The Pennsylvania State University

The Graduate School

Eberly School of Science

A PROBABILISTIC EXPLANATION OF A NATURAL PHENOMENON

A Thesis in

Statistics

by

Andreas A. Artemiou

Submitted in Partial Fullfillment
of the Requirements
for the degree of

Master of Science

May 2008

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

Bing Li
Professor of Statistics
Thesis Adviser

Runze Li
Assistant Professor of Statistics
Graduate Studies Chair

Donald St. P. Richards
Professor of Statistics
Acting Department Head

*Signatures are on file with Graduate School

Abstract

Regression is the procedure that attempts to relate a p-dimensional vector of predictors $\boldsymbol{X}$ with a response variable Y. Frequently, we deal with regression problems that have a large amount of predictors. In those cases, we try to reduce the dimension of our predictor vector. The reason we are trying to reduce the dimension, is the necessity to find the predictors that will affect our response the most. One of the most widely used methods is the Principal Components Analysis. With this analysis, I try to find the first few d ($\ll p$) principal components, that are generally believed to better describe the relationship between predictors $\boldsymbol{X}$ and response Y.

This procedure however has not been appropriately justified. In practice, it often occurs that the first few principal components are more highly correlated with the response variable, and better describe the relationship between the predictors and the response variable than the other principal components. However, there seems no logical reason for this tendency, and there are cases - albeit less often - where the first few principal components have weaker correlation with the response. There is a long standing debate on this issue among statisticians, and, todate, it has not been adequately resolved.

In this thesis I ask, and attempt to answer, the following questions: Is there a tendency for the first few principal components of the predictor to be more strongly related with the response? If so, what is the reason behind this tendency? And how strong is this tendency?

Key Words and Phrases Principal components; Regression; Correlation; Eigenpairs; Orientationally Uniform distribution; Random Covariance Matrix; Dimension Reduction.

Table of Contents

List of Figures

Acknowledgemet

I would also like to thank the two reviewers for their useful feedback and my collegues in STAT 590 class for their useful comments, after the presentation of part of this work to them.

I would also like to thank my friend Snow Efi, for a thorough reading of my thesis and her useful grammar and spelling corrections.

Finally, I would like to thank my advisor, Professor of Statistics, Bing Li, for his help and guidance, but mostly for his patience.

Dedication

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

"BE ON THE RIGHT WAY MY SON,
WITH A POWERFUL PEN"
IOANNIS CHARALAMBOUS APOSTOLIDES

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 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 they have been used in, 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 as 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.

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$^{\text{th}}$ principal component can be found by multiplying the eigenvector corresponding to λ_i (the 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$^{\text{th}}$ 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 correspond to the eigenvalues that are larger than a certain cutoff point (usually 1). There are many other subjective and inferential methods for determining 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 the response. Although there can be a multivariate response, for the purpose of this thesis, I will focus my analysis 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 therefore 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 his work, Hotelling 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 $\text{cov}\left(\boldsymbol{a}_i^T \boldsymbol{X}, \boldsymbol{a}_j^T \boldsymbol{X}\right) = 0, j = 1, ...i - 1$. Also, Kendall (1957), explained why doing regression using the principal components instead of the original predictors helps us towards 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 the 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 method of doing this is by using variance inflation factors (VIF's) for the p predictor variables. If VIF's are close to 1, it 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

In order for someone to completely understand the problem I address with this work, I will present a small and interesting piece of the debate that is actually still going on, between some scientists, mainly in Statistics. This debate was presented in Cook (2007) in greater detail.

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 say that they believe that although:

> "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 in most of the cases, regressing on 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, while still using this procedure. It is really interesing that there is very little work done in identifying how often we get the wrong answer.

The reason why 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 a direct or indirect way in calculating the 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 this thesis, I will try to show that, under mild assumptions, there is a higher probability to get the principal components that are more highly correlated with the response than the ones that are less correlated. Even though what I will prove is not definite and specific, it will give partial justification that principal components analysis can be used in a dimension reduction problem as a first step for further and more careful analysis.

Already, many scientists are working towards other dimension reduction methods, that are based on finding the central dimension reduction subspace for the regression of Y on $\boldsymbol{X}$. These methods are much more effective and they perform better in reducing the dimension of $\boldsymbol{X}$ using information of Y. Extensive research in sufficient dimension reduction can be found in the works by Li (1991) and (1992), Cook (1994) and (1996) and Li and Wang (2007) who present those methods in detail.

Before continuing with the details, I will present in the next section a conjecture by Li (2007) which incentivized me to work on this idea.

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.

Li (2007) argued intuitively 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 effects 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.

In this thesis I will present a precise formulation of the conjecture and a rigorous proof that will show that the above conjecture holds under some mild assumptions. In Chapter 2, I will demonstrate this phenomenon using examples. In Chapter 3, I will present the formulation when we have 2 dimensional predictor vector $\boldsymbol{X}$. In Chapter 4, I will present the formulation for a general p dimensional predictor vector $\boldsymbol{X}$. In Chapter 5, I will relate the results to Stochastic ordering. Chapter 6 presents some conclusions.

Chapter 2

Motivating examples

Before rigorously formulating and proving the conjecture I will present in this chapter an example in which some randomly chosen data sets show the property that the conjecture describes.

From a collection of 80 datasets, which can be found in *Arc* software database, which can be found at http://www.stat.umn.edu/arc/software.html, I have chosen 33 datasets that satisfy the following conditions:

1. Have a univariate response variable Y,
2. Have two or more predictors,
3. Do not have categorical response or predictors,
4. Are not simulated data.

Conditions 1 and 3 are to satisfy the assumptions in this work. Multivariate Y can be explored in future research. Condition 2 is essential in order to have a

principal component analysis and condition 4 is to ensure the data arise naturally from practice.

The selected datasets have a variable number of predictors, ranging from 2 to 12. The procedure followed is described herewith: For each dataset, I have calculated the eigenvalues and the corresponding eigenvectors. In each case I have calculated the principal components using those eigenvectors and finally the square correlation coefficient of each principal component with the response. I worked with the square correlations coefficients to avoid any confusion from negative correlation. For each case I found the principal component with the highest square correlation coefficient. Not suprisingly, among the 33 datasets, in 24 cases the first principal component is the one with the highest square correlation coefficient with the response, that is a percentage which is close to 73%. In the remaining 9 sets, there were 6 times that the second principal component had the highest square correlation coefficient with the response, 2 times the third principal component is the one with the highest square correlation coefficient and finally, in the last one the fifth principal component had the highest correlation coefficient.

While the above results verify, at least, datawise, the conjecture by Li (2007), there are obvious cases where the concerns of some researches are also verified. That is, there are cases where the least principal components have higher square correlation coefficient with the response than the square correlation coefficient of the first principal component with the response.

In order to strengthen this, I am including Figure 2.1 in which I present the squared correlation coefficient between the responses and the first principal components of the predictors (left) and between the responses and the second principal components of the predictors (right) for the 33 data sets, which does indicate the

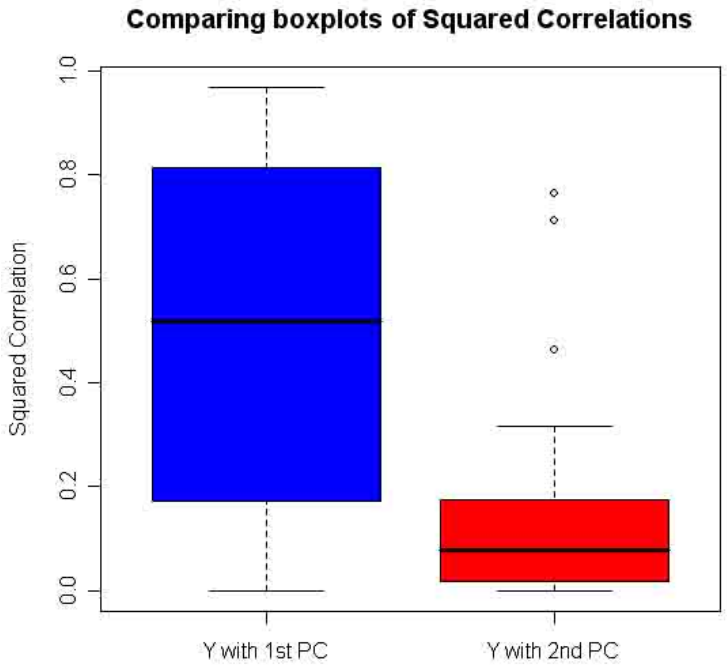

Figure 2.1: Box plots of the squared correlations between the response and the first principal components and the response and the second principal component.

tendency for the response to have higher squared correlation with the first principal component of the predictor. I am using only the first two principal component's correlation coefficients because not all the datasets had more than 2 predictors.

It is easy to see that tendency, since the quartiles of the distribution of the squared correlation coefficient of the response with the first principal component, are much higher than the respective quartiles of the distribution of the squared correlation coefficient of the response with the second principal component.

Chapter 3

Preliminary Results

In this chapter I will provide the proofs of some lemmas that will be helpful in proving the conjecture. Those results will guide me through the required path of finding the least possible assumptions required to prove our final theorem. They are being given here to help the reader to better understand the details of the problem and make an easy transition to the final version of our theorem.

I am also working with only 2 dimensional predictor vector, in order to explore all the necessary conditions that I need for the conjecture to holds. Finally, working in two dimensions will lead us into the complicate case where we have p dimensional predictor vector $\boldsymbol{X}$. The p dimensional case will be explored in the next Chapter.

