5-3-2019

Some Statistical Properties of Spectral Regression Estimators

Nawal Hassan
nh4821@rit.edu

Rochester Institute Of Technology

Master's Thesis

Some Statistical Properties of Spectral Regression Estimators

---

Author:
Nawal Hassan

Supervisor:
Dr. Ernest Fokoue

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

in the

Applied Statistics Graduate Program
School of Mathematical Sciences
College of Science

May 3, 2019

Committee Approval:

Ernest Fokoue, Associate Professor, School Of Mathematical Sciences Thesis Advisor Signature:	Date:
Robert Parody, Associate Professor, School Of Mathematical Sciences Committee Member Signature:	Date:
Joseph Voelkel, Professor, School Of Mathematical Sciences Committee Member Signature:	Date:

ROCHESTER INSTITUTE OF TECHNOLOGY

Abstract

School of Mathematical Sciences
College of Science

Master of Science in Applied Statistics

Some Statistical Properties of Spectral Regression Estimators

by Nawal Hassan

In this thesis we explore different Spectral Regression Estimators in order to solve the problem in regression where we have multiple columns that are linearly dependent: We explore two scenarios

- Scenario 1: $p << n$ where there exists at least two columns; x_j and x_k that are nearly linearly dependent which indicates co-linearity and $\boldsymbol{X}^\top \boldsymbol{X}$ becomes near singular.
- Scenario 2: $n << p$ since there are more predictors than observations so some columns must be a linear combination of another column which indicates linear dependence.

The scenarios give us an ill conditioned matrix of $X^\top X$ (when solving the normal equation) due to collinearity issues and the matrix becomes singular and makes the least squares estimate unstable and impossible to compute. In the paper, we explore different methods (variable selection, regularization, compression and dimensionality reduction) that solves the above issue. For variable selection techniques, we use Stepwise Selection Regression as well as the method of Best Subset Selection regression. Two approaches for Stepwise Selection regression are assessed in the paper: Forward Selection and Backward Elimination. Performance assessment of our regression models will be made based on criterion based procedures like AIC,BIC,R^2,R^2 adjusted and the Mallow's C_P statistic. In chapter three of this paper we introduce the concepts of General Regularization, Ridge Regression as well as subsequent shrinkage methods such as the Lasso, Bayesian Lasso and the Elastic net. Chapter five will look at Compression and Dimensionality reduction procedures which are outlined via SVD (Singular Value Decomposition) and Eigenvector Decomposition. Hard thresholding is subsequently introduced via SPCA (Sparse Principle Component Analysis) and a novel approach using RPCA (Robust Principle Component Analysis). Furthermore, RPCA also shows how it can aid with data and image compression. The basis of this study is concluded with an empirical exploration of all the methods outlined above using several performance indicators on simulated data and real data sets. Assessment of the data sets is done via cross-validation. We determine the optimal values of the settings and then evaluate the predictive and explanatory performance.

Acknowledgements

The writing of this thesis was made possible by some of the most wonderful people that I have ever had the chance to work with and be around. First and foremost, I would like to thank my thesis advisor Professor Ernest Fokoue for his full support, patience, wisdom and guidance throughout this thesis and throughout my entire Masters program. His presence in my life was undoubtedly the key to my ability to come this far and I feel extremely blessed to have had the privilege of his mentorship.

I would like to thank my committee for their great insight, constructive criticism and constant support and motivation. I owe a special thank you to the entire department and staff for their patience and cooperation with me; no question was left unanswered and for that I owe them my most sincere gratitude. Last but not least I would like to thank my family for being there for me throughout this entire journey and for always having my back at the lowest and highest of times. It is their love and unconditional support that kept me going.

Contents

List of Figures

List of Tables

To my dear family

Chapter 1

Introduction and Problem Specification

1.1 Summary of Properties of Multiple Regression and Data Sets

We will first start off with defining the multiple linear regression model in which many statistical learning approaches are based on. This model is defined as follows:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \qquad \text{for } i = 1, 2, \cdots, n \tag{1.1}$$

This multiple linear regression model studies the relationship between a response variable Y_i and its corresponding predictors $x_i = (x_{i1}, \cdots, x_{ip})$ for a given sample n.

The regression coefficients, $\beta_0, \beta_1, \cdots, \beta_p$, of the predictors are unknown and are necessary for the model. A work around this is to estimate them and this will be shown later using the least square estimator. Lastly the errors are random and identically distributed (i.i.d.) with mean 0 and variance σ^2.

An alternative form for the regression model can be written in matrix form in the following way:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.2}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{1.3}$$

We have from the model assumptions:

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \text{ and } Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n \tag{1.4}$$

Where $\boldsymbol{I}_n$ is the identity matrix of order n where:

$$\boldsymbol{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n}$$

We now define the objective function or the least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ to be able to get the estimated coefficients and in turn, the estimated response. In order to do this we must first define the residual sum of squares (RSS) as follows:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}))^2 \tag{1.5}$$

If we take the minimizer of (1.5) we yield the ordinary least squares (OLS) estimate:

We can now get the predicted or fitted values of the response variable which is given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip,} \quad \text{for } i = 1, 2, \cdots, n. \tag{1.6}$$

Another way to define the RSS is:

$$RSS = \sum (y_i - \hat{y}_i)^2 \tag{1.7}$$

The RSS is a measure of the discrepancy between the data and an estimation model. A small RSS indicates a tight fit of the model to the data. It is used as an optimality criterion in parameter selection and model selection. The ESS is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, as seen in (1.8). In general, the greater the ESS the better the estimated model performs. The ESS is the sum of the squares of the differences of the predicted values and the mean value of the response variable and is defined as:

$$ESS = \sum (\hat{y}_i - \bar{y})^2 \tag{1.8}$$

In statistical linear models the SST is the sum of the squares of the difference of the dependent variable and its mean:

$$SST = RSS + ESS = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{\boldsymbol{y}})^2 \tag{1.9}$$

$$= \sum (y_i - \bar{y})^2 \tag{1.10}$$

For wide classes of linear models, the total sum of squares equals the explained sum of squares plus the residual sum of squares. The SST tells us how much variation there is in our response. The variance of the residuals using the regression equation is given by:

$$MSE = \frac{SSE}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1} \tag{1.11}$$

Lastly the OLS has the following properties based on the matrices that we have defined in (1.3):

- The LSE is unbiased $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
- The covariance of $\hat{\boldsymbol{\beta}}$ is given by $= \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$
- The residuals sum up to zero $\sum_{i=1}^{n} e_i = 0$ where $e_i = y_i - \hat{y}_i$
- The residuals and predictors are orthogonal $\sum_{i=1}^{n} x_i e_i = 0$
- The predicted values and residuals are orthogonal $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

We now introduce the five real data sets we will apply to our analysis throughout the paper (which can be found in the archive(WhoisGuard 2017)) and the following articles (Sakaluk 2019),(Tilburg 2019),(Heller 2019):

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Age	Gender	Accuracy	Country
4	4	2	2	3	5	4	4	2	2	2	2	48	2	90	US
5	5	1	1	1	5	4	5	5	5	5	4	63	1	80	US
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	4	1	1	5	5	5	2	1	1	1	3	21	2	90	US
2	2	4	4	4	4	4	4	4	4	4	4	21	2	80	US

Table 1.1: Sample Observations of the Data Set T CFCS

Q1	Q2	Q3	⋯	Q40	Elapse	Gender	Age
2	2	2	⋯	2	211	1	50
2	2	2	⋯	1	149	1	40
⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮
2	2	1	⋯	1	167	1	24
1	2	1	⋯	1	291	1	36

Table 1.2: Sample Observations of the Data Set NPI

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Gender	Age	Source	Country
3	3	1	4	3	4	3	2	3	3	1	40	1	US
4	4	1	3	1	3	3	2	3	2	1	36	1	US
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3	3	2	3	3	3	3	2	3	2	1	31	1	AU
3	4	2	2	3	2	2	4	4	4	2	40	1	US

Table 1.3: Sample Observations of the Data Set T RSE

E1	E2	E3	⋯	S59	S60	EQ	SQ	Accuracy	Gender	Age
3	1	3	⋯	4	3	37	49	80	1	39
4	3	3	⋯	4	2	54	14	98	2	21
⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	3	2	⋯	3	1	44	14	85	2	56
2	1	3	⋯	3	4	22	26	87	1	17

Table 1.4: Sample Observations of the Data Set T EQSQ

Q1	Q2	Q3	⋯	PQ91	PQ92	PQ93
1	1	1	⋯	-1	-1	-1
1	1	1	⋯	-1	-1	-1
⋮	⋮	⋮	⋯	⋮	⋮	⋮
1	1	1	⋯	-1	-1	-1
1	1	1	⋯	-1	-1	-1

Table 1.5: Sample Observations of the Data Set T MSSCQ

Variables	Description	Units
Q1	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q2	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q3	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q4	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q5	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q6	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q7	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q8	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q9	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Q10	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-4
Gender	Gender of Participant 0,1,2,3 (No response, Male, Female and Other.)	0-3
Age	Age of Participant.	Years
Source	1,2,3 (Front page of personality test, google search and other means of search.)	1-3
Country	Country of Participant's Origin.	A-Z,A-Z

Table 1.6: T RSE Data Set Description

Variables	Description	Units
Q1	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q2	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q3	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q4	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q5	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q6	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q7	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q8	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q9	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q10	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q11	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Q12	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-5
Age	Age of Participant over 13 years of age.	Years
Gender	Gender of Participant 0,1,2,3 (No response, Male, Female and Other.)	0-3
Accuracy	Score degree of inclination towards statement in ascending order.	0-100
Country	Country of Participant's Origin.	A-Z,A-Z

Table 1.7: T CFCS Data Set Description

Variables	Description	Units
Q1	Answer Related to Question being asked coded into 1 or 2.	1-2
Q2	Answer Related to Question being asked coded into 1 or 2.	1-2
⋮	⋮	⋮
Q40	Answer Related to Question being asked coded into 1 or 2.	1-2
Elapse	The time elapsed when giving a response (time submitted-time loaded.)	Seconds
Gender	Gender of Participant 0,1,2,3 (No response, Male, Female and Other.)	0-3
Age	Age of Participant over 13 years of age.	Years

Table 1.8: NPI Data Set Description

Variables	Description	Units
E1	Score degree of inclination towards statement in ascending order (0 indicates no response).	0-3
E2	Score degree of inclination towards statement in ascending order(0 indicates no response).	0-3
⋮	⋮	⋮
E60	Score degree of inclination towards statement in ascending order(0 indicates no response).	0-3
S1	Score degree of inclination towards statement in ascending order(0 indicates no response).	0-3
S2	Score degree of inclination towards statement in ascending order(0 indicates no response).	0-3
⋮	⋮	⋮
S60	Score degree of inclination towards statement in ascending order(0 indicates no response).	0-3
EQ	Score degree of inclination towards statement in ascending order(0 indicates no response).	0-3
SQ	Score degree of inclination towards statement in ascending order(0 indicates no response).	0-3
Accuracy	How accurate participants thought their answer were	0-100
Gender	Gender of Participant 0,1,2,3 (No response, Male, Female and Other.)	0-3
Age	Age of Participant over 13 years of age.	Years

Table 1.9: T EQSQ Data Set Description

Variables	Description	Units
Q1	Score degree of inclination towards statement in ascending order	1-4
Q2	Score degree of inclination towards statement in ascending order.	1-4
⋮	⋮	⋮
Q75	Score degree of inclination towards statement in ascending order.	1-4
LAPSE1	Age of Participant over 13 years of age.	Years
LAPSE2	Age of Participant over 13 years of age.	Years
⋮	⋮	⋮
LAPSE74	Age of Participant over 13 years of age.	Years
LAPSE75	Age of Participant over 13 years of age.	Years
introelapse	Age of Participant over 13 years of age.	Years
testelapse	Age of Participant over 13 years of age.	Years
IP_country	Age of Participant over 13 years of age.	Years
engnat	Age of Participant over 13 years of age.	Years
age	Age of Participant over 13 years of age.	Years
education	Age of Participant over 13 years of age.	Years
gender	Age of Participant over 13 years of age.	Years
urban	Age of Participant over 13 years of age.	Years
orientation	Age of Participant over 13 years of age.	Years
race	Age of Participant over 13 years of age.	Years
religion	Age of Participant over 13 years of age.	Years
hand	Age of Participant over 13 years of age.	Years
PQ1	Gender of Participant 0,1,2,3 (No response, Male, Female and Other.)	0-3
PQ2	Gender of Participant 0,1,2,3 (No response, Male, Female and Other.)	0-3
⋮	⋮	⋮
PQ93	Gender of Participant 0,1,2,3 (No response, Male, Female and Other.)	0-3

Table 1.10: T EBFMT Data Set Description

The number of predictors (p+q) varies from one data set to another. Some data sets are

modified irrelevant variables are removed and the data set is made a bit smaller. Afterwards, one of the variables is used as the dependent variable and the rest is used as predictors. The remaining variables in each data sets are the candidate predictors which is $p+q$. We do not know which candidate predictors have a linear relationship with the dependent variable when we work with real data. We do not know which ones are true and which ones are false. The desired sample size is $N = 20$. Each data sets response variable is shown below:

- The response variables for Data Set NPI is: Score.
- The response variables for Data Set T CFCS is: Accuracy.
- The response variables for Data Set T EBFMT is: Score.
- The response variables for Data Set T EQSQ is: Age.
- The response variables for Data Set T RSE is: Age.

1.2 Problem Statement

In Linear Regression as well as Multiple Regression we are sometimes faced with the problem where we have multiple columns that are linearly dependent:

We have the following normal equation that we need to solve which is dependent on the state of $\boldsymbol{X}^\top \boldsymbol{X}$:

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{Y} \tag{1.12}$$

Two scenarios happen when we have linear dependence:

- Scenario 1: $p << n$ where there exists at least two columns; x_j and x_k that are nearly linearly dependent which indicates co-linearity and $\boldsymbol{X}^\top \boldsymbol{X}$ becomes near singular.
- Scenario 2: $n << p$ since there are more predictors than observations so some columns must be a linear combination of another column which indicates linear dependence.

So then the rank of matrix X is less than $p+1$ so the inverse of $\boldsymbol{X}^\top \boldsymbol{X}$ matrix does not exist. But we know that the OLS equation needs $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ which will in turn make the least squares estimate $\hat{\boldsymbol{\beta}}$ impossible to compute. As a result, we explore different methods (Variable Selection, Regularization, Compression and Dimensionality Reduction) that solves the aforementioned issue.

1.3 Chronological Order of Thesis

We introduce the first chapter of this thesis with a review of Multiple Regression and introduce the data sets that we will use throughout this paper. We then present the problem that we are trying to solve or at least work around. In the second chapter, we will explore Variable Selection techniques. We start off with Stepwise Selection Regression which is when we build our regression model from a set of candidate predictor variables by entering and removing predictors in a stepwise manner into our model until there is no justifiable reason to enter or remove any more. Two approaches are studied which are forward selection and backward elimination.

We then move on to Best Subset Selection Regression. The general idea behind the best subsets method is that we select the subset of predictors that do the best at meeting some well-defined objective criterion such as having the largest R^2 value or the smallest MSE. The last section of the chapter shows an empirical exploration of the performance of the RMSE (which we can make inferences on the mean square error) and the signal to noise ratio for seven methods that involve both Stepwise Regression and best subset selection regression. The models used in our regression study need to be assessed on goodness of fit. Some of the procedures explored are the AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion). Another good criterion used as a performance indicator of the methods we will see in this study are the R^2, R^2 adjusted and the Mallow's C_p statistic and its ability to be utilized in a mixed integer programming context to perform subset selection.

The third chapter focuses on the methods of regularization and how it can be used to solve the problem of the ill conditioned matrix we have in our original statement. We start off with general regularization. We define the empirical risk minimization for linear regression form and then define the three general regularization techniques used in the paper which are Tikhonov, Ivanov and Morozov. We present the Tikhonov and Ivanov regularization in both the Ridge regression and Lasso forms. What follows is a study of three theorems that eventually tie up the pieces to show the result that they are all equivalent. Another three theorems are then presented which show that the regularization methods can be used to solve and to design learning algorithms.

A special Tikhonov regularization method is then studied which is Ridge regression (also known as L2- regularization). The estimator's properties such as the geometric interpretation, bias, variance, the mean square error and their derivations are then outlined. Doing this gives us further properties of the trace of the MSE (Mean Square Error). Lastly K-fold cross validation theory is shown by first defining the steps for LOOCV (Leave One Out Cross Validation) which is later extended to K-fold cross validation the derivation of its main properties follows. The second subsection of the chapter goes into the concept of LASSO. We show some properties of the Lasso as well as the sparsity concept that gives this technique an edge over other similar techniques. We try and answer when the solution of the Lasso is unique. We revisit the concept of the Lasso but with an alteration we interpret the Lasso estimate for linear regression parameters as a Bayesian posterior mode estimate when the regression parameters have independent Laplace priors. The Gibbs sampling from this posterior is utilized using an expanded hierarchy with conjugate normal priors for the regression parameters and independent exponential priors on their variances.

A connection with the inverse Gaussian distribution is made to provide full conditional distributions. Moreover, we show the structure of the hierarchical model to provide both Bayesian and likelihood methods for selecting the Lasso parameter. More Bayesian versions of other Lasso related estimation methods including Bridge Regression are then grasped through slight modifications. We then enter the last section of the chapter where we propose the elastic net a unique shrinkage and variable selection method. This is penalized by both the L1-norm (Lasso penalty) and L2-norm (Ridge Regression penalty). The elastic net often outperforms the Lasso, while enjoying a similar sparsity of representation. In addition the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in the model together. We find that the elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the Lasso is not a very satisfactory variable selection method in the $p >> n$ case. Again, a brief empirical exploration of the Ridge, Lasso and LARS methods are shown and analysis of the results is reviewed.

In the fourth chapter PCA (Principle Component Analysis) compression and Dimensionality Reduction Methods are explored within the realm of Eigenvector Decomposition as well as Singular Value Decomposition. Once we have decomposed our matrix using either method, we choose the optimal number of Eigenvectors using the total variance explained and Scree plots. We then propose a novel approach to robust PCA (RPCA) to aid us for data compression when solving the problem of outliers when they arise. In the remaining part of the chapter, we develop the idea of hard thresholding and its role to produce better results via SPCA in interpreting the derived principal components. The approach is a strong competitor to the existing SPCA since the method shows it's superiority over the L_1 penalized method. Hard thresholding is also explored within the realm of non convex RPCA using a novel approach.

For the fifth chapter we present the theory that we will explore empirically. We define the PCR estimator and show the five performance characteristics that we will show empirically instead of theoretically. We then explicitly show the methodology and the outline of our empirical study. We are interested in five characteristic performances and they are: Out of Sample Prediction Error (which will be evaluated on a simulated test data set), Root Mean Square Error of Coefficients Estimates (which is the average of the coefficients), Bias of Coefficients Estimates, Variance of Coefficients Estimates and Percentage of Captured True Variation: R^2 in the regression of the true signal on the estimated signal. We consider the following parameter combinations and each exploration will be a variation of one base specification:

$r = 1, 2, \cdots, 14, 15$,$p = 5$,$\rho_X = 0.5$,$rho\_0.5$,$q = 10$,$S/N = 1$,$\nu = +\inf$. For each of the last five configurations we simulate $1,000$ independent models. For each simulated model we simulate one training sample of size 100 and one test sample of size $10,000$. We analyze the simulated data set and we discuss the results we yield.

In chapter six, we perform an extended simulation and real study to analyze the performance of the methods studied earlier on the basis of the optimal values obtained. We first explain the methodology of the empirical exploration that we will be performing on all 14

methods which are: Best Subset Selection, Forward Stepwise selection, Backward Stepwise Selection, Forward Stepwise Selection governed by AIC, Backward Stepwise Selection governed by AIC, forward Stepwise selection governed by Bayesian Information Criterion, Backward Stepwise selection governed by Cross-Validation, Forward Stepwise selection governed by Cross-Validation (CV), Backward Stepwise selection governed by Cross-Validation, Lasso, Ridge Regression, Elastic Net (a hybrid between Lasso and Ridge Regression), Least Angle Regression (LAR) and Principal Components Regression (PCR).

We use seven distinct characteristic performances that are: Out Of Sample Prediction Error (evaluated on an independently simulated test data set or via Cross- Validation if working with real data), Root Mean Square Error of Coefficients Estimates, Bias of coefficients Estimates, Variance of Coefficients Estimates, Percentage of Selected True Predictors 'X', Percentage of Selected False Predictors 'Z' and Percentage of Captured True Variation. To analyze and compare the aforementioned methods. We then outline how we would carry the empirical study using simulated data and cross validation technique for real data. We first apply the characteristic performances on the fourteen methods for the simulated data set in which we take two cases where $N < p(N = 15)$ and $N > p(N = 100)$ where 1000 models would be used. Lastly we apply the same method to five real data sets that were introduced in the first chapter. We then compare, contrast and explain our findings. We sum up the paper by forming a conclusion and discussing our analysis of the preceding five chapters.

Chapter 2

Variable Selection for Some Spectral Regression Estimators

2.1 Step Wise Selection Regression Techniques Utilizing Forward Selection and Backward Elimination Approaches

Variable Selection Method is when we have at least two variables that are redundant and we are trying to select the one that is significant. We will use the following methods to implement variable selection techniques:

- **Step Wise Selection Regression**.
- **Best Subset Selection**.
- **Backward Elimination Forward Selection**.

Prior to variable selection:

1. We identify outliers and influential points and we may exclude them at least temporarily.
2. We can add any transformation of the variables that is appropriate.

Redundant predictors should be removed. According to Occam's Razor among the plausible explanations for a phenomenon, the simplest is best. For Regression Analysis this implies that the smallest model that fits the data is the best model. Unnecessary predictors will add noise to the estimation of other and Co-linearity is caused by having too many redundant variables. As for the cost; we can save time and or money by not measuring redundant predictors.

First we start with no predictors in the "Step Wise Model." We stop when no more predictors can be entered or removed from our step wise model which would lead us to a final

model.

The Starting Procedure would be:

We set a significance level (denoted by α_E) to decide when to let a predictor into the step-wise model. The default setting would be ($\alpha_E = 0.15$)

We also set a significance level to remove a predictor from the step-wise model. The default setting would be $\alpha_R = 0.15$

Step 1:

1. Fit each predictor individually and regress it on y i.e. regress y on x_1, then regress y on $x_2 \cdots$ and lastly regress y on x_{p-1}.

2. Look for the predictors whose t-test $p-value < \alpha_E = 0.15$ the first predictor put in the step-wise model is the predictor that has the smallest t-test p-value we suppose that this predictor is x_1.

3. If no predictor has a $t-$test $P-$value less than $\alpha_E = 0.15$, stop.

Step 2:

1. Fit each of the two predictor models that include x_1 as a predictor so we regress y on x_1,x_2,$\cdots$ and regress y on x_1 and x_{p-1}.

2. Check for the predictors whose t-test $p-value < \alpha_E = 0.15$ The smallest p-value would be the second predictor put in this step-wise model.

3. If no predictor has a t-test $p-value < \alpha_E = 0.15$. The model with the one predictor obtained from the first step is your final model. We will suppose x_2 was the second best predictor.

4. Now we step back and see if entering x_2 affects x_1 into the step-wise model somehow affected the significance of the x_1 predictor.

We test $\beta_1 = 0$ has become significant that is $p-value > \alpha_R = 0.15$.

Step 3:

1. Suppose both x_1 and x_2 made it into the two-predictor step-wise model and remained there.

2. Now fit each of the three predictor models that include x_1 and x_2 as predictors- that is regress y on x_1,x_2 and x_3 regress y on x_1, x_2, and x_4,$\cdots$, and regress y on x_1, x_2, and x_{p-1}.

3. Of those predictors whose $t-$ test $P-$ value is less than $\alpha_E = 0.15$, the third predictor put in the step-wise model is the predictor that has the smallest $t-$ test $P-$value.

4. If no predictor has a $t-$ test $P-$ value less than $\alpha_E = 0.15$, stop. The model containing the two predictors obtained from the second step is your final model.

5. Suppose instead that x_3 was deemed the "best"third predictor and it is therefore entered into the step-wise model.

6. Now since x_1 and x_2 were the first predictors in the model, step back and see if entering x_3 into the step-wise model somehow affects the significance of the x_1 and x_2 predictors. Check the $t-$ test $P-$ values for testing $\beta_1 = 0$ and $\beta_2 = 0$. If the $t-$ test $P-$ value for either $\beta_1 = 0$ or $\beta_2 = 0$ has become not significant that is the $P-$ value is greater than $\alpha_R = 0.15$ remove the predictor from the step-wise model.

Stopping Procedure:

We continue the steps above until adding any additional predictor does not yield a t-test $p - value < \alpha_E = 0.15$.

Given a matrix $\boldsymbol{X}$, its QR-decomposition is a matrix decomposition of the form

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} \tag{2.1}$$

Where $\boldsymbol{R}$ is an upper triangular matrix and $\boldsymbol{Q}$ is an orthogonal matrix which satisfies the following property:

$$\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I} \tag{2.2}$$

Where $\boldsymbol{Q}^\top$ is the transpose of $\boldsymbol{Q}$ and $\boldsymbol{I}$ is the identity matrix. This matrix decomposition can be used to solve linear systems of equations.

Stepwise selection consists of two methods: Forward Step-wise selection and the Backward Elimination Method. Forward step-wise selection is a greedy algorithm producing a nested sequence of models. In this sense it might seem sub optimal compared to the best subset selection. However there are several reasons why it might be preferred. Forward step-wise produces a sequence of models indexed by k the subset size which must be determined.

Forward- step-wise selection starts with the intercept and then sequentially adds the predictor that most improves the fit. With many candidate predictors, this might seem like a lot of computation; however clever updating algorithms can exploit the QR decomposition for the current fit to rapidly establish the next candidate. There are two disadvantages that come with forward step-wise selection:

- Computational: for large p we cannot compute the best subset sequence but we can always compute the forward step-wise sequence (even when p»n).

- Statistical: a price is paid in variance for selecting the best subset of each size. Forward step-wise is a more constrained search and will have lower variance but perhaps more bias.

Backward step-wise selection starts with the full model and sequentially deletes the predictor that has the least impact on the fit. The candidate for dropping is the variable with the smallest $Z-$ score. Backward selection can only be used when $n > p$ while Forward step-wise can always be used.

Lastly it is worth noting that variables that come in groups (such as the dummy variables that get coded as multi-level categorical predictors). Smart step-wise procedures (such as step in R) will add or drop whole groups at a time, taking proper account of their degrees of freedom.

2.2 Best Subset Selection Regression Method

A fundamental rule of the best subsets regression procedure is that the predictor variable must include all of the variables that predict the response. Otherwise, we end up with a regression model that is under specified and hence misleading.

The Procedure:

Step 1: We identify all combinations of all regression models (2^p). This can be a huge number of possible models.

For e.g.: If we have three predictors that would give us $2^3 = 8$ possibilities of models.

- One model with no predictors (1) ().
- Three models each with one predictor (3) $(x_1)(x_2)(x_3)$.
- Three models each with two predictors (3) $(x_1, x_2)(x_1, x_3)(x_2, x_3)$.
- One model with all three predictors (1) (x_1, x_2, x_3).

Step 2:

Once the models from the first step are identified, We form a well defined criteria. then the one-predictor model that would do best in meeting that criteria. We apply the same thing for the two predictor models we look for the best in meeting the criteria and so on.

If we take our above example (the three predictors x_1, x_2, x_3) when choosing from the three models where each model has one predictor, we look for the one that does best. We do the same thing with the three models that have two predictors and we choose one or two that meet the best criteria. The word "best" is subjective and the following conditions help in choosing the right model:

- The model with the largest adjusted R^2, an even better measure would be the predicted R^2.
- The model with the smallest MSE $\left(\text{or } S = \sqrt{MSE}\right)$.

Step 3:

We refine the models identified in step two. To do this, we can perform several methods such as residual analyses transformations of the predictor and or response, adding interaction terms and so on. We do this until we find a model that answers our research question and does a good job of summarizing the trend in data.

2.3 Criterion Based Procedures (The AIC and BIC Criterion)

To compare regression models, some statistical software may also give values of statistics referred to as information criterion statistics. For regression models, these statistics combine information about the SSE number of parameters in the model and the sample size. A low value compared to values for other possible models is good.

If we have p potential predictors then we will get 2^p models. We fit all of these models and pick the best one according to the following criteria: **The Akaike Information Criterion (AIC):**

$$n\ ln(SSE) - n\ ln(n) + 2p \tag{2.3}$$

The Bayes Information Criterion (BIC):

$$n\ ln(SSE) - n\ ln(n) + p\ ln(n) \tag{2.4}$$

The best model would be the one with the smallest AIC and or BIC criterion. The BIC penalizes the larger models more than the AIC criterion and so will prefer the smaller models. The larger models will use more parameters but in contrast it will fit better and yields a smaller RSS, thus the best choice of model will balance fit with model size.

Some statisticians believe that these information criteria give a more realistic comparison of models than the Mallow C_p statistic because C_p tends to make models seem more different than they actually are.

We notice that the only difference between AIC and BIC is the multiplier of p the number of parameters. When comparing the two models, the model with the lower value is always preferred. Moreover, the BIC places a higher penalty on the number of parameters in the model so will tend to favor more parsimonious models. This culminates in the criticism of AIC in that it tends to over-fit models.("Information Criteria and PRESS" 2019)

2.4 Criterion Based Procedures (R^2, R^2 adjusted, Mallows C_P and Subset Selection Via Mixed Integer Programming Approach)

The R^2- value are defined as:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{2.5}$$

This value can only increase as more variables are added. It makes no sense to define the "best" model as the model with the largest R^2- value. If we did follow this process, then the model with the largest number of predictors would always win. We can however instead use the R^2 values to find the point where adding more predictors is not worthwhile; this is because it will yield a very small increase in the R^2 value. In other words, we look at the size of the increase in R^2 not just its magnitude alone. Since this is not a very reliable criterion, it is used more often in combination with other criteria.

The adjusted R^2 value which is defined as:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SSE}{SSTO}\right) = 1 - \left(\frac{n-1}{SSTO}\right)MSE = \frac{\frac{SSTO}{n-1} - \frac{SSE}{n-p}}{\frac{SSTO}{n-1}} \tag{2.6}$$

Makes us pay a penalty for adding more predictors to the model. So now we can just use the R^2 adjusted value. According to the R^2 adjusted value criterion, the best regression model is the one with the largest adjusted R^2 value. As we can see in the above equation the R^2 adjusted value is a function of the mean square error (MSE) and according to the MSE criterion, the best regression model is the one with the smallest MSE. The two criteria are equivalent in a sense that if we look at the formula again for the adjusted R^2 value, we can see that the adjusted R^2 value increases only if MSE decreases. That is the R^2 adjusted value and MSE criteria always yield the same best models.

When we have an underspecified model in which important predictors are missing, we yield biased regression coefficients and biased predictions of the response. This is where the Mallow's Cp-statistic is useful where it estimates the size of the bias that is introduced into the predicted responses by having an underspecified model.

Any regression model will be faced with the following two issues:

- The bias in the predicted responses.
- The variation in the predicted responses.

If there is no bias in the predicted responses then the average of the observed responses $\mathbb{E}(y_i)$ and the average of the predicted responses $\mathbb{E}(\hat{y}_i)$ both equal the average of the responses in the population $\mu_{Y|x}$. Conversely, if there is bias in the predicted responses, then $\mathbb{E}(y_i) = \mu_{Y|x}$ and $\mathbb{E}(\hat{y}_i)$ do not equal each other. The difference between $\mathbb{E}(y_i) = \mu_{Y|_x}$ and $\mathbb{E}(\hat{y}_i)$ is the bias B_i in the predicted response. Which is defined as follows:

$$\text{Bias: } \mathbb{E}(\hat{y}_i) - \mathbb{E}(y_i) \tag{2.7}$$

When bias exists in the predicted responses the variance in the predicted responses for a

data point i is due to the random sampling variation ($\sigma^2_{\hat{y}_i}$). If our regression model is biased, it will not make sense to consider the bias at just one data point i. So we will need to consider the bias that exists for all data points n. The same thing applies for the variation in the predicted responses; we cannot just consider the variation in the predicted responses at one data point i. We need to consider the total variation in the predicted responses. To do this and to quantify the total variation in the predicted responses we consider the following standardized measure of the total variation in the predicted responses Γ_p("Information Criteria and PRESS" 2019):

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}\sigma^2_{\hat{y}_i} + \sum_{i=1}^{n}\left[\mathbb{E}(\hat{y}_i) - \mathbb{E}(y_i)\right]^2\right\} \tag{2.8}$$

Where $\sum_{i=1}^{n}\sigma^2_{\hat{y}_i}$ quantifies the random sampling variation summed over all n data points. And $\sum_{i=1}^{n}\left[\mathbb{E}(\hat{y}_i) - \mathbb{E}(y_i)\right]^2$ quantifies the amount of bias squared summed over all n data points. Because the size of the bias depends on the measurement units used we need to get a standardized unit-less measure which can be found by dividing by σ^2.

It can be shown that if there is no bias in the predicted responses i.e. if the bias $= 0$ then Γ_p achieves its smallest possible value of p parameters:

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}\sigma^2_{\hat{y}_i} + 0\right\} = p \tag{2.9}$$

Γ_p quantifies the amount of bias and variance in the predicted responses so it seems to be a good measure of an under specified model:

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}\sigma^2_{\hat{y}_i} + \sum_{i=1}^{n}\left[\mathbb{E}(\hat{y}_i) - \mathbb{E}(y_i)\right]^2\right\} \tag{2.10}$$

The best model is simply the model with the smallest value of Γ_p. We even know that the theoretical minimum of Γ_p is the number of parameters p.("Best Subsets Regression, Adjusted R-Sq, Mallows Cp" 2007) Since we cannot know the value of Γ_p, we must estimate and this is where we use the Mallow's C_p statistics. If the population variance σ^2 is known, then we can estimate Γ_p using the following equation:

$$C_p = p + \frac{\left(MSE_p - \sigma^2\right)(n-p)}{\sigma^2} \tag{2.11}$$

Where MSE_p is the mean squared error from fitting the model which contains the subset of $p-1$ predictors (if we include the intercept we get p predictors). Since we don't know

σ^2, we need to estimate using MSE_{all} which is the mean squared error obtained from fitting the model containing all of the candidate predictors. This yields the following equation:

$$C_p = p + \frac{(MSE_p - MSE_{all})(n-p)}{MSE_{all}} = \frac{SSE_p}{MSE_{all}} - (n-2p) \tag{2.12}$$

We must note two things when we estimate σ^2 using MSE_{all} :

- We assume that there are no biases in the full model with all of the predictors.
- Since $MSE_p - MSE_{all} = 0$ guarantees that $C_p = p$ for the full model.

The following are a few facts about Mallow's C_p- statistic that are essential in determining the "best" model.

Subset models with small C_p values have a small total (standardized) variance of prediction.

- When the C_p value is near p the bias is small near 0.
- When the C_p value is much greater than p the bias is substantial.
- When the C_p value is below p it is due to sampling error; interpret as no bias.

Furthermore, the largest model containing all of the candidate predictors $C_p = p$ always holds. Therefore we should not use C_p as a measure to evaluate the fullest model.

The following strategy is useful for using C_p to identify the "best" model:

If possible, we identify subsets of predictors for which the C_p value is near p. The full model always yields $C_p = p$ so we don't select the full model based on C_p. Following that, if all the models except for the full model yield a large C_p not near p, this suggests some important predictor(s) are missing from the analysis. In this case, we are well advised to identify the predictors that are missing.

We then check a number of models that have C_p near p to ensure that the combination of the bias and the variance is at a minimum. We choose the model with the smallest C_p. Lastly, when more than one model has a small value of C_p value near p, in general we choose the simpler model or the model that meets our research goal.

The Mallow C_p Statistic can also be written in the following form:

$$C_P^{\text{Full}} = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\hat{\sigma}^2} + 2(p+1) - n \tag{2.13}$$

Where $\hat{\sigma}^2$ is an estimator of the residual variance, σ^2. This estimator is usually set to the unbiased estimator of the full regression model in (1.1). Where $\hat{\sigma}^2$ is defined as follows:

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-p-1} \tag{2.14}$$

$2(p+1)$ represents the model complexity to be decreased. Occam's Razor principle helps to avoid over fitting and computational error which in turn improves the generalization capability of the predictive model. (2.13) can be converted to the following equation:

$$C_P^{\text{Full}} = \min_{\boldsymbol{\beta}} \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{\hat{\sigma}^2} + 2(p+1) - n \tag{2.15}$$

This is because the OLS (Ordinary Lease Squares) estimator $\hat{\beta}$ minimizes $S(\beta)$. We can go further with the C_p expression when we do best subset regression; we can use (2.15) to eliminate the explanatory variable x_j and show it is equivalent to fixing its coefficient β_j to zero.

So C_p for the subset model is:

$$C_p(S) = \min_{\beta} \left\{ \frac{(\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta)}{\hat{\sigma}^2} \beta_j = 0 \ (j \notin S) \right\} + 2(|S|+1) - n \tag{2.16}$$

Where $|S|$ is the number of elements of the set S or the number of selected variables. Substituting (2.14) into C_p^{Full}, we see that $C_p^{Full} = p+1$. Hence, if $C_p(S)$ is minimized with respect to $S \subset \{1, 2, \cdots, p\}$, it will not be more than $p+1$. If we consider the case in which the number $k = |S|$ is given, k explanatory variables are to be selected from p candidate ones. In this case by omitting constant terms, the minimization of $C_p(S)$ reduces to the following RSS minimization problem:

$$\min_{\boldsymbol{\beta}, S} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) | \beta_j = 0 \ (j \notin S), |S| = k, S \subseteq \{1, 2, \cdots, p\} \right\} \tag{2.17}$$

This subset selection problem can be shown as a mixed integer quadratic programming (MIQP) problem. (Arthanari and Dodge 1981), (Bertsimas and Shioda 2009) and (Konno and Yamamoto 2009) If we introduce $0-1$ decision variables z_j for $j = 1, 2, \cdots, p$ to determine whether the $j-$th candidate explanatory variable is selected or not selected. $z_j = 1$ if the $j-$th variable is selected $z_j = 0$ otherwise. By using the big-M formulation the subset selection problem with the fixed k can be expressed as an MIQP problem.

$$\underset{\boldsymbol{\beta},\boldsymbol{z}}{\text{minimize}} \quad \sum_{i=1}^{n}\left(y_i-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \tag{2.18}$$

$$\text{subject to } -Mz_j \le \beta_j \le Mz_j (j=1,2,\cdots,p), \tag{2.19}$$

$$\sum_{j=1}^{p} z_j = k \tag{2.20}$$

$$z_j \in \{0,1\} (j=1,2,\cdots p) \tag{2.21}$$

Where M is a sufficiently large positive constant. Constraint (2.19) is called a big-M constraint. If $z_j = 0$, the $j-$th candidate explanatory variable is eliminated from the regression model, because its coefficient a_j has to be 0 from constraint (2.19). If the interval $[-M, M]$ is sufficiently large, $z_j = 1$ implies that a_j can take an arbitrary value. (2.20) forces the number of selected explanatory variables to be k. Consequently problems $(2.18) - (2.21)$ is equivalent to problem (2.17).

Problems $(2.18) - (2.21)$ enables one to find k explanatory variables that minimize RSS. However we might want to determine a certain number $k = |S|$ simultaneously. To accomplish this, we shall use Mallow's C_p as a GOF measure. If we consider the representation (2.16) of $C_p(S)$ the subset selection problem of minimizing Mallows' C_p can be formulated as an MIQP problem:

$$\underset{\boldsymbol{\beta},\boldsymbol{z}}{\text{minimize}} \quad \frac{\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2}{\hat{\sigma}^2} + 2\left(\sum_{j=1}^{p} z_j + 1\right) - n \tag{2.22}$$

$$\text{subject to} \quad -Mx_j \le \beta_j \le Mz_j \quad (j=1,2,\cdots,p), \tag{2.23}$$

$$z_j \in \{0,1\} \quad (j=1,2,\cdots,p). \tag{2.24}$$

We notice the number of selected explanatory variables $\sum_{j=1}^{p} z_j$ is not pre-specified whereas it is a given constant k in problems$(2.18) - (2.21)$. We can select the best subset of explanatory variables according to C_p by solving problem $(2.22) - (2.24)$. In problem $(2.22) - (2.24)$, the positive constant M needs to be sufficiently large. If M is not sufficiently large, $(2.22) - (2.24)$ cannot guarantee the optimality of the selected explanatory variables. On the other hand, it is known that a large M can cause numerical instabilities in computations. Mixed logical programming is a remedy for this problem and it is supported by several mathematical programming solvers. We can replace the big M constraint (2.23) with its logical implication:

$$\underset{\boldsymbol{\beta},\boldsymbol{z}}{\text{minimize}} \quad \frac{\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2}{\hat{\sigma}^2}+2\left(\sum_{j=1}^{p}z_j+1\right)-n \tag{2.25}$$

$$\text{subject to} \quad z_j=0 \quad \Rightarrow \quad \beta_j=0 \quad (j=1,2,\cdots,p), \tag{2.26}$$

$$z_j\in\{0,1\} \quad (j=1,2,\cdots,p). \tag{2.27}$$

The logical implications (2.26) mean that if $z_j=0$ the $j-$th candidate explanatory variable is eliminated from the regression model. This sort of logical implication can be efficiently handled in a branch and bound procedure for MIP problems.

Chapter 3

Regularization for Some Spectral Regression Estimators

3.1 Ridge Regression Regularization Properties and Analysis Using K-Fold Cross Validation

Ridge Regression or also known as Tikhonov regularization helps solve the problem imposed in chapter one. Similar to the least squared estimator $\hat{\boldsymbol{\beta}}$, the ridge regression coefficients are estimated by minimizing:

$$S(\boldsymbol{\beta}, \lambda) = S(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.1}$$

over $\boldsymbol{\beta}$ for a given λ, where λ is called the tuning parameter and $\lambda \geq 0$. And $S(\boldsymbol{\beta})$ is defined in (1.5). The Ridge regression estimator is denoted as $\hat{\boldsymbol{\beta}}^R(\lambda)$. The goal is to shrink $S(\boldsymbol{\beta})$ as much as possible to be able to find estimates that fit the data reasonably well. When this happens $(\hat{\beta}_1, \cdots, \hat{\beta}_p)$ are also shrunken in the process as they are dependent on the minimization of the residual sum of squares and so in turn this affects the term $\lambda \sum_{j=1}^{p} \beta_j^2$ which becomes smaller. When we constrain the coefficient estimates we reduce the ridge estimator's variance and introduce some bias.

Various regularization techniques can be applied to these cases such as ridge regression. We can increase the stability of the solution if further information about the parameters is known; for example a range of possible values of the $\hat{\boldsymbol{\beta}}$ then various techniques can be used to increase the stability of the solution. Such a system usually has no solution, so the goal is instead to find the coefficients $\boldsymbol{\beta}$ which fit the equations "best" in the sense of solving the quadratic minimization problem.

We can show various properties of the ridge regression estimator. Starting with the geometric representation of the ridge regression estimator which we show below in Figure 3.2:

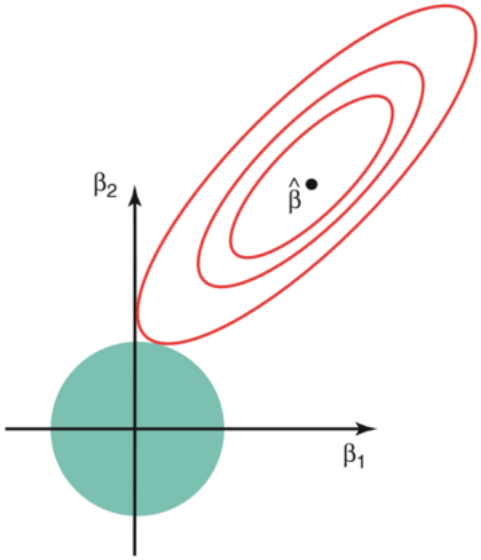

FIGURE 3.1: Geometric Interpretation of the Ridge Regression Estimator in 2-Dimensional Space.

The ellipses in Figure 3.2 correspond to the contours of residual sum of squares (RSS). The inner ellipse has smaller RSS and RSS is minimized at ordinal least square (OLS) estimates. For $p = 2$, the constraint in ridge regression corresponds to a circle. $\sum_{j=1}^{p} \beta_j^2 < c$.

We are trying to minimize the ellipse size and circle simultaneously in the ridge regression. The ridge estimate is given by the point at which the ellipse and the circle touch. There is a trade-off between the penalty term and RSS. Perhaps a large $\boldsymbol{\beta}$ would give you a better residual sum of squares but then it will push the penalty term higher. This is why you might actually prefer smaller $\boldsymbol{\beta}$'s with worse residual sum of squares. From an optimization perspective, the penalty term is equivalent to a constraint on the $\boldsymbol{\beta}$'s. The function is still the residual sum of squares but now we constrain the norm of the β_j 's to be smaller than some constant c. There is a correspondence between λ and c. The larger the λ is, the more we prefer the β_j's close to zero. In the extreme case when $\lambda = 0$, then we would simply be doing a normal linear regression. On the other hand as λ approaches infinity, we set all the $\boldsymbol{\beta}$'s to zero.

If we look at the objective function of the ridge regression estimator, the shrinkage penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ is only applied to the estimates $\beta_1, \beta_2, \cdots, \beta_p$ and not to β_0 the intercept. The intercept term is a measure of the mean value of the response if $x_{i1} = x_{i2} = \cdots = x_{ip} = 0$. To omit the intercept β_0, we standardize the predictors which means that each column is centered so that $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$, for $j = 1, 2, \cdots, p$. We also center the outcome values such that $\frac{1}{n}\sum_{i=1}^{n} y_i = 0$. Once the ridge regression coefficient estimates are obtained, we can recover the intercept term by:

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \bar{x}_j \hat{\beta}_j^R(\lambda) \tag{3.2}$$

where $\bar{y}$ and $\bar{x}_j$ for $j = 1, 2, \cdots, p$ are the original means. Another way we could write the ridge regression problem in the Lagrangian form:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} (\boldsymbol{\beta}, \lambda) = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \text{ for some } \lambda \geq 0 \tag{3.3}$$

Hoerl and Kennard (Hoerl and Kennard 1970) proposed that potential instability in the LS estimator (**??**) could be improved by adding a small constant value λ (The Tuning Parameter) to the diagonal entries of the matrix $X^\top X$ before taking its inverse. This yields the following ridge regression estimator:

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \tag{3.4}$$

The following constraint is placed on the parameters $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ to minimize the penalized sum of squares:

$$\begin{aligned} &\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta^2 \\ \hat{\boldsymbol{\beta}}_{ridge} = S(\boldsymbol{\beta}, \lambda) &= RSS(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \end{aligned} \tag{3.5}$$

over $\boldsymbol{\beta}$ for a given λ, where λ is called a tuning parameter and $\lambda \geq 0$. We have the following properties for the ridge regression estimator:("Ridge Regression" 2018)

- $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{\boldsymbol{R}}(\lambda)$ is a biased estimator of $\boldsymbol{\beta}$.
- For orthogonal covariates $\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}_p, \hat{\boldsymbol{\beta}}^{\boldsymbol{R}}(\lambda) = \frac{n}{n+\lambda}\hat{\boldsymbol{\beta}}$. So in this case, we have the ridge estimator always producing shrinkage towards 0. λ controls the amount of shrinkage.

  A concept that is important in shrinkage is the effective degrees of freedom associated with a set of parameters. In a ridge regression setting:

- If we choose $\lambda = 0$ we have p parameters since there is no penalization.
- If λ is large, the parameters are heavily constrained and the degrees of freedom will be lower tending to 0 as $\lambda \longrightarrow \infty$.

