ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

PhD THESIS

Murat GENÇ

RIDGE, LASSO AND ELASTIC NET METHODS IN HIGH DIMENSIONAL DATA

DEPARTMENT OF STATISTICS

ADANA-2020

ABSTRACT

PhD THESIS

RIDGE, LASSO AND ELASTIC NET METHODS IN HIGH DIMENSIONAL DATA

Murat GENÇ

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Mahmude Revan ÖZKALE
Year: 2020, Pages: 159
Jury : Prof. Dr. Mahmude Revan ÖZKALE
: Prof. Dr. Sadullah SAKALLIOĞLU
: Assoc. Prof. Dr. Hüseyin GÜLER
: Prof. Dr. Çağdaş Hakan ALADAĞ
: Assoc. Prof. Dr. Duygu İÇEN

Penalized linear regression methods that yield interpretable models with high prediction accuracy have gained great importance with the emergence of big and high-dimensional data sets recently.

In this study, firstly, the ridge, LASSO, two-parameter estimation and elastic net methods which are popular methods in the literature are discussed in the context of penalized linear regression methods. In the next stage, the GO estimator is proposed as an alternative to these methods in case there is multicollinearity in the data set. After that, the shift-ridge and shift-LASSO estimators are proposed which can be used in modeling if the data set is high-dimensional in addition to having multicollinearity. Finally, the stochastic restricted LASSO estimator is proposed, which can be used as an alternative estimator if there are convenient stochastic restrictions for the data set. Real data set analysis and simulation studies are performed for each estimator. The proposed estimators are compared with other estimators in the literature by making use of C++ and R programming languages according to criteria such as test mean squared error and number of variables selected.

Key Words: High-Dimensionality, Multicollinearity, Penalized Regression, Shift-LASSO estimator, Stochastic restricted LASSO

ÖZ

DOKTORA TEZİ

YÜKSEK BOYUTLU VERİLERDE RIDGE, LASSO VE ELASTİK NET YÖNTEMLERİ

Murat GENÇ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK

Danışman : Prof. Dr. Mahmude Revan ÖZKALE
Yıl: 2020, Sayfa: 159
Jüri : Prof. Dr. Mahmude Revan ÖZKALE
: Prof. Dr. Sadullah SAKALLIOĞLU
: Doç. Dr. Hüseyin GÜLER
: Prof. Dr. Çağdaş Hakan ALADAĞ
: Doç. Dr. Duygu İÇEN

Günümüzde büyük ve yüksek boyutlu veri kümelerinin ortaya çıkması ile yüksek ön tahmin doğruluğuna sahip olmasının yanı sıra kolay yorumlanabilen modeller veren cezalandırılmış doğrusal regresyon yöntemleri büyük önem kazanmıştır.

Bu çalışmada öncelikle literatürde yaygın olarak bilinen ridge, LASSO ve iki parametreli tahmin ve elastik net yöntemleri, cezalandırılmış doğrusal regresyon yöntemleri bağlamında ele alınmış ve bu yöntemlere dair temel bilgiler verilmiştir. Bir sonraki aşamada veri kümesinde çoklu iç ilişki olması halinde bu yöntemlere alternatif olarak GO tahmin edicisi önerilmiştir. Daha sonra veri kümesinde çoklu iç ilişki olmasının yanı sıra veri kümesinin yüksek boyutlu olması halinde de model oluşturmada kullanılabilen shift-ridge ve shift-LASSO tahmin edicileri önerilmiştir. Son olarak veri kümesine uygun stokastik kısıtların olması durumunda alternatif bir tahmin edici olarak kullanılabilen stokastik kısıtlı LASSO tahmin edicisi önerilmiştir. Her bir tahmin edici için gerçek veri kümesi analizi ve simülasyon çalışmaları yapılmıştır. Önerilen tahmin ediciler, test kümesi hata kareler ortalaması ve seçilen değişken sayısı gibi kriterlere göre C++ ve R programlama dilleri kullanılarak literatürde bulunan diğer tahmin ediciler ile karşılaştırılmıştır.

Anahtar Kelimeler: Yüksek Boyutluluk, Çoklu İç İlişki, Cezalandırılmış Regresyon, Shift-LASSO tahmin edicisi, Stokastik kısıtlı LASSO

EXTENDED ABSTRACT

The statistical models are useful tools for simplifying and interpreting complex phenomena with uncertainty. Linear regression models are special types of statistical models to analyze the relationship between the response variable and explanatory variables. These models aim to determine which explanatory variables affect the response variable and how the magnitude and direction of this effect.

The prediction accuracy and the model interpretability are two important concepts in the linear regression modeling. Prediction accuracy is a measure of the closeness between the estimated response values and the observed response values while the model interpretability is about determining the explanatory variables that have a strong effect on the response variable. Therefore, a linear regression model should have high prediction accuracy and model interpretability.

The ordinary least squares method (OLS) is a commonly used method to obtain the coefficient estimates of a linear regression model. The OLS is the best linear unbiased estimator (BLUE) among all the linear unbiased estimators. However, the OLS is not satisfactory with the situations of multicollinearity and the high-dimensionality. Also, the OLS can not perform automatic variable selection. Various alternatives to the OLS have been proposed to improve the OLS estimates in the case of multicollinearity, high-dimensionality and model interpretability.

To improve the model interpretability of the OLS, many methods are proposed such as adjusted R^2 (Edwards, 1969; Seber, 1977), Mallows C_p (Mallows, 1964, 1966, 1973, 1995), Akaike's information criterion (AIC) (Akaike, 1973, 1974) and Bayesian information criterion (BIC) (Schwartz, 1978). All of these methods have their own advantages and disadvantages.

The multicollinearity implies the high linear dependencies among the explana-

tory variables. There are different sources of the multicollinearity in the data set such as the data collection method, the specification of the model and redundant variables that express the same information in a different form. When there is multicollinearity in the data set, the coefficient estimates of the OLS have high variance. The multicollinearity can also lead to incorrect signs and implausible magnitudes of the coefficients. The multicollinearity can be detected by diagnostic measures such as the correlation matrix of the data, variance inflation factors (VIF) (Marquardt, 1970), eigenvalues (Kendall, 1957) and variance decomposition proportions (Belsley et al., 1980).

The high-dimensionality indicates that the number of the explanatory variables is greater than the number of the observations in a data set. As a result of arising large data sets as a consequence of advances in technology, the high-dimensional data become very important recently. These data sets may contain a very large number of observations or explanatory variables.

The OLS does not exist for a high-dimensional data set. To deal with this problem, penalized regression methods can be used. The penalized regression methods do coefficient shrinkage and variable selection depending on the penalization type. Some special penalized regression methods are mentioned as follows.

The best subset selection method has an aim to model the explanatory variables that have a strong effect on the response variable which corresponds to ℓ_0-norm penalization. Hence, the best subset selection provides interpretable models. However, this method yields quite unstable coefficient estimates. This situation leads to poor prediction accuracy for the new observations. Also, the number of possible models increases exponentially with the number of the explanatory variables which indicates that the computational cost of the best subset selection is very high.

The ridge estimator proposed by Hoerl and Kennard (1970) corresponds to ℓ_2-norm penalization. According to the authors, the ridge estimator can deal with the

multicollinearity and high-dimensionality with better prediction performance than the OLS estimator by applying a penalty term. As the value of the penalty term increases, the variance of the ridge coefficient estimates decreases and the coefficient estimates are shrunken towards zero. However, the ridge estimator does not set any coefficient to zero. Therefore, the ridge estimator does not perform automatic variable selection.