3.1 Simplest case

The following lemma give us the case where we assume non-random covariance matrix $\boldsymbol{\Sigma}$ and normally distributed regression coefficient $\boldsymbol{\beta}$ in a regression function

that does not include the error term.

Lemma 3.1.1 *Let*

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

$\boldsymbol{\beta} \sim N(0, I_2)$ *satisfying* $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$, *and* $Y = \beta^T \boldsymbol{X}$. *Let*

$$\rho_1(\boldsymbol{\beta}) = \operatorname{corr}^2(Y, X_1|\boldsymbol{\beta}) \quad \textit{and} \quad \rho_2(\boldsymbol{\beta}) = \operatorname{corr}^2(Y, X_2|\boldsymbol{\beta}).$$

Then $P(\rho_1(\boldsymbol{\beta}) > \rho_2(\boldsymbol{\beta})) > \frac{1}{2}$ *when* $\sigma_1^2/\sigma_2^2 > 1$.

Proof. By definition

$$\rho_i(\boldsymbol{\beta}) = \frac{\operatorname{cov}^2(Y, X_i|\boldsymbol{\beta})}{\operatorname{var}(Y|\boldsymbol{\beta}) \operatorname{var}(X_i)}, \quad i = 1, 2.$$

Hence

$$P(\rho_1(\boldsymbol{\beta}) > \rho_2(\boldsymbol{\beta})) = P\left(\frac{\operatorname{cov}^2(Y, X_1|\boldsymbol{\beta})}{\operatorname{var}(Y|\boldsymbol{\beta}) \operatorname{var}(X_1)} > \frac{\operatorname{cov}^2(Y, X_2|\boldsymbol{\beta})}{\operatorname{var}(Y|\beta) \operatorname{var}(X_2)}\right). \tag{3.1}$$

Because $E(X_1) = 0$, we have $\operatorname{cov}(Y, X_1|\boldsymbol{\beta}) = E(YX_1|\boldsymbol{\beta})$ in the numerator of the left fraction in (3.1). So

$$\begin{aligned} \operatorname{cov}(Y, X_1|\boldsymbol{\beta}) =& \boldsymbol{\beta}^T E(\boldsymbol{X} X_1|\boldsymbol{\beta}) \\ =& \boldsymbol{\beta}^T E\left[\begin{pmatrix} X_1^2 \\ X_1 X_2 \end{pmatrix} \middle| \boldsymbol{\beta}\right] = \boldsymbol{\beta}^T \begin{pmatrix} E(X_1^2|\boldsymbol{\beta}) \\ E(X_1 X_2|\boldsymbol{\beta}) \end{pmatrix} = \boldsymbol{\beta}^T \begin{pmatrix} \sigma_1^2 \\ 0 \end{pmatrix} = \beta_1 \sigma_1^2. \end{aligned}$$

Similarly, in the numerator of the right fraction in (3.1), $\operatorname{cov}(Y, X_2|\boldsymbol{\beta}) = \beta_2 \sigma_2^2$.

In both denominators in (3.1) we have

$$\text{var}\left(Y|\boldsymbol{\beta}\right) = \text{var}\left(\boldsymbol{\beta}^T\boldsymbol{X}|\boldsymbol{\beta}\right) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}$$

which is a scalar and so we can cancel it from both sides of the inequality in (3.1). Also we have that $\text{var}\left(X_1|\boldsymbol{\beta}\right) = \sigma_1^2$ and $\text{var}\left(X_2|\boldsymbol{\beta}\right) = \sigma_2^2$. So (3.1) becomes

$$P\left(\frac{\beta_1^2\sigma_1^4}{\sigma_1^2} > \frac{\beta_2^2\sigma_2^4}{\sigma_2^2}\right) = P\left(\beta_1^2\sigma_1^2 > \beta_2^2\sigma_2^2\right) = P\left(\frac{\beta_1^2}{\beta_2^2} > \frac{\sigma_2^2}{\sigma_1^2}\right) > \frac{1}{2} \tag{3.2}$$

Because β_1 and β_2 are iid $N(0,1)$, the ratio β_1^2/β_2^2 is distributed as $F_{(1,1)}$. Hence (3.2) holds when $\frac{\sigma_2^2}{\sigma_1^2} < \text{median}\left(F_{1,1}\right) = 1$.

By symmetry, $P\left(\rho_2\left(\boldsymbol{\beta}\right) > \rho_1\left(\boldsymbol{\beta}\right)\right) > \frac{1}{2}$ holds when $\frac{\sigma_1^2}{\sigma_2^2} < \text{median}\left(F_{1,1}\right) = 1$. □

The above Lemma gives the proof in a very simple case. There is an observation though that will lead me to the next couple of lemmas. The distribution of the coefficients β_1 and β_2 is not being used in the proof until after expression (3.2). Actually, I can remove the distribution assumption on $\boldsymbol{\beta}$ and replace it with the assumption that β_1^2/β_2^2 and β_2^2/β_1^2 have the distribution have the same distribution. One extension is shown in the next Lemma, which is similar to Lemma (3.1.1). I am just replacing the assumption on the distribution of $\boldsymbol{\beta}$.

Lemma 3.1.2 *Let*

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

Let $\boldsymbol{\beta}$ be a 2-dimensional random vector satisfying $\beta_1^2/\beta_2^2 \overset{\mathcal{D}}{=} \beta_2^2/\beta_1^2$, $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$, and let

$Y = \boldsymbol{\beta}^T \boldsymbol{X}$. *Assume, furthermore, that the median of* β_1^2/β_2^2 *is unique. Let*

$$\rho_1(\boldsymbol{\beta}) = \mathrm{corr}^2(Y, X_1|\boldsymbol{\beta}) \quad \textit{and} \quad \rho_2(\boldsymbol{\beta}) = \mathrm{corr}^2(Y, X_2|\boldsymbol{\beta}).$$

Then $P(\rho_1(\boldsymbol{\beta}) \gtrless \rho_2(\boldsymbol{\beta})) > \frac{1}{2}$ *whenever* $\sigma_1^2 \gtrless \sigma_2^2$.

In the above, $\stackrel{\mathcal{D}}{=}$ means that the random variables have the same distribution.

The proof of this lemma is essentially the same as that of Lemma 3.1.1 since in the proof of Lemma 3.1.1 we do not use the fact that $\boldsymbol{\beta}$ follows the normal distribution before the line following expression (3.2). So the only thing we have to change after this is the following.

PROOF. Because $\beta_1^2/\beta_2^2 \stackrel{\mathcal{D}}{=} \beta_2^2/\beta_1^2$, they have the same median. Let m be this common median. Then

$$P(\beta_1^2/\beta_2^2 < m) = 1/2, \quad P(\beta_2^2/\beta_1^2 < m) = 1/2 \ \ (\text{because } \beta_1^2/\beta_2^2 \stackrel{\mathcal{D}}{=} \beta_2^2/\beta_1^2).$$

At the same time, from the first equality,

$$P(\beta_2^2/\beta_1^2 > 1/m) = 1/2 \Rightarrow P(\beta_2^2/\beta_1^2 < 1/m) = 1/2.$$

Because the median is unique, we have $m = 1/m$. So $m = 1$. By the definition of median, (3.2) holds whenever $\sigma_2^2/\sigma_1^2 < \mathrm{median}(D) = 1$ where D denotes the distribution of β_1^2/β_2^2. This completes the proof. □

The most important result that the above Lemmas present is that the larger the ratio between the two eigenvalues the larger the probability is. That is, the

larger the ratio between the two eigenvalues the higher the probability the first principal component will have higher correlation with the response variable than the correlation of the second principal component with the response variable. This means, the higher the probability the Principal Component Analysis will give you the component with the highest correlation with the response variable. This is true though only in case that we have, two independent predictors and a non-random predictor vector $\boldsymbol{X}$. We will see later than this is not true in the general case where matrix $\boldsymbol{\Sigma}$ is considered a random matrix.

3.2 Including error into regression functions

As one can see from the Lemmas in the previous subsections I have assumed the simplest form of regression form, that is one that includes no error term. This leads to easier calculations, but it is actually not useful to work with a function for the regression that does not include any error term. In this section I will show that Lemma 3.1.2 holds if I have an additive error term in my regression function.

Lemma 3.2.1 *Let*

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

Let $\boldsymbol{\beta}$ be a 2-dimensional random vector satisfying $\beta_1^2/\beta_2^2 \overset{\mathcal{D}}{=} \beta_2^2/\beta_1^2$, $\boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{X}$, and let $Y = \boldsymbol{\beta}^T\boldsymbol{X} + \delta$, where $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$. Assume, furthermore, that the median of β_1^2/β_2^2 is unique. Let

$$\rho_1(\boldsymbol{\beta}) = \operatorname{corr}^2(Y, X_1|\boldsymbol{\beta}) \quad \textit{and} \quad \rho_2(\boldsymbol{\beta}) = \operatorname{corr}^2(Y, X_2|\boldsymbol{\beta}).$$

Then $P\left(\rho_1\left(\boldsymbol{\beta}\right) \gtrless \rho_2\left(\boldsymbol{\beta}\right)\right) > \frac{1}{2}$ *whenever* $\sigma_1^2 \gtrless \sigma_2^2$.

