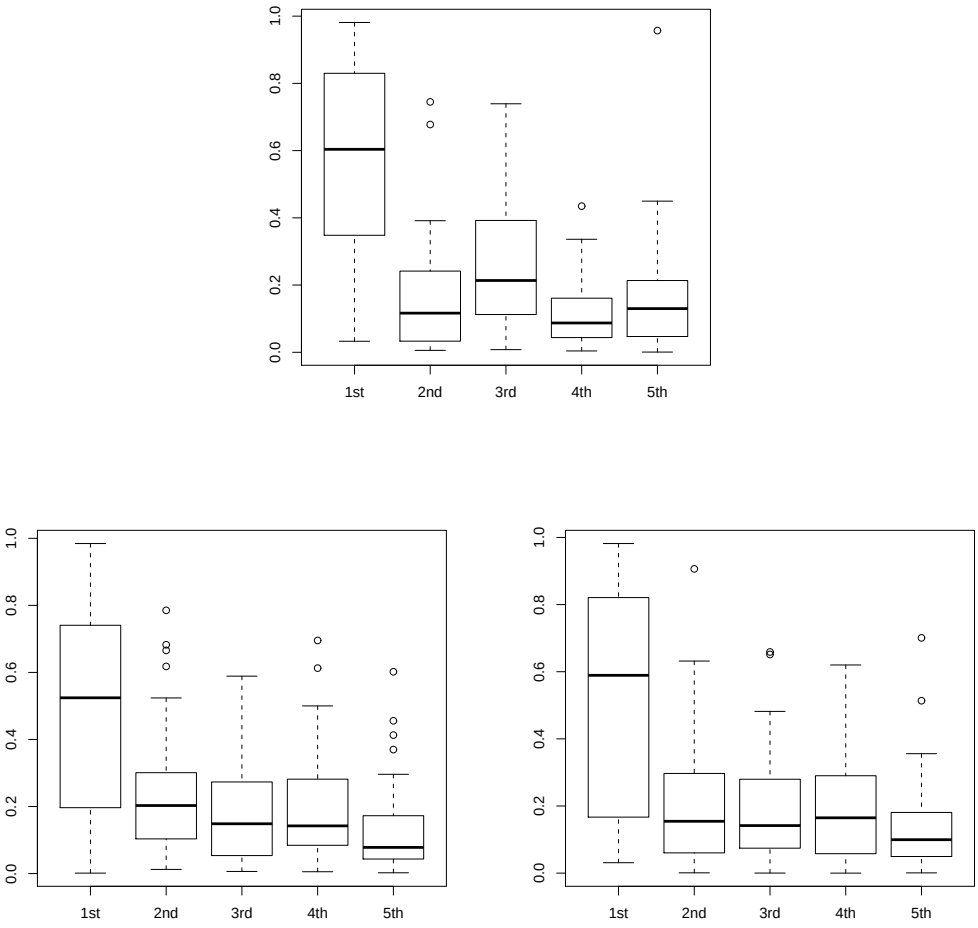

Figure 5.1: Boxplots for the absolute correlations between the response and the first 5 kernel principal components of the predictors in three databases using Gaussian kernel. Upper panel: 33 data sets from the *Arc* database. Lower-left panel: 53 data sets from Johnson and Wichern (2007). Lower-right panel: 54 data sets from CMU StatLib database.

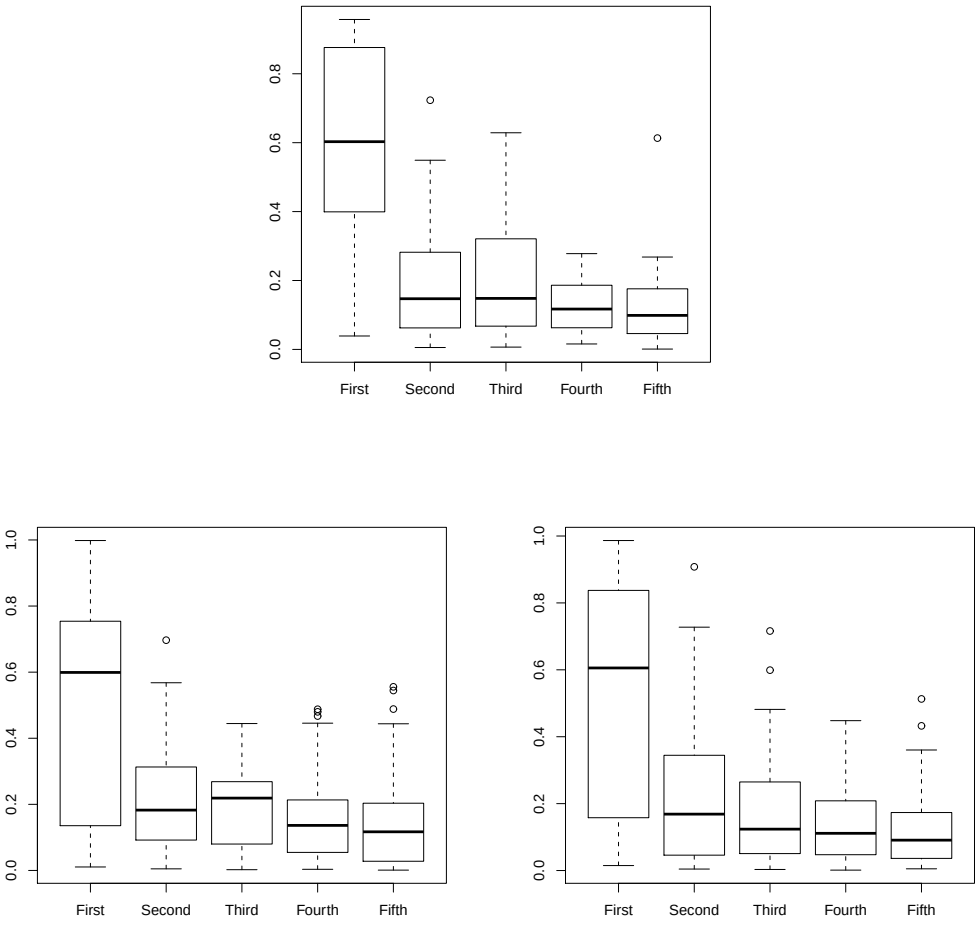

Figure 5.2: Boxplots for the absolute correlations between the response and the first 5 kernel principal components of the predictors in three databases using exponential kernel. Upper panel: 33 data sets from the *Arc* database. Lower-left panel: 53 data sets from Johnson and Wichern (2007). Lower-right panel: 54 data sets from CMU StatLib database.

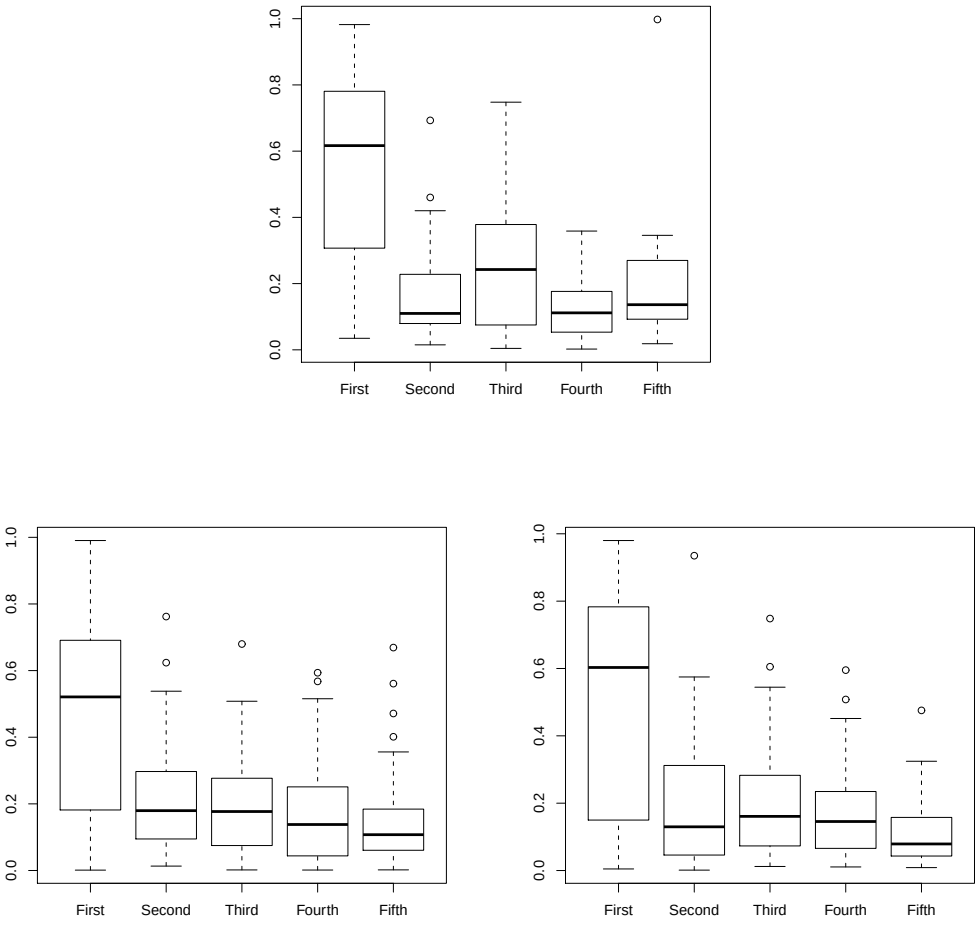

Figure 5.3: Boxplots for the absolute correlations between the response and the first 5 kernel principal components of the predictors in three databases using laplace kernel. Upper panel: 33 data sets from the *Arc* database. Lower-left panel: 53 data sets from Johnson and Wichern (2007). Lower-right panel: 54 data sets from CMU StatLib database.

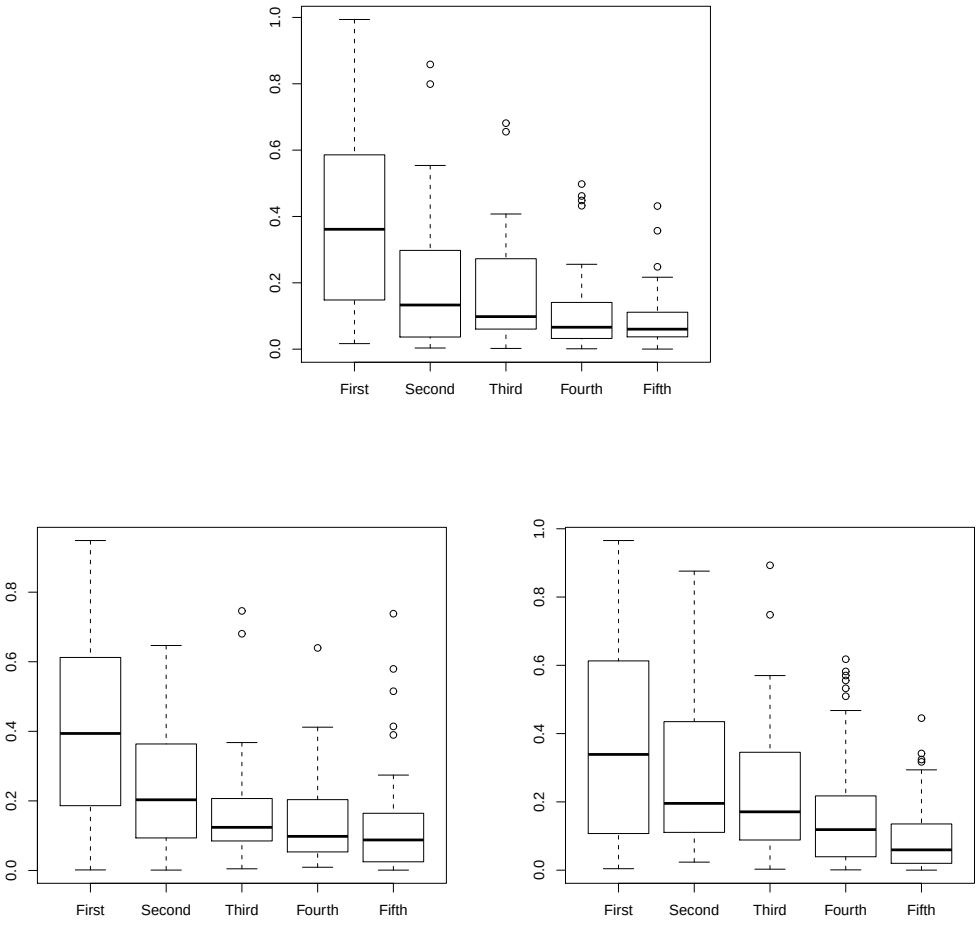

Figure 5.4: Boxplots for the absolute correlations between the response and the first 5 kernel principal components of the predictors in three databases using sigmoid kernel. Upper panel: 33 data sets from the *Arc* database. Lower-left panel: 53 data sets from Johnson and Wichern (2007). Lower-right panel: 54 data sets from CMU StatLib database.

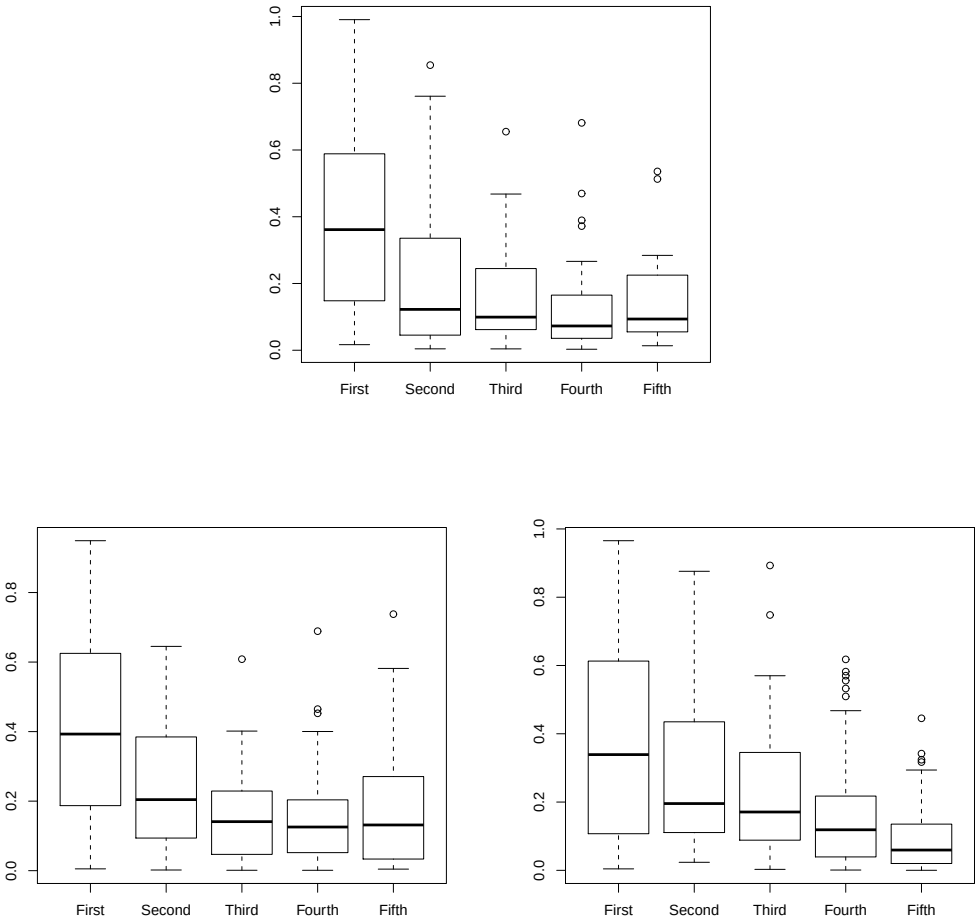

Figure 5.5: Boxplots for the absolute correlations between the response and the first 5 kernel principal components of the predictors in three databases using second order polynomial kernel with offset equal to 1. Upper panel: 33 data sets from the *Arc* database. Lower-left panel: 53 data sets from Johnson and Wichern (2007). Lower-right panel: 54 data sets from CMU StatLib database.

follows. Let $X_1, \ldots, X_n$ represent the observed predictors of a data set. We use the average of the Euclidean distances of $\{\|X_i - X_j\| : i, j = 1, \ldots, n, i < j\}$ as the value of σ. For the second degree polynomial kernel the offset parameter is is set equal to 1 and the scale parameter for the sigmoid and polynomial kernel is $(\|X_i\|\|X_j\|)^{-1}$. The absolute values of the sample correlation between each kernel principal component and the response is then calculated. Thus, for example, the *Arc* database yields 5 groups of correlations each having 33 correlations corresponding to a kernel principal component. The boxplots for the absolute correlations corresponding to the first 5 kernel principal components in the 3 databases are presented in Figure 5.1 for the Gaussian kernel, in Figure 5.2 for the exponential kernel, in Figure 5.3 for the Laplace kernel, in Figure 5.4 for the Sigmoid kernel and in Figure 5.5 for the Polynomial kernel .

It is evident from the boxplots in Figures 5.1 through 5.5 that higher-ranking kernel principal components tend to have stronger correlations with the response. In particular, in all three bases and with any type of kernel the first kernel principal components have considerably stronger correlations with the response than the other kernel principal components. Another point to note is the probabilistic nature of the tendency. For example, in each panel in all Figures 5.1 through 5.5 , there is a fair percentage of small correlations even for the first kernel principal component.

5.4 Unitarily invariant functions and operators

As in Artemiou and Li (2009) the covariance matrix $\mathbf{\Sigma}$ was defined to be orientationally uniform, a definition that describes the arbitrary orientation of the cloud of the points nd was used in earlier Chapters of this work. Similarly, in this Chapter we need to define the concept of an arbitrary covariance operator on a Hilbert

space, that is, $\mathbf{\Sigma} : \mathcal{H} \to \mathcal{H}$. Also, Artemiou and Li (2009) assumed random regression coefficients, that defined an arbitrary linear relationship. Since we will be talking about the general nonparametric model $Y = f(\boldsymbol{X}) + \epsilon$, we need to define what an arbitrary function $f \in \mathcal{H}$ is. Both definitions will be based on the definition of unitary invariance.

5.4.1 Arbitrary function in a Hilbert space

For simplicity, we assume that, like $\boldsymbol{X}$ and Y, all the other random elements are defined on the probability space $(\Omega, \mathfrak{F}, P)$. Let $\mathfrak{G}$ be the σ-field of Borel sets in $\mathcal{H}$. An $\mathcal{H}$-valued random element is a mapping $f : \Omega \to \mathcal{H}$ that is measurable $\mathfrak{F}/\mathfrak{G}$. A unitary operator $U : \mathcal{H} \to \mathcal{H}$ is an invertible linear operator such that $U^{-1} = U^*$, where $U^* : \mathcal{H} \to \mathcal{H}$ is the adjoint operator of U. This means that for a unitary operator the following reltionship s true,

$$\langle g, U(h)\rangle = \langle U^{-1}(g), h\rangle = \langle U^*(g), h\rangle$$

where g and h functions in $\mathcal{H}$. Since $\mathcal{H}$ is separable it has a countable orthonormal basis, say $\{u_i : i \in \mathbb{N}\}$. The sequence of Fourier coefficients, or Fourier sequence, of an element $f \in \mathcal{H}$ with respect to an orthonormal basis $\{u_i : i \in \mathbb{N}\}$ of $\mathcal{H}$ is the sequence $\{\langle f, u_i\rangle_{\mathcal{H}} : i \in \mathbb{N}\}$. For simplicity we will abbreviate sequences such as $\{a_i : i \in \mathbb{N}\}$ by $\{a_i\}$.

Any $f \in \mathcal{H}$ can be expressed as $\sum_{i\in\mathbb{N}} a_i u_i$, where $\sum_{i\in\mathbb{N}} a_i^2 < \infty$. Intuitively, an arbitrary function in $\mathcal{H}$ should have equal probability of assigning any coefficient sequence $\{a_i\}$ to the basis $\{u_i\}$, as long as $\sum_{i\in\mathbb{N}} a_i^2 < \infty$. Furthermore, since we will be concerned with quantities such as $\text{corr}(Y, u(\boldsymbol{X})|f)$, and not f itself, the

magnitude of f is irrelevant; what matters is the relative weights that f gives to each u_i. Finally, we want a definition of a random function f to be independent of any basis of the Hilbert space, although we have considered arbitrariness of f in term of a basis.

Definition 5.4.1 *An $\mathcal{H}$-valued random element f is said to be unitarily invariant if, for any unitary operator $U : \mathcal{H} \to \mathcal{H}$ we have $f \stackrel{\mathcal{D}}{=} U(f)$.*

An example of a unitarily invariant random function in $\mathcal{H}$, consider the standard Gaussian random function $f \in \mathcal{H}$, defined by the characteristic function

$$\int e^{i\langle g,f\rangle_{\mathcal{H}}} \nu(df) = e^{-\frac{1}{2}\langle g,g\rangle_{\mathcal{H}}},$$

where $i = \sqrt{-1}$ and $\nu = P \circ f^{-1}$ is the probability measure on $\mathcal{H}$ induced by f (Kannan and Bharucha-Reid (1970)). For any unitary operator U, the characteristic function of Uf is

$$\int e^{i\langle g,Uf\rangle_{\mathcal{H}}} \nu(df) = \int e^{i\langle U^{-1}g,f\rangle_{\mathcal{H}}} \nu(df) = e^{-\frac{1}{2}\langle U^{-1}g,U^{-1}g\rangle_{\mathcal{H}}} = e^{-\frac{1}{2}\langle g,g\rangle_{\mathcal{H}}}.$$

Hence f is unitarily invariant. In fact, one can show that any random element whose characteristic function depends on $g \in \mathcal{H}$ only through $\|g\|_{\mathcal{H}}$ is unitarily invariant. In this sense, a unitarily invariant random function in $\mathcal{H}$ is a generalization of spherically distributed random vector.

To understand why a unitarily invariant function can be regarded as "arbitrary", let ℓ^∞ be the Hilbert space of the sequences $\{c_i\}$ satisfying $\sum_{i\in\mathbb{N}} c_i^2 < \infty$, with respect to the inner product $\langle\{c_i\},\{d_i\}\rangle_{\ell^\infty} = \sum_{i\in\mathbb{N}} c_i d_i$. Then, for any member h of $\mathcal{H}$, its Fourier sequence $\{\langle h, u_i\rangle_{\mathcal{H}}\}$ belongs to ℓ^∞; conversely, for any member

$\{c_i\}$ of ℓ^∞, the function $\sum_{i\in\mathbb{N}} c_i u_i$ belongs $\mathcal{H}$.

Proposition 5.4.1 *If f is a unitarily invariant random function in $\mathcal{H}$, then its Fourier sequence with respect to any orthonormal basis $\{u_i\}$ is a unitarily invariant random element in ℓ^∞.*

PROOF. Let $T : \ell^\infty \to \ell^\infty$ be a unitary operator, and let h be a member of $\mathcal{H}$. Then $h = \sum_{i\in\mathbb{N}} c_i u_i$ for some $\{c_i\} \in \ell^\infty$. Let $T_j(\{c_i\})$ denote the jth entry of the sequence $T(\{c_i\})$. Let $U : \mathcal{H} \to \mathcal{H}$ be defined as

$$U(h) = \sum_{j\in\mathbb{N}} T_j(\{\langle h, u_i\rangle_{\mathcal{H}}\})u_j.$$

It is easy to see that U is invertible and

$$U^{-1}(h) = \sum_{j\in\mathbb{N}} T_j^{-1}(\{\langle h, u_i\rangle_{\mathcal{H}}\})u_j.$$

For any $g \in \mathcal{H}$,

$$\begin{aligned}
\langle g, U(h)\rangle_{\mathcal{H}} =& \sum_{j\in\mathbb{N}} T_j(\{\langle h, u_i\rangle_{\mathcal{H}}\})\langle g, u_j\rangle_{\mathcal{H}} \\
=& \langle T(\{\langle h, u_i\rangle_{\mathcal{H}}\}), \{\langle g, u_i\rangle_{\mathcal{H}}\}\rangle_{\ell^\infty} \\
=& \langle (\{\langle h, u_i\rangle_{\mathcal{H}}\}), T^{-1}\{\langle g, u_i\rangle_{\mathcal{H}}\}\rangle_{\ell^\infty} \\
=& \sum_{j\in\mathbb{N}} \{\langle h, u_j\rangle_{\mathcal{H}}\} T_j^{-1}(\langle g, u_i\rangle_{\mathcal{H}}) = \langle U^{-1}(g), h\rangle_{\mathcal{H}}.
\end{aligned}$$

Thus $U : \mathcal{H} \to \mathcal{H}$ is unitary. Since f is unitarily invariant, we have $f \overset{\mathcal{D}}{=} U(f)$.

Hence, for any $i \in \mathbb{N}$,

$$\langle f, u_i\rangle_{\mathcal{H}} \overset{\mathcal{D}}{=} \langle U(f), u_i\rangle_{\mathcal{H}} = \sum_{j\in\mathbb{N}} T_j(\{\langle f, u_k\rangle_{\mathcal{H}}\})\langle u_j, u_i\rangle_{\mathcal{H}} = \sum_{j\in\mathbb{N}} T_j(\{\langle f, u_k\rangle_{\mathcal{H}}\})\delta_{ij}$$
$$= T_i(\{\langle f, u_k\rangle_{\mathcal{H}}\}),$$

where $(i, j) \mapsto \delta_{ij}$ is the Kronecker δ function. The above equality implies that the random sequence $\{\langle f, u_i\rangle_{\mathcal{H}}\}$ is a unitarily invariant element in ℓ^∞. □

This proposition tells us that a unitarily invariant random function f has equal probability to assign any weights $\{c_i\}$ to $\{u_i\}$, so long as $\|\{c_i\}\|_{\ell^\infty}$ remains constant. Thus, taking into consideration what we said earlier that the norm of f is irrelevant for our discussion, a unitarily invariant random element is fully arbitrary. It is important to note that f is defined without reference to any basis of $\mathcal{H}$.

5.4.2 Arbitrary covariance operator $\boldsymbol{\Sigma}$

Now, we define an arbitrary covariance operator $\boldsymbol{\Sigma} : \mathcal{H} \to \mathcal{H}$. As mentioned earlier, the $\boldsymbol{\Sigma}$ in Artemiou and Li (2009) is defined as a random matrix which has orientationally uniform distribution (see Definition 1.6.2). Here, we give a technically different definition of an arbitrary operator $\boldsymbol{\Sigma}$. The advantage of this modification is that it is more compact and intuitively appealing, and it is easier to work with in an infinite dimensional setting.

Let $\mathfrak{R}$ be the σ-field of Borel sets in $\mathbb{R}$ and $\mathcal{L}(\mathcal{H})$ be the space of linear operators on $\mathcal{H}$. A random linear operator A is a mapping from Ω to $\mathcal{L}(\mathcal{H})$ such that, for any $f_1, f_2 \in \mathcal{H}$, the function $\omega \mapsto \langle A(\omega)f_1, f_2\rangle_{\mathcal{H}}$ is measurable with respect to $\mathfrak{F}/\mathfrak{R}$. For more on random linear operators the reader is referred to Skorohod (1976, 1984).

We define a random covariance operator $\mathbf{\Sigma} : \mathcal{H} \to \mathcal{H}$ as a bounded and self-adjoint random linear operator such that, for any $f_1, f_2 \in \mathcal{H}$, $\langle \mathbf{\Sigma}(\omega) f_1, f_2 \rangle_{\mathcal{H}} \geq 0$ almost surely P.

Definition 5.4.2 *A random covariance operator $\mathbf{\Sigma} : \mathcal{H} \to \mathcal{H}$ is said to be unitarily invariant if, for any unitary operator $U : \mathcal{H} \to \mathcal{H}$, we have $\mathbf{\Sigma} \stackrel{\mathcal{D}}{=} U\mathbf{\Sigma}U^{-1}$.*

Intuitively, this definition means that the operator $\mathbf{\Sigma}$, observed from any orthonormal system in $\mathcal{H}$, is the same random object. A consequence is that the ith kernel principal component of the operator $\mathbf{\Sigma}$ is equally likely to be any function in $\mathcal{H}$ of unit length. This is a fitting description that the distribution of $\boldsymbol{X}$ is chosen without regard to any response variable Y. This assumption is neither stronger nor weaker than the orientationally uniform assumption in Artemiou and Li (2009): notice that we do not require that the eigenvalues and eigenfunctions of $\mathbf{\Sigma}$ to be independent. In a finite dimensional setting, the above definition implies that $\mathbf{\Sigma} \stackrel{\mathcal{D}}{=} U\mathbf{\Sigma}U^{-1}$ for any orthogonal matrix U. Deift (1999, page 21) gave a more detailed description of this type of random matrices in the context of unitary ensembles.

Besides unitary invariance, we impose two additional technical assumption on $\mathbf{\Sigma}$ when it is treated as a unitarily invariant operator. The first one is that $\mathbf{\Sigma}$ is compact with probability 1 which ensures that is has a countable spectral decomposition. The second is that with probability 1, each nonzero eigenvalue of $\mathbf{\Sigma}$ has multiplicity 1 which ensures that nonzero eigenvalues and the corresponding eigenvectors are uniquely determined by $\mathbf{\Sigma}$. This is assumed for simplicity although we believe it can be avoided by a more elaborate analysis that will not be presented in this work.

Now that we have defined all the necessary tools to ensure randomness in a Hilbert space of the covariance operator $\mathbf{\Sigma}$ and a random function f we will try to answer in the next three sections the three problems we stated earlier. First we will see if similar results in a nonparametric setting where $Y = f(\boldsymbol{X}) + \epsilon$. Then we will see what happens if we select just a random measure for the relationship of Y on $\boldsymbol{X}$. At last we will develop the results in the supervised dimension reduction setting.

5.5 Predictive potential of Kernel PCA in nonparametric regression

In this section we tackle the first problem stated in Section 5.2. We first derive the distribution of the ratio of two Fourier coefficients of a unitarily invariant random function. A special case of this result in the finite-dimensional setting is given in Ni (2010). Lemma 5.5.1, basically, replaces the uniqueness of median lemma that was shown in Artemiou and Li (2009) and the results that were presented in earlier Chapters here. The Lemma in Artemiou and Li (2009) gave exactly the necessary conditions to satisfy a certain inequality. Lemma 5.5.1 along with the Lemma in Ni (2010) gives us the standard Cauchy distribution which is necessary in order to find an exact equality of the probability that a higher ranked principal component will have stronger correlation with the response than a lower order principal component would.

Lemma 5.5.1 *If f is a unitarily invariant random function in $\mathcal{H}$, then the ratio between two Fourier coefficients of f has a standard Cauchy distribution.*

PROOF. Let $\{u_i\}$ be any orthonormal basis of $\mathcal{H}$ and let A be a 2×2 orthogonal matrix. Let $k < \ell$ be two integers in $\mathbb{N}$. We will show that $\langle f, u_k\rangle_{\mathcal{H}}/\langle f, u_\ell\rangle_{\mathcal{H}}$ has a standard Cauchy distribution. Define $T : \ell^\infty \to \ell^\infty$ as the operator that maps $\{c_i\}$ to $\{d_i\}$, where $d_i = c_i$ if $i \notin \{k, \ell\}$ and $(d_k, d_\ell)^\intercal = A(c_k, c_\ell)^\intercal$. Then it is easy to verify that T is unitary.

From the proof of Proposition 5.4.1, T induces a unitary operator U on $\mathcal{H}$: $U(f) = \sum_{i\in\mathbb{N}} T_i(\{\langle f, u_m\rangle_{\mathcal{H}}\})u_i$. Since f is unitarily invariant, we have $U(f) \stackrel{\mathcal{D}}{=} f$. That is,

$$\sum_{i\in\mathbb{N}} T_i(\{\langle f, u_m\rangle_{\mathcal{H}}\})u_i \stackrel{\mathcal{D}}{=} \sum_{i\in\mathbb{N}} \langle f, u_i\rangle_{\mathcal{H}} u_i.$$

Hence

$$(T_k(\{\langle f, u_m\rangle_{\mathcal{H}}\}), T_\ell(\{\langle f, u_m\rangle_{\mathcal{H}}\}))^\intercal \stackrel{\mathcal{D}}{=} (\langle f, u_k\rangle_{\mathcal{H}}, \langle f, u_\ell\rangle_{\mathcal{H}})^\intercal,$$

which is equivalent to $A(\langle f, u_k\rangle_{\mathcal{H}}, \langle f, u_\ell\rangle_{\mathcal{H}})^\intercal \stackrel{\mathcal{D}}{=} (\langle f, u_k\rangle_{\mathcal{H}}, \langle f, u_\ell\rangle_{\mathcal{H}})^\intercal$. In other words, $(\langle f, u_k\rangle_{\mathcal{H}}, \langle f, u_\ell\rangle_{\mathcal{H}})^\intercal$ has a spherically contoured distribution. The desired result follows now from Theorem 1 of Arnold and Brockett (1992). □

The next theorem assumes f to be a unitarily invariant random function and the covariance operator $\mathbf{\Sigma}$ to be fixed.

Theorem 5.5.1 *Suppose that $\mathbf{\Sigma}$ is a compact operator. Let $\lambda_1 > \lambda_2 > \cdots$ be the distinct eigenvalues of $\mathbf{\Sigma}$. For each i, let u_i be any member of the linear subspace*

$\ker(\boldsymbol{\Sigma} - \lambda_i)$. *Suppose the nonparametric regression model*

$$Y = f(\boldsymbol{X}) + \varepsilon, \tag{5.4}$$

holds, where f is a unitarily invariant random element in $\mathcal{H}$ and $f \perp\!\!\!\perp \boldsymbol{X}$. Moreover, suppose $\epsilon \perp\!\!\!\perp (\boldsymbol{X}, f)$, $E(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \tau^2 < \infty$. Then, whenever $i < j$ and $\lambda_j > 0$, we have

$$P\left\{\mathrm{corr}^2[Y, u_i(\boldsymbol{X})|f] \geq \mathrm{corr}^2[Y, u_j(\boldsymbol{X})|f]\right\} = (2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}].$$

Proof. Because $\epsilon \perp\!\!\!\perp (\boldsymbol{X}, f)$,

$$\mathrm{cov}[Y, u_i(\boldsymbol{X})|f] = \mathrm{cov}[f(\boldsymbol{X}), u_i(\boldsymbol{X})|f].$$

Moreover, because $f \perp\!\!\!\perp \boldsymbol{X}$, for any fixed $f_0 \in \mathcal{H}$ we have

$$\mathrm{cov}[f(\boldsymbol{X}), u_i(\boldsymbol{X})|f = f_0] = \mathrm{cov}[f_0(\boldsymbol{X}), u_i(\boldsymbol{X})] = \langle f_0, \boldsymbol{\Sigma}\, u_i\rangle_{\mathcal{H}} = \lambda_i\langle f_0, u_i\rangle_{\mathcal{H}}. \tag{5.5}$$

That (5.5) holds for any fixed $f_0 \in \mathcal{H}$ implies

$$\mathrm{cov}[f(\boldsymbol{X}), u_i(\boldsymbol{X})|f] = \lambda_i\langle f, u_i\rangle_{\mathcal{H}}. \tag{5.6}$$

In the meantime, by $f \perp\!\!\!\perp \boldsymbol{X}$ again,

$$\mathrm{var}[u_i(\boldsymbol{X})|f] = \mathrm{var}[u_i(\boldsymbol{X})] = \langle u_i, \boldsymbol{\Sigma}\, u_i\rangle_{\mathcal{H}} = \lambda_i. \tag{5.7}$$

From (5.6) and (5.7) we see that

$$\text{corr}^2[Y, u_i(\boldsymbol{X})|f] = \lambda_i \langle f, u_i \rangle_{\mathcal{H}}^2 / \text{var}(Y|f),$$

which implies

$$\frac{\text{corr}^2[Y, u_i(\boldsymbol{X})|f]}{\text{corr}^2[Y, u_j(\boldsymbol{X})|f]} = \frac{\lambda_i}{\lambda_j} \frac{\langle f, u_i \rangle_{\mathcal{H}}^2}{\langle f, u_j \rangle_{\mathcal{H}}^2}. \tag{5.8}$$

Hence

$$\begin{aligned} &P\left\{\text{corr}^2[Y, u_i(\boldsymbol{X})|f] > \text{corr}^2[Y, u_j(\boldsymbol{X})|f]\right\} \\ &= P\left\{\frac{\langle f, u_i \rangle_{\mathcal{H}}^2}{\langle f, u_j \rangle_{\mathcal{H}}^2} > \frac{\lambda_j}{\lambda_i}\right\} = P\left\{-(\lambda_i/\lambda_j)^{\frac{1}{2}} < \frac{\langle f, u_j \rangle_{\mathcal{H}}}{\langle f, u_i \rangle_{\mathcal{H}}} < (\lambda_i/\lambda_j)^{\frac{1}{2}}\right\}. \end{aligned} \tag{5.9}$$

Since, by Lemma 5.5.1, the ratio $\langle f, u_j \rangle_{\mathcal{H}} / \langle f, u_i \rangle_{\mathcal{H}}$ has a standard Cauchy distribution, the right hand side of (5.9) is $(2/\pi) \arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]$. □

The interpretation of this theorem is that if nature chooses an arbitrary function f for the nonparametric regression model (5.4), then the correlation between Y and u_i tends to be larger than the correlation between Y and u_j in $(2/\pi) \arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}] \times 100$ percent times. We now extend this result to the situation where $\boldsymbol{\Sigma}$ is also random.