The effective degrees of freedom associated with $\beta_1, \beta_2, \cdots, \beta_p$ is defined as:

$$df(\lambda) = tr\left(\boldsymbol{X}\left(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_p\right)\boldsymbol{X}^\top\right) = \sum_{j=1}^{p} \frac{\boldsymbol{d}_j^2}{\boldsymbol{d}_j^2 + \lambda} \tag{3.6}$$

where d_j are the singular values of $\boldsymbol{X}$. We notice that $\lambda = 0$ and when $\boldsymbol{X}^\top \boldsymbol{X}$ is non-singular,

the expression corresponds to no shrinkage which gives us $df(\lambda) = p$. There is also 1:1 mapping between λ and the degrees of freedom so we can simply pick the effective degrees of freedom that we like associated with the fit, and solve for λ.

Furthermore, Cross validation can be used in choosing λ as an alternative to manually choosing a λ. In the cross validation process λ is selected based on the smallest cross validation error. Also since $\boldsymbol{Y}$ has been centered, the intercept β_0 has been left out. If the intercept is penalized, then that would make the procedure depend on the origin chosen for $\boldsymbol{Y}$. Calculating the variance covariance matrix is a simple process since the ridge regression estimator is linear.

We now introduce the Bias, variance and Mean Squared Error properties of the estimator:

$$\begin{aligned} \text{Bias: } bias\ [\hat{\boldsymbol{\beta}}^{\boldsymbol{R}}(\lambda)] &= \mathbb{E}\left[\hat{\boldsymbol{\beta}}^{\boldsymbol{R}}(\lambda) - \boldsymbol{\beta}\right] \\ &= -\lambda W^{-1}(\lambda)\boldsymbol{\beta} \qquad (3.7) \\ \text{Variance: } \mathbb{V}\left[\hat{\boldsymbol{\beta}}^{R}(\lambda)\right] &= \sigma^2 W^{-1}(\lambda)\boldsymbol{X}^\top \boldsymbol{X} W^{-1}(\lambda) \qquad (3.8) \end{aligned}$$

$$\text{Where} \quad W^{-1}(\lambda) = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda I_p\right)^{-1}$$

We notice that:

$$\lim_{\lambda\to\infty} \mathbb{V}[\hat{\boldsymbol{\beta}}^{R}(\lambda)] = \lim_{\lambda\to\infty} \sigma^2 W^{-1}(\lambda)\boldsymbol{X}^\top \boldsymbol{X} W^{-1}(\lambda) = \mathbf{0}_{p\times p}$$

The proof of the bias of the ridge regression (3.7) estimator is now outlined: (Wieringen 2015) and (AlNasser 2017).

Proof.

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda)\right] &= \mathbb{E}\left[(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{pp})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{pp}\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{pp}\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\hat{\boldsymbol{\beta}}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\right]\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] \\
&= \left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}
\end{aligned}$$

So now:

$$\begin{aligned}
\text{bias}\left(\hat{\boldsymbol{\beta}}_{\text{Ridge}}\right) &= \mathbb{E}\left[\hat{\boldsymbol{\beta}}_{\text{Ridge}}(\lambda)-\boldsymbol{\beta}\right] \\
&= \left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\beta} \\
&= \left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{pp}\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{p}-\lambda I_p\right)\boldsymbol{\beta}-\boldsymbol{\beta} \\
&= \left(\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{p}\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{p}\right)-\lambda\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{p}\right)^{-1}\right)\boldsymbol{\beta}-\boldsymbol{\beta} \\
&= -\lambda\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{p}\right)^{-1}\boldsymbol{\beta}
\end{aligned}$$

□

The remaining properties and its proof derivations of the ridge regression estimator are now shown (AlNasser 2017) starting with the variance (3.8):

Proof.

$$\begin{aligned}
\mathbb{V}\left[\hat{\boldsymbol{\beta}}(\lambda)\right] &= \mathbb{V}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}_{p}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\right] \\
&= \mathbb{V}\left[W^{-1}(\lambda)\boldsymbol{X}^\top\boldsymbol{Y}\right] \\
&= W^{-1}(\lambda)\boldsymbol{X}^\top
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}\left[\boldsymbol{Y}\right]\left(W^{-1}(\lambda)\boldsymbol{X}^\top\right)^\top &= W^{-1}(\lambda)\boldsymbol{X}^\top\mathbb{V}\left[\boldsymbol{\epsilon}\right]\boldsymbol{X}W^{-1}(\lambda) \\
&= \sigma^2 W^{-1}(\lambda)\boldsymbol{X}^\top\boldsymbol{X}W^{-1}(\lambda)
\end{aligned}$$

□

In which the following property was used:

$$\mathbb{V}\left(c\boldsymbol{Y}\right) = c\mathbb{V}(\boldsymbol{Y})c^\top$$

Where c is a constant and $\boldsymbol{Y}$ is the variable that we are interested in.

The Mean Squared Error is defined as follows:

$$\begin{aligned} MSE\left[\hat{\boldsymbol{\beta}}(\lambda)\right] &= \mathbb{V}\left[\hat{\boldsymbol{\beta}}(\lambda)\right] + bias\left[\hat{\boldsymbol{\beta}}(\lambda)\right]^2 \\ &= \mathbb{V}\left[\hat{\boldsymbol{\beta}}(\lambda)\right] + bias\left[\hat{\boldsymbol{\beta}}(\lambda)\right] bias\left[\hat{\boldsymbol{\beta}}(\lambda)\right]^\top \text{ (By using the property of a vector } A^\top A = A^2) \\ &= W^{-1}(\lambda)\boldsymbol{X}^\top \boldsymbol{X} W^{-1}(\lambda) + \lambda^2 W^{-1}(\lambda)\boldsymbol{\beta}\boldsymbol{\beta}^{-1} W^{-1}(\lambda) \end{aligned} \tag{3.9}$$

The proof of (3.9) is outlined as follows:

Proof.

$$\begin{aligned} MSE\left(\hat{\boldsymbol{\beta}}^R(\lambda)\right) &= \mathbb{V}\left[\hat{\boldsymbol{\beta}}(\lambda)\right] + bias\left[\hat{\boldsymbol{\beta}}(\lambda)\right] + bias^\top\left[\hat{\boldsymbol{\beta}}(\lambda)\right] \\ &= \sigma^2 W^{-1}(\lambda)\boldsymbol{X}^\top \boldsymbol{X} W^{-1}(\lambda)\left(-\lambda W^{-1}(\lambda)\right)\left(-\lambda W^{-1}(\lambda)\boldsymbol{\beta}\right)^\top \\ &= \sigma^2 W^{-1}(\lambda)\boldsymbol{X}^\top \boldsymbol{X} W^{-1}(\lambda) + \lambda^2 W^{-1}(\lambda)\boldsymbol{\beta}\boldsymbol{\beta}^\top W^{-1}(\lambda) \end{aligned}$$

□

The following properties of the trace of the Mean Square Error of the Ridge regression estimator are:

$$\frac{d_{mv}(\lambda)}{d(\lambda)} < 0 \text{ for all } \lambda > 0 \tag{3.10}$$

$$\frac{d^2_{mv}(\lambda)}{d\lambda^2} > 0 \text{ for all } \lambda > 0 \tag{3.11}$$

$$\frac{d_{mb}(\lambda)}{d(\lambda)} \geq 0 \text{ for all } \lambda > 0 \tag{3.12}$$

The proofs of $(3.10) - (3.12)$ respectively are now shown:

Proof. First Let $u_1 \geq u_2 \geq \cdots \geq u_p$ be the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$. Since $\boldsymbol{X}^\top \boldsymbol{X}$ is positive semi definite we have $u_i \geq 0$ for all $i = 1, 2, \cdots, p$. From $W(\lambda) = \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p$, $W^{-j}(\lambda)$ has eigenvalues $(u_1+\lambda)^{-j}, (u_2+\lambda)^{-j}, \cdots, (u_p+\lambda)^{-j}$, for $j = 1, 2$.

$$\begin{aligned} mv(\lambda) &= \sigma^2\left[tr\left(W^{-1}(\lambda)\right) - \lambda tr\left(W^{-2}(\lambda)\right)\right] \\ &= \sigma^2\left[\sum_{i=1}^{p} \frac{1}{u_i+\lambda} - \lambda \sum_{i=1}^{p} \frac{1}{(u_i+\lambda)^2}\right] \\ &= \sigma^2 \sum \frac{u_i}{(u_i+\lambda)^2} \end{aligned}$$

$$\text{We get: } \frac{dm_v(\lambda)}{d\lambda} = \sigma^2 \sum_{i=1}^{p} \frac{-2u_i}{(u_i+\lambda)^3} < 0, \text{ for all } \lambda > 0$$

□

Proof.

$$\begin{aligned}\frac{d^2 m_v(\lambda)}{d\lambda^2} &= (-2u_i(u_i+\lambda)^{-3})' \text{ for all } \lambda > 0 \\ &= 6u_i(u_i+\lambda)^{-4} \\ &= \frac{6u_i}{u_i+\lambda^4} > 0 \text{ for all } \lambda > 0\end{aligned}$$ □

To prove (3.12) we will use the following fact:

$W(\lambda)W^{-1}(\lambda) = \boldsymbol{I}$ to obtain the expression $\frac{dW^{-1}(\lambda)}{d\lambda}$.

Proof.

$$\begin{aligned}\frac{dW(\lambda)}{d\lambda}W^{-1}(\lambda) + \frac{W(\lambda)dW^{-1}(\lambda)}{d\lambda}\frac{d\boldsymbol{I}}{d\lambda} &= \boldsymbol{0} \\ \boldsymbol{I}_p W^{-1}(\lambda) + \frac{W(\lambda)dW^{-1}(\lambda)}{d\lambda} &= \boldsymbol{0} \\ \frac{dW^{-1}(\lambda)}{d\lambda} &= -W^{-1}(\lambda)W^{-1}(\lambda) = -W^{-2}(\lambda) \\ m_b(\lambda) = \lambda^2\boldsymbol{\beta}^\top = W^{-2}(\lambda)\boldsymbol{\beta} &= \frac{dm_b(\lambda)}{d\lambda} \\ &= 2\lambda\boldsymbol{\beta}^\top W^{-2}(\lambda)\boldsymbol{\beta} + \lambda^2\frac{d}{d\lambda}(\boldsymbol{\beta}^\top W^{-1}(\lambda)W^{-1}(\lambda)\boldsymbol{\beta}) \\ &= 2\lambda\boldsymbol{\beta}^\top W^{-2}(\lambda)\boldsymbol{\beta} + \lambda^2\boldsymbol{\beta}^\top\frac{dW^{-1}(\lambda)}{d\lambda}W^{-1}(\lambda)\boldsymbol{\beta} + \lambda^2\boldsymbol{\beta}^\top W^{-1}(\lambda)\frac{dW^{-1}(\lambda)}{d\lambda}\boldsymbol{\beta} \\ &= 2\lambda\boldsymbol{\beta}^\top W^{-2}(\lambda)\boldsymbol{\beta} - 2\lambda^2\boldsymbol{\beta}^\top W^{-3}(\lambda)\boldsymbol{\beta} \\ &= 2\lambda\boldsymbol{\beta}^\top\left[W^{-2}(\lambda) - \lambda W^{-3}(\lambda)\right]\boldsymbol{\beta} \\ &= 2\lambda\boldsymbol{\beta}^\top T(\lambda)\boldsymbol{\beta},\end{aligned}$$

With $T(\lambda) = W^2(\lambda) - \lambda$ and $W^{-3}(\lambda)$.

And (λ) has eigenvalues:

$$\frac{1}{(u_i+\lambda)^2} - \frac{\lambda}{(u_i+\lambda)^3} = \frac{u_i}{(u_i+\lambda)^3} \geq 0,\ i = 1, 2, \cdots, p \text{ for } \lambda > 0$$

$$\therefore \frac{d_{mb}(\lambda)}{d\lambda} \geq 0 \text{ for all } \lambda > 0 \text{ and } \boldsymbol{\beta}.$$ □

Further properties of the Trace Of The Mean Square Error are now introduced:

$$m_v(\lambda) \longrightarrow 0 \text{ as } \lambda \longrightarrow \infty \tag{3.13}$$

$$m_b(\lambda) = \lambda^2\boldsymbol{\beta}^\top W^{-2}(\lambda)\boldsymbol{\beta} \longrightarrow \boldsymbol{\beta}^\top\boldsymbol{\beta} \text{ as } \lambda \longrightarrow \infty \tag{3.14}$$

$$m_b(\lambda) \text{ is a concave function for } \lambda \in \left(\frac{u_1}{2}, \infty\right) \tag{3.15}$$

We now illustrate the proofs of $(3.13) - (3.15)$ respectively:

Proof.

$$m_v(\lambda) = \sigma^2 \sum_{i=1}^{p} \frac{u_i}{(u_i+\lambda)^2} \longrightarrow 0 \text{ as } \lambda \longrightarrow \infty$$

□

Proof.

$$m_b(\lambda) = \lambda^2 \boldsymbol{\beta}^\top W^{-2}(\lambda)\boldsymbol{\beta}$$

Where $\lambda^2 W^{-2}(\lambda)$ has eigenvalues:

$$\frac{\lambda^2}{(u_i+\lambda)^2} \longrightarrow 1 \text{ as } \lambda \longrightarrow \infty \text{ for all } i = 1, 2, \cdots, p$$

$$\therefore\ m_b(\lambda) \longrightarrow \boldsymbol{\beta}^\top \boldsymbol{\beta} \text{ as } \lambda \longrightarrow \infty$$

□

Proof.

$$\begin{aligned}
\frac{d^2 m_b(\lambda)}{d\lambda^2} &= 2\boldsymbol{\beta}^\top W^{-2}(\lambda)\boldsymbol{\beta} - 4\lambda\boldsymbol{\beta}^\top W^{-3}(\lambda)\boldsymbol{\beta} - 4\lambda\boldsymbol{\beta}^\top W^{-3}(\lambda)\boldsymbol{\beta} + 6\lambda^2\boldsymbol{\beta}^\top W^{-4}(\lambda)\boldsymbol{\beta} \\
&= 2\boldsymbol{\beta}^\top \left[W^{-2}(\lambda) - 2\lambda W^{-3}(\lambda) - 2\lambda W^{-3}(\lambda) + 3\lambda^2 W^{-4}(\lambda)\right]\boldsymbol{\beta} \\
&\quad 2\boldsymbol{\beta}^\top \left[S^{-2}(\lambda) - 4\lambda W^{-3}(\lambda) + 3\lambda^2 W^{-4}(\lambda)\right]\boldsymbol{\beta} \\
&= 2\boldsymbol{\beta}^\top P(\lambda)\boldsymbol{\beta} \\
\text{Where } P(\lambda) &= W^{-2}(\lambda) - 4\lambda W^{-3}(\lambda) + 3\lambda^2 W^4(\lambda)
\end{aligned}$$

The eigenvalues of $P(\lambda)$ are:

$$\begin{aligned}
\text{eig } Q(\lambda) &= \frac{1}{(u_i+\lambda)^2} - \frac{-4\lambda}{(u_i+\lambda)^3} + \frac{3\lambda^2}{(u_i+\lambda)^4} \\
&= \frac{1}{(u_i+\lambda)^4}[(u_i+\lambda)^2 - 4\lambda(u_i+\lambda) + 3\lambda^2] \\
&= \frac{1}{(u_i+\lambda)^4} u_i^2 + 2\lambda u_i + \lambda^2 - 4\lambda u_i - 4\lambda^2 + 3\lambda^2 \\
&= \frac{1}{(u_i+\lambda)^4} u_i^2 - 2\lambda u_i \\
&= \frac{u_i(u_i - 2\lambda)}{(u_i+\lambda)^4} < 0 \text{ if } \lambda > \frac{u_i}{2} \text{ for all } i = 1, 2, \cdots, p
\end{aligned}$$

This shows $P(\lambda)$ is negative definite for $\lambda > \frac{u_1}{2}$. Thus, $\frac{d^2 m_b(\lambda)}{d\lambda^2} = 2\boldsymbol{\beta}^\top P(\lambda)\boldsymbol{\beta} \leq 0$ for all $\boldsymbol{\beta}$ when $\lambda > \frac{u_1}{2}$ which implies $m_b(\lambda)$ is a concave function of λ. □

To analyze the performance of Ridge Regression on simulated and observational data sets, we use a technique called K-Fold Cross Validation. This function is used to select the tuning parameter λ that minimizes the MSE. We will now outline how Leave One Out Cross

Validation (LOOCV) is carried out and then show that it extends to K-fold Cross Validation. The following method is first applied LOOCV and can then be extended to K-Fold Cross Validation with a different criterion and different estimators:

Step 1: Define a range of interest for the penalty parameter. Divide the data set into training and test set comprising samples $\{1, \cdots, n\}$ and $\{i\}$ respectively.

Step 2: Fit the linear regression model by means of ridge estimation for each λ in the grid using the training set. This yields:

$$\hat{\boldsymbol{\beta}}_{-i}(\lambda) = \left(\boldsymbol{X}_{-i,*}^{\top}\boldsymbol{X}_{-i,*} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{-i,*}^{\top}\boldsymbol{Y}_{-i} \tag{3.16}$$

and this corresponding estimate of the error variance $\hat{\sigma}^2_{-i}(\lambda)$

Step 3:

Evaluate the prediction performance of these models on the the test set by $\log\left\{L\left[Y_i, \boldsymbol{X}_{i,*}; \hat{\boldsymbol{\beta}}_{-i}(\lambda), \hat{\sigma}^2_{-i}(\lambda)\right]\right\}$
Or by the prediction error $|Y_i - \boldsymbol{X}_{i,*}\hat{\boldsymbol{\beta}}_{-i}(\lambda)|$ possibly squared.

Step 4:

We repeat steps 1)-3) such that each sample plays the role of the test set once.

Step 5:

The average prediction performances of the test sets at each grid point of the penalty parameter:

$$\frac{1}{n}\sum_{i=1}^{n} \log\left\{L\left[Y_i, \boldsymbol{X}_{i,*}; \hat{\boldsymbol{\beta}}_{-i}(\lambda), \hat{\sigma}^2_{-i}(\lambda)\right]\right\} \tag{3.16}$$

The quantity above is called the cross-validated log likelihood. It is an estimate of the prediction performance of the model corresponding to this value of the penalty parameter on our data.

Step 6: The value of the penalty parameter that maximizes the cross validated log likelihood is the value of choice.

In the LOOCV procedure above, resampling can be avoided when the prediction performance is measured by Allen's PRESS (Predicted Residual Error Sum of Squares) statistic (Allen 1974). This is because LOOCV prediction performance can be expressed analytically in terms of the known quantities derived from the design matrix and response (Golub, Heath, and Wahba 1979).

We define the optimal penalty parameter to minimize Allen's PRESS statistic as follows:

$$\lambda_{\text{opt}} = \underset{\lambda}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}_{-i}(\lambda) \right]^2 \tag{3.17}$$

We define $\boldsymbol{A}_{\boldsymbol{\nu}}, \boldsymbol{U}_{\boldsymbol{\nu}}, \boldsymbol{V}_{\boldsymbol{\nu}}$ to be $(p \times p)$, $(p \times n)$ and $(n \times p)$ dimensional matrices respectively. The simplified form of the Woodbury identity then is:

$$(\boldsymbol{A}_{\boldsymbol{\nu}} + \boldsymbol{U}_{\boldsymbol{\nu}} \boldsymbol{V}_{\boldsymbol{\nu}})^{-1} = \boldsymbol{A}_{\boldsymbol{\nu}}^{-1} \boldsymbol{U}_{\boldsymbol{\nu}} \left(\boldsymbol{I}_{nn} + \boldsymbol{V}_{\boldsymbol{\nu}} \boldsymbol{A}_{\boldsymbol{\nu}}^{-1} \boldsymbol{U}_{\boldsymbol{\nu}} \right)^{-1} \boldsymbol{V}_{\boldsymbol{\nu}} \boldsymbol{A}_{\boldsymbol{\nu}}^{-1}. \tag{3.18}$$

Proof.

To derive an analytic expression for the right-hand side first, rewrite $\left(\boldsymbol{X}_{-i,*}^{\top} \boldsymbol{X}_{-i,*} + \lambda \boldsymbol{I}_{pp} \right)^{-1}$ by means of the Woodbury identity as:

$$\begin{aligned}
\left(\boldsymbol{X}_{-i,*}^{\top} \boldsymbol{X}_{-i,*} + \lambda \boldsymbol{I}_{pp} \right)^{-1} &= \left(\boldsymbol{X}^{\top} \boldsymbol{X} + \lambda \boldsymbol{I}_{pp} - \boldsymbol{X}_{i,*}^{\top} \boldsymbol{X}_{i,*} \right)^{-1} \\
&= \left(\boldsymbol{X}^{\top} \boldsymbol{X} + \lambda \boldsymbol{I}_{pp} \right)^{-1} + (\boldsymbol{X}^{\top} \boldsymbol{X} + \lambda \boldsymbol{I}_{pp})^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{X} \left(\boldsymbol{X}^{\top} \boldsymbol{X}_{i,*} + \lambda \boldsymbol{I}_{pp} \right)^{-1} \boldsymbol{X} \right]^{-1} \\
\boldsymbol{X}_{i,*} \left(\boldsymbol{X}^{\top} \boldsymbol{X} + \lambda \boldsymbol{I}_{pp} \right)^{-1} &= \left(\boldsymbol{X}^{\top} \boldsymbol{X} + \lambda \boldsymbol{I}_{pp} \right)^{-1} + \left(\boldsymbol{X}^{\top} \boldsymbol{X} + \lambda \boldsymbol{I}_{pp} \right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda) \right]^{-1} \boldsymbol{X}_{i,*} \left(\boldsymbol{X}^{\top} \boldsymbol{X} + \lambda \boldsymbol{I}_{pp} \right)^{-1}
\end{aligned}$$

With $\boldsymbol{H}_{ii}(\lambda) = \boldsymbol{X}_{i,*} \left(\boldsymbol{X}^{\top} \boldsymbol{X} + \lambda \boldsymbol{I}_{pp} \right)^{-1} \boldsymbol{X}_{i,*}^{\top}$.

Furthermore, $\boldsymbol{X}_{-i}^{\top} Y_{-i} = \boldsymbol{X}^{\top} \boldsymbol{Y} - \boldsymbol{X}_{i,*}^{\top} Y_i$.

If we substitute both in the leave one out ridge regression estimator and manipulate:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{-i}(\lambda) &= \left(\boldsymbol{X}_{-i,*}\boldsymbol{X}_{-i,*} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{-i,*}^{\top} \boldsymbol{Y}_{-i} \\
&= \left\{ \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} + \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \boldsymbol{X}_{i,*} \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \right\} \\
&\times \left(\boldsymbol{X}^{\top}\boldsymbol{Y} - \boldsymbol{X}_{i,*}^{\top} Y_i\right) \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}^{\top}\boldsymbol{Y} - \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} Y_i \\
&+ \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \boldsymbol{X}_{i,*} \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}^{\top}\boldsymbol{Y} \\
&- \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \boldsymbol{X}_{i,*} \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} Y_i \\
&= \hat{\boldsymbol{\beta}}(\lambda) - \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right] Y_i \\
&+ \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda) \\
&- \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \boldsymbol{H}_{ii}(\lambda) Y_i \\
&= \hat{\boldsymbol{\beta}}(\lambda) - \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \left\{ \left[1 - \boldsymbol{H}_{ii}\right] Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda) + \boldsymbol{H}_{ii}(\lambda) Y_i \right\} \\
&= \hat{\boldsymbol{\beta}}(\lambda) - \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \left[Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda) \right]
\end{aligned}$$

The latter enables the reformulation of the prediction error as:

$$\begin{aligned}
Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}_{-i}(\lambda) &= Y_i - \boldsymbol{X}_{i,*} \left\{ \hat{\boldsymbol{\beta}}(\lambda) - \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \left[Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda) \right] \right\} \\
&= Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda) + \boldsymbol{X}_{i,*} \left(\boldsymbol{X}^{\top}\boldsymbol{X} + \lambda \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}_{i,*}^{\top} \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \left[Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda) \right] \\
&= Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda) + \boldsymbol{H}_{ii}(\lambda) \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \left[Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}(\lambda) \right] \\
&= \left[1 - \boldsymbol{H}_{ii}(\lambda)\right]^{-1} \left[Y_i - \boldsymbol{X}_{i,*}^{\top} \hat{\boldsymbol{\beta}}(\lambda) \right]
\end{aligned}$$

Which yields the re-expression of Allen's PRESS statistic:

$$\lambda_{\text{opt}} = \underset{\lambda}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \boldsymbol{X}_{i,*} \hat{\boldsymbol{\beta}}_{-i}(\lambda) \right]^2 = \underset{\lambda}{\operatorname{argmin}} \frac{1}{n} \left\| \boldsymbol{B}(\lambda) \left[\boldsymbol{I}_{nn} - \boldsymbol{H}(\lambda) \right] \boldsymbol{Y} \right\|_F^2, \qquad \square \tag{3.21}$$

Where $\boldsymbol{B}(\lambda)$ is diagonal with $[\boldsymbol{B}(\lambda)]_{ii} = [1 - \boldsymbol{H}_{ii}(\lambda)]^{-1}$. So the prediction performance for a given λ can be assessed directly from the hat matrix and the response vector without having to recalculate the n leave-one-out ridge estimators. This is a considerable gain from a computational perspective.

3.2 General Regularization Methods Encompassing Tikhonov, Ivanov and Morozov Solvers

In general regularization, it is convenient to work with a nested sequence for spaces:

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_n \cdots \subset \mathcal{F} \tag{3.19}$$

Where:

$\mathcal{F} = \{\text{all polynomial functions}\}$

$\mathcal{F}_d = \{\text{all polynomials of degree} \leq p\}$

With complexity measure $\Omega : \mathcal{F} \longrightarrow [0, \infty)$.

In all the above, the functional $\Omega(f)$ is called the regularization functional. $\Omega(f)$ is defined in such a way that it controls the complexity of the function f.

$$\Omega(f) = \|f\|^2 = \int_a^b \Big(f''(t)\Big)^2 dt \tag{3.20}$$

We consider all functions in $\mathcal{F}$ with complexity at most r in other words:

$$\mathcal{F}_r = \Big\{ f \in \mathcal{F} | \Omega(f) \leq r \Big\} \tag{3.21}$$

Define for a collection $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ of *i.i.d.* observations as follows:

$x_i \in \mathcal{X} \subset \mathbb{R}^2, i = 1, \cdots, n$ where $\mathcal{X}$ is the input space.

$y_i \in \{-1 + 1\}$ where $\mathcal{Y} = \{-1, +1\}$ is the output space.

If Ω is a norm on $\mathcal{F}$, this would be a ball of radius r in $\mathcal{F}$.

We shall now define the Ivanov and Tikhonov regularization methods:

The Constrained ERM (Ivanov regularization) for complexity measure $\Omega : \mathcal{F} \longrightarrow [0, \infty)$ and fixed $r \geq 0$ s.t. $\Omega(f) \leq r$ is:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i) \tag{3.22}$$

$$\text{s.t. } \Omega(f) \leq r$$

Where r is chosen using validation data or cross validation. Each r corresponds to a different hypothesis spaces.

The Penalized ERM (Tikhonov regularization) for complexity measure $\Omega : \mathcal{F} \longrightarrow \mathbb{R}^{\geq 0}$ and fixed $\lambda \geq 0$ is:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega(f) \tag{3.23}$$

We choose λ using validation data or cross validation.

Next we consider regression models corresponding to it's ERM function. Such models include the linear least squares regression, Ridge regression for both the Tikhonov and Ivanov regularization and finally the Lasso method for both the Tikhonov and Ivanov methods.

First, let us consider the following linear models:

$$\mathcal{F} = \left\{ f : R^p \longrightarrow R | f(x) = \boldsymbol{\beta}^\top x \text{ for } \boldsymbol{\beta} \in R^p \right\} \tag{3.24}$$

Where the Loss function is:

$$\ell(\hat{y}, y) = (y - \hat{y})^2 \tag{3.25}$$

And the training data is :

$$\mathcal{D}_n = ((x_1, y_1), \cdots, (x_n, y_n))$$

The linear least squares regression is ERM for ℓ over $\mathcal{F}$:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in R^p}{\arg\min} \ \frac{1}{n} \sum_{i=1}^{n} \left\{ \boldsymbol{\beta}^\top x_i - y_i \right\}^2 \tag{3.26}$$

This equation can overfit when p is large compared to n or $p >> n$.

The Ridge Regression (Tikhonov Form) solution for regularization parameter $\lambda \geq 0$ is:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{\beta}^\top x_i - y_i\right\}^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{3.27}$$

Where $\|\boldsymbol{\beta}\|_2^2 = \beta_1^2 + \cdots + \beta_d^2$ is the square of the ℓ_2- norm.

The Ridge Regression (Ivanov Form) solution for complexity parameter $r \geq 0$ is:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\|\boldsymbol{\beta}\|_2^2\leq r^2} \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{\beta}^\top x_i - y_i\right\}^2. \tag{3.28}$$

The Lasso Regression (Tikhonov Form) solution for regularization parameter $\lambda \geq 0$ is:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{\beta}^\top x_i - y_i\right\}^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{3.29}$$

where $\|\boldsymbol{\beta}\|_1 = |\beta_1| + \cdots + |\beta_d|$is the ℓ_1- norm

The Lasso Regression (Ivanov Form) solution for complexity parameter $r \geq 0$ is:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\|\boldsymbol{\beta}\|_1\leq r} \frac{1}{n}\sum_{i=1}^{n}\left\{\boldsymbol{\beta}^\top x_i - y_i\right\}^2 \tag{3.30}$$

We will now redefine the class of functions so that we can explore further properties of the Tikhonov and Ivanov. We will also define a new regularization function called the Mozorov regularization. So now, we parametrize the class of functions as follows:

$$h(\boldsymbol{x}) = \boldsymbol{\omega}\cdot\boldsymbol{\phi}(\boldsymbol{x}) + b \tag{3.31}$$

Where $\boldsymbol{\phi} : \mathbb{R}^d \longrightarrow \mathbb{R}^D$ is a mapping function $\boldsymbol{\omega} \in \mathbb{R}^D$ and $b \in \mathbb{R}$. The naïve approach to learning specifically the Empirical Risk Minimization (ERM) consists in searching for the function h that minimizes the empirical error:

$$h : \arg\min_{\boldsymbol{\omega},b} \hat{L}(h) \tag{3.32}$$

Unfortunately ERM is well known to lead to a severe over fitting and then to poor performance in classifying new data originated by the same distribution τ but previously unseen.

So in order to avoid the over fitting issue that afflicts the ERM procedure, the Tikhonov regularization technique can be used which was suggested to solve ill-posed problems:

$$h : \arg\min_{\boldsymbol{\omega},b} \ \hat{L}(h) + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2 \quad \text{or} \quad \arg\min_{\boldsymbol{\omega},b} \ \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\hat{L}(h) \tag{3.33}$$

It should be noted that the solution $\{\boldsymbol{\omega}^*, b^*\}$ in (3.33) is not unique in general (Boyd and Vandenberghe 2004). To eliminate such ambiguity we can opt for the function $h(\boldsymbol{x})$ characterized by the minimum $\|\boldsymbol{\omega}|$ namely the smoothest possible solution. In order to do this without modifying the nature of the regularization procedure, we propose the following equivalent formulation to (3.33):

$$h : \arg\min_{\boldsymbol{\omega},b} \|\boldsymbol{\omega}\| \quad s.t. \ h \in \mathcal{S} \tag{3.34}$$

$$\mathcal{S} = \left\{ h : \hat{L}(h) = \arg\min_{\boldsymbol{\omega},b} \hat{L}(h) \ s.t. \ \|\boldsymbol{\omega}\|^2 \leq \omega_{\text{MAX}}^2 \right\}$$

To simplify the notation further (3.34) we simply add $\|\boldsymbol{\omega}\|$ to the argument of the minimum in (3.33)

$$h : \arg\min_{\boldsymbol{\omega},b,\|\boldsymbol{\omega}\|} \hat{L}(h) \quad s.t. \ \|\boldsymbol{\omega}\|^2 \leq \omega_{\text{MAX}}^2 \tag{3.35}$$

Where $\|\boldsymbol{\omega}\|$ is the Euclidean norm of $\boldsymbol{\omega}$ and implements an under fitting tendency so that the regularization parameter $\lambda \in [0, \infty)$ or equivalently $C = \frac{1}{\lambda} \in [0, \infty)$ balances the influence of the under fitting and the over fitting terms. A consequence of this formulation is that λ implicitly defines the class of functions $\mathcal{H}$ from which the models $h(\boldsymbol{x})$ are selected by the optimization procedure. However, the relation between the regularization parameter and the size of the hypothesis space is not evident at all.

In contrast to Tikhonov scheme, the method of quasi solutions which was first proposed by Ivanov (also known as Ivanov regularization) allows to explicitly control the size of $\mathcal{H}$ by upper bounding the square norm of the admissible hypotheses:

$$h : \arg\min_{\boldsymbol{\omega},b} \hat{L}(h) \ s.t. \ \|\boldsymbol{\omega}\|^2 \leq \omega_{MAX}^2 \tag{3.36}$$

by the means of the regularization parameter $\omega_{MAX}^2 \in [0, \infty)$. A third way to write the regularization problem is the less known approach suggested by Morozov:

$$h : \arg\min_{\boldsymbol{\omega},b} \frac{1}{2}\|\boldsymbol{\omega}\|^2 \; s.t. \; \hat{L}(h) \; \leq \hat{L}_{MAX} \tag{3.37}$$

Again we note that the solution $\{\boldsymbol{\omega}^*, b^*\}$ for (3.37) is not unique (Boyd and Vandenberghe 2004). To eliminate potential ambiguity we opt for the function $h(\boldsymbol{x})$ again, which is characterized by the minimum $\hat{L}(h)$ specifically the solution with minimum error. In order to highlight this without modifying the nature of the regularization procedure, we formulate an equivalent expression to (3.37)

$$h : \arg\min_{\boldsymbol{\omega},b} \hat{L}(h) \quad s.t. \; h \in \mathcal{S} \tag{3.38}$$

$$\mathcal{S} = \left\{ h : \|\boldsymbol{\omega}\|^2 = \arg\min_{\boldsymbol{\omega},b} \|\boldsymbol{\omega}\|^2 s.t. \; \hat{L}(h) \leq \hat{L}_{\text{MAX}} \right\}$$

To simplify (3.38) we add $\hat{L}(h)$ to the argument of the minimum in (3.37)

$$h : \quad \arg\min_{\boldsymbol{\omega},b,\hat{L}(h)} \frac{1}{2}\|\boldsymbol{\omega}\|^2 \quad s.t. \; \hat{L}(h) \leq \hat{L}_{\text{MAX}} \tag{3.39}$$

The philosophy underlying the Morozov regularization approach consists of choosing the simplest function by minimizing $\|\boldsymbol{\omega}\|^2$ which performs better than a pre determined performance threshold on the training set. If the threshold $\hat{L}_{MAX}$ is too small, a solution cannot exist. Therefore for the sake of simplicity, we will assume that $\hat{L}_{MAX}$ is large enough so that a solution can be found. This hypothesis does not modify the nature of Morozov regularization while it helps simplify the subsequent analysis.

According to the ERM principle which was formulated by Vapnik in the Statistical Learning Theory (SLT) framework (Vapnik 2013), learning can easily be implemented by an Ivanov regularization approach. A Tikhonov formulation has been usually preferred as it is easier to solve. Throughout the years, numerous effective methods have been developed to do this. We will show now in the following theorems that the Tikhonov, Ivanov and Morozov regularization approaches are equivalent or are three similar faces of the same problem.

Specifically we show how the Ivanov and Morozov problems can be solved through the procedures originally designed for the Tikhonov based formulation (Oneto, Ridella, and Anguita 2016). First we show that the value of the Tikhonov regularization parameter exists s.t. all three problems are equivalent.

Theorem 1. *Let us consider an Ivanov (or Morozov) regularization problem as formulated in* (3.35) *and* (3.39) *then, there exists a value of* $C = \frac{1}{\lambda}$ *for the Tikhonov regularization problem* (3.33) *such that the formulations are equivalent.*

Theorem 2. *Let us consider the Tikhonov and Ivanov formulations. Let* $\left(\|\boldsymbol{\omega}_T^*\|, \hat{L}_T^*\right)$ *and*

$\left(\|\boldsymbol{\omega}_I^*\|, \hat{L}_I^*\right)$ *be the solutions of respectively the Tikhonov and the Ivanov problem. If* $\|\boldsymbol{\omega}_T^*\| = \|\boldsymbol{\omega}_I^*\|$ *for a given* $C = \frac{1}{\lambda}$ *and for a given* ω_{MAX} *then* $\hat{L}_T^* = \hat{L}_1^*$ *and vice versa.*

Theorem 3. *Let us consider the Tikhonov and Morozov formulations. Let* $\left(\|\boldsymbol{\omega}_T^*\|, \hat{L}_T^*\right)$ *and* $\left(\|\boldsymbol{\omega}_M^*\|, \hat{L}_M^*\right)$ *be the solutions of respectively the Tikhonov and the Morozov problems. If* $\|\boldsymbol{\omega}_T^*\| = \|\boldsymbol{\omega}_M^*\|$ *for a given* $C = \frac{1}{\lambda}$ *and for a given* $\hat{L}_{MAX}$ *then* $\hat{L}_T^* = \hat{L}_M^*$ *and vice versa.*

In the next 3 theorems we show some properties that allow us to define general procedures for solving either an Ivanov or Morozov problem through designed techniques from Tikhonov formulations (Oneto, Ridella, and Anguita 2016). Theorem 4 can be exploited to prove theorem 5 which in turn can further be exploited to design actual learning algorithms. Theorem 6 proves that if $\|\boldsymbol{\omega}_C^*\|$ stops increasing as C increases, it will remain the same regardless of the value assumed by the regularization parameter. We now present the theorems:

Theorem 4. *Let us consider the Tikhonov formulation. Let us solve* (3.33) *for two given values of the regularization parameter* C_1 *and* $C_2 > C_1$*. In particular let the solutions of the problem be respectively* $\left(\|\boldsymbol{\omega}_{C_1}^*\|, \hat{L}_{C_1}^*\right)$ *for* C_1 *and* $\left(\|\boldsymbol{\omega}_{C_2}^*\|, \hat{L}_{C_2}^*\right)$ *for* C_2 *so that the corresponding values of the objective functions are:*

$$K^{C_1} = \frac{1}{2}\left(\|\boldsymbol{\omega}_{C_1}^*\|\right)^2 + C_1\hat{L}_{C_1}^*, \qquad K^{C_2} = \frac{1}{2}\left(\|\boldsymbol{\omega}_{C_2}^*\|\right)^2 + c_2\hat{L}_{C_2}^* \tag{3.40}$$

Then:

$$K^{C_2} \geq K^{C_1} \tag{3.41}$$

Theorem 5. *Let us consider the Tikhonov formulation. Given* $C_1, C_2 \in [0, +\infty]$ *such that* $C_2 > C_1$ *let us solve* (3.33) *and let* K^{C_1} *and* K^{C_2} *be the corresponding values of the objective functions, then:*

$$\left(\|\boldsymbol{\omega}_{C_2}^*\| > \|\boldsymbol{\omega}_{C_1}^*\| \Longrightarrow \hat{L}_{C_2}^* < \hat{L}_{C_1}^*\right) \vee \left(\|\boldsymbol{\omega}_{C_2}^*\| = \|\boldsymbol{\omega}_{C_1}^*\| \Longrightarrow \hat{L}_{C_2}^* = \hat{L}_{C_1}^*\right) \tag{3.42}$$

Theorem 6. *Let us consider the Tikhonov formulation. Let* $\left\|\boldsymbol{\omega}_{C_\infty}^*\right\|$ *be the solution to the regularization problem for a given value of* C_∞*. If* $\exists\ C > C_\infty$ *such that* $\|\boldsymbol{\omega}_C^*\| = \|\boldsymbol{\omega}_{C_\infty}^*\|$ *then* $\|\boldsymbol{\omega}_C^*\|$ *will not vary* $\forall\ C \geq C_\infty$

3.3 Lasso / L1 Regularization Methods Sparsity, Smoothness and Uniqueness

The Lasso (Tibshirani 1996) is introduced as follows:

$$L(\boldsymbol{\beta}, \lambda) = RSS(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\boldsymbol{\beta}_j| \tag{3.43}$$

The lasso estimator $\hat{\boldsymbol{\beta}}^L(\lambda)$ minimizes $L(\boldsymbol{\beta}, \lambda)$ over $\boldsymbol{\beta}$ for a given λ where λ increases as $\left\|\hat{\boldsymbol{\beta}}^L(\lambda)\right\|$ decreases. We yield the following expression of the intercept term after standardization and obtain the Lasso coefficient estimates:

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \bar{x}_j \hat{\beta}_j^L(\lambda) \tag{3.44}$$

From duality and KKT conditions we can rewrite (1) and (2) in the following form:

$$\text{Ridge Regression: } \min_{\beta \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda |\boldsymbol{\beta}|_2^2 \tag{3.45}$$

$$\text{LASSO: } \min_{\beta \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda + \lambda \|\boldsymbol{\beta}\|_1 \tag{3.46}$$

Throughout the paper we will use the penalized form of the Lasso problem and a combination of optimization tools and statistics tools to describe some of its favorable properties. These will apply to the more general problem.

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f(\boldsymbol{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \tag{3.47}$$

This form is called the penalized or Lagrange form of the problem where it is the analog of the constrained form where for every $\alpha > 0$ there is a corresponding $\lambda > 0$ with the same solutions.

1. There need not always be a unique solution in $\hat{\boldsymbol{\beta}}$ since the relation is not always strictly convex in $\boldsymbol{\beta}$ This happens in scenario 2 where $p > n$.

2. However there is in a unique fitted value $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ this is because the least squares loss function is strictly convex in $\boldsymbol{X}\boldsymbol{\beta}$

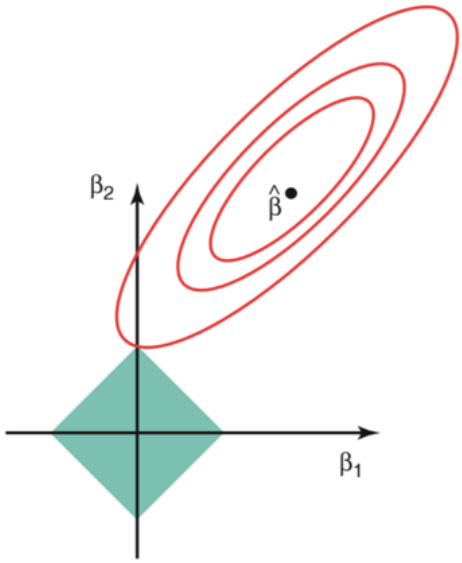

FIGURE 3.2: Geometric Interpretation of the Lasso in 2-Dimensional Space.

The Lasso performs $L1$ shrinkage so that there are "corners" in the constraint, which in two dimensions corresponds to a diamond. If the sum of squares "hits" one of these corners, then the coefficient corresponding to the axis is shrunk to zero. The contours shown are of the error and constraint functions. The solid diamond shape area is the constraint regions $|\beta_1| + |\beta_2| \leq t$ while the ellipses are the contours of the least squares error functions. As p increases, the multidimensional diamond has an increasing number of corners and so it is highly likely that some coefficients will be set equal to zero. Hence the Lasso performs shrinkage and effectively subset selection.

In contrast with subset selection, Lasso performs a soft thresholding. As the smoothing parameter is varied, the sample path for the estimates moves continuously to zero. We would be worried if we encountered the same problems with interpretation as we did with OLS regression. However, this won't be a problem as can be seen shortly. The Lasso is useful when dealing with an ill-conditioned model matrix $\boldsymbol{X}$ for $p > n$. When $p > n$, the Lasso provides a better variable selection method than ridge regression. The Lasso provides a sparse solution by penalizing the sum of the absolute values of the estimates. As λ increases, the number of significant coefficients decreases. Hence this makes the Lasso for variable selection and interpretation of the results a more plausible method than ridge regression.

We now give a partial answer to when the Lasso solution is unique. We provide sufficient conditions for a unique minimizer of the Lasso criterion. We start by using the following lemmas (R. J. Tibshirani et al. 2013):

Lemma 1. *For any* $\boldsymbol{y}, \boldsymbol{X}$, *and* $\lambda \geq 0$, *the Lasso problem a* (3.43) *has the following properties:*

- *There is either a unique Lasso solution or an (uncountably) infinite number of solutions*
- *Every Lasso solution* $\hat{\boldsymbol{\beta}}$ *gives the same fitted values* $\boldsymbol{X}\hat{\boldsymbol{\beta}}$
- *If* $\lambda > 0$ *then every Lasso solution* $\hat{\boldsymbol{\beta}}$ *has the same* ℓ_1 *norm* $\left\|\hat{\boldsymbol{\beta}}\right\|_1$

To go beyond the basics we turn to the Karush-Kuhn-Tucket (KKT) optimality conditions for the Lasso problem (3.43). These conditions can be written as:

$$\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \lambda\gamma \tag{3.48}$$

$$\gamma_i \in \begin{cases} \left\{\text{sign}(\hat{\beta}_i)\right\} & \text{if } \hat{\beta}_i \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_i = 0 \end{cases} \quad \text{for } i = 1, \cdots, p. \tag{3.49}$$

Here $\gamma \in \mathbb{R}^p$ is called a subgradient of the function $f(x) = \|\boldsymbol{x}\|_1$ evaluated at $\boldsymbol{x} = \hat{\boldsymbol{\beta}}$. Therefore, $\hat{\boldsymbol{\beta}}$ is a solution in (3.44) iff $\hat{\boldsymbol{\beta}}$ satisfies (3.48) and (3.49) for some γ. We now use the KKT conditions to write the Lasso fit and solutions in a more explicit form. In what follows, we assume that $\lambda > 0$ to simplify things. We start by defining the equicorrelation set $\mathcal{E}$ by:

$$\mathcal{E} = \left\{i \in \{1, \cdots, p\} : \left|X_i^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right| = \lambda\right\} \tag{3.50}$$

The equicorrelation set $\mathcal{E}$ is named this way since when $\boldsymbol{y}, \boldsymbol{X}$ have been standardized, $\mathcal{E}$ contains the variables that have equal and maximal absolute correlations with the residual. We define the equicorrelation signs s as follows:

$$s = \text{sign}\left(X_{\mathcal{E}}^\top\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right). \tag{3.51}$$

If we recall (3.48) we note that the optimal subgradient γ is unique by the uniqueness of the fit $\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Similarly, we can also define $\mathcal{E}, s$ in terms of γ, as in $\mathcal{E} = \{i \in \{1, \cdots, p \,|\gamma_i| = 1\}\}$ and $s = \gamma_{\mathcal{E}}$. The uniqueness of $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ implies the uniqueness of $\mathcal{E}, s$. We know that $\hat{\beta}_{-\mathcal{E}} = 0$ for any Lasso solution $\hat{\boldsymbol{\beta}}$ by definition of the subgradient γ in (3.49). Thus the $\mathcal{E}$ block of (3.48) can be written as:

$$X_{\mathcal{E}}^\top\left(y - X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}}\right) = \lambda_s. \tag{3.52}$$

This shows us that $\lambda_s \in \text{row}\,(X_{\mathcal{E}})$, so $\lambda_s = X_{\mathcal{E}}^\top\left(X_{\mathcal{E}}^\top\right)^+\lambda_s$. We use this fact and rearrange (3.52) to yield:

$$X_{\mathcal{E}}^\top X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}} = X_{\mathcal{E}}^\top\left(y - \left(X_{\mathcal{E}}^\top\right)^+\lambda s\right). \tag{3.53}$$

Therefore the (unique) Lasso fit $\boldsymbol{X}\hat{\boldsymbol{\beta}} = X_{\mathcal{E}}\hat{\beta}_{\mathcal{E}}$ is:

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = X_{\mathcal{E}}(X_{\mathcal{E}})^{+}\left(\boldsymbol{y} - (X_{\mathcal{E}})^{+}\lambda s\right), \tag{3.54}$$

And any Lasso solution $\hat{\boldsymbol{\beta}}$ is of the form:

$$\hat{\beta}_{-\mathcal{E}} = 0 \quad \text{and} \quad \hat{\beta}_{\mathcal{E}} = (X_{\mathcal{E}})^{+}\left(y - (X_{\mathcal{E}}^{\top})^{+}\lambda s\right) + b \tag{3.55}$$

Where $b \in \text{null}\,(X_{\mathcal{E}})$. Specifically any $b \in \text{null}\,(X_{\mathcal{E}})$ produces a Lasso solution $\hat{\beta}$ in (3.55) provided that $\hat{\beta}$ has the correct signs over its nonzero coefficients that is sign $(\hat{\beta}_i) = s_i$ for all $\hat{\beta}_i \neq 0$. We can write these conditions together as:

$$b \in \text{null}(X_{\mathcal{E}}) \quad \text{and} \quad s_i \cdot \left(\left[(X_{\mathcal{E}})^{+}\left(y - (X_{\mathcal{E}}^{\top})^{+}\lambda s\right)\right]_i + b_i\right) \geq 0 \;\; \text{for } i \in \mathcal{E}, \tag{3.56}$$

And thus, any b satisfying (3.56) gives a Lasso solution $\hat{\boldsymbol{\beta}}$ in (3.55). In the following part we use a sequence of straightforward arguments to prove that the Lasso solution is unique under somewhat general conditions. From the work developed in the previous part, we can see that if null $(X_{\mathcal{E}}) = \{0\}$ then the Lasso solution is unique and is given by (3.55) with $b = 0$. (We note that $b = 0$ necessarily satisfies the sign condition in (3.56) because a Lasso solution is guaranteed to exist by Lemma 1. If we then rearrange (3.55) to emphasize the rank of $X_{\mathcal{E}}$, we yield the following result.