PROOF. By definition

$$\rho_i(\boldsymbol{\beta}) = \frac{\operatorname{cov}^2\left(Y, X_i|\boldsymbol{\beta}\right)}{\operatorname{var}\left(Y|\boldsymbol{\beta}\right)\operatorname{var}\left(X_i\right)}, \quad i = 1, 2.$$

Hence

$$P\left(\rho_1\left(\boldsymbol{\beta}\right) > \rho_2\left(\boldsymbol{\beta}\right)\right) = P\left(\frac{\operatorname{cov}^2\left(Y, X_1|\boldsymbol{\beta}\right)}{\operatorname{var}\left(Y|\boldsymbol{\beta}\right)\operatorname{var}\left(X_1\right)} > \frac{\operatorname{cov}^2\left(Y, X_2|\boldsymbol{\beta}\right)}{\operatorname{var}\left(Y|\boldsymbol{\beta}\right)\operatorname{var}\left(X_2\right)}\right). \tag{3.3}$$

Because $E(X_1) = 0$, we have $\operatorname{cov}\left(Y, X_1|\boldsymbol{\beta}\right) = E\left(YX_1|\boldsymbol{\beta}\right)$ in the numerator of the left fraction in (3.3). So

$$\begin{aligned}
\operatorname{cov}\left(Y, X_1|\boldsymbol{\beta}\right) =& \boldsymbol{\beta}^T E\left(\boldsymbol{X}X_1|\boldsymbol{\beta}\right) + E\left(\delta X_1|\boldsymbol{\beta}\right) \\
=& \boldsymbol{\beta}^T E\left(\boldsymbol{X}X_1|\boldsymbol{\beta}\right) + E\left(\delta X_1\right) \\
=& \boldsymbol{\beta}^T E\left(\boldsymbol{X}X_1|\boldsymbol{\beta}\right) + E\left(\delta\right) E\left(X_1\right) \\
=& \boldsymbol{\beta}^T E\left[\begin{pmatrix} X_1^2 \\ X_1X_2 \end{pmatrix} \middle| \boldsymbol{\beta}\right] = \boldsymbol{\beta}^T \begin{pmatrix} E\left(X_1^2|\boldsymbol{\beta}\right) \\ E\left(X_1X_2|\boldsymbol{\beta}\right) \end{pmatrix} = \boldsymbol{\beta}^T \begin{pmatrix} \sigma_1^2 \\ 0 \end{pmatrix} = \beta_1\sigma_1^2.
\end{aligned}$$

Similarly, in the numerator of the right fraction in (3.3), $\operatorname{cov}\left(Y, X_2|\boldsymbol{\beta}\right) = \beta_2\sigma_2^2$.

In both denominators in (3.3) we have

$$\operatorname{var}\left(Y|\boldsymbol{\beta}\right) = \operatorname{var}\left(\boldsymbol{\beta}^T\boldsymbol{X}|\boldsymbol{\beta}\right) + \operatorname{var}\left(\delta|\boldsymbol{\beta}\right) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + \operatorname{var}\left(\delta\right)$$

which is a scalar and so we can cancel it from both sides of the inequality in (3.3). Now we have that $\text{var}\,(X_1|\boldsymbol{\beta}) = \sigma_1^2$ and $\text{var}\,(X_2|\boldsymbol{\beta}) = \sigma_2^2$. So (3.3) becomes

$$P\left(\frac{\beta_1^2\sigma_1^4}{\sigma_1^2} > \frac{\beta_2^2\sigma_2^4}{\sigma_2^2}\right) = P\left(\beta_1^2\sigma_1^2 > \beta_2^2\sigma_2^2\right) = P\left(\frac{\beta_1^2}{\beta_2^2} > \frac{\sigma_2^2}{\sigma_1^2}\right) > \frac{1}{2} \tag{3.4}$$

which is the same as (3.2) and so everything follows from the proofs of Lemmas 3.1.1 and 3.1.2. □

The above lemma has non-random predictor covariance matrix $\boldsymbol{\Sigma}$. But in the conjecture of Chapter 1, we can see that Li (2007) assumes the covariance matrix to be random. That's another assumption I add in the lemma in the next section in order to satisfy the assumptions on the conjecture.

3.3 Random predictor variances

In the Lemma that follows I am trying to add randomness on the way matrix $\boldsymbol{\Sigma}$ is formed. I am not adding any specific distribution on the variances σ_1^2 and σ_2^2. Mainly I am interested in introducing this distributional assumption, because in the stated conjecture Li (2007) assumes a ranodm ovariance matrix. This result is proven using a similar procedure as in the previous Lemmas of this Chapter. By assuming that σ_1^2 and σ_2^2 are random the proof gets more complicated as you have to condition on σ_1^2 and σ_2^2, in order to derive the desired outcome.

Lemma 3.3.1 *Let*

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

where σ_1^2 *and* σ_2^2 *are iid* G*. Let* $\boldsymbol{\beta}$ *be a 2-dimensional random vector satisfying* $\beta_1^2/\beta_2^2 \stackrel{\mathcal{D}}{=} \beta_2^2/\beta_1^2$, $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\Sigma})$*. Let* $Y = \boldsymbol{\beta}^T \boldsymbol{X} + \delta$*, where* $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$*. Furthermore, assume that the median of* β_1^2/β_2^2 *is unique. Suppose that* $P(\sigma_1^2 = \sigma_2^2) = 0$*. Let*

$$\rho_1(\boldsymbol{\beta}) = \begin{cases} \operatorname{corr}^2(Y, X_1|\boldsymbol{\beta}) & \text{if } \sigma_1^2 > \sigma_2^2 \\ \operatorname{corr}^2(Y, X_2|\boldsymbol{\beta}) & \text{if } \sigma_1^2 \leq \sigma_2^2 \end{cases} \text{ and } \rho_2(\boldsymbol{\beta}) = \begin{cases} \operatorname{corr}^2(Y, X_1|\boldsymbol{\beta}) & \text{if } \sigma_1^2 \leq \sigma_2^2 \\ \operatorname{corr}^2(Y, X_2|\boldsymbol{\beta}) & \text{if } \sigma_1^2 > \sigma_2^2 \end{cases}.$$

Then

$$P(\rho_1(\boldsymbol{\beta}) > \rho_2(\boldsymbol{\beta})) > P(\rho_1(\boldsymbol{\beta}) < \rho_2(\boldsymbol{\beta})). \tag{3.5}$$

Proof. Let $\boldsymbol{\eta} = (\boldsymbol{\beta}, \sigma_1^2, \sigma_2^2)$. By definition

$$\rho_i(\boldsymbol{\beta}) = \frac{\operatorname{cov}^2(Y, X_i|\boldsymbol{\beta})}{\operatorname{var}(Y|\boldsymbol{\beta}) \operatorname{var}(X_i)}, \quad i = 1, 2.$$

First, consider the case $\sigma_1^2 > \sigma_2^2$. We have

$$\begin{aligned} &P\left(\rho_1(\boldsymbol{\beta}) > \rho_2(\boldsymbol{\beta}) | \sigma_1^2, \sigma_2^2\right) = \\ &P\left(\frac{\operatorname{cov}^2(Y, X_1|\boldsymbol{\eta})}{\operatorname{var}(Y|\boldsymbol{\eta}) \operatorname{var}(X_1|\boldsymbol{\eta})} > \frac{\operatorname{cov}^2(Y, X_2|\boldsymbol{\eta})}{\operatorname{var}(Y|\boldsymbol{\eta}) \operatorname{var}(X_2|\boldsymbol{\eta})} \middle| \sigma_1^2, \sigma_2^2\right) \end{aligned} \tag{3.6}$$

Because $E(X_1|\boldsymbol{\eta}) = 0$, we have $\text{cov}\,(Y, X_1|\boldsymbol{\eta}) = E\,(YX_1|\boldsymbol{\eta})$ in the numerator of the left fraction in (3.6). So

$$\begin{aligned}
\text{cov}\,(Y, X_1|\boldsymbol{\eta}) =& \boldsymbol{\beta}^T E\,(\boldsymbol{X}X_1|\boldsymbol{\eta}) + E\,(\delta X_1|\boldsymbol{\eta}) \\
=& \boldsymbol{\beta}^T E\,(\boldsymbol{X}X_1|\boldsymbol{\eta}) + E\,(\delta|\boldsymbol{\eta})\,E\,(X_1|\boldsymbol{\eta}) \quad (\text{because } \delta \perp\!\!\!\perp X|\boldsymbol{\eta}). \\
=& \boldsymbol{\beta}^T E\left[\left.\begin{pmatrix} X_1^2 \\ X_1X_2 \end{pmatrix}\right| \boldsymbol{\eta}\right] = \boldsymbol{\beta}^T \begin{pmatrix} E\left(X_1^2|\boldsymbol{\eta}\right) \\ E\,(X_1X_2|\boldsymbol{\eta}) \end{pmatrix} \\
=& \boldsymbol{\beta}^T \begin{pmatrix} \sigma_1^2 \\ 0 \end{pmatrix} = \beta_1\sigma_1^2.
\end{aligned}$$

Similarly, in the numerator of the right fraction in (3.6), $\text{cov}\,(Y, X_2|\boldsymbol{\eta}) = \beta_2\sigma_2^2$.