As an alternative to the ridge estimator, Tibshirani (1996) proposes the LASSO estimator. The LASSO estimator corresponds to ℓ_1-norm penalization. The estimator reduces the number of unnecessary coefficients by setting their coefficients to zero as a result of the ℓ_1-norm. The estimator performs automatic variable selection, hence, overcomes the disadvantage of the ridge estimator with respect to the model interpretability. Although the LASSO estimator has become tremendously popular, it has some drawbacks. For some data structures, unique LASSO estimates do not exist. The LASSO can not select more variables than the number of the observations in a high-dimensional data set. The ridge estimator yields better prediction performance than the LASSO estimator if there is multicollinearity among the explanatory variables.

To deal with the deficiencies of the LASSO, the elastic net estimator proposed by Zou and Hastie (2005) uses a penalization function based on a linear combination of ℓ_1 and ℓ_2-norm penalization. The elastic net estimator can select all the variables even in the high-dimensional case. The estimator has a grouping property which implies that the estimator extracts the grouping pattern of the highly correlated explanatory variables.

In this thesis, we discuss the methods that perform both shrinkage and variable selection simultaneously. The thesis consists of seven chapters. After mentioning some preliminary concepts, we give some explanations about some penalized regression methods such as the ridge, LASSO, elastic net estimators in the literature.

We introduce three novel penalized regression methods in the scope of this

thesis. The first estimator is the GO[1] estimator. We investigate the properties of the GO estimator in detail. We propose the coordinate descent algorithm to estimate the coefficients of the GO estimator. We also give a theorem about the grouping property of the estimator which is a generalization of the grouping property of the elastic net estimator.

While the GO estimator has prominent properties, it can not be applied to the high-dimensional data sets. To deal with this situation, we propose the shift-ridge and shift-LASSO estimators. We propose the coordinate descent and ADMM algorithms for the coefficient estimation of the shift-ridge and shift-LASSO estimators. We also give a theorem about the grouping property of these estimators. The grouping property is a very important property for a microarray data set in medicine. In a microarray data set, the explanatory variables correspond to the genes. In this data set, the correlations between the genes from the same biological pathway can be very high. Hence, these genes form a group of explanatory variables. In this case, the shift-ridge and shift-LASSO estimators can include the group of variables in the model in a similar way to the elastic net estimator. Therefore, the shift-ridge and shift-LASSO estimators are alternatives to the elastic net estimator in a microarray data set with respect to the variable selection and grouping property.

We also propose the stochastic restricted LASSO estimator by combining the ideas of the mixed and LASSO estimators. We examine the properties of the estimator in detail. We also give the coordinate descent algorithm for stochastic restricted LASSO estimator.

We perform real data analyses from the field of medicine to compare the performances of the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators on three different data sets. We used the prediction mean squared error

[1]The word GO represents the first letters of the names of the authors.

and selected number of variables as comparison criteria for the GO, shift-ridge and shift-LASSO estimators. The scalar mean squared error is also used as a comparison criterion for the stochastic restricted LASSO estimator. According to the results of the analyses, the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators give better results compared to the alternative methods.

We carry out simulation studies by using different data structures to compare the performances of the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators with other penalized regression methods. The results from each simulation study are supported by tables and plots. According to the simulation studies, the new estimators yield better results than the alternative estimators depending on the simulation study.

As a general statement, all the theoretical and practical results given in the thesis show that the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators are very useful for high-dimensional and ill-conditioned data.

ACKNOWLEDGEMENTS

Foremost, I would like to express my gratitude to my supervisor Prof. Dr. M. Revan Özkale, who provide continuous support for my thesis study and research. Her guidance helped me in all the time of research and writing of this thesis. Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Dr. Sadullah Sakallıoğlu and Assoc. Prof. Dr. Hüseyin Güler, for their encouragement and insightful comments. Last but not the least, I would like to thank my family for their support throughout my life.

CONTENTS **PAGE**

ABSTRACT I
ÖZ II
EXTENDED ABSTRACT III
ACKNOWLEDGEMENTS VIII
CONTENTS IX
LIST OF TABLES XIII
LIST OF FIGURES XVII
LIST OF ABBREVIATONS XXI
1. INTRODUCTION 1
2. PRELIMINARY CONCEPTS 9
2.1. Comparison Criteria 10
2.2. Degrees of Freedom and Model Selection Criteria 14
2.3. Resampling Methods (Out-of-Sample Testing) 16
2.3.1. The Validation Set Method 17
2.3.2. Leave-One-Out Cross-Validation (LOOCV) 17
2.3.3. k-Fold Cross-Validation 18
2.3.4. The Bootstrap Method 19
2.4. Optimization Methods 19
2.4.1. Iterative Algorithms 22
2.4.1.1. The Coordinate Descent Algorithm 23
2.4.1.2. The Alternating Direction Method of Multipliers 24
3. PENALIZED REGRESSION 27
3.1. Penalized Regression Methods 31
3.1.1. The Ridge Estimator 32

3.1.1.1. Constrained and Bayesian Form of the Ridge Estimator . 33
3.1.1.2. The Ridge Estimator in Orthonormal Design 35
3.1.1.3. Degrees of Freedom of the Ridge Estimator 36
3.1.1.4. Choosing the Tuning Parameter of the Ridge Estimator . 37
3.1.2. A Two-Parameter Estimator 38
3.1.3. The LASSO Estimator 40
3.1.3.1. Constrained and Bayesian Form of the LASSO Estimator 42
3.1.3.2. The LASSO Estimator in Orthonormal Design 43
3.1.3.3. Degrees of Freedom of the LASSO Estimator 45
3.1.3.4. Choosing the Tuning Parameter of the LASSO Estimator 45
3.1.4. The Elastic Net Estimator 46
3.1.4.1. Constrained Form of the Elastic Net Estimator 48
3.1.4.2. The Elastic Net Estimator in Orthonormal Design 50
3.1.4.3. Relation of the Elastic Net Estimator with the LASSO Estimator 50
3.1.4.4. Reparametrization of the Tuning Parameters of the Elastic Net Estimator 51
3.1.4.5. Grouping Effect of the Elastic Net Estimator 52
3.2. Restricted-Penalized Regression Methods 53
4. NOVEL PENALIZED REGRESSION METHODS 57
4.1. The GO Estimator 57
4.1.1. ℓ_1 Adjustment for the GO Estimator 62
4.1.2. The Coordinate Descent Algorithm for the GO Estimator 63
4.1.3. A New Grouping Property for the GO Estimator 68
4.2. The Shift-ridge and Shift-LASSO Estimators 72
4.2.1. The New Methods 72

4.2.2. Algorithms for Fitting the Shift-ridge and Shift-LASSO Estimators 76
4.2.2.1. The ADMM Algorithm for the Shift-ridge and Shift-LASSO Estimators 77
4.2.2.2. The Coordinate Descent Algorithm for the Shift-ridge and Shift-LASSO Estimators 77
4.2.3. A New Grouping Property for the Shift-ridge and Shift-LASSO Estimators 80
4.3. The Stochastic Restricted LASSO Estimator 83
4.3.1. The Coordinate Descent Algorithm for the Stochastic Restricted LASSO Estimator 84
5. REAL DATA ANALYSES 87
5.1. Data Sets 87
5.1.1. The Prostate Data 87
5.1.2. The Diabetes Data 88
5.1.3. The Riboflavin Data 88
5.2. Real Data Analysis for the GO Estimator 89
5.2.1. The Prostate Data Analysis for the GO Estimator 90
5.2.2. The Diabetes Data Analysis for the GO Estimator 94
5.3. Real Data Analysis for the Shift-ridge and Shift-LASSO Estimators . . 98
5.4. Real Data Analysis for the Stochastic Restricted LASSO Estimator . . 101
5.4.1. The Prostate Data Analysis for the Stochastic Restricted LASSO Estimator 102
5.4.2. The Riboflavin Data Analysis for the Stochastic Restricted LASSO Estimator 107
6. SIMULATION STUDIES 113
6.1. Simulation Studies for the GO Estimator 114