Lemma 2. *For any $\boldsymbol{y}, \boldsymbol{X}$, and $\lambda > 0$ if null$(X_{\mathcal{E}}) = \{0\}$ or equivalently if rank $(X_{\mathcal{E}}) = |\mathcal{E}|$ then the Lasso solution is unique and is given by:*

$$\hat{\beta}_{\mathcal{E}} = 0 \quad \textit{and} \quad \hat{\beta}_{\mathcal{E}} = \left(X_{\mathcal{E}}^{\top}X_{\mathcal{E}}\right)^{-1}\left(X_{\mathcal{E}}^{\top}y - \lambda s\right) \tag{3.57}$$

where $\mathcal{E}$ and s are the equicorrelation set and signs as defined in (3.50) *and* (3.51)*. Note that this solution has at most min $\{n, p\}$ nonzero components.*

We will show later that the same condition is actually also necessary for almost every $\boldsymbol{y} \in \mathbb{R}^n$. Note that $\mathcal{E}$ depends on the Lasso solution at $\boldsymbol{y}, \boldsymbol{X}, \lambda$ and hence the condition null $(X_{\mathcal{E}}) = \{0\}$ is somewhat circular. There are more natural conditions depending on $\boldsymbol{X}$ alone that imply null$(X_{\mathcal{E}}) = \{0\}$. To see this suppose that null$(X_{\mathcal{E}}) \neq \{0\}$ then for some $i \in \mathcal{E}$ we can write:

$$X_i = \sum_{j \in \mathcal{E}\setminus\{i\}} c_j X_j, \tag{3.58}$$

Where $c_j \in \mathbb{R}, j \in \mathcal{E}\setminus\{i\}$. Hence:

$$s_i X_i = \sum_{j \in \mathcal{E}\setminus\{i\}} (s_i s_j c_j) \cdot (s_j X_j). \tag{3.59}$$

By definition of the equicorrelation set $X_j^{\top} r = s_j \lambda$ for any $j \in \mathcal{E}$ where $r = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the Lasso residual. Taking the inner product of both sides above with r, we get:

$$\lambda = \sum_{j \in \mathcal{E}\setminus\{i\}} (s_i s_j c_j)\, \lambda \tag{3.60}$$

or:

$$\sum_{j\in\mathcal{E}\backslash\{i\}}(s_i s_j c_j) = 1, \tag{3.61}$$

Assuming that $\lambda > 0$. Therefore we have shown that if null $(X_{\mathcal{E}}) \neq \{0\}$ then for some $i \in \mathcal{E}$

$$s_i X_i = \sum_{j\in\mathcal{E}\backslash\{i\}} a_j \cdot s_j X_j \tag{3.62}$$

With $\sum_{j\in\mathcal{E}\backslash\{i\}} a_j = 1$ which means that $s_i X_i$ lies in the affine span of $s_j X_j, j \in \mathcal{E}\backslash\{i\}$. Note that we can assume, without a loss of generality that $\mathcal{E}\backslash\{i\}$ has at most n elements since otherwise we can simply repeat the above arguments replacing $\mathcal{E}$ by any one of its subsets with $n+1$ elements; hench the affine span of $s_j X_j, j \in \mathcal{E}\backslash\{i\}$ is at most $n-1$ dimensional. We say that the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has columns in general position if no $k-$ dimensional subspace $L \subset \mathbb{R}^n$, for $k <$min $\{n,p\}$, contains more than $k+1$ elements of the set $\{\pm X_1, \cdots, \pm X_p\}$ excluding antipodal pairs. Another way of saying this L the affine span of any $k+1$ point $\sigma_1 X_{i_1}, \cdots, \sigma_{k+1} X_{i_{k+1}}$, for arbitrary signs $\sigma_1, \cdots, \sigma_{k+1} \in \{-1, 1\}$ does not contain any element of $\{\pm X_i : i \neq i_1, \cdots, i_{k+1}\}$. From what we have just shown, the predictor matrix $\boldsymbol{X}$ having columns in general position is enough to ensure uniqueness.

Lemma 3. *If the columns of $\boldsymbol{X}$ are in general position then for any $\boldsymbol{y}$ and $\lambda > 0$ the Lasso solution is unique and is given by* (3.57)

Although the definition of general position may seem somewhat technical, this condition is naturally satisfied when the entries of the predictor matrix $\boldsymbol{X}$ are drawn from a continuous probability distribution. More precisely, if the entries of $\boldsymbol{X}$ follow a joint distribution that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{np}$, then the columns of $\boldsymbol{X}$ are in general position with probability one. To see this first, consider the probability $P(X_{k+2} \in aff\{X_1, \cdots, X_{k+1}\})$ where aff$\{X_1, \cdots, X_{k+1}\}$ denotes the affine span of $X_1, \cdots, X_{k+1}$. Note that by continuity:

$$\mathrm{P}\left(X_{k+2} \in \mathrm{aff}\{X_1, \cdots, X_{k+1}\} \,|X_1, \cdots, X_{k+1}\right) = 0 \tag{3.63}$$

because (for fixed $X_1, \cdots, X_{k+1}$) the set aff $\{X_1, \cdots, X_{k+1}\} \subseteq \mathbb{R}^n$ has Lebesgue measure zero. Therefore, integrating over $X_1, \cdots, X_{k+1}$ we get that $\mathrm{P}(X_{k+2} \in \mathrm{aff}\{X_1, \cdots, X_{k+1}\}) = 0$. Taking a union over all subsets of $k+2$ columns, all combinations of $k+2$ signs and all $k <$ min$\{n,p\}$, we conclude that with probability zero, the columns are not in general position. This leads us to our final sufficient condition for uniqueness of the Lasso solution.

Lemma 4. *If the entries of $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ are drawn from a continuous probability distribution on $\mathbb{R}^{np}$ then for any $\boldsymbol{y}$ and $\lambda > 0$ the Lasso solution is unique and is given by* (3.57) *with probability one.*

According to this result we essentially never have to worry about uniqueness when the predictor variables come from a continuous distribution regardless of the sizes of n and p. Actually, there is nothing really special about ℓ_1 penalized linear regression. Next we show that the same uniqueness result hold for ℓ_1 penalized minimization with any differentiable

strictly convex loss function. We consider the more general minimization problem:

$$\hat{\boldsymbol{\beta}} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; f(\boldsymbol{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \tag{3.64}$$

Where the loss function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and strictly convex. To be clear, we mean that f is strictly convex in its argument. For instance, the function $f(u) = \|\boldsymbol{y} - u\|_2^2$ is strictly convex even though $f(\boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ may bot be strictly convex in $\boldsymbol{\beta}$. (i) There is either a unique solution or uncountably many solutions; (ii) every solution $\hat{\boldsymbol{\beta}}$ gives the same fit $\boldsymbol{X}\hat{\boldsymbol{\beta}}$;$(iii)$ if $\lambda > 0$ then every solution $\hat{\boldsymbol{\beta}}$ has the same ℓ_1 norm. The KKT conditions for (3.64) can be expressed as:

$$\boldsymbol{X}^\top (-\nabla f)\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \lambda\gamma \tag{3.65}$$

$$\gamma_i \in \begin{cases} \{\operatorname{sign}\left(\hat{\beta}_i\right)\} & \text{if } \hat{\beta}_i \neq 0 \\ [-1,1] & \text{if } \hat{\beta}_i = 0 \end{cases}, \quad \text{for } i = 1, \cdots, p \tag{3.66}$$

Where $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is the gradient of f and we can define the equicorrelation set and signs in the same way as before:

$$\mathcal{E} = \left\{ i \in \{1, \cdots, p\} : \left| X_i^\top (-\nabla f)\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \right| = \lambda \right\} \tag{3.67}$$

and

$$s = \operatorname{sign}\left(X_{\mathcal{E}}^\top (-\nabla f)\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right) \tag{3.68}$$

The subgradient condition (3.66) implies that $\hat{\beta}_{-\mathcal{E}} = 0$ for any solution $\hat{\boldsymbol{\beta}}$ in (3.64). For squared error loss recall that we then explicitly solved for $\hat{\beta}_{\mathcal{E}}$ as a function of $\mathcal{E}$ and s. This is not possible for a general loss function f; but given $\mathcal{E}$ and s, we can rewrite the minimization problem (3.64) over the coordinates in $\mathcal{E}$ as:

$$\hat{\beta}_{\mathcal{E}} \in \underset{\beta_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}}{\operatorname{argmin}} \; f\left(X_{\mathcal{E}}\beta_{\mathcal{E}}\right) + \lambda \left\|\beta_{\mathcal{E}}\right\|_1 \tag{3.69}$$

Now if $\operatorname{null}(X_{\mathcal{E}}) = \{0\}$ (equivalently $\operatorname{rank}(X_{\mathcal{E}})$), then the criterion in (3.69) is strictly convex as f itself is strictly convex. This implies that there is a unique solution $\hat{\beta}_{\mathcal{E}}$ in (3.69) and therefore a unique solution $\hat{\boldsymbol{\beta}}$ in (3.64). Hence we arrive at the same conclusions as those made in adsadsa that there is a unique solution in (3.64) if the columns of $\boldsymbol{X}$ are in general position and ultimately the following result.

Lemma 5. *If $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ has entries drawn from a continuous probability distribution on $\mathbb{R}^{np}$ then for any differentiable strictly convex function g and for any $\lambda > 0$ the minimization problem* (3.64) *has a unique solution with probability one. This solution has at most $\min\{n, p\}$ nonzero components.*

This general result applies to any differentiable strictly convex loss function f which is quite a broad class. For example it applies to logistic regression loss:

$$f(u) = \sum_{i=1}^{n} \Big[-y_i u_i + \log\left(1 + \exp\left(u_i\right)\right) \Big] \tag{3.70}$$

Where typically (but not necessarily) each $y_i \in \{0,1\}$ and Poisson regression loss:

$$f(u) = \sum_{i=1}^{n} \Big[-y_i u_i + \exp(u_i) \Big] \tag{3.71}$$

Where typically (but again not necessarily) each $y_i \in \mathbb{N} = \{0,1,2,\cdots\}$. We shift our focus in the next section and without assuming any conditions for uniqueness we show how to compute a solution path for the Lasso problem (over the regularization parameter λ).

3.4 The Bayesian Lasso and Bridge

We have seen from Tibshirani's Lasso (Tibshirani 1996) that it estimates linear regression coefficients through L_1 constrained least squares. They achieve:

$$\min_{\beta} \ (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^{\top} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.72}$$
$$\text{For some } \lambda \geq 0 \text{ where } \tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbb{1}_n.$$

The whole path of Lasso estimates for all tuning paramaeters that λ can be computed by modifying the LARS algorithm of (Efron et al. 2004). Noting the form of the Lasso penalty in (3.72) suggested that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (double exponential) priors. Through this connection, several other authors (Figueiredo and Gomes 2003), (Bae and Mallick 2004), (Yuan and Lin 2005) subsequently suggested using Laplace like priors. For instance, we can consider a fully Bayesian analysis using a conditional Laplace prior specification of the form:

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \tag{3.73}$$

And the noninformative scale invariant marginal prior $\pi(\sigma^2) = 1/\sigma^2$ on σ^2. The conditioning on σ^2 is a must because it gaurantees a unimodal full posterior. This is needed since lack of unimodality slows convergence of the Gibbs sampler and makes point estimates less meaningful. The Gibbs sampler for the Bayesian Lasso exploits the following representation of the Laplace distribution as a scale mixture of normals (with an exponential mixing density):

$$\frac{a}{2} e^{-a|z|} = \int_0^{\infty} \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds, \qquad a > 0 \tag{3.74}$$

This suggests the following hierarchical representation of the full model:

$$
\begin{aligned}
\boldsymbol{y}|\mu,\boldsymbol{X},\boldsymbol{\beta},\sigma^2 &\sim N_n(\mu\mathbb{1}_n+\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n),\\
\boldsymbol{\beta}|\sigma^2,\tau_1^2,\cdots,\tau_p^2 &\sim N_p(\boldsymbol{0}_p,\sigma^2\boldsymbol{D}_\tau),\\
\boldsymbol{D}_\tau &= \mathrm{diag}(\tau_1^2,\cdots,\tau_p^2),\\
\sigma^2,\tau_1^2,\cdots,\tau_p^2 &\sim \pi(\sigma^2)d\sigma^2\prod_{j=1}^{p}\frac{\lambda^2}{2}e^{-\lambda^2\tau_j^2/2}d\tau_j^2,\\
\sigma^2,\tau_1^2,\cdots,\tau_p^2 &> 0.
\end{aligned}
\tag{3.75}
$$

Similar hierarchies based on (3.74) have been used by other authors. Since the columns of $\boldsymbol{X}$ are centered it is easy to analytically integrate μ from the joint posterior under its independent flat prior. Since μ is rarely of interest, we marginalize it out in the interest of simplicity and speed. If desired it can be reintroduced with a full conditional distribution that is normal with mean $\bar{y}$ and variance σ^2/n.

Marginalizing over μ does not affect conjugacy. The full conditional distributions of $\boldsymbol{\beta},\sigma^2$, and $\tau_1^2,\cdots,\tau_p^2$ are still easy to sample, and they depend on the centered response vector $\tilde{\boldsymbol{y}}$. The full conditional for $\boldsymbol{\beta}$ is multivariate normal with mean $\boldsymbol{A}^{-1}\boldsymbol{X}^\top\tilde{\boldsymbol{y}}$ and variance $\sigma^2\boldsymbol{A}^{-1}$, where $\boldsymbol{A}=\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{D}_\tau^{-1}$. The full conditional for σ^2 is inverse-gamma with shape parameter (n-1)/2+p/2 and scale parameter $(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})^\top(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})/2+\boldsymbol{\beta}^\top\boldsymbol{D}_\tau^{-1-1}\boldsymbol{\beta}/2$ and $\tau_1^2,\cdots,\tau_p^2$ are conditionally independent with $1/\tau_j^2$ conditionally inverse-Gaussian with parameters:

$$
\mu' = \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}} \quad \text{and} \quad \lambda' = \lambda^2 \tag{3.76}
$$

In the parameterization of the inverse-Gaussian density given by:

$$
f(x) = \sqrt{\frac{\lambda'}{2\pi}}x^{-3/2}\exp\left\{-\frac{\lambda'(x-\mu')^2}{2(\mu')^2x}\right\}, \quad x>0 \tag{3.77}
$$

(Chhikara and Folks 1989). These full conditionals form the basis for an efficient Gibbs sampler with block updating of $\boldsymbol{\beta}$ and $(\tau_1^2,\cdots,\tau_p^2)$ the convergence is really fast. The parameter of the ordinary Lasso can be chosen by cross validation generalized cross-validation and ideas based on The Bayesian Lasso also offers some uniquely Bayesian alternatives: empirical Bayes through marginal maximum likelihood and use of an appropriate hyper prior.

(Casella 2001) proposed a Monte Carol EM algorithm that complements a Gibbs sample and provides marginal maximum likelihood estimates of hyper parameters. For the Bayesian Lasso, each iteration of the algorithm involves running the Gibbs sampler using a λ value estimated from the sample of the previous iteration. Specifically iteration k uses

the Gibbs sampler in (Park and Casella 2008) with hyper parameter $\lambda^{(k-1)}$ (i.e., the estimate from iteration $k-1$) to approximate the ideal updated estimate:

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\lambda^{(k-1)}}[\tau_j^2|\tilde{\boldsymbol{y}}]}} \tag{3.78}$$

by replacing the conditional expectations with averages from the Gibbs sample. We suggest the initial value:

$$\lambda^{(0)} = p\sqrt{\hat{\sigma}_{LS}^2} \Big/ \sum_{j=1}^{p} |\hat{\beta}_j^{LS}| \tag{3.79}$$

Where $\hat{\sigma}_{LS}^2$ and $\hat{\beta}_j^{LS}$ are estimates from the usual least squares procedure. This empirical estimate tends to be smaller than the maximizing λ, but our experience suggests that only extreme initial overestimates of λ lead to slow convergence. Because the expectations are estimated from the Gibbs sampler the successive λ estimates will not quite converge, but will eventually drift randomly about the true maximum likelihood estimate, with less drift if more Gibbs samples are taken in each iteration.

An alternative to choosing λ explicitly is to give it a diffuse hyper prior. We consider the class of gamma priors on λ^2 (not λ) of the form:

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2}, \quad \lambda^2 > 0 \ (r > 0, \delta > 0), \tag{3.80}$$

because the resulting conjugacy allows easy extension of the Gibbs sampler. The improper scale invariant prior $1/\lambda^2$ for λ^2 ($r = 0, \delta = 0$) is tempting but it leads to an improper posterior. Moreover, scale invariance is not a very compelling criterion because λ is unit-less. When prior (3.80) is used in the hierarchy of (3.75) the full conditional distribution of λ^2 is gamma with shape parameter $p + r$ and rate parameter $\sum_{j=1}^{p} \tau_j^2/2 + \delta$. With this specification λ^2 can simply join the other parameters in the Gibbs sampler, because the full conditional distributions of the other parameters do not change. The prior density for λ^2 should approach 0 sufficiently fast as $\lambda^2 \longrightarrow \infty$ but should be relatively flat and place high probability near the maximum likelihood estimate.

One direct generalization of the Lasso and ridge regression is penalized regression by solving:

$$\min_{\boldsymbol{\beta}} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^\top (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^q \tag{3.81}$$

For some $q \geq 0$ with the $q = 0$ corresponding to best subset regression, $q = 1$ corresponds is the ordinary Lasso and $q = 2$ is the ridge regression. For $q \geq 1$ the expression is called "Bridge Regression". Ofcourse $q = 1$ is the ordinary Lasso and $q = 2$ is Ridge Regression. The Bayesian analog of this penalization involves using a prior on $\boldsymbol{\beta}$ of the form:

$$\pi(\boldsymbol{\beta}) \propto \prod_{j=1}^{p} e^{-\lambda|\beta_j|^q} \tag{3.82}$$

We can also keep up with (3.73) by using:

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \prod_{j=1}^{p} e^{-\lambda(|\beta_j|/\sqrt{\sigma^2})^q} \tag{3.83}$$

Thus the elements of $\boldsymbol{\beta}$ have independent priors from the exponential power distribution. Although this term is usually reserved for the case where $q \geq 1$. Whenever $0 < q \leq 2$ this distribution is represented by a scale mixture of normals for $0 < q < 2$:

$$e^{-|z|^q} \propto \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{1}{s^3/2} g_{q/2}\left(\frac{1}{2s}\right) ds \tag{3.84}$$

Where $g_{q/2}$ is the density of a positive stable random variable with index $q/2$ ((West 1987), (Gneiting 1997)) which generally does not has a closed form expression. A hierarchy of the type discussed earlier is applicable by placing appropriate independent distributions on $\tau_1^2, \cdots, \tau_p^2$. Their full conditional distributions are closely related to certain exponential dispersion models (Jørgensen 1987). It is not clear if an efficient Gibbs sampler can be based on this hierarchy.

3.5 Elastic Net Regularization and Variable Selection Techniques

The Lasso faces the following issues:

- In the $p > n$ case because of the nature of the convex optimization problem the Lasso selects at most n variables before it saturates. This is a limiting feature for a variable selection method. The Lasso is not well defined unless the bound on the $L2$-norm of the coefficients is smaller than a certain value.
- If the pairwise correlations among a group of variables is quite high, then the Lasso tends to select only one variable and it will not care as to which variable is selected.

- For the other scenario where $n > p$, if there are high correlations between predictors, it has been observed empirically that the Lasso's prediction performance is dominated by the ridge regression.

To solve the above issue we introduce a new regularization technique which is called the elastic net. This is similar to the Lasso in a sense that the elastic net procedure does automatic variable selection and continuous shrinkage and it can select groups of correlated variables. To illustrate further the uses of this regularization technique, we start with the definition of the naïve elastic net. First we must define our predictors and response for the model.

We suppose that $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$ be the response and $\boldsymbol{X} = [x_1|\cdots|x_p]$ be the model matrix, where $x_j = (x_{1j}, \cdots, x_{nj})^\top, j = 1, \cdots, p$ are the predictors. After a scale transformation, we can assume, that the response is centred and the predictors are standardized.

$$\sum_{i=1}^{n} y_i = 0 \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for } j = 1, 2, \cdots, p \tag{3.85}$$

So now for any fixed non-negative λ_1 and λ_2 we define the naïve elastic net criterion:

$$L(\lambda_1, \lambda_2, , \boldsymbol{\beta}) = |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|^2 + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1 \tag{3.86}$$

Where:

$$|\boldsymbol{\beta}|^2 = \beta_j^2$$

The ridge regression penalty part is:

$$\lambda_2 |\boldsymbol{\beta}|^2 \tag{3.87}$$

The Lasso penalty part is:

$$\lambda_1 |\boldsymbol{\beta}|_1 \tag{3.88}$$

The naiïve elastic net estimator $\hat{\boldsymbol{\beta}}^E$ is the minimizer of equation (3.86) :

$$\hat{\boldsymbol{\beta}}^E = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{L(\lambda_1, \lambda_2, \boldsymbol{\beta})\} \tag{3.89}$$

This procedure can be viewed as a penalized least squares method. Let $\alpha = \frac{\lambda_2}{(\lambda_1+\lambda_2)}$ then solving $\hat{\boldsymbol{\beta}}$ in equation (3) is equivalent to the optimization problem:

$$\hat{\boldsymbol{\beta}}^E = \underset{\beta}{\operatorname{argmin}} \ |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|^2, \text{subject to } (1-\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2 \leq t \text{ for some } t. \tag{3.90}$$

The elastic net penalty, which is a convex combination of the Lasso and Ridge penalty. When $\alpha = 1$, the naive elastic net becomes simple ridge regression. In this paper, we only consider $\alpha < 1$. $\forall \ \alpha \in [0,1)$, the elastic net penalty function is singular (without first derivative) at 0 and it is strictly convex $\forall \ \alpha > 0$, thus possessing the characteristics of both the Lasso and ridge. Note that the Lasso penalty ($\alpha = 0$) is convex but not strictly convex. These arguments can be seen clearly from:

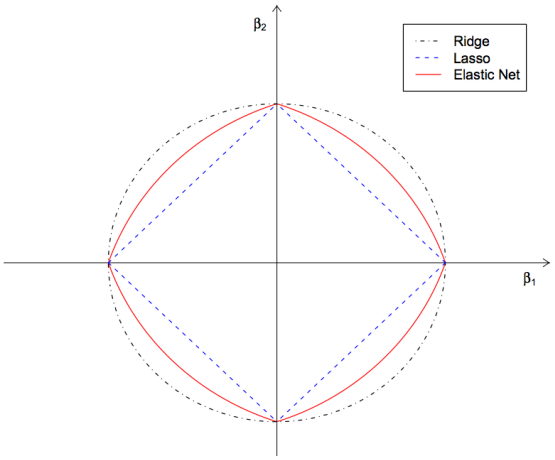

FIGURE 3.3: Geometric Interpretation of the Elastic Net Penalty in 2-Dimensional Space.

(3.3) shows 2-dimensional contour plots of the geometry of the elastic net penalty. The out most contour shows the shape of the ridge penalty, the diamond shaped contour is of the Lasso penalty. The red solid contour is for the elastic net penalty with $\alpha = 0.5$.. We see singularities at the vertexes and the edges are strictly convex. The strength of convexity varies with α.

We can develop a method to solve the naive elastic net problem in an adequate way. We find that the solution is equivalent to a Lasso type optimization problem. This also means that the naive elastic net also shares the computational advantage that the Lasso method has.(R. J. Tibshirani et al. 2013)

Lemma 6. *Given data set* $(\boldsymbol{y}, \boldsymbol{X})$ *and* (λ_1, λ_2) *define an artificial data set* $(\boldsymbol{y}^*, \boldsymbol{X}^*)$ *by*

$$\boldsymbol{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-\frac{1}{2}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I}, \end{pmatrix} \quad \boldsymbol{y}^*_{(n+p)} = \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix}. \tag{3.91}$$

Let $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$ *and* $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$. *Then the naive elastic net criterion can be written as:*

$$L(\gamma, \boldsymbol{\beta}) = L(\gamma, \boldsymbol{\beta}^*) = |\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*|^2 + \gamma\,|\boldsymbol{\beta}^*|_1 \tag{3.92}$$

Let

$$\hat{\boldsymbol{\beta}}^{E*} = \underset{\boldsymbol{\beta}^*}{arg\ min}\ L(\gamma, \boldsymbol{\beta}^*), \tag{3.93}$$

then

$$\hat{\boldsymbol{\beta}}^{E} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^{E*}. \tag{3.94}$$

Lemma 6 states that we can transform the naive elastic net problem to a Lasso problem on augmented data. We note that the sample size is $n+p$ and that $\boldsymbol{X}^*$ has rank p in the augmented data set. This means that the naive elastic net can select all p predictors in all situations. This overcomes the limitation for the Lasso we have in scenario a). Lemma 1 also shows that the naive elastic net can perform an automatic variable selection in a similar way to the Lasso. Next we will show how the naive elastic net has the ability of selecting grouped variables, a property that the Lasso does not have.

In the $p >> n$ problem (West et al.) (MikeWest et al. 2001) the grouped variables scenario is a particularly important concern. Another careful study by Segal and Conklin (Segal, Dahlquist, and Conklin 2003) gives a strong motivation for the use of regularized regression procedure to find the grouped genes. We consider the following generic penalization method:

$$\hat{\boldsymbol{\beta}}^{E} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|^2 + \lambda J(\boldsymbol{\beta}) \tag{3.95}$$

Where J(.) is positive valued for $\beta \neq 0$. A regression method exhibits the grouping effect if the coefficients of a group of highly correlated variables tend to be equal up to a sign of change if negatively correlated. In extreme situations, where some variables might be exactly identical, the regression method used should assign identical coefficients to these identical variables.

Lemma 7. *Assume* $x_i = x_j,\ i,j \in \{1,\cdots,p\}$

1. *If* $J(\cdot)$ *is strictly convex, then* $\hat{\beta}_i^E = \hat{\beta}_j^E\ \forall\ \lambda > 0$
2. *If* $J(\boldsymbol{\beta}) = |\beta|_1$, *then* $\hat{\beta}_i^E\hat{\beta}_j^E \geq 0$ *and* $\hat{\boldsymbol{\beta}}^{E*}$ *is another minimizer of* (3.95)

for any $s \in [0,1]$.

$$\hat{\beta}_k^{E*} = \begin{cases} \hat{\beta}_k^E & \text{if } k \neq i \text{ and } k \neq j \\ \left(\hat{\beta}_i^E + \hat{\beta}_j^E\right) \cdot (s) & \text{if } k = i \\ \left(\hat{\beta}_i^E + \hat{\beta}_j^E\right) \cdot (1-s) & \text{if } k = j \end{cases}$$

We can see that in Lemma 2 that there is a clear distinction between the strictly convex penalty function in 1. and the Lasso penalty shown in 2. The strict convexity property in 1. guarantees the grouping effect in the extreme case where the predictors are identical. We can see in 2. that the Lasso does not even have a unique solution.The elastic net penalty with $\lambda_2 > 0$ is strictly convex, thus enjoying the property in assertion (1).

Theorem 7. *Given data (**y**,**X**) and parameters* (λ_1, λ_2) *the response* $\boldsymbol{y}$ *is centered and the predictors* $\boldsymbol{X}$ *are standardized. Let* $\hat{\boldsymbol{\beta}}^{\boldsymbol{E}}(\lambda_1, \lambda_2)$ *be the naive elastic net estimate. Suppose* $\hat{\beta}_i^E(\lambda_1, \lambda_2)\hat{\beta}_j^E(\lambda_1, \lambda_2) > 0$. *Define:*

$$D_{\lambda_1,\lambda_2}(i,j) = \frac{1}{|y|_1}\left|\hat{\beta}_i^E(\lambda_1, \lambda_2) - \hat{\beta}_j^E(\lambda_1, \lambda_2)\right|, \tag{3.96}$$

then

$$D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)}, \text{ where } \rho = x_i^\top x_j, \text{ the sample correlation.} \tag{3.97}$$

The unit-less quantity $D_{\lambda_1,\lambda_2}(i,j)$ shows the difference between the coefficients of predictors i and j. If x_i and x_j are highly correlated then $\rho = 1$ if we have $\rho = -1$ then we consider $-x_j$ so theorem 1 states that the difference between the coefficient paths of predictor i and predictor j is almost 0. Where the upper bound of the inequality gives a qualitative description for the grouping effect of the naive elastic net.

We have talked about Bridge regression in an earlier section. We also have $J(\boldsymbol{\beta}) = |\boldsymbol{\beta}|^q$ in (3.95), which is a generalization of both the Lasso ($q = 1$) and ridge ($q = 2$). The bridge estimator can be viewed as the Bayes posterior mode under the prior:

$$P_{\lambda q}(\boldsymbol{\beta}) = C(\lambda, q)\exp\left(-\lambda|\boldsymbol{\beta}|^q\right) \tag{3.98}$$

Ridge regression ($q = 2$) corresponds to a Gaussian prior and the Lasso ($q = 1$) a Laplacian (or double exponential) prior. The elastic net penalty corresponds to a new prior given by:

$$P_{\lambda,\alpha}(\boldsymbol{\beta}) = C(\lambda, \alpha)\exp\left(-\lambda(|\boldsymbol{\beta}|^2 + (1-\alpha)|\boldsymbol{\beta}|_1)\right) \tag{3.99}$$

A compromise between the Gaussian and Laplacian priors. Although bridge with $1 < q < 2$ will have many similarities with the elastic net, there is a fundamental difference between them. The elastic nets produces sparse solutions, while the bridge does not. Fan & Li (Fan and Li 2001) prove that in the Lq ($q \geq 1$) penalty family, only the Lasso penalty ($q = 1$) can produce a sparse solution. Bridge ($1 < q < 2$) always keeps all predictors in the model, as does ridge. Since automatic variable selection via penalization is a primary objective of this article, $Lq(1 < q < 2)$ penalization is not a candidate.

Chapter 4

SVD and PCA

4.1 Eigenvector Decomposition (Spectral) and Principal Component Analysis

We will first explore the following five theorems to be able to explain the subsequent methodology.

Theorem 8. *The inverse of an orthogonal matrix is its transpose.*

Proof. Let $\boldsymbol{A}$ be an $m \times n$ orthogonal matrix where $\boldsymbol{a_i}$ is the i^{th} column vector. The ij^{th} element of $\boldsymbol{A}^\top \boldsymbol{A}$ is:

$$(\boldsymbol{A}^\top \boldsymbol{A})_{ij} = \boldsymbol{a_i}^\top \boldsymbol{a_j} = \begin{cases} 1 & if\ i = j \\ 0 & otherwise \end{cases} \tag{4.1}$$

Therefore, since $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$ it follows that $\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$. □

Theorem 9. *If A is any matrix then matrices $A^\top A$ and $AA^\top$ are both symmetric.*

Proof.

$$(\boldsymbol{A}\boldsymbol{A}^\top)^\top = \boldsymbol{A}^{\top\top}\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top \tag{4.2}$$

$$(\boldsymbol{A}^\top \boldsymbol{A})^\top = \boldsymbol{A}^\top \boldsymbol{A}^{\top\top} = \boldsymbol{A}^\top \boldsymbol{A} \tag{4.3}$$

□

Theorem 10. *A matrix is symmetric iff it is orthogonally diagonalizable.*

Proof. This statement is two directional, so it requires a two-part iff proof. We start with the forward case; if $\boldsymbol{A}$ is orthogonally diagonalizable then $\boldsymbol{A}$ is a symmetric matrix. Orthogonally diagonalizable means that there exists some $\boldsymbol{E}$ s.t. $\boldsymbol{A} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^\top$, where $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{E}$ is some special matrix which diagonalizes $\boldsymbol{A}$. Let use compute $\boldsymbol{A}^\top$.

$$\boldsymbol{A}^\top = (\boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^\top)^\top = \boldsymbol{E}^{\top\top}\boldsymbol{D}^\top \boldsymbol{E}^\top = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^\top = \boldsymbol{A} \tag{4.4}$$

If $\boldsymbol{A}$ is orthogonally diagonalizable, it must also be symmetric. The reverse case is more involved and less clean so is omitted from this paper. The forward case is suggestive and somewhat convincing. □

Theorem 11. *A symmetric matrix is diagonalized by a matrix of its orthonormal eigenvectors.*

Proof. Let $\boldsymbol{A}$ be a square $n \times n$ symmetric matrix with associated eigenvectors $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_n}\}$. We let $\boldsymbol{E} = [\boldsymbol{e_1}\ \boldsymbol{e_2}\ \cdots\ \boldsymbol{e_n}]$ where the i^{th} column of $\boldsymbol{E}$ is the eigenvector $\boldsymbol{e_i}$. This theorem asserts that there exists a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{EDE}^\top$. The proof is composed into two parts. In the first part, we find that any matrix can be orthogonally diagonalized iff that matrix eigenvectors are all linearly independent. In the second part of the proof, we see that a symmetric matrix has the unique property that all of its eigenvectors are not just linearly independent but also orthogonal which completes our proof.

We now start the first part of the proof. Let matrix $\boldsymbol{A}$ be any matrix not necessarily symmetric and let it have independent eigenvectors. Moreover, let $\boldsymbol{E} = [\boldsymbol{e_1}\ \boldsymbol{e_2}\ \cdots\ \boldsymbol{e_n}]$ be the matrix of eigenvectors placed in the columns. We let $\boldsymbol{D}$ be a diagonal matrix where the i^{th} eigenvalue is placed in the ii^{th} position. We now show that $\boldsymbol{AE} = \boldsymbol{ED}$. We examine the columns of the right hand and left hand sides of the equation which are:

$$\text{Left hand side: } \boldsymbol{AE} = \begin{bmatrix} \boldsymbol{Ae_1} & \boldsymbol{Ae_2} & \cdots & \boldsymbol{Ae_n} \end{bmatrix} \tag{4.5}$$

$$\text{Right hand side: } \boldsymbol{ED} = \begin{bmatrix} \boldsymbol{\lambda e_1} & \boldsymbol{\lambda e_2} & \cdots & \boldsymbol{\lambda e_n} \end{bmatrix} \tag{4.6}$$

We find that if $\boldsymbol{AE} = \boldsymbol{ED}$ then $\boldsymbol{Ae_i} = \lambda_i e_i$ for all i. This equation is the definition of the eigenvalue equation. So it must be that $\boldsymbol{AE} = \boldsymbol{ED}$. If we rearrange the equation a little, we find that $\boldsymbol{A} = \boldsymbol{EDE}^{-1}$ which completes the first part the proof. We now show the second part of the proof. Essentially we show that a symmetric matrix always has orthogonal eigenvectors $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$.

$$\begin{aligned} \boldsymbol{\lambda_1 e_1 \cdot e_2} &= (\boldsymbol{\lambda_1 e_1})^\top \boldsymbol{e_2} \\ &= (\boldsymbol{Ae_1})^\top \boldsymbol{e_2} \\ &= \boldsymbol{e_1}^\top \boldsymbol{A}^\top \boldsymbol{e_2} \\ &= \boldsymbol{e_1}^\top (\boldsymbol{\lambda_2 e_2}) \\ \boldsymbol{\lambda_1 e_1 \cdot e_2} &= \boldsymbol{\lambda_2 e_1 \cdot e_2} \end{aligned} \tag{4.7}$$

We can use the last relation to show the equality $(\boldsymbol{\lambda_1} - \boldsymbol{\lambda_2})\boldsymbol{e_1} \cdot \boldsymbol{e_2} = 0$. Since we hypothesized that the eigenvalues are in fact unique, it must be the case that $\boldsymbol{e_1} \cdot \boldsymbol{e_2} = 0$. Therefore the eigenvectors of a symmetric matrix are orthogonal. If we back up to our original postulate that $\boldsymbol{A}$ is a symmetric matrix. By the second part of the proof we know that the eigenvectors of $\boldsymbol{A}$ are all orthonormal. We choose the eigenvectors to be normalized. This means that $\boldsymbol{E}$ is an orthogonal matrix; hence by theorem 8 $\boldsymbol{E}^\top = \boldsymbol{E}^{-1}$ and we can rewrite the final result.

$$\boldsymbol{A} = \boldsymbol{EDE}^\top \tag{4.8}$$

Thus a symmetric matrix is diagonalized by a matrix of eigenvectors. □

Theorem 12. *For any arbitrary $m \times n$ matrix $\boldsymbol{X}$ the symmetric matrix $\boldsymbol{X}^\top \boldsymbol{X}$ has a set of orthonormal eigenvectors of $\{\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_2, \cdots, \hat{\boldsymbol{v}}_n\}$ and a set of associated eigenvalues $\{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_n\}$. The set of vectors $\{\boldsymbol{X}\hat{v}_1, \boldsymbol{X}\hat{v}_2, \cdots, \boldsymbol{X}\hat{\boldsymbol{v}}_n\}$ then form an orthogonal basis where each vector $\boldsymbol{X}\hat{\boldsymbol{v}}_i$ is of length $\sqrt{\boldsymbol{\lambda}_i}$.*

Proof. The aforementioned properties arise from the dot product of any two vectors from this set.

$$\begin{aligned}
(\boldsymbol{X}\hat{\boldsymbol{v}}_i) \cdot (\boldsymbol{X}\hat{\boldsymbol{v}}_j) &= (\boldsymbol{X}\hat{\boldsymbol{v}}_i)^\top (\boldsymbol{X}\hat{\boldsymbol{v}}_j) \\
&= \hat{\boldsymbol{v}}_i^\top \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{v}}_j \\
&= \hat{\boldsymbol{v}}_i^\top (\boldsymbol{\lambda}_j \hat{\boldsymbol{v}}_j) \\
&= \boldsymbol{\lambda}_j \hat{\boldsymbol{v}}_i \cdot \hat{\boldsymbol{v}}_j \\
(\boldsymbol{X}\hat{\boldsymbol{v}}_i) \cdot (\boldsymbol{X}\hat{\boldsymbol{v}}_j) &= \boldsymbol{\lambda}_j \boldsymbol{\delta}_{ij}
\end{aligned} \tag{4.9}$$

The last relation holds since the set of eigenvectors of $\boldsymbol{X}$ is orthogonal resulting in the kronecker delta. In simpler terms the last relation states:

$$(\boldsymbol{X}\hat{\boldsymbol{v}}_i) \cdot (\boldsymbol{X}\hat{\boldsymbol{v}}_j) = \begin{cases} \lambda_j & i = j \\ 0 & i \neq j \end{cases} \tag{4.10}$$

(4.10) states that any two vectors in the set are orthogonal. The second property arises from (4.10) by realizing that the length squared of each vector is defined as:

$$\|\boldsymbol{X}\hat{\boldsymbol{v}}_i\|^2 = (\boldsymbol{X}\hat{\boldsymbol{v}}_i) \cdot (\boldsymbol{X}\hat{\boldsymbol{v}}_i) = \boldsymbol{\lambda}_i \tag{4.11}$$

□

Now let $\boldsymbol{X}$ be a $m \times n$ matrix. We also define another matrix $\boldsymbol{Y}$ that is of dimension $(m \times n)$ that is formed by a data transformation $\boldsymbol{P}$.

$$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{Y} \tag{4.12}$$

We also define the following quantities:

- $\boldsymbol{p}_i$ are the rows of $\boldsymbol{P}$
- $\boldsymbol{x}_i$ are the columns of $\boldsymbol{X}$.
- $\tilde{\boldsymbol{s}}$ are the columns of $\tilde{\boldsymbol{S}}$.

(4.12) is a representation of a change of basis and can have one or more of the following interpretations:

- $\boldsymbol{P}$ is a matrix that transforms $\boldsymbol{X}$ into $\tilde{\boldsymbol{S}}$
- Geometrically $\boldsymbol{P}$ is a stretch and rotation which transforms $\boldsymbol{X}$ into $\tilde{\boldsymbol{S}}$
- The rows of $\boldsymbol{P}$, $\{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_m\}$ are a set of new basis vectors for expressing the columns of $\boldsymbol{X}$.

$$\boldsymbol{PX} = \begin{pmatrix} \boldsymbol{P}x_1 & \boldsymbol{P}x_2 & \cdots & \boldsymbol{P}x_n \end{pmatrix} = \begin{pmatrix} p_1 \cdot x_1 & p_1 \cdot x_2 & \cdots & p_1 \cdot x_n \\ p_2 \cdot x_1 & p_2 \cdot x_2 & \cdots & p_2 \cdot x_n \\ \vdots & \vdots & \ddots & \vdots \\ p_m \cdot x_1 & p_m \cdot x_2 & \cdots & p_m \cdot x_n \end{pmatrix} = \boldsymbol{\tilde{S}} \tag{4.13}$$

$$\tilde{\boldsymbol{s}} = \begin{bmatrix} \boldsymbol{p_1} \cdot \boldsymbol{x_i} \\ \vdots \\ \boldsymbol{p_m} \cdots \boldsymbol{x_i} \end{bmatrix} \tag{4.14}$$

We note that $p_i, x_j \in \mathbb{R}^m$. We also notice each coefficient of $\tilde{\boldsymbol{s}}$ is a dot product of $\boldsymbol{x_i}$ with the corresponding row in $\boldsymbol{P}$. Moreover, the j^{th} coefficient of $\tilde{\boldsymbol{s}}$ is a projection on to the basis of $\{\boldsymbol{p_1}, \cdots, \boldsymbol{p_m}\}$. Therefore, the rows of $\boldsymbol{P}$ are a new set of basis vectors for representing of columns of $\boldsymbol{X}$. To be able to extract information about the signal regardless of what analysis technique is used, the measurement noise in the data set must be low. All noise is quantified relative to the signal strength as there exists no absolute scale for noise. A common measure is the signal to noise ratio (SNR) or a ratio of variances σ^2.

$$SNR = \frac{\boldsymbol{\sigma^2_{signal}}}{\boldsymbol{\sigma^2_{noise}}} \tag{4.15}$$

A high SNR ($>> 1$) indicates high precision data while a low SNR indicates noise contaminated data. In a regression model, it is simple to identify redundant cases if we only have two variables. We would simply find the slope of the line of best fit and we subsequently judge the quality of the fit. We can generalize these notions to higher dimensions by using the covariance matrix. First consider the following two sets of measurements with zero means:

$$\boldsymbol{A} = \{a_1, a_2, \cdots, a_n\}, \quad \boldsymbol{B} = \{b_1, b_2, \cdots, b_n\} \tag{4.16}$$

where the subscript denotes the sample number. The variance of A and B are defined individually as:

$$\boldsymbol{\sigma^2_A} = \frac{1}{n}\sum_{\boldsymbol{i}} a_i^2, \quad \boldsymbol{\sigma^2_B} = \frac{1}{n}\sum_{\boldsymbol{i}} b_i^2 \tag{4.17}$$

The covariance between A and B is a straight forward generalization.

$$Covariance\ of\ A\ and\ B \equiv \boldsymbol{\sigma^2_{AB}} = \frac{1}{n}\sum_{\boldsymbol{i}} a_i b_i \tag{4.18}$$

The covariance measures the degree of the linear relationship between two variables. A large positive value indicates positively correlated data. Similarly, a large negative value denotes negatively correlated data. The absolute magnitude of the covariance measures the

degree of redundancy. Furthermore, additional facts that we may wish to consider are:

- $\boldsymbol{\sigma^2_{AB}} \geq 0$ (non negative). $\boldsymbol{\sigma_{AB}}$ is zero iff $\boldsymbol{A}$ and $\boldsymbol{B}$ are entirely uncorrelated.
- $\boldsymbol{\sigma^2_{AB}} = \boldsymbol{\sigma^2_A}$ if $\boldsymbol{A} = \boldsymbol{B}$.

$$\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \tag{4.19}$$

$$\boldsymbol{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \tag{4.20}$$

so that we can express the covariance as a dot product matrix:

$$\boldsymbol{\sigma^2_{ab}} \equiv \frac{1}{n}\boldsymbol{a}\boldsymbol{b}^\top \tag{4.21}$$

where $\frac{1}{n}$ is a constant for normalization. We can now generalize from two vectors to an arbitrary number. We rename the row vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ as $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ respectively and consider additional indexed row vectors $\boldsymbol{x_3} \cdots \boldsymbol{x_m}$. Define a new $m \times n$ matrix $\boldsymbol{X}$.

$$\boldsymbol{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \in \mathbb{R}^{m \times n}, \qquad x_i^\top \in \mathbb{R}^n \tag{4.22}$$

We can interpret $\boldsymbol{X}$ as follows. Each row of $\boldsymbol{X}$ corresponds to all measurements of a particular type. Each column of $\boldsymbol{X}$ corresponds to a set of measurements from one particular trial. We now define the covariance matrix $\boldsymbol{C_X}$:

$$\boldsymbol{C_X} \equiv \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top \tag{4.23}$$

If we consider the matrix $\boldsymbol{C_X} \equiv \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$. The ij^{th} element of $\boldsymbol{C_X}$ is the dot product between the vector of the i^{th} measurement type with the vector of the j^{th} measurement type.

i.e.

$$\boldsymbol{C}_X = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top = \frac{1}{n}\begin{pmatrix} x_1 x_1^\top & x_1 x_2^\top & \cdots & x_1 x_m^\top \\ x_1 x_1^\top & x_1 x_2^\top & \cdots & x_1 x_m^\top \\ \vdots & \vdots & \ddots & \vdots \\ x_m x_1^\top & x_m x_2^\top & \cdots & x_m x_m^\top \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{4.24}$$

We can summarize several properties of $\boldsymbol{C_X}$:

- $\boldsymbol{C_X}$ is a square symmetric $m \times m$ matrix.
- The diagonal terms of $\boldsymbol{C_X}$ are the variance of particular measurement types.
- The off-diagonal terms of $\boldsymbol{C_X}$ are the covariance between measurement types.

  $\boldsymbol{C_X}$ captures the correlations between all possible pairs of measurements. The correlation values reflect the noise and redundancy in our measurements.
- In the diagonal terms by assumption large values correspond to interesting dynamics (or noise)
- In the off-diagonal terms large values correspond to high redundancy.

So our goals are to first minimize redundancy, measured by covariance and then maximize the signal measured by variance. It should be noted that the $\boldsymbol{C_{\tilde{S}}}$ matrix must be diagonal and all off-diagonal terms in $\boldsymbol{C_{\tilde{S}}}$ would be zero or in other words $\tilde{\boldsymbol{S}}$ would be decorrelated. Furthermore, each successive dimension in $\tilde{\boldsymbol{S}}$would be rank ordered according to variance. There are many methods for diagonalizing $\boldsymbol{C_{\tilde{S}}}$, PCA arguably selects the easiest method. PCA assumes that all basis vectors $\{\boldsymbol{p_1}, \cdots, \boldsymbol{p_m}\}$ are orthonormal (*i.e.* $\boldsymbol{p_i} \cdot \boldsymbol{p_j} = \boldsymbol{\delta}_{ij}$). In the language of linear algebra, PCA assumes the directions with the largest variances. In other words they are the most principal. The following simple algorithm shows us how PCA works for multiple dimensions.