In both denominators in (3.6) we have

$$\begin{aligned}
\text{var}\,(Y|\boldsymbol{\eta}) =& \text{var}\left(\boldsymbol{\beta}^T\boldsymbol{X} + \delta|\boldsymbol{\eta}\right) \\
=& \boldsymbol{\beta}^T \text{var}\,(\boldsymbol{X}|\boldsymbol{\eta}) + \text{var}\,(\delta|\boldsymbol{\eta}) + 2\text{cov}\left(\boldsymbol{\beta}^T\boldsymbol{X}\delta|\boldsymbol{\eta}\right) \\
=& \boldsymbol{\beta}^T \text{var}\,(\boldsymbol{X}|\boldsymbol{\eta}) + \text{var}\,(\delta|\boldsymbol{\eta}) \\
=& \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + \text{var}\,(\delta|\boldsymbol{\eta})
\end{aligned}$$

which is a scalar and so we can cancel it from both sides of the inequality in (3.6). Also we have that $\text{var}\,(X_1|\boldsymbol{\eta}) = \sigma_1^2$ and $\text{var}\,(X_2|\boldsymbol{\eta}) = \sigma_2^2$. So the left side of inequality (3.5) becomes

$$P\left(\frac{\beta_1^2\sigma_1^4}{\sigma_1^2} > \frac{\beta_2^2\sigma_2^4}{\sigma_2^2}\middle|\, \sigma_1^2, \sigma_2^2\right) = P\left(\beta_1^2\sigma_1^2 > \beta_2^2\sigma_2^2\middle|\, \sigma_1^2, \sigma_2^2\right) = P\left(\frac{\beta_1^2}{\beta_2^2} > \frac{\sigma_2^2}{\sigma_1^2}\middle|\, \sigma_1^2, \sigma_2^2\right).$$

Similarly the right side becomes

$$P\left(\frac{\beta_1^2\sigma_1^4}{\sigma_1^2} < \frac{\beta_2^2\sigma_2^4}{\sigma_2^2}\middle|\, \sigma_1^2, \sigma_2^2\right) = P\left(\beta_1^2\sigma_1^2 < \beta_2^2\sigma_2^2 \middle|\, \sigma_1^2, \sigma_2^2\right) = P\left(\frac{\beta_1^2}{\beta_2^2} < \frac{\sigma_2^2}{\sigma_1^2}\middle|\, \sigma_1^2, \sigma_2^2\right),$$

and so we have to prove the following inequality.

$$P\left(\frac{\beta_1^2}{\beta_2^2} > \frac{\sigma_2^2}{\sigma_1^2}\middle|\, \sigma_1^2, \sigma_2^2\right) > P\left(\frac{\beta_1^2}{\beta_2^2} < \frac{\sigma_2^2}{\sigma_1^2}\middle|\, \sigma_1^2, \sigma_2^2\right) \tag{3.7}$$

.

Now we know from Lemma 3.1.2 that the median of the distribution of β_1^2/β_2^2 is equal to 1. Also at the beginning of the proof we have assumed that $\sigma_1^2 > \sigma_2^2$ which means $\sigma_2^2/\sigma_1^2 < 1$. By definition of the median the left hand side of inequality in (3.7) is

$$P\left(\frac{\beta_1^2}{\beta_2^2} > \frac{\sigma_2^2}{\sigma_1^2}\middle|\, \sigma_1^2, \sigma_2^2\right) > 1/2$$

and the right side is

$$P\left(\frac{\beta_1^2}{\beta_2^2} < \frac{\sigma_2^2}{\sigma_1^2}\middle|\, \sigma_1^2, \sigma_2^2\right) < 1/2$$

. So we have that (3.7) holds. Similarly we can prove the case when $\sigma_1^2 < \sigma_2^2$. Since now we have proved the theorem conditioned on the values of σ_1^2 and σ_2^2 we have that

$$\begin{aligned} &E\left(P\left(\rho_1\left(\boldsymbol{\beta}\right) > \rho_2\left(\boldsymbol{\beta}\right)\middle|\, \sigma_1^2, \sigma_2^2\right) > P\left(\rho_1\left(\boldsymbol{\beta}\right) < \rho_2\left(\boldsymbol{\beta}\right)\middle|\, \sigma_1^2, \sigma_2^2\right)\right) \\ &= P\left(\rho_1\left(\boldsymbol{\beta}\right) > \rho_2\left(\boldsymbol{\beta}\right)\right) > P\left(\rho_1\left(\boldsymbol{\beta}\right) < \rho_2\left(\boldsymbol{\beta}\right)\right) \end{aligned}$$

which means expression (3.5) holds. □

Until now I have assumed uncorrelated predictors. But this is not the case in the majority of the experiment that have the need of using principal component analysis. So in the next section I will add the last assumption needed to complete the list of assumptions stated in the conjecture.

3.4 Correlated predictors

In this section I will present the final version of the theorem in the case of a 2-dimensional predictor vector $\boldsymbol{X}$. Until now I have shown that the conjecture holds for the lemmas above which are using independent predictors X_1 and X_2. I will show that the conjecture holds for the case where we have dependent predictors X_1 and X_2. The appropriate randomness of the selection of dependent predictors is introduced by rotating the random matrix $\boldsymbol{\Sigma}_0$ with another random matrix Γ. Since this completes the list of assumptions in the conjecture by Li (2007) this is the final result of this Chapter.

Theorem 3.4.1 *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 $\mathbf{\Sigma} = \Gamma \mathbf{\Sigma}_0 \Gamma^T$. *Let* v_1 *and* v_2 *be the first and the second eigenvectors of* $\mathbf{\Sigma}$, *in the sense that* v_1 *corresponds to the larger eigenvalue, and* v_2 *corresponds to the smaller eigenvalue. Let* $\boldsymbol{\beta}$ *be a 2-dimensional random vector satisfying* $\left(\boldsymbol{\beta}^T v_1\right)^2 / \left(\boldsymbol{\beta}^T v_2\right)^2 \overset{\mathcal{D}}{=} \left(\boldsymbol{\beta}^T v_2\right)^2 / \left(\boldsymbol{\beta}^T v_1\right)^2$, $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{X}, \mathbf{\Sigma})$. *Let* $Y = \boldsymbol{\beta}^T \boldsymbol{X} + \delta$ *where* $\delta \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{\beta}, \mathbf{\Sigma})$. *Furthermore, assume that the median of* $\left(\boldsymbol{\beta}^T v_1\right)^2 / \left(\boldsymbol{\beta}^T v_2\right)^2$ *is unique.*

$$\rho_1(\boldsymbol{\beta}, \mathbf{\Sigma}) = \mathrm{corr}^2(Y, v_1^T \boldsymbol{X} | \boldsymbol{\beta}, \mathbf{\Sigma}), \quad \rho_2(\boldsymbol{\beta}, \mathbf{\Sigma}) = \mathrm{corr}^2(Y, v_2^T \boldsymbol{X} | \boldsymbol{\beta}, \mathbf{\Sigma}).$$

Then

$$P(\rho_1(\boldsymbol{\beta}, \mathbf{\Sigma}) > \rho_2(\boldsymbol{\beta}, \mathbf{\Sigma})) > P(\rho_1(\boldsymbol{\beta}, \mathbf{\Sigma}) < \rho_2(\boldsymbol{\beta}, \mathbf{\Sigma})). \tag{3.8}$$

Proof. Let $\boldsymbol{\eta} = \left(\boldsymbol{\beta}, \sigma_1^2, \sigma_2^2\right)$. Now assuming $\sigma_1^2 > \sigma_2^2$ then by definition we have the following equations:

$$\rho_1(\boldsymbol{\beta}, \mathbf{\Sigma} | \boldsymbol{\eta}) = \frac{\mathrm{cov}^2\left(Y v_1^T \boldsymbol{X} | \boldsymbol{\eta}, \mathbf{\Sigma}\right)}{\mathrm{var}(Y | \boldsymbol{\eta}, \mathbf{\Sigma}) \, \mathrm{var}\left(v_1^T \boldsymbol{X} | \boldsymbol{\eta}, \mathbf{\Sigma}\right)}$$

where

$$\begin{aligned}
\mathrm{var}(Y | \boldsymbol{\eta}, \mathbf{\Sigma}) =& \mathrm{var}\left(\boldsymbol{\beta}^T \boldsymbol{X} + \delta | \boldsymbol{\eta}, \mathbf{\Sigma}\right) \\
=& \mathrm{var}\left(\boldsymbol{\beta}^T \boldsymbol{X} | \boldsymbol{\eta}, \mathbf{\Sigma}\right) + \mathrm{var}(\delta | \boldsymbol{\eta}, \mathbf{\Sigma}) + 2\mathrm{cov}\left(\boldsymbol{\beta}^T \boldsymbol{X} \delta | \boldsymbol{\eta}, \mathbf{\Sigma}\right) \\
=& \mathrm{var}\left(\boldsymbol{\beta}^T \boldsymbol{X} | \boldsymbol{\eta}, \mathbf{\Sigma}\right) + \mathrm{var}(\delta | \boldsymbol{\eta}, \mathbf{\Sigma}) = \boldsymbol{\beta}^T \mathbf{\Sigma} \boldsymbol{\beta} + var(\delta | \boldsymbol{\eta}, \mathbf{\Sigma})
\end{aligned}$$

and

$$\mathrm{var}\left(v_1^T \boldsymbol{X} | \boldsymbol{\eta}, \mathbf{\Sigma}\right) = v_1^T \mathbf{\Sigma} v_1 = \sigma_1^2$$

Finally

$$\begin{aligned}\operatorname{cov}\left(Yv_1^T\boldsymbol{X}|\boldsymbol{\eta},\boldsymbol{\Sigma}\right) =& E\left(Yv_1^T\boldsymbol{X}|\boldsymbol{\eta},\boldsymbol{\Sigma}\right) - E\left(Y|\boldsymbol{\eta},\boldsymbol{\Sigma}\right)E\left(v_1^T\boldsymbol{X}|\boldsymbol{\eta},\boldsymbol{\Sigma}\right)\\ =& E\left(\boldsymbol{\beta}^T\boldsymbol{X}v_1^T\boldsymbol{X}|\boldsymbol{\eta},\boldsymbol{\Sigma}\right) + E\left(\delta v_1^T\boldsymbol{X}|\boldsymbol{\eta},\boldsymbol{\Sigma}\right) = \boldsymbol{\beta}^T E\left(\boldsymbol{X}\boldsymbol{X}^T|\boldsymbol{\eta},\boldsymbol{\Sigma}\right)v_1\\ =& \boldsymbol{\beta}^T\boldsymbol{\Sigma}v_1 = \boldsymbol{\beta}^T v_1\sigma_1^2\end{aligned}$$