Corollary 5.5.1 *Suppose that model (5.4) holds, where X is a random vector whose covariance operator is $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}$ is a random covariance operator satisfying Assumption* **??**. *Suppose f is a random element in $\mathcal{H}$, and $\varepsilon \perp\!\!\!\perp X|(f, \boldsymbol{\Sigma})$, $(f, \boldsymbol{\Sigma}) \perp\!\!\!\perp \varepsilon$,*

$f \perp\!\!\!\perp (X, \boldsymbol{\Sigma})$. *Moreover, assume that* $E(\varepsilon) = 0$ *and* $\mathrm{var}(\varepsilon) = \tau^2 < \infty$. *Then*

$$P\left\{\mathrm{corr}^2[Y, u_i(\boldsymbol{X})|f, \boldsymbol{\Sigma}] \geq \mathrm{corr}^2[Y, u_j(\boldsymbol{X})|f, \boldsymbol{\Sigma}]\right\} = (2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\}.$$

Proof. We have

$$E(Y|f, \boldsymbol{\Sigma}) = E[f(\boldsymbol{X})|f, \boldsymbol{\Sigma}] + E(\varepsilon|f, \boldsymbol{\Sigma}).$$

Since $\varepsilon \perp\!\!\!\perp (f, \boldsymbol{\Sigma})$, the second term on the right hand side is 0. Moreover, $f \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$ implies that $f \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{\Sigma}$. Hence, for any $f_0 \in \mathcal{H}$, $E[f(\boldsymbol{X})|f = f_0, \boldsymbol{\Sigma}] = E[f_0(\boldsymbol{X})|\boldsymbol{\Sigma}]$. It follows that

$$\begin{aligned}\mathrm{cov}[Y, u_i(\boldsymbol{X})|\boldsymbol{\Sigma}, f = f_0] =& \mathrm{cov}[E(Y|\boldsymbol{\Sigma}, f = f_0), u_i(\boldsymbol{X})|\boldsymbol{\Sigma}, f = f_0] \\ =& \mathrm{cov}\{E[f_0(\boldsymbol{X})|\boldsymbol{\Sigma}], u_i(\boldsymbol{X})|\boldsymbol{\Sigma}\} = \langle f_0, \boldsymbol{\Sigma}\, u_i\rangle_{\mathcal{H}} = \lambda_i\langle f_0, u_i\rangle_{\mathcal{H}}.\end{aligned}$$

In other words,

$$\mathrm{cov}[Y, u_i(\boldsymbol{X})|\boldsymbol{\Sigma}, f] = \lambda_i\langle f, u_i\rangle_{\mathcal{H}}.$$

Similarly, since $f \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$,

$$\mathrm{var}[u_i(\boldsymbol{X})|f, \boldsymbol{\Sigma}] = \mathrm{var}[u_i(\boldsymbol{X})|\boldsymbol{\Sigma}] = \lambda_i.$$

Thus the situation is identical to Theorem 5.5.1 except now we have conditioned, everywhere, on $\boldsymbol{\Sigma}$. Apply Theorem 5.5.1 to the conditional probability $P(\cdot|\boldsymbol{\Sigma})$ to obtain

$$P\left\{\mathrm{corr}^2[Y, u_i(\boldsymbol{X})|f, \boldsymbol{\Sigma}] \geq \mathrm{corr}^2[Y, u_j(\boldsymbol{X})|f, \boldsymbol{\Sigma}]|\boldsymbol{\Sigma}\right\} = (2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}].$$

Now take expectation on both sides the equality to complete the proof. □

5.6 Predictive potential of Kernel PCA in arbitrary $\boldsymbol{X} - Y$ relation

In this section we will explore the general situation where $\boldsymbol{X}$ and Y are dependent but the dependence is not restricted to a certain model, parametric or nonparametric. This requires a different set of conditions from those assumed in Theorem 5.5.1. Instead of assuming f is unitarily invariant, as we did in Section 5.5, we assume $\boldsymbol{\Sigma}$ to be a unitarily invariant random operator. This set of conditions is similar to the conditions assumed in Artemiou and Li (2009) in the linear regression setting. It is important to note that although the result in this section is more general result than Theorem 5.5.1 we need to make clear that Theorem 5.5.1 is not a special case of the theorems we will prove in this section, as the assumption are different.

Although we assume no model for the relation between $\boldsymbol{X}$ and Y, we do need the following conditional independence

$$Y \perp\!\!\!\perp \boldsymbol{\Sigma} | \boldsymbol{X}.$$

That is, Y depends on $\boldsymbol{X}$ only through the value of $\boldsymbol{X}$ itself, and not its covariance operator. This is a very mild assumption. As an example, consider the following scenario:

$$Y = g(\boldsymbol{X}, \varepsilon),$$

where g is an unknown function and $\epsilon \perp\!\!\!\perp \boldsymbol{X}$. In this case, conditioning on $\boldsymbol{X}$, the distribution of Y depends only on the distribution of ϵ and the value of $\boldsymbol{X}$; the operator $\boldsymbol{\Sigma}$ does not appear in the conditional distribution of $Y|\boldsymbol{X}$ except through $\boldsymbol{X}$. The nonparametric mean regression model (5.4) clearly satisfies this condition. Another example is

$$Y = \mu(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon, \quad \varepsilon \perp\!\!\!\perp \boldsymbol{X},$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are unknown functions.

As in the previous section we will prove a lemma that proves that a ratio of inner products between a fixed and a random function in the Hilbert space follows a standard Cauchy distribution. Although, it looks like the one in Section 5.5, one needs to be careful as the vector $(\langle f, u_1\rangle_{\mathcal{H}}, \langle f, u_2\rangle_{\mathcal{H}})^{\mathsf{T}}$ is not spherically contoured distributed in general and so we cannot use the result by Arnold and Brocket (1992) directly. The idea of the proof is to introduce an artificial random function $\tilde{f}$, and then condition on u_1, u_2, so that we "transfer" randomness from (u_1, u_2) to $\tilde{f}$. Then one can follow the same method as in Lemma 5.5.1 to prove the theorem.

Lemma 5.6.1 *Suppose that u_1, u_2 are random functions in $\mathcal{H}$ such that*

- $\langle u_1, u_2\rangle_{\mathcal{H}} = 0$, *and*
- *for any unitary operator U in $\mathcal{H}$, $(u_1, u_2) \stackrel{\mathcal{D}}{=} (U(u_1), U(u_2))$.*

Then, for any (nonrandom) function $f \in \mathcal{H}$, $f \neq 0$, the ratio $\langle f, u_1\rangle_{\mathcal{H}}/\langle f, u_2\rangle_{\mathcal{H}}$ has a standard Cauchy distribution.

PROOF. Since U^{-1} is also a unitary operator we have $(u_1, u_2) \stackrel{\mathcal{D}}{=} (U^{-1}(u_1), U^{-1}(u_2))$.

Consequently,

$$\begin{aligned}(\langle f, u_1\rangle_{\mathcal{H}}, \langle f, u_2\rangle_{\mathcal{H}}) &\stackrel{\mathcal{D}}{=} (\langle f, U^{-1}(u_1)\rangle_{\mathcal{H}}, \langle f, U^{-1}(u_2)\rangle_{\mathcal{H}}) \\ &= (\langle U(f), u_1\rangle_{\mathcal{H}}, \langle U(f), u_2\rangle_{\mathcal{H}})\,.\end{aligned}$$

Thus the distribution of $(\langle f, u_1\rangle_{\mathcal{H}}, \langle f, u_2\rangle_{\mathcal{H}})$ depends on f only through $\|f\|_{\mathcal{H}} \equiv a > 0$. Let $\tilde{f}$ be a random element in $\mathcal{H}$ that is independent of (u_1, u_2) and uniformly distributed on the sphere $\mathcal{S}(a) = \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = a\}$. Then, for any Borel subset A of $\mathbb{R}$, and any nonrandom function $f_0 \in \mathcal{S}(a)$,

$$P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A | \tilde{f} = f_0) = P(\langle f_0, u_1\rangle_{\mathcal{H}}/\langle f_0, u_2\rangle_{\mathcal{H}} \in A). \tag{5.10}$$

This implies

$$P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A | \tilde{f}) = P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A). \tag{5.11}$$

The right hand side can be rewritten as

$$E[P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A | u_1, u_2)].$$

Because $\tilde{f} \perp\!\!\!\perp (u_1, u_2)$, $\tilde{f}$ is unitarily invariant conditioning on (u_1, u_2). Moreover, $\langle u_1, u_2\rangle_{\mathcal{H}} = 0$. Then, by Lemma 5.5.1, conditioning on (u_1, u_2), the ratio $\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}}$ has a standard Cauchy distribution, regardless of the value of (u_1, u_2). But this means that the ratio $\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}}$ is independent of (u_1, u_2), and therefore has a standard Cauchy distribution unconditionally. Hence

$$P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A) = P_C(A),$$

where $P_C(A)$ is the probability of A under the standard Cauchy distribution. However, by equality (5.10) and (5.11) and the discussion preceding them, we have

$$\begin{aligned} P(\langle f, u_1\rangle_{\mathcal{H}}/\langle f, u_2\rangle_{\mathcal{H}} \in A) =& P(\langle f_0, u_1\rangle_{\mathcal{H}}/\langle f_0, u_2\rangle_{\mathcal{H}} \in A) \\ =& P(\langle \tilde{f}, u_1\rangle_{\mathcal{H}}/\langle \tilde{f}, u_2\rangle_{\mathcal{H}} \in A) = P_C(A). \end{aligned}$$

That is, $\langle f, u_1\rangle_{\mathcal{H}}/\langle f, u_2\rangle_{\mathcal{H}}$ has a standard Cauchy distribution. □

We now establish the main result of this section.

Theorem 5.6.1 *Suppose that $\boldsymbol{\Sigma}$ is a unitarily invariant variance operator that is compact with probability one and each nonzero eigenvalue has multiplicity 1 with probability 1. Suppose $Y \perp\!\!\!\perp \boldsymbol{\Sigma}|\boldsymbol{X}$. Let $g(Y)$ be any measurable function of Y such that the function $x \mapsto E[g(Y)|\boldsymbol{X} = x]$ belongs to $\mathcal{H}$. Then, for any two eigen-pairs (λ_i, u_i) and (λ_j, u_j) of $\boldsymbol{\Sigma}$ satisfying $i < j$ and*

$$\operatorname{cov}[g(Y), u_i(\boldsymbol{X})|\boldsymbol{\Sigma}] \neq 0, \ \operatorname{cov}[g(Y), u_j(\boldsymbol{X})|\boldsymbol{\Sigma}] \neq 0, \tag{5.12}$$

with probability 1, we have

$$P\left\{\operatorname{corr}^2[g(Y), u_i(\boldsymbol{X})|\boldsymbol{\Sigma}] \geq \operatorname{corr}^2[g(Y), u_j(\boldsymbol{X})|\boldsymbol{\Sigma}]\right\} = E\left\{(2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\right\}.$$

Proof. First we note that

$$\operatorname{cov}[g(Y), u_i(\boldsymbol{X})|\boldsymbol{\Sigma}] = \operatorname{cov}\{E[g(Y)|\boldsymbol{X}, \boldsymbol{\Sigma}], u_i(\boldsymbol{X})|\boldsymbol{\Sigma}\} = \operatorname{cov}\{E[g(Y)|\boldsymbol{X}], u_i(\boldsymbol{X})|\boldsymbol{\Sigma}\},$$

where the second equality follows from the conditional independence $Y \perp\!\!\!\perp \boldsymbol{\Sigma}|\boldsymbol{X}$.

Let $f(x) = E[g(Y)|\boldsymbol{X} = x]$. Note that condition (5.12) implies, with probability 1,

$$\lambda_i > 0,\ \lambda_j > 0,\ \langle f,\ u_i\rangle_{\mathcal{H}} \neq 0,\ \langle f, u_j\rangle_{\mathcal{H}} \neq 0.$$

Then, by the similar calculation to that which leads to (5.8) in the proof of Theorem 5.5.1, we have

$$P\left\{\text{corr}^2[f(\boldsymbol{X}), u_i(\boldsymbol{X})|\boldsymbol{\Sigma}] > \text{corr}^2[f(\boldsymbol{X}), u_j(\boldsymbol{X})|\boldsymbol{\Sigma}]\right\} = P\left(\frac{\langle f, u_j\rangle^2_{\mathcal{H}}}{\langle f, u_i\rangle^2_{\mathcal{H}}} < \frac{\lambda_i}{\lambda_j}\right). \tag{5.13}$$

By the assumption that nonzero eigenvalues have multiplicity 1 with probability 1, and ignoring a probability null set, $(\lambda_i, \lambda_j, u_i, u_j)$ is uniquely determined by $\boldsymbol{\Sigma}$. That is, $(\lambda_i, \lambda_j, u_i, u_j)$ is a function of $\boldsymbol{\Sigma}$. Write this function as $(\lambda_i(\boldsymbol{\Sigma}), \lambda_j(\boldsymbol{\Sigma}), u_i(\boldsymbol{\Sigma}), u_j(\boldsymbol{\Sigma}))$. By the unitary invariance of $\boldsymbol{\Sigma}$, we have $U\boldsymbol{\Sigma}U^{-1} \overset{\mathcal{D}}{=} \boldsymbol{\Sigma}$. Therefore,

$$\begin{aligned}(\lambda_i(U\boldsymbol{\Sigma}U^{-1}), \lambda_j(U\boldsymbol{\Sigma}U^{-1}), u_i(U\boldsymbol{\Sigma}U^{-1}), u_j(U\boldsymbol{\Sigma}U^{-1}))& \\ \overset{\mathcal{D}}{=} (\lambda_i(\boldsymbol{\Sigma}), \lambda_j(\boldsymbol{\Sigma}), u_i(\boldsymbol{\Sigma}), u_j(\boldsymbol{\Sigma})).&\end{aligned} \tag{5.14}$$

Because

$$\begin{aligned}&\lambda_i(U\boldsymbol{\Sigma}U^{-1}) = \lambda_i(\boldsymbol{\Sigma}),\ \lambda_j(U\boldsymbol{\Sigma}U^{-1}) = \lambda_j(\boldsymbol{\Sigma}),\\ &u_i(U\boldsymbol{\Sigma}U^{-1}) = U(u_i(\boldsymbol{\Sigma})),\quad u_j(U\boldsymbol{\Sigma}U^{-1}) = U(u_j(\boldsymbol{\Sigma})).\end{aligned}$$

equality (5.14) reduces to

$$\{\lambda_i(\boldsymbol{\Sigma}), \lambda_j(\boldsymbol{\Sigma}), U(u_i(\boldsymbol{\Sigma})), U(u_j(\boldsymbol{\Sigma}))\} \overset{\mathcal{D}}{=} \{\lambda_i(\boldsymbol{\Sigma}), \lambda_j(\boldsymbol{\Sigma}), u_i(\boldsymbol{\Sigma}), u_j(\boldsymbol{\Sigma})\}.$$

Now that the argument for all random elements is $\boldsymbol{\Sigma}$, we drop it and rewrite the

above as $\{\lambda_i, \lambda_j, U(u_i), U(u_j)\} \stackrel{\mathcal{D}}{=} (\lambda_i, \lambda_j, u_i, u_j)$. This implies

$$(u_i, u_j)|(\lambda_i, \lambda_j) \stackrel{\mathcal{D}}{=} [U(u_i), U(u_j)]|(\lambda_i, \lambda_j).$$

By Lemma 5.6.1, as applied to the conditional probability given (λ_i, λ_j), the conditional distribution of ratio $\langle f, u_j\rangle_{\mathcal{H}}/\langle f, u_i\rangle_{\mathcal{H}}|(\lambda_i, \lambda_j)$ has a standard Cauchy distribution. Hence

$$P\left(\frac{\langle f, u_j\rangle_{\mathcal{H}}^2}{\langle f, u_i\rangle_{\mathcal{H}}^2} < \frac{\lambda_i}{\lambda_j}\middle| \lambda_i, \lambda_j\right) = (2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]. \tag{5.15}$$

Now take the unconditional expectation on both sides to complete the proof. □

Thus, if nature selects an arbitrary covariance operator for $\boldsymbol{X}$, then, regardless of the form of dependence between $\boldsymbol{X}$ and Y, any measurable function $g(Y)$ tends to have a larger correlation (in absolute value) with u_i than with u_j. The relative frequency of this tendency is $(2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\} \times 100$ percent.

Next, we consider the situation that nature also chooses a relation between $\boldsymbol{X}$ and Y, in addition to choosing a covariance operator $\boldsymbol{\Sigma}$ for X. Because no specific model is given to the $\boldsymbol{X}$-Y relation, the randomness has to be imposed directly on the conditional distribution of $Y|\boldsymbol{X}$ itself, rather than on some aspect of it, such as the regression function f in model (5.4). For this reason we need to introduce the notion of a random conditional distribution of Y given $\boldsymbol{X}$.

Recall that a conditional distribution of $Y|\boldsymbol{X}$ is a mapping

$$\kappa : \mathfrak{R} \times \Omega_{\boldsymbol{X}} \to [0, 1]$$

such that (i) for each $\omega \in \Omega$, the function $A \mapsto \kappa(A, \boldsymbol{X}(\omega))$, $\mathfrak{R} \to [0,1]$ is a probability measure on $\mathfrak{R}$; (ii) for each $A \in \mathfrak{R}$, the function $\omega \mapsto \kappa(A, \boldsymbol{X}(\omega))$, $\Omega \to [0,1]$ is a version of the conditional probability $P(Y \in A|\boldsymbol{X})$. Let $\mathcal{K}$ be the collection of all such mappings κ. For simplicity, assume that $\mathcal{H}$ is rich enough to contain all bounded measurable functions of $\boldsymbol{X}$, so that, for each $\kappa \in \mathcal{K}$, and each $A \in \mathfrak{R}$, $\kappa(A, \cdot) \in \mathcal{H}$. Let $\mathfrak{R}^p$ denote the σ-field of Borel sets in $\mathbb{R}^p$. We define a random element in $\mathcal{K}$, or a random conditional distribution of $Y|\boldsymbol{X}$, to be a mapping

$$\nu : \Omega \to \mathcal{K}, \quad \omega \mapsto \nu_\omega(\cdot, \cdot),$$

such that, for each $A \in \mathfrak{R}$, the function $\Omega \to \mathcal{H}$, $\omega \mapsto \nu_\omega(A, \cdot)$ is measurable $\mathfrak{R}^p/\mathfrak{G}$. Note that, if $\mathcal{H}$ is a set of numbers rather than a set of functions, then our definition reduces to the classical definition of a random probability measure. See, for example, Kingman (1967). We use the notation $Y|(\boldsymbol{X}, \nu) \sim \nu$ to indicate that a ν is chosen from $\mathcal{K}$ to be the conditional distribution of $Y|\boldsymbol{X}$.

If, for each $A \in \mathfrak{R}$, $\kappa(A, \boldsymbol{X})$ is almost surely constant, then κ represents the conditional distribution under which $\boldsymbol{X}$ and Y are independent. Let $\mathcal{K}_0$ be the collection of all such κ. Since the tendency described in this paper occurs only when $\boldsymbol{X}$ and Y are related in a way, we obviously would like to exclude independence from consideration. In the present context this is formulated as $P(\nu \in \mathcal{K}_0) = 0$. This assumption is reasonable. For example, consider the simple case where $\boldsymbol{X}$ and Y are standard normal variables. Then the dependence of $\boldsymbol{X}$ and Y is completely determined by their correlation ρ. If we assume ρ to have a continuous distribution, then the probability for Y and $\boldsymbol{X}$ to be independent is 0.

Corollary 5.6.1 *Suppose that covariance operator* $\boldsymbol{\Sigma}$ *of* $\boldsymbol{X}$ *is unitarily invariant,*

compact with probability 1 and each nonzero eigenvalue has multiplicity 1 with probability 1. Suppose that ν is a random element of $\mathcal{K}$ such that $P(\nu \in \mathcal{K}_0) = 0$ and

$$Y|(\boldsymbol{X}, \nu) \sim \nu, \quad \nu \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma}), \quad Y \perp\!\!\!\perp \boldsymbol{\Sigma}|(\boldsymbol{X}, \nu). \tag{5.16}$$

Let g be any measurable function of Y such that the random function $m_\nu(\cdot) = \int g\, \nu(d\omega, \cdot)$ belongs to $\mathcal{H}$ almost surely and, with probability 1,

$$\text{cov}[g(Y), u_i(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}] \neq 0, \ \text{cov}[g(Y), u_j(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}] \neq 0.$$

Then, for any $i < j$ we have

$$P\{\text{corr}^2[g(Y), u_i(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}] \geq \text{corr}^2[g(Y), u_j(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}]\} \\ = (2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\}.$$

The independence and conditional independence in (5.16) have the similar interpretation as those in Corollary 5.5.1: $Y \perp\!\!\!\perp \boldsymbol{\Sigma}|(\boldsymbol{X}, \nu)$ means that the distribution of $Y|(\boldsymbol{X}, \nu)$ does not depend on $\boldsymbol{\Sigma}$; $\nu \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$ means that the relation between $\boldsymbol{X}$ and Y does not depend on $\boldsymbol{X}$ or its covariance operator $\boldsymbol{\Sigma}$.

Proof of Corollary 5.6.1. Note that

$$\text{cov}[g(Y), u_i(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}] = \text{cov}\{E[g(Y)|\nu, \boldsymbol{\Sigma}, \boldsymbol{X}], u_i(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}\}$$

Since $Y \perp\!\!\!\perp \boldsymbol{\Sigma}|(\boldsymbol{X}, \nu)$, we have

$$E[g(Y)|\nu, \boldsymbol{\Sigma}, \boldsymbol{X}] = E[g(Y)|\nu, \boldsymbol{X}] = m_\nu(\boldsymbol{X}).$$

Since $\nu \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$ we have $m_\nu \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$. Hence, for any $\kappa \in \mathcal{K}$, we have

$$\mathrm{cov}[m_\nu(\boldsymbol{X}), u_i(\boldsymbol{X})|\nu = \kappa, \boldsymbol{\Sigma}] = \mathrm{cov}[m_\kappa(\boldsymbol{X}), u_i(\boldsymbol{X})|\boldsymbol{\Sigma}] = \langle m_\kappa, \boldsymbol{\Sigma} u_i\rangle_{\mathcal{H}} = \lambda_i \langle m_\kappa, u_i\rangle_{\mathcal{H}}.$$

This implies

$$\mathrm{cov}[m_\nu(X), u_i(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}] = \lambda_i \langle m_\nu, u_i\rangle_{\mathcal{H}}.$$

Similarly, by $\nu \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$ we have

$$\mathrm{var}[u_i(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}] = \mathrm{var}[u_i(\boldsymbol{X})|\boldsymbol{\Sigma}] = \lambda_i.$$

It follows that

$$\frac{\mathrm{corr}^2[g(Y), u_i(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}]}{\mathrm{corr}^2[g(Y), u_j(\boldsymbol{X})|\nu, \boldsymbol{\Sigma}]} = \frac{\lambda_i \langle m_\nu, u_i\rangle_{\mathcal{H}}}{\lambda_j \langle m_\nu, u_i\rangle_{\mathcal{H}}}.$$

Since $m_\nu \perp\!\!\!\perp (u_i, u_j, \lambda_i, \lambda_j)$, we have $m_\nu \perp\!\!\!\perp (u_i, u_j)|(\lambda_i, \lambda_j)$. Hence, for any $\kappa \in \mathcal{K}$,

$$P\left(\frac{\langle m_\nu, u_j\rangle^2_{\mathcal{H}}}{\langle m_\nu, u_i\rangle^2_{\mathcal{H}}} < \frac{\lambda_i}{\lambda_j}\middle| \nu = \kappa, \lambda_i, \lambda_j\right) = P\left(\frac{\langle m_\kappa, u_j\rangle^2_{\mathcal{H}}}{\langle m_\kappa, u_i\rangle^2_{\mathcal{H}}} < \frac{\lambda_i}{\lambda_j}\middle| \lambda_i, \lambda_j\right).$$

By (5.15) the right hand side is $(2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]$. Thus we have proved

$$P\left(\frac{\langle m_\nu, u_j\rangle^2_{\mathcal{H}}}{\langle m_\nu, u_i\rangle^2_{\mathcal{H}}} < \frac{\lambda_i}{\lambda_j}\middle| \nu, \lambda_i, \lambda_j\right) = (2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}].$$

Now take the conditional expectation on both sides of the above equality to complete the proof. □

5.6.1 Data analysis

To test how our theory holds up in real data sets we now compare the estimated values of

$$\Pi_{ij} = (2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\}, \quad P_{ij} = P\{\text{corr}^2(Y, u_i|\nu, \Sigma) \geq \text{corr}^2(Y, u_j|\nu, \Sigma)\}$$

for each of the three databases described in Section 5.3. According to Corollary 5.6.1 these two values should be the same. The Π_{ij} and P_{ij} are estimated as follows. Let $D_1, \ldots, D_m$ represent the data sets in each database. Thus $m = 33, 53, 54$ for the three databases, respectively. For each D_k, we compute the ith eigenvalues of the centered Gram matrix from derived from the Gaussian kernel, Exponential kernel, Laplace kernel, Sigmoid kernel and second order polynomial kernel with offset equal to 1 (see description in Section **??**). Denote these eigenvalues as $\hat{\lambda}_{ik}$. The value Π_{ij} is then estimated by

$$\hat{\Pi}_{ij} = \frac{2}{\pi m} \sum_{k=1}^{m} \arctan[(\hat{\lambda}_{ik}/\hat{\lambda}_{jk})^{\frac{1}{2}}].$$

The probability P_{ij} is estimated similarly. For each data set D_k, we compute the sample correlation between the ith kernel principal component and the response. Denote this correlation by $\hat{\rho}_{ik}$. Then P_{ij} is estimated by

$$\hat{P}_{ij} = \frac{1}{m} \sum_{k=1}^{m} I(\hat{\rho}_{ik}^2 \geq \hat{\rho}_{jk}^2).$$

The results are presented in Table 5.1 for the Gaussian kernel, Table 5.2 for the Exponential kernel, Table 5.3 for the Laplace kernel, Table 5.4 for the Sigmoid kernel and Table 5.5 for the second degree polynomial kernel.

Table 5.1: Comparison of $\hat{\Pi}_{ij}$ and $\hat{P}_{ij}$ for three databases using Gaussian kernel.

	Arc		J & W (2007)		CMU StatLib	
(i,j)	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$
1 vs 2	0.643	0.818	0.619	0.755	0.644	0.815
2 vs 3	0.678	0.303	0.640	0.547	0.657	0.426
3 vs 4	0.648	0.727	0.614	0.434	0.623	0.537
4 vs 5	0.644	0.485	0.606	0.642	0.614	0.630

Table 5.2: Comparison of $\hat{\Pi}_{ij}$ and $\hat{P}_{ij}$ for three databases using Exponential kernel.

	Arc		J & W (2007)		CMU StatLib	
(i,j)	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$
1 vs 2	0.669	0.818	0.645	0.735	0.793	0.796
2 vs 3	0.606	0.394	0.591	0.528	0.688	0.537
3 vs 4	0.589	0.545	0.570	0.547	0.641	0.500
4 vs 5	0.566	0.545	0.555	0.585	0.610	0.574

Table 5.3: Comparison of $\hat{\Pi}_{ij}$ and $\hat{P}_{ij}$ for three databases using Laplace kernel.

	Arc		J & W (2007)		CMU StatLib	
(i,j)	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$
1 vs 2	0.606	0.848	0.588	0.773	0.692	0.796
2 vs 3	0.592	0.364	0.581	0.509	0.669	0.444
3 vs 4	0.566	0.727	0.557	0.566	0.612	0.556
4 vs 5	0.562	0.364	0.541	0.566	0.593	0.685

Table 5.4: Comparison of $\hat{\Pi}_{ij}$ and $\hat{P}_{ij}$ for three databases using the Sigmoid kernel.

	Arc		J & W (2007)		CMU StatLib	
(i,j)	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$
1 vs 2	0.863	0.758	0.791	0.642	0.824	0.611
2 vs 3	0.785	0.606	0.821	0.585	0.817	0.481
3 vs 4	0.764	0.697	0.749	0.623	0.846	0.574
4 vs 5	0.747	0.485	0.668	0.528	0.751	0.667

Table 5.5: Comparison of $\hat{\Pi}_{ij}$ and $\hat{P}_{ij}$ for three databases using second degree polynomial kernel with offset equal to 1.

	Arc		J & W (2007)		CMU StatLib	
(i,j)	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$	$\hat{\Pi}_{ij}$	$\hat{P}_{ij}$
1 vs 2	0.825	0.758	0.771	0.660	0.644	0.815
2 vs 3	0.847	0.545	0.795	0.660	0.657	0.426
3 vs 4	0.752	0.667	0.795	0.528	0.623	0.537
4 vs 5	0.781	0.424	0.753	0.453	0.614	0.630

Tables 5.1 through 5.5 show reasonable agreements between $\hat{\Pi}_{ij}$ and $\hat{P}_{ij}$, at least in overall trends. It is interesting to see that $\hat{P}_{ij}$ seems to fluctuate more than $\hat{\Pi}_{ij}$ does, which is perhaps to be expected because, intuitively, Π_{ij} acts as a theoretical expectation of the relative predictive potentials of u_i and u_j based purely on the properties of the predictors themselves. It should also be noted that equality $P_{ij} = \Pi_{ij}$ is marginal in nature. That is, a pair of eigenfunctions u_i, u_j are considered without reference to the other eigenfunctions. Perhaps this explains why, in Tables 5.1 through 5.5 , a relatively good agreement is sometimes followed by a relatively poor agreement, and nonadjacent pairs seem to agree better. Within our current theoretical framework, we believe it is possible to compute probabilities such as

$$P\{\text{corr}^2(Y, u_i|\nu, \Sigma) \geq \text{corr}^2(Y, u_j|\nu, \Sigma) \geq \text{corr}^2(Y, u_k|\nu, \Sigma)\}$$

for $i < j < k$, and such joint probabilities might improve the agreement.

5.7 Linear PCA and sufficient dimension reduction

In this section, we try to connect linear principal component analysis with sufficient dimension reduction. In sufficient dimension reduction (see Li (1991, 1992), Cook

and Weisberg (1991), Cook (1994, 1998), and Li, Zha, and Chiaromonte (2005)) the $\boldsymbol{X}$-Y relation is specified by the conditional independence

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}, \tag{5.17}$$

where $\boldsymbol{\beta}$ is a matrix in $\mathbb{R}^{p \times d}$, $d \leq p$. Also in the second part of this work we propose a new method to estimate $\boldsymbol{\beta}$ based on support vector machine algorithms. In this part of this work we only consider the special case of $d = 1$. For $d > 1$ similar issues as the ones we had before for multivariate response variables appear. Let $\boldsymbol{\Sigma}$ denote the covariance matrix of $\boldsymbol{X}$. Although relation (5.17) is a special case of the $\boldsymbol{X}$-Y relation in Section 5.6, which postulates no model at all, the conditions used in this section are different from the previous sections. Here, we assume $\boldsymbol{\beta}$ or $\boldsymbol{\Sigma}$, or both, to be randomly selected by nature.

Recall that a p-dimensional random vector has a spherical distribution if, for any $p \times p$ orthogonal matrix A, $AV \stackrel{\mathcal{D}}{=} V$. As a special case of Definition 5.4.2, we say that $p \times p$ random covariance matrix $\boldsymbol{\Sigma}$ is unitarily invariant if $A\boldsymbol{\Sigma}A^{-1} \stackrel{\mathcal{D}}{=} \boldsymbol{\Sigma}$. The next lemma summarizes the special cases of Lemmas 5.5.1 and 5.6.1 in a finite-dimensional setting.

Lemma 5.7.1 *Suppose that v_1, v_2 are nonrandom vectors in $\mathbb{R}^p$ and u_1, u_2 are random vectors in $\mathbb{R}^p$, that $v_1^{\mathsf{T}} v_2 = 0$ and $u_1^{\mathsf{T}} u_2 = 0$, and that, for any orthogonal matrix A, $A(u_1, u_2) \stackrel{\mathcal{D}}{=} (u_1, u_2)$. Then the ratios*

$$v_1^{\mathsf{T}} u_1 / v_2^{\mathsf{T}} u_1, \quad v_1^{\mathsf{T}} u_1 / v_1^{\mathsf{T}} u_2$$

each follows a standard Cauchy distributions.

Let (λ_1, v_1), ..., (λ_p, v_p) be the eigen-pairs of $\mathbf{\Sigma}$, so ordered that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $U_1, \ldots, U_p$ be the 1st, ..., pth principal components of X. That is, $U_i = v_i^{\mathsf{T}} \boldsymbol{X}$. Let A be a p-row matrix. In the following $P_A(\mathbf{\Sigma})$ denotes the projection on to span(A) with respect to the $\mathbf{\Sigma}$-inner product. That is, $P_A(\mathbf{\Sigma}) = A(A^{\mathsf{T}}\mathbf{\Sigma}A)^{-1}A^{\mathsf{T}}\mathbf{\Sigma}$. Let $Q_A(\mathbf{\Sigma}) = I_p - P_A(\mathbf{\Sigma})$ be the projection onto the orthogonal complement of span(A). In the next lemma, $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ are assumed nonrandom.

In the following lemma we will need the assumption $E(X|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X})$ is linear in $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$. This is commonly used in the sufficient dimension reduction literature. See, for example, Li (1991) and Cook (1998). It implies that

$$E(\boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}) = P_{\boldsymbol{\beta}^{\mathsf{T}}}(\mathbf{\Sigma})\boldsymbol{X} + Q_{\boldsymbol{\beta}^{\mathsf{T}}}(\mathbf{\Sigma})E(\boldsymbol{X}). \tag{5.18}$$

Lemma 5.7.2 *Suppose that the conditional independence (5.17) holds for some $\boldsymbol{\beta} \in \mathbb{R}^p$, and $E(X|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X})$ is a linear function of $\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$, then for any $i \neq j$, and any measurable function $g(Y)$ with finite variance and $\operatorname{cov}[g(Y), \boldsymbol{X}] \neq 0$, we have*

$$\frac{\operatorname{corr}^2[g(Y), U_i]}{\operatorname{corr}^2[g(Y), U_j]} = \frac{\lambda_i(v_i^{\mathsf{T}}\boldsymbol{\beta})^2}{\lambda_j(v_j^{\mathsf{T}}\boldsymbol{\beta})^2}. \tag{5.19}$$

Proof. Note that

$$\operatorname{cov}[g(Y), U_i] = \operatorname{cov}\{E[g(Y)|\boldsymbol{X}], U_i\} = \operatorname{cov}\{E[g(Y)|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}], U_i\}.$$

Since conditional expectation is a self-adjoint operator, we also have

$$\operatorname{cov}\{E[g(Y)|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}], U_i\} = \operatorname{cov}[g(Y), E(U_i|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X})]. \tag{5.20}$$

By (5.18),

$$E(U_i|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}) = v_i^\mathsf{T}E(\boldsymbol{X}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}) = v_i^\mathsf{T}P_{\boldsymbol{\beta}^\mathsf{T}}(\boldsymbol{\Sigma})\boldsymbol{X} + v_i^\mathsf{T}Q_{\boldsymbol{\beta}^\mathsf{T}}(\boldsymbol{\Sigma})E(\boldsymbol{X}).$$

Substitute this into the right hand side of (5.20) to obtain

$$\mathrm{cov}[g(Y), U_i] = v_i^\mathsf{T}P_{\boldsymbol{\beta}^\mathsf{T}}(\boldsymbol{\Sigma})\mathrm{cov}[g(Y), \boldsymbol{X}] = \lambda_i v_i^\mathsf{T}\boldsymbol{\beta}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\mathsf{T}\mathrm{cov}[g(Y), \boldsymbol{X}].$$

In the meantime we note that $\mathrm{var}(U_i) = \lambda_i$. Hence

$$\mathrm{corr}^2(Y, U_i) = \lambda_i\{v_i^\mathsf{T}\boldsymbol{\beta}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\mathsf{T}\mathrm{cov}[g(Y), \boldsymbol{X}]\}^2/\mathrm{var}(Y).$$

Now take the ratio of $\mathrm{corr}^2(Y, U_i)$ and $\mathrm{corr}^2(Y, U_j)$, which implicitly evokes the assumption that $\mathrm{cov}[g(Y), \boldsymbol{X}] \neq 0$, to complete the proof. □

The key point of this lemma is that $g(Y)$ disappears from the ratio on the left hand side of (5.19), so that the ratio is completely determined by the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$. This is what gives linear PCA its predictive potential.