1. Select a normalized direction in m dimensional space along which the variance in $\boldsymbol{X}$ is maximized. Save this vector as $\boldsymbol{p_1}$.
2. We then find another direction along which variance is maximized. The orthonormality condition we first restrict the search to all directions perpindicular to all previous selected directions. Save this vector as $\boldsymbol{p_i}$.
3. Repeat this procedure until m vectors are selected.

   The resulting ordered set of $\boldsymbol{p}'s$ are the principal components.

The following show when PCA might perform poorly:

- **Linearity**.

  The linearity assumption shapes the problem as a change of basis. Extensive research have focused on extending these notions to nonlinear regimes.
- **Large variances have important structure**.

  This assumption also encompasses the belief that the data has a high SNR. Moreover, principal components with larger associated variances represent interesting structure while those with lower variances represent noise.
- **The principal components are orthogonal**.

  This assumption provides an intuitive simplification that makes PCA feasible with linear algebra decomposition techniques. This is highlighted in what follows in our

study.

The above show all aspects of deriving PCA. What remains is the linear algebra solutions. The first solution is somewhat straightforward while the second solution involves understanding an important algebraic decomposition. We now derive our first algebraic solution to PCA using linear algebra. This solution is based on an important property of eigenvector decomposition. Once again, the data set is $\boldsymbol{X}$ an $m \times n$ matrix where m is the number of measurement types and n is the number of samples. The first solution is somewhat straight forward while the second solution involves understanding an important algebraic decomposition. We derive our first algebraic solution to PCA using linear algebra. The solution is based on an essential property of eigenvector decompostion. We use the data set $\boldsymbol{X}$ again where it is an $m \times n$ matrix where m is the number of measurement types and n is the number of samples. The process is summarized as follows:

We start by finding some orthonormal matrix $\boldsymbol{P}$ where $\boldsymbol{Y} = \boldsymbol{PX}$ s.t. $\boldsymbol{C}_{\tilde{\boldsymbol{S}} \equiv \frac{1}{n}\tilde{\boldsymbol{S}}^\top}$ is diagonalized. The rows of $\boldsymbol{P}$ are the principal components of $\boldsymbol{X}$. We now begin by rewriting $\boldsymbol{C}_{\tilde{\boldsymbol{S}}}$ in terms of our variable of choice $\boldsymbol{P}$.

$$\begin{aligned}
\boldsymbol{C}_{\tilde{\boldsymbol{S}}} &= \frac{1}{n}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^\top \\
&= \frac{1}{n}(\boldsymbol{PX})(\boldsymbol{PX})^\top \\
&= \frac{1}{n}\boldsymbol{PXX}^\top\boldsymbol{P}^\top \\
&= \boldsymbol{P}(\frac{1}{n}\boldsymbol{XX}^\top)\boldsymbol{P}^\top \\
\boldsymbol{C}_{\tilde{\boldsymbol{S}}} &= \boldsymbol{PC}_{\boldsymbol{X}}\boldsymbol{P}^\top
\end{aligned} \tag{4.25}$$

For a symmetric matrix $\boldsymbol{A}$ theorem 11:

$$\boldsymbol{A} = \boldsymbol{EDE}^\top \tag{4.26}$$

Where $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{E}$ is a matrix of eigenvectors of $\boldsymbol{A}$ arranged as columns. The matrix $\boldsymbol{A}$ has $r \leq m$ orthonormal eigenvectors where r is the rank of the matrix. The rank of $\boldsymbol{A}$ is less than m when $\boldsymbol{A}$ is degenerate or all data occupy a subspace of dimension $r \leq m$. Maintaining the constraint of orthogonality, we can remedy this situation by selecting $(m - r)$ additional orthonormal vectors to fill out the matrix $\boldsymbol{E}$. These additional vectors do not effect the final solution because the variances associated with these directions are zero. We select the matrix $\boldsymbol{P}$ to be a matrix where each row $\boldsymbol{p_i}$ is an eigenvector of $\frac{1}{n}\boldsymbol{XX}^\top$. With this relation and Theorem 8 ($\boldsymbol{P}^{-1} = \boldsymbol{P}^\top$) we can finalize the evaluation of $\boldsymbol{C}_{\tilde{\boldsymbol{S}}}$.

$$\begin{aligned}
\boldsymbol{C}_{\tilde{\boldsymbol{S}}} &= \boldsymbol{PC}_{\boldsymbol{X}}\boldsymbol{P}^\top \\
&= \boldsymbol{P}\left(\boldsymbol{E}^\top\boldsymbol{DE}\right)\boldsymbol{P}^\top \\
&= \boldsymbol{P}\left(\boldsymbol{P}^\top\boldsymbol{DP}\right)\boldsymbol{P}^\top \\
&= (\boldsymbol{PP}^\top)\boldsymbol{D}(\boldsymbol{PP}^\top) \\
&= (\boldsymbol{PP}^{-1})\boldsymbol{D}(\boldsymbol{PP}^{-1}) \\
\boldsymbol{C}_{\tilde{\boldsymbol{S}}} &= \boldsymbol{D}
\end{aligned} \tag{4.27}$$

It is evident that the choice of $\boldsymbol{P}$ diagonalizes $\boldsymbol{C}_{\tilde{\boldsymbol{S}}}$. This was the goal for PCA. We can summarize the results of PCA in the matrices $\boldsymbol{P}$ and $\boldsymbol{C}_{\boldsymbol{Y}}$.

- The principal components of $\boldsymbol{X}$ are the eigenvectors of $\boldsymbol{C}_{\boldsymbol{X}} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$.
- The i^{th} diagonal value of $\boldsymbol{C}_{\boldsymbol{Y}}$ is the variance of $\boldsymbol{X}$ along $\boldsymbol{p}_{\boldsymbol{i}}$.

When computing PCA of a data set $\boldsymbol{X}$ we need to do two things, first we need to subtract off the mean of each measurement type and then we need to compute the eigenvectors of $\boldsymbol{C}_{\boldsymbol{X}}$. Let $\boldsymbol{X}$ be an arbitrary $n \times m$ matrix and $\boldsymbol{X}^\top\boldsymbol{X}$ be a rank r, square symmetric $m \times m$ matrix. In a seemingly unmotivated fashion, we now shall define all the quantities of interest.

4.2 Singular Value Decomposition and Principal Component Analysis

We now derive another algebraic solution for PCA and in the process find out that PCA is closely related to singular value decomposition (SVD). The two are so intimately related that the names are often used interchangeably. An essential concept that we shall see is that SVD is a more general method of understanding change of basis.

- $\{\hat{\boldsymbol{v}}_{\boldsymbol{1}}, \hat{\boldsymbol{v}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{v}}_{\boldsymbol{r}}\}$ is the set of orthonormal $m \times 1$ eigenvectors with associated eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_r\}$ for the symmetric matrix $\boldsymbol{X}^\top\boldsymbol{X}$.

$$\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\hat{\boldsymbol{v}}_{\boldsymbol{i}} = \lambda_i\hat{\boldsymbol{v}}_{\boldsymbol{i}}. \tag{4.28}$$

- $\sigma_i \equiv \sqrt{\lambda_i}$ are positive real and termed the singular values.
- $\{\hat{\boldsymbol{u}}_{\boldsymbol{1}}, \hat{\boldsymbol{u}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{u}}_{\boldsymbol{r}}\}$ is the set of orthonormal $n \times 1$ vectors defined by $\hat{\boldsymbol{u}}_{\boldsymbol{i}} \equiv \frac{1}{\sigma_1}\boldsymbol{X}\hat{\boldsymbol{v}}_{\boldsymbol{i}}$.

We obtain the final definition by referring to theorem 12 which includes two new and unexpected properties.

- $$\hat{\boldsymbol{u}}_{\boldsymbol{i}} \cdot \hat{\boldsymbol{u}}_{\boldsymbol{j}} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{4.29}$$

- $$\|\boldsymbol{X}\hat{\boldsymbol{v}}_{\boldsymbol{i}}\| = \boldsymbol{\sigma}_{\boldsymbol{i}} \tag{4.30}$$

These properties are both proven in Theorem 12. We now have all of the pieces to construct the decomposition. The scalar version of singular value decomposition is just a restatement of the third definition.

$$\boldsymbol{X}\hat{\boldsymbol{v}}_{\boldsymbol{i}} = \boldsymbol{\sigma}_{\boldsymbol{i}}\hat{\boldsymbol{u}}_{\boldsymbol{i}} \tag{4.31}$$

This result gives us a lot of information. $\boldsymbol{X}$ multiplied by an eigenvector of $\boldsymbol{X}^\top\boldsymbol{X}$ is the same as a scalar times another vector. The set of eigenvectors $\{\hat{\boldsymbol{v}}_{\boldsymbol{1}}, \hat{\boldsymbol{v}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{v}}_{\boldsymbol{r}}\}$ and the set of vectors $\{\hat{\boldsymbol{u}}_{\boldsymbol{1}}, \hat{\boldsymbol{u}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{u}}_{\boldsymbol{r}}\}$ are both orthonormal sets or bases in r dimensional space. We also construct accompanying orthogonal matrices $\boldsymbol{V}$ and $\boldsymbol{U}$.

$$\boldsymbol{V} = \begin{bmatrix} \hat{\boldsymbol{v}}_{\tilde{1}} & \hat{\boldsymbol{v}}_{\tilde{2}} & \cdots & \hat{\boldsymbol{v}}_{\tilde{m}} \end{bmatrix} \tag{4.32}$$

$$\boldsymbol{U} = \begin{bmatrix} \hat{\boldsymbol{u}}_{\tilde{1}} & \hat{\boldsymbol{u}}_{\tilde{2}} & \cdots & \hat{\boldsymbol{u}}_{\tilde{m}} \end{bmatrix} \tag{4.33}$$

Where we added an additional $(m-r)$ and $(n-r)$ orthonormal vectors to "fill up" the matrices for $\boldsymbol{V}$ and $\boldsymbol{U}$ respectively.

$$\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \tag{4.34}$$

Where each column of $\boldsymbol{V}$ and $\boldsymbol{U}$ perform the value version of the decomposition. Because $\boldsymbol{V}$ is orthogonal we can multiply both sides by $\boldsymbol{V}^{-1} = \boldsymbol{V}^\top$ to arrive at the final form of the decomposition.

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \tag{4.35}$$

This decomposition is very powerful. (4.35) states that any arbitrary matrix $\boldsymbol{X}$ can be converted to an orthogonal matrix, a diagonal matrix and another orthogonal matrix. We now study (4.35) in more depth. The final form of SVD is a concise and thick statement. If we interpret (4.31) as:

$$\boldsymbol{X}\boldsymbol{a} = k\boldsymbol{b} \tag{4.36}$$

Where $\boldsymbol{a}$ and $\boldsymbol{b}$ are column vectors and k is a scalar constant. The set $\{\hat{\boldsymbol{v}}_{\boldsymbol{1}}, \hat{\boldsymbol{v}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{v}}_{\boldsymbol{m}}\}$ is analogous to $\boldsymbol{a}$ and the set $\{\hat{\boldsymbol{u}}_{\boldsymbol{1}}, \hat{\boldsymbol{u}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{u}}_{\boldsymbol{n}}\}$ is analogous to $\boldsymbol{b}$. The unique thing is $\{\hat{\boldsymbol{v}}_{\boldsymbol{1}}, \hat{\boldsymbol{v}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{v}}_{\boldsymbol{m}}\}$ and $\{\hat{\boldsymbol{u}}_{\boldsymbol{1}}, \hat{\boldsymbol{u}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{u}}_{\boldsymbol{n}}\}$ are orthonormal sets of vectors which span an m or n dimensional space respectively. Loosely speaking, these sets appear to span all possible inputs ($\boldsymbol{a}$) and outputs ($\boldsymbol{b}$). We can manipulate equation (4.35) to make this hypothesis more precise:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \\ \boldsymbol{U}^\top\boldsymbol{X} &= \boldsymbol{\Sigma}\boldsymbol{V}^\top \\ \boldsymbol{U}^\top\boldsymbol{X} &= \boldsymbol{Z} \end{aligned} \tag{4.37}$$

Where we have defined $\boldsymbol{Z} \equiv \boldsymbol{\Sigma}\boldsymbol{V}^\top$. Note that the previous columns $\{\hat{\boldsymbol{u}}_{\boldsymbol{1}}, \hat{\boldsymbol{u}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{u}}_{\boldsymbol{n}}\}$ are rows in $\boldsymbol{U}^\top$. Comparing this to (4.12) $\{\hat{\boldsymbol{u}}_{\boldsymbol{1}}, \hat{\boldsymbol{u}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{u}}_{\boldsymbol{n}}\}$ perform the same role as $\{\hat{\boldsymbol{p}}_{\boldsymbol{1}}, \hat{\boldsymbol{p}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{p}}_{\boldsymbol{m}}\}$. Thus $\boldsymbol{U}^\top$ is a change of basis from $\boldsymbol{X}$ to $\boldsymbol{Z}$. This is the same process as before where we were transforming column vectors. Since the orthonormal basis $\boldsymbol{U}^\top$ transforms column vectors means that $\boldsymbol{U}^\top$ is a basis that spans the columns of $\boldsymbol{X}$. The basis that span the columns are termed the column space of $\boldsymbol{X}$. The column space formalizes the notion of

what are the possible outputs of any matrix.

$$\begin{aligned} \boldsymbol{X}\boldsymbol{V} &= \boldsymbol{\Sigma}\boldsymbol{U} \\ (\boldsymbol{X}\boldsymbol{V})^\top &= (\boldsymbol{\Sigma}\boldsymbol{U})^\top \\ \boldsymbol{V}^\top\boldsymbol{X}^\top &= \boldsymbol{U}^\top\boldsymbol{\Sigma} \\ \boldsymbol{V}^\top\boldsymbol{X}^\top &= \boldsymbol{Z} \end{aligned} \tag{4.38}$$

Where we have defined $\boldsymbol{Z} \equiv \boldsymbol{U}^\top\boldsymbol{\Sigma}$. Again the rows/columns of $\boldsymbol{V}^\top$ are an orthonormal basis for transforming $\boldsymbol{X}^\top$ into $\boldsymbol{Z}$. Because of the transpose on $\boldsymbol{X}$, it follows that $\boldsymbol{V}$ is an orthonormal basis spanning the row space of $\boldsymbol{X}$. The row space forms the notion of what are the possible inputs into an arbitrary matrix. It is clear that PCA and SVD are very related. If we return to the original $m \times n$ data matrix $\boldsymbol{X}$. Then we can define a new matrix $\tilde{\boldsymbol{S}}$ as an $n \times m$ matrix.

$$\tilde{\boldsymbol{S}} \equiv \frac{1}{\sqrt{n}}\boldsymbol{X}^\top \tag{4.39}$$

Where each column of $\tilde{\boldsymbol{S}}$ has zero mean. The definition of $\tilde{\boldsymbol{S}}$ becomes clear by analyzing $\tilde{\boldsymbol{S}}^\top\tilde{\boldsymbol{S}}$

$$\begin{aligned} \tilde{\boldsymbol{S}}^\top\tilde{\boldsymbol{S}} &= (\frac{1}{\sqrt{n}}\boldsymbol{X}^\top)^\top(\frac{1}{\sqrt{n}}\boldsymbol{X}^\top) \\ &= \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top \\ \tilde{\boldsymbol{S}}^\top\tilde{\boldsymbol{S}} &= \boldsymbol{C}_{\boldsymbol{X}} \end{aligned} \tag{4.40}$$

By construction $\tilde{\boldsymbol{S}}^\top\tilde{\boldsymbol{S}}$ is equal to the covariance matrix of $\boldsymbol{X}$. We know from earlier results that the principal components of $\boldsymbol{X}$ are the eigenvectors of $\boldsymbol{C}_{\boldsymbol{X}}$. If we calculate the SVD of $\tilde{\boldsymbol{S}}$ the columns of matrix $\boldsymbol{V}$ contain the eigenvectors of $\tilde{\boldsymbol{S}}^\top\tilde{\boldsymbol{S}} = \boldsymbol{C}_{\boldsymbol{X}}$. Therefore the columns of $\boldsymbol{V}$ are the principal components of $\boldsymbol{X}$. This means that $\boldsymbol{V}$ spans the row space of $\tilde{\boldsymbol{S}} \equiv \frac{1}{\sqrt{n}}\boldsymbol{X}^\top$. Therefore, $\boldsymbol{V}$ must also span the column space of $\frac{1}{\sqrt{n}}\boldsymbol{X}$. We conclude that finding the principal components leads to finding an orthonormal basis that spans the column space of $\boldsymbol{X}$.

The importance of an eigenvector is measured by the percentage of total variance explained by the corresponding eigenvalue. All eigenvectors are arranged according to their eigenvalues in descending order. Now we have to decide how many eigenvectors to retain. We will be discussing two methods: **Total variance explained** and **Scree Plot**. We start with the Total Variance Explained. If we suppose we have a vector of n eigenvalues $(e_0, \cdots, e_n)$ sorted in descending order. We take the cumulative sum of eigenvalues at every index until the sum is greater than 95% of the total variance. We then reject all eigenvalues and eigenvectors after that index. Similarly for the Scree Plot, we have to arrange the eigenvalues in descending order. We plot the eigenvalues against its index. An ideal scree plot is a steep curve which is followed by a sharp bend and a straight line. We reject all the eigenvalues after the sharp bend and their corresponding eigenvectors.

4.3 Image and Data Compression (Sparse Face Recognition) Utilizing Robust Principal Component Analysis

In the previous subsection, we developed a method for principal component analysis which used the SVD of an $m \times m$ matrix $\boldsymbol{Z}^\top \boldsymbol{Z}$ where $\boldsymbol{Z} = \frac{1}{\sqrt{n}} \boldsymbol{X}^\top$ and $\boldsymbol{X}$ was an $m \times n$ data matrix. Since $\boldsymbol{Z}^\top \boldsymbol{Z} \in \mathbb{R}^{m \times m}$ the matrix $\boldsymbol{V}$ obtained in the singular value decomposition of $\boldsymbol{Z}^\top \boldsymbol{Z}$ must also be of dimensions $m \times m$. We recall also that the columns of $\boldsymbol{V}$ are the principal component directions and that the SVD automatically sorts these components in decreasing order of importance or principality so that the principality component is the first column of $\boldsymbol{V}$. Let's suppose that before we project the data using $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$, we truncate the matrix $\boldsymbol{V}$ so that we kept only the first $r < m$ columns. We would end up with a matrix $\tilde{\boldsymbol{V}} \in \mathbb{R}^{m \times r}$. The projection $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{V}}^\top \boldsymbol{X}$ is still dimensionally consistent and the result of the product is a matrix X and obtain $\tilde{\boldsymbol{X}} \in \mathbb{R}^{m \times n}$.

The matrices, $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ have the same dimensions but they are different matrices. This is because we truncated the matrix of principal components $\boldsymbol{V}$ in order to obtain $\tilde{\boldsymbol{X}}$. It is therefore reasonable to conclude that the matrix $\tilde{\boldsymbol{X}}$ has in some sense less information in it than the matrix $\boldsymbol{X}$. Of course, in terms of memory allocation on a computer, this is certainly not the case since both matrices have the same dimensions and would therefore allotted the same amount of memory. This, together with the fact that the 'important' information in the matrix is captured by the first principal components suggests a possible method for image compression. The matrix $\tilde{\boldsymbol{X}}$ can be computed as the product of two smaller matrices ($\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{Y}}$. This together with the fact that the important information in the matrix is captured by the first principal components suggests a possible method for image compression. We first briefly review regression type optimization framework of PCA. We suppose the data matrix is $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, where n is the number of observations or samples p is the number of features or variables. The singular value decomposition of $\boldsymbol{X}$ is $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ the first k PCs of $\boldsymbol{X}$ are defined as $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{V}_k$, where $\boldsymbol{V}_k \in \mathbb{R}^{p \times k}$ is called a loading matrix or projecting matrix. $\boldsymbol{P} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k] \in \mathbb{R}^{n \times k}$ usually $k << p$. thus dimensionality reduction is achieved. Furthermore, the uncorrelated PCs capture the maximum variability of $\boldsymbol{X}$ which guarantees minimal information loss. In regression type optimization frameworks, PCA can be formulated as a ridge regression problem:

For any $\lambda_1 > 0, j = 1, 2, \cdots, k$, let:

$$\begin{aligned} (\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) &= \underset{\boldsymbol{A},\boldsymbol{B}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{B}^\top \boldsymbol{x}_i \right\|_2^2 + \lambda_1 \left\| \boldsymbol{B} \right\|_2^2 \\ &= \underset{\boldsymbol{A}\boldsymbol{B}}{\operatorname{argmin}} \left\{ \left\| (\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top) \right\|_2^2 + \lambda_1 \left\| \boldsymbol{B} \right\|_2^2 \right\} \\ &\quad \text{s.t.} \quad \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}. \end{aligned} \tag{4.41}$$

Then $\hat{\beta}_j \propto \boldsymbol{v}_j$. Where x_i is the ith column of $\boldsymbol{X}^\top, \boldsymbol{A} = [\alpha_1, \cdots, \alpha_k] \in \mathbb{R}^{p \times k}, \boldsymbol{B} = [\beta_1, \cdots, \beta_k] \in \mathbb{R}^{p \times k}, \boldsymbol{V}_k = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k]$ is the loading matrix of $\boldsymbol{X}$. We now discuss PCA with outliers. We define $\epsilon_i = \boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{B}^\top \boldsymbol{x}_i$ and assume that its elements $e_{i,j} (j = 1, \cdots, p)$ are independently identically distributed $(i.i.d)$ with probability density function $(p.d.f)$ $f(e_{i,j})$ then the likelihood function is $\boldsymbol{L}(e_{i,1}, \cdots, e_{i,p}) = \prod_{j=1}^{p} f(e_{i,j})$. We minimize the objective function by

using maximum likelihood estimation (MLE).

$$-\ln \boldsymbol{L} = \sum_{j=1}^{p} -lnf(e_{ij}) = \sum_{j=1}^{p} \rho(e_{i,j}) = \boldsymbol{F}(\epsilon_i), \tag{4.42}$$
$$\text{Where } \rho(e_{i,j}) = -\ln f(e_{i,j})$$

We approximate $\boldsymbol{F}(\epsilon_i)$ by its first order Taylor series expansion in the neighborhood of ϵ_0 we have:

$$\tilde{\boldsymbol{F}}(\epsilon_i) = \boldsymbol{F}(\epsilon_0) + (\epsilon_i - \epsilon_0)^\top \boldsymbol{F}^{`}(\epsilon_0) + \boldsymbol{R}_1(\epsilon_i) \tag{4.43}$$

Where $\boldsymbol{R}_1(\epsilon_i)$ is the higher order residual term, it can be approximated as:

$$\boldsymbol{R}_1(\epsilon_i) = \frac{1}{2}(\epsilon_i - \epsilon_0)^\top \boldsymbol{\Omega}_i (\epsilon_i - \epsilon_0) \tag{4.44}$$

Where $\boldsymbol{\Omega}_i$ is a diagonal matrix for the elements in ϵ_i that are independent and there is no cross term between $e_{i,j}$ and $e_{i,k}$, $(j \neq k)$. Since $\boldsymbol{F}(\epsilon_i)$ reaches its minimal value at $\epsilon_i = \boldsymbol{0}$, We also require that $\tilde{\boldsymbol{F}}(\epsilon_i)$ has its minimal value at $\epsilon_i = \boldsymbol{0}$. We let $\tilde{\boldsymbol{F}}(\boldsymbol{0}) = \boldsymbol{0}$, we see that we have the diagonal elements of $\boldsymbol{\Omega}_i$ as:

$$\omega_i(e_0, j) = \rho^{'}(e_{0,j}/e_{0,j}). \tag{4.45}$$

Then $\tilde{\boldsymbol{F}}(\epsilon_i)$ can be written as:

$$\tilde{\boldsymbol{F}}(\epsilon_i) = \frac{1}{2}\left\|\boldsymbol{\Omega}_i^{1/2}\epsilon_i\right\|^2 + b \tag{4.46}$$

Where b is a scalar value determined by ϵ_0.

Then (4.41) can be approximated by:

$$\begin{aligned}(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) &= \underset{\boldsymbol{A},\boldsymbol{B}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\|\boldsymbol{\Omega}_i^{1/2}(\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{B}^\top \boldsymbol{x}_i)\right\|_2^2 + \lambda_1 \left\|\boldsymbol{B}\right\|_2^2 \\ &= \underset{\boldsymbol{A},\boldsymbol{B}}{\operatorname{argmin}} \left\{ \left\|\boldsymbol{W}^{1/2} \circ (\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top)^\top \right\|_2^2 + \lambda_1 \left\|\boldsymbol{B}\right\|_2^2 \right\} \\ &s.t.\ \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}\end{aligned} \tag{4.47}$$

Where:

$$\boldsymbol{W} \ = \ \text{unrvec}\Big[\text{diag}(\boldsymbol{\Omega}_1), \cdots, \text{diag}(\boldsymbol{\Omega}_n)\Big] \in \mathbb{R}^{n \times p} \tag{4.48}$$

Where unrvec$(\cdot)$ denotes the matrix of a row vector and "$\circ$" denote Hadamard product. Furthermore diag$(\boldsymbol{\Omega}_i) = [\omega_i(e_{i,1}), \cdots, \omega_i(e_{i,p})]$. The (4.41) be formulated as a weighted ridge regression problem. Now we define the Geman-McClure function:

$$\ominus(e) = \frac{e^2}{\sigma^2 + e^2} \tag{4.49}$$

We now suggest an iterative approach to solve the weighted ridge regression problem. Since the Geman-McClure function has properties similar to the hinge loss function in SVM, we

choose it as the cost function:

$$\rho(e) = \frac{e^2}{(\sigma + e^2)^2} \tag{4.50}$$

For the relation in (4.45) we have:

$$\omega_i(e) = \frac{2\sigma}{(\sigma + e^2)^2} \tag{4.51}$$

Where σ is a positive scalar, which controls the location of demarcation point. Beyond some threshold τ the outlier would be adaptively assigned with low weights to reduce their affects on the regression estimation thus resulting in more robust dimension reduction. The point where the influence of outliers first begins to decrease as the magnitude of the residuals increases from zero occurs when the second partial derivative of the $\rho-$ function is zero. For the Geman-McClure function the second partial derivative is:

$$\frac{\partial^2 \rho}{\partial e^2} = \frac{2\sigma(\sigma - 3e^2)}{(\sigma + e^2)^3} \tag{4.52}$$

Equals zero when $\tau = \pm\sqrt{\sigma/3}$. In other words a residual e is an outlier if $|e| \geq \sqrt{\sigma/3}$.

Theorem 13. *For any $\lambda_1 > 0, j = 1, 2, \cdots, k$, let*

$$(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}) = \underset{\boldsymbol{A},\boldsymbol{B}}{argmin} \left(\left\| \boldsymbol{W}^{1/2} \circ \left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top \right) \right\|_2^2 + \lambda_1 \left\| \boldsymbol{B} \right\| \right) \tag{4.53}$$

Then $\hat{\beta}_j \propto \boldsymbol{v}_j$

Suppose the $\boldsymbol{A}_\perp$ be the orthonormal matrix of $\boldsymbol{A}$ we have:

$$\begin{aligned} \left\| \boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top \right\|_2^2 &= \left\| \boldsymbol{X}\boldsymbol{A}_\perp \right\|_2^2 + \left\| \boldsymbol{X}\boldsymbol{A} - \boldsymbol{X}\boldsymbol{B} \right\|_2^2 \\ &= \left\| \boldsymbol{X}\boldsymbol{A}_\perp \right\|_2^2 + \sum_{j=1}^{k} \left\| \boldsymbol{X}\alpha_j - X\beta_j \right\|_2^2 \end{aligned} \tag{4.54}$$

If $\boldsymbol{A}$ is given then the optimal $\boldsymbol{B}$ minimizing (4.53) can be written as:

$$\hat{\boldsymbol{B}}_{opt} = \arg\min_{\boldsymbol{B}} \sum_{j=1}^{k} \left\{ \left\| \boldsymbol{W}^{1/2} \circ (\boldsymbol{X}\alpha_j - \boldsymbol{X}\beta_j) \right\|_2^2 + \lambda_1 \left\| \beta_j \right\|_2^2 \right\} \tag{4.55}$$

On the other hand if $\boldsymbol{B}$ is fixed we have following theorem.

Theorem 14. *Let $\boldsymbol{M}_{n\times p} = \boldsymbol{W}^{1/2} \circ \boldsymbol{X}$, $\boldsymbol{N}_{n\times k} = \boldsymbol{W}^{1/2} \circ \boldsymbol{X}\boldsymbol{B}$. Consider the constrained minimization problem*

$$\hat{\boldsymbol{A}} = \underset{A}{argmin} \left\| \boldsymbol{M} - \boldsymbol{N}\boldsymbol{A}^\top \right\|_2^2 \quad s.t. \quad \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}. \tag{4.56}$$

Suppose the SVD of $\boldsymbol{M}^\top \boldsymbol{N} = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top$, then $\hat{\boldsymbol{A}} = \boldsymbol{U}_1 \boldsymbol{V}_1^\top$.

If we suppose the SVD of $\boldsymbol{M}^\top \boldsymbol{N} = \boldsymbol{U}_1 \boldsymbol{\Sigma}_1 \boldsymbol{V}_1^\top$, then we have $\hat{\boldsymbol{A}} = \boldsymbol{U}_1 \boldsymbol{V}_1^\top$ In order to initialize the weight we should firstly estimate the residual ϵ_i of x_i. For $\epsilon_i = x_i - AB^\top x_i$ we let $\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{V}_k$ where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}, \boldsymbol{V}_k$ is the first k columns of $\boldsymbol{V}$. The scale parameter σ of ρ, which control the shape of $\rho-$ function and hence determines what residual errors are treated as outliers. We estimated the σ by using median absolute deviation (MED) method which can be viewed as a robust statistical estimation of the standard deviation we compute it as:

$$\sigma = 1.468 \text{ med } (|\boldsymbol{e} - \text{med}(\boldsymbol{e})|) \tag{4.57}$$

Where med indicates the median of a vector $\boldsymbol{e} = \text{vect}(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^\top)$. The convergence is achieved when the difference of $\boldsymbol{B}$ between adjacent iteration is small enough. In this paper we will use the angle between $\boldsymbol{B}^{(n)}$ and $\boldsymbol{B}^{(n+1)}$ as criterion we stop the iteration if the following holds:

$$\text{angle}\left(\boldsymbol{B}^{(n)}, \boldsymbol{B}^{(n+1)}\right) < \epsilon \tag{4.58}$$

Where $\epsilon > 0$ is a scalar.

In face recognition, the traditional sparse coding can be formulated as:

$$\min_{\gamma} \left\| y - \boldsymbol{X}^\top \gamma \right\|_2^2 + \lambda_2 \left\| \gamma \right\|_1 \tag{4.59}$$

Where $y = [y_1, y_2, \cdots, y_p] \in \mathbb{R}^{p \times 1}$ is a new test sample belong to $i-$th class, $\boldsymbol{X}^\top = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m] \in \mathbb{R}^{p \times n}$ is the matrix of training samples $\boldsymbol{x}_i$ is the $i-$th object class $\gamma = [0, \cdots, 0, z_1, z_2, \cdots, z_m, 0, \cdots, 0] \in \mathbb{R}^n$ is a coefficient vector whose entries are zero expect those associated with the $i-$th class $\lambda_2 \geq 0$ Lagrange multipliers. The (4.59) is a convex optimization problem, which can be solved by Lasso algorithm or Elastic Net algorithm. Suppose the test sample y is approximated as $\hat{y}_i = \boldsymbol{X}^\top \hat{\gamma}_i$, we can classify y to the object class with the largest entry in $\hat{\gamma}$, or classify y based on these $\hat{y}_i$ by assigning it to the object class that minimizes the residual:

$$\min_i \, r_i(y) \doteq \left\| y - \hat{y}_i \right\|_2 \tag{4.60}$$

For raw image data y, the dimension p usually is very large which will cause large data dimension and computation cost. The many feature extraction methods be used. Most of the feature transformations were linear or approximately operations $\boldsymbol{V}^\top \in \mathbb{R}^{k \times n}$, which project the image space to the feature space usually the $k << p$. Applying $\boldsymbol{V}^\top$ to (4.59) yields:

$$\hat{\gamma} = \min_{\gamma} \left\| \boldsymbol{V}^\top y - \boldsymbol{V}^\top \boldsymbol{X}^\top \gamma \right\|_2^2 + \lambda_2 \left\| \gamma \right\|_1 \tag{4.61}$$

In this paper we use the above robust PCA to extract the main feature. Then the classification be expressed in feature space as following (Wang and Cheng 2013):

$$\min_i \, r_i(y) \doteq \left\| \boldsymbol{V}^\top y - \boldsymbol{V}^\top \hat{y}_i \right\|_2 \tag{4.62}$$

4.4 Non-Convex Robust and Sparse PCA Via Hard Thresholding

The derived PCs the linear combinations of the original p variables are hard to be interpreted.(Jolliffe, Trendafilov, and Uddin 2003) and (Zou, Hastie, and Tibshirani 2006). One of the reasons is that PCs over-fit to noise and so almost all the elements of $\boldsymbol{V}$ are non zero. The other reason is that the loadings are orthogonal. Some methods modify PCA by introducing regularizations such that the derived loading is sparse (Ulfarsson and Solo 2011). A pioneer approach would be to directly apply the Lasso or L_1 penalty on the loadings(Jolliffe, Trendafilov, and Uddin 2003). The L_1 penalty shrinks the entries of loadings to zero until a sparse solution is derived as in the lasso regression(Tibshirani 1996). A more sophisticated and novel approach has been introduced where PCA can be used as a regression type optimization (elastic net optimization) which gives us sparse loadings by solving a L_1 and L_2 penalized regression(Zou, Hastie, and Tibshirani 2006). The resulting sparse principal component analysis is a promising method and is considered the benchmark. The SPCA requires additional algorithm to handle the elastic net optimization L_1 and L_2 penalized regression in each iteration(Zou and Hastie 2005). Furthermore, there are k tuning parameters each one controls the sparsity of a single PC.

There are 2 draw backs to sparse PCA:

- The choice of tuning parameters usually requires subjective domain knowledge or further searching steps.
- L_1- penalty introduces distortions by shrinking the entries of the loading matrix to zero.

Based on the above 2 drawbacks, we introduce a novel method substitute (SPCA-HT). The L_1- penalty introduces additional distortions by shrinking the elements of $\boldsymbol{V}$ to zero. Therefore the suggested SPCA-HT uses hard thresholding to regularize PCA. This approach only requires one tuning parameter which is the hard threshold and can be determined by following statistical decision theory. Moreover, there are two benefits arising from the relief of L1-penalty.

- SPCA-HT only requires linear operations and thus computational efficient even in $p >> n$.
- Simulations show that the SPCA-HT better estimates principal directions and thus explains more variance of the data since it does not shrink the coefficient of $\boldsymbol{V}$ to zero as the Lasso based methods do.

The first k columns of $\boldsymbol{V}$ in (4.35) can be obtained by solving the L_2 penalized optimization described in the below theorem.

Theorem 15. (PCA and Regularized Optimization) *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ both be $p \times k$ matrix such that $\boldsymbol{A} = [\boldsymbol{a}_1 \cdots \boldsymbol{a}_k]$ and $\boldsymbol{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_k]$. For any $\lambda > 0$ let $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ be the solution*

of the optimization:

$$\underset{\boldsymbol{A},\boldsymbol{B}}{minimize} \sum_{j=1}^{k} \left\{ \|\boldsymbol{X}\boldsymbol{a}_j - \boldsymbol{X}\boldsymbol{b}_j\|^2 + \lambda\|\boldsymbol{b}_j\|^2 \right\} \tag{4.63}$$
$$subject\ to\ \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}.$$

$$Then\ \boldsymbol{b}_j^*/\|\boldsymbol{b}_j^*\| = \boldsymbol{v}_j,\ \forall j = 1,2,\cdots,k. \tag{4.64}$$

The SPCA introduces the L_1-penalty into (4.63).

We now define what SPCA is and what the SPCs are we follow up the definition with an algorithm of how to perform SPCA.

Definition 1. (Sparse PCA) *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ both be $p \times k$ matrix such that $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots \boldsymbol{a}_k]$ and $\boldsymbol{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_k]$. Given $\lambda > 0$ nonnegative constants $\lambda_j, j = 1,2,\cdots k$, and let $(\hat{\boldsymbol{A}},\hat{\boldsymbol{B}})$ be the solution of the optimization:*

$$\underset{\boldsymbol{A},\boldsymbol{B}}{minimize} \sum_{j=1}^{k} \left\{ \|\boldsymbol{X}\boldsymbol{a}_j - \boldsymbol{X}\boldsymbol{b}_j\|^2 + \lambda\|\boldsymbol{b}_j\|^2 + \lambda_j\|\boldsymbol{b}_j\|_1 \right\} \tag{4.65}$$
$$subject\ to\ \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I}$$

$$Then,\ the\ sparse\ loadings\ are\ \hat{\boldsymbol{v}}_j = \hat{\boldsymbol{b}}_j/\|\hat{\boldsymbol{b}}_j\|, \forall j \tag{4.66}$$

The following is an iterative algorithm to show how SPCA is done.(Wu and Chen 2016)

Algorithm 1: SPCA

1. Initialize with $\boldsymbol{A} = [\boldsymbol{v}_1\ \boldsymbol{v}_2 \cdots \boldsymbol{v}_k]$ where $\boldsymbol{X} = \boldsymbol{U} \wedge \boldsymbol{V}^\top$.
2. **Update $\boldsymbol{B}$ given $\boldsymbol{A}$**: For $j = 1,2,\cdots,k$ update $\boldsymbol{b}_j$ by solving the elastic net optimization

$$\underset{b}{\text{minimize}}\ \|\boldsymbol{X}\boldsymbol{a}_j - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda_2\|\boldsymbol{b}\|^2 + \lambda_{1,j}\|\boldsymbol{b}\|_1. \tag{4.67}$$

3. **Update $\boldsymbol{A}$ given $\boldsymbol{B}$**: Derive the SVD, $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{V}}^\top$, then $\boldsymbol{A} \longleftarrow \tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}}^\top$.
4. Repeat step 2 and 3 until convergence.
5. $\boldsymbol{v}_j \longleftarrow \boldsymbol{b}_j/\|\boldsymbol{b}_j\|, \forall j = 1,2,\cdots,k$

The L_1 penalty plays an important role in the SPCA. In the $L_q, q \in \mathbb{N}$ penalty family (Fan and Li 2001) shows that only the Lasso ($L_1 - penalty$) can produce a sparse solution and thus is capable of de-noising. In other words, the L_1- penalty prevents PCA from overexplaining the variance of noise. Moreover there are benefits after substituting hard thresholding for the L_1 penalty. Two of these benefits that arise are:

- The elastic net (4.67) is reduced to a ridge regression (Hoerl and Kennard 1970) and can be solved by linear operations.

- Hard thresholding would not introduce additional distortions of $\boldsymbol{V}$ by shrinking the elements to zero as the Lasso does. Thus we use hard thresholding for the proposed method.

Now we use a regularization matrix to control the sparsity of $\boldsymbol{V}$.

Definition 2. (Regularization Matrix $\boldsymbol{G}$.) *Any matrix $\boldsymbol{G} \in \{0,1\}^{p\times k}$ can be a valid regularization matrix and $[\boldsymbol{V}]_{ij} = 0$ if $[\boldsymbol{G}]_{ij} = 0$*

With a well- designed regularization matrix $\boldsymbol{G}$, we can replace the L_1 penalty in (4.65) by the zero constraints introduced by $\boldsymbol{G}$ and still be able to derive sparse loadings.

Definition 3. (Sparse PCA via Hard Thresholding) *Given a sparse regularization matrix $\boldsymbol{G}$. For any $\lambda > 0$ Let $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}})$ be the solution of the optimization*

$$\begin{aligned} &\underset{\boldsymbol{A},\boldsymbol{B}}{minimize} \sum_{j=1}^{k} \left\{ \|\boldsymbol{X}\boldsymbol{a}_j - \boldsymbol{X}\boldsymbol{b}_j\|^2 + \lambda\|\boldsymbol{b}_j\|^2 \right\} \\ &subject\ to\ \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I},\ \ [\boldsymbol{B}]_{ij} = 0\ if\ [\boldsymbol{G}]_{ij} = 0 \end{aligned} \tag{4.68}$$

$$Then,\ the\ sparse\ loadings\ are: \tilde{\boldsymbol{v}}_j = \tilde{\boldsymbol{b}}_j / \|\tilde{\boldsymbol{b}}_j\|,\ \forall j. \tag{4.69}$$

Furthermore we show that the SPCA HT can be obtained by an algorithm using linear operations which is essential for computational efficiency.

Theorem 16. (Equivalent Form of the SPCA-HT:) *Let $\boldsymbol{D}_j$ be the diagonal matrix with $[\boldsymbol{D}_j]_{ii} = [\boldsymbol{G}]_{ij}$ that is $\boldsymbol{D}_j = diag(\boldsymbol{g}_j)$. Then $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}})$ in Definition 3 can be obtained by solving:*

$$\begin{aligned} &\underset{\boldsymbol{A},\boldsymbol{B}}{minimize} \sum_{j=1}^{k} \left\{ \|\boldsymbol{X}\boldsymbol{a}_j - \boldsymbol{X}\boldsymbol{D}_j\boldsymbol{b}_j\|^2 + \lambda\|\boldsymbol{b}_j\|^2 \right\} \\ &subject\ to\ \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I} \end{aligned} \tag{4.70}$$

Therefore the elastic net (4.67) *of Algorithm 1 is reduced to a ridge regression with solution:*

$$\boldsymbol{b}_j = (\boldsymbol{D}_j\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{D}_j + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}_j\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a}_j \tag{4.71}$$

$$Moreover\ ,\ when\ p >> n,\ (4.71)\ is\ further\ simplified\ to:$$

$$\boldsymbol{b}_j = \boldsymbol{D}_j\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a}_j \tag{4.72}$$

We can show how SPCA-HT is done using the following algorithm (Wu and Chen 2016):

Algorithm 2: Proposed SPCA-HT

1. Given $\boldsymbol{G} = [\boldsymbol{g}_1\ \boldsymbol{g}_2 \cdots \boldsymbol{g}_k]$ and initialize with $\boldsymbol{A} = [\boldsymbol{v}_1\ \boldsymbol{v}_2 \cdots \boldsymbol{v}_k]$ where $\boldsymbol{X} = \boldsymbol{U} \wedge \boldsymbol{V}^\top$.
2. Update $\boldsymbol{B}$ given $\boldsymbol{A}:\ \forall\ j = 1, 2, \cdots, k, \boldsymbol{D}_j \leftarrow diag(\boldsymbol{g}_j)$.

$$\boldsymbol{b}_j \leftarrow \begin{cases} \boldsymbol{D}_j \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}_j & ,\text{if } p >> n. \\ (\boldsymbol{D}_j \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{D}_j + \lambda \boldsymbol{I})^{-1} \boldsymbol{D}_j \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{a}_j & ,\text{ otherwise.} \end{cases}$$

3. **Update $\boldsymbol{A}$ given $\boldsymbol{B}$:** Derive the SVD, $\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{B} = \tilde{\boldsymbol{U}} \tilde{\wedge} \tilde{\boldsymbol{V}}^\top$ then $\boldsymbol{A} \leftarrow \tilde{\boldsymbol{U}} \tilde{\boldsymbol{V}}^\top$.
4. Repeat step 2 and 3 until convergence.
5. $\boldsymbol{v}_j \leftarrow \boldsymbol{b}_j / \|\boldsymbol{b}_j\|, \forall j = 1, 2, \cdots, k$.

We summarize with the following three points how SPCA-HT outdoes the LASSO with the following 3 benefits:

- It is easy to implement since only linear operations are required in the algorithm.
- The threshold tuning parameter ρ can be objectively chosen based on statistical decision theory.
- The hard thresholding does not introduce additional distortions of the principal directions by shrinking them to zero as the Lasso does.

We now introduce a few essential notations needed in order to comprehend the coming notions. Lowercase letters denote scalars and uppercase letters denote matrices unless otherwise stated. $\boldsymbol{A}_i \cdot$ and $\boldsymbol{A} \cdot j$ represent the i^{th} row and the j^{th} column of $\boldsymbol{A}$. Projection onto support set Ω is given by $\prod_\Omega$. $|\boldsymbol{A}|$ is the element wise absolute value of matrix $\boldsymbol{A}$. For norms of matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_F$ is the Frobenius norm; $\|\boldsymbol{A}\|_*$ is the nuclear norm; $\|\boldsymbol{A}\|_2$ is the largest singular value, $\|\boldsymbol{A}\|_p$ is the l_p- norm of vectorized $\boldsymbol{A}$ and $\|\boldsymbol{A}\|_{2,\infty}$ is the maximum of matrix row l_2- norms. Moreover $\langle \boldsymbol{A}, \boldsymbol{B} \rangle$ represents $\text{tr}(\boldsymbol{A}^\top \boldsymbol{B})$ for real matrices $\boldsymbol{A}, \boldsymbol{B}$. Additionally σ_i is the i^{th} largest singular value of a matrix. The Euclidean metric is not applicable here because of the non-uniqueness of the bi- factorisation $\boldsymbol{L}^* = \boldsymbol{A}^* \boldsymbol{B}^{*\top}$ which corresponds to a manifold rather than a point. So we define the following distance between $(\boldsymbol{A}, \boldsymbol{B})$ and any of the optimal pair $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ such that $\boldsymbol{L}^* = \boldsymbol{A}^* \boldsymbol{B}^{*\top}$:

$$d(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{A}^*, \boldsymbol{B}^*) = \min_{\boldsymbol{R}} \sqrt{\|\boldsymbol{A} - \boldsymbol{A}^* \boldsymbol{R}\|_F^2 + \|\boldsymbol{B} - \boldsymbol{B}^* \boldsymbol{R}\|_f^2} \tag{4.73}$$

Where $\boldsymbol{R}$ is an $r \times r$ orthogonal matrix.

We now show a novel non-convex optimization approach to decompose a given observation matrix into a low-rank core and the corresponding sparse residual. We suppose that there is a known data matrix $\boldsymbol{M} \in {}^{n_1 \times n_2}$ which can be decomposed into a low-rank component $\boldsymbol{L}^*$ and a sparse error matrix $\boldsymbol{S}^*$ of compatible dimensions. Our aim is to identify these underlying matrices and hence robustly recover the low-rank component with the help of available side information in the form of feature matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. Concretely let $\boldsymbol{L}^* = \boldsymbol{U}^* \boldsymbol{\Sigma}^* \boldsymbol{V}^{*\top}$ be the singular value decomposition and $\boldsymbol{P}^* = \boldsymbol{X}^\top \boldsymbol{U}^* \boldsymbol{\Sigma}^{\frac{1}{2}}$ and $\boldsymbol{Q}^* = \boldsymbol{Y}^\top \boldsymbol{V}^* \boldsymbol{\Sigma}^{*\frac{1}{2}}$. $\boldsymbol{S}^*$ follows the random sparsity model. That is the support of $\boldsymbol{S}^*$ is chosen

uniformly at random from the collection of all support sets of the same size. Now assume that there are also available features $\boldsymbol{X} \in \mathbb{R}^{n_1 \times d_1}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times d_2}$ s.t. they are feasible i.e. $\text{col}(\boldsymbol{X}) \supseteq \text{col}\,(\boldsymbol{U}^*)$ and $\text{col}(\boldsymbol{Y}) \supseteq \text{col}\,(\boldsymbol{V}^*)$ where $\text{col}\,(\boldsymbol{A})$ is the column space of $\boldsymbol{A}$ and $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{Y}^\top \boldsymbol{Y} = \boldsymbol{I}$. We now discuss robust low rank recovery using the above mentioned features and three different incoherence conditions: $(i)\, \|\boldsymbol{U}^*\|_{2,\infty} \leq \sqrt{\frac{\mu_1 r}{n_1}}$ and $\|\boldsymbol{V}^*\|_{2,\infty} \leq \sqrt{\frac{\mu_1 r}{n_2}}$; $(ii)\, \left\|\boldsymbol{X}_{2,\infty} \leq \sqrt{\frac{\mu_2 d_1}{n_1}}\right\|$ and $\|\boldsymbol{Y}\|_{2,\infty} \leq \sqrt{\frac{\mu_2 d_2}{n_2}}$; (iii) both (i) and (ii) where r is the given rank of $\boldsymbol{L}^*$ and μ_1, μ_2 are constants.

$$\mathcal{T}_\theta(\boldsymbol{A})_{ij} = \begin{cases} 0 & \text{if } |\boldsymbol{A}_{ij}| \leq \boldsymbol{A}^\theta i\cdot \text{ or } |\boldsymbol{A}_{ij}| \leq \boldsymbol{A}^\theta \cdot j, \\ \boldsymbol{A}_{ij} & \text{otherwise,} \end{cases}$$

Where $\boldsymbol{A}^\theta i\cdot, \boldsymbol{A}^\theta \cdot j$ are the $(n_2\theta)^{\text{th}}$ and $(n_1\theta)^{\text{th}}$ largest element in absolute value in row i and column j respectively.