Combining the four equations above we have that

$$\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}|\,\boldsymbol{\eta}\right) = \frac{\sigma_1^2\left(\boldsymbol{\beta}^T v_1\right)^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + var\left(\delta|\boldsymbol{\eta},\boldsymbol{\Sigma}\right)} \tag{3.9}$$

Similarly we have that:

$$\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}|\,\boldsymbol{\eta}\right) = \frac{\sigma_2^2\left(\boldsymbol{\beta}^T v_2\right)^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} + var\left(\delta|\boldsymbol{\eta},\boldsymbol{\Sigma}\right)} \tag{3.10}$$

By replacing equations (3.9) and (3.10) into expression (3.8) and by canceling from everywhere the denominator $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}$ since it is a positive scalar, we have that expression (3.8) simplifies to:

$$\begin{aligned}P\left(\sigma_1^2\left(\boldsymbol{\beta}^T v_1\right)^2 > \sigma_2^2\left(\boldsymbol{\beta}^T v_2\right)^2\middle|\,\sigma_1^2,\sigma_2^2\right) > P\left(\sigma_1^2\left(\boldsymbol{\beta}^T v_1\right)^2 < \sigma_2^2\left(\boldsymbol{\beta}^T v_2\right)^2\middle|\,\sigma_1^2,\sigma_2^2\right)\\ P\left(\frac{\left(\boldsymbol{\beta}^T v_1\right)^2}{\left(\boldsymbol{\beta}^T v_2\right)^2} > \frac{\sigma_2^2}{\sigma_1^2}\middle|\,\sigma_1^2,\sigma_2^2\right) > P\left(\frac{\left(\boldsymbol{\beta}^T v_1\right)^2}{\left(\boldsymbol{\beta}^T v_2\right)^2} < \frac{\sigma_2^2}{\sigma_1^2}\middle|\,\sigma_1^2,\sigma_2^2\right)\end{aligned} \tag{3.11}$$

Now from the assumption that $\left(\boldsymbol{\beta}^T v_1\right)^2/\left(\boldsymbol{\beta}^T v_2\right)^2 \overset{\mathcal{D}}{=} \left(\boldsymbol{\beta}^T v_2\right)^2/\left(\boldsymbol{\beta}^T v_1\right)^2$ and since the median of $\left(\boldsymbol{\beta}^T v_1\right)^2/\left(\boldsymbol{\beta}^T v_2\right)^2$ must be unique we have that the median of $\left(\boldsymbol{\beta}^T v_1\right)^2/\left(\boldsymbol{\beta}^T v_2\right)^2 = 1$. This proves expression (3.11).

The case where $\sigma_1^2 < \sigma_2^2$ is symmetric and proved in a similar way. With this, expression (3.8) is proved. □

This theorem is very important as it shows that in the general regression context that has an additive error term, the probability, that the squared correlation between the response variable and the first principal component is larger than the squared correlation between the response variable and the second principal component, is greater than 1/2. This result holds in very mild assumptions. This important theorem gives us the proof that in the case of regression with two predictors and an additive error term, one can be confident that using the Principal Components Analysis to reduce the dimension of the problem will get a meaningful result. That is, the resulting first principal component has higher probability of having higher squared correlation with the response, than the second principal component.

Of course, this theorem is almost useless, since the real reason one might need to do dimension reduction using principal component analysis, is when the dimension of the predictor vector is large. Reducing a two dimension vector to one, is desirable in most cases, but the real issue is when you need to reduce a dimension in the order of the tenths, hundreds or even thousands. This is the work presented in the next Chapter of this thesis.

Chapter 4

General Result for p-dimensional vector $\boldsymbol{X}$

Until now I have proved that the conjecture stated by Li B.(2007) can be proven for a 2-dimensional predictor vector $\boldsymbol{X}$. If all the regression problems in real life were with only two predictors that will be great. First, the curse of dimensionality would have been non-existent. Second, we wouldn't have the need to use procedures that reduce the dimension, thus, anything we have discussed until now would not have been useful. Although everything that have been discussed before is important, as they can be used as guiding results, they can be useless if they cannot get extended to a general p-dimensional predictor vector $\boldsymbol{X}$, since the larger the dimension of our predictor vector, the larger the need to have a reliable procedure to reduce the dimension. In this Chapter my main objective is to prove the conjecture stated by Li B.(2007) for a p-dimensional predictor vector $\boldsymbol{X}$.

4.1 Orientationally uniform

One of the main assumptions of the theory I am describing in this work is that the variance of the predictor vector $\boldsymbol{X}$, which is the matrix $\boldsymbol{\Sigma}$ is uniformly distributed among all the possible positive definite *p*x*p* matrices. In this section we give the definition of orientationally uniform matrix, that is, a matrix that makes the random ellipsoid

$$\{\boldsymbol{x} : \boldsymbol{x}^T \boldsymbol{\Sigma}\, \boldsymbol{x} \leq c\}$$

to have any orientation with equal probability.

Before giving the definition, I will define another term that we will see in the definition later.

Definition 4.1.1 *Let say $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$, are p random elements. We say that they are exchangeable if, for any permutation $(i_1, \ldots, i_p)$ of $(1, \ldots, p)$, we have*

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

In other words they are exchangable if I can change their order and still get the same distribution.

Now I have all the components needed to give a definition of what orientationally uniform is.

Definition 4.1.2 *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.*

4.2 Preliminary result

Before I present the main result of this Chapter I need to prove a lemma that will be helpful. Like in the previous Chapter, we need to see under which conditions the distribution of $(\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2$ has unique median that is equal to 1. Next lemma will show that under mild conditions this is satisfied for p-dimensional vectors $\boldsymbol{\beta}$ and $\boldsymbol{v}_1, \boldsymbol{v}_2$ under mild conditions.

Lemma 4.2.1 *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.*

PROOF. First, we shown that 1 is a median of $(\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2$; that is,

$$P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 < 1) \le 1/2 \le P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 \le 1). \tag{4.1}$$

Because $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ are exchangeable and $\boldsymbol{\beta} \perp\!\!\!\perp (\boldsymbol{v}_1, \boldsymbol{v}_2)$, the random variables $(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2$ and $(\boldsymbol{\beta}^T\boldsymbol{v}_2)^2$ are exchangeable. Hence

$$\begin{aligned} P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 \le 1) =& P((\boldsymbol{\beta}^T\boldsymbol{v}_1)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_2)^2 \le 1) \\ =& 1 - P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 < 1). \end{aligned}$$

So

$$\begin{aligned} P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 < 1) &\le 1 - P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 < 1), \\ P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 \le 1) &\ge 1 - P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 \le 1), \end{aligned}$$

which imply (4.1).

Now we need to show that 1 is the unique median. That is, (4.1) is satisfied by no other numbers. In other words, for any $0 < c_1 < 1$ and $c_2 > 1$ we have

$$P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 \le c_1) < 1/2 \quad \text{and} \quad P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 < c_2) > 1/2.$$

We will only show the first inequality; the second can be shown similarly. Since

$$P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 \le c_1) = E[P((\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2 \le c_1 | \boldsymbol{v}_1, \boldsymbol{v}_2)],$$

it suffices to show that for any nonrandom, linearly independent $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, we have

$$P((\boldsymbol{\beta}^T \boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T \boldsymbol{v}_1)^2 \leq c_1 | (\boldsymbol{v}_1, \boldsymbol{v}_2) = (\boldsymbol{a}, \boldsymbol{b})) < 1/2.$$

However, because $(\boldsymbol{v}_1, \boldsymbol{v}_2) \perp\!\!\!\perp \boldsymbol{\beta}$, the above inequality is equivalent to

$$P((\boldsymbol{\beta}^T \boldsymbol{b})^2/(\boldsymbol{\beta}^T \boldsymbol{a})^2 \leq c_1) < 1/2. \tag{4.2}$$

Let $c_3 \in (c_1, 1)$. Since $(\boldsymbol{a}, \boldsymbol{b})$ has full column rank, the following system of equations

$$\begin{cases} \boldsymbol{\beta}^T \boldsymbol{b} = \sqrt{c_3} \\ \boldsymbol{\beta}^T \boldsymbol{a} = 1 \end{cases}$$

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

$$\boldsymbol{\beta} \in G \Rightarrow (\boldsymbol{\beta}^T \boldsymbol{b})^2/(\boldsymbol{\beta}^T \boldsymbol{a})^2 \in (c_1, 1).$$

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

$$P((\boldsymbol{\beta}^T \boldsymbol{b})^2/(\boldsymbol{\beta}^T \boldsymbol{a})^2 \in (c_1, 1)) > 0,$$

which, combined with (4.1), implies (4.2). □

The above result might not be clear why is useful, but reading the main theorem of this Chapter that follows in the next section, the reader should understand why

we need the median to be unique.

4.3 Main theorem

Until now, we have shown all the useful tools that we need in order to state and proof the main theorem of this work. In this section, I am going to precisely state and rigorously proof the theorem that has been the main objective of this work.