Theorem 5.7.1 *Suppose $\boldsymbol{\Sigma}$ is a random matrix in $\mathbb{R}^{p\times p}$ and $\boldsymbol{\beta}$ is a random vector in $\mathbb{R}^p$ such that $Y \perp\!\!\!\perp \boldsymbol{\Sigma}|(\boldsymbol{X}, \boldsymbol{\beta})$, $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$. Suppose, furthermore,*

$$Y \perp\!\!\!\perp \boldsymbol{X}|(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \tag{5.21}$$

and $E(\boldsymbol{X}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ is linear in $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$. Suppose $g(Y)$ is any measurable function such that

$$P\{E[g^2(Y)|\boldsymbol{\beta}, \boldsymbol{\Sigma}] < \infty\} = 1, \quad P\{\mathrm{cov}[g(Y), \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}] \neq 0\} = 1. \tag{5.22}$$

If either of the following conditions holds:

1. $\boldsymbol{\Sigma}$ *is a unitarily invariant operator with each of its eigenvalues having multiplicity 1 almost surely,*

2. $\boldsymbol{\beta}$ *spherically distributed,*

then, for any $i < j$*,*

$$P\left\{\text{corr}^2[g(Y), U_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}] \geq \text{corr}^2[g(Y), U_j|\boldsymbol{\beta}, \boldsymbol{\Sigma}]\right\} = (2/\pi)E\{\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]\}.$$

Condition (5.21) is the same in spirit as the dimension reduction relation (5.17), except that we have taken into account the randomness of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Condition (5.22) ensures that $\text{cov}[g(Y), \boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}]$ exists and is nonzero almost surely.

Proof of Theorem 5.7.1. Conditioning on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, the situation is identical to Lemma 5.7.2. Hence

$$\frac{\text{corr}^2[g(Y), U_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}]}{\text{corr}^2[g(Y), U_j|\boldsymbol{\beta}, \boldsymbol{\Sigma}]} = \frac{\lambda_i(v_i^\intercal\boldsymbol{\beta})^2}{\lambda_j(v_j^\intercal\boldsymbol{\beta})^2}.$$

If assumption 1 is satisfied, then we first condition on $\boldsymbol{\Sigma}$ to compute

$$P\left(\frac{(v_j^\intercal\boldsymbol{\beta})^2}{(v_i^\intercal\boldsymbol{\beta})^2} < \frac{\lambda_i}{\lambda_j}\middle|\boldsymbol{\Sigma}\right).$$

Since $\boldsymbol{\beta}$ is independent of $\boldsymbol{\Sigma}$, it has a spherical distribution conditioning on $\boldsymbol{\Sigma}$. Hence, by Lemma 5.7.1 (the result for the first ratio), this conditional probability is $(2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]$. Now take unconditional expectation to prove the desired equality under scenario 1.

If assumption 2 is satisfied, then we first condition on $(\boldsymbol{\beta}, \lambda_i, \lambda_j)$ to compute

$$P\left(\frac{(v_j^\intercal \boldsymbol{\beta})^2}{(v_i^\intercal \boldsymbol{\beta})^2} < \frac{\lambda_i}{\lambda_j} \middle| \boldsymbol{\beta}, \lambda_i, \lambda_j\right). \tag{5.23}$$

By the similar argument to that used in proving Theorem 5.6.1 we can show that, for any orthogonal matrix A, $A(v_i, v_j)|(\lambda_i, \lambda_j) \stackrel{\mathcal{D}}{=} (v_i, v_j)|(\lambda_i, \lambda_j)$. Since $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{\Sigma}$, this implies

$$A(v_i, v_j)|(\boldsymbol{\beta}, \lambda_i, \lambda_j) \stackrel{\mathcal{D}}{=} (v_i, v_j)|(\boldsymbol{\beta}, \lambda_i, \lambda_j)$$

Applying Lemma 5.7.1 (the result for the second ratio) to the conditional probability $P(\cdot|\boldsymbol{\beta}, \lambda_i, \lambda_j)$, we see that conditional probability (5.23) is $(2/\pi)\arctan[(\lambda_i/\lambda_j)^{\frac{1}{2}}]$. Now take the unconditional expectation to complete the proof. □

5.7.1 Central mean subspace for sufficient dimension reduction

Cook and Li (2002, 2004) introduced the notion of the central mean subspace for sufficient dimension reduction to deal with the situation where the conditional mean $E(Y|\boldsymbol{X})$, rather than the full conditional distribution of $Y|\boldsymbol{X}$, is of primary interest. See also Yin and Cook (2002). Suppose

$$E(Y|\boldsymbol{X}) = E(Y|\beta^\intercal \boldsymbol{X}), \tag{5.24}$$

for some matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$, $d \leq p$. Then the subspace of $\mathbb{R}^p$ spanned by the columns of $\boldsymbol{\beta}$ is a mean dimension reduction subspace. The intersection of all such subspaces is called the central mean subspace. A related concept is the single index model: $Y = f(\boldsymbol{\beta}^\intercal \boldsymbol{X}) + \varepsilon$, where $\varepsilon \perp\!\!\!\perp \boldsymbol{X}$ and $\boldsymbol{\beta}$ is a vector in $\mathbb{R}^p$ (Ichimura (1993)). It is

easy to see that if $\boldsymbol{\beta}$ satisfies this relation then the linear subspace spanned by $\boldsymbol{\beta}$ is a 1-dimensional central mean subspace.

Theorem 5.7.1 can be modified in an obvious way to cover this case. Assuming again $d = 1$, if we replace $g(Y)$ by Y, and replace condition (5.21) by

$$E(Y|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = E(Y|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

then the conclusion of Theorem 5.7.1 still holds.

Chapter 6

Discussion

In this first part of this work we study a phenomenon that in the context of linear regression and classical principal component analysis, this phenomenon has long been noticed, and was a focal point of a historical debate. This problem is even more important today because, in its most general form, it lies at the intersection of supervised, unsupervised, and semi-supervised dimension reductions, three rapidly advancing areas in statistics and machine learning.

This wok is a continuation of Li (2007), Artemiou and Li (2009) and Ni (2010). In the first Chapter we presented this past work and a small piece of the historical debate on the issue. In Chapter 2, we present how one can extend the results presented in Artemiou and Li (2009) for fixed covariance matrix (a similar result was simultaneously developed by Ni (2010)) and in the case of multivariate response variable Y. In Chapter 3 we show our first attempt to go beyond the linear model. So under certain conditions we expand the result of Artemiou and Li (2009) and Ni (2010) in two important directions. The results are presented for both fixed and random covariance matrices as well as for univariate and multivariate response

variables. Those results show that linear PCA hold some predictive power even if the underline model is not linear. In Chapter 4 we present an effort to go around the correlation in finding the predictive power of linear PCA, by introducing an information criterion, and showing a result for normally distributed predictors in the linear model. We believe that there is some potential to that result that we need to explore in the future. In Chapter 5, the more interesting and convincing results are presented. The predictive power of linear PCA is shown in connection with sufficient dimension reduction under the assumption that $E(\boldsymbol{X}|\boldsymbol{\beta}^{\intercal}\boldsymbol{X})$ is linear in $\boldsymbol{\beta}^{\intercal}\boldsymbol{X}$. A similar result was shown in Chapter 3 under different assumptions. Also we investigate the predictive power of kernel PCA in infinite dimensional Hilbert spaces. We show that under the assumption of $\boldsymbol{\Sigma}$ being a unitarily invariant covariance operator (Section 5.6) and for a general arbitrary relationship for Y on $\boldsymbol{X}$, the higher ranked kernel principal components tend to have more correlation with the response than the lower ranked one. This is also shown in the case of the nonparametric setting where $Y = f(\boldsymbol{X}) + \epsilon$ in section 5.5 under the assumption of a unitarily invariant random function f.

We must emphasize that the tendency studied here is probabilistic. If we have only one dataset we do not expect it to be true. In the case that we have a collection of datasets we expect that this tendency will be clearer, as in the case with the three datasets presented in Figures 5.1 through 5.5. Those boxplots show the tendency that higher ranked kernel principal components tend to be more correlated with the response than lower ranked ones do, a tendency that can be quantified using Cauchy distribution as we have shown in the proofs of our theoretical results. Given the fact that this tendency is probabilistic, the results presented in this part of this work, does not prove that linear principal components or kernel principal components gives you always the correct results or that it is always good to use

them. They just show a tendency but there exists an unmeasurable risk that this relationship will not hold true for some datasets. So, these results present the need of other methods to be developed, that incorporate the information of Y when extracting directions for dimension reduction in regression as the extensive literature in sufficient dimension reduction suggests. The reader is referred to Li (1991), Li (1992), Cook (1994), (1996), (1998a), (1998b), Li, Zha and Chiaromonte (2005), Li and Wang (2007), as well as the next part of this work which deals with the proposal of another method to perform sufficient dimension reduction using machine learning algorithms.

6.1 Future work

The results developed in this work are somewhat incomplete, in the sense that there are many interesting results not developed to the full potential.

One possible extension that we are interesting to investigate more is the case where we have multivariate response Y. As we saw in the results presented in Chapters 2 and 3, there is a need to find the exact assumptions under which certain ratios between random variables have unique median that is equal to 1. Those cases where assumed to be true, but there is no indication if there are real life reasonable examples where those assumptions are true. Further investigation is needed towards that direction.

A second possible extension is the information criterion presented in Chapter 4. The information criterion, was an idea that we were trying to use to see if linear PCA have predictive power beyond the linear regression setting. Unfortunately, the result we were able to develop and show in the course of this work was very

limited, as it is just for the case where we have a linear regression model with normal predictors, which of course is very restrictive. Much more work is needed towards this direction and we will try to develop something in the future.

Moreover, in the linear regression model (and even the nonparametric case presented in Section 5.5) one might think to relax the assumption of independence between the regression coefficients and the predictors (or in the case of the nonparametric case the independence between f and the predictors). This might result in different (and maybe better) lower bound in the inequality presented by Artemiou and Li (2009) or even another equality at the one presented in Ni (2010).

Furthermore, the last result that relates sufficient dimension reduction with linear principal component analysis, was proved only for $d = 1$. It will be interesting to extend it for cases where $d > 1$ as this is the case often in real datasets.

Finally it is very interesting to see, how strong is the assumption of unitary invariance. There is an indication that this might be very strong in the infinite dimensional Hilbert spaces, in which case our proof is only true for the finite dimensional Hilbert spaces, but further work is needed towards this direction.

PART II

SUFFICIENT DIMENSION REDUCTION USING SUPPORT VECTOR MACHINES

Chapter 7

Introduction on sufficient dimension reduction

Sufficient Dimension Reduction is a field developed mainly in the last two decades and especially after a breakthrough work of Li K. C. (1991). The main idea of Sufficient Dimension Reduction and the fact that makes it different and more effective procedure than PCA, is the involvement of the response variable Y in the information used to calculate the axis which Y has the most variation, and hence the axis that will show the most information for the regression of Y on $\boldsymbol{X}$.

7.1 General On Sufficient Dimension Reduction

In this section we give the main definitions that will be used in the description of the different Sufficient Dimension Reduction methods that were proposed over the years. For more details on the subject the reader is referred to Cook (1998a),

where we borrow most of the notation.

It is important to clearly state that in all those methods we assume that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}^T \boldsymbol{X}. \tag{7.1}$$

That means that the response variable Y depends on the predictor matrix $\boldsymbol{X}$ only through $\boldsymbol{\beta}^T \boldsymbol{X}$. This implies that there will be no loss of information in the regression if $\boldsymbol{X}$ is replaced by $\boldsymbol{\beta}^T \boldsymbol{X}$. In this case, $S(\boldsymbol{\beta})$ is considered the dimension reduction subspace (DRS) for the regression of Y on $\boldsymbol{X}$, where with $S(\boldsymbol{\beta})$ we define the subspace that is spanned by the column vectors of matrix $\boldsymbol{\beta}$.

One can define the minimum DRS to be the dimension reduction subspace S of the regression of Y on $\boldsymbol{X}$, if the $\dim(S) \leq \dim(S_{DRS})$ for all possible DRSs, which are denoted with S_{DRS}. With "dim" the dimension of a subspace is denoted. The minimum DRS of a regression may not be unique, but all the minimum DRS by definition should have the same dimension.

Furthermore, one can define the central dimension reduction subspace (CDRS) to be the subspace $S_{Y|\boldsymbol{X}}$ such that $S \subset S_{DRS}$ for all possible DRSs, which are denoted with S_{DRS}. A CDRS exists only if the $\cap S_{DRS}$ is itself a DRS and $S_{Y|\boldsymbol{X}} = \cap S_{DRS}$.

The relationship between minimum DRS and CDRS in a regression problem is given in the following proposition(Cook (1998a)):

Proposition 7.1.1 *If* $\mathrm{S}_{Y|\boldsymbol{X}}$ *is the CDRS for the regression of y on x then* $\mathrm{S}_{Y|\boldsymbol{X}}$ *is the unique minimum DRS.*

7.2 Sufficient Dimension Reduction Methods

In this section an overview of the main methods of Sufficient Dimension Reduction that have been presented over the years is given. The main idea behind every method is given and the advantages and their limitations are explained. The reader is referred to the referenced works for further reading.

7.2.1 Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) was the first method for reducing the dimension and it was proposed by Li K. C. and Duan (1989). The method is based on minimizing the loss function

$$L(\alpha, \boldsymbol{\beta}) = E(Y - \alpha - \boldsymbol{\beta}^T \boldsymbol{X})^2. \tag{7.2}$$

The major assumption of this work, which is also an assumption on most of the following methods is the assumption of linear conditional mean (LCM), that is:

Assumption 7.2.1 *For any $\boldsymbol{\beta} \in \mathbb{R}^p$, the conditional expectation*

$$E\left(\boldsymbol{\beta}\boldsymbol{X} | \beta_1 \boldsymbol{X}, \ldots, \beta_K \boldsymbol{X}\right) \tag{7.3}$$

is linear in $\beta_1 \boldsymbol{X}, \ldots, \beta_K \boldsymbol{X}$; that is, for constants $c_0, \ldots, c_K$,

$$E\left(\boldsymbol{\beta}\boldsymbol{X} | \beta_1 \boldsymbol{X}, \ldots, \beta_K \boldsymbol{X}\right) = c_0 + c_1 \beta_1 \boldsymbol{X} + \ldots + c_K \beta_K \boldsymbol{X} \tag{7.4}$$

It was shown by Eaton (1986) that Assumption 7.2.1 is satisfied if and only if the distribution of the predictor vector $\boldsymbol{X}$ is elliptically symmetric.

Denoting with $\boldsymbol{Z}$ the standardized version of the predictors $\boldsymbol{X}$ it is easy to show that the solution of the loss function

$$L(\alpha, \boldsymbol{\beta}) = E(Y - \alpha - \boldsymbol{\beta}^T \boldsymbol{Z})^2. \tag{7.5}$$

is $E(\boldsymbol{Z}Y)$.

Under Assumption 7.2.1 it can be shown that $E(\boldsymbol{Z}Y) \in S_{Y|\boldsymbol{Z}}$.

7.2.2 Sliced Inverse Regression (SIR)

The main problem with OLS was that it can detect only one direction. If there are two or more directions that are needed it will not detect those. To avoid this Li K. C.(1991) introduced the Sliced Inverse Regression (SIR) algorithm.

The author doesn't assume any parametric or nonparametric model fitting process. The dimension reduction is defined under the following regression model:

$$Y = f\left(\beta_1 \boldsymbol{X}, \ldots, \beta_K \boldsymbol{X}, \boldsymbol{\epsilon}\right) \tag{7.6}$$

which is more restrictive than the model in (7.1). The goal is to find as small K as possible. They define effective dimension reduction (EDR) space to be the linear space generated by the EDR directions where EDR directions are the estimated directions where $\boldsymbol{X}$ has the greatest variability. The EDR directions are denoted in the above model by $\beta_i, i = 1, ..., K$ and K is the dimension of the EDR space.

To estimate the EDR directions the author is using the inverse regression, that

is regressing $\boldsymbol{X}$ against Y instead of Y on $\boldsymbol{X}$. Using inverse regression, one can regress each predictor in $\boldsymbol{X}$ on Y, thus reducing the problem into multiple one dimensional problems.

The main theorems that set the theoretical formulation of the above are the following:

Theorem 7.2.1 *Under Assumption 7.2.1 and model (7.6) the centered inverse regression curve* $E(\boldsymbol{X}|Y) - E(\boldsymbol{X})$ *is contained in the linear subspace spanned by* $\beta_i \boldsymbol{\Sigma}$ *where* $i = 1, \ldots, K$ *and* $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X})$ *a* $p \times p$ *matrix.*

The above theorem, of course is true when we standardize $\boldsymbol{X}$ to $\boldsymbol{Z}$. That is:

Corollary 7.2.1 *Under Assumption 7.2.1 and model (7.6) the standardized regression curve* $E(\boldsymbol{Z}|Y)$ *is contained in the linear space generated by the standardized EDR directions* $\eta_1, \ldots, \eta_K$. *Consequently, the column space of* $\mathrm{cov}(E(\boldsymbol{Z}|Y))$ *is a subspace of the EDR subspace.*

One important step in SIR is the slicing of the domain of response variable Y into H slices. That means, that in fact when we perform SIR we are using the discretized version od the above theorems.

Another issue of the SIR algorithm is finding the number of directions that are necessary to completely describe the relationship between Y and $\boldsymbol{X}$. For $\boldsymbol{X}$ having a multivariate normal distribution, the solution is in the following theorem,

Theorem 7.2.2 *If* $\boldsymbol{X}$ *is normally distributed, let* $\lambda_1 \geq \ldots \geq \lambda_p$ *denote the eigenvalues of the matrix* $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X})$ *and* $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_p$ *their estimators . Under the hypothesis* $H_0 : \lambda_{p-K+1} = \ldots = \lambda_p = 0$ *the test statistic* $\sum_{i=p-K+1}^{p} \widehat{\lambda}_i$ *follows*

asymptotically χ^2 distribution with $(p-K)(H-K-1)$ degrees of freedom, where H is the number of slices we use in the SIR algorithm.

Using the above theorem one can perform a sequential test to for $K = 1, \ldots, p-1$ and the number of of EDR directions that are needed is the smallest K for which the hypothesis in Theorem 7.2.2 is rejected.

The problem with the above work is that the ideas of EDR space and EDR directions were proposed without addressing existence or uniqueness issues and as Cook (1998a) say, "they can be elusive".

After the appearance of the theoretical background on Dimension Reduction Subspace (DRS) and the Central Dimension Reduction Subspace (CDRS), by Cook (1994a), Cook (1994b), Cook (1996) the ideas were proven to work in that framework that was presented in section 7.1. The theorems in this section can be restated to fit that framework as follows:

Theorem 7.2.3 *Under Assumption 7.2.1 and model (7.1) the centered inverse regression curve* $E(\boldsymbol{X}|Y) - E(\boldsymbol{X}) \in \mathrm{S}_{Y|\boldsymbol{X}}$

The above theorem, of course is true when we standardize $\boldsymbol{X}$ to $\boldsymbol{Z}$. That is:

Corollary 7.2.2 *Under Assumption 7.2.1 and model (7.1) the standardized regression curve* $E(\boldsymbol{Z}|Y) \in \mathrm{S}_{Y|\boldsymbol{Z}}$. *Consequently, the column space of* $\mathrm{cov}(E(\boldsymbol{Z}|Y)) \in \mathrm{S}_{Y|\boldsymbol{Z}}$.

7.2.3 Sliced Average Variance Estimates (SAVE)

This method was presented by Cook and Weisberg (1991) in the discussion of the paper by Li K. C. (1991) presenting the SIR. The authors note that SIR can fail to estimate efficiently the CDRS directions in case $E(\boldsymbol{Z}|Y = y) = 0 \; \forall y$. So, by observing that sometimes the dependence can be through higher moments they proposed the Sliced Average Variance Estimates (SAVE) algorithm that instead of calculating $E(\boldsymbol{Z}|Y)$ in each slice they are calculating the $\text{var}(\boldsymbol{Z}|Y)$. The directions needed to estimate the CDRS is the eigenvectors corresponding to the largest eigenvalues of the matrix $\sum_{h=1}^{H} (I - \text{var}(\boldsymbol{Z}|y \in L_h))^2$ where H denotes the number of slices of the domain of Y and L_h denotes the h^{th} slice.

There is an extra assumption associated with the use of SAVE algorithm, which is known as the Constant Conditional Variance Assumption.

Assumption 7.2.2 *The conditional variance of* $\text{var}(\boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X})$ *is a non-ranom matrix.*

The theorem associated with the SAVE algorithm is again the discretized version of:

Theorem 7.2.4 *Suppose Assumptions 7.2.1 and 7.2.2 hold and* $\boldsymbol{Z}$ *the standardized version of the predictors* $\boldsymbol{X}$*. Then the column space of the matrix* $I - \text{var}(\boldsymbol{Z}|Y)$ *is a subset of* $\text{S}_{Y|\boldsymbol{Z}}$*, the central space. Consequently the column space of the matrix* $E\left(I - \text{var}(\boldsymbol{Z}|Y)\right)^2$ *is a subspace of* $\text{S}_{Y|\boldsymbol{Z}}$*.*

Originally the work was presented in the context of EDR space, but since it has been shown that can be extended into the more general case of CDRS, which was presented in section 7.1, only the more general case is shown.

7.2.4 Principal Hessian Directions (pHds)

The idea of finding the Principal Hessian Directions (pHds) to achieve dimension reduction was first proposed by Li K. C. (1992) and was further extended, generalized and refined by Cook (1998b). When Li K. C. (1992) first presented it, was under the EDR directions and EDR space formulation but here only the results that are using the more general concept of DRS and CDRS will be presented, since they are more general. Also the results are shown under the standardized version $\boldsymbol{Z}$ of the predictors $\boldsymbol{X}$, since they apply without loss of generality.

In mathematics the Hessian matrix H is the square matrix of second-order partial derivatives of a function. In the presented work the function is $E(Y|\boldsymbol{Z})$. We have that:

$$H(\boldsymbol{Z}) = \frac{\partial^2 E(Y|\boldsymbol{Z})}{\partial \boldsymbol{Z} \partial \boldsymbol{Z}^T} = \frac{\partial^2 E(Y|\boldsymbol{\beta}^T \boldsymbol{Z})}{\partial \boldsymbol{Z} \partial \boldsymbol{Z}^T} = \boldsymbol{\beta} \frac{\partial^2 E(Y|\boldsymbol{\beta}^T \boldsymbol{Z})}{\partial (\boldsymbol{\beta}^T \boldsymbol{Z}) \partial (\boldsymbol{Z}^T \boldsymbol{\beta})} \boldsymbol{\beta}^T \tag{7.7}$$

where $\boldsymbol{\beta}$ is the basis of the CDRS $S_{Y|\boldsymbol{Z}}$. The motivation for this work for Li K. C. (1992) was the fact that the Hessian matrix is degenerate along the directions that are orthogonal to the EDR space (and later it was shown that is also orthogonal to CDRS $S_{Y|\boldsymbol{Z}}$). Also, Li K.C. (1992) uses Stein's Lemma to effectively estimate the average Hessian matrix, $\overline{H}_{\boldsymbol{Z}} = E(H(\boldsymbol{Z}))$, and the pHds.

Before giving more results, it is important to note that $H_Y = E\left(Y \boldsymbol{Z} \boldsymbol{Z}^T\right)$ is called the Y-based Hessian matrix and $H_e = E\left(e \boldsymbol{Z} \boldsymbol{Z}^T\right)$ is called the e-based Hessian matrix, where e is the residual of the simple linear regression $e = Y - w^T \boldsymbol{Z}$. Also the eigenvectors of the Hessian matrix are called pHds.

The following theorems are very important because they show that if one can estimate the average Hessian matrix well, then the associated pHds with significant

nonzero eigenvalues can be used to find the basis for the CDRS $S_{Y|\boldsymbol{Z}}$.

Theorem 7.2.5 *Suppose Assumptions 7.2.1 and 7.2.2 holds. Then the column space of H_Y is a subspace of $\mathrm{S}_{Y|\boldsymbol{Z}}$.*

Theorem 7.2.6 *Suppose Assumptions 7.2.1 and 7.2.2 holds. Then the column space of H_e is a subspace of $\mathrm{S}_{Y|\boldsymbol{Z}}$.*

Let $\lambda_1 \geq \ldots \geq \lambda_p$ the eigenvalues of $H_e H_e^T$ and $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_p$ their sample estimates. In order to determine the number of directions K that we need to estimate we need to perform the sequence of tests

$$H_0 : \lambda_{j+1} = \ldots = \lambda_p = 0 j = 0, 1, \ldots, p-1 \tag{7.8}$$

The rank K of H_e which also denotes the direction needed is the smallest j for which the above hypothesis holds. The test statistic is the following:

$$\sum_{i=j+1}^{p} \frac{\widehat{\lambda}_i}{2\mathrm{var}(e)} \tag{7.9}$$

The following theorem is true for the hypothesis test procedure that is performed in order to find the number of directions K. (A series of results that prove this theorem can be found in Li B. (2003)).

Theorem 7.2.7 *Suppose that:*

1. *The column space of H_e exhausts the CDRS: that is, $span(H_e) = \mathrm{S}_{Y|\boldsymbol{Z}}$*
2. *The predictor vector $\boldsymbol{Z}$ has a p-dimensional standard multivariate distribution.*

Then under the hypothesis (7.8) the test statistic (7.9) converges to a χ^2 distribution with $j(j+1)/2$ degrees of freedom.

The results on this issue can be achieved with the use of matrix H_Y but as Li K. C. (1992) argues the results with H_e are more powerful since usually $\text{var}(Y) \geq \text{var}(e)$.

While the idea of pHd algorithm is very attractive it was shown that it can fail to detect linear trends.

7.2.5 Central Mean Subspace and Iterative Hessian Transformations (IHTs)

Iterative Hessian Transformation algorithm was proposed by Cook and Li (2002). In this work, the authors introduced the concept of Central Mean Subspace (CMS) which is a natural inferential object for dimension reduction when the mean function $E(Y|\boldsymbol{X})$ is of interest.

Development of Central Mean Subspace

The basic definitions and theorems that were used to develop the ideas of CMS are presented in this subsection.

Definition 7.2.1 *If*

$$Y \perp\!\!\!\perp E(Y|\boldsymbol{X})|\boldsymbol{\beta}^T\boldsymbol{X} \tag{7.10}$$

then $\text{S}(\boldsymbol{\beta})$ *is a mean dimension reduction subspace for the regression of* Y *on* $\boldsymbol{X}$.

Since model 7.1 implies model 7.10 it means that a dimension reduction subspace is also a mean dimension reduction subspace.

Parallel with the definition of Central Dimension Reduction Subspace given in section 7.1 one can define $S_{E(Y|\boldsymbol{X})} = \cap S$ where S is used to denote all mean dimension reduction subspaces. If $S_{E(Y|\boldsymbol{X})}$ exists then it is called the Central Mean Subspace (CMS). CMS does not always exist, but the existence and uniqueness of it can be guaranteed under similar conditions as the CDRS in section 7.1. Also, $S_{E(Y|\boldsymbol{X})} \subseteq S_{Y|\boldsymbol{X}}$. For location regressions where $Y \perp\!\!\!\perp \boldsymbol{X}|E(Y|\boldsymbol{X})$ we have that $S_{E(Y|\boldsymbol{X})} = S_{Y|\boldsymbol{X}}$.

Directions in the Central Mean Subspace

First of all, the authors revisit all the methods that have been presented (OLS, SIR, SAVE, pHd) to see under which conditions the estimated directions that were shown to be in $S_{Y|\boldsymbol{X}}$ using those methods, are also in $S_{E(Y|\boldsymbol{X})}$. For easiness, the results involving the standardized predictors $\boldsymbol{Z}$ are shown.

For the OLS, it is noted that if one restricts the attention to objective functions based on the natural exponential family

$$L\left(a + \boldsymbol{b}^T\boldsymbol{Z}, Y\right) = -Y\left(a + \boldsymbol{b}^T\boldsymbol{Z}\right) + \phi\left(a + \boldsymbol{b}^T\boldsymbol{Z}\right) \tag{7.11}$$

for some strictly convex function ϕ, then the directions $\boldsymbol{\beta}$ are always in $S_{E(Y|\boldsymbol{Z})}$, where $\boldsymbol{\beta}$ the population minimizers of

$$(\alpha, \boldsymbol{\beta}) = \arg\min_{a,\boldsymbol{b}} R\left(a, \boldsymbol{b}\right) \tag{7.12}$$

For SIR and SAVE, it is noted that both can find vectors in $S_{Y|\boldsymbol{Z}} \backslash S_{E(Y|\boldsymbol{Z})}$. That is, in general $E(\boldsymbol{Z}|Y)$ is in $S_{Y|\boldsymbol{Z}}$ but not in $S_{E(Y|\boldsymbol{Z})}$.

For pHd, it was shown that using H_Y for the estimation of the directions then the algorithm is actually estimating the directions in CMS.

Theorem 7.2.8 *Let $\boldsymbol{\gamma}$ be the basis for $\mathrm{S}_{E(Y|\boldsymbol{Z})}$. If Assumption 7.2.1 holds and if $\mathrm{var}(\boldsymbol{Z}|\boldsymbol{\gamma}^T\boldsymbol{Z})$ is uncorrelated with Y, then $\mathrm{S}(\boldsymbol{\beta}_{yz}, H_Y) \subseteq \mathrm{S}_{E(Y|\boldsymbol{Z})}$ where $\boldsymbol{\beta}_{yz}$ is the OLS coefficient vector $E(Y\boldsymbol{Z})$*

Instead of using the H_Y one can use H_e as was suggested by the derivation of pHd in Li K.C. (1992) and Cook (1998). So:

Proposition 7.2.1 *Assume that Assumption 7.2.1 holds. Then:*

$$\mathrm{S}_{E(Y|\boldsymbol{Z})} = \mathrm{S}_{E(e|\boldsymbol{Z})} + \mathrm{S}(\boldsymbol{\beta}_{yz}) \tag{7.13}$$

where the summation of the two subspaces means the collection of vectors of the form $\boldsymbol{\beta} + \boldsymbol{\beta}'$ with $\boldsymbol{\beta} \in \mathrm{S}_{E(r|\boldsymbol{Z})}$ and $\boldsymbol{\beta}' \in \mathrm{S}(\boldsymbol{\beta}_{yz})$

Finally the authors, show that under the assumptions

1. $E(\boldsymbol{Z}|\boldsymbol{\beta}_{yz}^T\boldsymbol{Z})$ is linear
2. $\mathrm{var}(\boldsymbol{Z}|\boldsymbol{\beta}_{yz}^T\boldsymbol{Z})$ is constant

the following equation holds

$$S(\boldsymbol{\beta}_{yz}, H_e) = S(\boldsymbol{\beta}_{yz}, H_Y) \tag{7.14}$$

Iterative Hessian Transformation

The results of the previous section for the already presented methods, can be summarized as follows:

1. SIR requires Assumption 7.2.1 and find vectors in CDRS
2. SAVE requires Assumptions 7.2.1 and 7.2.2 and find vectors in CDRS
3. pHd requires Assumptions 7.2.1 and 7.2.2 and finds vectors in CMS
4. OLS requires Assumption 7.2.1 and finds vectors in CMS

The problem is that while SIR, SAVE and pHd can find multiple directions in the space they are effective into finding directions (CDRS or CMS), OLS can find only one direction. Cook and Li (2002) propose a method which will require only Assumption 7.2.1 and will be able to find multiple directions in CMS.

First the following theorem is proven.

Theorem 7.2.9 *Under Assumption 7.2.1 the central mean space is an invariant subspace of the linear transformation* $\boldsymbol{v} \mapsto H\boldsymbol{v}$ *where* H *can be replaced with any of* H_Y *or* H_e*. That means:*

$$H_Y \mathrm{S}_{E(Y|\boldsymbol{Z})} \subset \mathrm{S}_{E(Y|\boldsymbol{Z})}, H_e \mathrm{S}_{E(Y|\boldsymbol{Z})} \subset \mathrm{S}_{E(Y|\boldsymbol{Z})} \tag{7.15}$$

This theorem basically says that one can find one vector in $S_{E(Y|\boldsymbol{Z})}$ and all the rest will be produced by multiplying with the Hessian matrices H_Y or H_e. Since the vectors $\boldsymbol{\beta}_{yz}$ as was defined in the OLS procedure is in $S_{E(Y|\boldsymbol{Z})}$ that brings the following result.

Corollary 7.2.3 *Under Assumption 7.2.1*

1. $span\{H_Y^j \boldsymbol{\beta}_{yz} : j = 0, 1, ...\} \subseteq \mathrm{S}_{E(Y|\boldsymbol{Z})}$
2. $span\{H_e^j \boldsymbol{\beta}_{yz} : j = 0, 1, ...\} \subseteq \mathrm{S}_{E(Y|\boldsymbol{Z})}$

This iteration process is exactly what gave the name to this procedure as the Iterative Hessian Transformations (IHT).

This procedure cannot be done infinitely many times but since the CMS space we are estimating has finite dimension there is a need for a rule that will define a stopping rule, that is a way to ensure exhaustiveness of the CMS that is estimated has already been achieved. This rule is given by the following Proposition.

Proposition 7.2.2 *Let $\boldsymbol{A}$ be a $p \times p$ matrix and $\boldsymbol{\beta}$ a p-dimensional vector. If $\boldsymbol{A}^j\boldsymbol{\beta}$ belongs in the subspace spanned by $\boldsymbol{\beta}, \ldots, \boldsymbol{A}^{j-1}\boldsymbol{\beta}$ then so does $\boldsymbol{A}^s\boldsymbol{\beta}$ for $s > j$.*

7.2.6 Consistency and exhaustiveness of OLS, SIR, SAVE, pHd, IHT

The five methods that have been discussed until now, they have the following two properties:

1. The estimated directions are $\sqrt{n}$ consistent of the population parameters
2. The estimated directions do not exhaust the CDRS or the CMS.

7.2.7 Structure Adaptive Estimation

This approach was presented by Hristache et al (2001) and it is based on iterative improvement of the family of average derivatives. If we denote with $F(X_i) = \nabla f(X_i)$ of the regression function f at every point X_i, then $F(X_i)$ belongs to the index space I. Then one can perform Principal Component Analysis (PCA) to estimate the space I, by computing the matrix $M = \frac{1}{n}\sum_{i=1}^{n} F(X_i)F^T(X_i)$.

This approach was developed for the model 7.6. The problem with this method is the fact that when the effective dimension of the index space is greater than 3, it doesn't achieve $\sqrt{n}$ consistency.

7.2.8 Minimum Average Variance Estimation (MAVE)

Minimum Average Variance Estimation (MAVE) was presented by Xia et al (2002). It was proposed under model 7.6 and it is another method that doesn't achieve $\sqrt{n}$ consistency of the estimators.

The idea here, is that by removing Assumption 7.2.1 one can accommodate applications in data sets like time series data sets where the assumption is violated. So they try to estimate the EDR directions by minimizing $E(Var(Y|\boldsymbol{\gamma}^T\boldsymbol{X}))$ over all $\boldsymbol{\gamma} \in \mathbb{R}^{p\times q}$. To minimize this expectation they employ multivariate kernels which can be very complicated.

7.2.9 Contour Regression

In search of a method that will ensure exhaustiveness of the CDRS and at the same time $\sqrt{n}$ consistency of the estimates Li, Zha and Chiaromonte (2005) proposed

two different methods Simple Contour Regression (SCR) and General Contour Regression (GCR). The two methods target the contour directions of the response surface. Contour directions are those along which the response has small variance, that is they span the complement of the CDRS. The two methods are based on two different measures of the variation of the response. Again, the results are presented considering $\boldsymbol{Z}$, the standardized version of the predictors $\boldsymbol{X}$. The results on the non-standardized predictors $\boldsymbol{X}$ follows similarly.

Simple Contour Regression

Simple Contour Regression (SCR) needs an additional assumption to be established.

Assumption 7.2.3 *For any choice of vectors* $\boldsymbol{v} \in \mathrm{S}_{Y|\boldsymbol{Z}}$ *and* $\boldsymbol{\omega} \in \left(\mathrm{S}_{Y|\boldsymbol{Z}}\right)^{\perp}$ *such that* $\|\boldsymbol{v}\| = \|\boldsymbol{\omega}\| = 1$*, and any sufficiently small* $c > 0$*, we have*

$$\mathrm{var}\left(\boldsymbol{\omega}^T\left(\tilde{\boldsymbol{Z}} - \boldsymbol{Z}\right)\middle|\left|\tilde{Y} - Y\right| \leq c\right) > \mathrm{var}\left(\boldsymbol{v}^T\left(\tilde{\boldsymbol{Z}} - \boldsymbol{Z}\right)\middle|\left|\tilde{Y} - Y\right| \leq c\right) \tag{7.16}$$

where $\left(\tilde{\boldsymbol{Z}}, \tilde{Y}\right)$ *is an independent copy of the random pair* $(\boldsymbol{Z}, Y)$*.*

In order to establish the main theory, the matrix

$$K(c) = E\left(\left(\tilde{\boldsymbol{Z}} - \boldsymbol{Z}\right)\left(\tilde{\boldsymbol{Z}} - \boldsymbol{Z}\right)^T\middle|\left|\tilde{Y} - Y\right| \leq c\right) \tag{7.17}$$

is being consider and it is shown that the eigenvectors of $K(c)$ corresponding to the smallest q eigenvalues span the CDRS as the following theorem states.