$\boldsymbol{S}$ is first initialized as $\boldsymbol{S}_0 = \mathcal{T}_\alpha(\boldsymbol{M})$. We then obtain $\boldsymbol{U}_0 \boldsymbol{\Sigma}_0 \boldsymbol{V}_0^\top$ as the $r-$ truncated SVD of $\boldsymbol{L}_0$ which is calculated via $\boldsymbol{L}_0 = \boldsymbol{M} - \boldsymbol{S}_0$. We then construct $\boldsymbol{P}_0 = \boldsymbol{X}^\top \boldsymbol{U}_0 \boldsymbol{\Sigma}_0^{\frac{1}{2}}$ and $\boldsymbol{Q}_0 = \boldsymbol{Y}^\top \boldsymbol{V}_0 \boldsymbol{\Sigma}_0^{\frac{1}{2}}$. Such an initialization scheme gives $\boldsymbol{P}, \boldsymbol{Q}$ the desirable properties for use in the subsequent phase.

$$\mathcal{P} = \left\{ \boldsymbol{A} \in \mathbb{R}^{d_1 \times r} \, \|\boldsymbol{X}\boldsymbol{A}\|_{2,\infty} \leq \sqrt{\frac{2\mu_1 r}{n_1}} \|\boldsymbol{P}_0\|_2 \right\} \tag{4.74}$$

$$\mathcal{Q} = \left\{ \boldsymbol{A} \in \mathbb{R}^{d_2 \times r} \, \|\boldsymbol{Y}\boldsymbol{A}\|_{2,\infty} \leq \sqrt{\frac{2\mu_1 r}{n_1}} \|\boldsymbol{Q}_0\|_2 \right\} \tag{4.75}$$

We can simply take $\mathcal{P}$ as $\mathbb{R}^{d_1 \times r}$ and $\mathcal{Q}$ as $\mathbb{R}^{d_2 \times r}$. To proceed we first regularise $\boldsymbol{P}_0$ and $\boldsymbol{Q}_0$:

$$\boldsymbol{P} = \boldsymbol{\Pi}_{\mathcal{P}}(\boldsymbol{P}_0), \boldsymbol{Q} = \boldsymbol{\Pi}_{\mathcal{Q}}(\boldsymbol{Q}_0) \tag{4.76}$$

At each iteration we start by updating $\boldsymbol{S}$ with the sparse estimator using a threshold of $\alpha + \min(10\alpha + 0.1)$:

$$\boldsymbol{S} = \mathcal{T}_{\alpha + \min(10\alpha, 0.1)} \left(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top \boldsymbol{Y}^\top \right) \tag{4.77}$$

For $\boldsymbol{P}, \boldsymbol{Q}$ we define the following objective function:

$$\mathcal{L}(\boldsymbol{P}, \boldsymbol{Q}) = \frac{1}{2} \left\| \boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top \boldsymbol{Y}^\top + \boldsymbol{S} - \boldsymbol{M} \right\|_F^2 + \frac{1}{64} \left\| \boldsymbol{P}^\top \boldsymbol{P} - \boldsymbol{Q}^\top \boldsymbol{Q} \right\|_F^2 \tag{4.78}$$

$\boldsymbol{P}$ and $\boldsymbol{Q}$ are updated by minimizing the above function subject to the constraints imposed by the sets $\mathcal{P}$ and $\mathcal{Q}$:

$$\boldsymbol{P} = \boldsymbol{\Pi}_{\mathcal{P}} \left(\boldsymbol{P} - \eta \boldsymbol{\nabla}_{\boldsymbol{P}} \mathcal{L} \right) \tag{4.79}$$

$$\boldsymbol{Q} = \boldsymbol{\Pi}_{\mathcal{Q}} \left(\boldsymbol{Q} - \eta \boldsymbol{\nabla}_{\boldsymbol{Q}} \mathcal{L} \right) \tag{4.80}$$

Where the step size η is determined analytically below and $\boldsymbol{P}$ and $\boldsymbol{Q}$ are properly initialized. Such an optimization design converges to $\boldsymbol{P}^*$ and $\boldsymbol{Q}^*$. The procedure is summarized in algorithm 3. The former initialization phase provides us with the following guarantees on $\boldsymbol{P}$ and $\boldsymbol{Q}$.(Xue et al. 2017)

Theorem 17. *In cases (i) and (iii) if* $\alpha \leq \frac{1}{16\kappa r \mu_1}$ *we have:*

$$d(\boldsymbol{P}_0, \boldsymbol{Q}_0, \boldsymbol{P}^*, \boldsymbol{Q}^*) \leq 18\alpha r \mu_1 \sqrt{r\kappa\sigma_1^*} \tag{4.81}$$

In case (ii),if $\alpha \leq \frac{1}{16\kappa\mu_2\sqrt{d_1 d_2}}$ *we have:*

$$d(\boldsymbol{P}_0, \boldsymbol{Q}_0, \boldsymbol{P}^*, \boldsymbol{Q}^*) \leq 18\alpha \mu_2 \sqrt{r d_1 d_2 \kappa \sigma_1^*} \tag{4.82}$$

where κ *is the condition number of* $\boldsymbol{L}^*$ *and* d *is a distance metric.*

Theorem 18. *For* $\eta \leq \frac{1}{192\|L_0\|_2}$ *there exist constants* $c_1 > 0, c_2 > 0, c_3 > 0, c_4 > 0, c_5 > 0, c_6 > 0$ *such that in case (i) when* $\alpha \leq \frac{c_1}{\mu_1(\kappa r)^{\frac{3}{2}}}$ *we have the following relationship:*

$$d(\boldsymbol{P}_t, \boldsymbol{Q}_t, \boldsymbol{P}^*, \boldsymbol{Q}^*)^2 \leq (1 - c_2\eta\sigma_r^*)^t d(\boldsymbol{P}_0, \boldsymbol{Q}_0, \boldsymbol{P}^*, \boldsymbol{Q}^*)^2 \tag{4.83}$$

in case (ii) when $\alpha \leq \frac{c_3}{\mu_2 d r^{\frac{1}{2}} \kappa^{\frac{3}{2}}}$, *we have:*

$$d(\boldsymbol{P}_t, \boldsymbol{Q}_t, \boldsymbol{P}^*, \boldsymbol{Q}^*)^2 \leq (1 - c_4\eta\sigma_r^*)^t d(\boldsymbol{P}_0, \boldsymbol{Q}_0, \boldsymbol{P}^*, \boldsymbol{Q}^*)^2 \tag{4.84}$$

and in case (iii), when $\alpha \leq c_5$ *min* $\left(\frac{1}{\mu_2 d_\kappa}, \frac{1}{\mu_1(\kappa r)^{\frac{3}{2}}}\right)$ *we have:*

$$d(\boldsymbol{P}_t, \boldsymbol{Q}_t, \boldsymbol{P}^*, \boldsymbol{Q}^*)^2 \leq (1 - c_6\eta\sigma_r^*)^t d(\boldsymbol{P}_0, \boldsymbol{Q}_0, \boldsymbol{P}^*, \boldsymbol{Q}^*)^2 \tag{4.85}$$

Algorithm 3: Non-convex solver for robust principal component analysis with features

Input: Observation $\boldsymbol{M}$, features $\boldsymbol{X}$, $\boldsymbol{Y}$, rank r, corruption approximation α and step size η.

Initialization:

1. $\boldsymbol{S} = \mathcal{T}_\alpha(\boldsymbol{M})$
2. $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = r - \text{SVD}(\boldsymbol{M} - \boldsymbol{S})$
3. $\boldsymbol{P} = \boldsymbol{X}^\top \boldsymbol{U}\boldsymbol{\Sigma}^{\frac{1}{2}}$
4. $\boldsymbol{Q} = \boldsymbol{Y}^\top \boldsymbol{V}\boldsymbol{\Sigma}^{\frac{1}{2}}$

Gradient Descent:

5. $\boldsymbol{P} = \boldsymbol{\Pi}_{\mathcal{P}}(\boldsymbol{P})$
6. $\boldsymbol{Q} = \boldsymbol{\Pi}_{\mathcal{Q}}(\boldsymbol{Q})$
7. **while** not converged **do**
8. $\quad \boldsymbol{S} = \mathcal{T}_{\alpha+\min(10\alpha,0.1)}(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top\boldsymbol{Y}^\top)$
9. $\quad \boldsymbol{P} = \boldsymbol{\Pi}_{\mathcal{P}}(\boldsymbol{P} - \eta\boldsymbol{\nabla}_{\boldsymbol{P}}\mathcal{L})$
10. $\quad \boldsymbol{Q} = \boldsymbol{\Pi}_{\mathcal{Q}}(\boldsymbol{Q} - \eta\boldsymbol{\nabla}_{\boldsymbol{Q}}\mathcal{L})$
11. **end while**

Return: $\boldsymbol{L} = \boldsymbol{X}\boldsymbol{P}\boldsymbol{Q}^\top\boldsymbol{Y}^\top, \boldsymbol{S}$

Chapter 5

Compression for Some Spectral Regression Estimators

5.1 The PCR Estimator and Properties That it Entails

In this chapter of the thesis I will empirically explore the bias, variance and MSE as well as other properties of the PCR estimator. We first start with defining our data set and mapping of the space respectively:

$$\mathcal{D} = \Big\{(X_i, Y_i) \overset{iid}{\sim} \mathcal{P}_{XY}(x, y),\ i = 1, \cdots, n\Big\} \tag{5.1}$$

$$\text{Where: } \boldsymbol{X} \in \mathbb{R}^p,\ \boldsymbol{Y} \in \mathbb{R}^r$$

$$\phi : \begin{cases} \mathbb{R}^p \longrightarrow \mathbb{R}^r \qquad r <<< p \\ x_i \longrightarrow z_i \in \mathbb{R}^r, \quad \phi(x_i) = z_i \text{ (This matrix is orthogonal)} \end{cases}$$

Where the individual principal components are defined as follows:

$$\begin{pmatrix} z_{i1} & z_{i2} & \cdots & z_{iq} \end{pmatrix}^\top = z_i \tag{5.2}$$

Another way to define each principal component is to take $i \in \{1, \cdots, k\}$ and let x_i denote the $k \times i$ matrix with orthonormal columns consisting of the first i columns of X then:

Let:

$$z_i \overset{LPCA}{=} Vx_i = \begin{bmatrix} Vx_1 & \cdots & Vx_k \end{bmatrix} \tag{5.3}$$

Where z_i denotes the $n \times i$ matrix having the first i principal components as its columns.

We define The PCR estimator as follows:

$$\begin{aligned} \hat{f}_{PCR}(x) &= x^\top v^\top (z^\top z)^{-1} z^\top \boldsymbol{Y} \\ &= z^\top (z^\top z)^{-1} z^\top \boldsymbol{Y} \end{aligned} \tag{5.4}$$

The regression model for our simulated data is defined as:

$$Y_i = f^\star(x_i) + \epsilon_i \tag{5.5}$$

It would be nice to derive the following properties theoretically but in the context of this work we will show the results empirically. We will study the following the properties:

$$\text{Bias}(\hat{f}_{PCR}(x)) = \mathbb{E}[\hat{f}_{PCR}(x)] - f^\star(x) \tag{5.6}$$

$$\mathbb{V}(\hat{f}_{PCR}(x)) = \mathbb{E}\left\{(\hat{f}_{PCR}(x) - \mathbb{E}[\hat{f}_{PCR}(x)])^2\right\} \tag{5.7}$$

$$\begin{aligned} \text{MSE } (\hat{f}_{PCR}(x)) &= \text{ Bias}(\hat{f}_{PCR}(x))^2 + \text{Variance}(\hat{f}_{PCR}(x)) \\ &= \left(\mathbb{E}[\hat{f}_{PCR}(x)] - f^\star(x)\right)^2 + \mathbb{E}\left\{(\hat{f}_{PCR}(x) - \mathbb{E}[\hat{f}_{PCR}(x)])^2\right\} \end{aligned} \tag{5.8}$$

$$\text{PE } (\hat{f}_{PCR}(x)) = \mathbb{E}[||\hat{f}_{PCR} - f||^2_{\mathcal{H}}] \tag{5.9}$$

And the percentage of true variation captured (2.5) for the PCR estimator.

The variance of the principal component estimator can also be written as:

$$\mathbb{V}(\hat{f}_{PCR}) = \int_{\mathcal{X}} \mathbb{V}(\hat{f}_{PCR}(x)) \mathcal{P}_X(x) dx \tag{5.10}$$

5.2 Methodology and Outline of Empirical Exploration of Distinct Performance Characteristics for PCR Method on Simulated Data for Case $N > P$ ($N = 100$)

This study focuses on properties of principal components regression (PCR) in different settings. We are mostly concerned with what happens when the number of principal components (r) increases from 1 to a large number. Our simulation experiment largely follows the methodology and specifications of the experiment in the main study (comparing 14 different estimation methods). In particular, the considered parameter combinations are variations of one base specification:

1. $r = 1, 2, \cdots, 14, 15$.
2. $p = 5$.
3. $\rho_X = 0.5$.
4. $q = 10$.
5. $\rho_Z = 0.5$.
6. $S/N = 0.1$ and 1.
7. $\nu = +Inf$.

Furthermore:

(A) For each configuration $(p, \rho_X, q, \rho_Z, S/N, \nu)$, we simulate 1000 independent models $(\beta_1, \cdots, \beta_p)$;

(B) For each simulated model $(\beta_1, \cdots, \beta_p)$, we simulate one training sample of size 100 and one test sample of size 10000.

Since PCR gives a non-zero weight to each candidate predictor, it formally "selects" each candidate predictor. So the following performance measures are no longer interesting to us:

- Percentage of (correctly) selected true predictors X.
- Percentage of (mistakenly) selected false predictors Z,

We look only at the following measures:

1. Out-Of-Sample Prediction Error (evaluated on an independently simulated test data set).
2. Root-Mean-Square Error of coefficients estimates (averaged over the coefficients, which is fine since all of them are of the same order of magnitude)
3. Bias of coefficients estimates.

4. Variance of coefficients estimates.

5. Percentage of captured true variation: R^2 in the regression of the true signal on the estimated signal.

5.3 Results of Empirical Exploration of Distinct Performance Characteristics for PCR method.

In this document we only state the main findings and illustrate our thoughts with the most important graphs. The complete set of results can be found in the appendix section. PCR with only 1 principal component does not perform well, which will be explained later. However, the first thing we notice is that the estimation and prediction performance of PCR improves with r increasing if the signal-to-noise ratio is substantial, e.g. 1 (Figures 5.1 - 5.3). The estimation and prediction performance of PCR reaches its peak at $r = 2$ and then deteriorates with r increasing if the signal-to-noise ratio is low, e.g. 0.1 (Figures 5.2 and 5.4).

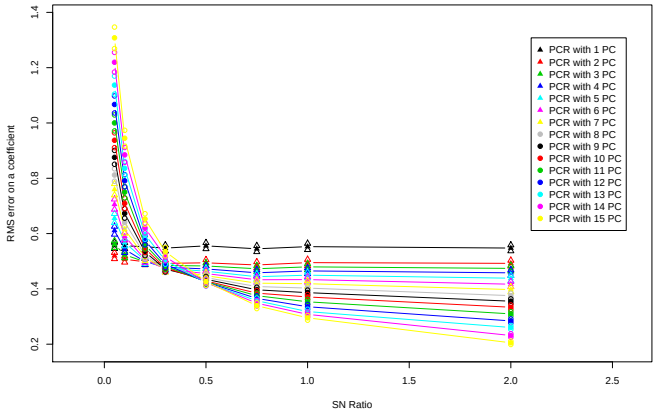

FIGURE 5.1: Performance of the Root-Mean-Square Error of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, \rho_X = 0.5, q = 10, \rho_Z = 0.5, \nu = +inf)$

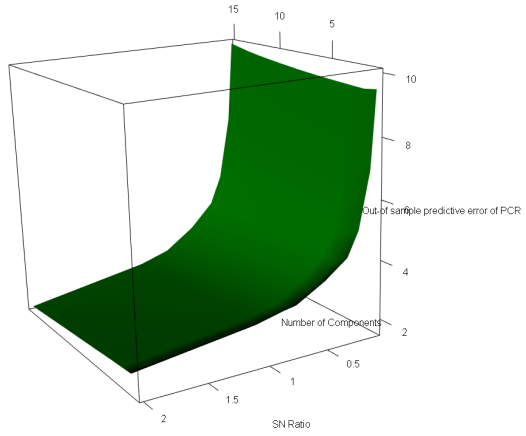

FIGURE 5.2: Performance of the Out-Of-Sample-Prediction- Error of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, \rho_X = 0.5, q = 10, \rho_Z = 0.5, \nu = +\text{inf})$

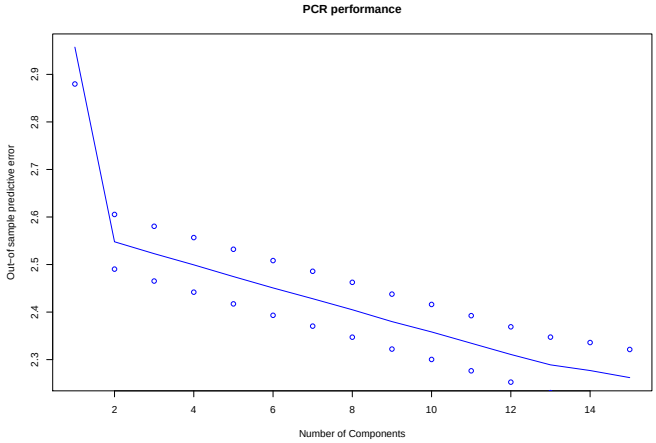

FIGURE 5.3: Performance of the Out-Of-Sample-Prediction- Error of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, \rho_X = 0.5, q = 10, \rho_z = 0.5, S/N = 1, \nu = +\text{inf})$

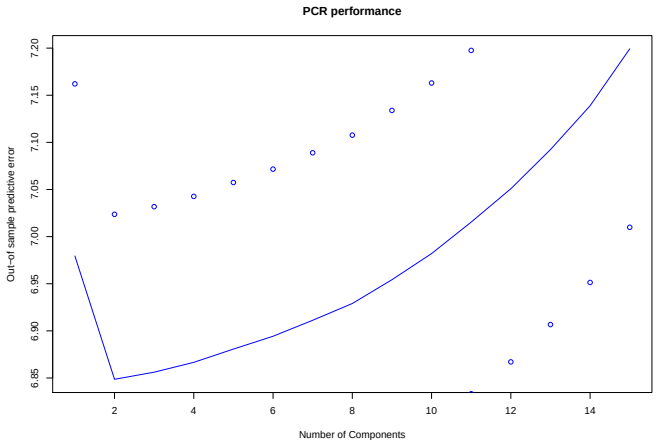

FIGURE 5.4: Performance of the Out-Of-Sample-Prediction- Error of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, \rho_X = 0.5, q = 10, \rho_z = 0.5, S/N = 0.1, \nu = +\text{inf})$

By construction, there are two independent themes going on: the correlated behavior of X's and the correlated behavior of Z's. PCR cannot capture the two themes with just 1 principal component. It needs at least two. For that reason, PCR with only one principal component performs much worse than its competitors, overall. Figures 5.5 - 5.7 highlight the issue. Note that, as ρ_Z increases, the correlated behavior of Z's become more and more important. At some point, the only allowed principal component becomes dedicated to Z's almost fully; almost completely ignoring the information contained in the true predictors (X's). As a result, the estimated model does not have much of explanatory or predictive power.

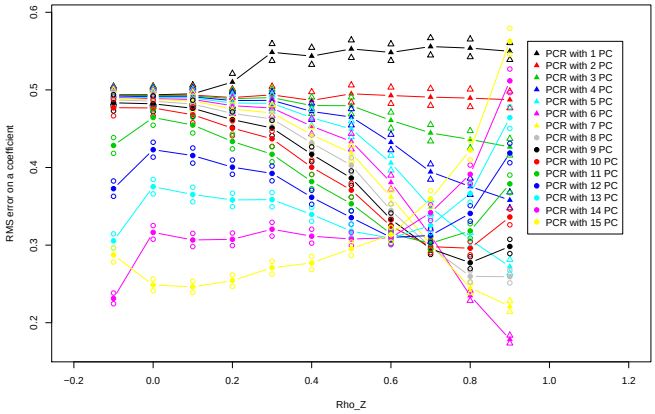

FIGURE 5.5: Performance of the Root-Mean-Square Error of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, \rho_X = 0.5, q = 10, S/N = 1, \nu = +\inf)$

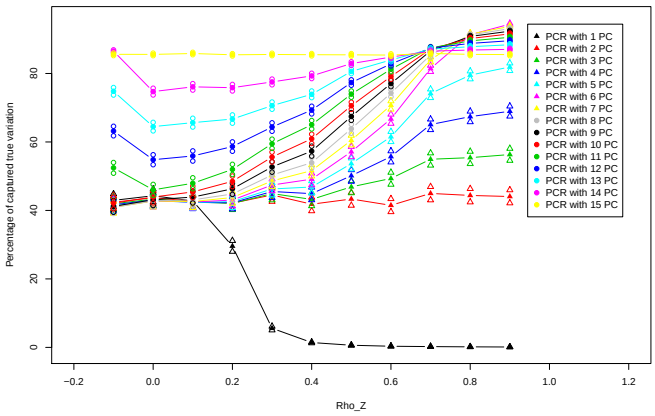

FIGURE 5.6: Performance of the Percentage of Captured True Variation of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, \rho_X = 0.5, q = 10, S/N = 1, \nu = +\inf)$

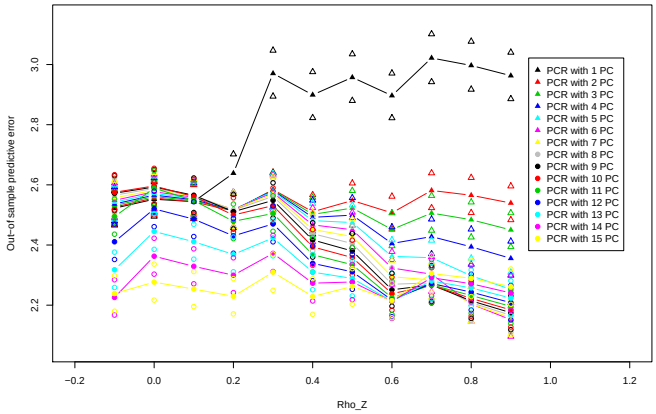

FIGURE 5.7: Performance of the Out-Of-Sample-Prediction- Error of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, \rho_X = 0.5, q = 10, S/N = 1, \nu = +\text{inf})$

The higher is the correlation among X's and/or Z's, the fewer components we need to do a relatively good job (Figures 5.5, 5.7 and 5.10).

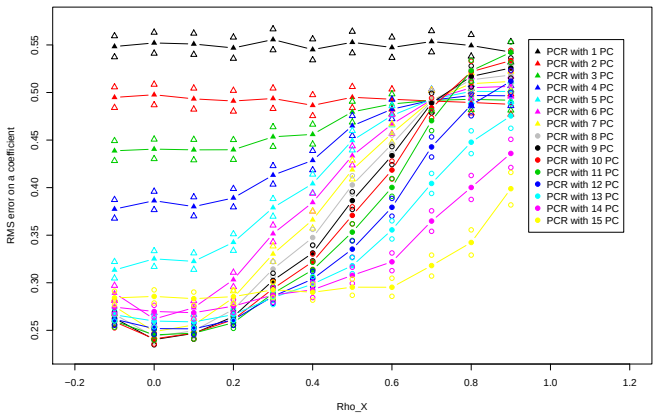

FIGURE 5.8: Performance of the Out-Of-Sample-Prediction- Error of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, q = 10, \rho_Z = 0.5, S/N = 1, \nu = +\text{inf})$

The bigger the true model is, the harder it is to estimate and forecast (Figure 5.11). All the specifications of PCR have trouble.

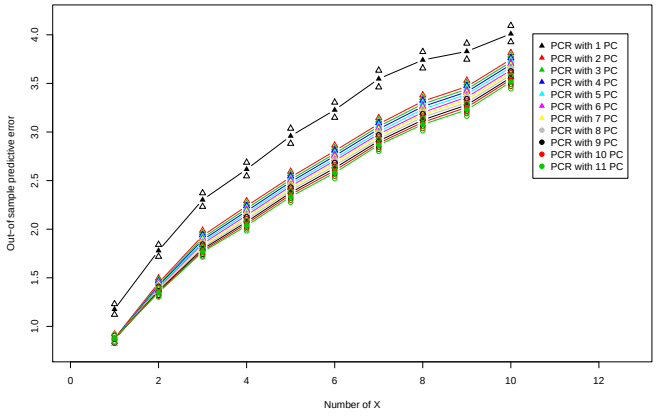

FIGURE 5.9: Performance of the Out-Of-Sample-Prediction- Error of Coefficients Estimates for PCR Method with Different PCs for the Case $(\rho_X = 0.5, q = 10, \rho_Z = 0.5, S/N = 1, \nu = +\text{inf})$

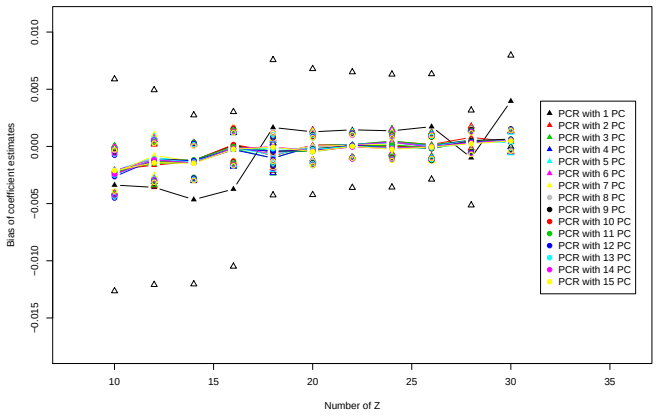

FIGURE 5.10: Performance of the Bias of Coefficients Estimates for PCR Method with Different PCs for the Case $(p = 5, \rho_X = 0.5, \rho_Z = 0.5, S/N = 1, \nu = +\text{inf})$

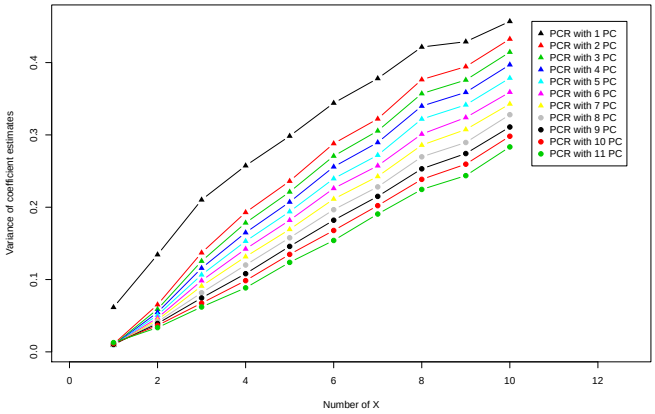

FIGURE 5.11: Performance of the Variance of Coefficients Estimates for PCR Method with Different PCs for the Case ($p = 5, \rho_X = 0.5, \rho_Z = 0.5, S/N = 1, \nu = +\inf$)

SN Ratio	PCR with 1 PC	PCR with 2 PC	PCR with 3 PC	PCR with 4 PC	PCR with 5 PC	PCR with 6 PC	PCR with 7 PC	PCR with 8 PC
0.05	9.5 (0.13)	9.4 (0.12)	9.4 (0.13)	9.5 (0.13)	9.5 (0.13)	0.5 (0.13)	9.6 (0.13))	9.6 (0.13)
0.1	7.0 (0.09)	6.8 (0.09)	6.9 (0.09)	6.9 (0.09)	6.9 (0.09)	6.9 (0.09)	6.9 (0.09)	6.9 (0.09)
0.2	5.1 (0.07)	4.9 (0.06)	4.9 (0.06)	4.9 (0.06)	4.9 (0.06)	4.9 (0.06)	4.9 (0.06)	4.9 (0.06)
0.3	4.3 (0.06)	4.0 (0.05)	4.0 (0.05)	4.0 (0.05)	4.0 (0.05)	4.0 (0.05)	4.0 (0.05)	4.0 (0.05)
0.5	3.7 (0.05)	3.4 (0.04)	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)
0.75	3.1 (0.04)	2.8 (0.03)	2.8 (0.03)	2.8 (0.03)	2.7 (0.03)	2.7 (0.03)	2.7 (0.03)	2.7 (0.03)
1	3.0 (0.04)	2.5 (0.03)	2.5 (0.03)	2.5 (0.03)	2.5 (0.03)	2.5 (0.03)	2.4 (0.03)	2.4 (0.03)
2	2.5 (0.03)	2.0 (0.02)	2.0 (0.02)	2.0 (0.02)	2.0 (0.02)	1.9 (0.02)	1.9 (0.02)	1.8 (0.02)

SN Ratio	PCR with 9 PC	PCR with 10 PC	PCR with 11 PC	PCR with 12 PC	PCR with 13 PC	PCR with 14 PC	PCR with 15 PC
0.05	9.6 (0.13)	9.7 (0.13)	9.7 (0.13)	9.8 (0.13)	9.8 (0.13)	9.9 (0.13)	10.0 (0.13)
0.1	7.0 (0.09)	7.0 (0.09)	7.0 (0.09)	7.1 (0.09)	7.1 (0.09)	7.1 (0.10)	7.2 (0.10)
0.2	4.9 (0.06)	4.9 (0.06)	4.9 (0.06)	4.9 (0.06)	5.0 (0.07)	5.0 (0.07)	5.0 (0.07)
0.3	4.0 (0.05)	4.0 (0.05)	4.0 (0.05)	4.0 (0.05)	4.0 (0.05)	4.1 (0.05)	4.1 (0.05)
0.5	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)	3.3 (0.04)
0.75	2.6 (0.03)	2.6 (0.03)	2.6 (0.03)	2.6 (0.03)	2.59 (0.03)	2.58 (0.03)	2.58 (0.03)
1	2.4 (0.03)	2.4 (0.03)	2.4 (0.03)	2.3 (0.03)	2.3 (0.03)	2.3 (0.03)	2.3 (0.03)
2	1.8 (0.02)	1.7 (0.02)	1.7 (0.02)	1.7 (0.02)	1.6 (0.02)	1.6 (0.02)	1.6 (0.02)

TABLE 5.1: SNR of Coefficients Estimates for PCR Method for the Case ($p = 5, \rho_X = 0.5, q = 10, \rho_z = 0.5, \nu = +\inf$); Standard Error is in the Brackets.

Chapter 6

Comparison of Data Sets for Variable Selection, Regularization and Compression of Some Spectral Regression Estimators

6.1 Methodology and Outline of Empirical Exploration Comparison of Distinct Performance Characteristics for Variable Selection, Regularization and Compression methods on Simulated Data and Observational Data Sets

In this study, we compare fourteen different methods of estimating linear models. The methods are:

1. Best Subset Selection
2. Forward Stepwise Selection governed by p-values and the significance level of 5%
3. Backward Stepwise Selection governed by p-values and the significance level of 5%
4. Forward Stepwise Selection governed by Akaike Information Criterion (AIC)
5. Backward Stepwise Selection governed by Akaike Information Criterion
6. Forward Stepwise Selection governed by Bayesian Information Criterion (BIC)
7. Backward Stepwise Selection governed by Bayesian Information Criterion
8. Forward Stepwise Selection governed by Cross-Validation (CV)

9. Backward Stepwise selection governed by Cross-Validation (CV)
10. The method of the Lasso
11. The method of Ridge Regression
12. The method of Elastic Net (a hybrid between the Lasso and Ridge Regression methods)
13. The method of Least Angle Regression (LAR)
14. Principal Components Regression (PCR)

The comparison is done on thousands of simulated data sets and five real data sets. We consider a number of different specifications. In particular, we run simulations for the cases when the number of candidate predictors is smaller than the sample size and when it is larger than the sample size. The simulation part is the most interesting and insightful part of the study. Calculations on real data sets purpose is illustrations mostly, as almost anything is possible on any given data set out of sampling variability (randomness).

The simulation study is structured as follows. For each set of parameters, we simulate $1,000$ independent models. Each model has the form:

$$\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \tag{6.1}$$

Where $X_1, \cdots, X_p$ are standard normal variables and each two of them have correlation ρ_X. Note that it is possible for the correlation between each two variables in the set to be negative. Residual ϵ has either t-distribution with ν degrees of freedom or standard normal distribution (the same as t-distribution with ∞ degrees of freedom). The true signal is defined as the systematic part of (6.1), so we have:

$$\text{True Signal } = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{6.2}$$

The signal-to-noise ratio is defined as:

$$S/N = \mathbb{V}\left[\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right] / \mathbb{V}[\epsilon]. \tag{6.3}$$

We call variables $X_1, \cdots, X_p$ the "true" predictors. Unfortunately, when building a linear model for $\boldsymbol{Y}$, the researcher does not know the exact list of determinants of $\boldsymbol{Y}$. At best, variables $X_1, \cdots, X_p$ are only part of the universe of "candidate" predictors. In addition to X's, the universe has variables $Z_1, \cdots, Z_q$, which are completely useless. They are uncorrelated with $\boldsymbol{Y}$ (so useless in the linear framework, to be exact). Z's represent research noise, giving way to potential errors in model selection. We call $Z_1, \cdots, Z_q$ "false" predictors. In our little simulation study, we assume that each of Z's has standard normal distribution and each two of them have correlation ρ_Z.

Altogether, parameters of $p, \rho_X, q, \rho_Z, S/N$ and $\boldsymbol{\nu}$ constitute the specification of any given random model. For each such model, we simulate true regression coefficients $(\beta_1, \cdots, \beta_p)$ as independent standard normal variables. The estimation methods are run on centered versions of $Y, X_1, \cdots, X_p, Z_1, \cdots, Z_q$. Therefore, the value of β_0 is set to 0 without loss of generality. For each model we simulate realizations of $(\boldsymbol{Y}, X_1, \cdots, X_p, Z_1, \cdots, Z_q)$ to create a sample of size N. The simulation is done independently over the models.

One may wonder why we estimate many models (many values of vector $(\beta_1, \cdots, \beta_p)$) for each configuration of parameters $(p, \rho_X, q, \rho_Z, S/N, \boldsymbol{\nu})$ instead of simulating many random samples per model. The answer to that would be estimation methods may get "lucky" or "unlucky". If $\beta_1, \cdots, \beta_p$ are simulated with comparable absolute values, ridge regression is favored over Lasso. If only a few of $\beta_1, \cdots, \beta_p$ have big absolute values and the rest are close to 0, Lasso is favored over ridge regression. Also, overall, estimation methods have more difficulties in estimating the signal when the signal is composed of many influential predictors instead of just a few.

The fourteen estimation methods are contrasted and compared with the help of seven distinct performance characteristics:

1. Out-Of-Sample Prediction Error (evaluated on an independently simulated test data set or via cross-validation if working with real data)

2. Root-Mean-Square Error of coefficients estimates (averaged over the coefficients, which is fine since all of them are of the same order of magnitude)

3. Bias of coefficients estimates

4. Variance of coefficients estimates

5. Percentage of (correctly) selected true predictors $\boldsymbol{X}$

6. Percentage of (mistakenly) selected false predictors $\boldsymbol{Z}$

7. Percentage of captured true variation: R^2 in the regression of the true signal on the estimated signal.

Characteristics 2-7 can be evaluated only if the true values of coefficients are known. Therefore, they are used during the simulation study only. This is no concern to us, the simulation study is more informative anyway. Out of the seven characteristics, we view 1. and 2. as most important. For diagnostic purposes, characteristics 5. and 6. are interesting as well.

We now outline how we will conduct our study on our observational data sets. Observational or real data we do not know the truth. That is in fact, the primary reason why we do statistics: to get to the truth as closely as possible. We would ask ourselves "how do we know if we are right or wrong?" or "How do we know whether the quality of our expertise is high or low?". There are several ways to look at ourselves from a distance. A good one as mentioned before is to use cross-validation.

We split the data into K blocks and approach model assessment in K steps. At each step,

we estimate the model on K-1 out of K blocks and see how well the estimated model predicts the dependent variable on the remaining K-th block. It goes without saying that this one testing block is different over the K steps. This way, the statistical power of model assessment is commensurate with the sample size of the whole data set. This way, each block gets to participate in both: training the model and testing the model. At the end, the root-mean-square prediction error is aggregated over the K steps. This is the metric which serves as an equivalent of the out of sample prediction error calculated during the simulation stage. This is the metric which is used to compare the performance of 14 estimation methods in real setting.

In our analysis we exploit the five real data sets that were introduced in chapter 1. To ensure relative homogeneity of experimental settings, we sample the same number of observations from each data set. In addition to that, we look at only a subset of available variables. This is done to mitigate computational issues: memory, speed, stability etc. The philosophy and implementation of most estimation methods is reasonable. However, the philosophy of the best subset selection is quite inefficient. Most analyses containing best subset selection cannot handle huge numbers of potential predictors. Best subset selection is too greedy of a method.

6.2 Results of Empirical Comparison of Distinct Performance Characteristics for Variable Selection, Regularization and Compression Methods on Simulated Data When Sample Size > Number of Predictors (N>p (N=100))

We start with the simulation results. Two types of specifications have been tried:

(A) Sample size bigger than the number of predictors: training sample size = 100; $p = 1, 2, \cdots, 10$; $q = 1, 3, \cdots, 13, 15$;

(B) Sample size smaller than the number of predictors: training sample size = 15; $p = 1, 2, \cdots, 10$; $q = 10, 12, \cdots, 18, 20$; $p + q > 15$.

In each case:

- For each configuration $(p, \rho_X, q, \rho_Z, S/N, \nu)$, we simulate 1,000 independent models $(\beta_1, \cdots, \beta_p)$;
- For each simulated model $(\beta_1, \cdots, \beta_p)$, we simulate one training sample and one test sample of size 10,000.

Below we display only the most important and illustrative results. For the complete set of graphs and tables please refer to appendices A and B.

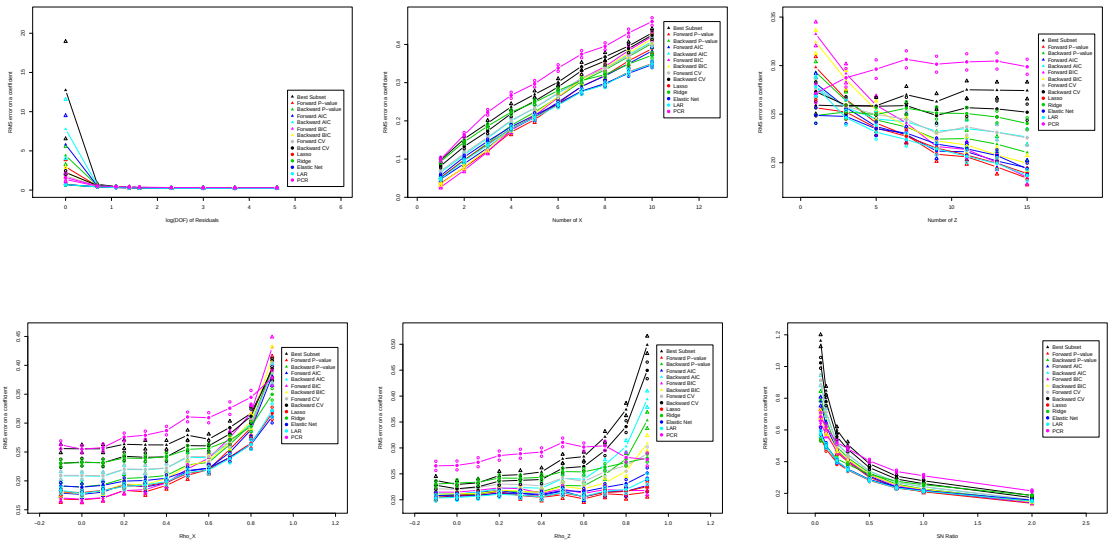

FIGURE 6.1: Comparison of Performance of the RMSE for Methods on Simulated Data of 1000 Models and when $N = 100$.

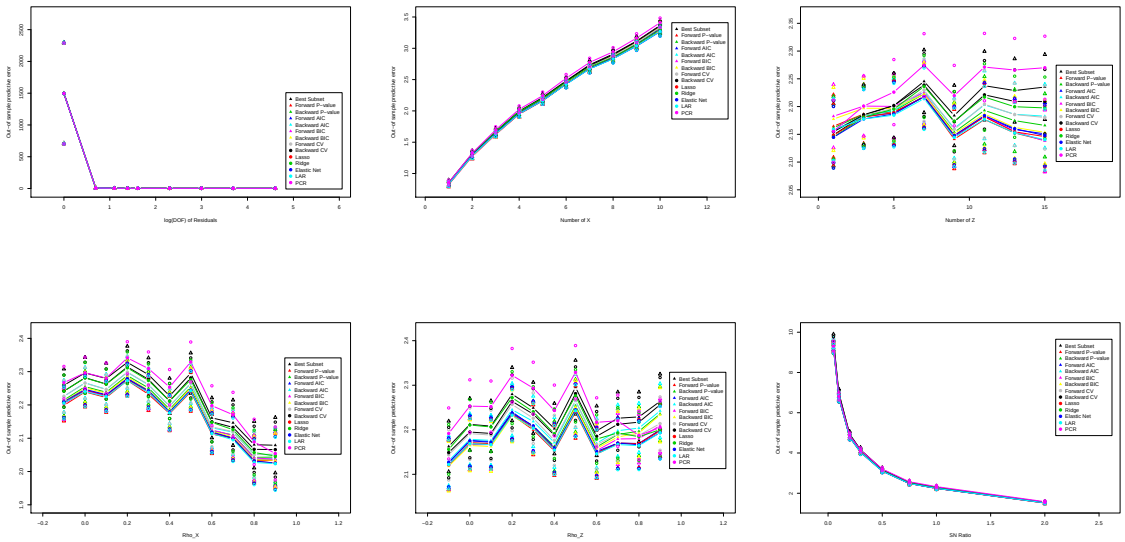

FIGURE 6.2: Comparison of Performance on the Out-Of-Sample-Predictive-Error for Methods on Simulated Data of 1000 Models and when $N = 100$.

Number Of Xs	Best Subset	Forward P-value	Backward P-value	Forward AIC	Backward AIC	Forward BIC	Backward BIC
1	0.10 (0.003)	0.04 (0.002)	0.05 (0.002)	0.05 (0.002)	0.07 (0.002)	0.03 (0.001)	0.03 (0.002)
2	0.15 (0.003)	0.08 (0.002)	0.10 (0.003)	0.09 (0.003)	0.11 (0.003)	0.07 (0.002)	0.08 (0.002)
3	0.19 (0.004)	0.12 (0.03)	0.14 (0.003)	0.13 (0.003)	0.15 (0.003)	0.12 (0.03)	0.13 (0.003)
4	0.24 (0.004)	0.17 (0.003)	0.19 (0.004)	0.18 (0.004)	0.20 (0.004)	0.18 (0.004)	0.18 (0.004)
5	0.27 (0.005)	0.20 (0.004)	0.22 (0.004)	0.21 (0.004)	0.23 (0.004)	0.21 (0.004)	0.22 (0.004)
6	0.30(0.004)	0.25 (0.004)	0.26 (0.004)	0.25 (0.004)	0.27 (0.004)	0.26 (0.004)	0.26 (0.005)
7	0.34 (0.005)	0.29 (0.005)	0.30 (0.005)	0.29 (0.005)	0.31 (0.005)	0.31 (0.005)	0.31 (0.005)
8	0.37 (0.005)	0.32 (0.005)	0.33 (0.005)	0.32 (0.005)	0.34 (0.005)	0.34 (0.005)	0.34 (0.005)
9	0.40 (0.005)	0.36 (0.005)	0.37 (0.005)	0.35 (0.005)	0.37 (0.005)	0.38 (0.005)	0.38 (0.005)
10	0.43 (0.006)	0.39 (0.005)	0.40 (0.005)	0.38 (0.005)	0.41 (0.006)	0.42 (0.006)	0.41 (0.006)

Table 6.1: Root-Mean-Square Error of Coefficient Estimates for Best Subset - Backward BIC methods; Standard Error is in the Brackets.

Number Of Xs	Forward CV	Backward CV	Lasso	Ridge	Elastic Net	LAR	PCR
1	0.07 (0.002)	0.08 (0.002)	0.05 (0.002)	0.09 (0.002)	0.06 (0.002)	0.05 (0.002)	0.10 (0.003)
2	0.11 (0.003)	0.13 (0.003)	0.10 (0.002)	0.14 (0.003)	0.11 (0.002)	0.10 (0.002)	0.16 (0.003)
3	0.16 (0.003)	0.17 (0.03)	0.14 (0.002)	0.19 (0.003)	0.15 (0.003)	0.14 (0.03)	0.22 (0.004)
4	0.20 (0.004)	0.22 (0.004)	0.18 (0.003)	0.22 (0.003)	0.18 (0.003)	0.18 (0.003)	0.27 (0.004)
5	0.23 (0.004)	0.25 (0.004)	0.21 (0.003)	0.25 (0.003)	0.21 (0.003)	0.21 (0.003)	0.30 (0.004)
6	0.27(0.004)	0.29 (0.005)	0.24 (0.003)	0.28 (0.003)	0.25 (0.003)	0.24 (0.003)	0.34 (0.004)
7	0.31 (0.005)	0.33 (0.005)	0.28 (0.004)	0.31 (0.004)	0.28 (0.004)	0.28 (0.004)	0.37 (0.005)
8	0.34 (0.005)	0.36 (0.005)	0.30 (0.004)	0.32 (0.004)	0.30 (0.004)	0.29 (0.004)	0.40 (0.005)
9	0.37 (0.005)	0.39 (0.005)	0.33 (0.004)	0.35 (0.004)	0.32 (0.004)	0.33 (0.004)	0.43 (0.005)
10	0.41 (0.005)	0.42 (0.006)	0.35 (0.004)	0.37 (0.004)	0.35 (0.004)	0.35 (0.004)	0.46 (0.005)

Table 6.2: Root-Mean-Square Error of Coefficient Estimates for Forward CV - PCR; Standard Error is in the Brackets.