Theorem 4.3.1 *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* $i < j$*,* $P(\rho_i \geq \rho_j) > 1/2$*.*

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

$$\begin{aligned}\mathrm{cov}(Y, \boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) =& \mathrm{cov}(\boldsymbol{\beta}^T\boldsymbol{X} + \delta, \boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ =& \boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{v}_{(i)} + \mathrm{cov}(\delta, \boldsymbol{v}^T_{(i)}\boldsymbol{X}|\boldsymbol{\beta}, \boldsymbol{\Sigma}). \qquad (4.3)\end{aligned}$$

Because $\delta \perp\!\!\!\perp (\boldsymbol{\Sigma}, \boldsymbol{X}, \boldsymbol{\beta})$, we have $\delta \perp\!\!\!\perp (\boldsymbol{v}_{(i)}^T \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$. This implies $\delta \perp\!\!\!\perp \boldsymbol{v}_{(i)}^T \boldsymbol{X} | (\boldsymbol{\beta}, \boldsymbol{\Sigma})$, and hence that the second term in (4.3) is zero. Because $(\sigma_{(i)}^2, \boldsymbol{v}_{(i)})$ is an eigen pair of $\boldsymbol{\Sigma}$, we have $\boldsymbol{\Sigma}\boldsymbol{v}_{(i)} = \sigma_{(i)}^2 \boldsymbol{v}_{(i)}$. Hence

$$\mathrm{cov}^2(Y, \boldsymbol{v}_{(i)}^T \boldsymbol{X} | \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sigma_{(i)}^4 (\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2. \tag{4.4}$$

In the meantime

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

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

$$\mathrm{var}(Y | \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2. \tag{4.5}$$

Moreover,

$$\mathrm{var}(\boldsymbol{v}_{(i)}^T \boldsymbol{X} | \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \boldsymbol{v}_{(i)}^T \boldsymbol{\Sigma} \boldsymbol{v}_{(i)} = \sigma_{(i)}^2. \tag{4.6}$$

Now combine (4.4), (4.5), and (4.6) to obtain

$$\rho_i = \mathrm{corr}(Y, \boldsymbol{v}_{(i)}^T \boldsymbol{X}) = \frac{\sigma_{(i)}^2 (\boldsymbol{\beta}^T \boldsymbol{v}_{(i)})^2}{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} + \tau^2}. \tag{4.7}$$

Let $i < j$. Then, using (4.7),

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

The right hand side can be written as

$$\sum_{k\neq l} P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_{(i)})^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_{(j)})^2} \geq \frac{\sigma^2_{(j)}}{\sigma^2_{(i)}}\middle| \sigma^2_{(i)}=\sigma^2_k, \sigma^2_{(j)}=\sigma^2_\ell\right) P\left(\sigma^2_{(i)}=\sigma^2_k, \sigma^2_{(j)}=\sigma^2_\ell\right).$$

Because $\sigma^2_1,\ldots,\sigma^2_p$ are exchangeable, $(\sigma^2_{(i)},\sigma^2_{(j)})$ has equal probability to be any pair $(\sigma^2_k,\sigma^2_\ell)$ for any $k\neq\ell$, and that probability is $\binom{p}{2}$. Hence the above reduces to

$$\begin{aligned}&\binom{p}{2}^{-1}\sum_{k\neq l} P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_{(i)})^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_{(j)})^2} \geq \frac{\sigma^2_{(j)}}{\sigma^2_{(i)}}\middle| \sigma^2_{(i)}=\sigma^2_k, \sigma^2_{(j)}=\sigma^2_\ell\right)\\ &=\binom{p}{2}^{-1}\sum_{k\neq l} P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{\sigma^2_\ell}{\sigma^2_k}\middle| \sigma^2_{(i)}=\sigma^2_k, \sigma^2_{(j)}=\sigma^2_\ell\right),\end{aligned} \tag{4.8}$$

where the equality follows from the fact that, conditioning on the event $(\sigma^2_{(i)},\sigma^2_{(j)}) = (\sigma^2_k,\sigma^2_\ell)$, one has $(\boldsymbol{v}^2_{(i)},\boldsymbol{v}^2_{(j)}) = (\boldsymbol{v}^2_k,\boldsymbol{v}^2_\ell)$.

Reexpress each term in the summation in (4.8) as

$$E\left[P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{\sigma^2_\ell}{\sigma^2_k}\middle| \sigma^2_{(i)}=\sigma^2_k, \sigma^2_{(j)}=\sigma^2_\ell, \sigma^2_k, \sigma^2_\ell\right)\middle| \sigma^2_{(i)}=\sigma^2_k, \sigma^2_{(j)}=\sigma^2_\ell\right]. \tag{4.9}$$

By part 3 of Definition 4.1.2 we have

$$\begin{aligned}(\boldsymbol{v}_k,\boldsymbol{v}_\ell)\perp\!\!\!\perp(\sigma^2_1,\ldots,\sigma^2_p) &\Rightarrow (\boldsymbol{v}_k,\boldsymbol{v}_\ell)\perp\!\!\!\perp(\sigma^2_1,\ldots,\sigma^2_p;\sigma^2_{(1)},\ldots,\sigma^2_{(p)})\\ &\Rightarrow (\boldsymbol{v}_k,\boldsymbol{v}_\ell)\perp\!\!\!\perp(\sigma^2_k,\sigma^2_\ell,\sigma^2_{(i)},\sigma^2_{(j)})\\ &\Rightarrow (\boldsymbol{v}_k,\boldsymbol{v}_\ell)\perp\!\!\!\perp(\sigma^2_{(i)},\sigma^2_{(j)})|(\sigma^2_k,\sigma^2_\ell).\end{aligned}$$

So the event $\{\sigma^2_k=\sigma^2_{(i)},\sigma^2_\ell=\sigma^2_{(j)}\}$ can be removed from the conditional probability inside the conditional expectation (4.9), which then reduces to

$$E\left[P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{\sigma^2_\ell}{\sigma^2_k}\middle| \sigma^2_k,\sigma^2_\ell\right)\middle| \sigma^2_{(i)}=\sigma^2_k, \sigma^2_{(j)}=\sigma^2_\ell\right]. \tag{4.10}$$

Because $(\boldsymbol{\beta}, \boldsymbol{v}_k, \boldsymbol{v}_\ell) \perp\!\!\!\perp (\sigma_k^2, \sigma_\ell^2)$, for each fixed $0 < s < t$,

$$P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{s}{t}\middle|\, \sigma_k^2 = t, \sigma_\ell^2 = s\right) = P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{s}{t}\right) > 1/2$$

where the inequality follows from Lemma 4.2.1. By part 1 of Definition 4.1.2, the event $\{\sigma_k^2 = \sigma_\ell^2\}$ has probability 0. It follows that

$$P\left(\frac{(\boldsymbol{\beta}^T\boldsymbol{v}_k)^2}{(\boldsymbol{\beta}^T\boldsymbol{v}_\ell)^2} \geq \frac{\sigma_\ell^2}{\sigma_k^2}\middle|\, \sigma_k^2, \sigma_\ell^2\right) > 1/2$$

almost surely on the event $\{\sigma_{(i)}^2 = \sigma_k^2, \sigma_{(j)}^2 = \sigma_\ell^2\}$. Therefore (4.10), and hence (4.8), are strictly greater than 1/2. □

The proof of the theorem above, completes the objective of proving the conjecture stated by Li (2007). It proves that the 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 is proving that, the probability that we will get the principal components that are more correlated with the responce variable, is greater than the probability to get the principal components that are less correlated.

In other words this result is proving the conjecture by assuming that you have an orientationally uniform distribution for $\boldsymbol{\Sigma} = \text{var}(\boldsymbol{X}|\boldsymbol{\Sigma})$ and a unique median for $(\boldsymbol{\beta}^T\boldsymbol{v}_2)^2/(\boldsymbol{\beta}^T\boldsymbol{v}_1)^2$ in the usual regression context where $Y = \boldsymbol{\beta}^T\boldsymbol{X} + \delta$, 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 $\text{var}(\delta) < \infty$. Satisfying those assumptions we have proved that the probability the principal component analysis will give you the

highest correlated principal components with the response variable is greater than the probability it will give you the less correlated principal components with the response. This probability comparing any two principal component is always in favor of the principal component that corresponds to the higher eigenvalue.

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 us 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, we have proved that one can confidently use principal component analysis to find the principal components and being confident he will get the mostly correlated principal components with the response.

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 already use this procedure in their work. The proof, is based on probabilities, so the ones that were critical against the use of principal component analysis, will probably continue to be thinking critically against it. On the other hand, those that are in favor of the principal component analysis, 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.

Chapter 5

Stochastic ordering

In this Chapter, I will present my work, that tries to connect the inequality in the conjecture with stochastic ordering. People are usually more familiar with stochasting ordering that the inequality in Theorem 4.3.1, so I make a try to find a relation between the two.

Generally, I believe that stochasting ordering is neither stronger or weaker than the inequality we have proved in Theorem 4.3.1 . I will show that there are some cases were stochasting ordering is stronger under certain assumptions.

5.1 Definition

In the work by Li, Zha, and Chiaromonte, (2005) we can find the definition of stochastic ordering.

Definition 5.1.1 *If A and B are two random variables, and for any real number c, $P(A \leq c) \geq P(B \leq c)$ then A is said to be stochastically less than or equal to B,*

and is denoted with $A \overset{\mathcal{D}}{\leq} B$. *If the inequality is strict on a subset of the real line with positive Lebesque measure then* A *is said to be stochastically strictly less than* B, *and is denoted with* $A \overset{\mathcal{D}}{<} B$.