6.2. Simulation Studies for the Shift-ridge and Shift-LASSO Estimators . . 127
6.2.1. Simulation Studies for the Shift-ridge and Shift-LASSO Estimators on Correlated Data 128
6.2.2. Simulation Studies for the Shift-ridge and Shift-LASSO Estimators on High-Dimensional Data 135
6.3. Simulation Studies for the Stochastic Restricted LASSO Estimator . . 141
7. CONCLUSION 149
REFERENCES 151
CURRICULUM VITAE 159

LIST OF TABLES **PAGE**

Table 5.1. Correlation table for prostate data 88

Table 5.2. Correlation table for diabetes data 89

Table 5.3. Comparison of the OLS, ridge, OK, LASSO, elastic net and GO estimators on prostate data set with parameter values and selected variables 90

Table 5.4. Coefficient estimates, length of the coefficient vectors and PMSE values for the OLS, ridge, OK, LASSO, elastic net and GO estimators in prostate data 92

Table 5.5. Comparison of the OLS, ridge, OK, LASSO, elastic net and GO estimators on diabetes data with parameter values and selected variables 95

Table 5.6. Coefficient estimates, length of the coefficient vectors and PMSE values for the OLS, ridge, OK, LASSO, elastic net and GO estimators in diabetes data 96

Table 5.7. Comparison of the ridge, LASSO, elastic net shift-ridge and shift-LASSO estimators on riboflavin data with tuning parameters, PMSE, active set sizes and reduce rates 99

Table 5.8. Coefficient estimates, SMSE and PMSE values for the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators in prostate data 104

Table 5.9. Comparison of the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators on riboflavin data where the tuning parameters are selected by ten-fold cross-validation . 110

Table 5.10. Comparison of the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators on riboflavin data where the tuning parameters are selected by BIC 111

Table 6.1. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A1 with $\rho = 0.5$ and $\sigma = 3, 5, 15, 25$ 116

Table 6.2. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A1 with $\rho = 0.7$ and $\sigma = 3, 5, 15, 25$ 117

Table 6.3. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A1 with $\rho = 0.9$ and $\sigma = 3, 5, 15, 25$ 118

Table 6.4. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A1 with $\rho = 0.99$ and $\sigma = 3, 5, 15, 25$ 119

Table 6.5. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A2 with $\rho = 0.5$ and $\sigma = 3, 5, 15, 25$ 120

Table 6.6. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A2 with $\rho = 0.7$ and $\sigma = 3, 5, 15, 25$ 121

Table 6.7. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A2 with $\rho = 0.9$ and $\sigma = 3, 5, 15, 25$ 122

Table 6.8. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A2 with $\rho = 0.99$ and $\sigma = 3, 5, 15, 25$ 123
Table 6.9. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A3 124
Table 6.10. Comparisons of the OLS, ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 on the simulated correlated and grouping effected data sets for $\sigma = 3$ 130
Table 6.11. Comparisons of the OLS, ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 on the simulated correlated and grouping effected data sets for $\sigma = 15$ 131
Table 6.12. Comparisons of the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 on the simulated high-dimensional data sets for $\sigma = 3$ 137
Table 6.13. Comparisons of the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 on the simulated high-dimensional data sets for $\sigma = 15$ 138
Table 6.14. Comparisons of the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Study C1 145
Table 6.15. Comparisons of the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Study C2 145

Table 6.16. Comparisons of the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Study C3 146
Table 6.17. Comparisons of the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Study C4 146

LIST OF FIGURES **PAGE**

Figure 2.1. The PMSE, squared bias, variance curves and irreducible error line for a typical model. The vertical line denotes the optimal complexity point 12

Figure 3.1. (a) Constraint region and contours for the ridge estimator: the circular region corresponds to the constraint $\beta_1^2 + \beta_2^2 \leq t$, the contours correspond to the residual sum of squares (b) Density plot of the normal distribution $\mathcal{N}\left(0, \sigma_r^2\right)$ 34

Figure 3.2. (a) Constraint region and contours for the LASSO estimator: the rotated square region corresponds to the constraint $|\beta_1| + |\beta_2| \leq t$, the contours correspond to the residual sum of squares (b) Density plot of the Laplace distribution with location parameter zero and scale parameter σ_ℓ 43

Figure 3.3. The contours of residuals and region of constraints of the elastic net estimator 48

Figure 3.4. Two-dimensional contour plots of the ridge, LASSO and elastic net penalization functions 49

Figure 4.1. (a) Two dimensional contour plots of the ridge, OK, LASSO, elastic net and GO penalties with $\alpha = 0.5$, $d = 0.9$ (b) Two dimensional contour plots of the GO penalty for $\alpha = d = 0.5$ (c) Two dimensional contour plots of the GO penalty for $d = 0.2$, $d = 0.5$, $d = 0.9$ with $\alpha = 0.5$ 61

Figure 4.2. Exact solutions for the ridge, LASSO, elastic net and GO estimators in the orthonormal design where the shrinkage parameters are $\lambda = 2.5, \alpha = 0.5, d = 0.8$ 66
Figure 4.3. The LASSO (a), elastic net (b) and GO (c, d) coefficient estimates as a function of $\|\boldsymbol{\beta}\|_1 \big/ \max \|\boldsymbol{\beta}\|_1$ for the mpg data set when $\alpha = 0.5$, $d = 0.2$ and $d = 0.9$ for 100 different values of λ 67
Figure 4.4. The LASSO (a), elastic net (b) and GO (c, d) coefficient estimates as a function of $\|\boldsymbol{\beta}\|_1$ for the simulated data set when $\alpha = 0.3$, $d = 0.2$ and $d = 0.9$ for 100 different values of λ 73
Figure 4.5. Solutions to the ridge, LASSO, elastic net, GO, shift-ridge, shift-LASSO estimators in the orthonormal design where the tuning parameters are $\lambda = 2.5, \alpha = 0.5, d = 0.8$ 79

Figure 5.1. Box plots of bootstrap coefficient estimates of the prostate data for the ridge, OK, LASSO, elastic net and GO estimators 93
Figure 5.2. PMSE values of the prostate data set for the ridge, OK, LASSO, elastic net and GO estimators 94
Figure 5.3. Box plots of bootstrap coefficient estimates of the diabetes data for the ridge, OK, LASSO, elastic net and GO estimators 97
Figure 5.4. PMSE values of the diabetes data set for the ridge, OK, LASSO, elastic net and GO estimators 98
Figure 5.5. Coefficient paths of the ridge, LASSO, elastic net shift-ridge and shift-LASSO estimators for the riboflavin data 100
Figure 5.6. Box plots of bootstrap coefficient estimates of the prostate data for the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators obtained by cross-validation 105

Figure 5.7. Box plots of bootstrap coefficient estimates of the prostate data for the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators obtained by BIC criterion 106
Figure 5.8. SMSE plots as a function of λ for the ridge-stochastic restricted ridge estimators and the LASSO-stochastic restricted LASSO estimators 108
Figure 5.9. PMSE plots as a function of λ for the ridge-stochastic restricted ridge estimators and the LASSO-stochastic restricted LASSO estimators 109
Figure 5.10. Coefficient path of the LASSO and stochastic restricted LASSO estimators in riboflavin data. 111

Figure 6.1. Box plots to compare the accuracy of the ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 for $\sigma = 3$ 132
Figure 6.2. Box plots to compare the accuracy of the OLS, ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 for $\sigma = 15$ 133
Figure 6.3. Box plots to compare the accuracy of the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators in Simulation Studies B5-B8 for $\sigma = 3$ 139
Figure 6.4. Box plots to compare the accuracy of the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators in Simulation Studies B5-B8 for $\sigma = 15$ 140

Figure 6.5. Box plots of MSE values of the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Studies C1-C4 147

LIST OF ABBREVIATONS

ADMM	: Alternating Direction Method of Multipliers
AIC	: Akaike information criterion
BIC	: Bayesian information criterion
BLUE	: Best linear unbiased estimator
df	: Degrees of freedom
KKT	: Karush–Kuhn–Tucker
LASSO	: Least absolute shrinkage and selection operator
MSE	: Mean squared error
OLS	: Ordinary least squares
PMSE	: Prediction mean squared error
RSS	: Residual sum of squares
SMSE	: Scalar mean squared error
tr	: Trace of matrix
$\mathbf{I}_p$	: $p \times p$ identity matrix

1. INTRODUCTION

Statistics aim to learn from data. The statistical models are used to simplify and interpret complex phenomena with uncertainty based on the information in the observed data. A statistical model describes the data, determines the functional expression that best represents the data according to a certain criterion and depicts the relationship among the explanatory variables.