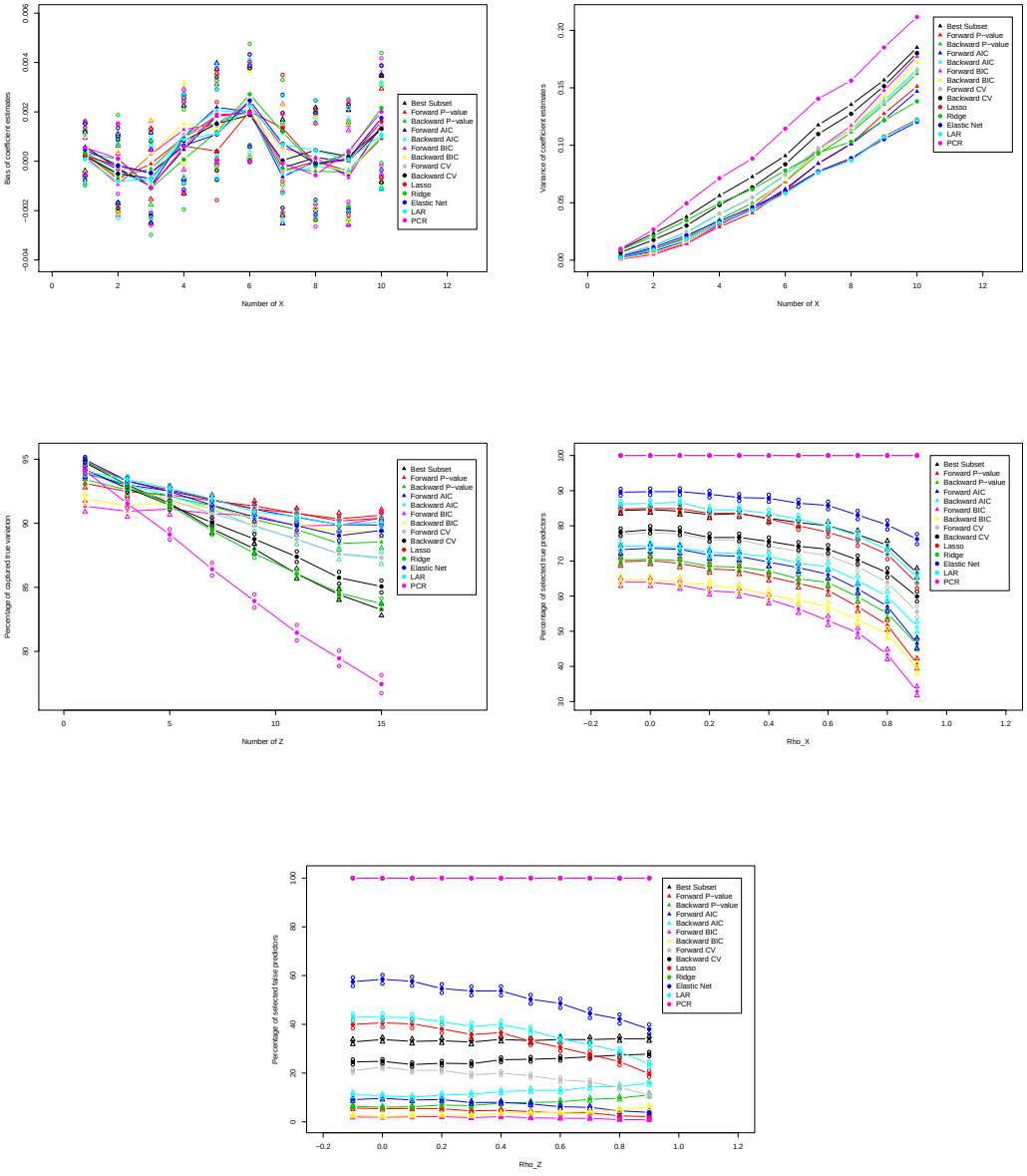

FIGURE 6.3: Comparison of Various Performance Techniques for Methods on Simulated Data of 1000 Models and when $N = 100$.

6.3 Results of Empirical Comparison of Distinct Performance Characteristics for Variable Selection, Regularization and Compression Methods on Simulated Data when Sample Size < Number of Predictors (N<p (N=15))

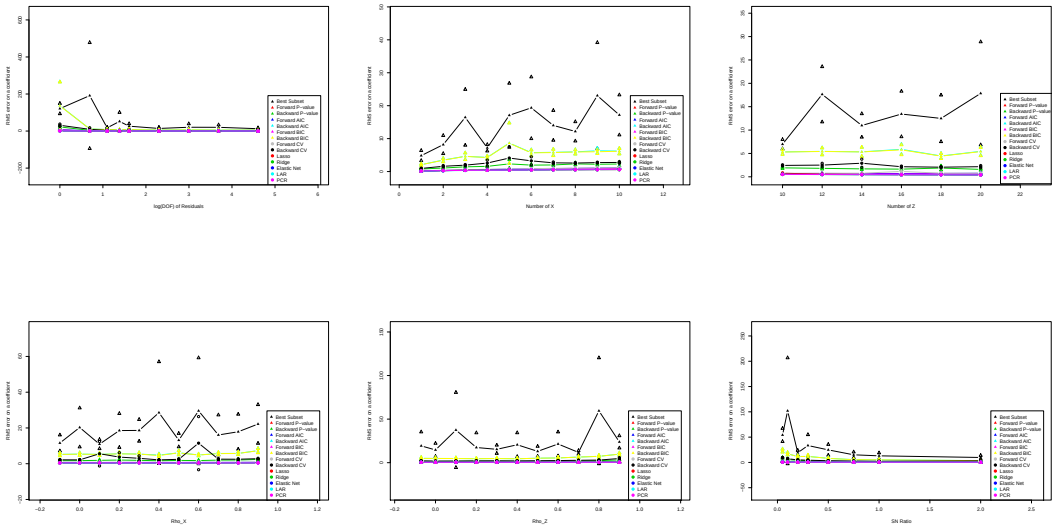

FIGURE 6.4: Comparison of Performance of the RMSE for Methods on Simulated Data of 1000 Models and when $N = 15$.

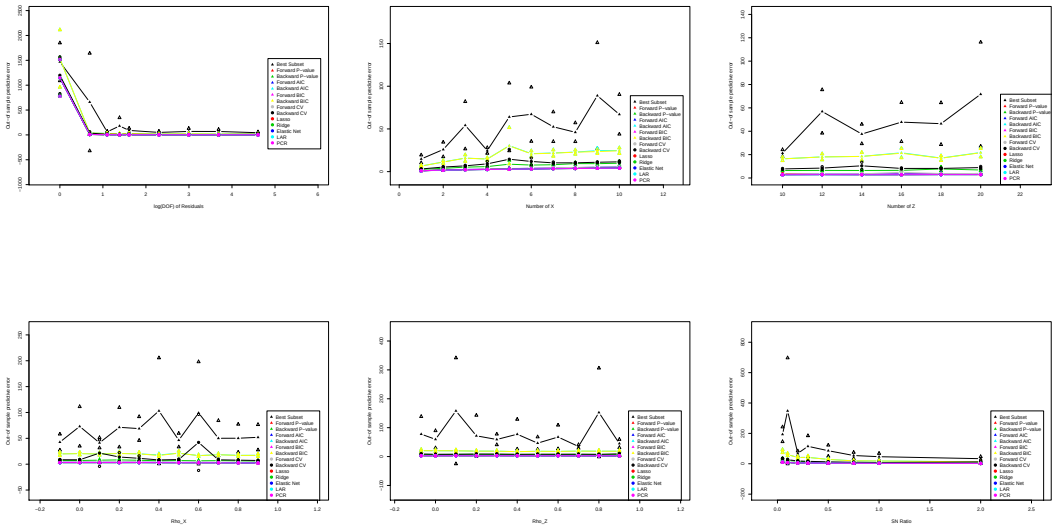

FIGURE 6.5: Comparison of Performance on the Out-Of-Sample-Predictive-Error. For Methods on Simulated Data of 1000 Models and when $N = 15$.

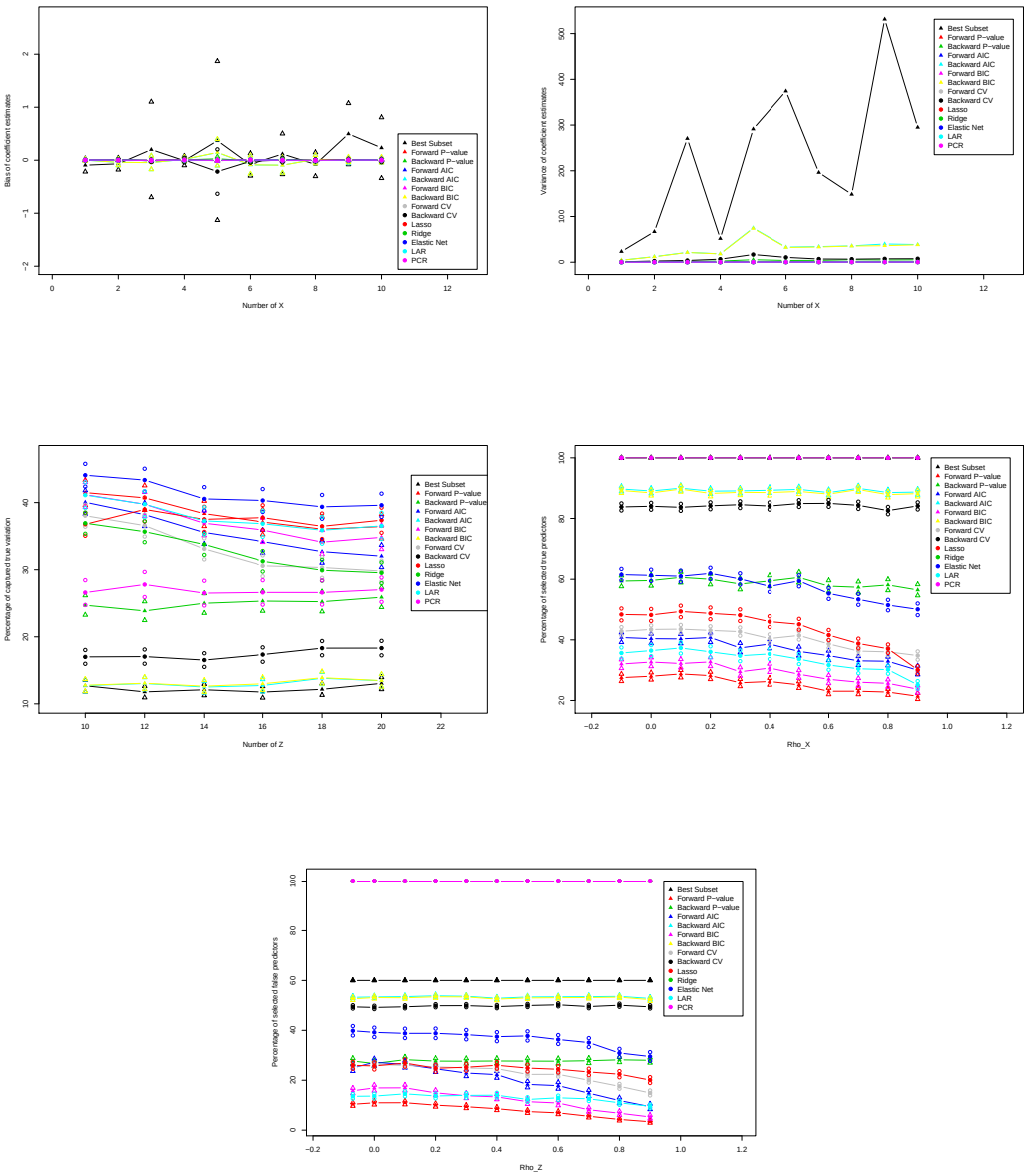

FIGURE 6.6: Comparison of Various Performance Techniques for Methods on Simulated Data Of 1000 Models and when $N = 15$.

It is wrong to analyze each block of results, corresponding to a specific setting separately. For each estimation method one can cook up an idealistic situation, where that method beats all the other methods. That is not the point, what matters is how the estimation methods perform overall: how accurate they are, how robust they are and how stable they are. So they should be compared in aggregate terms; using all the body of evidence at hand. Looking back at the generated results, we notice:

That the out of sample prediction errors of most methods are comparable. There are no clear winners, however there are some clear losers for e.g. the best subset selection method. The Lasso, Elastic Net and Least Angle Regression are good at identifying most true predictors, whenever that is possible at all. Also the Lasso, Elastic net and Least Angle Regression (LAR) are good at identifying most true predictors, whenever that is possible at all. We also notice that the Lasso, Elastic Net and LAR estimate regression coefficients relatively accurately. The principal components regression (PCR) method tends to be somewhere in the middle. Across various specifications, it exhibits inferior performance to Lasso, Elastic Net and LAR. The Best subset selection experiences complete fiasco on data sets where the sample size is small and the number of candidate predictors is bigger than the sample size.

Moving on to the "forward" and "backward" methods we see that unsurprisingly, "forward" methods perform much better than "backward" methods on data sets where the sample size is small and the number of candidate predictors is bigger than the sample size. Under extremely noisy circumstances, "forward" methods perform better than "backward" methods. However under normal circumstances,"forward" and "backward" methods are comparable. On small data sets BIC performs better than AIC. Looking at the Least Angle Regression it performs at an optimal level relative to the other methods. Lastly PCR and ridge regression should not be judged severely on the account of putting 100% of false predictors into the estimated model. By definition, they give a non-zero weight to every candidate predictor. So, formally speaking, each candidate predictor is "chosen" to be in the model. What is important is that PCR and ridge regression apply clever, soft regularization. This type of regularization shrinks the coefficients of the least informative predictors the most. And so, the resulting coefficients error and out of sample prediction error are not as bad as one would think.

We may wonder why PCR disappoints in this study. It is likely the consequence of our experiment offering an unsuitable climate to PCR. PCR assumes that there is one, two, three or more factors driving a big set of explanatory variables. In other words, there are several influential variables and almost every true predictor is highly correlated with at least one of them. On the other hand, the environment we simulate allows for quite individualistic preference. Yes, every two predictors X_i and X_j are correlated but that happens through their own, separate, individualistic mechanism. The relationship between X_i and X_j has little to do with the relationship between X_i and X_k. It does not exist because all of them are highly correlated with one major factor. In our world, X_i may "like" X_j for other reasons than why it may like X_k. Individualistic, high entropy.

6.4 Results of Empirical Exploration Comparison of Out of Sample Prediction Error for Variable Selection, Regularization and Compression Methods on Five Observational Data Sets.

Now, let us discuss illustrations based on the five real data sets. While processing real data, we sample 20 data points from each data set. This is a relatively small number, for three out of five data sets, the number of candidate predictors is larger than the sample size. For two out of five data sets, the situation is the opposite (however, the sample size is still quite small). 10-fold cross-validation is applied to estimate prediction error. Not surprisingly, the estimation performance on real data sets is largely consistent with the simulation results B (sample size $<$ number of predictors). The estimation performance is summarized in Figure 6.7 and Tables 6.3 and 6.4.

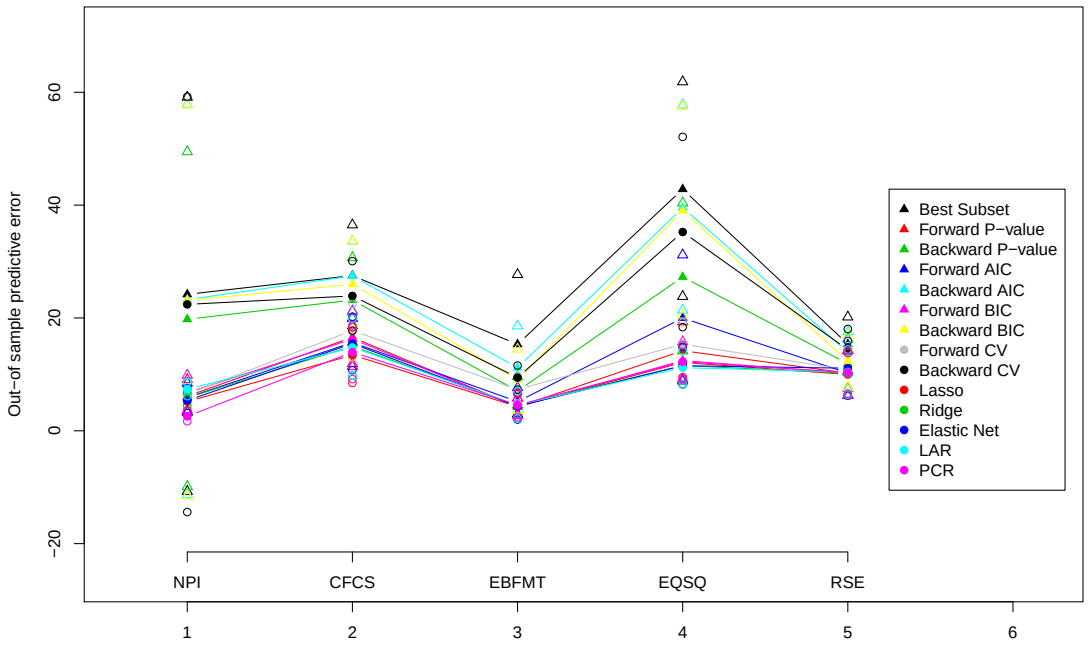

Figure 6.7: Comparison of Performance of the Out-Of-Sample-Predictive-Error for Methods on the 5 Data Sets for N=20.

Data Set	Best Subset	Forward P-value	Backward P-value	Forward AIC	Backward AIC	Forward BIC	Backward BIC
NPI	24.2 (15.4)	6.1 (1.3)	19.8 (13.1)	5.9 (1.1)	23.3 (15.3)	6.8 (1.4)	23.1 (15.3)
CFCS	27.5 (4.0)	16.6 (2.1)	23.2 (3.3)	15.6 (1.9)	27.4 (2.7)	16.3 (2.2)	26.0 (3.4)
EBFMT	15.3 (5.5)	4.2 (0.7)	7.1 (1.1)	5.2 (1.1)	11.2 (3.3)	4.2 (0.7)	9.1 (2.4)
EQSQ	42.8 (8.4)	14.2 (2.3)	27.3 (5.8)	20.0 (4.9)	39.6 (8.0)	12.5 (1.5)	39.0 (8.2)
RSE	15.4 (2.1)	10.2 (1.8)	11.9 (1.9)	10.2 (1.8)	13.9 (1.6)	10.2 (1.8)	12.3 (2.0)

Table 6.3: Cross-Validated Prediction Error on Real Data Sets for Best Subset - Backward BIC Methods; Standard Error is in the Brackets.

Data Set	Forward CV	Backward CV	Lasso	Ridge	Elastic Net	LAR	PCR
NPI	6.5 (0.9)	22.4 (16.3)	5.2 (0.7)	5.9 (0.6)	5.4 (1.0)	7.4 (0.5)	2.6 (0.4)
CFCS	17.8 (2.1)	23.9 (2.7)	13.4 (2.2)	15.0 (2.3)	15.5 (2.1)	14.8 (2.2)	13.9 (2.1)
EBFMT	7.3 (1.1)	9.4 (0.9)	4.3 (1.0)	4.5 (1.1)	4.4 (1.0)	4.6 (1.1)	4.6 (1.0)
EQSQ	15.4 (1.7)	35.2 (7.5)	12.1 (1.1)	11.6 (1.4)	11.5 (1.5)	11.2 (1.3)	12.2 (1.3)
RSE	10.8 (1.8)	14.4 (1.6)	10.0 (1.7)	10.2 (1.7)	11.1 (2.1)	10.4 (1.8)	10.3 (1.7)

Table 6.4: Cross-Validated Prediction Error on Real Data Sets for Forward CV - PCR; Standard Error is in the Brackets.

Overall, forward stepwise selection + p-value, forward stepwise selection + BIC, Lasso, ridge regression, elastic net, LAR and PCR are the best performers. For the most part, the difference in predictive performance of each two methods is not statistically significant. On the other end of spectrum, best subset selection fails completely.

Chapter 7

Conclusion and Discussion

We started off the paper by recapping the notion of multiple regression and formulated the linear regression model. From there, we introduced our problem statement and the outline of the thesis. In Variable Selection we show how we choose variables. We started off with the concept of Stepwise Selection regression which consisted of Forward Selection and Backward Elimination. The methodology behind the Stepwise regression procedure is that we build our regression model from a set of candidate predictor variables by entering and removing predictors with a critical value of entry of:$\alpha_E = 0.15$ and a critical value of exit of $\alpha_R = 0.15$. Forward Stepwise Selection is great to use since the method usually has less computation and in turn less models to analyze which is known explicitly as: $(\sum_{k=0}^{p-1}) = 1 + p(p+1)/2$ models). It has no problem for the first n-steps if $p > n$. However once an input is in, it does not get out. A second way to carry out the Stepwise selection regression is Backward Elimination. We see that again it does not require intensive computation which results in less models needed to be analyzed specifically $(\sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models).(If $p = 20$, only 211 models, are needed compared with more than 1 million models for the best subset selection). Obviously this is a very big advantage. Two drawbacks to the method is that once an input is out, it does not get it and the method is not applicable to the case with $p > n$.

Best Subset Selection is another good variable selection technique to pick the best model as it is straightforward to carry out and a conceptually clear method, However, it's often computationally infeasible since we have too many models to run. Moreover, the search space is too large ;we have (2^p models) which may lead to overfit. For instance, if $p = 20$ there are $2^{20} > 1000,000$ models. To compare regression models, some statistical software may also give values of statistics referred to as information criterion statistics. For regression models, these statistics combine information about the SSE, number of parameters in the model, and the sample size. A low value, compared to values for other possible models, is good. Some data analysts feel that these statistics give a more realistic comparison of models than the C_p statistic because C_p tends to make models seem more different than they actually are.

In our study we used the AIC, BIC and coefficient of determination and R^2 adjusted as well as the Mallow's C_p statistic via mixed integer programming approach. Which showed

us that it can be utilized to do best subset selection via a mixed integer programming approach. For the mallows C_p statistic we saw that the subset models with small C_p values have a small total (standardized) variance of prediction.

A reasonable strategy for using C_p to identify the best models is to identify subsets of predictors for which the C_p value is near p. We notice that the full model always yields $C_p = p$ so we don't select the full model based on C_p. If all models except the full model yield a large C_p not near p it suggests some important predictor(s) are missing from the analysis. In this case we are advised to identify the predictors that are missing. We then saw how the Mallow C_p statistic is useful for subset selection using a mixed integer quadratic programming approach. When $p < 30$, the method provides the best subset of variables. Moreover, when handling datasets consisting of a large number of samples it finds better quality solutions faster than stepwise regression methods do.

In the Regularization and Shrinkage chapter we started with the General Regularization techniques section which included the Tikhonov, Ivanov and Morozov methods. We started by first defining our empirical risk minimization function corresponding to our regression model. From there, we defined some regularization methods in relation to Ridge regression and the Lasso for both Tikhonov and Ivanov methods. We subsequently parameterized the functions to then show that they are equivalent. We can then formulate theory to be able to design more learning algorithms. Ridge Regression techniques then followed which explore the properties and the derivations of these properties of the estimator, The Bias, Variance, MSE etc. We see that the ridge regression estimator is biased and we also discover that the ridge estimator coefficients vanishes as the penalty parameter increases and tends to infinity. This holds for both cases in the problem statement for $p < N$ and $p > N$. Similarly, the variance of the ridge regression estimator vanishes as the tuning parameter goes to infinity. The trace of the MSE is realized to be a convex function. A theoretical exploration of the K fold cross validation follows which was then applied to 5 observational data sets to the 14 methods explored later in chapter 6.

We then explored 3 different shrinkage methods which were: The method of the Lasso, Bayesian Lasso and Elastic net. The method of the Lasso is suggested when dealing with an ill conditioned model matrix $\boldsymbol{X}$ for case $p > n$. When $p > n$ the Lasso provides a better variable selection method than Ridge regression. The Lasso provides a sparse solution by penalizing the sum of the absolute values of the estimates. As λ increases the number of significant coefficients decreases. Hence this makes the Lasso for variables selection and interpretation of the results a more plausible method than Ridge regression. We then partially answer the question "When is the lasso solution well-defined (unique)?" When reviewing results from the literature, we see that if the predictor variables are drawn from a continuous probability distribution then there is a unique Lasso solution with probability one regardless of the sizes of n and p We also show that this result extends easily to ℓ_1 penalized minimization problems over a wide range of loss functions. Another Shrinkage method explored is the Bayesian Lasso where a connection with the inverse Gaussian distribution is made which provides tractable full conditional distributions. The Bayesian Lasso provides interval estimates that can guide variable selection. Moreover the structure of the hierarchical model provides both Bayesian and likelihood methods for selecting the Lasso parameter. We then delved into the theory of bridge regression. In the elastic net section of the chapter we perceive the notion of the grouping effect and how it is useful when there are strongly correlated predictors in the model. The elastic net is theoretically presumed to

perform better than the Lasso for the $p >> n$ whilst still keeping a similar representation of sparsity.

In the compression chapter we investigate the theory of eigenvector decomposition as well as singular value decomposition and it's relation to PCA and dimensionality reduction. Starting with the SVD, one of the most common applications we find is obtaining a low rank approximation to a matrix. This is used for compression, speed up and also actual data analysis. Similarly, for Eigenvector (Spectral) Decomposition we see that it is useful because it gives us the ability to efficiently raise a matrix to a large power. For this and many other reasons it's used heavily in engineering to efficiently analyze and predict the behaviour of a linear dynamical system at a future point in time. The two are equal when $A \succeq 0 \wedge A^\top = A$ is symmetric positive semidefinite. This can be denoted as A which is the set of matrices $A \in \mathbb{R}^{n \times n}$ that satisfy $A = b^\top B$. For SPCA hard thresholding we find that the proposed method can be implemented by linear operators and thus is computationally efficient even in the case where $p >> n$ or large p scenarios. The method shows the superiority compared to the L_1 penalized method which makes the method a strong rival of the existing sparse PCA. Lastly, RPCA's use can be extended to data compression, more specifically the suggested weighted RPCA approach to estimate the data principal components under outlier sampling. By modeling the MLE of residues, the LS problem was formulated as a weighted ridge regression expression. With an iteratively regression processing, a better estimated performance was obtained compared with ℓ_2- PCA and ℓ_1-PCA methods. For the estimation of the loading matrix $\boldsymbol{B}$, we find that it is computationally expensive.

The Principal Component Regression method was explored empirically and we found in our simulated study that:

The PCR with only one principal component does not perform well. However, the first thing we notice is that the estimation and prediction performance of PCR improves with r increasing if the signal to noise ratio is substantial. By construction there are two independent themes going on: the correlated behavior of $X's$ and the correlated behavior of Z's. PCR cannot capture the two themes with just one principal component. It needs at least two. For that reason PCR with only one principal component performs much worse than its competitors overall. We noted that ρ_Z increases the correlated behavior of $Z's$ become more and more important. At some point, the only allowed principal component becomes dedicated to $Z's$ almost fully almost completely ignoring the information contained in the true predictors (X's). As a result, the estimated model does not have much of explanatory or predictive power. The higher is the correlation among $X's$ and or $Z's$ the fewer components we need to do a relatively good job. The bigger the true model is, the harder it is to estimate and forecast.

We then carried out an extensive simulation study for $n > p$, $n < p$ and for 5 real data sets comparing the performance of 7 characteristics performance methods on 14 methods. For the simulation study we carried out for 1000 models for the $n < p$ and $n > p$. We use simulated data for our simulation study and k fold cross validation for our real data sets. We concluded that the out of sample prediction error of most methods was comparable and that there were no clear winners but certainly there were clear losers. The best subset selection was the worst method out of the 14 methods. Best Subset Selection performed very poorly for case two where N<p where the sample size is small and the number of candidate

predictors is bigger than the sample size.

The best method in the study was the LARs in several categories, moreover Lasso, Elastic net and least angle regression (LAR) are good at identifying most true predictors whenever that is possible at all. The Lasso, Elastic net and LAR estimate regression coefficients relatively accurately. The Principal Components Regression (PCR) tends to be somewhere in the middle. Across various specifications it exhibits inferior performance to Lasso, elastic net and LAR. The PCR in our simulated experiment did not do well, this is because there are several influential variables and almost every true predictor is highly correlated with at least one of them. Every two predictors X_i and X_j are correlated in their own individualistic mechanism. For our 5 real data sets while processing real data, we sample 20 data points from each data set which is a relatively small number. 10-fold cross-validation is applied to estimate prediction error. As was expected, the estimation performance on real data sets is largely consistent with the simulation results for the case of $N < p$.

Appendix A

Complete Set of Graphs from the Simulation Study with $N = 100$

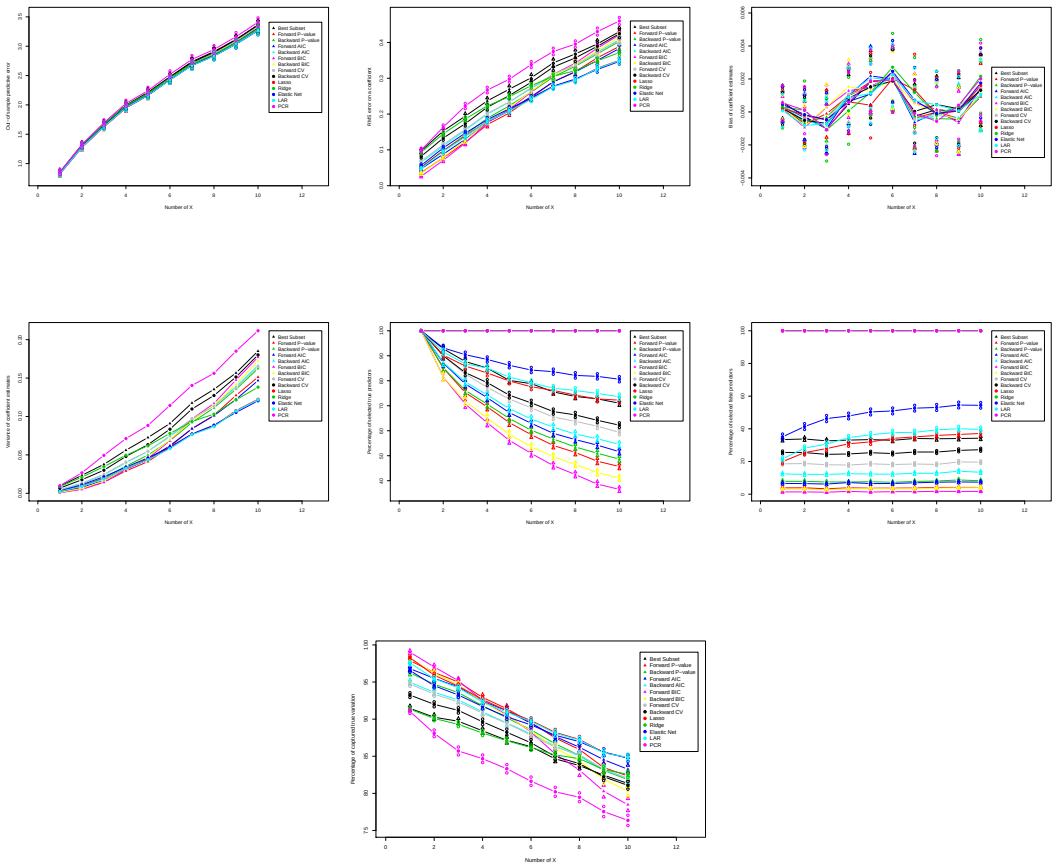

FIGURE A.1: Variability of the Seven Performance Characteristics Over the Number of True Predictors.

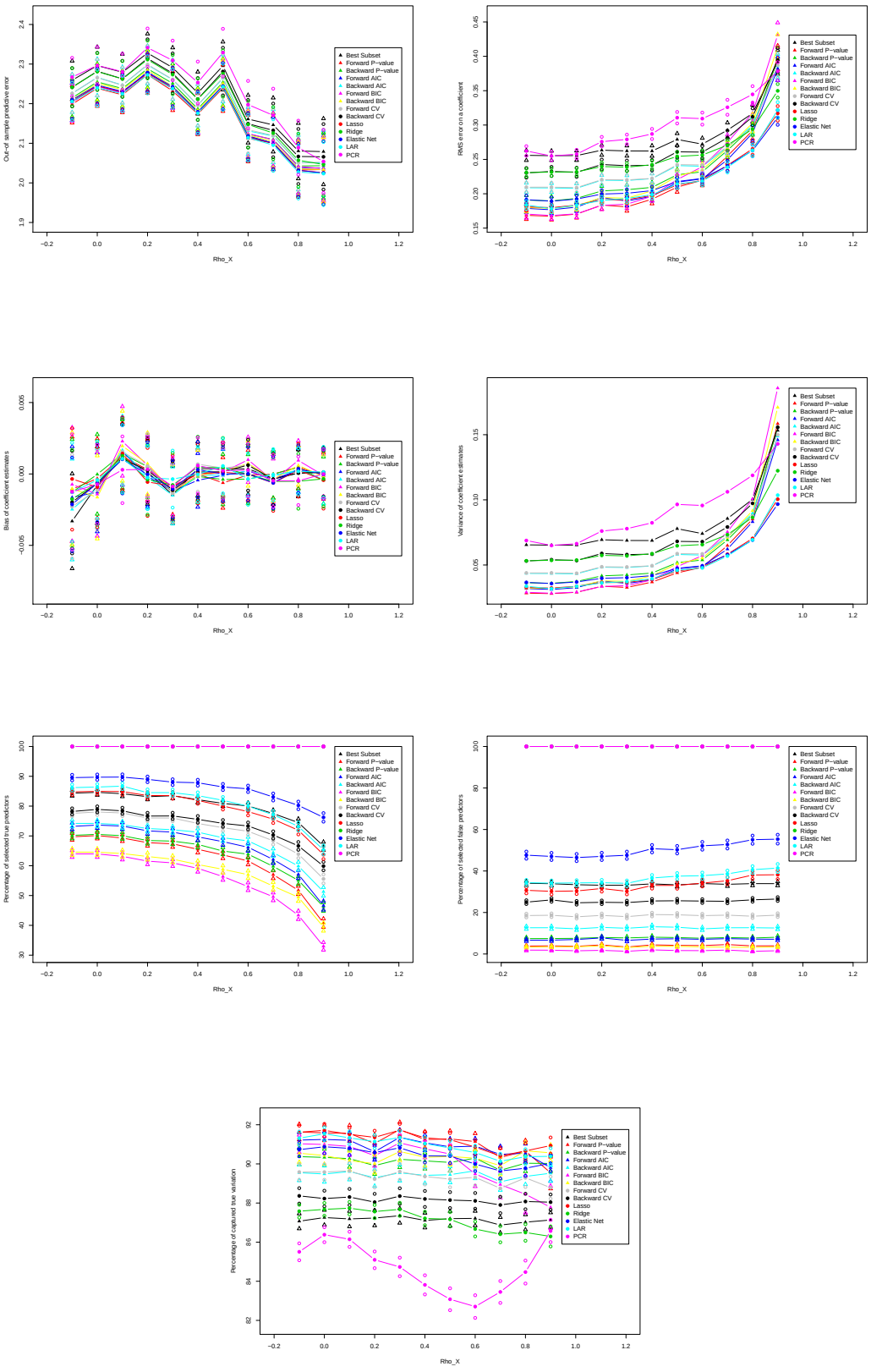

FIGURE A.2: Variability of the Seven Performance Characteristics Over the Correlation of True Predictors.

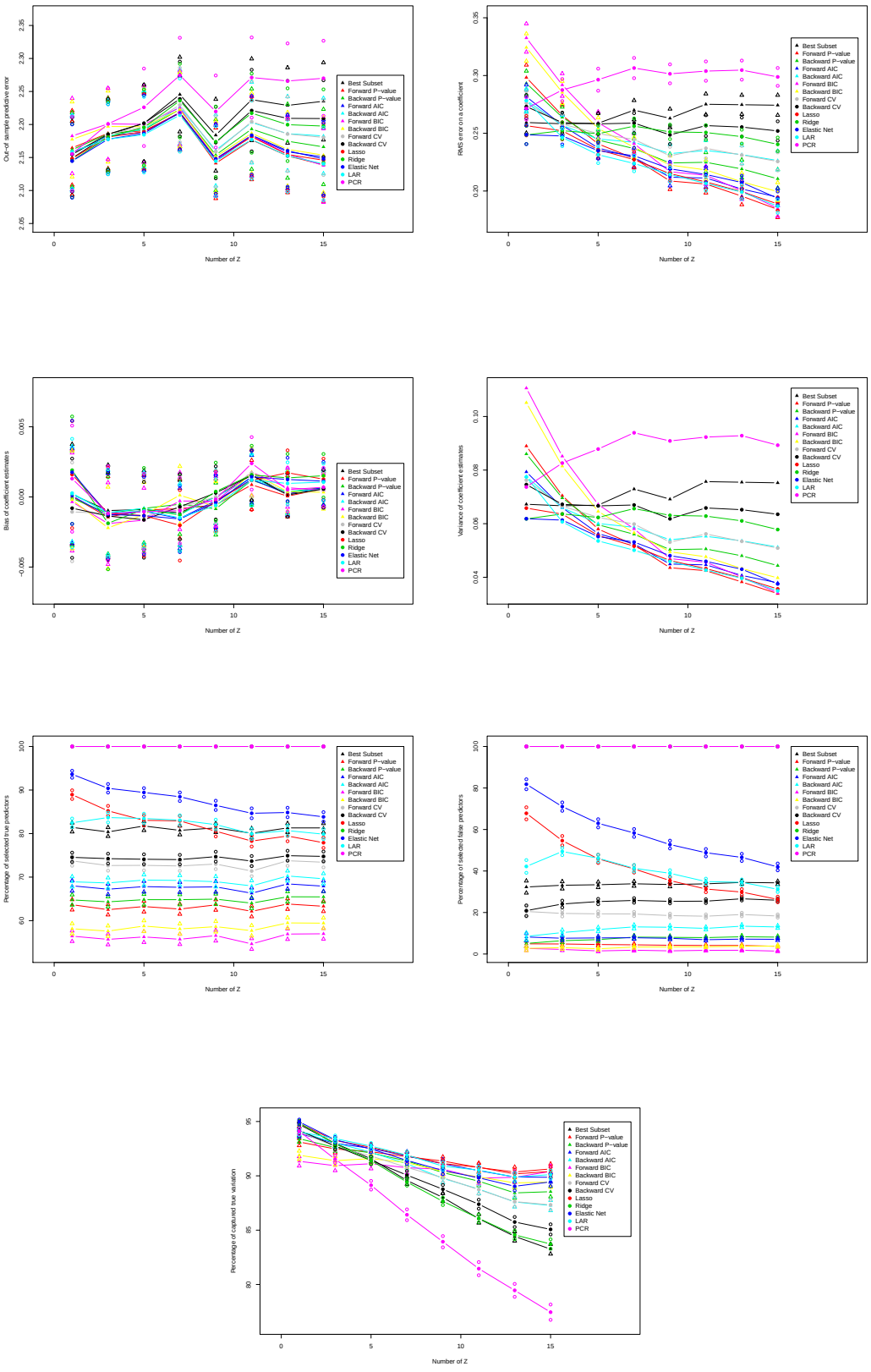

FIGURE A.3: Variability of the Seven Performance Characteristics Over the Number of False Predictors.

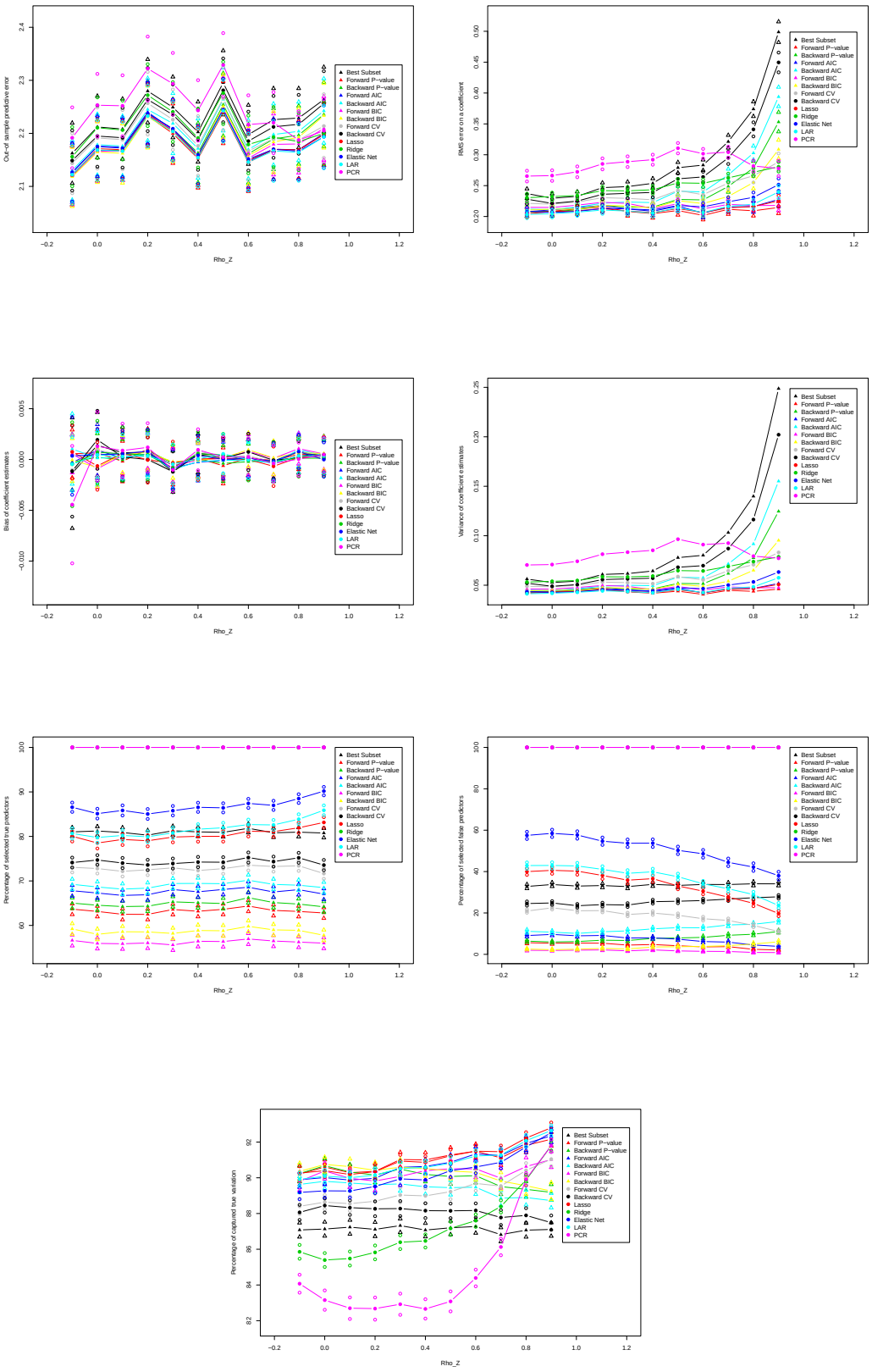

FIGURE A.4: Variability of the Seven Performance Characteristics Over the Correlation of False Predictors.

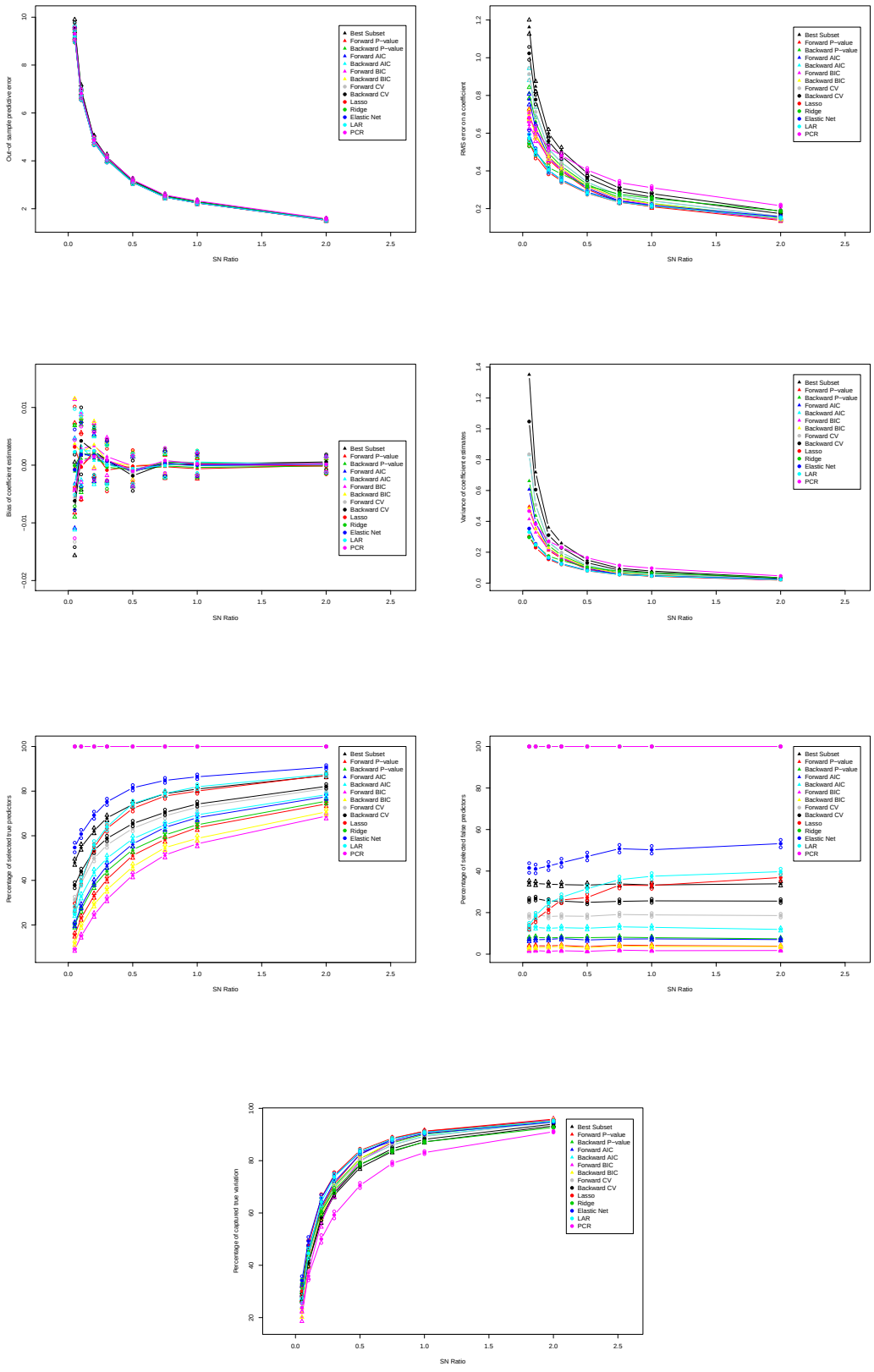

FIGURE A.5: Variability of the Seven Performance Characteristics Over the Signal-To-Noise Ratio.