Theorem 7.2.10 *If Assumption 7.2.3 holds, then the eigenvectors of* $K(c)$ *corresponding to the smallest* q *eigenvalues span the central subspace* $\mathrm{S}_{Y|\boldsymbol{Z}}$

It has been also shown that the estimation procedure is $\sqrt{n}$ consistent. Proving exhaustiveness and creating test procedures require Assumption 7.2.2 to hold.

General Contour Regression

The problem with SCR is the fact that using the inequality $\left|\tilde{Y} - Y\right| \leq c$ to pick up the contour directions creates problems when the function is not monotone. In this case, it can pick more directions that will affect the results of the estimation. In order to overcome this problem, the use of the following matrix (instead of $K(c)$) is proposed

$$G(c) = E\left(\left(\boldsymbol{Z} - \tilde{\boldsymbol{Z}}\right)\left(\boldsymbol{Z} - \tilde{\boldsymbol{Z}}\right)^T \middle| V\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \leq c\right) \tag{7.18}$$

where

$$V\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) = \text{var}\left(Y \middle| \boldsymbol{Z} = l\left(t; \boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) t \in \mathbb{R}\right) \tag{7.19}$$

and $l\left(t; \boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) = (1-t)\,\boldsymbol{Z} + t\tilde{\boldsymbol{Z}}, t \in \mathbb{R}$

In order to establish the theory we need an assumption similar to the one for the SCR, that is

Assumption 7.2.4 *For any choice of vectors* $\boldsymbol{v} \in \mathrm{S}_{Y|\boldsymbol{Z}}$ *and* $\boldsymbol{\omega} \in \left(\mathrm{S}_{Y|\boldsymbol{Z}}\right)^{\perp}$ *such that* $\|\boldsymbol{v}\| = \|\boldsymbol{\omega}\| = 1$*, and any sufficiently small* $c > 0$*, we have*

$$\text{var}\left(\boldsymbol{\omega}^T\left(\tilde{\boldsymbol{Z}} - \boldsymbol{Z}\right) \middle| V\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \leq c\right) > \text{var}\left(\boldsymbol{v}^T\left(\tilde{\boldsymbol{Z}} - \boldsymbol{Z}\right) \middle| V\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) \leq c\right) \tag{7.20}$$

where $\left(\tilde{\boldsymbol{Z}}, \tilde{Y}\right)$ *is an independent copy of the random pair* $(\boldsymbol{Z}, Y)$*.*

Using this assumption the following theorem can be established

Theorem 7.2.11 *If Assumption 7.2.4 holds, then the eigenvectors of* $G(c)$ *corresponding to the smallest* q *eigenvalues span the central subspace* $\mathrm{S}_{Y|\boldsymbol{Z}}$.

Finally it can be shown that the estimators are $\sqrt{n}$ consistent and that we can exhaustively estimate the CDRS. Also, it can be shown that GCR is robust against non-ellipticity of the predictors.

7.2.10 Directional Regression (DR)

All the methods, that were presented, and especially, SIR, SAVE and pHds are considered classical dimension reduction methods. Each one of them have its own advantages and disadvantages. That's the reason several authors like Gannoun and Saracco (2003) and Ye and Weis (2003) have proposed several combinations of the different methods to achieve better results.

Li and Wang (2006) proposed Directional Regression (DR) that synthesizes the dimension reduction methods based on the first two conditional moments. DR needs substantially less computation and achieve higher accuracy of the direction estimates. The method is based on the empirical directions $\boldsymbol{X}_i - \boldsymbol{X}_j : 1 \leq i < j \leq n$, that were introduced for Contour Regression and were the base of estimation in SCR and GCR.

The whole idea of DR is on the following theorem.

Theorem 7.2.12 *Suppose Assumptions 7.2.1 and 7.2.2 holds. Then* $2I_p - A\left(Y, \tilde{Y}\right) =$

$P\left(2I_p - A\left(Y, \tilde{Y}\right)\right)$ *where*

$$A\left(Y, \tilde{Y}\right) = E\left(\left(\boldsymbol{Z} - \tilde{\boldsymbol{Z}}\right)\left(\boldsymbol{Z} - \tilde{\boldsymbol{Z}}\right)^T \middle| Y, \tilde{Y}\right) \tag{7.21}$$

and P *the projection onto* $\mathrm{S}_{Y|\boldsymbol{Z}}$.

In other words this theorem tells us that the column space of $2I_p - A\left(Y, \tilde{Y}\right)$ is contained in $S_{Y|\boldsymbol{Z}}$.

It is also proved that DR achieves exhaustiveness of the CDRS and $\sqrt{n}$ consistency of the estimates. Moreover, is proved that if the moments involved in SAVE and in DR are finite then the subspace estimated by the two methods are the same. Finally, it is important to note that since DR is a second-moment based method, it will not perform as well as GCR in situations where the regression surfaces have higly fluctuating shapes like high frequency trigonometric functions.

7.2.11 Kernel Dimension Reduction

Fukumizu, Bach and Jordan (2009) proposed a sufficient dimension reduction method called the Kernel Dimension Reduction (KDR) method which involves the use of conditional covariance operators on reproducing kernel Hilbert spaces. The authors, basically, identify that since the "kernel trick" can be applied on reproducing kernel Hilbert spaces it makes them computationally and at the same time they can be used to capture nonparametric phenomena of interest, that makes them really attractive to be used to achieve sufficient dimension reduction. Basically, they show that the conditional covariance operators can be estimated using Gram matrices, and then those Gram matrices can be used to get estimates for the central dimension reduction subspace. That is, those covariance operators can be

used to measure departures of conditional independence.

The authors first explain that a reproducing kernel Hilbert space $\mathcal{H}$ is characteristic if and only if $\int f dP = \int f dQ$ for all $f \in \mathcal{H}$ means $P = Q$. Then they state the basic theorem which states the following:

Theorem 7.2.13 *Let*

- $\mathfrak{X}$ *be a closed ball* $D_m(r) = \{x \in \mathbb{R}^m | \|x\| \leq r\}$ *or an entire Euclidean space* $\mathbb{R}^m$
- $\mathcal{H}_{\mathfrak{X}}$ *is the Reproducing kernel space of functions on* $\mathfrak{X}$.
- $S_d^m(\mathbb{R})$ *is the set of all* d *orthonormal vectors in* $\mathbb{R}^m$
- $\mathcal{H}_{\mathfrak{X}}^B$ *be the reproducing kernel Hilbert space associated with the positive definite kernel* $k_{\mathfrak{X}^B}(x, \tilde{(x)}) = k_d\left(B^\intercal x, B^\intercal \tilde{(x)}\right)$ *where* d *is the minimum dimension of the sufficient dimension reduction subspace and* $B \in S_d^m(\mathbb{R})$.

Suppose the closure of $\mathcal{H}_{\mathfrak{X}}^B$ *in* $L^2(P_{\boldsymbol{X}})$ *is included in the closure of* $\mathcal{H}_{\mathfrak{X}}$ *in* $L^2(P_{\boldsymbol{X}})$ *for any* $B \in S_d^m(\mathbb{R})$. *If for* $(\mathcal{H}_{\mathfrak{X}}, P_{\boldsymbol{X}})$ $\mathcal{H}_{\mathfrak{X}} + \mathbb{R}$ *is dense in* $L^2(P_{\boldsymbol{X}})$ *and for* $(\mathcal{H}_{\mathfrak{X}}^B, P_B)$ $\mathcal{H}_{\mathfrak{X}}^B + \mathbb{R}$ *is dense in* $L^2(P_{\boldsymbol{X}})$ *for all* $B \in S_d^m(\mathbb{R})$ *(where "+" denotes the direct sum between two reproducing kernel Hilbert spaces) and also* $\mathcal{H}_{\mathfrak{Y}}$ *is characteristic then:*

$$\Sigma_{YY|\boldsymbol{X}} = \Sigma_{YY|\boldsymbol{X}}^B \Leftrightarrow Y \perp\!\!\!\perp \boldsymbol{X} | B^\intercal \boldsymbol{X}$$

where $\Sigma_{YY|\boldsymbol{X}}$ *is the conditional covariance operator.*

Two of the most important features of KDR is first, the fact that it does not impose any strong assumptions on the distribution of the predictors $\boldsymbol{X}$ and second,

that it can find directions even when the central dimension reduction subspace does not exist. As long as d is chosen to be large enough such that sufficient dimension subspaces with that dimension exist, then the algorithm will converge to those subspaces.

7.2.12 Sufficient Dimension Reduction without matrix inversion

All the methods that have been mentioned for inference about $S_{Y|\boldsymbol{Z}}$ require the inversion of the sample version $\hat{\boldsymbol{\Sigma}}$ of the $p \times p$ predictor covariance matrix $\boldsymbol{\Sigma}$. That means, they also require that $n > p$. If $n < p$ then the inversion will be impossible since the rank of $\hat{\boldsymbol{\Sigma}}$ is $\min(n, p)$.

Cook, Li and Chiaromonte (2007) propose a general approach that allows many methods to be adapted in regressions where $n < p$. The approach, is similar to IHT and requires computation of powers of $\hat{\boldsymbol{\Sigma}}$ instead of $\hat{\boldsymbol{\Sigma}}^{-1}$.

Let ν be any matrix such that $span(\nu) \subseteq \boldsymbol{\Sigma} S_{Y|\boldsymbol{X}}$. Such a matrix is called a seed matrix. The idea is that if we have a matrix $\mathbf{R}$ which columns form a basis for $\mathcal{M}$, where $\mathcal{M}$ is a subspace of $\mathbb{R}^p$ that contains $S_{Y|\boldsymbol{Z}}$, then

$$\boldsymbol{\Sigma}^{-1}\nu = \mathbf{R}\left(\mathbf{R}^T\boldsymbol{\Sigma}\mathbf{R}\right)^{-1}\mathbf{R}^T\nu \tag{7.22}$$

and consequently $\boldsymbol{\Sigma}^{-1}$ is not required.

By defining $\mathcal{M}_{Y|\boldsymbol{Z}}$ to be the intersection of all subspaces $\mathcal{M}$ that contain $S_{Y|\boldsymbol{Z}}$ then, $\mathcal{M}_{Y|\boldsymbol{Z}}$ is the smallest subspace that contains $S_{Y|\boldsymbol{Z}}$ and conforms the eigenstructure of $\boldsymbol{\Sigma}$. That means $\mathcal{M}_{Y|\boldsymbol{Z}}$ can be constructed without inverting $\boldsymbol{\Sigma}$.

Let the matrices $\mathbf{R}_u \equiv \left(\nu, \boldsymbol{\Sigma}\nu, \ldots, \boldsymbol{\Sigma}^{u-1}\nu\right)$ for $u = 1, 2, \ldots$. The following

theorem explains how one using matrices $\mathbf{R}_u$ can find the space $\mathcal{M}_{Y|\boldsymbol{Z}}$.

Theorem 7.2.14 *Suppose* $\boldsymbol{\Sigma}$ *is positive definite with* q *the distinct nonzero eigenvalues,* k *of which correspond to eigenspaces not orthogonal to* $\mathrm{S}_{Y|\boldsymbol{Z}}$*. Then there exists an integer* $1 \leq \tilde{u} \leq k$ *such that* $span(\mathbf{R}_u)$ *is strictly increasing until* $u = \tilde{u}$*, and settles upon* $\mathcal{M}_{Y|\boldsymbol{Z}}$ *thereafter:*

$$span(\mathbf{R}_1) \subset \ldots \subset span(\mathbf{R}_{\tilde{u}}) = \mathcal{M}_{Y|\boldsymbol{Z}} = span(\mathbf{R}_{\tilde{u}+1}) = \ldots \tag{7.23}$$

Further results to ensure the capture of $S_{Y|\boldsymbol{Z}}$, the choice of d the dimension of $S_{Y|\boldsymbol{Z}}$ and the choice of $\tilde{u}$ are given by Cook, Li, Chiaromonte (2007).

The importance of this work is that it transforms the already known algorithms to accommodate regressions where $n < p$, as long as $n > d$, by removing the requirement of matrix inversion.

7.3 Sufficient Dimension Reduction for non-linear feature extraction by applying existing methods in the feature space

Wang (2008) observed that model 7.1 is not suitable for dimension reduction in case we have interaction terms in our regression function. One can still apply the methods that were presented in section 7.1 but will lose power since to capture the interactions there is a need to find too many linear combinations. In this philosophy Wang (2008) proposed three different methods of using the feature space of support vector machines (SVM) (to be introduced later) to achieve Sufficient

Dimension Reduction. Those methods are based on the assumption of $\phi(\boldsymbol{X})$ and Y, or consequently $\boldsymbol{X}$ and Y, are independent given $\alpha^T\phi(\boldsymbol{X})$, that is

$$Y \perp\!\!\!\perp \phi(\boldsymbol{X})|\alpha^T\phi(\boldsymbol{X}) \text{ or } Y \perp\!\!\!\perp \boldsymbol{X}|\alpha^T\phi(\boldsymbol{X}) \tag{7.24}$$

where, for a degree 2 polynomial,

$$\phi(\boldsymbol{X})^T = \left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p, \boldsymbol{X}_1^2, \ldots, \boldsymbol{X}_p^2, \boldsymbol{X}_1\boldsymbol{X}_2, \ldots, \boldsymbol{X}_1\boldsymbol{X}_p, \boldsymbol{X}_2\boldsymbol{X}_3, \ldots, \boldsymbol{X}_{p-1}\boldsymbol{X}_p\right) \tag{7.25}$$

One can modify the definition accordingly for higher order polynomials.

7.3.1 General Estimation method for Dimension Reduction on the Feature space

The goal in the setting of model 7.24 is to estimate matrix α. There is a big issue associated with it though. It is not possible for the predictor space to have the elliptically contoured predictor distribution required by most dimension reduction methods.

To overcome this issue, Wang (2008) extends a result by Diaconis and Freedman (1984), to show that for degree two random polynomial feature vectors generated from multivariate normal distributions, the low dimensional projections of the feature vectors have an asymptotic multivariate normal distribution. Using this result, one can apply existing dimension reduction methods on the feature space directly. This method is called Global Estimation method (GE).

7.3.2 Local Estimation method for Dimension Reduction on the Feature space

The results of GE although can get results that other methods of Sufficient Dimension Reduction were not able to give, they are limited to the extend to which $\alpha^T\phi(\boldsymbol{X})$ is elliptical. In order to maximize performance one can refer to methods such as SAE and MAVE, that can be used without using Assumption 7.2.1.

In this case, one can use an Invariant Aggregation (IA) method proposed by Tang (2007) which requires weaker assumption on the distribution of $\boldsymbol{X}$. IA requires the joint distribution of $(\boldsymbol{X}, Y)$ to be symmetric about the central space and the dimension reduction vectors estimators to satisfy an equivariant assumption that is proposed by Theorem 5.5.2 in Tang (2007). Wang (2008) combine the SIR and IA methods to implement a new algorithm of Dimension Reduction on the feature space which performs better than the GE estimator, creating sharper images of the data, but the improvement is not very significant.

Another approach proposed by Wang (2008) is the Nonlinear Aggregation (NA). This method is based on the fact that

$$\phi(x) \approx \phi(b) - \phi(b)b + \phi(b)x \tag{7.26}$$

for any $b \in \Omega_{\boldsymbol{X}}$, the domain of $\boldsymbol{X}$. Then for model 7.24 we have approximately in a neighborhood of b

$$Y \perp\!\!\!\perp \boldsymbol{X} | \alpha^T\phi(b)\boldsymbol{X} \tag{7.27}$$

That means, that one can first establish a dimension reduction method at local regions of the original predictor $\boldsymbol{X}$ to get an estimator for $M_0(b;\rho)$ (this matrix

definition depends on the dimension reduction method that will be used). This matrix help estimate the matrix

$$M(b;\rho) = \phi(b)\left(\phi(b)^T\phi(b)\right)^{-1} M_0(b;\rho)\left(\phi(b)^T\phi(b)\right)^{-1}\phi^T(b) \tag{7.28}$$

whose columns span $S_{Y|\phi(b)\boldsymbol{X}(b;\rho)}$ where $\boldsymbol{X}(b;\rho) = \boldsymbol{X}I(\|\boldsymbol{X}-b\| \leq \rho)$. By taking the $E\left(M(b;\rho)\right)$, that is, aggregating the local information, one can obtain $S_{Y|\phi(\boldsymbol{X})}$. The estimation results are a little bit better compared to the IA method and a lot better compared to the GE method.

7.3.3 Kernel Slice Inverse Regression

More recently Wu (2008) and Yeh, Huang, and Lee (2009) used the "kernel trick" to extend the SIR to the nonlinear setting. They proposed the Kernel Slice Inverse Regression algorithm, in a reproducing kernel Hilbert space. First, they define the space spanned by the column vectors α in (7.24) as the effective dimension reduction subspace and then they extend the linear conditional mean assumption to the feature space which is a reproducing kernel Hilbert space. So using a finite basis in the feature space, denoted with $K(\cdot,A)$ one can express the following theorem

Theorem 7.3.1 *If the existence of a feature space can be assumed* $\mathcal{H} = \text{span}\{K(\mathbf{x},A)\alpha_1,\ldots,K(\mathbf{x},A)\alpha_d\}$ *and that the linear conditional mean assumption holds in the feature space, that is:*

$$E\left(\alpha^\intercal K(\mathbf{x},A)|\alpha_1^\intercal K(\mathbf{x},A),\ldots,\alpha_d^\intercal K(\mathbf{x},A)\right) = c_0 + c_1\alpha_1^\intercal K(\mathbf{x},A) + \ldots, + c_d\alpha_d^\intercal K(\mathbf{x},A)$$

for every $\alpha = (\alpha_1,\ldots,\alpha_n) \in \mathbb{R}^n$. *Then:*

$$E(K(\mathbf{x},A)|Y) - E(K(\mathbf{x},A)) \in \text{span}\{\Sigma_K\alpha_1,\ldots,\Sigma_K\alpha_d\}$$

where $\Sigma_K = \text{cov}\left(E(K(\mathbf{x}, A))\right)$.

The authors also proposed several algorithmic revisions that reduced the complexity of the implementation of the algorithm and cut down the computational load. The main problem is that since they needed matrix inversion, singularity issues led to reduced performance and also numerical instability.

In this Chapter we have given an overview of the most important methods developed for Sufficient Dimension Reduction and some recent works towards nonlinear dimension reduction. This is just an overview, and it does not exhaust the current literature. There is an abundance of methods not discussed here for several reasons, the most important one, is that it is quite impossible for an introduction into sufficient dimension reduction to cover all of them. Also, some methods that were developed for other type of data, for example, functional data (Amato, Antoniadis and De Feis (2006)) or survival data (Li, Wang and Chen (1999)), does not fit the context of this work. In the next Chapter 8 we will describe support vector machines and how we intend to use them to achieve sufficient dimension reduction. In Chapter 9 we give the asymptotic results for SVM, we describe the estimation procedure and we describe how one can achieve dimension reduction. In Chapter 10 we present a simulation analysis to compare the performance of our method to other methods and the performance of our method when we change some of the parameters. Finally, in Chapter 11 we show the effectiveness of our method in two real datasets and in Chapter 12 we discuss our work and implications for future work.

Chapter 8

Support Vector Machines for dimension reduction

In this Chapter we first introduce the idea of a separating hyperplane and the extension to support vector machines, for separable and non-separable data. Then we discuss how Support Vector Machines can be used for sufficient dimension reduction, and we present some modifications on the objective function that can lead to dimension reduction without matrix inversion.

8.1 Early Machine Learning Algorithms

Since the use of computer was widely used in the 1950's, the construction of machines that were capable of learning from experience has received a lot of attention by researchers in both aspects; philosophically and technically. The first algorithms that was developed, was the Rosenblatt's Perceptron algorithm which was

proposed by Rosenblatt (1962) and it's generalization, radial basis functions (or potential functions) that were proposed by Aizerman, Braverman and Rozonoer (1964a) and (1964b) and neural networks that were proposed by Rumelhart, Hinton and Williams (1986) and LeCun (1986). The reader is referred to Vapnik (1998) Chapter 9 for a complete introduction on these methods and for a presentation of more theoretical results.

While the learning machine algorithms were becoming more widely used, since the computer power increased significantly in the early 1990s the need for more accurate algorithms forced researches to explore more ideas. Support Vector Machines (SVMs) were introduced by Cortes and Vapnik (1995) and since then became one of the most widely used learning method for classification and regression.

8.2 The optimal hyperplane

In a typical two category classification problem in $\mathbb{R}^p$ each point is viewed as a p-dimensional vector. The objective of a classification algorithm is to find a $(p-1)$-dimensional hyperplane that divides the data in the two category clouds. This hyperplane is called the linear classifier, as data on one side of the hyperplane belong to one category and data on the other side of the hyperplane belong to the other category. The hyperplane is called the optimal hyperplane if it maximizes the margin between the two data clouds. The notation and formulation that are used in this subsection and next subsection follow closely the one by Vapnik (1998).

Let assume that we have a finite set of vectors x that form the training set $(y_1, x_1), \ldots, (y_n, x_n)$ where $x \in \mathbb{R}^p$ and $y \in \{-1, 1\}$. As one can see y is the indicator variable on which category each x vector belongs. Let $x \in I$ if $y = 1$ and

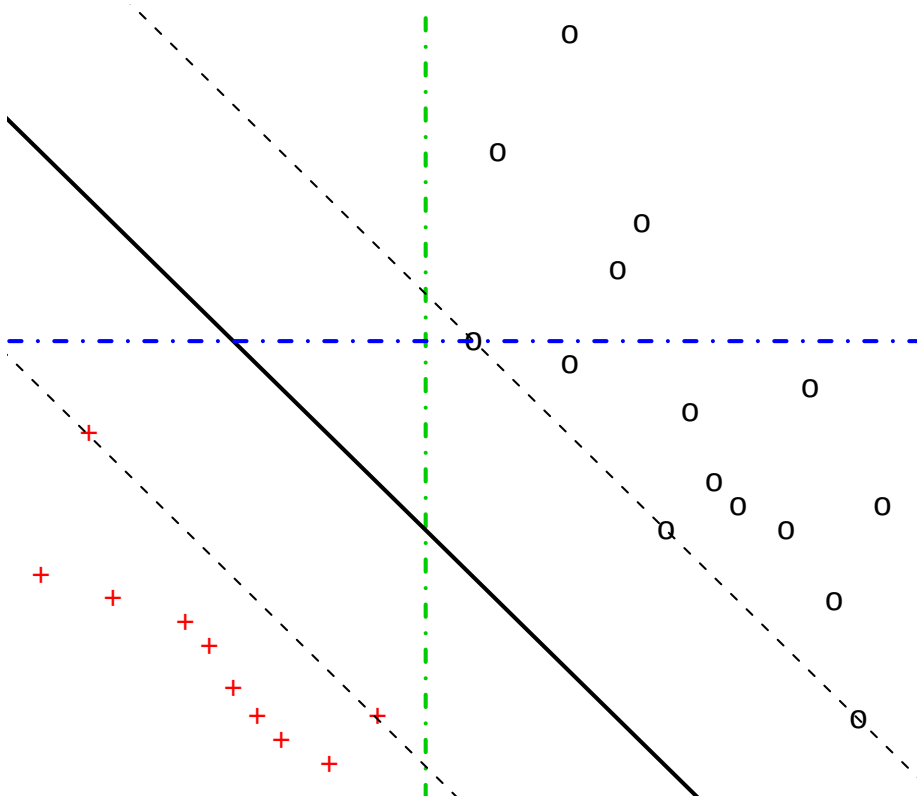

Figure 8.1: An example of two samples (black circles and red crosses) with possible separating hyperplanes. The blue dot-dash line is not a separating hyperplane. The green dot dash line is a separating hyperplane. The black solid line is the optimal hyperplane, as it achieves separation with maximum distance from the points. The points that fall on the two black dash lines are the support vectors.

$x \in II$ if $y = -1$. There exists a hyperplane

$$(x * \phi) = c \tag{8.1}$$

if there exists a unit vector ϕ and a constant c such that the following inequalities hold:

$$\begin{aligned} (x_i * \phi) > c, \text{ if } x_i \in I \\ (x_i * \phi) < c, \text{ if } x_i \in II \end{aligned} \tag{8.2}$$

For any unit vectors ϕ one can define the following:

$$\begin{aligned} c_1(\phi) = \min_{x_i \in I}(x_i * \phi), \\ c_2(\phi) = \max_{x_i \in II}(x_i * \phi), \end{aligned}$$

Then by taking the unit vector ϕ_0 that maximizes the function

$$\rho(\phi) = \frac{c_1(\phi) - c_2(\phi)}{2}, \ |\phi| = 1 \tag{8.3}$$

and the constant

$$c_0 = \frac{c_1(\phi_0) + c_2(\phi_0)}{2} \tag{8.4}$$

one can determine the optimal hyperplane that divides the data in the two predefined categories. The optimal hyperplane is also called the "maximal margin hyperplane", since equation (8.3) is called the margin of the separating hyperplane. Conceptually, margin is the distance of the hyperplane from the closest point, in each side of the hyperplane. It has been proved that this hyperplane is

unique.

Figure 8.1 shows a case where we try to find the optimal hyperplane. Our two populations are the black circles and the red crosses. The optimal hyperplane is shown with a black solid line, while the blue dot dash lines is not a separating hyperplane and the green dot dash line is a separating hyperplane which is not optimal, as the distance from the points is not as big as the distance the black solid line has.

A formal definition of the optimal hyperplane (Vapnik (1998)) that achieves a margin Δ is as follows:

Definition 8.2.1 *We call a hyperplane*

$$\langle \psi^*, x \rangle - b = 0, |\psi^*| = 1$$

the Δ***-separating hyperplane*** *if it classifies vectors* x *as:*

$$y = \begin{cases} 1 & \text{if} \langle \psi^*, x \rangle - b \geq \Delta \\ -1 & \text{if} \langle \psi^*, x \rangle - b \leq -\Delta \end{cases}$$

There is a very attractive feature on the calculation of the hyperplane. It was shown that the equation of the vector ψ^* that is associated with the optimal hyperplane is only associated with the points in the dataset that are closer to the hyperplane. On Figure 8.1 those are the points that are on the black dash lines. Those points are called the support vectors of the hyperplane.

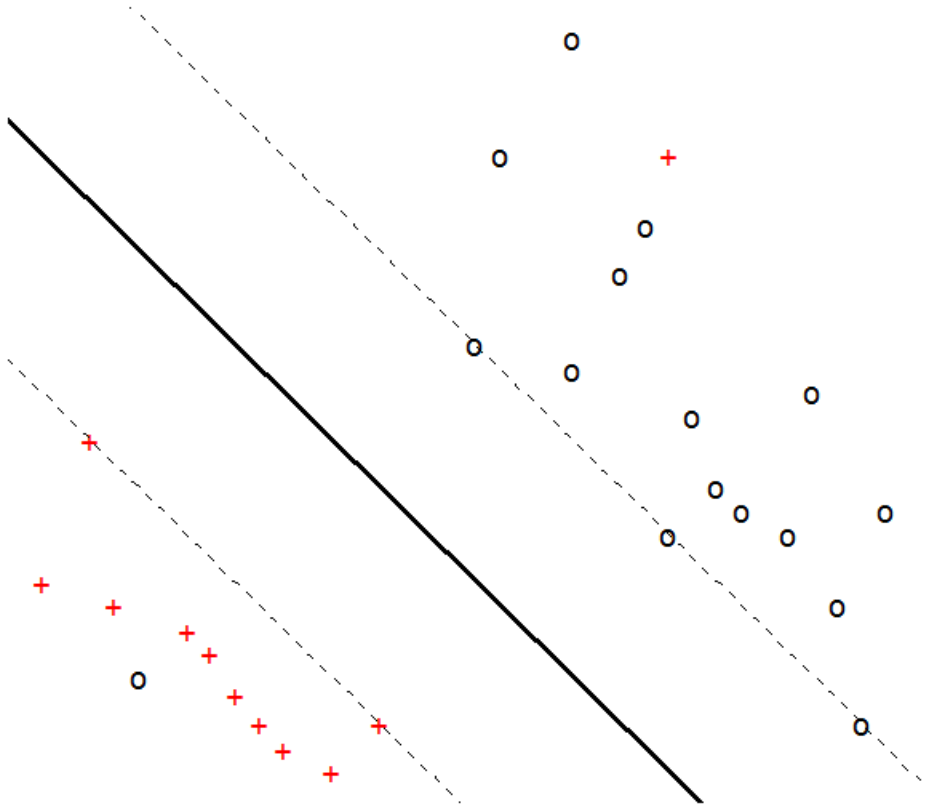

Figure 8.2: An example of two samples (black circles and red crosses) that are not linearly separable. We can see that one black circle and a red cross are incorrectly classified by this optimal hyperplane.

8.3 Duality of the problem

As we have said before the objective of the SVM algorithms is to maximize the margin Δ for all unit vectors that separate our dataset (see Definition 8.2.1). That is our objective is maximize the margin Δ among all $|\psi| = 1$ under constraints:

$$y_i \left(\psi^\intercal x_i - \beta\right) \geq \Delta$$

In the literature the dual problem is presented more often, due to the easier way of dealing with the calculations. The dual problem instead of trying to maximize the margin for all unit vectors, we minimize the length of the vector associated with the hyperplane, for a fixed unit margin. That is one can minimize $|\psi|^2 = \psi^\intercal \psi$ so that the margin $\Delta = 1$ under constraints

$$y_i \left(\psi^\intercal x_i - b\right) \geq 1$$

As we will see in the future developments, this expression of the problem, will make it easier for us to extend the problem in the nonseparable case as well as the population version of the problem.

This idea is easy to understand and implement if the data is separable, but that is not the case in the majority of real life problems (see Figure 8.2). In the following sections we describe how one can attack the nonseparable cases, using the optimal hyperplane idea, to find the optimal linear hyperplane (by minimizing the cost of misclassification) and how one can use the support vector machine algorithm to attack it and find a nonlinear separation of the dataset.

8.4 Nonseparable case

Figure 8.2 shows a non linearly separable case. The way to attack this problem is to introduce some variables, $\xi_i, i = 1, \ldots, n$ which denote the misclassification distance for each of the points. If the i^{th} point is correctly classified then $\xi_i = 0$, and if it is incorrectly classified then $\xi_i > 0$. In this case the problem of finding the optimal hyperplane is transformed into the one that tries to minimize:

$$\psi^{\mathsf{T}}\psi + c\sum_{i=1}^{n}\xi_i \tag{8.5}$$

under the modified constraints that $\xi_i \geq 0$ and:

$$y_i\left(\psi^{\mathsf{T}}x_i - b\right) \geq 1 - \xi_i \tag{8.6}$$

8.5 Support Vector Machines (SVMs)

Support Vector Machine maps the input vectors x into a high-dimensional space, which is called "the feature space", using a non-linear mapping which is chosen a priori. In the feature space the optimal hyperplane can be constructed.

If for example, someone needs to find a hyperplane that is a polynomial of second degree then the feature space is a $p(p+3)/2$ dimensional space, where p is the dimension of vector x. These dimensions break as follows:

1. p coordinates for each of the elements of vector x,
2. p coordinates for the square of each of the elements of vector x,
3. $p(p-1)/2$ for the all the interactions of between the elements of vector x

From the above example, one can easily see that the dimension of the feature space, increase dramatically, as the dimension of vector x increases and as the degree of the polynomial that constructs the hyperplane increases.

To reduce the computation, it was shown that one can use only the support vectors of a data set in order to find the optimal hyperplane using the SVM algorithm. Support vectors are the vectors in the data set that their distance from the hyperplane is exactly equal to the margin of the hyperplane. In Figure 8.1 the support vectors are the points lying on the black dash lines.

A strong feature of Support Vector Machines is the fact that you can use functions to map the nonseparable data into higher (possible infinite) dimensional spaces called Reproducing Kernel Hilbert spaces. The functions that derive those spaces are called kernels (see relationship (5.2) and the discussion that precedes it, for a definition of Reproducing kernel Hilbert spaces). The advantage of kernel functions is that one can use the "kernel trick"; that is if a computation depends only on inner products, one can extract the lower dimensional projections without calculating the projection coefficients, which reside in the feature space, which is a higher (sometimes infinite) dimensional space.

After their appearance in Cortes and Vapnik (1995), SVMs were expanded in very different directions, such as the estimation of real valued functions, pattern recognition and regression estimation. The interested reader is referred in Vapnik (1998) and Hastie, T., Tibishrani, R. and Friedman, J. (2009) for further reading.

8.6 Using Support Vector Machines for Dimension Reduction

In this section we describe how SVM works in the dimension reduction framework. This is basically a description of the theoretical results that follow in the next Chapter.

First, we revisit the SIR algorithm that was presented in section 7.2.2. As it is shown in Figure 8.3 SIR slices the response surfaces into a number of slices. In each slice we find the average of the all the x_i's that belong in the slice. Figure 8.3 shows that in the population level this average will be in the middle of the slice. Connecting E($\boldsymbol{X}|Y$) with the point of origin will give you a direction (denoted as $\boldsymbol{\beta}$ on the figure) in the Central Dimension Reduction Subspace (CDRS) denoted as $S_{Y|\boldsymbol{X}}$. SIR as well as other inverse regression methods like SAVE and DR depend on the sample moments, which is well known that they are not robust in the presence of outliers.

With support vector machines the procedure will be similar. We will divide the response surface into slices. But we will project only the response surface on the predictor space. The slice divisor will not be projected on the predictor space, so in a sense it will be unknown. Assuming that we have only two slices, on the predictor space we will have the points that belong in slice 1 and the points that belong in slice 2. Our objective will be to estimate the slice divisor as the optimal hyperplane dividing the points in the two slices using a SVM algorithm. The vector that is vertical to the optimal hyperplane is in the CDRS $S_{Y|\boldsymbol{X}}$.

This procedure will be helpful in three ways. First, as we have seen earlier, in Sections 8.2 and 8.5, the equation of the optimal hyperplane depends only on the

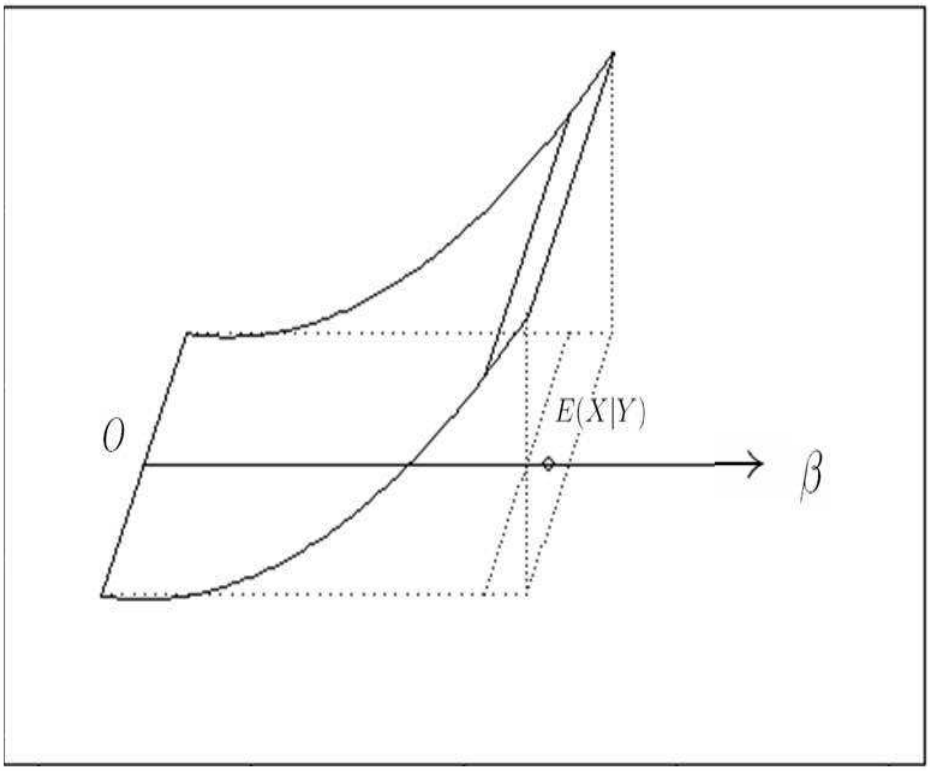

Figure 8.3: An example with two-dimensional predictor. A response surface is shown and one slice divisor is marked on the response surface. The response surface and the slice divisor is projected on the predictor surface. The population mean of the predictors in the slice is the point in the middle of the slice denoted as $\mathrm{E}(\boldsymbol{X}|Y)$

support vectors, those points which are closer to the separating hyperplane. Thus, if there are outliers in our samples, which are far away from the rest of the points in the two slices, there will not be much effect on the final equation of the hyperplane. This gives us a robust way to estimate the directions in the CDRS, something previous dimension reduction methods that depend on inverse sample moments did not address. Second, a slight modification in the objective function that we minimize when estimating the optimal hyperplane, enables dimension reduction without matrix inversion. Third, we can use any kernel to move into a higher dimensional feature space, something that allow us to extract nonlinear features in the CDRS. The first two objectives are explored in the next two Chapters in theory first and then through some simulations. The third one is addressed only at the end of the next Chapter by showing a brief theoretical extension. More theoretical work and simulations are left for the future.