Linear regression models are widely used special types of statistical models to analyze the relationship between the response variable (dependent variable) and explanatory variables (independent variables). These models aim to determine whether explanatory variables affect the response variable and, if so, try to explain how the magnitude and direction of this effect. These models have many applications in almost every field of science as well as economics, physics, engineering and chemistry.

Two important concepts in the linear regression modeling are the prediction accuracy and the interpretability of the model. These concepts are evaluated with respect to the model coefficients. Prediction accuracy is a measure of the closeness between the estimated response values and the actual response values. An estimator yielding low variance is preferable in that it exhibits high prediction accuracy. The model interpretability is related to the modeling of the explanatory variables that have the strongest effects on the response variable. When a linear regression model contains explanatory variables that best describe the response variable, the model has a high prediction performance which is a desirable situation.

The oldest known method to estimate the coefficient vector of a linear regression model is the ordinary least squares (OLS). The OLS is unbiased and yields minimum variance among all unbiased estimators. However, estimations and predictions based on the OLS are not satisfactory with the situations where the explanatory

variables exhibit extremely high linear dependencies among the explanatory variables, known as multicollinearity and the number of the explanatory variables exceeds the number of the observations, known as high-dimensionality. Also, the OLS can not discard variables from the model that are not related to the response variable. This leads to poor results in terms of model interpretability. Various methods have been proposed to improve the weakness of the OLS estimates in the case of multicollinearity, high-dimensionality and model interpretability. A brief explanation of these methods is given in this chapter.

A wide range of studies takes care of the model interpretability of the OLS in the sense of removing redundant variables by some criteria from the model. Along with traditional methods such as subset selection and stepwise regression (Miller, 2002), some examples are as follows: Adjusted R^2 (Edwards, 1969; Seber, 1977), Mallows C_p (Mallows, 1964, 1966, 1973, 1995), Akaike's information criterion (AIC) (Akaike, 1973, 1974) and Bayesian information criterion (BIC) (Schwartz, 1978). The mentioned variable selection methods have pros and cons depending on the data and the target of the analysis. Also, the methods can result in different best models. In these cases, a comprehensive analysis can be conducted to compare the models.

The multicollinearity mainly arises from the method of data collection, the restrictions on the model or the population, the specification of the model or the over-defined model (Montgomery et al., 2012). Another reason for multicollinearity in a classical data set is that the number of explanatory variables is close to the number of the observations which cause the matrix $\mathbf{X}^\top\mathbf{X}$ approaches singularity where $\mathbf{X}$ is the design matrix of the explanatory variables. Redundant variables expressing the same information in different forms and closely linked explanatory variables under the studied system can also be the reasons for the multicollinearity. The multicollinearity leads to high variance of coefficient estimates and high covariance between the coefficient

estimates. It may also yield coefficient estimates with incorrect signs and implausible magnitudes; the minor changes in the data end up with very different coefficient estimates (Greene, 2000). Seber (1977), Hoerl and Kennard (1970) and Frank and Friedman (1993) discuss the multicollinearity in detail. Many diagnostic measures are proposed to detect the multicollinearity. Some of these measures are the correlation matrix of the data, variance inflation factors (VIF) (Marquardt, 1970), eigenvalues (Kendall, 1957) and variance decomposition proportions (Belsley et al., 1980).

In statistics, a data set is called as high-dimensional if the number of the explanatory variables is greater than the number of the observations (James et al., 2013). The high-dimensional data attracts great attention recently, as a result of arising extremely large data sets as a consequence of advances in technology. These data sets may consist of an enormous number of observations or explanatory variables. In many disciplines, a great deal of high-dimensional data exists (Fan and Tang, 2013). Among many others, some examples of these data sets are financial market data to construct and assess portfolio (Jagannathan and Ma, 2003), gene expression data from the leukemia microarray study for molecular classification of cancer (Golub et al., 1999), spatial earthquake data set for geographical analysis (van der Hilst et al., 2007) and near-infrared (NIR) spectra data to model some chemical compositions as a function of their NIR spectra (Gusnanto and Pawitan, 2015).

When the data set is high-dimensional, the OLS does not exist because the matrix $\mathbf{X}^\top\mathbf{X}$ is not invertible. One way to deal with this issue is to penalize the unrealistic values of the unknown coefficients, which entails penalized regression methods. The penalized regression methods may perform coefficient shrinkage or variable selection depending on the type of the penalization.

The best subset selection method corresponds to ℓ_0-norm penalization. This method aims to include the explanatory variables that have a strong effect on the

response variable by considering the regression models of response on all the combinations of the explanatory variables. Therefore, this method provides interpretable models in the sense of variable selection. However, the coefficient estimates obtained by this method are quite unstable, i.e. a small change in the data set leads to a completely different model. This entails poor prediction performance for the new observations. Also, the number of candidate models grows exponentially as the number of the explanatory variables increases. To obtain the sub-models when $p \geq 40$ with the classical approach is very hard for even developed computers; however, hundreds of explanatory variables can be tackled down by mixed-integer optimization (MIO) method proposed by Bertsimas et al. (2016).

Hoerl and Kennard (1970) proposed the ridge estimator which corresponds to ℓ_2-norm penalization. The authors pointed out that the ridge estimator can overcome the multicollinearity and high-dimensionality with better prediction performance than the OLS estimator. This is because the ridge estimator does variance reduction by using a penalty term. As the value of the penalty increases, the variance of the ridge estimator decreases. As a result of this, the ridge estimates are shrunken towards zero. Although the introduced penalty shrinks the ridge estimates towards zero, it does not set any coefficient exactly to zero. Hence, the estimator includes all of the explanatory variables in the final model and does not have the advantage of performing the model selection. This is a problematic situation with respect to model interpretation when the number of the explanatory variables is large.

Tibshirani (1996) established the LASSO estimator which is based on ℓ_1-norm penalization. The LASSO estimator reduces the number of unnecessary coefficients directly to zero as a result of the nature of the ℓ_1-norm. The estimator realizes dimension reduction so that it overcomes the disadvantage of the ridge estimator's inability to reduce the number of explanatory variables in the final model. Although the LASSO

estimator has become a very popular method in the high-dimensional linear models, it has some downsides with respect to the continuous shrinkage and automatic variable selection simultaneously. The optimization problem which leads to the LASSO estimator is not strictly convex. Therefore unique LASSO solution can not be obtained for some data structures. The LASSO estimator can not deal with the grouping effect when the explanatory variables are highly correlated. The ridge estimator performs better than the LASSO estimator if the explanatory variables are highly correlated. To handle these insufficiencies, various approaches have been proposed. These approaches are based on some modifications of the ℓ_1-norm penalization term.

Zou and Hastie (2005) proposed the elastic net estimator considering a linear combination of ℓ_1 and ℓ_2-norm penalization that can select all the variables even in the high-dimensional case and that works when the LASSO estimator can not tackle the grouping effect, and the ridge estimator outperforms the LASSO estimator in prediction performance.