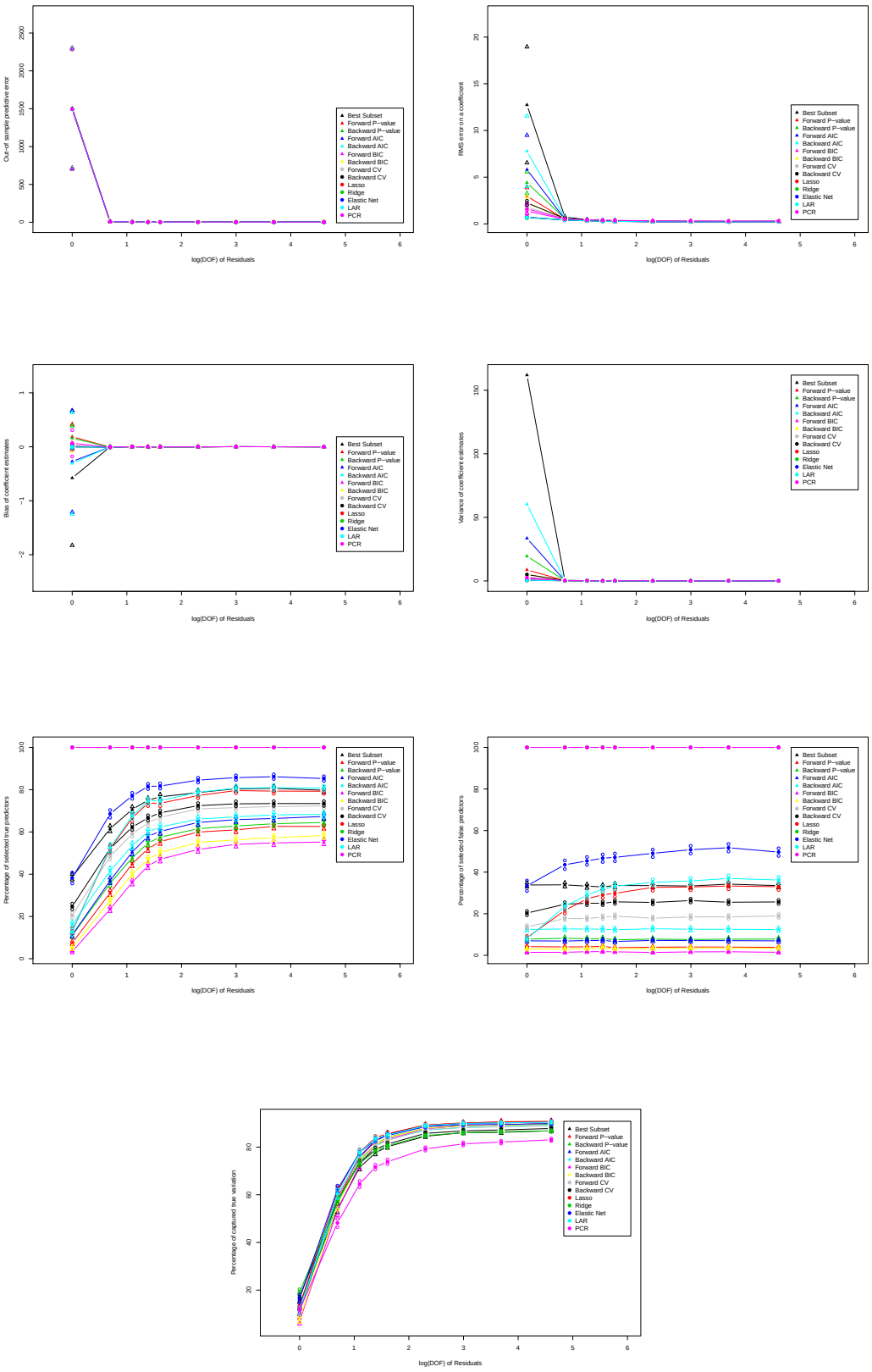

FIGURE A.6: Variability of the Seven Performance Characteristics Over the Degrees of Freedom of Residuals (Tail Fatness of Residuals).

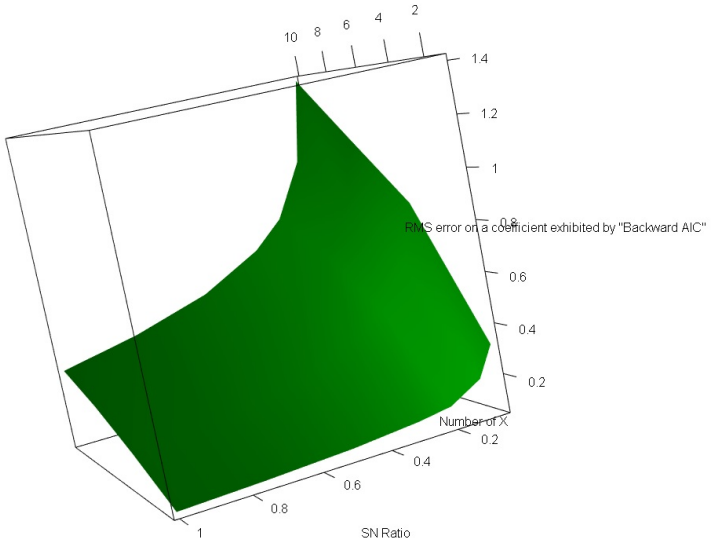

FIGURE A.7: Performance of the RMSE and SNR for the Backward Stepwise Regression Governed by the AIC Method on Simulated Data of 1000 Models and when $N = 100$

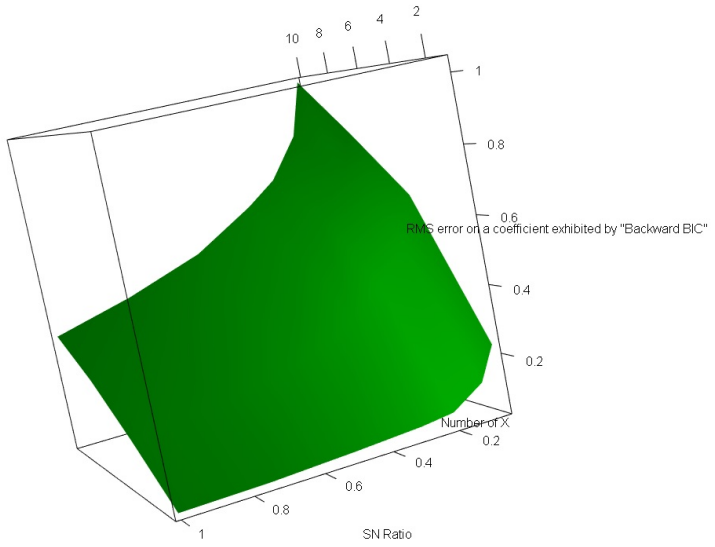

FIGURE A.8: Performance of the RMSE and SNR for the Backward Stepwise Regression Governed by the BIC Method on Simulated Data of 1000 Models and when $N = 100$

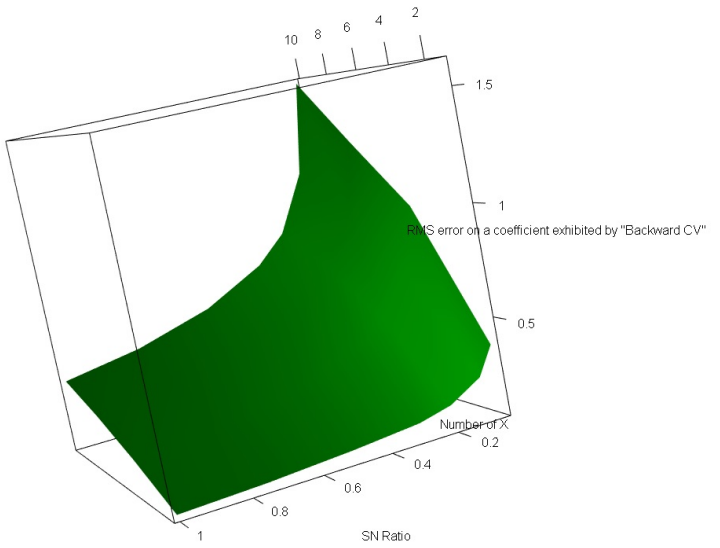

Figure A.9: Performance of the RMSE and SNR for the Backward Stepwise Regression Governed by the CV Method on Simulated Data of 1000 Models and when $N = 100$

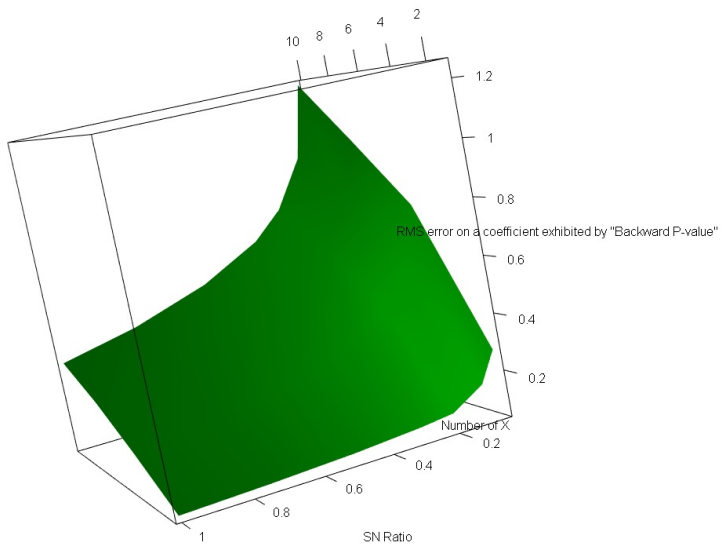

Figure A.10: Performance of the RMSE and SNR for the Backward Stepwise Regression Governed by p-values and the Significance Level of 5% on Simulated Data of 1000 Models and when $N = 100$

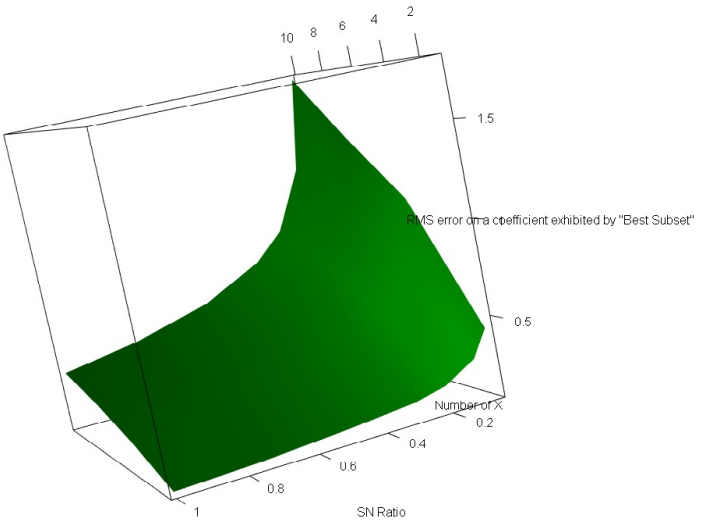

FIGURE A.11: Performance of the RMSE and SNR for the Best Subset Regression Governed by the Best Subset Selection Method on Simulated Data of 1000 Models and when $N = 100$

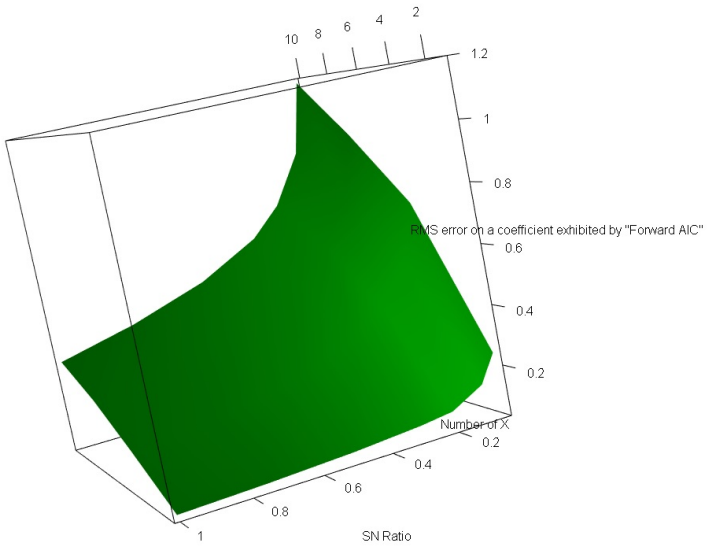

FIGURE A.12: Performance of the RMSE and SNR for the Forward Stepwise Regression Governed by the AIC Method on Simulated Data of 1000 Models and when $N = 100$

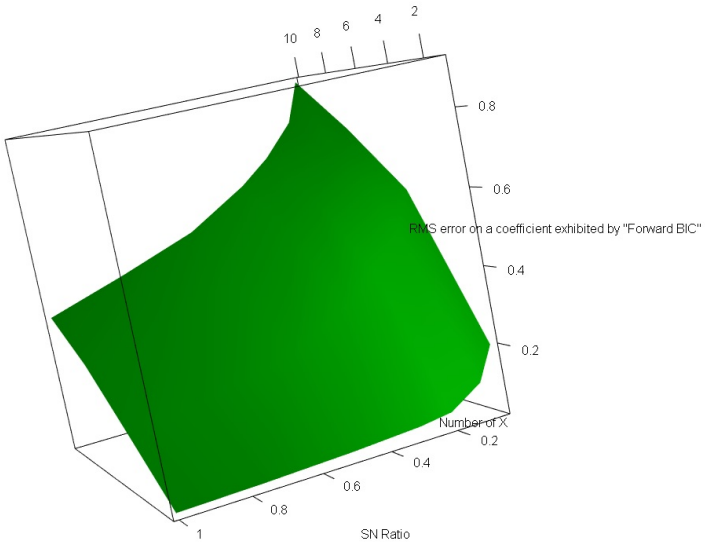

FIGURE A.13: Performance of the RMSE and SNR for the Forward Stepwise Regression Governed by the BIC Method on Simulated Data of 1000 Models and hen $N = 100$

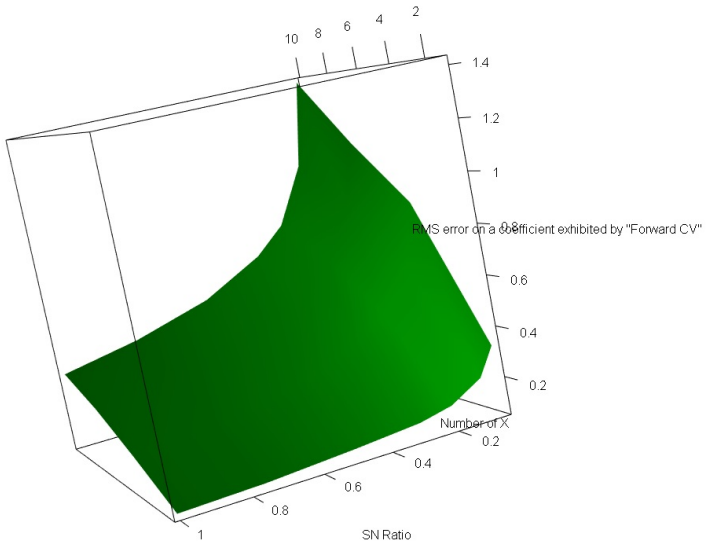

FIGURE A.14: Performance of the RMSE and SNR for the Forward Stepwise Regression Governed by the CV Method on Simulated Data of 1000 Models and when $N = 100$

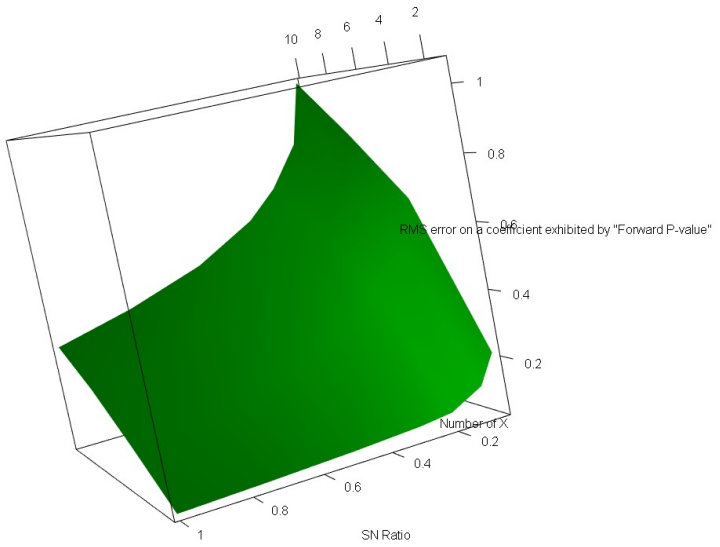

FIGURE A.15: Performance of the RMSE and SNR for the Forward Stepwise Regression Governed by p-values and the Significance Level of 5% on Simulated Data of 1000 Models and when $N = 100$

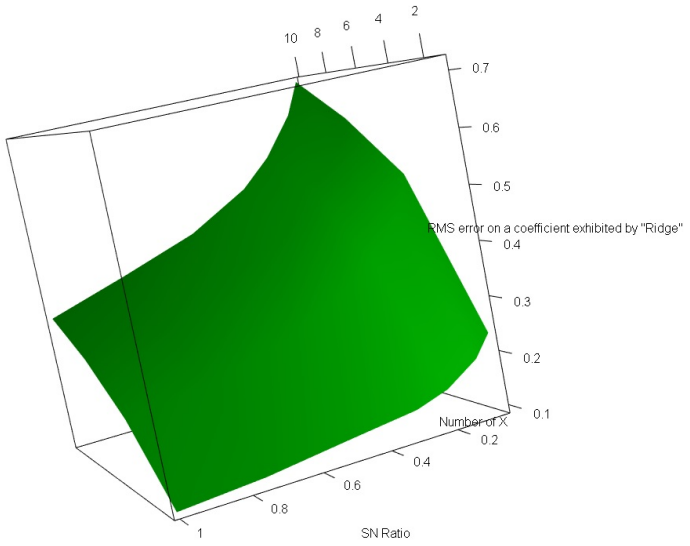

Figure A.16: Performance of the RMSE and SNR for the Ridge Regression Method on Simulated Data of 1000 Models and when $N = 100$.

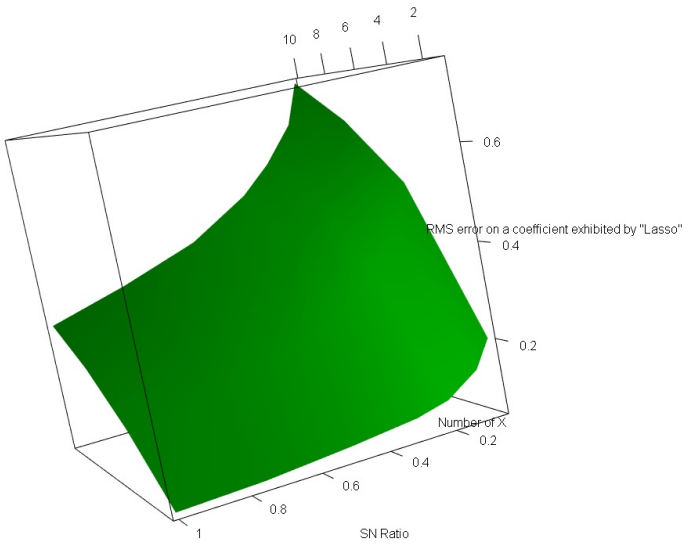

Figure A.17: Performance of the RMSE and SNR for the Lasso Method on Simulated Data of 1000 Models and when $N = 100$.

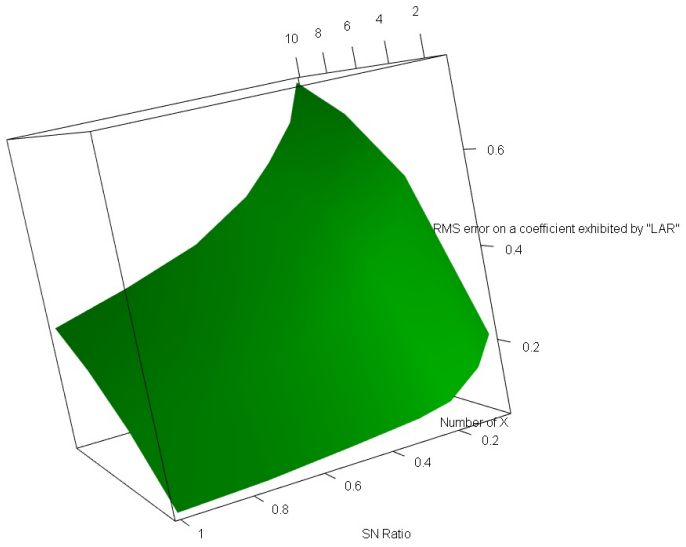

FIGURE A.18: Performance of the RMSE and SNR for the LAR Method on Simulated Data of 1000 Models and when $N = 100$.

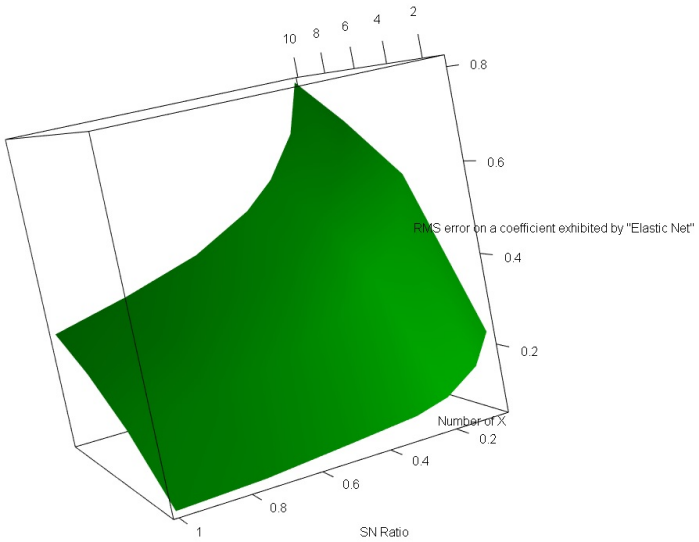

FIGURE A.19: Performance of the RMSE and SNR for the Elastic Net Method on Simulated Data of 1000 Models and when $N = 100$.

Appendix B

Complete Set of Graphs from the Simulation Study with $N = 15$

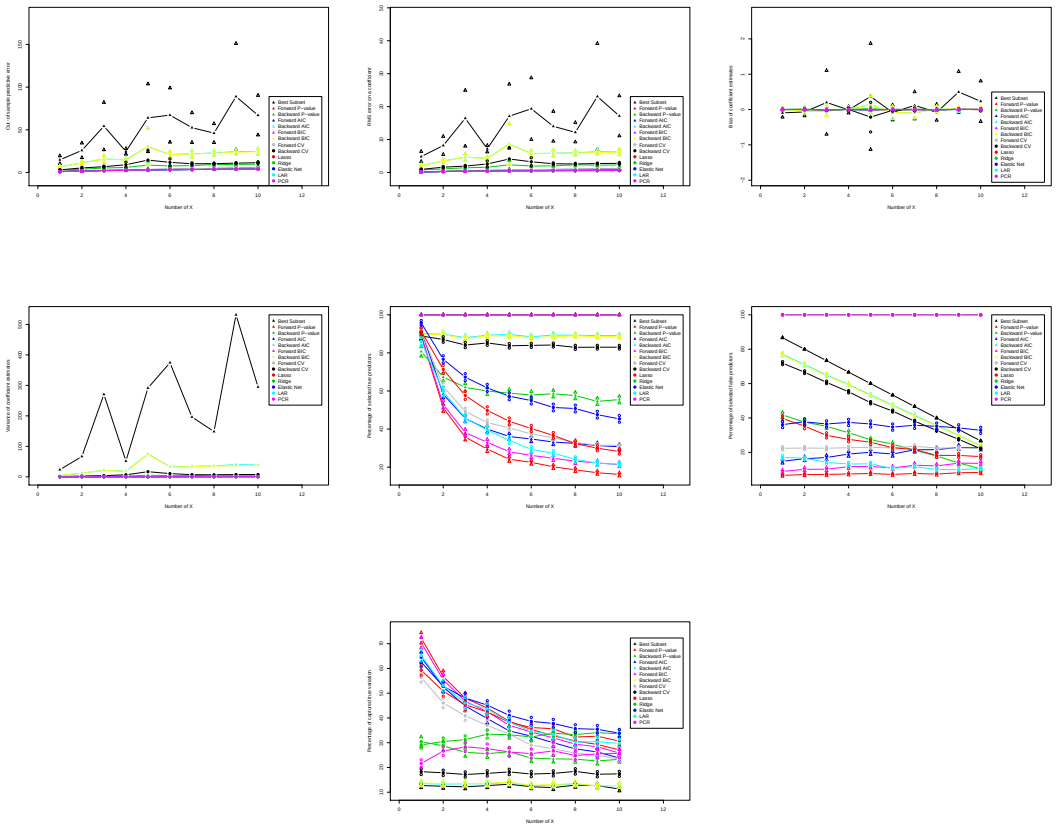

FIGURE B.1: Variability of the Seven Performance Characteristics Over the Number of True Predictors.

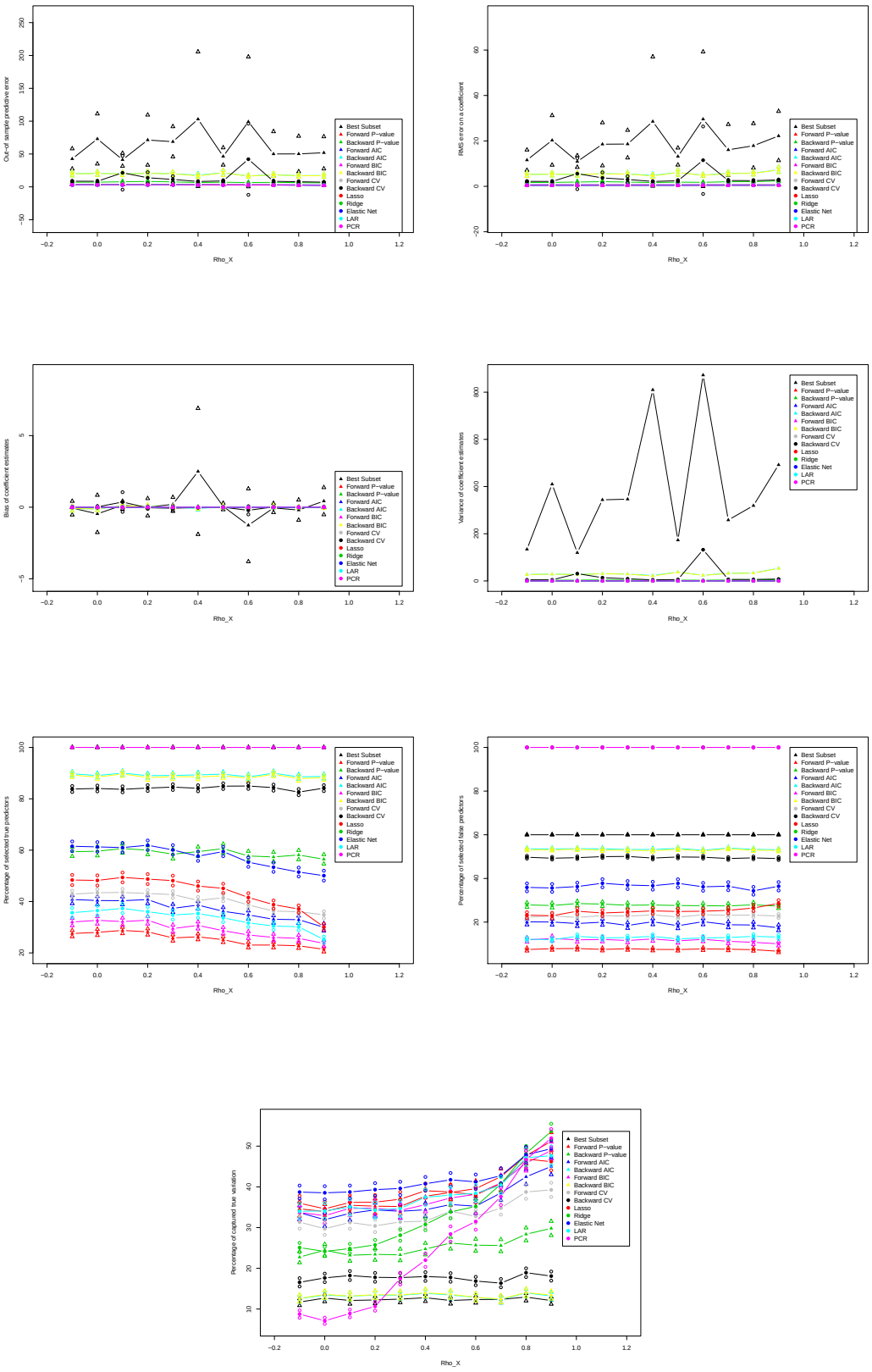

FIGURE B.2: Variability of the Seven Performance Characteristics Over the Correlation of True Predictors.

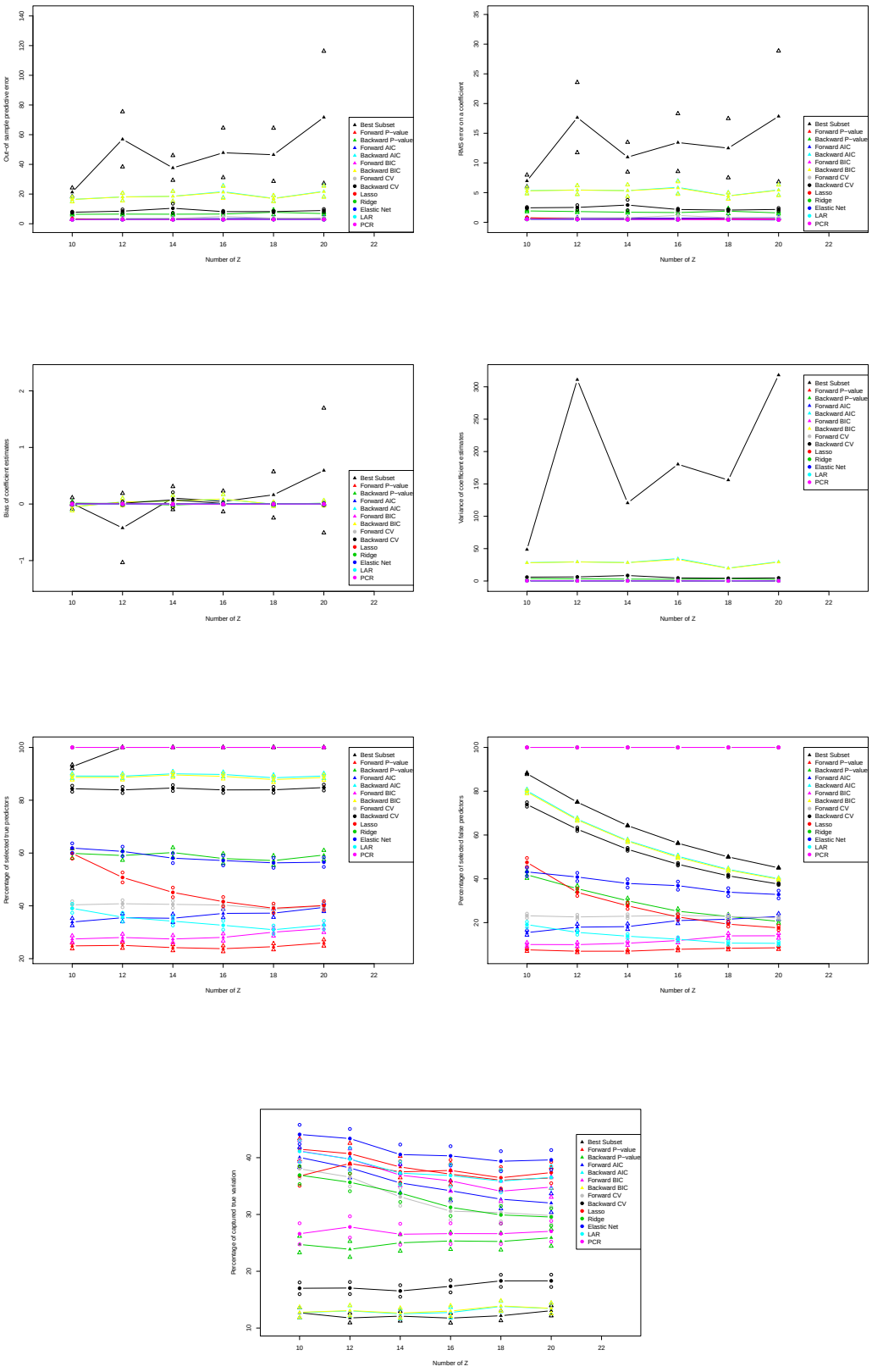

FIGURE B.3: Variability of the Seven Performance Characteristics Over the Number of False Predictors.

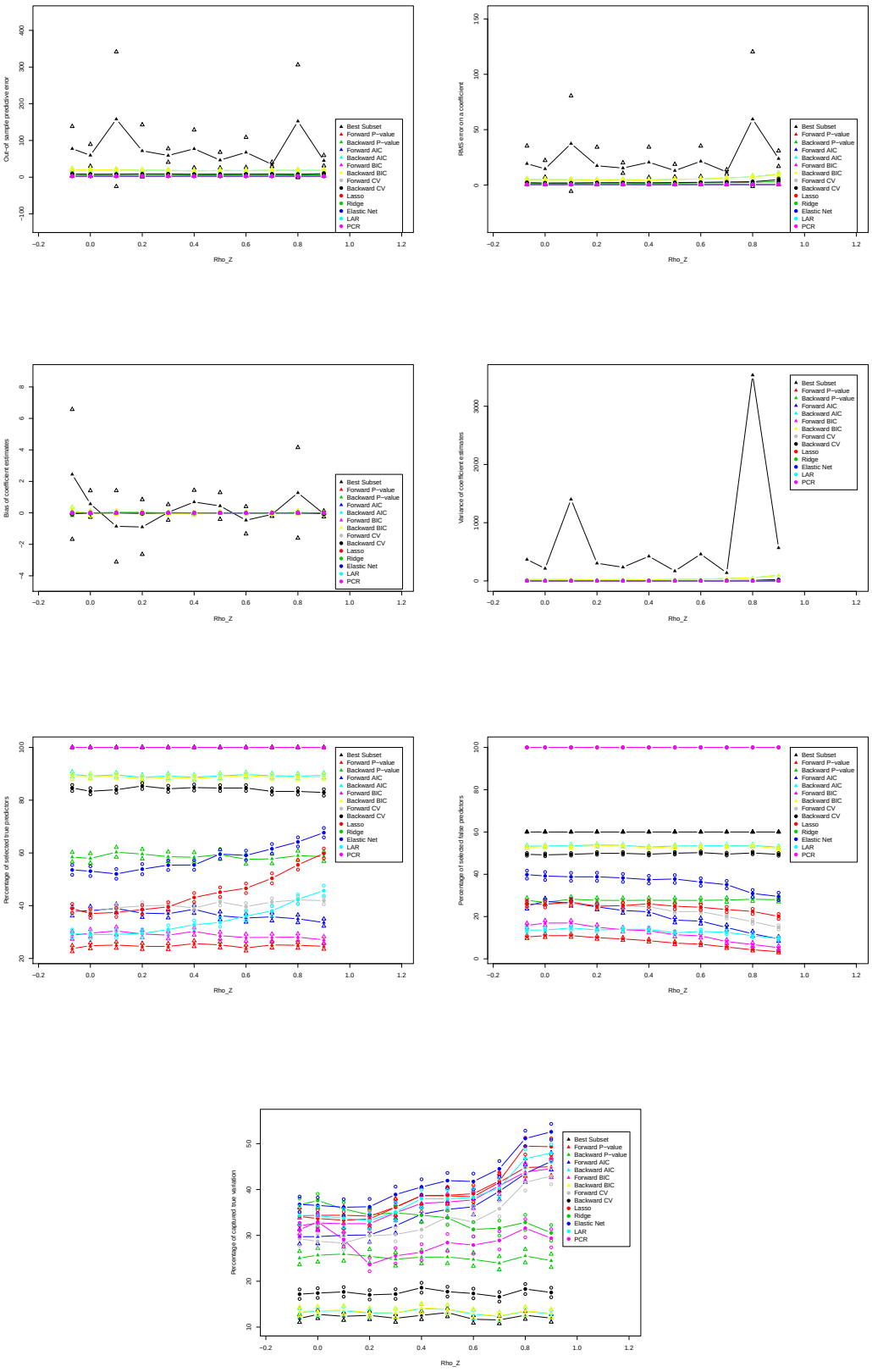

FIGURE B.4: Variability of the Seven Performance Characteristics Over the Correlation of False Predictors.

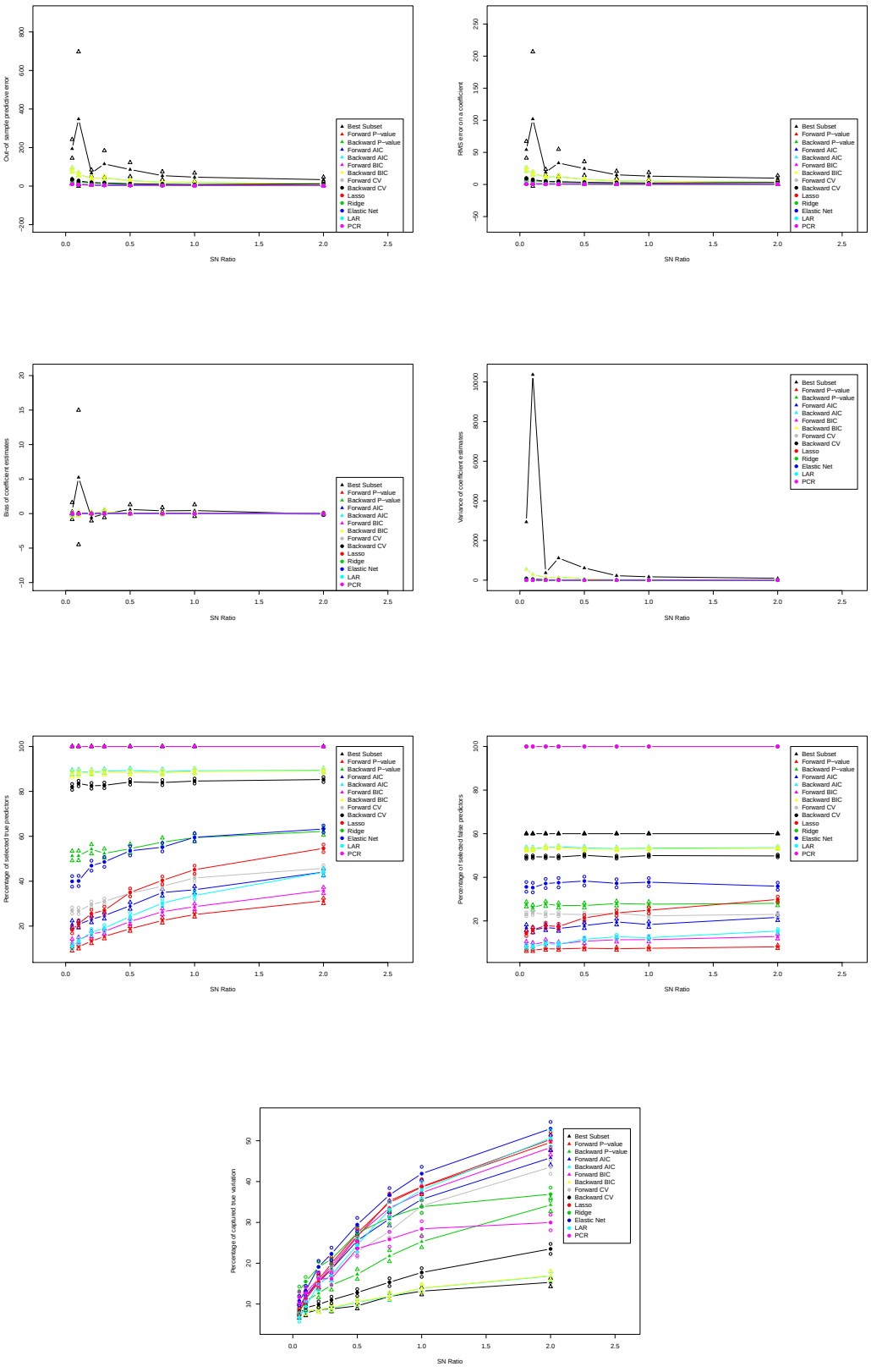

FIGURE B.5: Variability of the Seven Performance Characteristics Over the Signal-To-Noise Ratio.

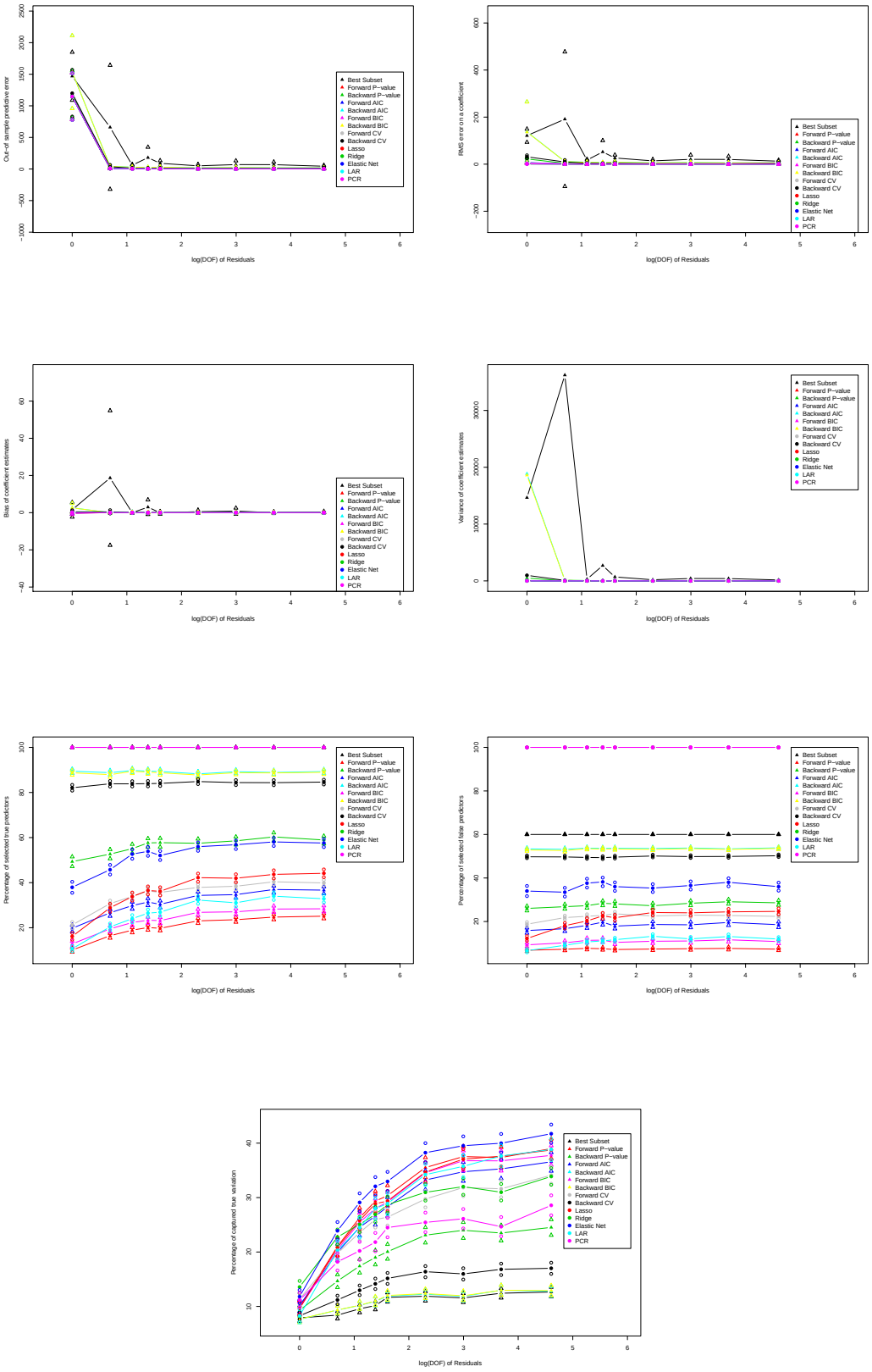

FIGURE B.6: Variability of the Seven Performance Characteristics Over the Degrees of Freedom of Residuals (Tail Fatness of Residuals).

Appendix C

Complete Set of Graphs from the PCR Simulation Study with $N = 100$

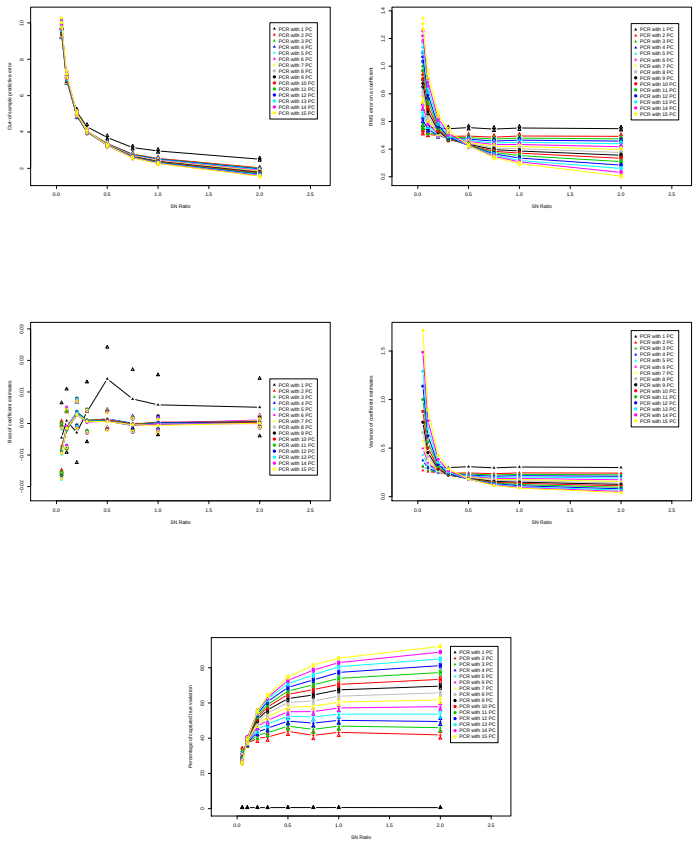

FIGURE C.1: Variability of the Five Performance Characteristics Over the SNR for Different Values of PCs.

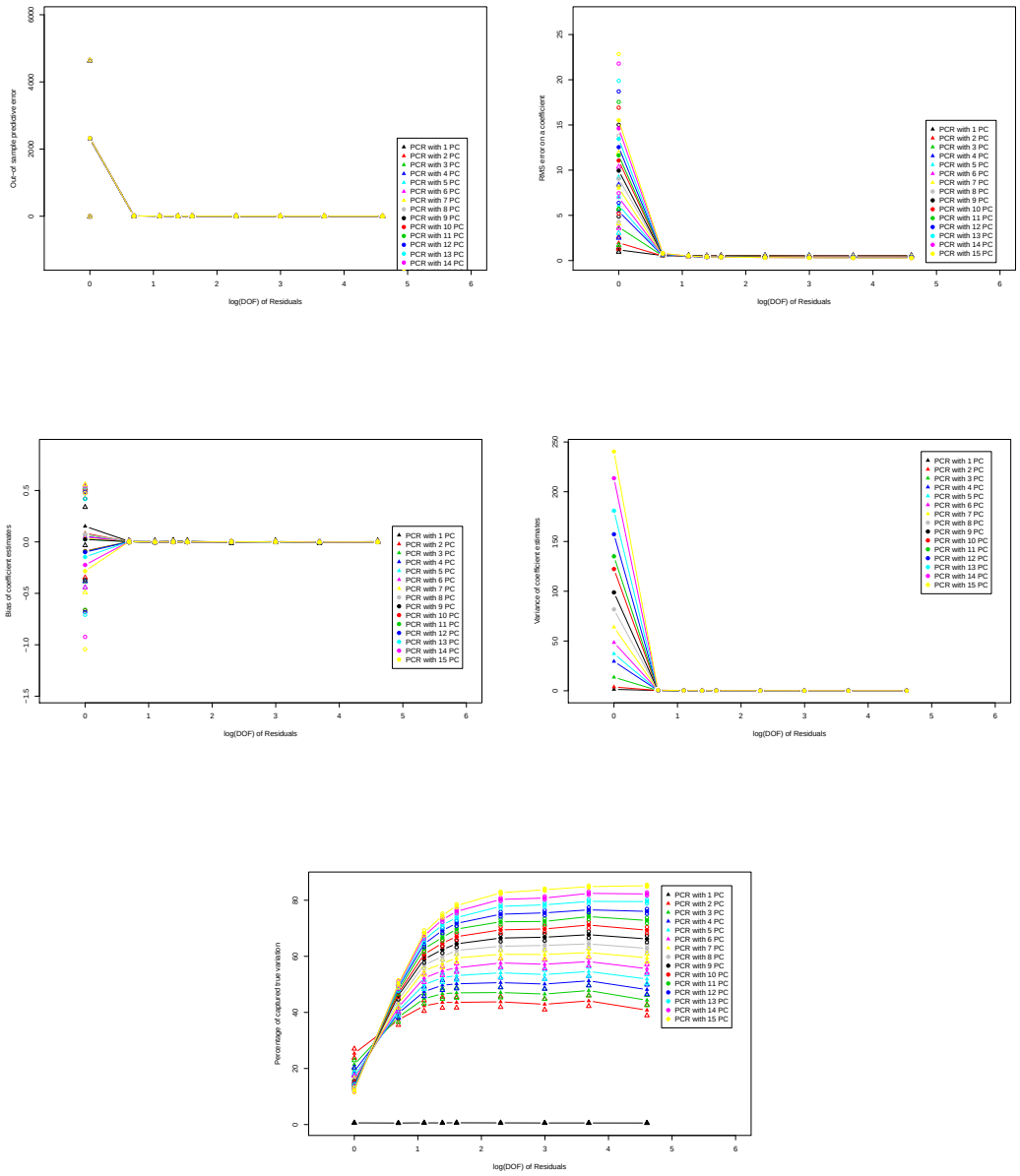

FIGURE C.2: Variability of the Five Performance Characteristics Over the log(DOF) for Different Values of PCs.