In this Chapter I will try to explore the sufficient conditions necessary where stochastic ordering implies

$$P\left(\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) > \rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)\right) > P\left(\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) < \rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)\right) \tag{5.1}$$

5.2 Results under the regression context

At first, I will prove a Lemma that assumes uncorrelated predictors with non-random covariance matrix $\boldsymbol{\Sigma}$ in a regression context without error term. Later, I will show a much simpler proof for the same result, which is based on the fact that $\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) + \rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) = 1$, and also I will generalize it for the case where $\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) + \rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) = k$, for any real number k.

Lemma 5.2.1 *Suppose that* $\boldsymbol{X}|\boldsymbol{\Sigma} \sim N(0, \boldsymbol{\Sigma})$, *and that*

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

where σ_1^2 *and* σ_2^2 *are iid* G *with* $P(\sigma_1^2 = \sigma_2^2) = 0$. *Let* $\boldsymbol{\beta} \perp\!\!\!\perp \left(\boldsymbol{X}, \boldsymbol{\Sigma}\right)$ *and* $Y = \boldsymbol{\beta}^T \boldsymbol{X}$. *Let*

$$\rho_1\left(\boldsymbol{\beta}\right) = \begin{cases} \operatorname{corr}^2\left(Y, X_1|\boldsymbol{\beta}\right) & \text{if } \sigma_1^2 > \sigma_2^2 \\ \operatorname{corr}^2\left(Y, X_2|\boldsymbol{\beta}\right) & \text{if } \sigma_1^2 \leq \sigma_2^2 \end{cases} \text{ and } \rho_2\left(\boldsymbol{\beta}\right) = \begin{cases} \operatorname{corr}^2\left(Y, X_1|\boldsymbol{\beta}\right) & \text{if } \sigma_1^2 \leq \sigma_2^2 \\ \operatorname{corr}^2\left(Y, X_2|\boldsymbol{\beta}\right) & \text{if } \sigma_1^2 > \sigma_2^2 \end{cases}.$$

If

$$\rho_1(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \overset{\mathcal{D}}{>} \rho_2(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \tag{5.2}$$

then

$$P\left(\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) > \rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)\right) > P\left(\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) < \rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)\right). \tag{5.3}$$

PROOF. By definition of stochastic ordering, expression (5.2) implies

$$P(\rho_2(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \leq r) \geq P(\rho_1(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \leq r) \tag{5.4}$$

for any real number r, with strict inequality on a subset of the real line with positive Lebesque measure. The above expression is meaningful when r is in the interval $(0, 1)$, since $\rho_1(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ and $\rho_2(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ can take values only in that interval. The above expression can be re-written as

$$\begin{aligned} &P\left(\rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \leq r | \sigma_1 > \sigma_2\right) P\left(\sigma_1 > \sigma_2\right) + P\left(\rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \leq r | \sigma_2 \geq \sigma_1\right) P\left(\sigma_2 \geq \sigma_1\right) \geq \\ &P\left(\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \leq r | \sigma_1 > \sigma_2\right) P\left(\sigma_1 > \sigma_2\right) + P\left(\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \leq r | \sigma_2 \geq \sigma_1\right) P\left(\sigma_2 \geq \sigma_1\right). \end{aligned}$$

Since $P\left(\sigma_2 \geq \sigma_1\right) = P\left(\sigma_1 > \sigma_2\right)$, the above simplifies to

$$\begin{aligned} &P\left(\rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \leq r | \sigma_1 > \sigma_2\right) + P\left(\rho_2\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \leq r | \sigma_2 \geq \sigma_1\right) \geq \\ &P\left(\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \leq r | \sigma_1 > \sigma_2\right) + P\left(\rho_1\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) \leq r | \sigma_2 \geq \sigma_1\right). \end{aligned} \tag{5.5}$$

Now by definition, for $i = 1, 2$,

$$\text{corr}^2\left(Y, X_i|\boldsymbol{\beta}\right) = \frac{\text{cov}^2\left(Y, X_i|\boldsymbol{\beta}\right)}{\text{var}\left(Y|\boldsymbol{\beta}\right)\text{var}\left(X_i\right)} = \frac{E^2\left(YX_i|\boldsymbol{\beta}\right)}{\sigma_i^2\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}} = \frac{\beta_i^2\sigma_i^4}{\sigma_i^2\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}} = \frac{\beta_i^2\sigma_i^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}$$

where $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} = \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2$. Substituting this into (5.5) to obtain

$$P\left(\frac{\beta_2^2\sigma_2^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}} \leq r|\sigma_1 > \sigma_2\right) + P\left(\frac{\beta_1^2\sigma_1^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}} \leq r|\sigma_2 \geq \sigma_1\right) \geq$$
$$P\left(\frac{\beta_1^2\sigma_1^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}} \leq r|\sigma_1 > \sigma_2\right) + P\left(\frac{\beta_2^2\sigma_2^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}} \leq r|\sigma_2 \geq \sigma_1\right) \Rightarrow$$

$$P\left(\beta_2^2\sigma_2^2 \leq r\left(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2\right)|\sigma_1 > \sigma_2\right) + P\left(\beta_1^2\sigma_1^2 \leq r\left(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2\right)|\sigma_2 \geq \sigma_1\right) \geq$$
$$P\left(\beta_1^2\sigma_1^2 \leq r\left(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2\right)|\sigma_1 > \sigma_2\right) + P\left(\beta_2^2\sigma_2^2 \leq r\left(\beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2\right)|\sigma_2 \geq \sigma_1\right) \Rightarrow$$

$$P\left((1-r)\,\beta_2^2\sigma_2^2 \leq r\beta_1^2\sigma_1^2|\sigma_1 > \sigma_2\right) + P\left((1-r)\,\beta_1^2\sigma_1^2 \leq r\beta_2^2\sigma_2^2|\sigma_2 \geq \sigma_1\right) \geq$$
$$P\left((1-r)\,\beta_1^2\sigma_1^2 \leq r\beta_2^2\sigma_2^2|\sigma_1 > \sigma_2\right) + P\left((1-r)\,\beta_2^2\sigma_2^2 \leq r\beta_1^2\sigma_1^2|\sigma_2 \geq \sigma_1\right)$$

Now since (5.4) is true for everyvalue of r, that means it is true for $r = 1/2$. So by replacing r in the above expression we have that

$$P\left(\frac{1}{2}\beta_2^2\sigma_2^2 \leq \frac{1}{2}\beta_1^2\sigma_1^2|\sigma_1 > \sigma_2\right) + P\left(\frac{1}{2}\beta_1^2\sigma_1^2 \leq \frac{1}{2}\beta_2^2\sigma_2^2|\sigma_2 \geq \sigma_1\right) \geq$$
$$P\left(\frac{1}{2}\beta_1^2\sigma_1^2 \leq \frac{1}{2}\beta_2^2\sigma_2^2|\sigma_1 > \sigma_2\right) + P\left(\frac{1}{2}\beta_2^2\sigma_2^2 \leq \frac{1}{2}\beta_1^2\sigma_1^2|\sigma_2 \geq \sigma_1\right) \Rightarrow$$

$$P\left(\beta_2^2\sigma_2^2 \leq \beta_1^2\sigma_1^2|\sigma_1 > \sigma_2\right) + P\left(\beta_1^2\sigma_1^2 \leq \beta_2^2\sigma_2^2|\sigma_2 \geq \sigma_1\right) \geq$$
$$P\left(\beta_1^2\sigma_1^2 \leq \beta_2^2\sigma_2^2|\sigma_1 > \sigma_2\right) + P\left(\beta_2^2\sigma_2^2 \leq \beta_1^2\sigma_1^2|\sigma_2 \geq \sigma_1\right) \tag{5.6}$$

Next, re-express inequality (5.3) as

$$P\left(\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)>\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)|\sigma_1>\sigma_2\right)P\left(\sigma_1>\sigma_2\right)+P\left(\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)>\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)|\sigma_2\geq\sigma_1\right)P\left(\sigma_2\geq\sigma_1\right)>$$
$$P\left(\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)<\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)|\sigma_1>\sigma_2\right)P\left(\sigma_1>\sigma_2\right)+P\left(\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)<\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)|\sigma_2\geq\sigma_1\right)P\left(\sigma_2\geq\sigma_1\right)\Rightarrow$$

$$P\left(\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)>\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)|\sigma_1>\sigma_2\right)+P\left(\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)>\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)|\sigma_2\geq\sigma_1\right)>$$
$$P\left(\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)<\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)|\sigma_1>\sigma_2\right)+P\left(\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)<\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)|\sigma_2\geq\sigma_1\right)\Rightarrow$$

$$P\left(\frac{\beta_1^2\sigma_1^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}>\frac{\beta_2^2\sigma_2^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}\middle|\,\sigma_1>\sigma_2\right)+P\left(\frac{\beta_2^2\sigma_2^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}>\frac{\beta_1^2\sigma_1^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}\middle|\,\sigma_2\geq\sigma_1\right)>$$
$$P\left(\frac{\beta_1^2\sigma_1^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}<\frac{\beta_2^2\sigma_2^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}\middle|\,\sigma_1>\sigma_2\right)+P\left(\frac{\beta_2^2\sigma_2^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}<\frac{\beta_1^2\sigma_1^2}{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}}\middle|\,\sigma_2\geq\sigma_1\right)\Rightarrow$$

$$P\left(\beta_1^2\sigma_1^2>\beta_2^2\sigma_2^2\middle|\,\sigma_1>\sigma_2\right)+P\left(\beta_2^2\sigma_2^2>\beta_1^2\sigma_1^2\middle|\,\sigma_2\geq\sigma_1\right)>$$
$$P\left(\beta_1^2\sigma_1^2<\beta_2^2\sigma_2^2\middle|\,\sigma_1>\sigma_2\right)+P\left(\beta_2^2\sigma_2^2<\beta_1^2\sigma_1^2\middle|\,\sigma_2\geq\sigma_1\right)$$

which is the same as expression (5.6). This completes our proof. □

This proof it is complete, but I figured out that there is an easier way to prove the same results and it us based on the fact that the sum of the two squared correlations is known to be equal to 1, that is, $\rho_1\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)+\rho_2\left(\boldsymbol{\beta},\boldsymbol{\Sigma}\right)=1$. This is how one can prove the above result much more easier.