Various generalizations of the LASSO estimator other than the elastic net estimator have been proposed depending on the properties of the data. In some linear regression problems, the explanatory variables have a natural group structure such as the levels of a categorical variable, in which levels are coded using a set of dummy variables and these variables form a group structure. In this case, it is appropriate that all the coefficients in a group are zero or nonzero, simultaneously. The group LASSO (Yuan and Lin, 2006) gets this objective by using the square root of ℓ_2-norm penalization instead of ℓ_1-norm penalization. Another generalization of the LASSO estimator is fused LASSO (Tibshirani et al., 2005) which is useful for the data sets such as time series or image-based data in which temporal or spatial structure must be taken into account during the analysis. The fused LASSO penalization is a linear combination of ℓ_1-norm term and a term based on the sum of the absolute values of

the differences of the successive coefficients.

In this thesis, methods that perform both shrinkage and variable selection simultaneously are discussed. The thesis consists of seven chapters which are organized as follows:

In Chapter 2, some preliminary concepts that are necessary and useful for penalized regression methods are mentioned such as comparison criteria and optimization techniques.

In Chapter 3, the concept of penalized regression and brief explanations about some well-known penalized regression methods in the literature that are the ridge, two-parameter, LASSO, and elastic net estimators are given.

In Chapter 4, novel estimators in penalized regression in the scope of the thesis are proposed. The estimators are called as the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators.

The relation of the GO estimator with the methods introduced in Chapter 3 is explained. The coordinate descent algorithm for the coefficient estimation of the GO estimator is proposed. Besides, a theorem about the grouping property for the GO estimator which includes the grouping property of the elastic net estimator is given.

The GO estimator can not be applied to the high-dimensional data sets. To deal with this deficiency, the shift-ridge and shift-LASSO estimators are proposed. The coordinate descent and ADMM algorithms for the coefficient estimation of the shift-ridge and shift-LASSO estimators are proposed. Also, a theorem about the grouping property of these estimators is given. The grouping property is a prominent property in a microarray data set. In medicine, a microarray data set has thousands of explanatory variables corresponding to the genes and a small number of observations. In this data set, the correlations between the genes from the same biological pathway can be high. Therefore, these genes correspond to a group of explanatory variables. In

this case, the shift-ridge and shift-LASSO estimators can eliminate the trivial genes and include the group of variables in the model in a similar way to the elastic net estimator. Hence, the shift-ridge and shift-LASSO estimators are alternatives to the elastic net estimator in the fields of genetics in medicine with respect to the variable selection and grouping property.

The stochastic restricted LASSO estimator is proposed by combining the ideas of the mixed and LASSO estimators. The coordinate descent algorithm for stochastic restricted LASSO estimator is given in this chapter.

In Chapter 5, real data analyses from the field of medicine for performance comparisons of the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators are performed on three different data sets. The prediction mean squared error and selected number of variables are used as comparison criteria for the GO, shift-ridge and shift-LASSO estimators. For the stochastic restricted LASSO estimator, the scalar mean squared error is also used as a comparison criterion. According to the results of the analyses, the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators yield better results compared to alternative methods. Therefore, the proposed estimators are of a qualification to shed light on data analysis in the field of medicine.

In Chapter 6, simulation studies are performed involving different data structures to compare the performances of the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators with other penalized regression methods. The results from each simulation study are commented and these comments are supported by tables and plots.

In Chapter 7, the findings obtained in the previous chapters of the thesis are summarized, the results are examined and finally the study is completed.

2. PRELIMINARY CONCEPTS

Consider the usual linear regression models. In these models, it is assumed that $n \times 1$ response vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ is a linear combination of $n \times 1$ explanatory variables $\mathbf{x}_j = \left(x_{1j}, x_{2j}, \ldots, x_{nj}\right)^\top$, $j = 1, 2, \ldots, p$ with an $n \times 1$ additive error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$. Errors are assumed to be independent, to have mean zero and constant variance, $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$ where $\mathbf{I}_n$ is an $n \times n$ identity matrix. Generally, a linear regression model has an intercept term corresponding to a vector of one's. Then, the linear regression model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1.}$$

where $\mathbf{X} = \left(\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\right)$ is called as design matrix of explanatory variables with $\mathbf{1}$ is an $n \times 1$ vector of one's, and $(p+1) \times 1$ vector $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^\top$ is the coefficient vector of the model.

In many studies, it is assumed that explanatory variables are in scaled form. Various scaling methods can be used to standardize the explanatory variables after the explanatory variables are centered to have mean zero, $n^{-1}\sum_{i=1}^{n} x_{ij} = 0$ for $j = 1, 2, \ldots, p$. Two of the commonly used scaling methods are unit-length scaling, $\sum_{i=1}^{n} x_{ij}^2 = 1$ and unit normal scaling, $n^{-1}\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, 2, \ldots, p$. Under the unit-length scaling, the matrix $\mathbf{X}^\top\mathbf{X}$ becomes the correlation matrix of the explanatory variables while under the unit-normal scaling, the explanatory variables have unit variance. The estimation of the intercept term is the mean of the observations of the response variable, $\widehat{\beta}_0 = \sum_{i=1}^{n} y_i/n$, when the explanatory variables are scaled. Generally, the response variable is centered to have mean zero, $\sum_{i=1}^{n} y_i/n = 0$ and the

intercept is omitted in the analysis. In the rest of this chapter, the explanatory variables are assumed to be scaled and the response variable is assumed to be centered. Therefore β_0 is omitted and, $\mathbf{X}$ and $\boldsymbol{\beta}$ are taken as $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\right)$ and $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^\top$, respectively.

The rest of this chapter is organized as follows. In Section 2.1 and Section 2.2, the concepts and methods used for comparing estimators and selecting a model from the candidate models are given. In Section 2.3, the resampling methods in the literature are mentioned. These methods are used for estimating the test error of a linear regression model or obtaining a measure of the accuracy of the model coefficients. In Section 2.4, some optimization concepts and methods are given. These concepts and methods are necessary for the coefficient estimation for the estimators that are proposed in Chapter 4.

2.1. Comparison Criteria

In linear regression analysis, there are alternative estimators or methods to carry out coefficient estimation of the model. In this situation, comparison criteria are used to select the best estimator or method. One of the criteria to compare the methods is the scalar mean squared error (SMSE) which considers both variance and bias of the estimator. The SMSE is defined for any estimator $\widehat{\boldsymbol{\beta}}$ of a coefficient vector $\boldsymbol{\beta}$ under the model given by Equation (2.1.) as

$$\text{SMSE} = \text{E}\left[\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_2^2\right] \tag{2.2.}$$

that is the expectation of the squared distance between the coefficients and their estimates where $\|\mathbf{a}\|_2^2$ represents the ℓ_2-norm of the vector $\mathbf{a}$ (Hastie et al., 2009). The

SMSE can be expressed in terms of bias and variance as

$$\text{SMSE} = \text{tr}\left(\text{var}\left(\widehat{\boldsymbol{\beta}}\right)\right) + \left\|\text{bias}\left(\widehat{\boldsymbol{\beta}}\right)\right\|_2^2. \tag{2.3.}$$

As can be seen from Equation (2.3.), the SMSE is equivalent to the variance of $\widehat{\boldsymbol{\beta}}$ if $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