8.7 Population level SVM

One of the most important features of the SVM is the fact that the development presented until now in this work, as well as the definition of the optimal hyperplane in Definition 8.2.1, is based on the sample level. To be able to derive asymptotic properties of our estimators, we need to extend the development to the population level.

Since the problems we deal with have nonseparable datasets we revisit the problem of finding the optimal hyperplane as it was presented in section 8.4. The way it was presented there is not helpful for our future developments and so we will present it a little bit differently. The first step will be to fix the hyperplane parameters ψ and β and try to find the equation of the ξ_i's that minimizes the

objective function (8.5). Using the constraints (8.6) one can see that the minimum of 8.5 is achieved when

$$\xi_i^* = [1 - y_i(x_i^\intercal \psi - b)]^+$$

Substituting this into the objective function (8.5) we get that our objective is to minimize:

$$\psi^\intercal \psi + C \sum_{i=1}^{n} [1 - y_i(x_i^\intercal \psi - b)]^+. \tag{8.7}$$

This version of the minimization problem can be expressed as an expectation in the population level. Let's assume that we have $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be a sample of points in $\mathbb{R}^p$ and $\{Y_1, \ldots, Y_n\}$ be a sample of labels in $\{-1, 1\}$. Then the above objective function can be re-expressed in the population level as follows:

$$\psi^\intercal \psi + C\mathrm{E}[1 - Y(\boldsymbol{X}^\intercal \psi - b)]^+. \tag{8.8}$$

The optimal hyperplane (ψ, b) that minimizes this objective function, is the optimal hyperplane that separates the conditional distributions of $\boldsymbol{X}|Y = 1$ and $\boldsymbol{X}|Y = -1$.

8.8 Achieving Dimension Reduction without matrix inversion

Finally, there is a way to address dimension reduction without the need of matrix inversion. This is very important as it enables us to attack dimension reduction in large p small n problems, that frequently appear in the literature.

The idea is that the minimizer of the objective function 8.8 is the same as the minimizer of the following objective function:

$$\psi^{\mathsf{T}}\boldsymbol{\Sigma}\psi + C\mathrm{E}[1 - Y(\boldsymbol{X}^{\mathsf{T}}\psi - b)]^{+}. \tag{8.9}$$

where $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X})$. This small modification that adds the covariance matrix in the first term enables us to perform dimension reduction without the need of inverting the matrix, that is without the need of standardizing the predictors.

It is important to note that the objective function (8.9) is different form the standard objective function (8.8), we can nevertheless use the standard SVM algorithm to solve our problem by first applying a linear transformation of $\boldsymbol{X}$. Note that

$$\psi^{\mathsf{T}}\boldsymbol{\Sigma}\psi = \langle\psi, \boldsymbol{\Sigma}\psi\rangle = \langle\boldsymbol{\Sigma}^{\frac{1}{2}}\psi, \boldsymbol{\Sigma}^{\frac{1}{2}}\psi\rangle, \quad \boldsymbol{X}^{\mathsf{T}}\psi = \langle\boldsymbol{X}, \psi\rangle = \langle\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}, \boldsymbol{\Sigma}^{\frac{1}{2}}\psi\rangle.$$

Thus, if we let

$$\boldsymbol{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}, \quad \phi = \boldsymbol{\Sigma}^{\frac{1}{2}}\psi,$$

then the objective function can be re-written as

$$\phi^{\mathsf{T}}\phi + CE\left[1 - Y\left(\boldsymbol{Z}^{\mathsf{T}}\phi - b\right)\right]^{+}.$$

This is the standard objective function. Thus we can apply the standard SVM to this problem to estimate ϕ^*, and then compute ψ^* using the relation $\psi^* = \boldsymbol{\Sigma}^{-\frac{1}{2}}\phi^*$.

Chapter 9

Estimation procedure and asymptotic results

In this Chapter we will discuss some theoretical results. Mainly we will show that there is a unique minimizer of the objective function 8.9.

First we refine more the objective function to fit the idea of slicing the range of values of Y. Let Ω_1 and Ω_2 be a partition of Ω. Let $\tilde{Y}$ be the discrete random variable defined by

$$\tilde{Y} = \begin{cases} -1 & Y \in \Omega_1 \\ 1 & Y \in \Omega_2 \end{cases} \tag{9.1}$$

We modify the objective function (8.9) as follows

$$\psi^{\intercal}\mathbf{\Sigma}\psi + CE[1 - \tilde{Y}(\boldsymbol{X}^{\intercal}\psi - b)]^{+}. \tag{9.2}$$

Now it is clear the objective function depends on ψ, b and the joint distribution of $(\boldsymbol{X}, \tilde{Y})$. We denote it with $L(\psi, b, \mathbf{P}_{\boldsymbol{X},\tilde{Y}})$, that is:

$$L(\psi, b, \mathbf{P}_{\boldsymbol{X},\tilde{Y}}) = \psi^{\mathsf{T}}\boldsymbol{\Sigma}\psi + CE[1 - \tilde{Y}(\boldsymbol{X}^{\mathsf{T}}\psi - b)]^{+}. \tag{9.3}$$

The following result shows that the value of the objective function stays the same, whether we multiply the predictors or the minimizer, with a matrix from the left.

Theorem 9.0.1 *Let $\boldsymbol{A}$ be a $p \times p$ matrix. Then*

$$L(\psi, b, P_{\boldsymbol{AX},\tilde{Y}}) = L(\boldsymbol{A}^{\mathsf{T}}\psi, b, P_{\boldsymbol{X},\tilde{Y}}).$$

Proof. It is easy to see that both the right- and the left-hand sides are

$$\psi^{\mathsf{T}}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}\psi + CE[1 - \tilde{Y}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}\psi - b)]^{+},$$

as desired. □

9.1 Unbiasedness of the normal vector of optimal hyperplane

In this section we show that, if (ψ^*, b^*) is the minimizer of $L(\psi, b, P_{\boldsymbol{X},\tilde{Y}})$ in (9.3) and if $\boldsymbol{X}$ has an elliptically-contoured distribution, then, under some additional mild conditions, ψ^* belongs to the central subspace $\mathcal{S}_{Y|\boldsymbol{X}}$. The proof relies on

a symmetric property of the joint distribution of $(\boldsymbol{X}, Y)$, which is derived from the elliptical distribution assumption on $\boldsymbol{X}$. This idea was first used in Li, Zha, and Chiaromonte (2005) to prove the unbiasedness of contour regression. It is developed more fully in Tang (2007), in the context of invariant aggregation of dimension reduction estimators.

Recall that a random vector $\boldsymbol{Z}$ has a spherical distribution if $\boldsymbol{Z} \stackrel{\mathcal{D}}{=} \boldsymbol{A}\boldsymbol{Z}$ for any orthogonal matrix $\boldsymbol{A}$. If $\boldsymbol{X} = \boldsymbol{\Sigma}\boldsymbol{Z}$ for some positive definite matrix $\boldsymbol{\Sigma}$, then $\boldsymbol{X}$ is said to have an elliptical distribution with shape matrix $\boldsymbol{\Sigma}$. If the components of $\boldsymbol{X}$ have finite variances, then the shape matrix is proportional to the covariance matrix of $\boldsymbol{X}$. Since we always assume $E(\boldsymbol{X}\boldsymbol{X}^\intercal)$ to have finite components, henceforth we take the shape matrix $\boldsymbol{\Sigma}$ to be the covariance of $\boldsymbol{X}$ without loss of generality. Consider the Hilbert space $\{\mathbb{R}^p, \langle\cdot,\cdot\rangle_{\boldsymbol{\Sigma}}\}$, where the inner product is defined by $\langle \mathbf{a}, \mathbf{b}\rangle_{\boldsymbol{\Sigma}} = \mathbf{a}^\intercal\boldsymbol{\Sigma}\mathbf{b}$.

The adjoint matrix of $\mathbf{U}$ with respect to the inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{\Sigma}}$ is the matrix $\mathbf{U}^*$ such that

$$\langle \mathbf{a}, \mathbf{U}\mathbf{b}\rangle_{\boldsymbol{\Sigma}} = \langle \mathbf{U}^*\mathbf{a}, \mathbf{b}\rangle_{\boldsymbol{\Sigma}}$$

for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$. It is easy to see that $\mathbf{U}^*$ satisfies the above relation if and only if $\mathbf{U}^* = \boldsymbol{\Sigma}^{-1}\mathbf{U}^\intercal\boldsymbol{\Sigma}$. We will say that $\mathbf{U}$ is a $\boldsymbol{\Sigma}$-orthogonal matrix if $\mathbf{U}^* = \mathbf{U}^{-1}$.

Lemma 9.1.1 *The following statements are equivalent:*

1. $\boldsymbol{X}$ *has an elliptical distribution with shape matrix* $\boldsymbol{\Sigma}$*;*
2. *for any* $\boldsymbol{\Sigma}$*-orthogonal matrix* $\mathbf{U}$, $\boldsymbol{X} \stackrel{\mathcal{D}}{=} (\mathbf{U}^*)^\intercal\boldsymbol{X}$*;*
3. *for any* $\boldsymbol{\Sigma}$*-orthogonal matrix* $\mathbf{U}$, $\boldsymbol{X} \stackrel{\mathcal{D}}{=} \mathbf{U}^\intercal\boldsymbol{X}$.

4. *for any* $\boldsymbol{\Sigma}$*-orthogonal matrix* $\mathbf{U}$, $\boldsymbol{X} \overset{\mathcal{D}}{=} \mathbf{U}^*\boldsymbol{X}$;

5. *for any* $\boldsymbol{\Sigma}$*-orthogonal matrix* $\mathbf{U}$, $\boldsymbol{X} \overset{\mathcal{D}}{=} \mathbf{U}\boldsymbol{X}$.

PROOF. $1 \Rightarrow 2$. Let $\mathbf{U}$ be a $\boldsymbol{\Sigma}$-orthogonal matrix. Then $\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{U}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ is an orthogonal matrix. Since $\boldsymbol{X}$ has an elliptical distribution with shape matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}$ has a spherical distribution. Hence

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X} \overset{\mathcal{D}}{=} \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \Rightarrow \boldsymbol{X} \overset{\mathcal{D}}{=} (\mathbf{U}^*)^\mathsf{T}\boldsymbol{X}.$$

$2 \Rightarrow 3$. Multiply both sides of the second equality above by $\mathbf{U}^\mathsf{T}$ from the left.

$3 \Rightarrow 1$. From 3 we deduce that $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X} \overset{\mathcal{D}}{=} \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{U}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}$, where $\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{U}^\mathsf{T}\boldsymbol{\Sigma}^{\frac{1}{2}}$ is an orthogonal matrix. Hence $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}$ has a spherical distribution.

$2 \Leftrightarrow 4$. $(\mathbf{U}^*)^\mathsf{T}$ is $\boldsymbol{\Sigma}$-orthogonal if and only if $\mathbf{U}^*$ is $\boldsymbol{\Sigma}$-orthogonal.

$3 \Leftrightarrow 5$. $\mathbf{U}^\mathsf{T}$ is $\boldsymbol{\Sigma}$-orthogonal if and only if $\mathbf{U}$ is $\boldsymbol{\Sigma}$-orthogonal.

$4 \Rightarrow 5$. Multiply both sides of the equality in 4 from the left by $\mathbf{U}$ to complete the proof. □

Let $\boldsymbol{\beta}$ be a basis matrix of $\mathcal{S}_{Y|\boldsymbol{X}}$, and let $\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})$ be the projection onto $\text{span}(\boldsymbol{\beta})$ with respect to the inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{\Sigma}}$; that is,

$$\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) = \boldsymbol{\beta}(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}.$$

Let $\mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) = \mathbf{I}_p - \mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})$.

The next theorem we will try to prove requires the assumption that $\mathbf{U}$ is a

$\boldsymbol{\Sigma}$-orthogonal matrix that satisfies equation (9.5). To better understand this assumption note that every vector $\mathbf{a} \in \mathbb{R}^p$ can be written as $\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{a} + \mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{a}$. The $\boldsymbol{\Sigma}$-orthogonal matrix $\mathbf{U}$ only rotates the component $\mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{a}$, while leaving the component $\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{a}$ intact. In other words, the central subspace $\mathcal{S}_{Y|\boldsymbol{X}}$ acts as an "axis" around which $\mathbf{U}$ rotates. Thus, this next theorem basically shows that if $\boldsymbol{X}$ has an elliptical distribution with shape matrix $\boldsymbol{\Sigma}$, then the joint distribution of $(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}, Y)$ is unaffected by any rotation of $\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ around the central subspace. Here, we should emphasize that the rotation is relative to the inner product $\langle \cdot, \cdot \rangle_{\boldsymbol{\Sigma}}$.

Also note that, by multiplying both sides of (9.5) by $\mathbf{U}^*$ from the right we have $\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) = \mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{U}^*$. Since the projection $\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})$ is self-adjoint, this is equivalent to

$$\mathbf{U}\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) = \mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}). \tag{9.4}$$

Theorem 9.1.1 *Suppose that $\boldsymbol{X}$ has an elliptical distribution with shape matrix $\boldsymbol{\Sigma}$. Let $\mathbf{U}$ be a $\boldsymbol{\Sigma}$-orthogonal matrix that satisfies*

$$\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{U} = \mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}). \tag{9.5}$$

Then

$$(\boldsymbol{X}, Y) \overset{\mathcal{D}}{=} (\mathbf{U}^\intercal \boldsymbol{X}, Y), \quad (\boldsymbol{\Sigma}^{-1}\boldsymbol{X}, Y) \overset{\mathcal{D}}{=} (\mathbf{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}, Y). \tag{9.6}$$

Proof. Let $\phi_{\boldsymbol{X},Y}(\mathbf{t}, \tau)$ and $\phi_{\mathbf{U}^\intercal \boldsymbol{X},Y}(\mathbf{t}, \tau)$ be the characteristic functions for $(\boldsymbol{X}, Y)$ and $(\mathbf{U}^\intercal \boldsymbol{X}, Y)$. That is,

$$\phi_{\boldsymbol{X},Y}(\mathbf{t}, \tau) = E\left(e^{i(\mathbf{t}^\intercal \boldsymbol{X} + \tau Y)}\right), \quad \phi_{\mathbf{U}^\intercal \boldsymbol{X},Y}(\mathbf{t}, \tau) = E\left(e^{i(\mathbf{t}^\intercal \mathbf{U}^\intercal \boldsymbol{X} + \tau Y)}\right).$$

We will show that $\phi_{\boldsymbol{X},Y}(\mathbf{t},\tau) = \phi_{\mathbf{U}^\mathsf{T}\boldsymbol{X},Y}(\mathbf{t},\tau)$ for all $\mathbf{t} \in \mathbb{R}^p$ and $\tau \in \mathbb{R}$.

We first note that, if $Y, \mathbf{T}_1, \mathbf{T}_2$ are random elements such that $Y \perp\!\!\!\perp \mathbf{T}_1 | \mathbf{T}_2$ and $\mathbf{T}_2$ is measurable with respect to the σ-field generated by $\mathbf{T}_1$, then

$$\begin{aligned} E\left(e^{i\mathbf{t}^\mathsf{T}\mathbf{T}_1 + i\tau Y}\right) =& E\left[e^{i\mathbf{t}^\mathsf{T}\mathbf{T}_1} E\left(e^{i\tau Y}|\mathbf{T}_1\right)\right] \\ =& E\left[e^{i\mathbf{t}^\mathsf{T}\mathbf{T}_1} E\left(e^{i\tau Y}|\mathbf{T}_2\right)\right] = E\left[E\left(e^{i\mathbf{t}^\mathsf{T}\mathbf{T}_1}|\mathbf{T}_2\right) e^{i\tau Y}\right]. \end{aligned} \tag{9.7}$$

Take $\mathbf{T}_1 = \mathbf{U}^\mathsf{T}\boldsymbol{X}$ and $\mathbf{T}_2 = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$. Then

$$\phi_{\mathbf{U}^\mathsf{T}\boldsymbol{X},Y}(\mathbf{t},\tau) = E\left[E\left(e^{i\mathbf{t}^\mathsf{T}\mathbf{U}^\mathsf{T}\boldsymbol{X}}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}\right) e^{i\tau Y}\right]. \tag{9.8}$$

By assumption (9.5) and its consequence (9.4), the matrix $\mathbf{U}$ can be rewritten as

$$\mathbf{U} = [\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) + \mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})]\mathbf{U}[\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) + \mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})] = \mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) + \mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{U}\mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}). \tag{9.9}$$

Substitute this into (9.8) and use the fact that $\mathbf{P}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}$ is a measurable function of $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$ to obtain

$$\phi_{\mathbf{U}^\mathsf{T}\boldsymbol{X},Y}(\mathbf{t},\tau) = E\left[e^{i\mathbf{t}^\mathsf{T}\mathbf{P}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}} E\left(e^{i\mathbf{t}^\mathsf{T}\mathbf{Q}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{U}^\mathsf{T}\mathbf{Q}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}\right) e^{i\tau Y}\right]. \tag{9.10}$$

Because $\boldsymbol{X}$ has an elliptical distribution with shape matrix $\boldsymbol{\Sigma}$, and $\mathbf{U}$ is a $\boldsymbol{\Sigma}$-orthogonal matrix, we have $\boldsymbol{X} \overset{\mathcal{D}}{=} \mathbf{U}^\mathsf{T}\boldsymbol{X}$, and consequently,

$$\begin{aligned} [\mathbf{P}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}, \mathbf{Q}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}] &\overset{\mathcal{D}}{=} [\mathbf{P}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{U}^\mathsf{T}\boldsymbol{X}, \mathbf{Q}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{U}^\mathsf{T}\boldsymbol{X}] \\ &= [\mathbf{P}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}, \mathbf{Q}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\mathbf{U}^\mathsf{T}\mathbf{Q}^\mathsf{T}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}], \end{aligned}$$

where the second equality follows from (9.9). This implies

$$E\left(e^{i\mathbf{t}^\mathsf{T}\mathbf{Q}_{\boldsymbol{\beta}}^\mathsf{T}(\boldsymbol{\Sigma})\mathbf{U}^\mathsf{T}\mathbf{Q}_{\boldsymbol{\beta}}^\mathsf{T}(\boldsymbol{\Sigma})\boldsymbol{X}}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}\right) = E\left(e^{i\mathbf{t}^\mathsf{T}\mathbf{Q}_{\boldsymbol{\beta}}^\mathsf{T}(\boldsymbol{\Sigma})\boldsymbol{X}}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}\right).$$

Substitute this relation into the right hand side of (9.10) to obtain

$$\begin{aligned}\phi_{\mathbf{U}^\mathsf{T}\boldsymbol{X},Y}(\mathbf{t},\tau) =& E\left[e^{i\mathbf{t}^\mathsf{T}\mathbf{P}_{\boldsymbol{\beta}}^\mathsf{T}(\boldsymbol{\Sigma})\boldsymbol{X}}E\left(e^{i\mathbf{t}^\mathsf{T}\mathbf{Q}_{\boldsymbol{\beta}}^\mathsf{T}(\boldsymbol{\Sigma})\boldsymbol{X}}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}\right)e^{i\tau Y}\right]\\ =& E\left[E\left(e^{i\mathbf{t}^\mathsf{T}\boldsymbol{X}}|\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}\right)e^{i\tau Y}\right] = E\left(e^{i\mathbf{t}^\mathsf{T}\boldsymbol{X}+i\tau Y}\right) = \phi_{\boldsymbol{X},Y}(\mathbf{t},\tau),\end{aligned}$$

where, to obtain the third equality we have again evoked (9.7), taking $\mathbf{T}_1$ and $\mathbf{T}_2$ therein as $\boldsymbol{X}$ and $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}$. This proves the first equality in (9.6).

Multiply both sides of the first equality in (9.6) by $(\mathbf{U}^*)^\mathsf{T}$ from the left to obtain $\boldsymbol{X} \overset{\mathcal{D}}{=} (\mathbf{U}^*)^\mathsf{T}\boldsymbol{X}$. This is equivalent to the second equality in (9.6). □

The next lemma gives a sufficient condition under which the objective function $L(\psi, b, P_{\boldsymbol{X},\tilde{Y}})$ has a unique minimizer. A proof can be found in Jiang, Zhang, and Cai (2008).

Lemma 9.1.2 *Suppose that for any (ψ_1, b_1) and (ψ_2, b_2) in $\mathbb{R}^p \times \mathbb{R}$ we have*

$$P\{[1 - \tilde{Y}(\psi_1^\mathsf{T}\boldsymbol{X} - b_1)][1 - \tilde{Y}(\psi_2^\mathsf{T}\boldsymbol{X} - b_2)] < 0\} > 0. \tag{9.11}$$

Then $L(\psi, b, P_{\boldsymbol{X},\tilde{Y}})$ has a unique minimizer in $\mathbb{R}^p \times \mathbb{R}$.

Assumption (9.11) is quite mild in the context of sufficient dimension reduction. The following lemma provides some intuition about this condition in nonparametric regression.

Lemma 9.1.3 *Suppose that*

$$Y = f(\boldsymbol{X}) + \sigma(\boldsymbol{X})\varepsilon,$$

where $f : \Omega_{\boldsymbol{X}} \to \mathbb{R}$, $\sigma : \Omega_{\boldsymbol{X}} \to \mathbb{R}^+$ *are measurable functions,* σ *is bounded away from 0,* $\varepsilon \perp\!\!\!\perp \boldsymbol{X}$. *Suppose, furthermore, that for any nonempty open sets* $G_1 \in \mathbb{R}^p$ *and* $G_2 \in \mathbb{R}$, *we have* $P(\boldsymbol{X} \in G_1) > 0$ *and* $P(\varepsilon \in G_2) > 0$. *Let* $c \in \mathbb{R}$ *and* $\tilde{Y}$ *be as defined in (9.1) with* $\Omega_1 = (-\infty, c)$. *Then condition (9.11) is satisfied.*

Proof. Let G be a bounded, nonempty open set in $\mathbb{R}^p$, and $\sigma(\mathbf{x}) > \sigma_0 > 0$ for all $\mathbf{x} \in \Omega_{\boldsymbol{X}}$. Consider the case $\tilde{y} = -1$. We have

$$\begin{aligned} P(\boldsymbol{X} \in G, \tilde{Y} = -1) \leq & P(\{\boldsymbol{X} \in G\} \cap \{f(\boldsymbol{X}) + \sigma_0\varepsilon < c\}) \\ = & \int_{\mathbb{R}} P(\{\boldsymbol{X} \in G\} \cap \{f(\boldsymbol{X}) + \sigma_0\varepsilon < c\} | \varepsilon)\phi(\varepsilon)d\varepsilon, \end{aligned}$$

where ϕ is the density of ε. Since $\varepsilon \perp\!\!\!\perp \boldsymbol{X}$, the above can be rewritten as

$$\begin{aligned} & \int_{\mathbb{R}} P(\{\boldsymbol{X} \in G\} \cap \{f(\boldsymbol{X}) + \sigma_0\varepsilon < c\})\phi(\varepsilon)d\varepsilon \\ & \qquad \geq \int_{\epsilon < -K} P(\{\boldsymbol{X} \in G\} \cap \{f(\boldsymbol{X}) + \sigma_0\varepsilon < c\})\phi(\varepsilon)d\varepsilon, \end{aligned}$$

where K is any positive constant. Since any bounded open set is contained in $\{\mathbf{x} : f(\mathbf{x}) < c + \sigma_0 K\}$ for sufficiently large K, the right hand side above is no smaller than $P(\boldsymbol{X} \in G)P(\epsilon < -K)$, which by assumption is greater than 0.

Now

$$\begin{aligned} & P\{[1 - \tilde{Y}(\psi_1^\intercal \mathbf{X} - b_1)][1 - \tilde{Y}(\psi_2^\intercal \boldsymbol{X} - b_2)] < 0\} \\ & \geq P\{[1 + (\psi_1^\intercal \boldsymbol{X} - _1)][1 + (\psi_2^\intercal \boldsymbol{X} - b_2)] < 0, \tilde{Y} = -1\}. \end{aligned} \tag{9.12}$$

Obviously $\{\mathbf{x} : [1 + (\psi_1^\intercal \mathbf{x} - b_1)][1 + (\psi_2^\intercal \mathbf{x} - b_2)] < 0\}$ contains a nonempty open set. Hence the probability on the right hand side of (9.12) is positive. □

We now establish the unbiasedness of normal vector in support vector machine as an estimator of the central subspace.

Theorem 9.1.2 *Suppose that $\boldsymbol{X}$ has an elliptical distribution with shape matrix $\boldsymbol{\Sigma}$, and that the uniqueness condition (9.11) is satisfied. Let (ψ^*, b^*) be the minimizer of $L(\psi, b, P_{\boldsymbol{X}, \tilde{Y}})$ over $\mathbb{R}^p \times \mathbb{R}$. Then $\psi^* \in \mathcal{S}_{Y|\boldsymbol{X}}$.*

Proof. Let $\mathbf{U}$ be a $\boldsymbol{\Sigma}$-orthogonal matrix that satisfies condition (9.5). Then, by Theorem 9.1.1, $P_{\boldsymbol{X}, \tilde{Y}} = P_{\mathbf{U}^\intercal \boldsymbol{X}, \tilde{Y}}$. Hence

$$L(\psi, b, P_{\boldsymbol{X}, \tilde{Y}}) = L(\psi, b, P_{\mathbf{U}^\intercal \boldsymbol{X}, \tilde{Y}}).$$

By Theorem 9.0.1, the right hand side is the same as $L(\mathbf{U}\psi, b, P_{\boldsymbol{X}, \tilde{Y}})$. Hence

$$L(\psi, b, P_{\boldsymbol{X}, \tilde{Y}}) = L(\mathbf{U}\psi, b, P_{\boldsymbol{X}, \tilde{Y}}).$$

Let (ψ^*, b^*) be the minimizer of $L(\psi, b, P_{\boldsymbol{X}, \tilde{Y}})$. Since, by Lemma 9.1.2, the minimizer is unique, we have

$$\psi^* = \mathbf{U}\psi^*.$$

Take $\mathbf{U} = \mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) - \mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})$. It is easy to check that this $\mathbf{U}$ is $\boldsymbol{\Sigma}$-orthogonal and satisfies condition (9.5). Then,

$$\psi^* = [\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) - \mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})]\psi^*.$$

This implies $\mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\psi^* = -\mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\psi^*$. Hence $\mathbf{Q}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\psi^* = 0$, and consequently

$$\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\psi^* = \psi^*.$$

This equality means that ψ^* belongs to the range of the projection $\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})$, which is $\mathcal{S}_{Y|\boldsymbol{X}}$. □

9.2 Estimation procedure

Theorem 9.1.2 in the last section tells us that, if $\boldsymbol{X}$ has an elliptical distribution, then the normal vector of the optimal hyperplane that separates any pair of conditional distributions $P_{\boldsymbol{X}|Y\in\Omega_1}$ and $P_{\boldsymbol{X}|Y\in\Omega_2}$ lies in the central subspace $\mathcal{S}_{Y|\boldsymbol{X}}$. This motivates us to divide the support of Y into two slices, and apply support vector machine to sets of $\boldsymbol{X}$ corresponding to the two slices of Y. We repeat this process for several pairs of slices, and combines the optimal normal vectors by principal components to recover the central subspace.

We propose two ways to generate the set of pairs. One, which we call "left versus right" (LVR), divides the observed predictors into two parts according to whether their responses are fall above or below a set of numbers. The other, which we call "one versus another" (OVA), pairs up every possible pair slices in a partition of Ω_X determined by the values of Y. The particulars are summarized in the following procedure.

1. Center $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ to

$$\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i - \bar{\boldsymbol{X}}, \quad \text{where} \quad \bar{\boldsymbol{X}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i.$$

Compute the shape matrix $\hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\intercal}$.

2. (LVR) Let $q_r, r = 1, \ldots, h$ be h dividing points. For example, q_r can be the $(100/r)$th sample quartile of $\{Y_1, \ldots, Y_n\}$. For each r, apply support vector machine to the two collection of $\boldsymbol{X}$'s

$$\{\boldsymbol{X}_i : Y_i \leq q_r\}, \quad \{\boldsymbol{X}_i : Y_i > q_r\}. \tag{9.13}$$

This process gives h normal vectors $\hat{\psi}_1, \ldots, \hat{\psi}_h$.

2′. (OVA) Alternatively, we can apply support vector machine to the following pairs of slices

$$\{\boldsymbol{X}_i : q_{r-1} < Y_i \leq q_r\}, \quad \{\boldsymbol{X}_i : q_{s-1} < Y_i \leq q_s\}, \quad 2 \leq r < s \leq h.$$

This process gives $\binom{h}{2}$ normal vectors $\hat{\psi}_{rs}$, $2 \leq r < s \leq h$.

3. Let $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d$ be the eigenvectors of one of the matrices corresponding to LVR or OVA:

$$\sum_{r=1}^{h} \hat{\psi}_r \hat{\psi}_r^{\intercal} \quad \text{or} \quad \sum_{r=2}^{h} \sum_{s=r+1}^{h} \hat{\psi}_{rs} \hat{\psi}_{rs}^{\intercal}$$

corresponding to its d largest eigenvalues. We use subspace spanned by $\hat{\boldsymbol{v}} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d)$ to estimate the central subspace $\mathcal{S}_{Y|\boldsymbol{X}}$.

Based on our experiences, LVR works best when the response is a continuous

variable, where Y being larger or small has a concrete physical meaning; whereas OVA works the best when the response is categorical, where the values of Y are simply labels of different classes subjects, such as different types of proteins in our example in Chapter 11. The implementation of our method requires an algorithm for support vector machine. Two packages are widely available: e1071 (Demetriadou et al., 2010) and kernlab (Karatzoglou et al., 2009). The results shown in this work are based on the e1071 package but more on this will be discusses in Chapter 10.

9.3 Asymptotic analysis

In this section we derive the asymptotic distribution of $\hat{\boldsymbol{v}}$, defined in the last section. This is divided into two steps. First, we derive the influence function of the normal vector $\hat{\psi}$ of the support vector machine estimate based on two slices (9.13) for a generic dividing point q. This step is largely similar to the development of Jiang, Zhang, and Cai (2008), except for some differences in details. In the second step we develop the asymptotic distribution of the eigenvectors of $\hat{\boldsymbol{v}}$ from the influence function of $\hat{\psi}$. In this step we use a recent work of Bura and Pfeiffer (2008), which studies the asymptotic distribution of left singular vectors in a general setting.

9.3.1 Gradient of support vector machines

The asymptotic results Jiang, Zhang, and Cai (2008) are largely applicable to the current setting except for three places. First, our support vector machine involves the covariance matrix $\boldsymbol{\Sigma}$; Second, our C is fixed but the C_n in their paper depends on n; Third, they did not provide the explicit form of the Hessian matrix (and

hence neither the asymptotic variance) but we are interested in the asymptotic variance. Among these, the first two points are minor but the third point needs nontrivial additional work.

Let $\boldsymbol{\theta} = (\psi^\intercal, b)^\intercal$, $\boldsymbol{Z} = (\boldsymbol{X}^\intercal, Y)^\intercal$, $\boldsymbol{X}^* = (\boldsymbol{X}^\intercal, -1)^\intercal$, and $\boldsymbol{\Sigma}^* = \mathrm{diag}(\boldsymbol{\Sigma}, 0)$. Let

$$m(\boldsymbol{\theta}, \boldsymbol{Z}) = \psi^\intercal \boldsymbol{\Sigma} \psi + C[1 - Y(\boldsymbol{X}^\intercal \psi - b)]^+ = \boldsymbol{\theta}^\intercal \boldsymbol{\Sigma}^* \boldsymbol{\theta} - C(1 - \boldsymbol{\theta}^\intercal \boldsymbol{X}^* \tilde{Y})^+. \quad (9.14)$$

We need a coherent notation system for differentiation. Let $\mathbf{h} : \Omega_{\boldsymbol{Z}} \times \Theta \to \mathbb{R}^r$ be a function of $(\boldsymbol{\theta}, \boldsymbol{Z})$. We use $D_{\boldsymbol{\theta}}$ to denote a $(p+1)$-dimensional column vector of differential operators $(\partial/\partial\theta_1, \ldots, \partial/\partial\theta_{p+1})^\intercal$. Thus, $D_{\boldsymbol{\theta}}^\intercal \mathbf{h}(\boldsymbol{\theta}, \boldsymbol{Z})$ denote the $r \times (p+1)$ matrix whose (i, j) entry is $\partial h_i(\boldsymbol{\theta}, \boldsymbol{Z})/\partial\theta_j$. We use $D_{\boldsymbol{\theta}}^2$ to denote the $D_{\boldsymbol{\theta}} D_{\boldsymbol{\theta}}^\intercal$. Thus $D_{\boldsymbol{\theta}}^2 m(\boldsymbol{\theta}, \mathbf{z})$ is the $(p+1) \times (p+1)$ matrix whose (i, j)th entry is $\partial^2 m(\boldsymbol{\theta}, \mathbf{z})/\partial\theta_i \partial\theta_j$. For each $\boldsymbol{\theta} \in \Theta$, let $N_\theta(m)$ be the set of $\mathbf{z}$ for which a function $m(\mathbf{z}, \cdot)$ is not differentiable at $\boldsymbol{\theta}$. That is,

$$N_{\boldsymbol{\theta}}(m) = \{\mathbf{z} : m(\cdot, \mathbf{z}) \text{ is not differentiable at } \boldsymbol{\theta}\}.$$

Lemma 9.3.1 *Suppose that $m : \Omega_{\boldsymbol{Z}} \times \Theta \to \mathbb{R}$ satisfying the following conditions*

1. *(almost surely differentiable) for each $\boldsymbol{\theta} \in \Theta$, $P[\boldsymbol{Z} \in N_{\boldsymbol{\theta}}(m)] = 0$.*
2. *(Lipschitz condition) there is an integrable function $c(\mathbf{z})$, independent of $\boldsymbol{\theta}$, such that for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,*

$$|m(\boldsymbol{\theta}_2, \mathbf{z}) - m(\boldsymbol{\theta}_1, \mathbf{z})| \leq c(\mathbf{z}) \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|.$$

Then $D_{\boldsymbol{\theta}} m(\boldsymbol{Z}, \boldsymbol{\theta})$ is integrable, $Es(\boldsymbol{\theta}, \boldsymbol{Z})$ is differentiable, and

$$D_{\boldsymbol{\theta}} E[m(\boldsymbol{\theta}, \boldsymbol{Z})] = E[D_{\boldsymbol{\theta}} m(\boldsymbol{\theta}, \boldsymbol{Z})]. \tag{9.15}$$

This result is rather standard and its proof is omitted. Roughly, Assumption 1 guarantees that $E[D_{\boldsymbol{\theta}} m(\boldsymbol{\theta}, \boldsymbol{Z})]$ is defined; Assumption 2 allows us to apply the Dominated Convergence Theorem to bring a limit inside an integral. The next theorem gives the gradient of the support vector machine objective function $E[m(\boldsymbol{\theta}, \boldsymbol{Z})]$.

Theorem 9.3.1 *Suppose*

1. *for each $y \in \{-1, 1\}$, the distribution of $\boldsymbol{X}|Y = y$ is dominated by the Lebesgue measure,*
2. *$E(\|\boldsymbol{X}\|^2) < \infty$.*

Then

$$D_{\boldsymbol{\theta}} E[m(\boldsymbol{\theta}, \boldsymbol{Z})] = (2\psi^{\mathsf{T}}, 0)^{\mathsf{T}} - CE[\boldsymbol{X}^* Y I(1 - \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{X}^* Y > 0)]. \tag{9.16}$$

Proof. We first verify the two assumptions in Lemma 9.3.1. In our case,

$$P[(\boldsymbol{X}, Y) \in N_{\boldsymbol{\theta}}(m)] = \sum_{y \in \{-1,1\}} P(Y = y) P[\boldsymbol{X} \in H(\psi, b + y) | Y = y].$$

Since the Lebesgue measure of $H(\psi, b + y)$ is 0 for $y \in \{-1, 1\}$, by assumption 1 the above probability is 0. Thus condition 1 of Lemma 9.3.1 is satisfied.