Lemma 5.2.2 *If ρ_1 and ρ_2 are random variables such that $\rho_1+\rho_2=1$, then $\rho_1\overset{\mathcal{D}}{\geq}\rho_2$ implies $P(\rho_1\geq\rho_2)\geq P(\rho_1\leq\rho_2)$.*

PROOF. Because $\rho_1 \overset{\mathcal{D}}{\geq} \rho_2$, we have

$$P(\rho_2 \leq 1/2) \geq P(\rho_1 \leq 1/2). \tag{5.7}$$

But because $\rho_1 + \rho_2 = 1$, we have

$$P(\rho_2 \leq 1/2) = P(\rho_2 \leq (\rho_1 + \rho_2)/2) = P(\rho_2 \leq \rho_1).$$

Similarly, $P(\rho_1 \leq 1/2) = P(\rho_1 \leq \rho_2)$. Substitute these into (5.7) to complete the proof. □

Now in the case of Lemma 5.2.1 it is easy to check that $\rho_1(\boldsymbol{\beta}, \boldsymbol{\Sigma}) + \rho_2(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = 1$ for all $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. So Lemma 5.2.1 follows from Lemma 5.2.2 much more easier than the way it was proven above.

But again the Lemma 5.2.2 is a special case of the Theorem that assumes $\rho_1(\boldsymbol{\beta}, \boldsymbol{\Sigma}) + \rho_2(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = k$ where k is any real number. The proof of the Theorem that generalizes Lemma 5.2.2 is as follows.

Lemma 5.2.3 *If ρ_1 and ρ_2 are random variables such that $\rho_1 + \rho_2 = k$, then $\rho_1 \overset{\mathcal{D}}{\geq} \rho_2$ implies $P(\rho_1 \geq \rho_2) \geq P(\rho_1 \leq \rho_2)$.*

PROOF. Because $\rho_1 \overset{\mathcal{D}}{\geq} \rho_2$, we have

$$P(\rho_2 \leq k/2) \geq P(\rho_1 \leq k/2). \tag{5.8}$$

But because $\rho_1 + \rho_2 = k$, we have

$$P(\rho_2 \leq k/2) = P(\rho_2 \leq (\rho_1 + \rho_2)/2) = P(\rho_2 \leq \rho_1).$$

Similarly, $P(\rho_1 \leq k/2) = P(\rho_1 \leq \rho_2)$. Substitute these into (5.8) to complete the proof. □

In the context of the previous Chapters this is how far one can connect stochastic ordering with the Inequality in Theorem 4.3.1. In the next section I will give a result in a more general measure theoretic context.

5.3 Results in more general context

Also, there is a more general case that this is true and it will be proved in this section as a proposition. Let me first describe the notation I will use in this section.

For $i = 1, 2$, let F_i be the distribution of U_i and f_i be the density of U_i with respect to μ; that is $f_i = dF_i/d\mu$. We say that U_1 and U_2 have a common support if $\{f_1 > 0\} = \{f_2 > 0\}$. It is easy to see that if $U_1 \overset{D}{<} U_2$ and U_1 and U_2 have a common support, then

$$F_1(\{c : F_1(c) > F_2(c)\}) > 0, \quad F_2(\{c : F_1(c) > F_2(c)\}) > 0. \tag{5.9}$$

The following proposition gives a sufficient condition for $U_1 \overset{\mathcal{D}}{<} U_2$ to imply $P(U_1 \leq U_2) > 1/2$.

Proposition 5.3.1 *Suppose U_1 and U_2 are random variables whose distributions*

are dominated by a common measure μ*;* $U_1 \overset{\mathcal{D}}{<} U_2$*;* $U_1 \perp\!\!\!\perp U_2$*; and* U_1 *and* U_2 *have a common support. Then* $P(U_1 \leq U_2) > 1/2$*.*

PROOF. By independence of U_1 and U_2 and by Fubini's theorem,

$$\begin{aligned} P(U_1 \leq U_2) &= \int_{\mathbb{R}} \left[\int_{u_1 \leq u_2} f_1(u_1)\mu(du_1) \right] f_2(u_2)\mu(du_2) \\ &= \int_{\mathbb{R}} F_1(u_2) f_2(u_2)\mu(du_2) = \int_{\mathbb{R}} F_1(u_2) dF_2(u_2). \end{aligned}$$

By the second inequality in (5.9) the right hand side above is (strictly) greater than

$$\int_{\mathbb{R}} F_2(u_2) dF_2(u_2) = \left[F_2^2(u_2)/2 \right]_{-\infty}^{\infty} = 1/2,$$

which completes the proof. □

So overall, in this Chapter I have presented some cases where the inequality in the in Theorem 4.3.1 is weaker than sochasting ordering. I believe that generally neither is weaker or stronger than the other, but in some special cases we can see that stochasting ordering is stronger.

Although, one may expand this section further and find other cases that stochastic ordering is stronger, it was not clear to me if we can find more general cases than the ones described above. Also, I believe it will not offer anything more in the main objective of this work, that was the proof of the conjecture as it was stated in Li (2007).

Chapter 6

Conclusion

In this thesis I have presented a probabilistic explanation as to why in the regression setting the response variable often tends to have stronger correlation with the first few principal components of the predictor.

This provides an answer to a historical debate among statisticians. Scientists were able to understand in general why this phenomenon emerges. They were also able to understand why there are cases where the least principal components are more correlated with the response than the primary principal components. The problem is that they couldn't predict when and under which conditions this event would happen. In this thesis, my main objective was to show that the response is more likely to have stronger correlation with the leading principal components than with the least principal components. This, however, doesn't answer under which assumptions principal component cnalysis might fail to give us the correct results. Moreover, this can be used as a satisfying condition from both groups of researchers. Those in favor of using PCA, can argue that since the probability of getting the correct principal components is higher than the probability of getting the wrong

ones using PCA, then PCA can be used effectively to reduce the dimension of a problem. On the other hand, those against the use of PCA, can argue that I have also shown that there is an unmeasurable risk in this procedure to get the wrong principal components. I need to emphasize, that the importance of the proof is exactly the fact that the probability of getting the wrong results is less than the probability of getting the correct results.

Also, this work provides an answer, by formulating and rigorously proving, a theorem that explains nothing more than a natural phenomenon. It proves that nature is fair in choosing the response variable for a set of predictors, as Mosteller and Tukey (1977) thought it is. Again, it is fair, but it is not always fair as there are cases where this might not work. This natural tendency is neither definite nor particularly strong. The inequality $P(\rho_i \geq \rho_j) > 1/2$ says nothing more than that it is more likely for ρ_i to be greater than ρ_j than to be less than ρ_j. In fact, Proposition 5.3.1 indicates that the inequality is weaker than the commonly used stochastic ordering in a special case. While this tendency does imply that performing principal component analysis on $\boldsymbol{X}$ produces a viable predictor of Y, it is much more effective to reduce the dimension of $\boldsymbol{X}$ using the information of Y, for which the extensive research in sufficient dimension reduction (see, for example, Li, 1991, 1992; Cook, 1994, 1996: Li and Wang, 2007) has provided ample evidence.

6.1 Future work

There are many ways that this work may extend. In this section I will list some of the proposals I have although future research may not be limited to only these ideas.

First of all, one can extend the results in the multivariate setting for the responce variable Y. It is important that many experiments today involve multivariate responses and the correlation of each predictor might be different for each component of the response variable. There is a possibility that this may make the analysis more complicated.

Secondly, this work can extend in other type of regression functions. Here we have assumed the normal model for the regression function with an additive error term that follows normal distribution. This is the simplest form of the regression and one might extend this work to more complicated functions of regressing the predictors to the response variable. There is a portion of experiments today that do not use this simple form of regression that I have used.

Finally, one may extend this research to the cases where we have categorical predictors, or a mixture of categorical and continuous predictors. Categorical predictors, might lead to complicated situations where under one value of the predictor the principal component analysis gives us the principal components that are higlhy correlated with the response and under another value of the predictor it gives us principal components that are least correlated with the response.

BIBLIOGRAPHY

Adcock, R. J. (1878). A problem in least squares. *The Analyst*, **5**, 53–54.

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.

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

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.

Cook, R.D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proceedings of the Section on Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, 18–25.

Cook, R. D. (1996). Graphics for regressions with a binary response.*Journal of the American Statistical Association*, **91**, 983–992.

Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science*, **22**, 1–40.

Cook, R. D., Li, B. and Chiaromonte F. (2007). Dimension reduction without matrix inversion. *Biometrika*, **94**, 569–584.

Cox, D. R. (1968).Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society, Ser. A*, **131**, 265–279.

Hadi, A. S. and Ling, R. F. (1998). Some cautionalry notes on the use pf principal components in regression. *The American Statistician*, **52**, 15–19.

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.

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.

Joliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, **31**, 300–303.

Joliffe, I. T. (2002). *Principsl 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.

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.

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, 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.

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.

Scott, D. (1992). *Multivariate Density Estimation*. New York: Wiley.

Tipping M. E. and Bishop, C. M. (1999). Probabilistic principal components. *Journal of the Royal Statistical Society, Series B*, **61**, 611–622.