The prediction mean squared error (PMSE) on the observation $\mathbf{x} = \mathbf{x_0}$ is defined as

$$\text{PMSE} = \text{E}\left[\left(y_0 - \mathbf{x}^\top \widehat{\boldsymbol{\beta}}\right)^2 \mid \mathbf{x} = \mathbf{x_0}\right] \tag{2.4.}$$

for the model given by Equation (2.1.) where y_0 is the corresponding response value (Hastie et al., 2009). Under the assumption that $\widehat{\boldsymbol{\beta}}$ and ε are independent, the decomposition of this expression is

$$\begin{aligned}
\text{PMSE} &= \text{E}\left(\mathbf{x_0}^\top \boldsymbol{\beta} + \varepsilon - \mathbf{x_0}^\top \widehat{\boldsymbol{\beta}}\right)^2 \\
&= \text{E}\left(\mathbf{x_0}^\top \boldsymbol{\beta} + \varepsilon - \mathbf{x_0}^\top \widehat{\boldsymbol{\beta}} + \text{E}\left(\mathbf{x_0}^\top \widehat{\boldsymbol{\beta}}\right) - \text{E}\left(\mathbf{x_0}^\top \widehat{\boldsymbol{\beta}}\right)\right)^2 \\
&= \text{E}\left(\text{E}\left(\mathbf{x_0}^\top \widehat{\boldsymbol{\beta}}\right) - \mathbf{x_0}^\top \widehat{\boldsymbol{\beta}}\right)^2 + \left(\mathbf{x_0}^\top \boldsymbol{\beta} - \text{E}\left(\mathbf{x_0}^\top \widehat{\boldsymbol{\beta}}\right)\right)^2 + \text{E}\left(\varepsilon^2\right) \\
&= \text{var}\left(\mathbf{x_0}^\top \widehat{\boldsymbol{\beta}}\right) + \text{bias}\left(\mathbf{x_0}^\top \widehat{\boldsymbol{\beta}}\right)^2 + \sigma^2.
\end{aligned}$$

The terms variance and bias of the right-hand side in the equation of PMSE stand for the reducible error while the variance σ^2 is the variance of the error, which represents the irreducible error. In general, as the complexity of a regression model increases, the

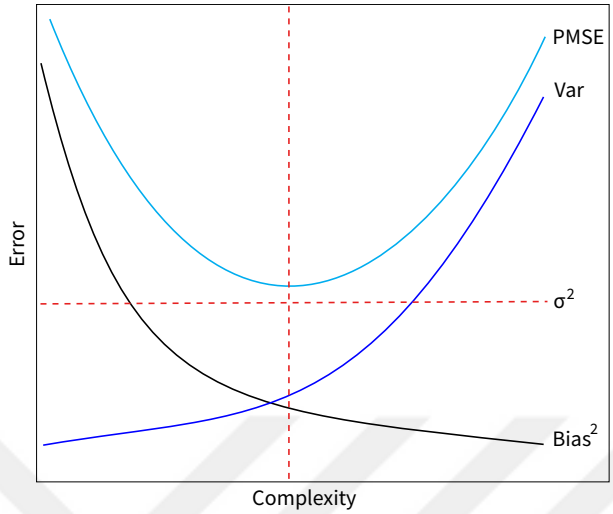

Figure 2.1. The PMSE, squared bias, variance curves and irreducible error line for a typical model. The vertical line denotes the optimal complexity point

squared bias decreases and the variance increases (James et al., 2013). This concept is known as bias-variance trade-off.

The analyst generally split the data set into the training and test sets. The model coefficients are estimated on the training set and the PMSE is computed on the test set. In this situation, the PMSE follows a U-shaped curve as a function of the complexity of the model. That means the PMSE has an optimal value at an optimal complexity point. Typical plots of the PMSE, squared bias, variance curves and irreducible error line on test set are given in Figure 2.1. The optimal complexity is represented by the vertical line.

The SMSE and PMSE are good measures to evaluate the performance of an estimator. In some situations, the analyst needs to evaluate the ability of an estimator in selecting the correct model with regard to variable selection. In these situations, there are several measures that can be used. Among many others such as true/false positive rates, true/false negative rates, precision, it can be used the accuracy and F

measures which are defined as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2.5.}$$

and

$$\text{F} = \frac{2}{\frac{1}{\text{pr}} + \frac{1}{\text{re}}},\ 0 \leq \text{F} \leq 1 \tag{2.6.}$$

where the precision (pr) and recall (re) are

$$\text{pr} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

and

$$\text{re} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP, TN, FP and FN represent true positives, true negatives, false positives and false negatives, respectively (see Chapter 8 of Manning et al., 2010 for details). An estimator is superior in the sense of variable selection if it has large values of accuracy and F measures.

2.2. Degrees of Freedom and Model Selection Criteria

The degrees of freedom is generally used to quantify the model complexity of a statistical modeling procedure (Hastie and Tibshirani, 1990). It can be used with an information criterion to decide the proper value of the tuning parameter. Let the response variable $\mathbf{y}$ has a distribution with the expectation $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$ and covariance matrix $\mathrm{cov}(\mathbf{y}) = \sigma^2 \mathbf{I}_n$, then the "degrees of the freedom" or "effective number of parameters" of a function $g : \mathbb{R}^n \to \mathbb{R}^n$ with $g(\mathbf{y}) = (g_1(\mathbf{y}), \ldots, g_n(\mathbf{y}))^\top$ is defined as

$$\mathrm{df}(g) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}(g_i(\mathbf{y}), y_i) \tag{2.7.}$$

where $g_i(\cdot) : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2, \ldots, n$ are some functions of $\mathbf{y}$ (Tibshirani and Taylor, 2012). If the fitted vector is a linear function of $\mathbf{y}$, that is $g(\mathbf{y}) = \mathbf{A}\mathbf{y}$ for some square matrix $\mathbf{A}$, then the degrees of freedom of the fitted vector can be written as

$$\mathrm{df}(g) = \mathrm{tr}(\mathbf{A}). \tag{2.8.}$$

For example, the degrees of freedom of the OLS fit on full column rank design matrix $\mathbf{X}$ is $\mathrm{df}(\hat{\mathbf{y}}) = \mathrm{df}(\mathbf{H}\mathbf{y}) = \mathrm{tr}(\mathbf{H}) = p$ where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$.

The information criteria are useful when the analyst would like to select a model from several candidate models to determine which of the candidate models fit the data best. These criteria are based on the log-likelihood of the response variable and degrees of the freedom of the model fit. The Akaike information criterion (AIC) proposed by Akaike (1973, 1974) is one of the best-known information criteria. The AIC is based

on the relationship between the Kullback-Leibler measure and the likelihood of the candidate model evaluated at the parameter vector. The mathematical expression of the AIC is

$$\mathrm{AIC} = -2\log L\left(\tilde{\boldsymbol{\beta}}\right) + 2\mathrm{df} \tag{2.9.}$$

where df is the degrees of freedom of the fitted model and $\tilde{\boldsymbol{\beta}}$ is estimated coefficient vector. The best model is the one that has the smallest AIC value when the AIC is the comparison criterion. As the degrees of freedom increases, the second term in Equation (2.9.) enlarges the AIC, that is the amount of penalization increases. This implies a trade-off between the better fit and the effective number of parameters. For the small samples, the corrected AIC can be used to prevent the AIC to overfit the data (Giraud, 2015):

$$\mathrm{AIC_c} = -2\log L\left(\tilde{\boldsymbol{\beta}}\right) + 2\mathrm{df} + \frac{2\mathrm{df}+1}{n-\mathrm{df}-1}$$

where n denotes the sample size.

The Bayesian information criterion (BIC) proposed by Schwartz (1978) is an alternative to the AIC for model selection. The BIC is defined as

$$\mathrm{BIC} = -2\log L\left(\tilde{\boldsymbol{\beta}}\right) + \mathrm{df}\cdot\log\left(n\right). \tag{2.10.}$$

As can be seen easily from Equations (2.9.) and (2.10.), the BIC imposes a greater

penalty for the effective number of the parameters than the AIC for the large values of n. Comparisons of the AIC and BIC in different perspectives can be found in Burnham and Anderson (2002) and Yang (2005).

Under the assumption that the errors have a normal distribution $\mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_p\right)$, the response variable has the distribution $\mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_p\right)$ when the design matrix $\mathbf{X}$ is fixed. Then the log-likelihood in Equation (2.9.) can be written in terms of RSS as

$$\log L\left(\tilde{\boldsymbol{\beta}}\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\text{RSS}$$

where $\text{RSS} = \left\|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right\|_2^2$. Considering the maximum likelihood estimate of σ^2, $\tilde{\sigma}^2 = \text{RSS}/n$ and canceling data-dependent terms, the log-likelihood becomes $\log L\left(\tilde{\boldsymbol{\beta}}\right) = -\frac{n}{2}\log \text{RSS}$. Therefore the AIC in Equation (2.9.) has the form $\text{AIC} = n \log \text{RSS} + 2\text{df}$ for the model given by Equation (2.1.). Similar statements can be given for the AIC_c and BIC based on the linear model given by Equation (2.1.).