Let $m_1(\boldsymbol{\theta}, \mathbf{z}) = \psi^{\mathsf{T}} \psi$ and $m_2(\boldsymbol{\theta}, \mathbf{z}) = [1 - Y(\psi^{\mathsf{T}} \boldsymbol{X} - b)]^+$. Then $m(\boldsymbol{\theta}, \mathbf{z})$ can be

written as $m_1(\boldsymbol{\theta}, \mathbf{z}) + Cm_2(\boldsymbol{\theta}, \mathbf{z})$. By the triangular inequality, it suffices to show that each m_i, $i = 1, 2$, is Lipschitz. The function m_1 is evidently Lipschitz. To verify that m_2 is Lipschitz, let $(\psi_1, b_1), (\psi_2, b_2) \in \mathbb{R}^{p+1}$. Then

$$m_2(\boldsymbol{\theta}_2, \mathbf{x}, y) - m_2(\boldsymbol{\theta}_1, \mathbf{x}, y) = [1 - y(\psi_2^\mathsf{T}\mathbf{x} - b_2)]^+ - [1 - y(\psi_1^\mathsf{T}\mathbf{x} - b_1)]^+.$$

Note that, for any two real numbers a_1 and a_2, $|a_2^+ - a_1^+| \leq |a_2 - a_1|$. Hence

$$\begin{aligned} |m_2(\boldsymbol{\theta}_2, \mathbf{x}, y) - m_2(\boldsymbol{\theta}_1, \mathbf{x}, y)| \leq & |\psi_1^\mathsf{T}\mathbf{x} - \psi_2^\mathsf{T}\mathbf{x} + b_2 - tb_1| \\ \leq & (1 + \|\mathbf{x}\|^2)^{\frac{1}{2}} \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|. \end{aligned}$$

Since $E(\|\boldsymbol{X}\|^2) < \infty$,

$$E(1 + \|\boldsymbol{X}\|^2)^{\frac{1}{2}} \leq [1 + E(\|\boldsymbol{X}\|^2)]^{\frac{1}{2}} < \infty. \tag{9.17}$$

This verifies condition 1 of Lemma 9.3.1.

Finally, by direct calculation, we find that, for $\mathbf{z} \notin N_{\boldsymbol{\theta}}(m)$,

$$\begin{aligned} D_\psi[m(\boldsymbol{\theta}, \mathbf{z})] = & 2\psi - C\mathbf{x}yI[1 - Y(\psi^\mathsf{T}\mathbf{x} - b) > 0], \\ D_b[m(\boldsymbol{\theta}, \mathbf{z})] = & CyI[1 - y(\psi^\mathsf{T}\mathbf{x} - b) > 0]. \end{aligned}$$

Hence

$$D_{\boldsymbol{\theta}}[m(\boldsymbol{\theta}, \mathbf{z})] = (2\psi^\mathsf{T}, 0)^\mathsf{T} - C\mathbf{x}^*yI(1 - \boldsymbol{\theta}^\mathsf{T}\mathbf{x}^*y > 0). \tag{9.18}$$

The corollary follows now from an application of Lemma 9.3.1. □

9.3.2 Hessian matrix for support vector machine

The next two Lemmas provide means of computing the derivative of an expectation of a non-Lipschitz function. Let $D_{\epsilon=0}$ denote the operation of first taking derivative with respect to ϵ and then evaluate the derivative at $\epsilon = 0$.

Lemma 9.3.2 *Suppose that U and V are random variables and $\mathbf{h}(u, v)$ is a measurable $\mathbb{R}^k$-valued function. Suppose, moreover,*

1. *the joint distribution of (U, V) is dominated by the Lebesgue measure;*
2. *for each v, the function $u \mapsto \mathbf{h}(u, v) f_{U|V}(u|v)$ is continuous;*
3. *for each component, say $h_i(u, v)$, of $\mathbf{h}(u, v)$, and any constant η, there is a function $c_i(v) \geq 0$ such that*

$$|(\eta - v)h_i(u, v)| f_{U|V}(u|v) \leq c_i(v), \quad E[c_i(V)] < \infty. \tag{9.19}$$

Then, for any constant a, the function $\epsilon \mapsto E[\mathbf{h}(U, V) I(U + \epsilon V < a + \epsilon\eta)]$ is differentiable at $\epsilon = 0$ with derivative

$$D_{\epsilon=0} E[\mathbf{h}(U, V) I(U + \epsilon V < a + \epsilon\eta)] = f_U(a) E[(\eta - V)\mathbf{h}(U, V)|U = a]. \tag{9.20}$$

Proof. We need to show that, for each $i = 1, \ldots, k$, the limit

$$\lim_{\epsilon \to 0} \int \left[\epsilon^{-1} \int_a^{a+\epsilon(\eta - v)} h(u, v) f_{U|V}(u|v) du \right] f_V(v) dv \tag{9.21}$$

exists and is equal to $f_U(a)E[(\eta - V)h_i(a,V)|U=a]$. By the mean value theorem and assumptions 2, 3, there is a $\xi \in (0,\epsilon)$ such that

$$\left|\epsilon^{-1}\int_a^{a-\epsilon v} h_i(u,v)f_{U|V}(u|v)du\right| = |h_i(a+\xi(\eta-v),v)f_{U|V}(a+\xi(\eta-v)|v)| \leq c(v).$$

Hence, by the Dominated Convergence Theorem, we can bring the limit in (9.21) to obtain

$$\begin{aligned}
&\int \lim_{\epsilon\to 0}\left[\epsilon^{-1}\int_a^{a+\epsilon(\eta-v)} h_i(u,v)f_{U|V}(u|v)du\right] f_V(v)dv \\
&\qquad = (\eta - v)\int h_i(a,v)f_{U|V}(a|v)du f_V(v)dv \\
&\qquad = f_U(a)\int (\eta-v)h_i(a,v)f_{V|U}(v|a)dv = f_U(a)E[(\eta-V)h_i(a,V)|U=a],
\end{aligned}$$

as desired. □

We also need to deal with the case where U and V are linearly independent. In this case $f_{U|V}$ is degenerate and the assumptions in Lemma 9.3.2 are not satisfied. Nevertheless, the formula (9.20) still holds, as shown in the next lemma.

Lemma 9.3.3 *Suppose that U and V are linearly dependent random variables and $\mathbf{h}(u)$ is a measurable $\mathbb{R}^k$-valued function.*

1. *the joint distribution of (U,V) is dominated by the Lebesgue measure;*
2. *$\mathbf{h}(u)f_U(u)$ is continuous.*

Then, for any constant a, the function $\epsilon \mapsto E[\mathbf{h}(U)I(U+\epsilon V < a+\epsilon\eta)]$ is differentiable at $\epsilon = 0$ with derivative given by (9.20).

PROOF. Suppose, without loss of generality, $V = \kappa U$ for some $\kappa > 0$. We have

$$E[h_i(U)I(U + \epsilon V < a + \epsilon\eta)] = \int_{-\infty}^{(a+\epsilon\eta)/(1+\epsilon\kappa)} E[h_i(U)|U = u]f_U(u)du.$$

Hence

$$D_{\epsilon=0}E[h_i(U)I(U + \epsilon V < a + \epsilon\eta)] = (\eta - \kappa a)E[h_i(U)|U = a]f_U(a).$$

Under the condition that $U = a$ and $V = \kappa U$, the right hand becomes (9.20). □

Theorem 9.3.2 *Suppose that $\boldsymbol{X}$ has a convex and open support and the distributions of $\boldsymbol{X}|Y = y$ for $y = -1, 1$ are dominated by the Lebesgue measure. Suppose, moreover:*

1. *for any linearly independent $\psi, \boldsymbol{\delta} \in \mathbb{R}^p$, $y \in \{-1, 1\}$, and $v \in \mathbb{R}$, the function*

$$u \mapsto E(\boldsymbol{X}^*|\psi^\top \boldsymbol{X} = u, \boldsymbol{\delta}^\top \boldsymbol{X} = v, Y = y)f_{\psi^\top \boldsymbol{X}|\boldsymbol{\delta}^\top \boldsymbol{X}, Y}(u|v, y)$$

   is continuous.

2. *For any $i = 1, \ldots, p$, and $y = -1, 1$, there is a nonnegative function $c_i(v, y)$ with $E[c_i(V, Y)|Y = y] < \infty$ such that*

$$vE(X_i|\psi^\top \boldsymbol{X} = u, \boldsymbol{\delta}^\top \boldsymbol{X} = v, Y = y)f_{\psi^\top \boldsymbol{X}|\boldsymbol{\delta}^\top \boldsymbol{X}, Y}(u|v, y) \leq c_i(v, y).$$

3. *For any $y = -1, 1$ there is a nonnegative function $c_0(v, y)$ with $E[c_0(V, Y)|Y =$*

$y] < \infty$ *such that*

$$f_{\psi^\mathsf{T}\boldsymbol{X}|\boldsymbol{\delta}^\mathsf{T}\boldsymbol{X},Y}(u|v,y) \le c_0(v,y) \quad \textit{and} \quad E[c(V,Y)|Y=y] \le \infty.$$

Then the function $\boldsymbol{\theta} \mapsto D_{\boldsymbol{\theta}} E[m(\boldsymbol{\theta},\boldsymbol{Z})]$ *is Gateaux differentiable with Gateaux derivative*

$$2\mathrm{diag}(\boldsymbol{\Sigma},0) + C\sum_{y=-1,1} P(Y=y) f_{\psi^\mathsf{T}\boldsymbol{X}|Y}(b+y|y) E(\boldsymbol{X}^*\boldsymbol{X}^{*\mathsf{T}}|\psi^\mathsf{T}\boldsymbol{X}=b+y). \quad (9.22)$$

Furthermore, if the function $(\psi,b) \mapsto f_{\psi^\mathsf{T}\boldsymbol{X}|Y}(b+y|y)E(\boldsymbol{X}^*\boldsymbol{X}^{*\mathsf{T}}|\psi^\mathsf{T}\boldsymbol{X}=b+y)$ *is continuous, then* $D_{\boldsymbol{\theta}}[m(\boldsymbol{\theta},\boldsymbol{Z})]$ *is differentiable with derivative matrix (9.22).*

PROOF. We need to show that the function (9.16) is differentiable with respect to (ψ,b). The first term is obviously differentiable with derivative $2\mathrm{diag}(\boldsymbol{\Sigma},0)$. Thus only need to consider the differentiability of $E[\boldsymbol{X}^*YI(1-\boldsymbol{\theta}^\mathsf{T}\boldsymbol{X}^*Y>0)]$, which is

$$\sum_{y=-1,1} P(Y=y)E[\boldsymbol{X}^*yI(1-\boldsymbol{\theta}^\mathsf{T}\boldsymbol{X}^*Y>0)|Y=y].$$

First, consider the case $y=1$ and verify Gateaux differentiability of the function $(\psi,b) \mapsto E[\boldsymbol{X}^*I(\psi^\mathsf{T}\boldsymbol{X} < b+1)|Y=1]$.

Let ψ and $\boldsymbol{\delta}$ be linearly independent vectors in $\mathbb{R}^p$. Let η be a number. Thus $(\boldsymbol{\delta}^\mathsf{T},\eta)^\mathsf{T}$ is an arbitrary vector in $\mathbb{R}^{p+1}$. The directional derivative along $(\boldsymbol{\delta}^\mathsf{T},\eta)^\mathsf{T}$ is the derivative of the following function with respect to ϵ at $\epsilon=0$:

$$\begin{aligned} &E[\boldsymbol{X}^*I(\psi^\mathsf{T}\boldsymbol{X}+\epsilon\boldsymbol{\delta}^\mathsf{T}\boldsymbol{X} < b+1+\epsilon\,\eta)|Y=1] \\ &\qquad = E[E(\boldsymbol{X}^*|\psi^\mathsf{T}\boldsymbol{X},\boldsymbol{\delta}^\mathsf{T}\boldsymbol{X},Y=1)I(\psi^\mathsf{T}\boldsymbol{X}+\epsilon\boldsymbol{\delta}^\mathsf{T}\boldsymbol{X} < b+1+\epsilon\,\eta)|Y=1]. \end{aligned}$$

Let $U = \psi^{\mathsf{T}}\boldsymbol{X}$, $V = \boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{X}$, $\mathbf{h}(U,V) = E(\boldsymbol{X}^*|U,V)$, $a = b+1$. Then, by Lemma 9.3.2, as applied to the probability measure $P(\,\cdot\,|Y=1)$, the above derivative is

$$\begin{aligned} & f_{\psi^{\mathsf{T}}\boldsymbol{X}|Y}(b+1|1)E[(\eta - V)E(\boldsymbol{X}^*|U,V)|U = b+1] \\ & \qquad = f_{\psi^{\mathsf{T}}\boldsymbol{X}|Y}(b+1|1)E[(\eta - V)\boldsymbol{X}^*|U = b+1]. \end{aligned}$$

Since this holds for all $(\boldsymbol{\delta}^{\mathsf{T}}, \eta)^{\mathsf{T}}$, the function $(\psi, b) \mapsto E[\boldsymbol{X}^* I(\psi^{\mathsf{T}}\boldsymbol{X} < b+1)|Y=1]$ is Gateaux differentiable with Gateaux derivative

$$-f_{\psi^{\mathsf{T}}\boldsymbol{X}|Y}(b+1|1)E(\boldsymbol{X}^*\boldsymbol{X}^{*\mathsf{T}}|\psi^{\mathsf{T}}\boldsymbol{X} = b+1, Y=1) \tag{9.23}$$

If $\boldsymbol{\delta}$ and ψ are linearly dependent, then $\psi^{\mathsf{T}}\boldsymbol{X}$ and $\psi^{\mathsf{T}}\boldsymbol{X}$ are linearly independent. We apply Lemma 9.3.3 in the similar fashion to arrive at the same Gateaux derivative (9.23).

The case for $y = -1$ can be proved similarly. Hence the Gateaux derivative of $E_{\boldsymbol{\theta}}E[m(\boldsymbol{\theta}, \boldsymbol{Z})]$ is given by (9.22). If $f_{\psi^{\mathsf{T}}\boldsymbol{X}|Y}(b+y|y)E(\boldsymbol{X}^*\boldsymbol{X}^{*\mathsf{T}}|\psi^{\mathsf{T}}\boldsymbol{X} = b+y)$ is continuous then the Gateaux derivative is continuous, and hence $D_{\boldsymbol{\theta}}E[m(\boldsymbol{\theta}, \boldsymbol{Z})]$ is differentiable (see, for example, Bickel, Klaassen, Ritov, and Wellner, 1993, page 453). □

9.3.3 Influence function for support vector machine

Theorem 9.3.3 *If the conditions in Theorems 9.3.1 and 9.3.2 are satisfied, then*

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - \mathbf{H}^{-1}\{(2\psi_0^{\mathsf{T}}\boldsymbol{\Sigma}, 0)^{\mathsf{T}} - CE_n[\boldsymbol{X}^* Y I(1 - Y\boldsymbol{\theta}_0^{\mathsf{T}}\boldsymbol{X}^* > 0)]\} + o_P(n^{-\frac{1}{2}}),$$

where $\mathbf{H}$ *is given by (9.22).*

The proof is similar to that of Jiang, Zhang, and Cai (2008) and is therefore omitted. Alternatively, one can apply Theorem 5.23 of van der Vaart (1998), which provides three sufficient conditions for asymptotic normality of an M-estimates with non-differentiable objective function: (i) $P((\boldsymbol{X}, Y) \in N_{\boldsymbol{\theta}}(m)) = 0$; (ii) a Lipschitz condition, and (iii) the population-level objective function has a second-order Taylor approximation. The first condition in Theorem 9.3.1 guarantees condition (i); the second condition in Theorem 9.3.1 guarantees condition (ii), and the conclusions of Theorems 9.3.1 and 9.3.2 guarantee condition (iii).

9.4 Nonlinear dimension reduction

In this section we outline an extension of the SVMIR to the sufficient nonlinear dimension reduction. Here, we are interested in estimating functions $\phi_1(\boldsymbol{X}), \ldots, \phi_d(\boldsymbol{X})$ such the following one that appeared in Cook (2007)

$$Y \perp\!\!\!\perp \boldsymbol{X} | \phi_1(\boldsymbol{X}), \ldots, \phi_d(\boldsymbol{X}). \tag{9.24}$$

We assume that $\phi_1, \ldots, \phi_d$ belong to a finite- or infinite-dimensional Hilbert space $\mathcal{H}$. Let $\{u_i : i = 1, 2, \ldots\}$ be a basis of $\mathcal{H}$, and let $\beta_{\ell i}$ be sequences such that $\phi_\ell = \sum_{i=1}^{\infty} \beta_{\ell i} u_i$. Then problem (9.24) becomes estimating $\{\beta_{\ell i} : i = 1, 2, \ldots, \ell = 1, \ldots, d\}$ in the relation

$$Y \perp\!\!\!\perp \boldsymbol{X} | \sum_{i=1}^{\infty} \beta_{1i}\, u_i(\boldsymbol{X}), \ldots, \sum_{i=1}^{\infty} \beta_{di}\, u_i(\boldsymbol{X}).$$

One can see that this is a generalization of problem (7.1), with $\boldsymbol{X}$ replaced by the

sequence of feature functions $\{u_i(\boldsymbol{X}) : i = 1, 2, \ldots\}$, and the matrix $\boldsymbol{\beta}$ replaced by the finite or infinite dimensional array $\{\beta_{\ell i}\}$.

The SVMIR can be naturally extended to this setting using the kernel SVM, which, at the population level can be described as follows. Let $K : \Omega_{\boldsymbol{X}} \times \Omega_{\boldsymbol{X}} \to \mathbb{R}$ be a positive definite bivariate function. We assume that $\mathcal{H}$ to be the reproducing kernel Hilbert space generated by K. That is, $\mathcal{H}$ is the closed linear span of functions of the form

$$c_1 K(\cdot, \mathbf{x}_1) + \cdots + c_m K(\cdot, \mathbf{x}_m), \quad c_1, \ldots, c_m \in \mathbb{R}, \quad \mathbf{x}_1, \ldots, \mathbf{x}_m \in \Omega_{\boldsymbol{X}}$$

equipped with the inner product determined by $\langle K(\cdot, \mathbf{x}_1), K(\cdot, \mathbf{x}_2)\rangle = K(\mathbf{x}_1, \mathbf{x}_2)$, $\mathbf{x}_1, \mathbf{x}_2 \in \Omega_{\boldsymbol{X}}$. Consider the following bilinear form from $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$:

$$(f_1, f_2) \mapsto \mathrm{cov}[f_1(\boldsymbol{X}), f_2(\boldsymbol{X})].$$

This induces a positive semi-definite and self adjoint operator $\boldsymbol{\Sigma} : \mathcal{H} \to \mathcal{H}$ such that

$$\langle f_1, \boldsymbol{\Sigma} f_2\rangle_{\mathcal{H}} = \langle \boldsymbol{\Sigma} f_1, f_2\rangle_{\mathcal{H}} = \mathrm{cov}[f_1(\boldsymbol{X}), f_2(\boldsymbol{X})].$$

We replace the vector $\boldsymbol{\psi}$ in (9.3) by a member $\boldsymbol{\psi}$ of $\mathcal{H}$, and the random vector $\boldsymbol{X}$ by the random function $\mathbf{x} \mapsto K(\boldsymbol{X}, \mathbf{x}) - E[K(\boldsymbol{X}, \mathbf{x})]$ so that the objective function (9.3) gets generalized in the infinite dimensional setting. Thus the inner product $\langle \boldsymbol{\psi}, \boldsymbol{X}\rangle$ in (9.3) is replaced by

$$\langle K(\boldsymbol{X}, \cdot) - EK(\boldsymbol{X}, \cdot), \boldsymbol{\psi}\rangle_{\mathcal{H}} = \boldsymbol{\psi}(\boldsymbol{X}) - E\boldsymbol{\psi}(\boldsymbol{X}),$$

where $\boldsymbol{\psi}$ is a member of $\mathcal{H}$, and $K(\boldsymbol{X}, \cdot)$ denotes the mapping $\mathbf{x} \mapsto E[K(\boldsymbol{X}, \mathbf{x})]$, and $\boldsymbol{\psi}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\psi} = \mathrm{var}(\boldsymbol{\psi}^\mathsf{T}\boldsymbol{X})$ in (9.3) is replaced by

$$\mathrm{var}(\boldsymbol{\psi}, \langle K(\boldsymbol{X}, \cdot) - EK(\boldsymbol{X}, \cdot)\rangle_\mathcal{H}) = \mathrm{var}[\boldsymbol{\psi}(\boldsymbol{X})] = \langle \boldsymbol{\psi}, \boldsymbol{\Sigma}\boldsymbol{\psi}\rangle_\mathcal{H}.$$

The objective function (9.3) then becomes

$$\langle \boldsymbol{\psi}, \boldsymbol{\Sigma}\boldsymbol{\psi}\rangle_\mathcal{H} + CE[1 - Y(\boldsymbol{\psi}(\boldsymbol{X}) - E\psi(\boldsymbol{X}) - b)]^+. \tag{9.25}$$

At the sample level, the covariance operator is defined by the bilinear form

$$(f_1, f_2) = \mathrm{cov}_n[f_1(\boldsymbol{X}), f_2(\boldsymbol{X})];$$

That is, $\langle f_1, \boldsymbol{\Sigma}_n f_2\rangle_\mathcal{H} = \langle \boldsymbol{\Sigma}_n f_1, f_2\rangle_\mathcal{H} = \mathrm{cov}_n[f_1(\boldsymbol{X}), f_2(\boldsymbol{X})]$. The function $\boldsymbol{\psi}$ is an arbitrary linear combination $\sum_{\mu=1}^n c_\mu\{K(\cdot, \boldsymbol{X}_\mu) - E_n[K(\cdot, \boldsymbol{X})]\}$, where $c_1, \ldots, c_n \in \mathbb{R}$. If we let $\boldsymbol{\kappa}_n(\cdot)$ denote the vector-valued function

$$\mathbf{x} \mapsto [K(\mathbf{x}, \boldsymbol{X}_1) - E_n K(\mathbf{x}, \boldsymbol{X}), \ldots, K(\mathbf{x}, \boldsymbol{X}_n) - E_n K(\mathbf{x}, \boldsymbol{X})],$$

then $\boldsymbol{\psi}$ can be written as $\mathbf{c}^\mathsf{T}\boldsymbol{\kappa}_n$. It is easy to see that

$$\langle \boldsymbol{\psi}, \boldsymbol{\Sigma}_n\boldsymbol{\psi}\rangle_\mathcal{H} = \mathrm{var}_n[\mathbf{c}^\mathsf{T}\boldsymbol{\kappa}_n(\boldsymbol{X})] = \mathbf{c}^\mathsf{T}\mathbf{G}_n\mathbf{c},$$

where $\mathbf{c} = (c_1, \ldots, c_n)^\mathsf{T}$ and $\mathbf{G}_n$ is the $p \times p$ Gram matrix whose (i, j)th entry is

$$K(\boldsymbol{X}_i, \boldsymbol{X}_j) - E_n[K(\boldsymbol{X}_i, \boldsymbol{X})] - E_n[K(\boldsymbol{X}_j, \boldsymbol{X})] + E_n[K(\boldsymbol{X}, \boldsymbol{X}')],$$

where quantities such as $E_n[K(\boldsymbol{X}_i, \boldsymbol{X})]$ and $E_n[K(\boldsymbol{X}, \boldsymbol{X}')]$ denotes the sample

means $n^{-1}\sum_{\mu=1}^{n} K(\boldsymbol{X}_j, \boldsymbol{X}_\mu)$ and $n^{-2}\sum_{\mu=1}^{n}\sum_{\nu=1}^{n} K(\boldsymbol{X}_\mu, \boldsymbol{X}_\nu)$. Thus the sample-level counterpart of objective function (9.25) is

$$\mathbf{c}^\mathsf{T}\mathbf{G}_n\mathbf{c} + CE_n[1 - Y(\mathbf{c}^\mathsf{T}\boldsymbol{\kappa}_n(\boldsymbol{X}) - b)]^+. \tag{9.26}$$

The construction of the objective function (9.26) determines that it is invariant under translation; that is, if $\mathbf{1}_n$ is the n-dimensional vector $(1, \ldots, 1)^\mathsf{T}$, then the above function is unchanged if we replace $\mathbf{c}$ by $\mathbf{c} + \tau\mathbf{1}_n$ for any constant τ. This is reflected in the fact that $\mathbf{1}_n^\mathsf{T}\mathbf{G}_n\mathbf{1}_n = 0$ and $\boldsymbol{\kappa}_n^\mathsf{T}\mathbf{1}_n = 0$. We minimize (9.26) over $\tilde{\mathbb{R}}^n \times \mathbb{R}$, where $\tilde{\mathbb{R}}^n$ is the orthogonal complement of the vector $\mathbf{1}_n$. Let $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{n-1}$ be the eigenvectors of $\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\mathsf{T}/n$ corresponding to its nonzero eigenvalues, and let $\boldsymbol{\Xi}_n$ be the $n \times (n-1)$ matrix $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{n-1})$. The vector $\mathbf{c}$ can be parameterized as $\mathbf{c} = \boldsymbol{\Xi}_n^\mathsf{T}\mathbf{d}$, where $\mathbf{d} \in \mathbb{R}^{n-1}$ is a free parameter. In this parametrization objective function (9.26) reduces to

$$\mathbf{d}^\mathsf{T}\tilde{\mathbf{G}}_n\mathbf{d} + CE_n[1 - Y(\mathbf{d}^\mathsf{T}\tilde{\boldsymbol{\kappa}}_n(\boldsymbol{X}) - b)]^+,$$

where $\tilde{\mathbf{G}}_n = \boldsymbol{\Xi}_n^\mathsf{T}\mathbf{G}_n\boldsymbol{\Xi}_n$ and $\tilde{\boldsymbol{\kappa}}_n = \boldsymbol{\Xi}_n^\mathsf{T}\boldsymbol{\kappa}_n$.

For the LVR scheme, we minimize the objective function (9.26) for h pairs of slices to obtain $\hat{\mathbf{d}}_1, \ldots \hat{\mathbf{d}}_h$. Let $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d$ be the first d eigenvectors of the matrix $\sum_{i=1}^{h} \hat{\mathbf{d}}_\ell\hat{\mathbf{d}}_\ell^\mathsf{T}$. We use

$$\hat{\boldsymbol{v}}_1^\mathsf{T}\tilde{\boldsymbol{\kappa}}_n(\cdot), \ldots, \hat{\boldsymbol{v}}_d^\mathsf{T}\tilde{\boldsymbol{\kappa}}_n(\cdot)$$

as the estimate of sufficient predictors $\phi_1, \ldots, \phi_d$ in (9.24). For the OVA scheme, we obtain $\binom{h}{2}$ vectors $\hat{\mathbf{d}}_{rs} : 1 \le r \le s \le p$. The vectors $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d$ are the eigenvectors of $\sum_{1\le r\le s\le p} \hat{\mathbf{d}}_{rs}\hat{\mathbf{d}}_{rs}^\mathsf{T}$.

Chapter 10

Simulation Results

In this Chapter we run some simulation analysis to compare our method with other methods surrounding the idea of inverse regression, such as SIR, SAVE and DR. We label our method as SVMIR to indicate the combination of the support vector machine and the idea of inverse regression.

10.1 Description

To evaluate the performance of each method we use the Frobenius norm (Golub and van Loan, 1996; page 55) of the difference between the projections on to the estimated and the true central subspaces. Specifically, let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of $\mathbb{R}^p$. Then

$$\text{dist}(\mathcal{S}_1, \mathcal{S}_2) = \|P_{\mathcal{S}_1} - P_{\mathcal{S}_2}\|, \tag{10.1}$$

where $P_{\mathcal{S}_1}$ and $P_{\mathcal{S}_2}$ are the orthogonal projections on to $\mathcal{S}_1$ and $\mathcal{S}_2$. This distance was used in Li, Zha, and Chiaromonte (2005).

We present the results for the following 5 models. The first three are taken from (Li, 1991), and the other two from the Li, Zha, and Chiaromonte (2005):

Model 1: $Y = X_1 + X_2 + X_3 + X_4 + \sigma\varepsilon$,

Model 2: $Y = X_1/[0.5 + (X_2 + 1)^2] + \sigma\varepsilon$,

Model 3: $Y = X_1(X_1 + X_2 + 1) + \sigma\varepsilon$,

Model 4: $Y = (3/2)\sin(X_1 + X_2 + X_3) + (3/4)\sin(X_1 + X_5 + 3X_6) + \sigma\varepsilon$,

Model 5: $Y = \sin^2(\pi X_2 + 1) + \sigma\varepsilon$.

In the above models, $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_p)$, where $p = 10$, and $\varepsilon \sim N(0, 1)$. The sample sizes n for all 5 models are taken to be 200. We use $h = 4$ dividing points, located at the 20th, 40th, 60th, 80th sample percentiles of $Y_1, \ldots, Y_n$. The misclassification penalty constant C is taken to be 1. The results for $\sigma = 0.2,\ 0.5,\ 1$ are presented in Table 10.1. The entries are of the form $a \pm b$ where a is the mean and b is the standard deviation of the distance criterion (10.1) calculated from 500 simulated samples.

10.2 Compare performance

We can see from Table 10.1 that in most models and for all the values of parameter σ the SVMIR compares at least as good as the best of the the other methods. Model 4 seems to that SAVE and DR perform slightly better than our method. The reason for this is due to the fact that the periodicity of the sin function favors

Table 10.1: Comparison of different methods in five models

Models	σ	SVMIR	SIR	SAVE	DR
1	0.2	0.09±0.021	0.11±0.030	0.13±0.035	0.11±0.029
	0.5	0.12±0.028	0.13±0.031	0.16±0.052	0.14±0.035
	1	0.20±0.050	0.19±0.050	0.34±0.173	0.21±0.052
2	0.2	0.50±0.118	0.55±0.127	1.41±0.143	0.68±0.167
	0.5	0.83±0.193	0.90±0.202	1.70±0.164	1.09±0.243
	1	1.33±0.189	1.30±0.197	1.72±0.156	1.44±0.219
3	0.2	0.77±0.223	0.92±0.218	1.35±0.151	0.82±0.244
	0.5	0.97±0.238	1.07±0.233	1.34±0.166	1.01±0.261
	1	1.18±0.218	1.29±0.218	1.36±0.163	1.21±0.227
4	0.2	1.29±0.148	1.32±0.124	1.06±0.252	1.06±0.249
	0.5	1.31±0.130	1.33±0.128	1.08±0.274	1.09±0.269
	1	1.32±0.129	1.33±0.111	1.17±0.234	1.17±0.230
5	0.2	1.34±0.101	1.33±0.104	1.34±0.103	1.34±0.104
	0.5	1.33±0.123	1.34±0.087	1.33±0.099	1.33±0.103
	1	1.33±0.100	1.33±0.111	1.33±0.112	1.33±0.113

the fact that SAVE and DR use second moments to find the directions that span the central dimension reduction subspace.

10.3 Robustness

10.3.1 Outliers with difference covariance matrix

One of the most important features of SVMIR is its robustness against outliers in the predictor. In the separable case, the optimal separating hyperplane is completely determined by a few support vectors, and is not affected by the rest of the predictor. This is also true, though to a lesser degree, for the non-separable case. The robustness property can also be seen from the influence function of SVMIR, as appear in Theorem 9.3.3 in section 9.3.3, which resembles the influence functions of quantiles or the median. To demonstrate the robustness of SVMIR we also

Table 10.2: Comparison of different methods for five models when predictors contain outliers with different covariance matrix

Models	σ	SVMIR	SIR	SAVE	DR
1	0.2	0.14±0.035	0.70±0.152	1.39±0.040	1.07±0.276
	0.5	0.15±0.042	0.70±0.150	1.39±0.036	1.15±0.253
	1	0.21±0.054	0.69±0.168	1.39±0.032	1.18±0.227
2	0.2	0.97±0.260	1.26±0.200	1.80±0.125	1.71±0.151
	0.5	1.26±0.228	1.39±0.188	1.80±0.119	1.76±0.134
	1	1.48±0.155	1.55±0.163	1.79±0.126	1.77±0.130
3	0.2	1.34±0.238	1.52±0.203	1.57±0.184	1.53±0.174
	0.5	1.40±0.233	1.58±0.191	1.58±0.173	1.53±0.148
	1	1.52±0.218	1.63±0.181	1.58±0.182	1.55±0.165
4	0.2	1.42±0.109	1.58±0.119	1.75±0.131	1.72±0.142
	0.5	1.43±0.113	1.58±0.122	1.74±0.134	1.73±0.138
	1	1.43±0.118	1.57±0.124	1.75±0.140	1.74±0.145
5	0.2	1.34±0.099	1.34±0.090	1.33±0.105	1.33±0.100
	0.5	1.34±0.095	1.34±0.099	1.33±0.102	1.33±0.099
	1	1.33±0.105	1.34±0.102	1.33±0.100	1.33±0.105

introduce a small contamination to the predictor $\boldsymbol{X}$ in the above models. That is,

$$\boldsymbol{X} \sim (1-\epsilon)N(\mathbf{0}, \mathbf{I}_p) + \epsilon N(\mathbf{0}, 100\mathbf{I}_p),$$

where $\epsilon = 0.05$. The results are presented in Table 10.2 where its obvious that in the presence of outliers SVMIR performs better than previously proposed methods.

In Table 10.3 we present the results of the simulation we run to see if the value of the variance of the outliers has an effect on the performance of our algorithm. It seems that there is a small effect when we increase the variance from 5 to 50, but there is not much effect beyond value 50 (or at least the effect is very small). Of course, variance equal to 5 might as well not give you many outliers, that's why in future runs when we want to put outliers with different variance we choose a value of 100.

Table 10.3: Comparison of the effect different values of the variance for the outliers, have on the performance of all methods for all five models when predictors contain outliers with different covariance matrix

Models	*var*	SVMIR	SIR	SAVE	DR
1	5	0.19±0.045	0.21±0.053	0.98±0.406	0.23±0.060
	50	0.19±0.051	0.55±0.146	1.39±0.032	1.04±0.280
	100	0.21±0.054	0.69±0.168	1.39±0.032	1.18±0.227
	200	0.24±0.063	0.81±0.163	1.38±0.042	1.27±0.179
2	5	1.31±0.206	1.30±0.205	1.76±0.142	1.57±0.194
	50	1.42±0.192	1.48±0.180	1.79±0.125	1.76±0.132
	100	1.48±0.155	1.55±0.163	1.79±0.126	1.77±0.130
	200	1.53±0.186	1.61±0.165	1.79±0.120	1.78±0.125
3	5	1.18±0.207	1.32±0.219	1.40±0.169	1.28±0.201
	50	1.41±0.213	1.56±0.194	1.57±0.169	1.52±0.155
	100	1.52±0.218	1.63±0.181	1.58±0.182	1.55±0.165
	100	1.57±0.215	1.68±0.176	1.61±0.168	1.58±0.165
4	5	1.32±0.128	1.35±0.121	1.38±0.211	1.28±0.190
	50	1.41±0.106	1.51±0.120	1.76±0.132	1.70±0.149
	100	1.43±0.118	1.57±0.124	1.75±0.140	1.74±0.145
	200	1.48±0.121	1.63±0.126	1.75±0.140	1.74±0.142
5	5	1.33±0.103	1.34±0.094	1.34±0.101	1.34±0.099
	50	1.34±0.098	1.34±0.091	1.33±0.103	1.33±0.106
	100	1.33±0.105	1.34±0.102	1.33±0.100	1.33±0.105
	200	1.34±0.097	1.34±0.103	1.33±0.101	1.33±0.101

10.3.2 Outliers with different mean

In this section we show that SVMIR is still robust in the case that the outliers have different mean instead of different variance So we introduce a small contamination to the predictor $\boldsymbol{X}$ as follows

$$\boldsymbol{X} \sim (1-\epsilon)N(\mathbf{0}, \mathbf{I}_p) + \epsilon N(10 \times \mathbf{1}_p, \mathbf{I}_p),$$

where $\epsilon = 0.05$ and $\mathbf{1}_p$ is a p-dimensional vector with all its entries equal to 1. The results are presented on Table 10.1 which show that SVMIR is robust to this type

Table 10.4: Comparison of different methods for five models when predictors contain outliers with different mean

Models	σ	SVMIR	SIR	SAVE	DR
1	0.2	0.39±0.050	0.82±0.019	1.35±0.026	0.40±0.082
	0.5	0.41±0.051	0.82±0.021	1.35±0.027	0.47±0.086
	1	0.43±0.067	0.82±0.024	1.36±0.027	0.65±0.096
2	0.2	0.66±0.120	0.80±0.092	1.61±0.159	1.13±0.246
	0.5	1.02±0.198	1.08±0.174	1.73±0.148	1.34±0.203
	1	1.41±0.179	1.39±0.171	1.73±0.147	1.56±0.182
3	0.2	0.85±0.220	0.97±0.187	1.48±0.111	1.40±0.050
	0.5	1.04±0.225	1.10±0.216	1.49±0.111	1.42±0.038
	1	1.22±0.197	1.29±0.205	1.53±0.122	1.43±0.045
4	0.2	1.48±0.097	1.60±0.106	1.60±0.144	1.49±0.148
	0.5	1.49±0.097	1.61±0.103	1.59±0.163	1.48±0.158
	1	1.51±0.097	1.61±0.095	1.66±0.165	1.52±0.138
5	0.2	1.34±0.106	1.33±0.111	1.34±0.105	1.34±0.098
	0.5	1.33±0.096	1.34±0.095	1.33±0.097	1.33±0.101
	1	1.34±0.092	1.34±0.097	1.34±0.103	1.34±0.101

of outliers. Actually for models 2 and 3 we can see that it is more robust to outliers in mean than outliers in variance. For models 1 and 4 it is not as robust in outliers in mean as it is in outliers for variance.