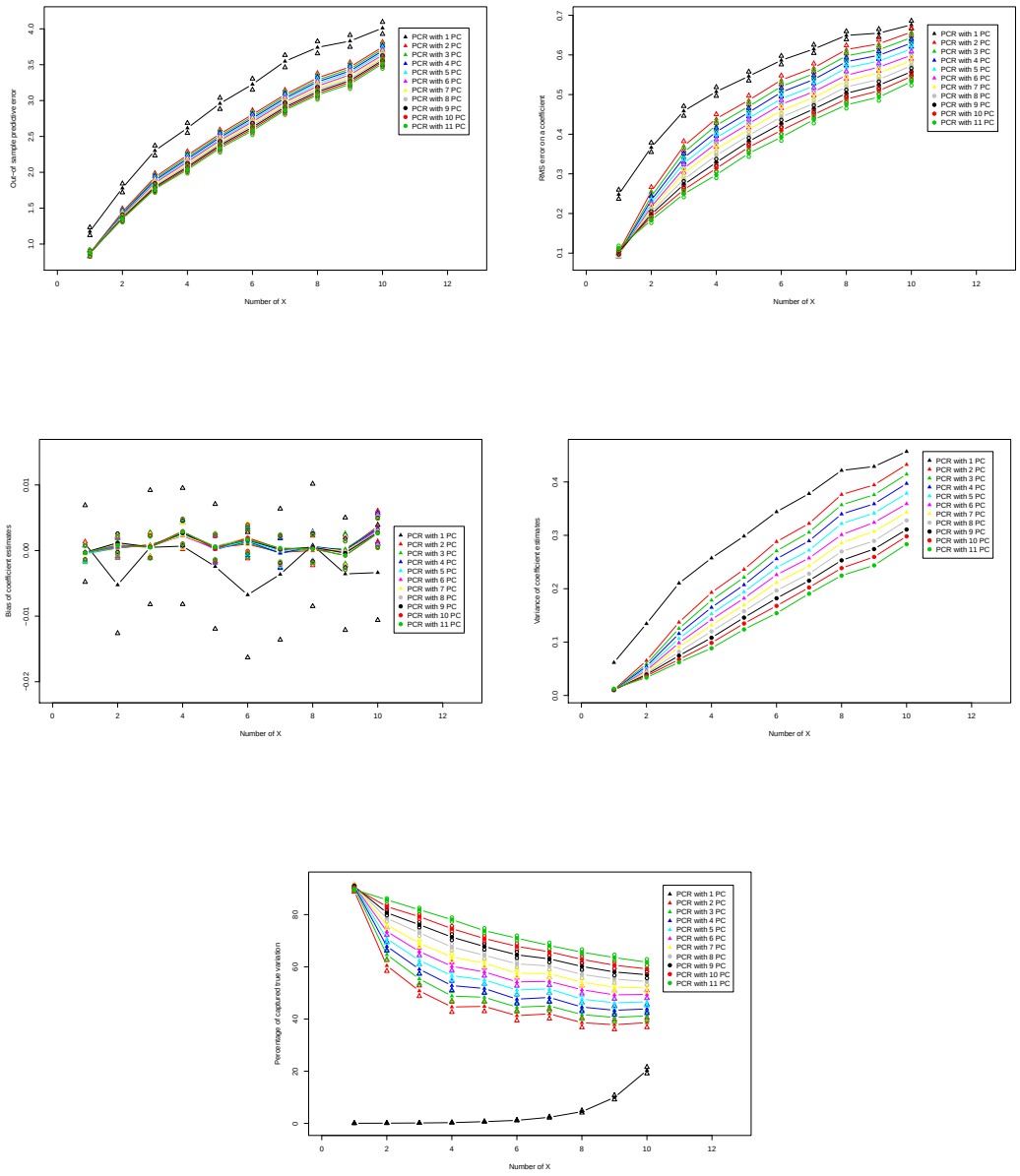

FIGURE C.3: Variability of the Seven Performance Characteristics Over the Number of True Predictors.

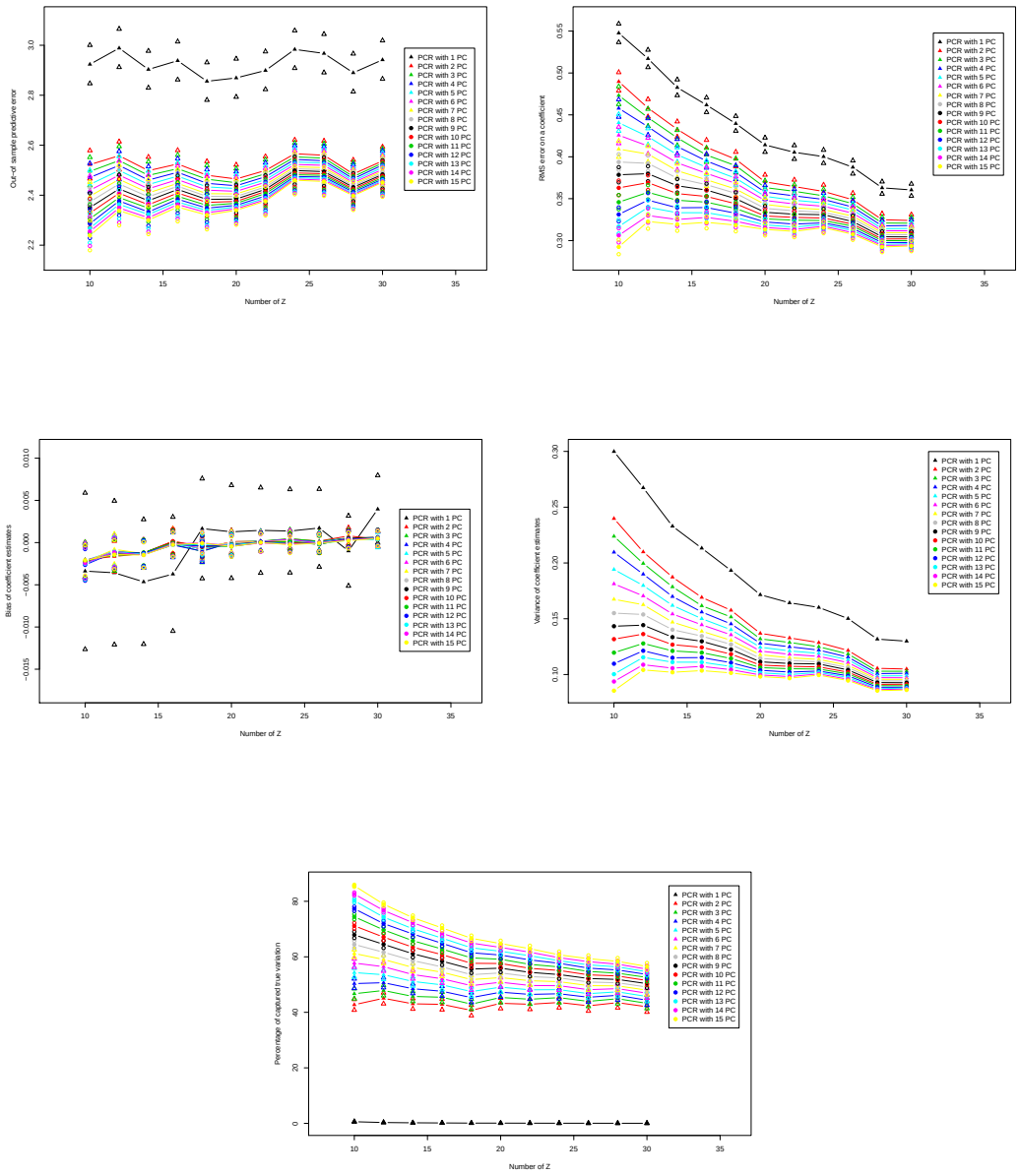

FIGURE C.4: Variability of the Five Performance Characteristics Over the Number of False Predictors for Different Values of PCs.

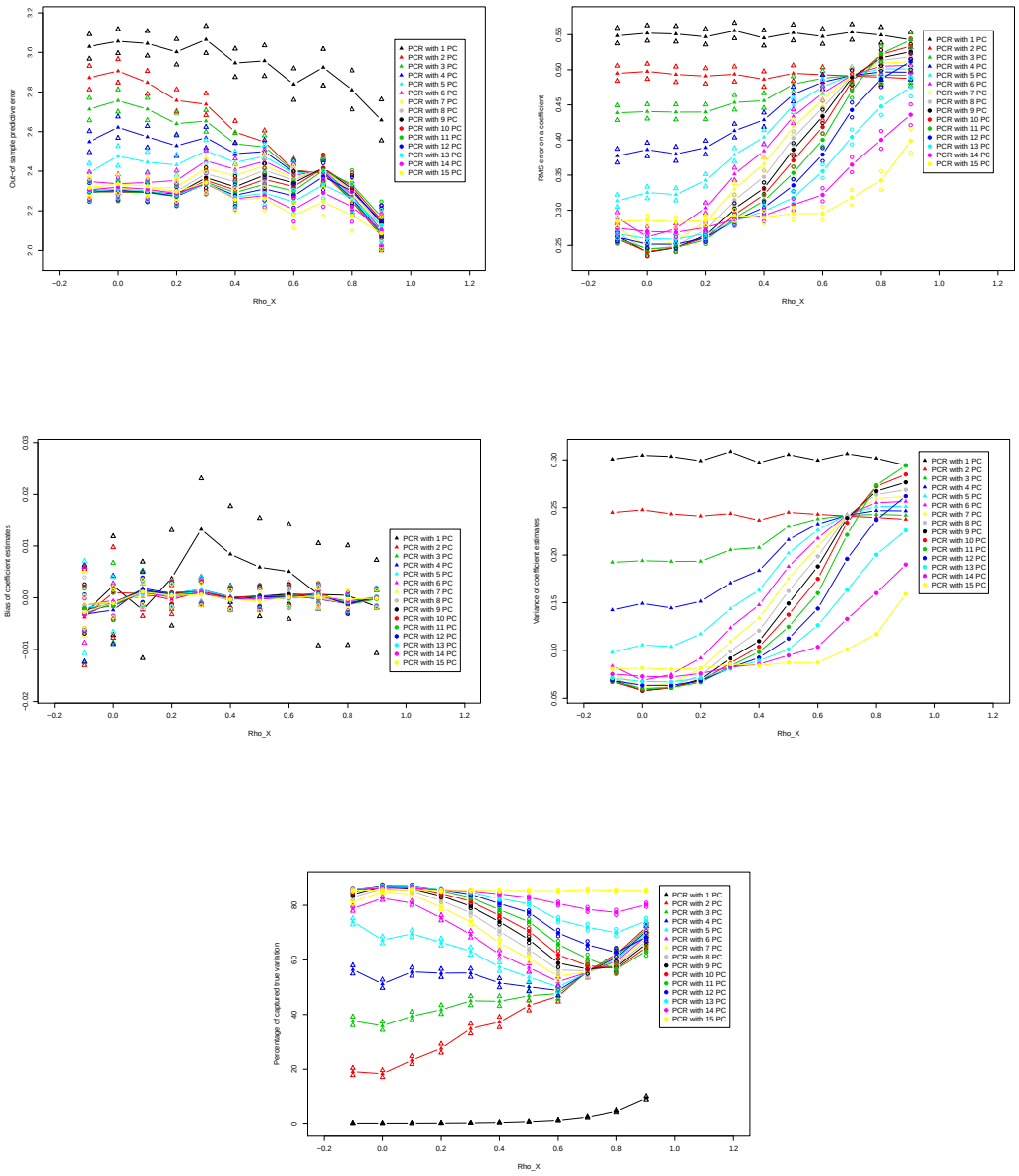

FIGURE C.5: Variability of the Five Performance Characteristics Over the Correlation of True Predictors for Different Values of PCs.

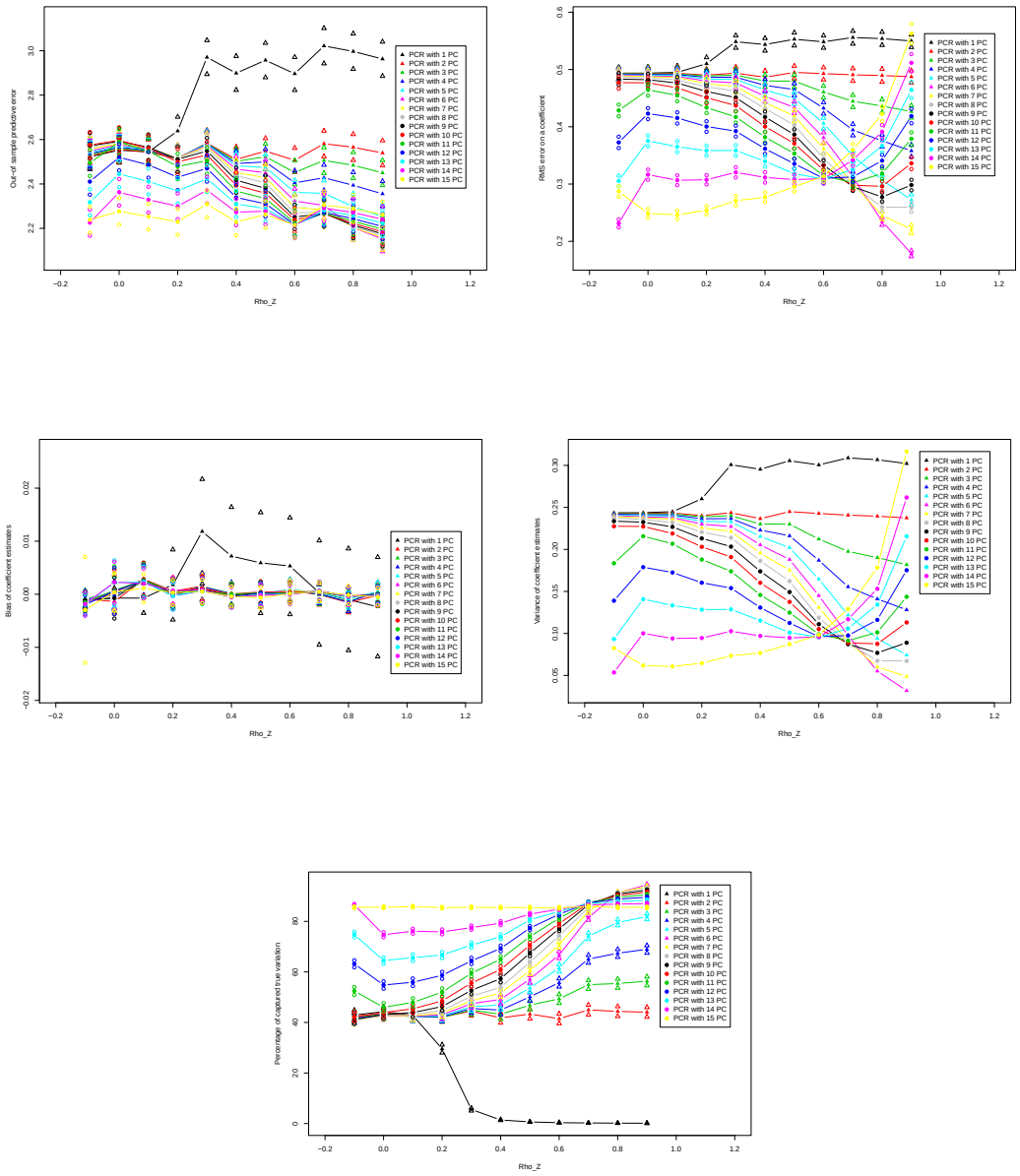

FIGURE C.6: Variability of the Five Performance Characteristics Over the Correlation of False Predictors for Different Values of PCs.

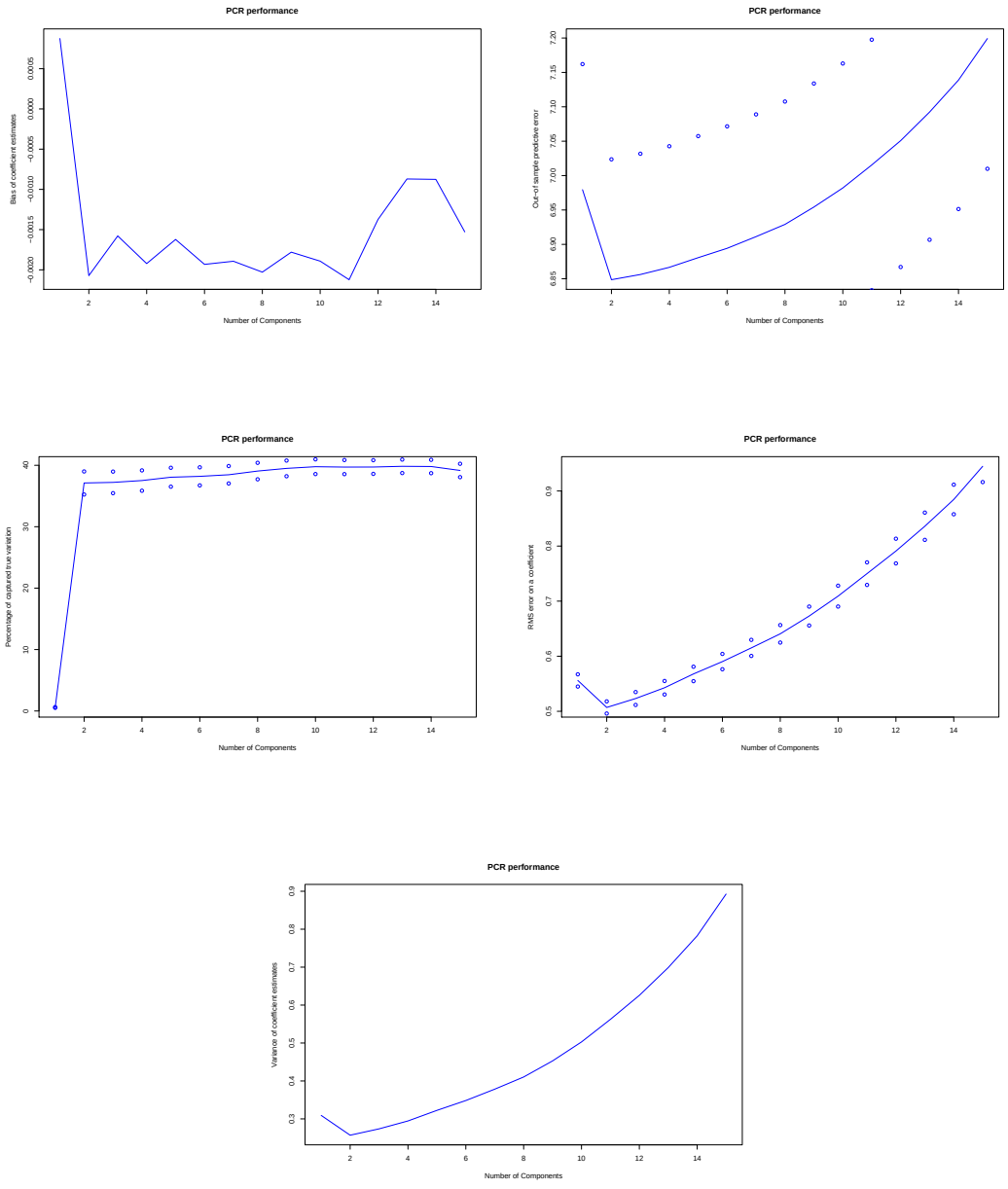

FIGURE C.7: Variability of the Five Performance Characteristics when the $SNR = 0.1$ for Different Values of PCs.

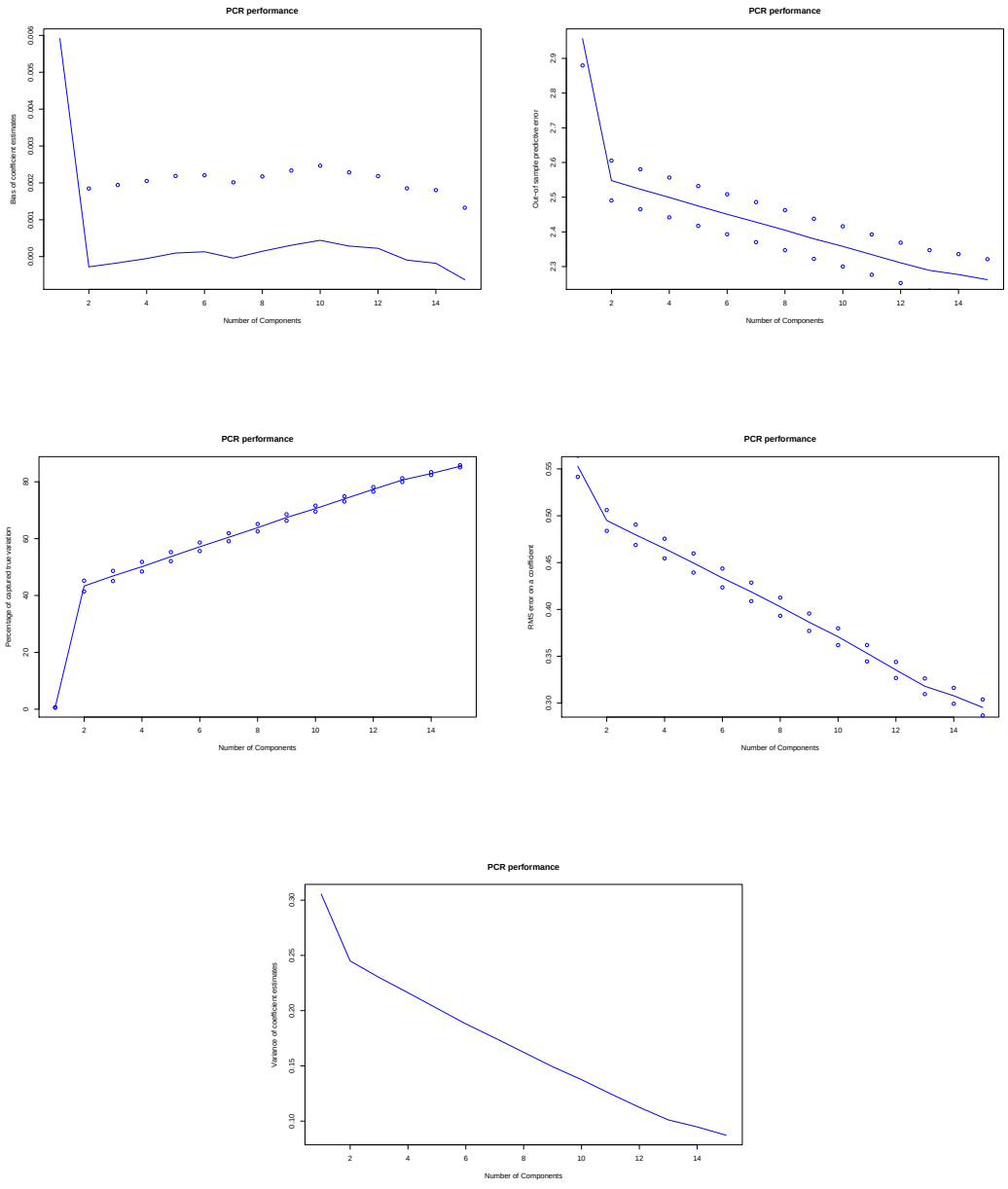

FIGURE C.8: Variability of the Five Performance Characteristics when the $SNR = 1.0$ for Different Values of PCs.

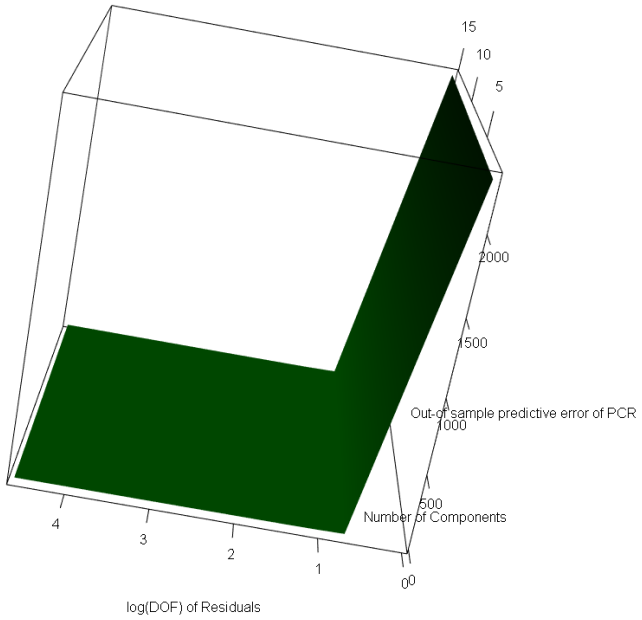

Figure C.9: Variability of the Degrees of Freedom of Residuals Performance Over the Out-Of-Sample-Predictive-Error for Different Values of PCs

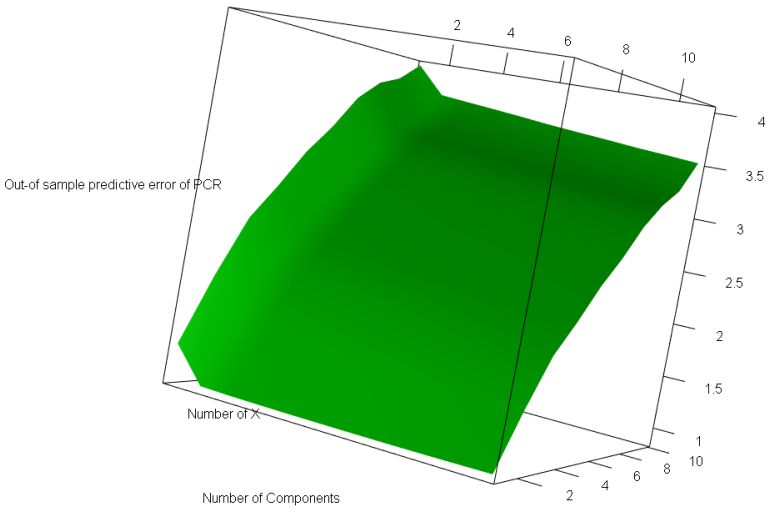

Figure C.10: Variability of the Number of True Predictors Performance Over the Out-Of-Sample-Predictive-Error for Different Values of PCs

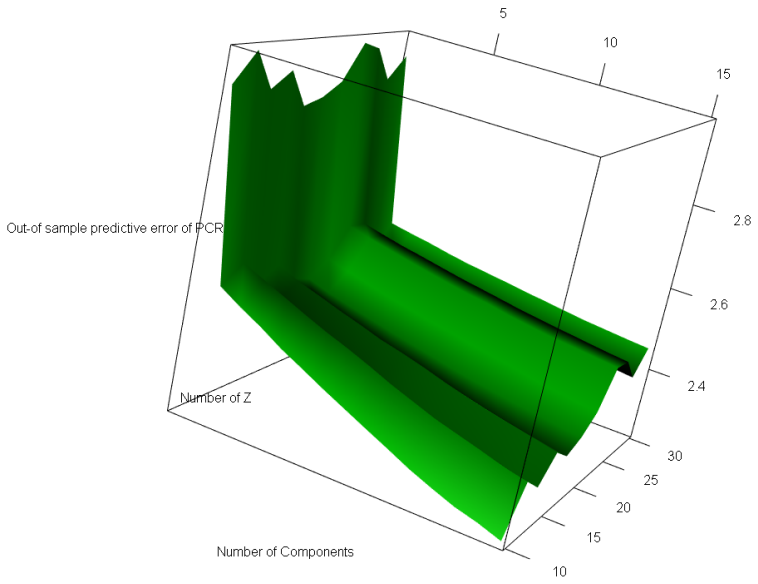

FIGURE C.11: Variability of the Number of False Predictors Performance Over the Out-Of-Sample-Predictive-Error for Different Values of PCs

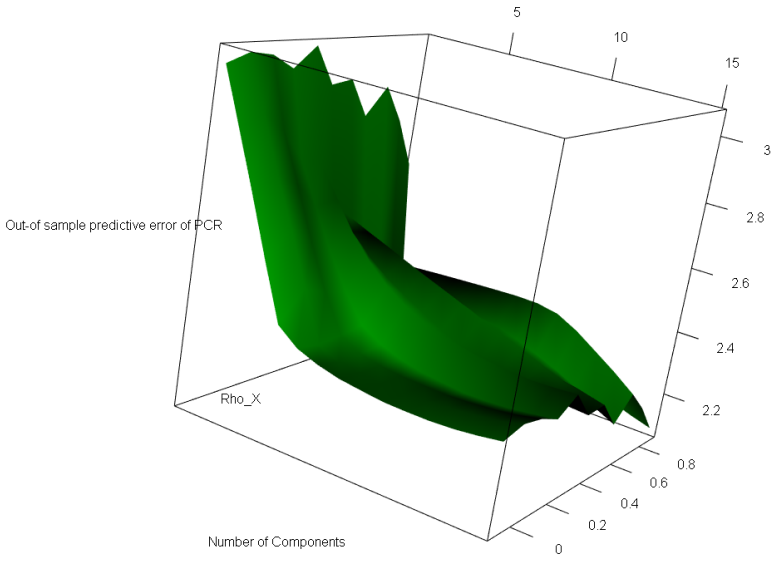

FIGURE C.12: Variability of the Correlation of True Predictors Performance Over the Out-Of-Sample-Predictive-Error for Different Values of PCs

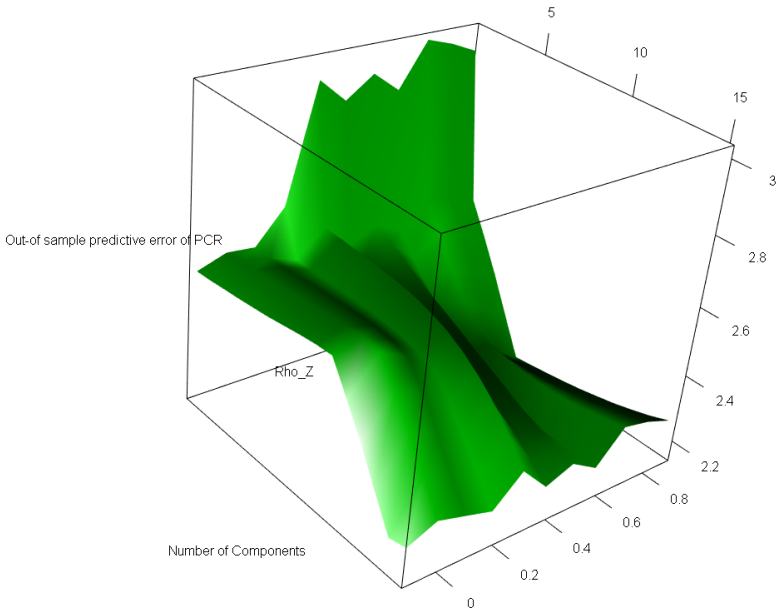

Figure C.13: Variability of the Correlation of False Predictors Performance Over the Out-Of-Sample-Predictive-Error for Different Values of PCs

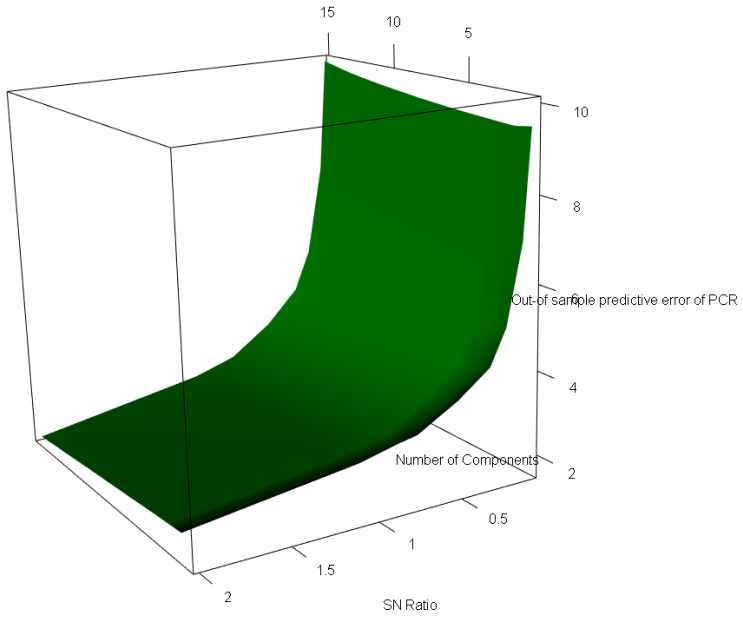

Figure C.14: Variability of the SNR Performance Over the Out-Of-Sample-Predictive-Error for Different Values of PCs

Appendix D

Complete Set of Graphs from the Real Data Sets Study with $N = 20$

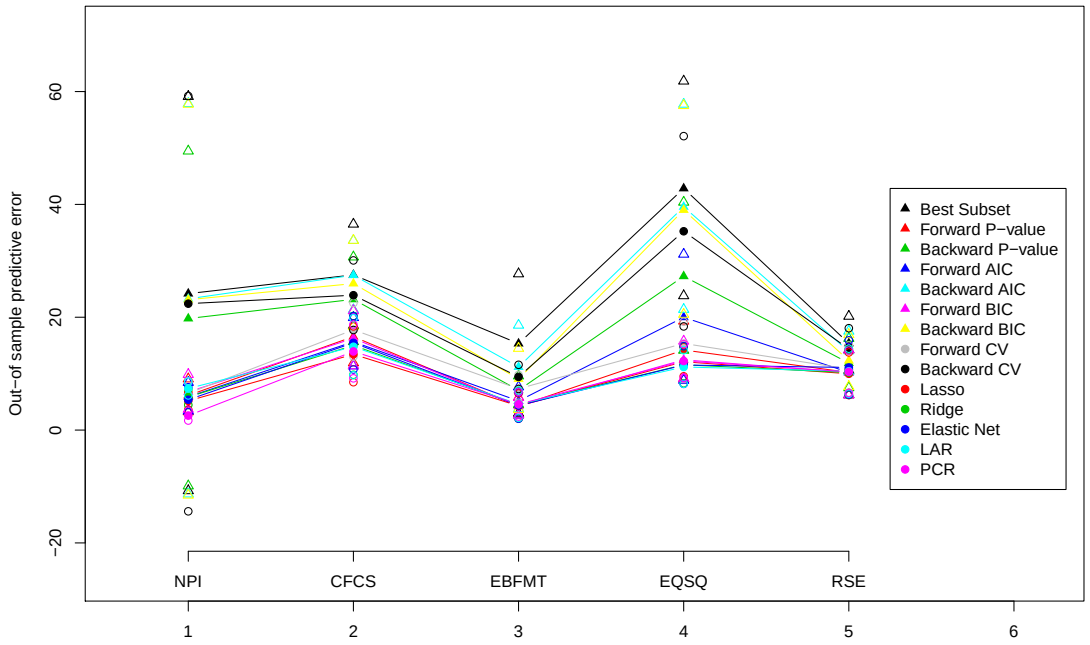

FIGURE D.1: Variability of the Out-Of-Sample-Predictive-Error on Five Real Sets when $N = 20$.

Bibliography

Aanæs, Henrik, Rune Fisker, Kalle Astrom, and Jens Michael Carstensen. 2002. "Robust factorization." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (9): 1215–1225.

Allen, David M. 1974. "The relationship between variable selection and data agumentation and a method for prediction." *Technometrics* 16 (1): 125–127.

AlNasser, Hassan. 2017. "On ridge regression and least absolute shrinkage and selection operator." PhD diss.

Andrews, David F, and Colin L Mallows. 1974. "Scale mixtures of normal distributions." *Journal of the Royal Statistical Society. Series B (Methodological):* 99–102.

Arthanari, TS, and Y Dodge. 1981. *Mathematical Programming in Statistics. A Wiley-Interscience Publication John Wiley and Sons.*

Bae, Kyounghwa, and Bani K Mallick. 2004. "Gene selection using a two-level hierarchical Bayesian model." *Bioinformatics* 20 (18): 3423–3430.

Basri, Ronen, and David W Jacobs. 2003. "Lambertian reflectance and linear subspaces." *IEEE Transactions on Pattern Analysis & Machine Intelligence,* no. 2: 218–233.

Bauschke, Heinz H, Patrick L Combettes, et al. 2011. *Convex analysis and monotone operator theory in Hilbert spaces.* Vol. 408. Springer.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57 (1): 289–300.

Bertsimas, Dimitris, Angela King, Rahul Mazumder, et al. 2016. "Best subset selection via a modern optimization lens." *The annals of statistics* 44 (2): 813–852.

Bertsimas, Dimitris, and Romy Shioda. 2009. "Algorithm for cardinality-constrained quadratic optimization." *Computational Optimization and Applications* 43 (1): 1–22.

"Best Subsets Regression, Adjusted R-Sq, Mallows Cp." 2007. University Lecture. https://newonlinecourses.science.psu.edu/stat501/node/330/.

Bishop, Chris M. 1995. "Training with noise is equivalent to Tikhonov regularization." *Neural computation* 7 (1): 108–116.

Bousquet, Olivier, Stéphane Boucheron, and Gábor Lugosi. 2003. "Introduction to statistical learning theory." In *Summer School on Machine Learning,* 169–207. Springer.

Boyd, Stephen, and Lieven Vandenberghe. 2004. *Convex optimization.* Cambridge university press.

Breiman, Leo, et al. 1996. "Heuristics of instability and stabilization in model selection." *The annals of statistics* 24 (6): 2350–2383.

Carbonneau, Réal A, Gilles Caporossi, and Pierre Hansen. 2011. "Globally optimal clusterwise regression by mixed logical-quadratic programming." *European Journal of Operational Research* 212 (1): 213–222.

Casella, George. 2001. "Empirical bayes gibbs sampling." *Biostatistics* 2 (4): 485–500.

Chen, Scott Shaobing, David L Donoho, and Michael A Saunders. 2001. "Atomic decomposition by basis pursuit." *SIAM review* 43 (1): 129–159.

Chhikara, RS, and JL Folks. 1989. *The Inverse Gaussian Distribution: Theory.*

Chiang, Kai-Yang, Cho-Jui Hsieh, and Inderjit S Dhillon. 2016. "Robust Principal Component Analysis with Side Information." In *ICML,* 1:4. 2.

De La Torre, Fernando, and Michael J Black. 2003. "A framework for robust subspace learning." *International Journal of Computer Vision* 54 (1-3): 117–142.

Dıaz-Uriarte, Ramón, and Melchor Fernández Almagro. 2003. "A simple method for finding molecular signatures from gene expression data." *arXiv preprint q-bio/0401043.*

Donoho, David L. 1995. "De-noising by soft-thresholding." *IEEE transactions on information theory* 41 (3): 613–627.

Donoho, David L, and Jain M Johnstone. 1994. "Ideal spatial adaptation by wavelet shrinkage." *biometrika* 81 (3): 425–455.

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. 2004. "Least angle regression." *The Annals of statistics* 32 (2): 407–499.

Fan, Jianqing, and Runze Li. 2001. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association* 96 (456): 1348–1360.

Figueiredo, Adelaide, and Paulo Gomes. 2003. "Power of tests of uniformity defined on the hypersphere." *Communications in Statistics-Simulation and Computation* 32 (1): 87–94.

Figueiredo, Mário AT. 2003. "Adaptive sparseness for supervised learning." *IEEE transactions on pattern analysis and machine intelligence* 25 (9): 1150–1159.

Frank, LLdiko E, and Jerome H Friedman. 1993. "A statistical view of some chemometrics regression tools." *Technometrics* 35 (2): 109–135.

Fu, Wenjiang J. 1998. "Penalized regressions: the bridge versus the lasso." *Journal of computational and graphical statistics* 7 (3): 397–416.

Gneiting, Tilmann. 1997. "Normal scale mixtures and dual probability densities." *Journal of Statistical Computation and Simulation* 59 (4): 375–384.

Golub, Gene H, Michael Heath, and Grace Wahba. 1979. "Generalized cross-validation as a method for choosing a good ridge parameter." *Technometrics* 21 (2): 215–223.

Heller, Jürgen. 2019. "Assessment structures in psychological testing." *Mathematical Psychology* 91:1–13.

Hochberg, Yosef. 1988. "A sharper Bonferroni procedure for multiple tests of significance." *Biometrika* 75 (4): 800–802.

Hoerl, Arthur E, and Robert W Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12 (1): 55–67.

Hojsgaard, Soren. n.d. "Examples of multivariate analysis Principal component analysis (PCA)." *Statistics and Decision Theory Research Unit, Danish Institute of Agricultural Sciences.*

Hooker, John N, and Maria A Osorio. 1999. "Mixed logical-linear programming." *Discrete Applied Mathematics* 96:395–442.

"Information Criteria and PRESS." 2019. University Lecture. https://newonlinecourses.science.psu.edu/stat501/node/334/.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning.* Vol. 112. Springer.

Jolliffe, Ian T, Nickolay T Trendafilov, and Mudassir Uddin. 2003. "A modified principal component technique based on the LASSO." *Journal of computational and Graphical Statistics* 12 (3): 531–547.

Jørgensen, Bent. 1987. "Exponential dispersion models." *Journal of the Royal Statistical Society: Series B (Methodological)* 49 (2): 127–145.

Konno, Hiroshi, and Rei Yamamoto. 2009. "Choosing the best set of variables in regression analysis using integer programming." *Journal of Global Optimization* 44 (2): 273–282.

Li, Guoying, and Zhonglian Chen. 1985. "Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo." *Journal of the American Statistical Association* 80 (391): 759–766.

Mallows, Colin L. 1973. "Some comments on C p." *Technometrics* 15 (4): 661–675.

Marquardt, Donald W, and Ronald D Snee. 1975. "Ridge regression in practice." *The American Statistician* 29 (1): 3–20.

MikeWest, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, Jeffrey R Marks, and Joseph R Nevins. 2001. "Predicting the Clinical Status of Human Breast Cancer using Gene Expression Profiles."

Miller, Alan. 2002. *Subset selection in regression.* Chapman / Hall/CRC.

Miyashiro, Ryuhei, and Yuichi Takano. 2015. "Subset selection by Mallows' Cp: A mixed integer programming approach." *Expert Systems with Applications* 42 (1): 325–331.

Montgomery, Douglas C, Elizabeth A Peck, and G Geoffrey Vining. 2012. *Introduction to linear regression analysis.* Vol. 821. John Wiley & Sons.

Morozov, Vladimir Alekseevich. 2012. *Methods for solving incorrectly posed problems.* Springer Science & Business Media.

Oneto, Luca, Sandro Ridella, and Davide Anguita. 2016. "Tikhonov, Ivanov and Morozov regularization for support vector machine learning." *Machine Learning* 103 (1): 103–136.

Osborne, Michael R, Brett Presnell, and Berwin A Turlach. 2000a. "A new approach to variable selection in least squares problems." *IMA journal of numerical analysis* 20 (3): 389–403.

———. 2000b. "On the lasso and its dual." *Journal of Computational and Graphical statistics* 9 (2): 319–337.

Park, Trevor, and George Casella. 2008. "The bayesian lasso." *Journal of the American Statistical Association* 103 (482): 681–686.

Pelckmans, Kristiaan, Johan AK Suykens, and Bart De Moor. 2004. "Morozov, ivanov and tikhonov regularization based LS-SVMs." In *International Conference on Neural Information Processing,* 1216–1222. Springer.

Richardson, Mark. 2009. "Principal component analysis." *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si* 6:16.

"Ridge Regression." 2018. University Lecture. https://newonlinecourses.science.psu.edu/stat508/lesson/5/5.1.

Sakaluk, John Kitchener. 2019. "Expanding Statistical Frontiers in Sexual Science: Taxometric, Invariance, and Equivalence Testing." PMID: 30793956, *The Journal of Sex Research:* 1–36. doi:10.1080/00224499.2019.1568377. eprint: https://doi.org/10.1080/00224499.2019.1568377. https://doi.org/10.1080/00224499.2019.1568377.

Segal, Mark R, Kam D Dahlquist, and Bruce R Conklin. 2003. "Regression approaches for microarray data analysis." *Journal of Computational Biology* 10 (6): 961–980.

Shen, Haipeng, and Jianhua Z Huang. 2008. "Sparse principal component analysis via regularized low rank matrix approximation." *Journal of multivariate analysis* 99 (6): 1015–1034.

Shlens, Jonathon. 2014. "A tutorial on principal component analysis." *arXiv preprint arXiv:1404.1100.*

Simon, Noah, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2013. "A sparse-group lasso." *Journal of Computational and Graphical Statistics* 22 (2): 231–245.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological):* 267–288.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (1): 91–108.

Tibshirani, Ryan J, et al. 2013. "The lasso problem and uniqueness." *Electronic Journal of Statistics* 7:1456–1490.

Tibshirani, Ryan, and Larry Wasserman. 2015. *Sparsity and the lasso.*

Tikhonov, Andrey N, and Vasilii Iakkovlevich Arsenin. 1977. *Solutions of ill-posed problems.* Vol. 14. Vh Winston.

Tilburg, Wijnand A.P.van. 2019. "It's not unusual to be unusual (or: A different take on multivariate distributions of personality)." *Personality and Individual Differences* 139:175–180.

Trevor, Hastie, Tibshirani Robert, and Friedman JH. 2009. *The elements of statistical learning: data mining, inference, and prediction.*

Ulfarsson, Magnus O, and Victor Solo. 2011. "Vector l_0 Sparse Variable PCA." *IEEE Transactions on Signal Processing* 59 (5): 1949–1958.

Vapnik, Vladimir. 2013. *The nature of statistical learning theory.* Springer science & business media.

Wang, Ling, and Hong Cheng. 2013. "Robust principal component analysis for sparse face recognition." In *2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP),* 171–176. IEEE.

West, Mike. 1984. "Outlier models and prior distributions in Bayesian linear regression." *Journal of the Royal Statistical Society: Series B (Methodological)* 46 (3): 431–439.

———. 1987. "On scale mixtures of normal distributions." *Biometrika* 74 (3): 646–648.

WhoisGuard, Inc. 2017. "Raw data from online personality tests." September. https://openpsychometrics.org/_rawdata/.

Wieringen, Wessel N van. 2015. "Lecture notes on ridge regression." *arXiv preprint arXiv:1509.09169.*

Williams, H Paul. 2013. *Model building in mathematical programming.* John Wiley & Sons.

Wu, Ming-Chun, and Kwang-Cheng Chen. 2016. "Sparse PCA via hard thresholding for blind source separation." In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on,* 2539–2543. IEEE.

Xue, Niannan, Jiankang Deng, Yannis Panagakis, and Stefanos Zafeiriou. 2017. "Informed Non-convex Robust Principal Component Analysis with Features." *arXiv preprint arXiv:1709.04836.*

Xue, Niannan, Yannis Panagakis, and Stefanos Zafeiriou. 2017. "Side information in robust principal component analysis: Algorithms and applications." In *Proceedings of the IEEE International Conference on Computer Vision,* 4317–4325.

Yang, Meng, Lei Zhang, Jian Yang, and David Zhang. 2011. "Robust sparse coding for face recognition." In *CVPR 2011,* 625–632. IEEE.

Yuan, Ming, and Yi Lin. 2005. "Efficient empirical Bayes variable selection and estimation in linear models." *Journal of the American Statistical Association* 100 (472): 1215–1225.

Zhang, Lingsong, JS Marron, Haipeng Shen, and Zhengyuan Zhu. 2007. "Singular value decomposition and its visualization." *Journal of Computational and Graphical Statistics* 16 (4): 833–854.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–320.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. "Sparse principal component analysis." *Journal of computational and graphical statistics* 15 (2): 265–286.