2.3. Resampling Methods (Out-of-Sample Testing)

Resampling methods are based on repeatedly drawing samples from a data set and refitting a model on each sample to obtain additional information about the fitted model (James et al., 2013). The resampling methods aim to reduce the effect of overfitting and provide intuitive way about the generalization of the model to an independent data (Cawley and Talbot, 2010).

In this section, the validation set, leave-one-out cross-validation, k-fold cross-validation and bootstrap resampling methods are discussed. The first three methods are used to estimate the test error associated with an estimator in order to evaluate its performance and the bootstrap method is used to provide a measure of the accuracy of

the coefficient estimates for the estimators in the scope of this thesis.

2.3.1. The Validation Set Method

The validation set method is a simple method to estimate the test error. To implement this approach, the data set is split randomly into a training set and an independent validation set. The model is fitted on the training set. The fitted model is made use of predicting the observations in the validation set. Then the validation set error is computed as an estimate of the test error.

The validation set method has two drawbacks (James et al., 2013):

(a) The test errors obtained by this method are highly variable depending on the position of the observations (in the training or validation set.)

(b) Since this method uses a subset of the data set as the training set, the parameter estimation method on the training set is expected to perform worse than that of the full data set. This results in an overestimate of the test error.

2.3.2. Leave-One-Out Cross-Validation (LOOCV)

LOOCV is based on a combination of a special type of the validation set method. LOOCV splits the data set into a training set with $n-1$ observations and a validation set with a single observation. This method estimates the parameters in the training set and computes the validation set error, PMSE, on the single observation as in the validation set method. Unlike the validation set method, LOOCV repeats this process n times in a cyclical manner such that all the observations form the validation set once. Then LOOCV estimate of the test error is obtained as the mean of these n

validation error estimates (James et al., 2013):

$$\mathrm{CV}_n = \frac{1}{n}\sum_{i=1}^{n}\mathrm{PMSE}_i.$$

LOOCV yields less bias compared to the validation set method because LOOCV uses more observations to estimate the parameters. Also, LOOCV has no randomness in the training and test splitting which means it always yields the same estimate of the test error. However, the cost of this method is very high especially when n is large and each model is slow to fit.

2.3.3. k-Fold Cross-Validation

An alternative method to the validation set method and LOOCV is k-fold cross-validation. The method is based on splitting the data into approximately equal-sized k folds. Similar to LOOCV, a fold is held out for validation set and the parameter estimates are obtained from the union of the remaining $k-1$ folds. In this way, k validation set errors, PMSE, are obtained. Then, k-fold cross-validation estimate of the test error is computed as the mean of these k validation set errors (James et al., 2013):

$$\mathrm{CV}_k = \frac{1}{k}\sum_{i=1}^{k}\mathrm{PMSE}_i.$$

This method increases the variability of the test error compared to LOOCV because of the randomness in the data set splittings. However, this variability in the test error estimates is much lower than that of the one obtained from the validation set method.

Also, k-fold cross-validation outperforms the LOOCV with regard to the computation time when n is large.

2.3.4. The Bootstrap Method

The bootstrap method proposed by Efron (1979) is a statistical tool that can be used to quantify the uncertainty of an estimator. When an estimator does not have a closed-form expression, the bootstrap method can be used to obtain the standard error estimates of the coefficients of the model obtained by using that estimator.

Suppose that a model is fitted to a data set. Then, the analyst needs to compute the standard error estimates of the model coefficients. In this case, the bootstrap method can be used as detailed by following. The first step is to randomly draw data sets with replacement from the original data set, such that each sample has the same size as the original data set. This procedure is done B times (say, $B = 1000$), which produces B bootstrap data sets. After that, the model is refitted to each of the bootstrap data sets. Let $\beta^*_{j1}, \beta^*_{j2}, \ldots, \beta^*_{jB}$ be the bootstrap estimates of the j^{th} coefficient of the model. Then the bootstrap estimate of the standard error of the j^{th} coefficient estimate of the model is obtained as

$$\text{se}\left(\hat{\beta}_j\right) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\beta^*_{jr} - \overline{\beta^*_j}\right)}$$

where $\overline{\beta^*_j} = \sum_{r=1}^{B} \beta^*_{jr}$ (Efron and Tibshirani, 1986).

2.4. Optimization Methods

In classical regression methods, the optimization problems that yield the coefficient estimates generally have a closed-form solution which is easily obtained by

calculus and matrix algebra. However, for some methods including ℓ_1-norm of the coefficients, the problems have no closed-form solution which leads the analyst to make use of more complicated optimization methods. In this section, first, some optimization concepts are mentioned. Then, some optimization methods in the literature which deal with the convex problems that have no closed-form solution are given.

Definition 2.1. (Boyd, 2004) Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are arbitrary $p \times 1$ vectors. If

$$f\left(w\boldsymbol{\theta} + (1-w)\boldsymbol{\theta}'\right) \leq wf\left(\boldsymbol{\theta}\right) + (1-w)f\left(\boldsymbol{\theta}'\right)$$

holds for any scalar $w \in (0,1)$ then the function $f : \mathbb{R}^p \to \mathbb{R}$ is called a convex function. Let f be a convex function. A convex optimization problem in constrained form with inequality constraints is defined as

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} f\left(\boldsymbol{\theta}\right) \quad \text{s.t.} \quad v_k\left(\boldsymbol{\theta}\right) \leq 0 \quad \text{for } k = 1,2,\ldots,m \tag{2.11.}$$

where $v_k,\ k = 1,2,\ldots,m$ are called constraint functions, which are convex functions representing constraints to be satisfied.

Definition 2.2. (Boyd, 2004) Let $\boldsymbol{\gamma}$ be an $m \times 1$ vector with nonnegative numbers and $v_k, k = 1,2,\ldots,m$ are constraint functions in the optimization problem given by Equation (2.11.). Then the function

$$L\left(\boldsymbol{\theta},\boldsymbol{\gamma}\right) = f\left(\boldsymbol{\theta}\right) + \sum_{k=1}^{m} \gamma_k v_k\left(\boldsymbol{\theta}\right) \tag{2.12.}$$

is called as the Lagrange function associated with the problem given by Equation (2.11.). The components of $\boldsymbol{\gamma}$ are called as Lagrange multipliers.

The Lagrange functions are used to solve the constrained optimization problems by reducing the problems to equivalent unconstrained problems. The Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient conditions for $\boldsymbol{\theta}$ to be the optimal parameter vector of the problem given by Equation (2.11.) if the optimization problem satisfies the strong duality condition (Hastie et al., 2015). The conditions are first published by Kuhn and Tucker (1951), however, it was discovered that the necessary conditions for this problem had been stated by Karush (1939). The KKT conditions relate the optimal Lagrange multiplier vector $\boldsymbol{\gamma}$ (dual vector) to the optimal vector $\boldsymbol{\theta}$ (primal vector). These conditions are given as follows (Boyd, 2004):

1. $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ satisfies $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \mathbf{0}$ (Stationary)
2. $\gamma_k v_k(\boldsymbol{\theta}) = 0$ (Complementary slackness)
3. $v_k(\boldsymbol{\theta}) \leq 0$ (Primal feasibility)
4. $\gamma_k \geq 0$ (Dual feasibility)

for $k = 1, 2, \ldots, m$.

In some optimization problems, the Lagrange function in Equation (2.12.) is non-differentiable. In these cases, the stationary condition of KKT can not be applied to the problem directly. One of the methods to deal with the non-differentiable functions in optimization problems is based on subdifferential given by Definition 2.3.