Another advantage that this results show is the fact the if there is a small divergence from the elliptical distribution for the predictors, SVMIR is robust to this divergence. In order to explore further the effect of the mean, we run the analysis for different values of the mean. We use, mean, **2**, **10**, **20**, **50**, and we can see from the resutls in Table 10.5 that as long as we increase the distance between the outliers mean and the true mean of the dataset, there is a decrease in the performance as expected. Interestingly in some cases, SAVE performs better as we increase the mean. This is probably due to the fact that SAVE uses only second moments and actually the more different the variances are from slice to slice the easiest will be to capture the direction, which is the case, if we have outliers much

Table 10.5: Comparison of how different outlier mean values affect the performance of all methods

Models	mean	SVMIR	SIR	SAVE	DR
1	2	0.20±0.048	0.33±0.057	1.21±0.145	0.42±0.091
	10	0.43±0.067	0.82±0.024	1.36±0.027	0.65±0.096
	20	0.71±0.047	0.88±0.022	1.23±0.035	0.61±0.129
	100	0.93±0.022	0.93±0.020	1.05±0.040	0.68±0.212
2	2	1.34±0.193	1.22±0.191	1.74±0.140	1.51±0.198
	10	1.41±0.179	1.39±0.171	1.73±0.147	1.56±0.182
	20	1.42±0.164	1.38±0.170	1.72±0.152	1.56±0.183
	100	1.43±0.175	1.39±0.172	1.74±0.141	1.57±0.168
3	2	1.15±0.214	1.23±0.227	1.51±0.127	1.41±0.085
	10	1.22±0.197	1.29±0.205	1.53±0.122	1.43±0.045
	20	1.23±0.188	1.30±0.206	1.50±0.135	1.44±0.042
	100	1.29±0.180	1.34±0.183	1.46±0.130	1.47±0.060
4	2	1.37±0.094	1.42±0.121	1.38±0.215	1.32±0.186
	10	1.51±0.097	1.61±0.095	1.66±0.165	1.52±0.138
	20	1.64±0.096	1.63±0.101	1.59±0.177	1.59±0.119
	100	1.64±0.097	1.64±0.099	1.64±0.156	1.64±0.108
5	2	1.34±0.098	1.34±0.097	1.34±0.090	1.34±0.089
	10	1.34±0.092	1.34±0.097	1.34±0.103	1.34±0.101
	20	1.34±0.095	1.34±0.092	1.33±0.102	1.33±0.104
	100	1.34±0.093	1.35±0.088	1.33±0.103	1.34±0.103

further away from the mean.

10.4 Robust covariance matrix

There is a technical detail that we incorporate in our algorithm. In all the algorithms we need to estimate the covariance matrix of our predictor $\text{cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Since the classical estimator for $\boldsymbol{\Sigma}$ is not robust to outliers, it is shown in our results that the SVMIR method doesn't perform as well as we expected it to work when we used this estimator in the presence of outliers. In order to improve performance in the presence of outliers we also used the idea of using a robust estimator for

Table 10.6: Comparison of different methods in five models using the robust covariance estimator

Models	σ	SVMIR	SIR	SAVE	DR
1	0.2	0.09±0.021	0.17±0.051	0.17±0.171	0.38±0.151
	0.5	0.12±0.028	0.19±0.055	0.22±0.209	0.42±0.160
	1	0.20±0.09	0.23±0.067	0.48±0.361	0.55±0.225
2	0.2	0.50±0.118	0.58±0.131	1.43±0.161	1.23±0.246
	0.5	0.83±0.194	0.92±0.203	1.72±0.165	1.54±0.184
	1	1.33±0.189	1.30±0.195	1.73±0.150	1.69±0.164
3	0.2	0.77±0.222	0.93±0.218	1.38±0.144	1.34±0.197
	0.5	0.97±0.237	1.08±0.229	1.38±0.147	1.40±0.166
	1	1.17±0.218	1.30±0.216	1.40±0.153	1.45±0.147
4	0.2	1.29±0.147	1.32±0.123	1.16±0.233	1.32±0.150
	0.5	1.31±0.131	1.33±0.132	1.16±0.244	1.31±0.160
	1	1.32±0.126	1.33±0.112	1.23±0.210	1.34±0.138
5	0.2	1.34±0.106	1.34±0.102	1.34±0.091	1.34±0.085
	0.5	1.33±0.112	1.34±0.089	1.34±0.100	1.34±0.103
	1	1.33±0.103	1.33±0.109	1.33±0.110	1.33±0.105

$\mathbf{\Sigma}$. We choose the estimator proposed by Rousseeuw (1985). In our results we show all four methods MLIR, SIR, SAVE, DR using the robust estimator for the covariance. One can see that all the methods perform similarly as the case when we had the classical estimator for the covariance matrix, in the case that there are no outliers in the sample. Those results are presented in Table 10.6. When there are outliers in the sample, we can see in Table 10.7 that the results for the SVMIR there is an increase in the performance for model 1, while for the rest of the models, the performance is similar as with the classical estimators. For the other methods, SIR, SAVE, DR, there is a decrease in performance when we use the robust covariance matric operator. For the robust estimation of the covariance matrix, there is a tuning parameter that we said equal to 0.75 and it represents the percentage of points we use in finding the covariance estimator in each step of the algorithm

Table 10.7: Comparison of different methods for five models when predictors contain outliers and we are using the robust covariance estimator

Models	σ	SVMIR	SIR	SAVE	DR
1	0.2	0.10±0.022	0.99±0.077	1.14±0.069	1.11±0.71
	0.5	0.13±0.031	0.99±0.079	1.14±0.067	1.11±0.070
	1	0.20±0.049	1.00±0.075	1.14±0.067	1.11±0.069
2	0.2	0.61±0.138	1.20±0.93	1.77±0.110	1.62±0.152
	0.5	1.00±0.207	1.32±0.123	1.75±0.110	1.71±0.203
	1	1.41±0.179	1.48±0.121	1.76±0.115	1.75±0.117
3	0.2	0.81±0.233	1.36±0.099	1.49±0.108	1.40±0.076
	0.5	1.01±0.240	1.45±0.133	1.51±0.113	1.45±0.038
	1	1.21±0.206	1.56±0.144	1.55±0.133	1.48±0.101
4	0.2	1.44±0.087	1.42±0.059	1.57±0.133	1.52±0.113
	0.5	1.45±0.093	1.42±0.067	1.57±0.135	1.53±0.114
	1	1.47±0.097	1.42±0.063	1.59±0.132	1.54±0.117
5	0.2	1.34±0.102	1.34±0.052	1.34±0.036	1.34±0.036
	0.5	1.34±0.096	1.34±0.049	1.34±0.035	1.34±0.035
	1	1.33±0.096	1.34±0.048	1.34±0.037	1.34±0.037

10.5 Dimension of predictors

In this section we are investigating how the dimension of the predictor might affect the performance of SVMIR. We run the same analysis as before with $\sigma = 1$ and we increased the dimension of the predictor to $p = 50$ and $p = 100$. For comparison purposes we report together with them the results for $p = 10$. The results are reported in Table 10.8 for the data without outliers and in Table 10.9 for the case when outliers are included. As we can see from those results, increasing the dimension of the predictors doesn't affect the performance of SVMIR, as in most of the cases is either the best method or at least comparable to the best of the other 3 methods.

Table 10.8: Comparison for all methods for all models with different dimension of predictors without outliers

Models	p	SVMIR	SIR	SAVE	DR
1	10	0.20±0.050	0.19±0.050	0.34±0.173	0.21±0.052
	50	0.54±0.057	0.49±0.054	1.41±0.007	0.57±0.066
	100	0.87±0.074	0.76±0.075	1.41±0.006	0.84±0.071
2	10	1.33±0.189	1.30±0.197	1.72±0.156	1.44±0.219
	50	1.77±0.071	1.78±0.072	1.96±0.029	1.90±0.065
	100	1.91±0.034	1.90±0.041	1.98±0.013	1.93±0.037
3	10	1.18±0.218	1.29±0.218	1.36±0.163	1.21±0.227
	50	1.74±0.095	1.78±0.092	1.88±0.091	1.69±0.077
	100	1.93±0.032	1.91±0.049	1.97±0.026	1.89±0.051
4	10	1.32±0.129	1.33±0.111	1.17±0.234	1.17±0.230
	50	1.47±0.029	1.46±0.028	1.97±0.027	1.48±0.038
	100	1.60±0.037	1.56±0.030	1.98±0.012	1.66±0.040
5	10	1.33±0.100	1.33±0.111	1.33±0.112	1.33±0.113
	50	1.40±0.018	1.40±0.020	1.40±0.020	1.40±0.017
	100	1.41±0.010	1.41±0.009	1.41±0.011	1.41±0.009

10.6 Misclassification Penalty

In this section we check what happens if we change the value of the misclassification penalty C in our function. We run our SVMIR algorithm in case there are no outliers, outliers to the mean and outliers in variance. As we can see in Table 10.10 there are not many differences although it seems that especially when there are outliers, there is a slight increase in the performance on some models as we increase the value of C. This increase though is very small. The only one that seems to have a significant increase in performance as we increase the value of C is the first model. This might be due to the fact that model 1 it's the simplest and has only one direction in the Central Dimension Reduction Subspace.

Table 10.9: Comparison for all methods for all models with different dimension of predictors with outliers

Models	p	SVMIR	SIR	SAVE	DR
1	10	0.43±0.067	0.82±0.024	1.36±0.027	0.65±0.096
	50	0.68±0.054	0.65±0.051	1.11±0.060	0.61±0.069
	100	0.95±0.069	0.86±0.065	1.29±0.048	0.92±0.071
2	10	1.41±0.179	1.39±0.171	1.73±0.147	1.56±0.182
	50	1.78±0.070	1.79±0.073	1.96±0.029	1.91±0.061
	100	1.92±0.033	1.91±0.040	1.98±0.013	1.93±0.036
3	10	1.22±0.197	1.29±0.205	1.53±0.122	1.43±0.045
	50	1.75±0.095	1.77±0.089	1.76±0.062	1.77±0.080
	100	1.93±0.032	1.90±0.052	1.91±0.045	1.90±0.050
4	10	1.51±0.097	1.61±0.095	1.66±0.165	1.52±0.138
	50	1.56±0.034	1.54±0.029	1.96±0.032	1.57±0.038
	100	1.67±0.042	1.63±0.034	1.98±0.012	1.66±0.041
5	10	1.34±0.092	1.34±0.097	1.34±0.103	1.34±0.101
	50	1.40±0.018	1.40±0.019	1.40±0.022	1.40±0.022
	100	1.41±0.011	1.41±0.009	1.41±0.010	1.41±0.010

10.7 Number of slices

We are also interested in learning how the number of slices will affect the performance of SVMIR. We do a simulation analysis, where we run the same 5 models, with the same parameters as they were described in section 10.1. Our results are shown in Table 10.11 where if we exclude models 4 and 5 there is an increase in performance of our algorithm as we increase the number of slices. Most of the increase comes up to number of slices around 10 for most models and then the increase is really small and insignificant.

Table 10.10: Performance of SVMIR for different values of misclassification penalty C with and without outliers. "with (v)" means there are outliers with different variance and "with (m)" means there are outliers with difference mean

Models	*outliers*	$C = 0.1$	$C = 0.5$	$C = 1$	$C = 2$	$C = 10$
1	without	0.19±0.045	0.19±0.045	0.20±0.050	0.20±0.050	0.19±0.050
	with (v)	0.39±0.109	0.24±0.063	0.21±0.054	0.19±0.049	0.19±0.050
	with (m)	0.80±0.032	0.57±0.061	0.43±0.067	0.33±0.068	0.22±0.056
2	without	1.32±0.196	1.33±0.189	1.33±0.189	1.33±0.197	1.29±0.205
	with (v)	1.49±0.171	1.49±0.163	1.48±0.155	1.48±0.169	1.47±0.178
	with (m)	1.42±0.159	1.42±0.164	1.41±0.179	1.42±0.161	1.40±0.185
3	without	1.17±0.216	1.17±0.209	1.18±0.218	1.16±0.204	1.17±0.216
	with (v)	1.52±0.210	1.50±0.217	1.52±0.218	1.50±0.208	1.48±0.203
	with (m)	1.20±0.199	1.20±0.204	1.22±0.197	1.18±0.205	1.19±0.216
4	without	1.33±0.122	1.32±0.125	1.32±0.129	1.31±0.138	1.32±0.124
	with (v)	1.46±0.121	1.44±0.114	1.43±0.118	1.43±0.116	1.43±0.116
	with (m)	1.59±0.099	1.53±0.103	1.51±0.097	1.49±0.107	1.46±0.095
5	without	1.34±0.101	1.35±0.091	1.33±0.100	1.34±0.094	1.34±0.092
	with (v)	1.33±0.103	1.34±0.096	1.33±0.105	1.34±0.105	1.34±0.097
	with (m)	1.34±0.102	1.34±0.102	1.34±0.092	1.35±0.085	1.34±0.102

Table 10.11: Performance of SVMIR for different values of misclassification penalty C with and without outliers. "with (v)" means there are outliers with different variance and "with (m)" means there are outliers with difference mean

Models	*outliers*	2	5	10	25	50
1	without	0.31±0.073	0.20±0.050	0.17±0.041	0.16±0.040	0.16±0.039
	with (v)	0.31±0.078	0.21±0.054	0.19±0.051	0.19±0.050	0.18±0.045
	with (m)	0.49±0.082	0.43±0.067	0.43±0.062	0.43±0.062	0.43±0.061
2	without	1.42±0.178	1.33±0.189	1.25±0.216	1.17±0.237	1.13±0.228
	with (v)	1.52±0.160	1.48±0.155	1.47±0.183	1.43±0.192	1.43±0.198
	with (m)	1.49±0.155	1.41±0.179	1.38±0.198	1.30±0.210	1.27±0.203
3	without	1.42±0.178	1.18±0.218	1.15±0.212	1.09±0.231	1.07±0.243
	with (v)	1.57±0.179	1.52±0.218	1.51±0.201	1.52±0.192	1.52±0.204
	with (m)	1.45±0.153	1.22±0.197	1.20±0.210	1.17±0.224	1.14±0.226
4	without	1.20±0.234	1.32±0.129	1.32±0.119	1.33±0.119	1.33±0.112
	with (v)	1.37±0.192	1.43±0.118	1.42±0.125	1.42±0.116	1.42±0.119
	with (m)	1.51±0.170	1.51±0.097	1.34±0.089	1.48±0.090	1.48±0.085
5	without	1.33±0.099	1.33±0.100	1.34±0.093	1.34±0.097	1.34±0.090
	with (v)	1.34±0.104	1.34±0.105	1.34±0.100	1.34±0.087	1.34±0.092
	with (m)	1.34±0.090	1.34±0.092	1.34±0.089	1.33±0.103	1.34±0.097

Chapter 11

Data Analysis

In this Chapter we perform the analysis of two datasets found in the UC Irvine depository (see Asuncion and Newman (2007)). The first dataset builds a regression model for the performance of computer hardware and the second is a classification problem of *E.coli* genes. We run the analysis for, SIR, SAVE, DR, SVMIR.

11.1 Computer Hardware

This dataset was first presented in Ein-Dor and Feldmesser (1987) and the objective of the authors is to create a regression model that estimates relative performance of the Central Processing Unit (CPU) of a computer using some of its characteristics, including cache memory size, cycle time, minimum and maximum input/output channels and minimum and maximum main memory. Relative performance was calculated using observations from users of different machines in the market. For machines not in the market the relative performance was not able to be calculated. The authors recognized that, collected a data of 209 models in the market in 1987

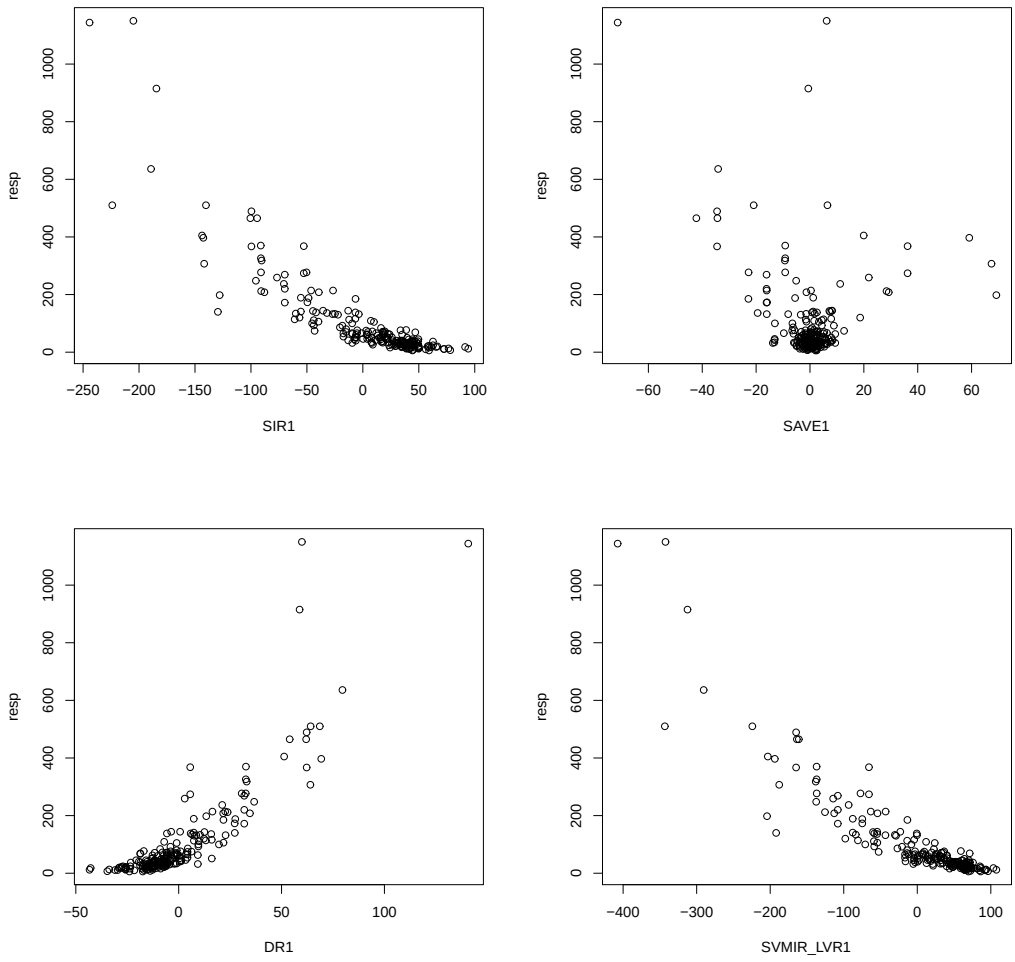

Figure 11.1: First direction for SIR and SAVE in the upper panel, DR and SVMIR in the lower panel.

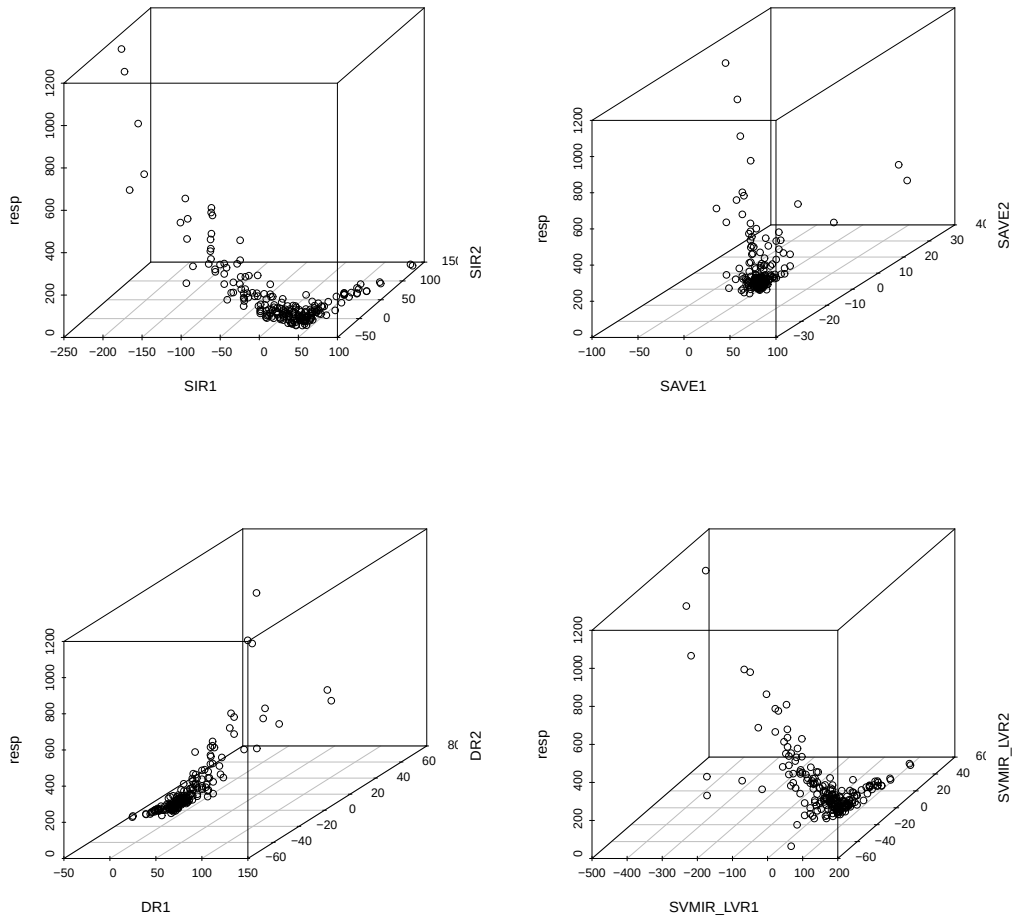

Figure 11.2: First direction for SIR and SAVE in the upper panel, DR and SVMIR in the lower panel.

and build a regression model for those machines.

Figure 11.1 shows the first direction for all the methods. It is clear that SAVE is the only method that fail to capture anything, while all other methods capture a nonlinear trend of the points. This nonlinear trend agrees with the comments of Ein-Dor and Feldmesser (1987) who proposed a linear model with response the square root of the relative performance. This is because interactions among the size of the main memory, the size of cache memory, the machine cycle time and the number of input/output channels affect the performance of the CPU. There are configurations that make communications among the components of the CPU faster, while other configurations are not as effective.

In Figure 11.2 one can see the the 3d plots with the second direction as well. We can see than SIR and SVMIR they have the nonlinear trend in the first direction and in the second direction there is a division among the points that have smaller relative frequency. We can see that this is not viewable using the DR plot and of course not in SAVE.

In this example since we have a continuous response, we used the LVR method for comparison between slices. OVA performs very similar.

11.2 *E.coli* Protein Dataset

11.2.1 Full dataset - all categories

This dataset was constructed and presented first in Horton and Nakai (1996). There are 336 proteins and are classified in 8 categories, based on the 7 predictors, as follows:

Table 11.1: Categories, number of data and color used in graphs for the *E.coli* dataset

Class	Number of points	Color in graphs
cytoplasm	143	black
inner membrane without signal sequence	77	red
perisplasm	52	light blue
inner membrane, uncleavable signal sequence	35	yellow
outer membrane	20	purple
outer membrane lipoprotein	5	grey
inner membrane lipoprotein	2	blue
inner membrane, cleavable signal sequence	2	green

- McGeoch's method for signal sequence recognition.
- von Heijne's method for signal sequence recognition.
- von Heijne's Signal Peptidase II consensus sequence score. Binary attribute.
- Presence of charge on N-terminus of predicted lipoproteins. Binary attribute.
- score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.
- score of the ALOM membrane spanning region prediction program.
- score of ALOM program after excluding putative cleavable signal regions from the sequence.

The 8 categories with the number of datapoints in each categories as well as the colors in the plots are shown in Table 11.1.

Since the response has no real ordering we found it useful to run the OVA method of comparing slices for SVMIR, in addition to LVR.

The dataset has two variables (the two binary attributes) that 326 out of the 336 points have the same values ("von Heijne's Signal Peptidase II consensus se-

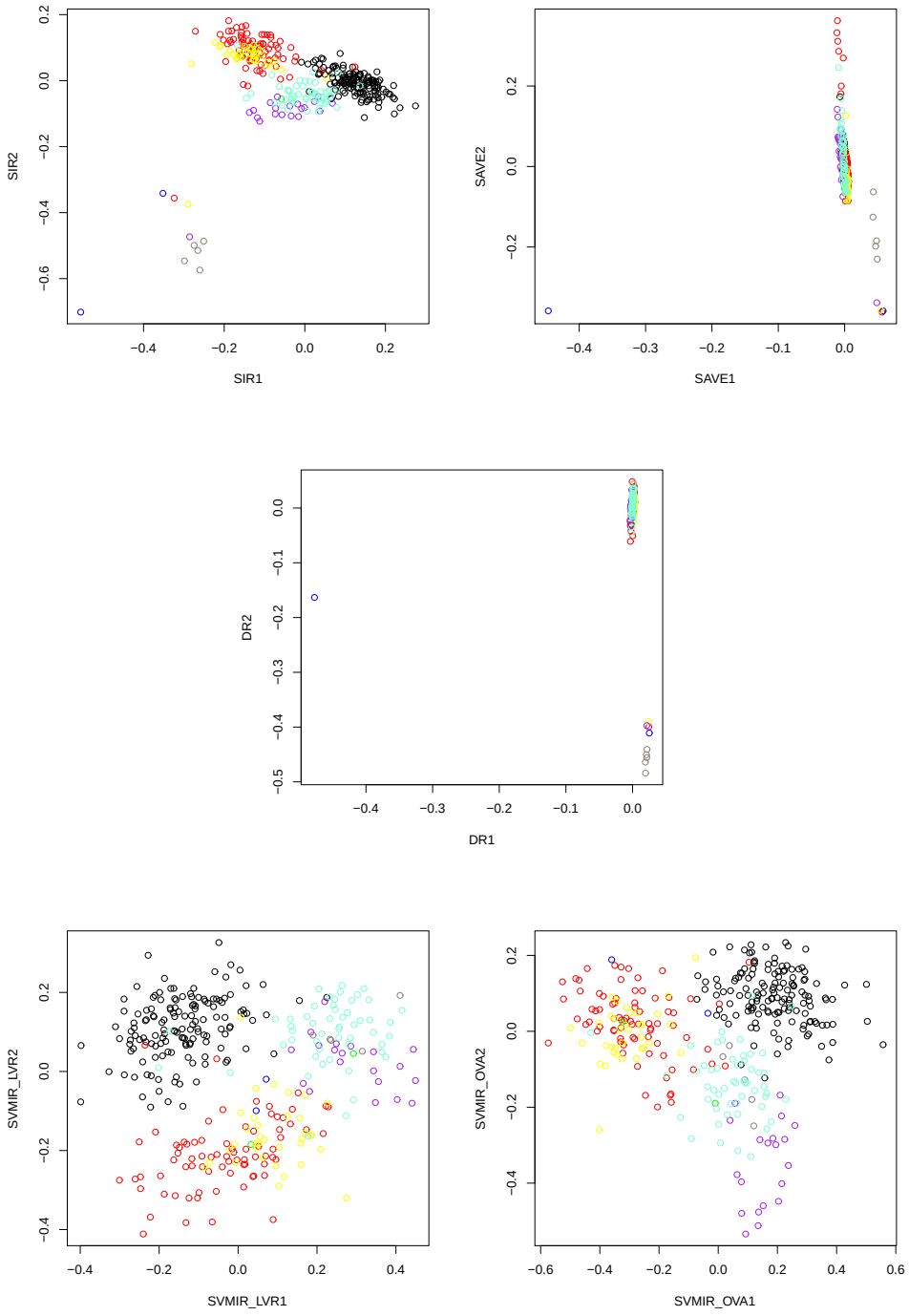

Figure 11.3: The first two directions for all the methods in the full *E.coli* dataset analysis. Upper panel is SIR and SAVE, middle panel is DR and in the lower panel is SVMIR using both LVR and OVA method of estimating the directions.

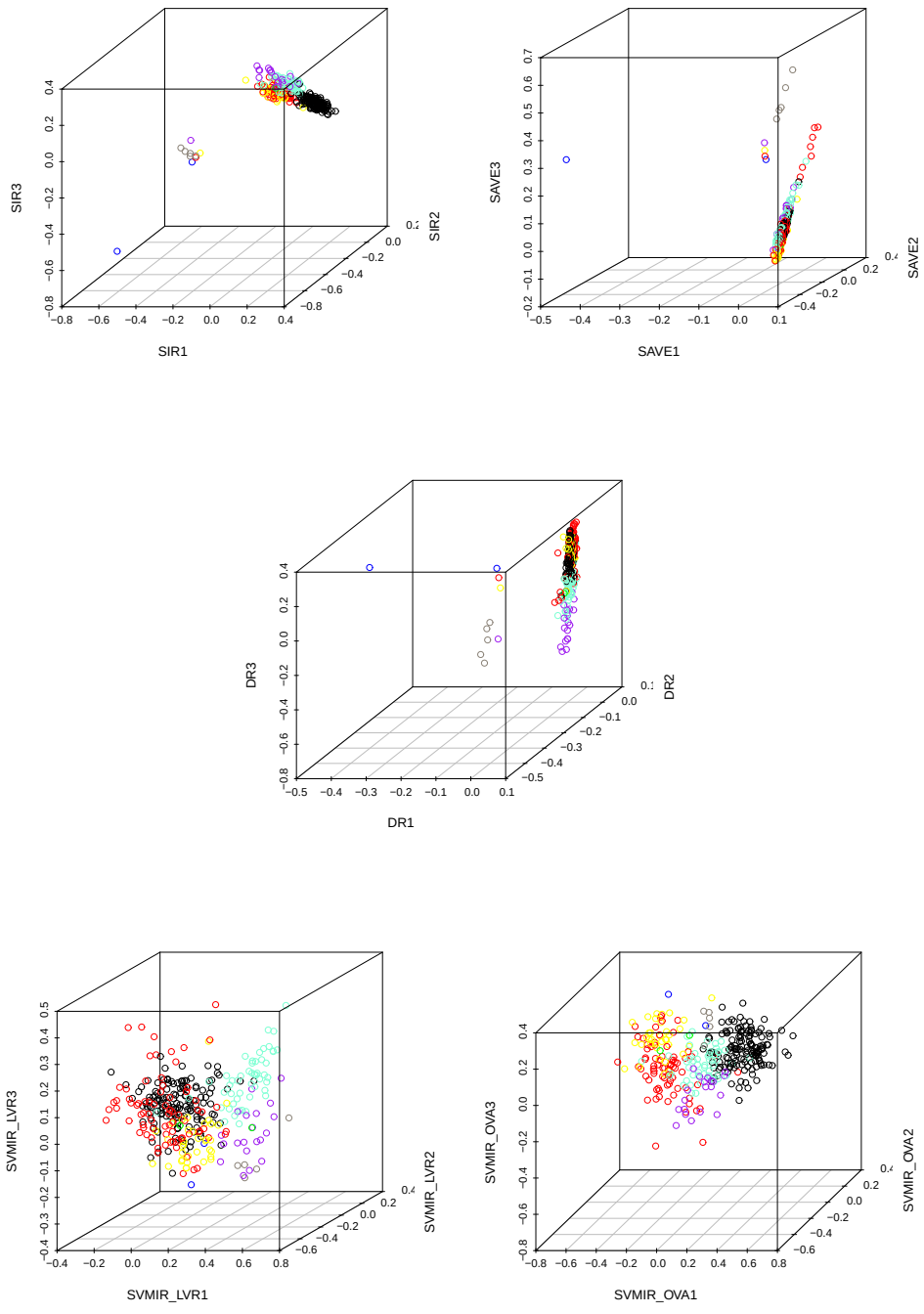

Figure 11.4: The first three directions for all the methods in the full *E.coli* dataset analysis. Upper panel is SIR and SAVE, middle panel is DR and in the lower panel is SVMIR using both LVR and OVA method of estimating the directions.

quence score"=0.48 and "Presence of charge on N-terminus of predicted lipoproteins=0.5"). There are nine points that gets values (1.00, 0.50) and one point with values (1.00, 1.00). It is clear from Figure 11.3 that the first two directions in SIR, SAVE and DR is highly driven by those two variables. SIR achieves some separation in the rest of the points, where in one cluster are the cytoplasm cells (black), the second cluster has "the inner membrane without signal sequence cells" (red) and the "inner membrane, uncleavable signal sequence cells" (yellow) and finally the third cluster has the "outer membrane cells" (purple) and the "periplasm cells" (light blue). This separation makes sense because cytoplasm is inside the membrane, then we have in one group two different cells in the inner membrane (red and yellow points) and then there is another group with the outer membrane cells and the periplasm cells (periplasm is the area between the inner membrane and the outer membrane).

SVMIR with LVR and OVA comparisons achieve the same type of separation as SIR without being driven from those ten points that have different values in the two binary attributes. The five "outer membrane lipoprotein cells" (grey) are in the same cluster with the other outer membrane cells and periplasm cells (purple and light blue, respectively) and the two "inner membrane lipoprotein" cells (blue) and the two "inner membrane, cleavable signal sequence" cells (green) are grouped with the other inner membrane cells. The separation is slightly better with the LVR comparison, but with OVA we are able to capture 4 clusters because the outer "membrane cells" (purple) and the "periplasm cells" (light blue) are more clearly divided with this method. Also with the OVA method, in the third direction we get a clearer separation of the cluster that has the inner membrane cells, since "the inner membrane without signal sequence cells" (red) have mostly negative values in the third direction and the "inner membrane, uncleavable signal sequence cells"

(yellow) have mostly positive cluster in the third direction (see Figure 11.4).

11.2.2 *E.coli* without the binary variables

When we remove the binary attributes, SIR and DR perform similar to SVMIR, although they seem to slightly underperform the SVMIR-OVA method. See Figures 11.5 and 11.6

11.2.3 Removing 3 small categories

Here we remove blue, green and grey points. If we remove this 9 points, one of the binary attributes ("Presence of charge on N-terminus of predicted lipoproteins") have the same value for the rest of the points, so we remove that attribute as well as well. So now we have 327 points, with 6 predictors, in 5 categories for the response.

There are 3 points now that have different value in the binary attribute. In Figures 11.7 and 11.8 we can see that SAVE's first direction is greatly influenced by those 3 points, Also DR's third direction is influenced by those three points. SVMIR we showed before is robust, in the presence of binary variables. On the other hand, SIR now that we have only one binary variable, seems to perform pretty good.

11.2.4 *E.coli*, no small categories, not binary attributes

In the last section here, we remove the last binary attribute. So now we have the 327 points with 5 categories in the response, with 5 predictors (excluding the binary attributes).

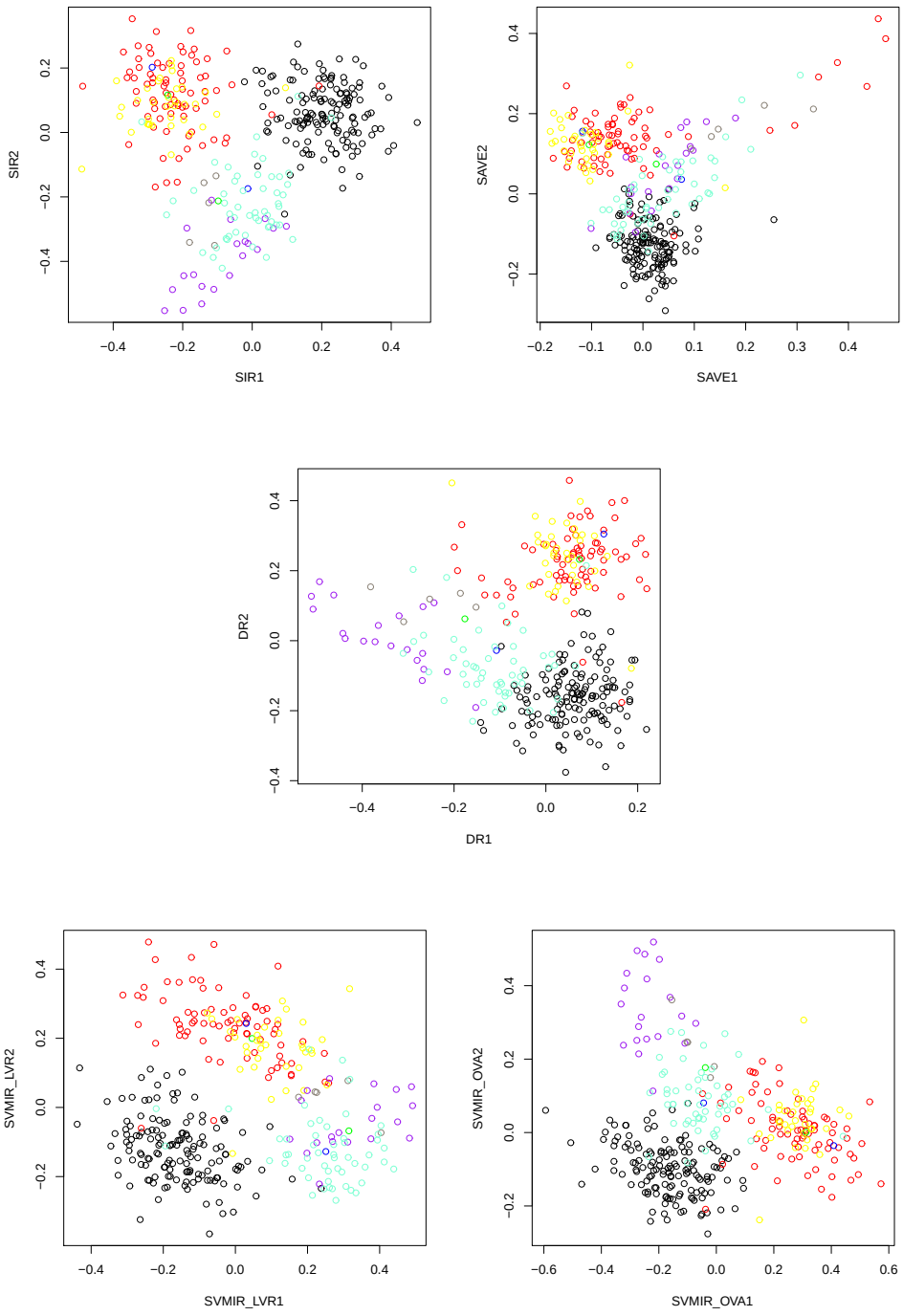

Figure 11.5: The first two directions for all the methods in the *E.coli* dataset analysis without binary predictors. Upper panel is SIR and SAVE, middle panel is DR and in the lower panel is SVMIR using both LVR and OVA method of estimating the directions.

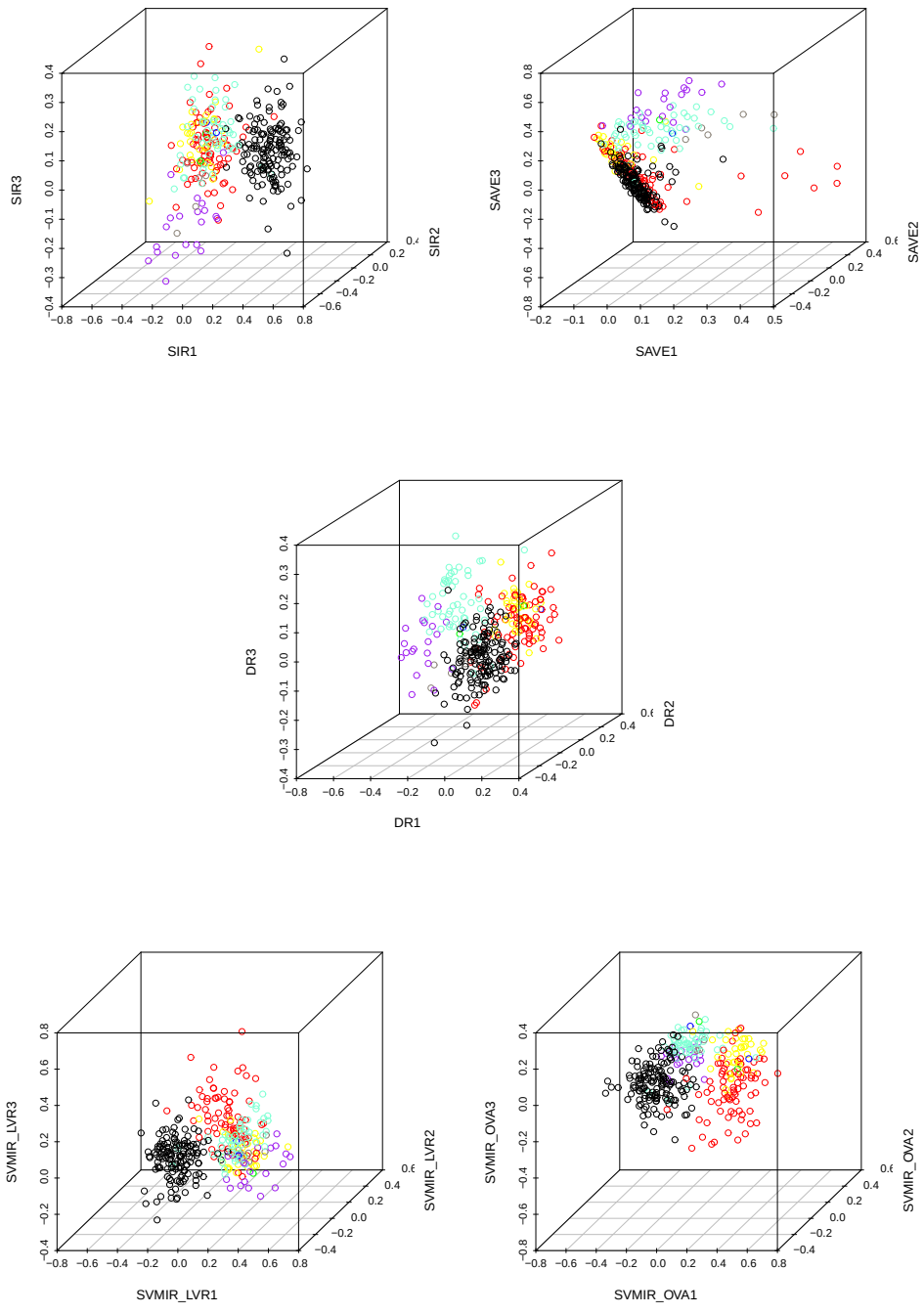

Figure 11.6: The first three directions for all the methods in the full *E.coli* dataset analysis. Upper panel is SIR and SAVE, middle panel is DR and in the lower panel is SVMIR using both LVR and OVA method of estimating the directions.

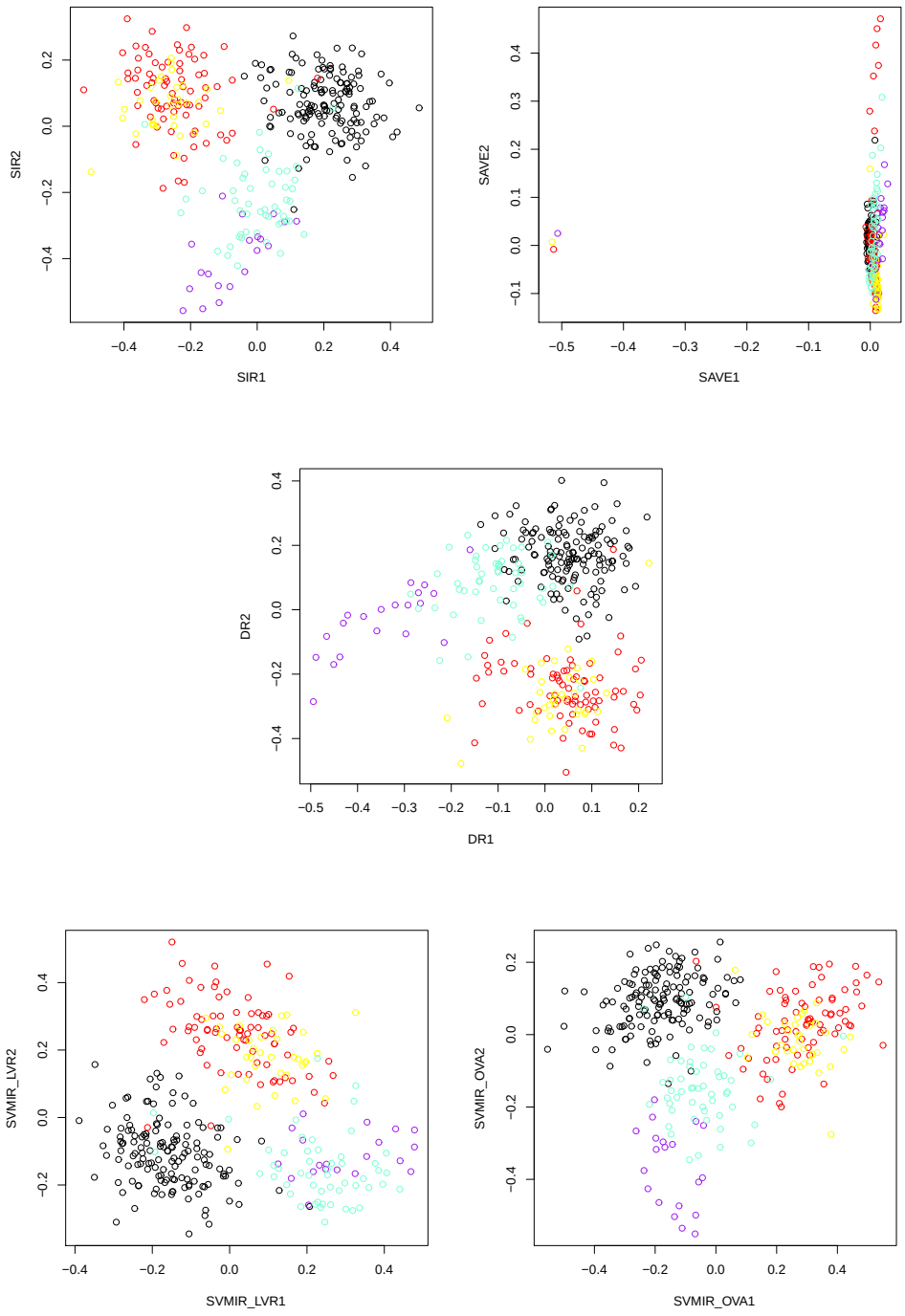

Figure 11.7: The first two directions for all the methods in the *E.coli* dataset analysis using only the 5 largest clusters of cells. Upper panel is SIR and SAVE, middle panel is DR and in the lower panel is SVMIR using both LVR and OVA method of estimating the directions.

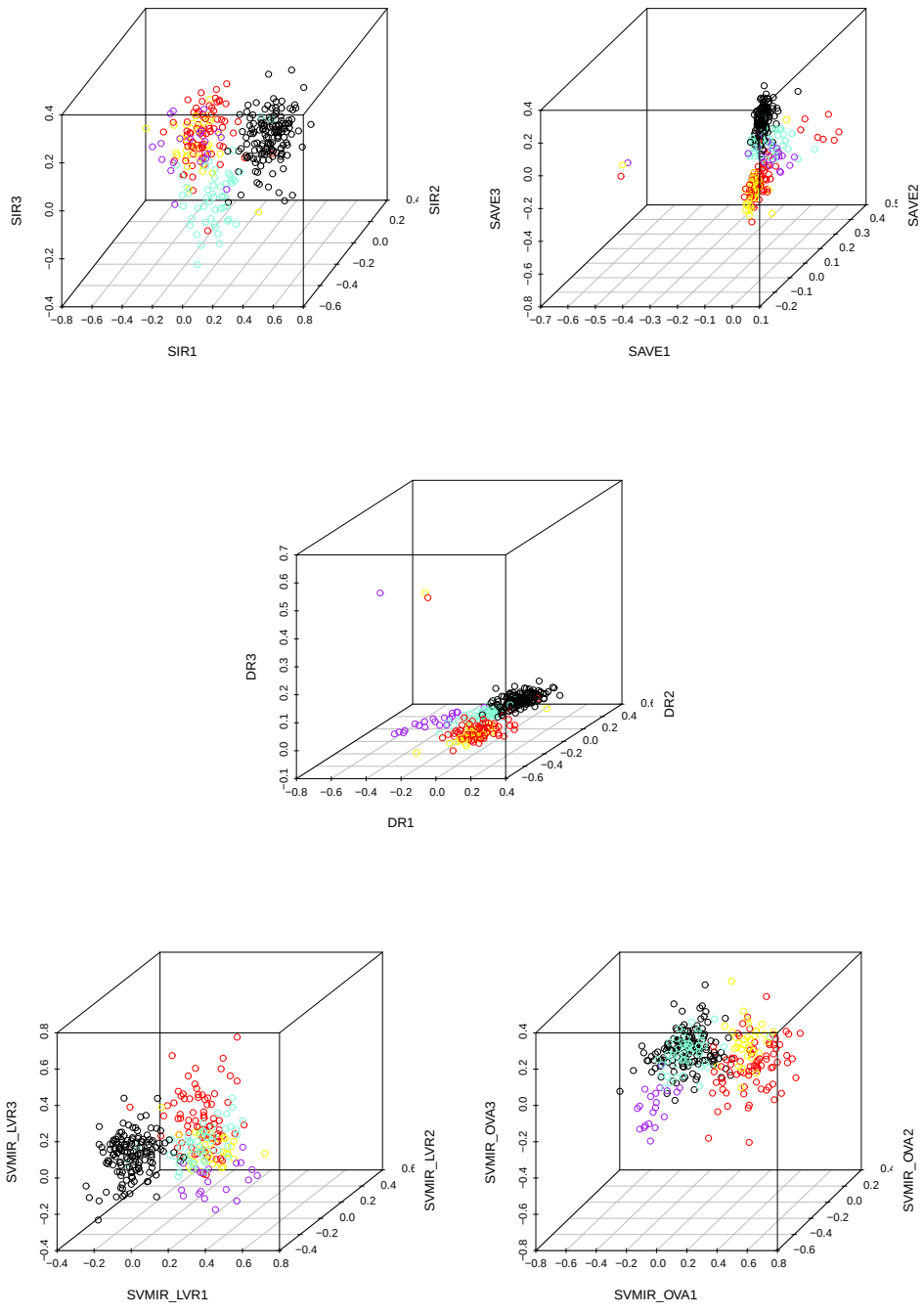

Figure 11.8: The first three directions for all the methods in the *E.coli* dataset analysis using only the 5 largest clusters of cells. Upper panel is SIR and SAVE, middle panel is DR and in the lower panel is SVMIR using both LVR and OVA method of estimating the directions.

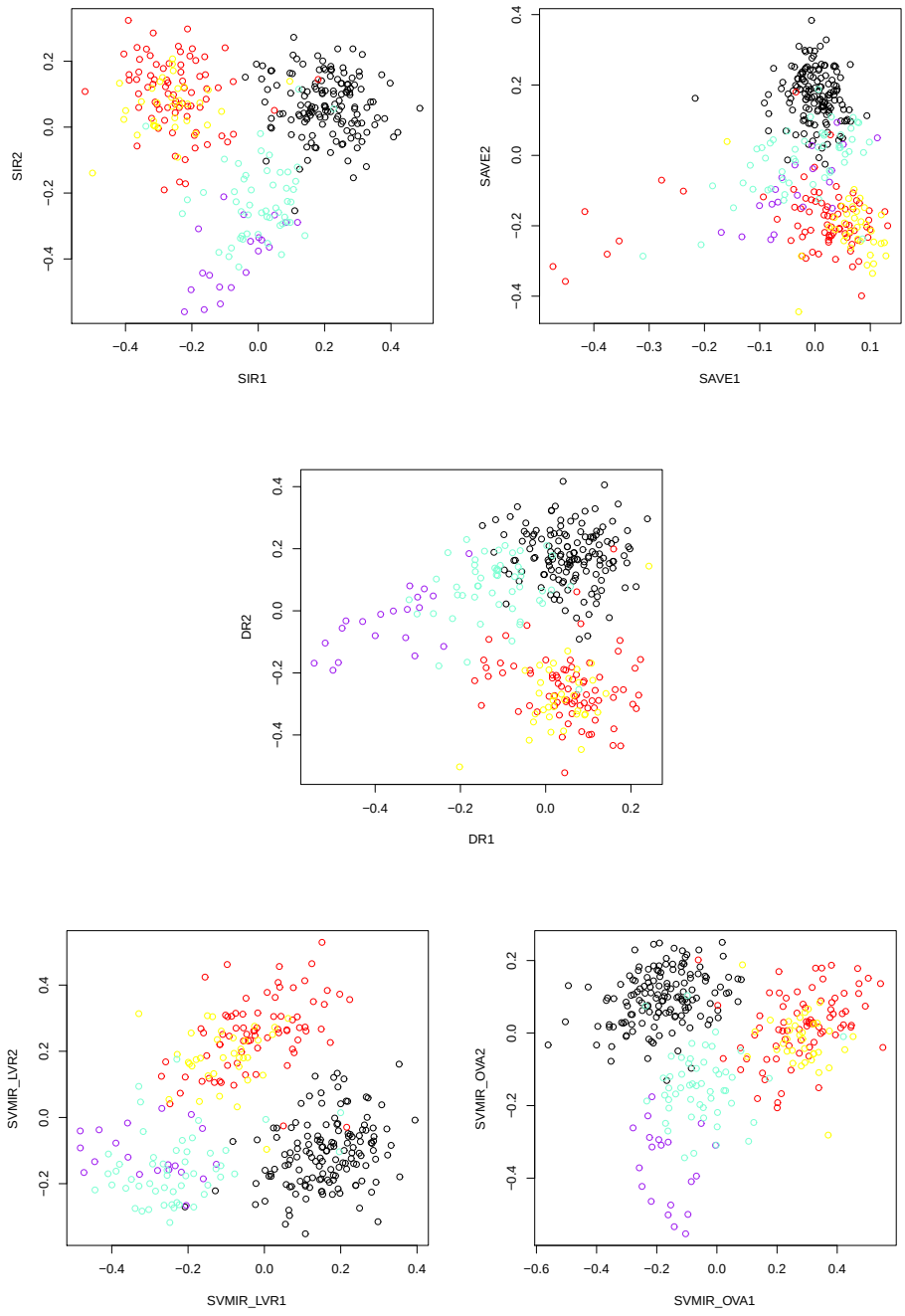

Figure 11.9: The first two directions for all the methods in the *E.coli* dataset analysis without binary predictors using only the 5 largest clusters of cells. Upper panel is SIR and SAVE, middle panel is DR and in the lower panel is SVMIR using both LVR and OVA method of estimating the directions

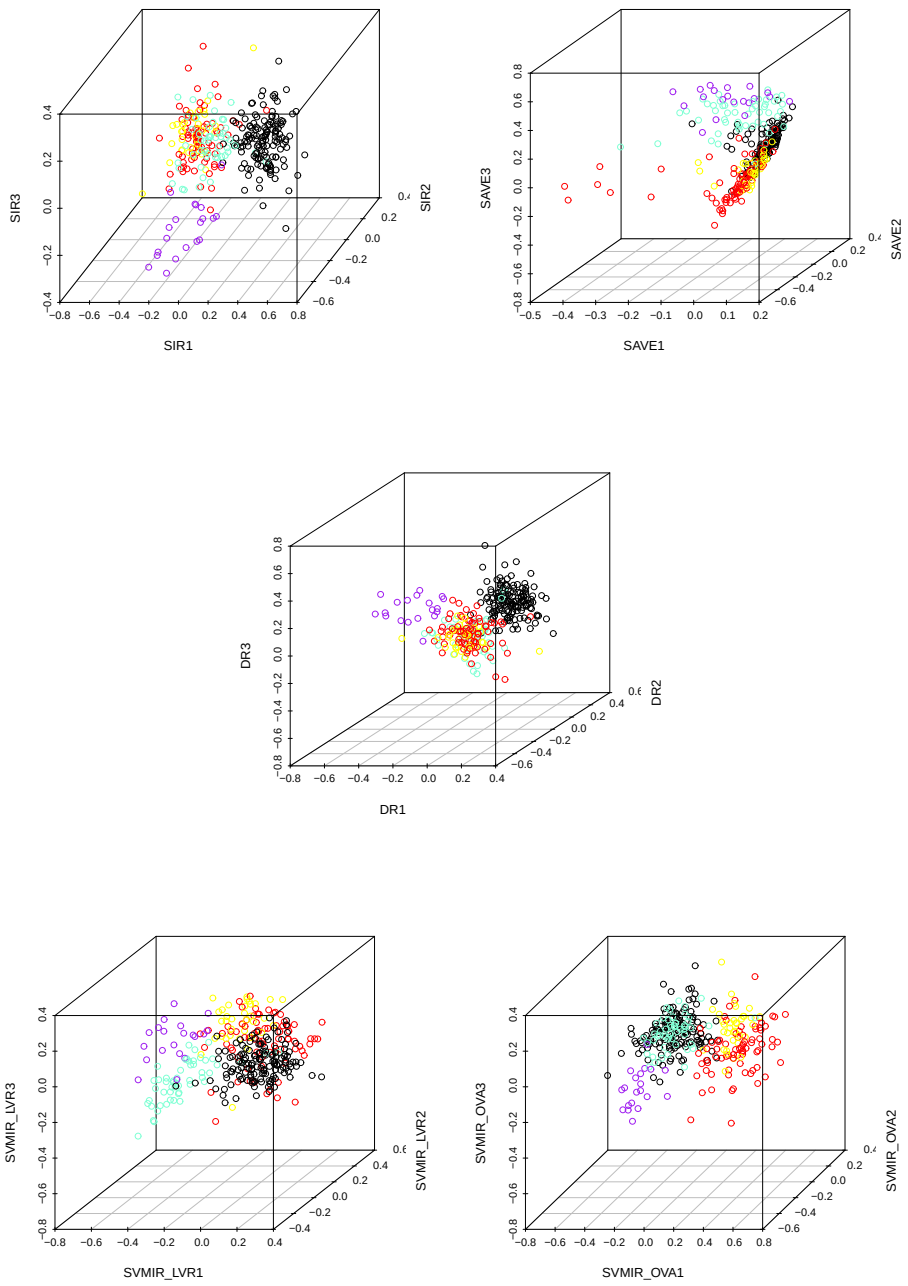

Figure 11.10: The first three directions for all the methods in the *E.coli* dataset analysis without binary predictors using only the 5 largest clusters of cells. Upper panel is SIR and SAVE, middle panel is DR and in the lower panel is SVMIR using both LVR and OVA method of estimating the directions

Figures 11.9 and 11.10 show that all methods perform similarly, except SAVE. DR regression seems to reverse the first and second direction compared to SIR and SVMIR methods. As we can see that SIR and DR were able to capture a picture when the binary predictors and the small categories are removed, while SVMIR was able to capture the same exact picture, before doing any kind of "manipulation" on the dataset.

Out of the analysis of the *E.coli* dataset, the most important observation is the fact that SVMIR, is not affected by the presence of binary/categorical predictors, while SIR, SAVE and DR are being affected. Also, the presence of smaller categories, seems to do not affect the dimension reduction process at all, for any of the methods we used. Finally, it is important to note that the two methods for SVMIR slice comparison LVR and OVA don't seem to affect the results a lot, but LVR performs slightly better when the response is continues and OVA performs slightly better when the response variable is categorical and there is no meaning in the ordering of the slices.

Chapter 12

Discussion

In this part of this work we proposed an inverse regression algorithm SVMIR for sufficient dimension reduction that has several advantages over previously proposed inverse regression algorithms. The advantages are listed as follows:

- It is robust against outliers
- It can perform dimension reduction without matrix inversion
- It is robust in violations of ellipticity of the predictors
- It is robust in the presence of binary/categorical variables in the predictors
- Using kernel functions we can extract nonlinear features, so we are able to perform nonlinear dimension reduction

The reason this method is robust to outliers is the fact that instead of depending on inverse moments to estimate the directions that span the Central Dimension Reduction Subspace for the regression of Y on $\boldsymbol{X}$, $S_{Y|\boldsymbol{X}}$, it depends on the optimal

separating hyperplane that is estimated, as a separator of the points in two slices. The vector that is orthogonal to the separating hyperplane, denoted as ψ is in the dimension reduction subspace. The equation of ψ, depends only on support vectors, the points that are closer to the optimal hyperplane. This, implies that if there are outliers in our samples, the estimation of $S_{Y|\boldsymbol{X}}$ will not be affected using SVMIR. This is proved in the derivation of asymptotic theory and the influence function in Chapter 9. It is also shown in Chapter 10 where simulations of several models show that indeed in the presence of outliers SVMIR performs better than other existing methods that uses inverse regression, like SIR, SAVE and DR.

Through a small modification in our objective function that we are minimizing in order to estimate the hyperplane, one can achieve dimension reduction without matrix inversion. Nowadays, many problems, especially in Biology and Genetics, where thousands or even hundreds of thousands of predictors are present and only a handful of observations are available (usually in the order of hundredths), there is a greater need of methods that address the large p small n issue. The fact that existing methods need matrix inversion, makes them unappealing to those problems. This method is the first method proposed to achieve dimension reduction, other than the work by Cook, Li and Chiaromonte (2007) where they propose rather a theoretical framework where any of the previous methods can be transformed to accommodate problems where $n < p$. The method proposed in this work is the first method that achieves that without the need of the trick proposed by Cook, Li and Chiaromonte (2007)

Through the simulations in Chapter 10 and the data analysis of the *E.coli* dataset in Chapter 11 we showed that our method is also robust in departures of ellipticity, and in the presence of categorical predictors. Since in real datasets the assumption of ellipticity is rarely true, having a method that is robust to deviations

of ellipticity, helps in finding better estimators of $S_{Y|\boldsymbol{X}}$.

Finally, it can be shown that using kernel functions, instead of the classic linear operator, one can extract nonlinear features in the central dimension reduction subspace $S_{Y|\phi(\boldsymbol{X})}$ using the "kernel trick". The theory was developed in section 9.4 and it shows, how kernel functions can be used to extract directions in the feature space of the kernel function. This method provides the first successful way of achieving nonlinear dimension reduction in the sufficient dimension reduction concept. The effort by Wang (2008) uses the idea of the feature space of a polynomial kernel, but she is not incorporating any support vector machine algorithms ideas and thus her results are not as good as one should expect and as the proposed method has.

12.1 Future work

This work opens the ground for better and more detailed work in dimension reduction. There are pieces that are missing though.

In the future we are interested in developing

- the structural dimension d for the number of significant directions
- the asymptotic distribution of $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d$ the first d eigenvectors of the matrices developed for LVR and OVA in section 9.2

$$\sum_{r=1}^{h} \hat{\psi}_r \hat{\psi}_r^{\mathsf{T}} \quad \text{or} \quad \sum_{r=2}^{h} \sum_{s=r+1}^{h} \hat{\psi}_{rs} \hat{\psi}_{rs}^{\mathsf{T}}.$$

Also, since the influence function of the optimal hyperplane seems to match

that of the median or quartiles, it should be interesting for someone to explore if this method can be applied to quantile regression, to achieve sufficient dimension reduction in that context. As far as we are concerned there is no literature on the sufficient dimension reduction for quantile regression.

Finally, further theoretical extension for nonlinear feature extraction as well as simulation results need to be developed, as we believe that our method will be very effective in extracting nonlinear features for sufficient dimension reduction. This will give maybe a very powerful method for nonlinear sufficient dimension reduction as the example with the vowel data showed

BIBLIOGRAPHY

Adcock, R. J. (1878). A problem in least squares. *The Analyst*, **5**, 53–54.

Aizerman, M. A., Braverman, E. M. and Rozonoer, L. I. (1964a). Theoretical foundation of the potential function method in pattern recognition learning. *Autom. Remote Control*, **25**, 821–837.

Aizerman, M. A., Braverman, E. M. and Rozonoer, L. I. (1964b). The problem of pattern recognition learning and the method of potential functipons. *Autom. Remote Control*, **25**, 1175–1193.

Alter, O., Brown, P. and Botstein, D. (2000). Singular value decomposition for gene-wide expression data processing and modelling. *Proceedings of the National Academy of Science*, **97**, 10101–10106.

Amato, U., Antoniadis, A. and De Feis, I. (2006). Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis*, **50**, 9, 2422 – 2446.

Anderson, T. W. and Rubin H. (1956). Statistical inference in fector analysis. In J. Neyman (Ed.) *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume V, U. Cal, Berkley, 111–150.

Arnold, B. C. and Brockett, P. L. (1992). On distributions whose component ratios are Cauchy. *American Statistician*, **46**, 25 – 26.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, **68**, 337 – 404.

Artemiou, A. (2008). *A probabilistic explanation of a natural phenomenon.* Unpublished Master Thesis. Pennsylvania State University, Department of Statistics.

Artemiou, A. and Li, B. (2009). On principal components and regression: A statistical explanation of a natural phenomenon. *Statistica Sinica*, **19**, 1557–1565.

Baker, C. R. (1973). Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, **186**, 273 – 289.

Asuncion, A. and Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

Bickel, P. J., Klaassen, C. A. J., Riton, Y., and Wellner, J. A. (1998) *Efficient and Adaptive Estimation for Semiparametric Models.* Springer.

Boser, B. E., Guyon, I. M. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Fifth Annual Workshop on COLT*, Pittsburgh, ACM

Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, **19**, 1252–1258.

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci.*, **176**, 123–144.

Conway, J. (1990). *A Course in Functional Analysis.* Second edition. Springer.

Cook, R. D. (1994a). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proceedings of the Section on Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, 18–25.

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

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

Cook, R. D. (1998a). *Regression Graphic. Ideas for Studying Regressions through Graphics* . Wiley Interscience.

Cook, R. D. (1998b). Principal Hessian directions revisited (with discussion) *Journal of the American Statistical Association*, **93**, 84–100.

Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**, 1–40.

Cook, R. D. and Li, B. (2002). Dimension Reduction for the conditional mean. *The Annals of Statistics*, **30**, 455–474.

Cook, R. D., Li, B. and Chiaromonte F. (2007). Dimension reduction without matrix inversion. *Biometrika*, **94**, 569–584.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 1–25.

Cox, D. R. (1968).Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society, Ser. A*, **131**, 265–279.

Deift, Percy (1999). *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach.* Courant Institute of Mathematical Sciences, New York University.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A. (2010). Package e1071. Available from http://cran.r-project.org

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

Eaton, M. L. (1986). A characterization of spherical distribtuions. *Journal of Multivariate Analysis*, **20**, 272–276.

Ein-Dor, P., Feldmesser, J. (1987). Attributes of the performance of central processing units: A relative performance prediction model *Communications of the ACM*, **30**, 4, 308–317.

Fukumizu, Bach, and Jordan (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *The Journal of Machine Learning Research*, **5**, 73–99.

Fukumizu, Bach, and Jordan (2009). Kernel dimension reduction in regression. *Annals of Statistics.* **4** 1871 – 1905.

Gannoun, A. and Saracco, J (2003). An asymptotic theory for SIR α method. *Statistica Sinica*, **13**, 297–310.

Golub,G. H. and van Loan, C. F. (1983) *Matrix Computations*, Baltimore, MD. John Hopkins University Press.

Hadi, A. S. and Ling, R. F. (1998). Some cautionary notes on the use of principal components in regression. *The American Statistician*, **52**, 15–19.

Hall, W. J. and Mathiason, D. J. (1990). On large-sample estimation and testing in parametric models. *International Statistics Review* **58**, 77–97

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association.* **84**, 502 – 516.

Hastie, T., Tibishrani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction* . 2nd Edition, Springer

Hawkins, D. M. and Fatti, L. P. (1984). Exploring multivariate data using the minor principal components. *The Statistician*, **33**, 325–338.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Horton, P., and Kenta, N. (1996). A Probablistic Classification System for Predicting the Cellular Localization Sites of Proteins. *Intelligent Systems in Molecular Biology*, 109–115.

Hotelling, H. (1933). Analysis of a complex statistical variable into its principal components *Journal of Educational Psychology*, **24**, 417–441.

Hotelling, H. (1957). The relationship of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, **10**, 69–79.

Hristache, M., Juditsky, A. Polzehl, J. and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics*, **29**, 1537–1566.

Hsing, T. and Ren, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *Annals of Statistics.* **37** 726–755.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58, 71–120.

Jiang, B., Zhang, X. and Cai, T. (2008). Estimating the confidence interval for prediction errors of support vector machines. *Journal of Machine Learning Research* , **9**, 521–540.

Joliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, **31**, 300–303.

Joliffe, I. T. (2002). *Principal Component Analysis, 2nd edition.* New York: Springer.

Jong, J. and Kotz, S. (1999). On a relation between principal components and regression analysis. *The American Statistician*, **53**, 349–352.

Kannan, D. and Bharucha-Reid, A. T. (1970). Note on covariance operators of probability measures on a Hilbert space.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A, (2004) Kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software*, **11**, 9

Kendall, M. G. (1957). *A course in Multivariate Analysis.* London: Griffin.

Lawley, D. N. (1953). A modified method of estimation in factor analysis and some large sample results. *Uppsala Symposium on Psychological Fector Analysis*, Number 3 in Nordisk Psykologi Monograph Series, 35–42. Uppsala: Almqvist and Wiksell.

LeCun, Y. (1986). Learning processes in an asymmetric threshold network. *Disordered Systems and Biological Organizations*, Springer, Les Houches, France, 233–240.

Li, B. (2003) *Dimension Reduction and regression analysis.* (Lecture Notes).

Li, B. (2007). Comment: Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**, 32–35.

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

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

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

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma. *Journal of the American Statistical Association*, **87**, 1025–1039.

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

Li, K. C., Wang, J. L. and Chen, C. H. (1999). Dimension Reduction for censored regression data. *The Annals of Statistics*, **27**, 1, 1 – 23.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 3406–3412.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression.* Reading, Massachusetts: Addison-Wesley.

Ni, L. (2010). Principal component regression revisited. *Statistica Sinica*, to appear.

Pearson, K (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine (6)*, **2**, 559–572.

Preisendorfer, R. W. and Mobley C. D. (188). *Principal Components Analysis in Meteorology and Oceanography*. Amsterdam: Elsevier.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya A*, **26**, 329–358.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of Royal Statistical Society, Series B*, **53**, 233 – 243.

Rosenblatt, F. (1956). *Principles of Neurodynamics: Perceptron and Theory of Brain Mechanisms*, Spartan Books, Washington D.C.

Rousseeuw, P. J. (1985). Multivariate Estimation with High-Breakdown Point. *Mathematical Statistics and Applications*, **B**, 283–297

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning internal representations by error propagation. *Parallel Distributed Processing: Exploration in the Macrostructure of Cognition*, Vol. I, Bradford Books, Cambridge, MA, 318–362.

Schölkopf, B., Smola, A., Müller, K.-R. (1997). Kernel principal component analysis. *Artificial Neural Networks*. 583 – 588.

Schölkopf, B., Smola, A., Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*. **10**. 1299 – 1319.

Scott, D. (1992). *Multivariate Density Estimation*. New York: Wiley.

Shi, T., Belkin, M., and Yu, B. (2009). Data Spectroscopy: Eigenspaces of Convolution Operators and Clustering. *Annals of Statistics*, **37**, 3960 – 3984.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, **24**, 1 – 24.

Skorohod, A. V. (1976). Random operators in a Hilbert space. *Lecture Notes in Mathematics*. **550**. 567 – 591.

Skorohod, A. V. (1984). *Random Linear Operators*. D. Reidel Publishing Company, Dordrecht, Holland.

Tang, Z. (2007). *Three topics on dimension reduction*. Unpublished Doctoral Dissertation. Pennsylvania State University, Department of Statistics.

Tipping M. E. and Bishop, C. M. (1999). Probabilistic principal components. *Journal of the Royal Statistical Society, Series B*, **61**, 611–622.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley Intersience.

Wang, Y. (2008) *Nonlinear Dimension Reduction in Feature Space*. Unpublished Doctoral Dissertation. Pennsylvania State University, Department of Statistics.

Wu, H. M. (2008). Kernel Sliced Inverse Regression with Applications on Classification. *Journal of Computational and Graphical Statistics*, **17**, 3, 590 – 610.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society, Series B*

(Methodological), **64**, 363–410.

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

Yeh, Y. R., Huang S. Y., and Lee Y. J. (2009). Nonlinear Dimension Reduction with Kernel Sliced Inverse Regression. *IEEE transactions on Knowledge and Data Engineering*. **21**. 1590–1603

VITA
Andreas A. Artemiou

Education

- Ph. D. Statistics Pennsylvania State University August 2010
- M. Sc. Statistics Pennsylvania State University May 2008
- B. Sc. Mathematics and Statistics University of Cyprus June 2005

Employment

- Teaching Assistant / Instructor Department of Statistics Penn State August 2005- June 2010
- Research Assistant Department of Statistics Pennsylvania State University May 2009 - July 2009
- Research Assistant Department of Computer Science University of Cyprus June 2003 - August 2005

Publications

- Artemiou, A. and Li, B. (2009). On principal components and regression: A statistical explanation of a natural phenomenon. *Statistica Sinica*, **19**, 1557–1565.
- Vonta F. and Artemiou, A. (2008). Hypothesis testing in frailty models for arbitrary censored and truncated data. *CDQM*, **10**, 1, 110-121.