Definition 2.3. (Hastie et al., 2015) Given a convex function f, a vector $\mathbf{z} \in \mathbb{R}^p$ is called as subgradient of f at $\boldsymbol{\theta} \in \mathbb{R}^p$ if

$$f\left(\boldsymbol{\theta}'\right) \geq f(\boldsymbol{\theta}) + \mathbf{z}^{\top}\left(\boldsymbol{\theta}' - \boldsymbol{\theta}\right)$$

for all $\boldsymbol{\theta}' \in \mathbb{R}^p$. The subdifferential denoted by $\partial f(\boldsymbol{\beta})$ is defined as the set of all subgradients of f at $\boldsymbol{\theta}$.

The subdifferential is always a nonempty convex set (Hiriart-Urruty, Lemaréchal, 2012). When the Lagrange function is non-differentiable, KKT conditions can be applied with the modified stationary condition

$$\mathbf{0} \in \partial f\left(\boldsymbol{\theta}^*\right) + \sum_{k=1}^{m} \gamma_k \cdot \partial v_k\left(\boldsymbol{\theta}\right)$$

under mild conditions (Hastie et al., 2015). Therefore, the optimization problem given by Equation (2.11.) can be solved by using the subdifferential of the Lagrange function given by Equation (2.12.) when the Lagrange function is non-differentiable.

2.4.1. Iterative Algorithms

Since the absolute value function is non-differentiable at zero, ℓ_1-norm problems do not have closed-form solutions except in some special cases such as single explanatory variable and orthogonal design cases. There are many optimization methods that can deal with non-differentiability of ℓ_1-norm such as local quadratic approximations (Fan and Li, 2001), least angle regression (LARS) (Efron et al., 2004), coordinate descent (Friedman et al., 2007), ADMM (Boyd et al., 2011) algorithms. In this section, the coordinate descent and ADMM algorithms are addressed to solve ℓ_1-norm problems. By using these algorithms, it is aimed to obtain the coefficient estimates of the estimators given in Chapter 4.

2.4.1.1. The Coordinate Descent Algorithm

The coordinate descent algorithm is a natural choice to solve an optimization problem involving ℓ_1-norm of the parameters. The coordinate descent algorithm can be used effectively when the problem has the separability property given by Definition 2.4.

Definition 2.4. (Hastie et al., 2015) A function $f : \mathbb{R}^p \to \mathbb{R}$ has separability property if it has an additive decomposition in the form of

$$f(\boldsymbol{\theta}) = u(\boldsymbol{\theta}) + \sum_{j=1}^{p} h_j(\theta_j)$$

where $u : \mathbb{R}^p \to \mathbb{R}$ is a convex, differentiable function, $h_j : \mathbb{R} \to \mathbb{R}$ are convex functions and $\boldsymbol{\theta}$ is a $p \times 1$ vector of unknown parameters.

For any convex function f with separability property, the coordinate descent algorithm is guaranteed to converge to the minimizer of the function (Tseng, 2001). The coordinate descent is an iterative algorithm that updates unknown vector $\boldsymbol{\theta}$ by choosing a single coordinate and then performs univariate minimization for this coordinate. Therefore, the update rule of the coordinate descent algorithm is given by

$$\begin{aligned}
\theta_1^{(k+1)} &= \underset{\beta_1}{\operatorname{argmin}} f\left(\theta_1, \theta_2^{(k)}, \theta_3^{(k)}, \ldots, \theta_p^{(k)}\right) \\
\theta_2^{(k+1)} &= \underset{\beta_2}{\operatorname{argmin}} f\left(\theta_1^{(k+1)}, \theta_2, \theta_3^{(k)}, \ldots, \theta_p^{(k)}\right) \\
&\vdots \\
\theta_p^{(k+1)} &= \underset{\beta_p}{\operatorname{argmin}} f\left(\theta_1^{(k+1)}, \theta_2^{(k+1)}, \theta_3^{(k+1)}, \ldots, \theta_p\right)
\end{aligned}$$

where $k = 0, 1, 2, \ldots$ is the iteration counter (Wright, 2015).

2.4.1.2. The Alternating Direction Method of Multipliers

The alternating direction method of multipliers (ADMM) proposed by Boyd et al. (2011) is one of the Lagrangian based methods. It has some attractive properties for large-scale applications. Following Boyd et al. (2011), the ADMM can be explained as follows:

Consider a problem of the form

$$\begin{aligned} &\underset{\boldsymbol{\theta},\boldsymbol{\xi}}{\text{argmin}} && u(\boldsymbol{\theta}) + h(\boldsymbol{\xi}) \\ &\text{subject to} && \mathbf{A}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\xi} = \mathbf{c} \end{aligned} \tag{2.13.}$$

over the parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ with sizes $p \times 1$ and $q \times 1$, respectively. The functions $u : \mathbb{R}^p \to \mathbb{R}$ and $h : \mathbb{R}^q \to \mathbb{R}$ are closed convex functions, $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $\mathbf{B} \in \mathbb{R}^{n \times q}$ are known matrices of constraints and $\mathbf{c}$ is an $n \times 1$ known vector. The corresponding augmented Lagrange function is

$$L_\tau(\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{u}) = u(\boldsymbol{\theta}) + h(\boldsymbol{\xi}) + \boldsymbol{\gamma}^\top (\mathbf{A}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\xi} - \mathbf{c}) + \frac{\tau}{2} \|\mathbf{A}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\xi} - \mathbf{c}\|_2^2$$

where $\boldsymbol{\gamma}$ is the vector of Lagrange multipliers associated with the constraints and $\tau > 0$ is a fixed penalty parameter. The ADMM introduces the augmented quadratic term $\|\mathbf{A}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\xi} - \mathbf{c}\|_2^2$ involving τ which makes constraint in a smoother fashion.

The ADMM iterations consist of primal updates over $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ and a dual update over $\boldsymbol{\gamma}$ which are given as follows:

$$
\begin{aligned}
\boldsymbol{\theta}^{(k+1)} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, L_{\rho}\left(\boldsymbol{\theta}, \boldsymbol{\xi}^{(k)}, \boldsymbol{\gamma}^{(k)}\right) \\
\boldsymbol{\xi}^{(k+1)} &= \underset{\boldsymbol{\xi}}{\operatorname{argmin}}\, L_{\rho}\left(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\xi}, \boldsymbol{\gamma}^{(k)}\right) \\
\boldsymbol{\gamma}^{(k+1)} &= \boldsymbol{\gamma}^{(k)} + \tau\left(\mathbf{A}\boldsymbol{\theta}^{(k+1)} + \mathbf{B}\boldsymbol{\xi}^{(k+1)} - \mathbf{c}\right)
\end{aligned}
\tag{2.14.}
$$

where $k = 0, 1, 2, \ldots$ is the iteration counter. Under slightly mild conditions, it can be shown that the procedure given by Equation (2.14.) converges to an optimal solution to the problem given by Equation (2.13.) (Boyd et al., 2011). Employing the separation of the parameters into $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the non-differentiable constraints can be tackled. The ADMM can also split up a large problem into smaller sub-problems with respect to the observations or explanatory variables.

It is generally convenient to use the scaled form of the ADMM since it is often simple to express. The iterations in this form are obtained by the transformation $\boldsymbol{\gamma}_{*}^{(k)} = (1/\tau)\,\boldsymbol{\gamma}^{(k)}$ on the dual variable. The scaled ADMM iterations are as follows:

$$
\begin{aligned}
\boldsymbol{\theta}^{(k+1)} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, u\left(\boldsymbol{\theta}\right) + \frac{\tau}{2}\left\|\mathbf{A}\boldsymbol{\theta} + \mathbf{B}\boldsymbol{\xi}^{(k)} - \mathbf{c} + \boldsymbol{\gamma}_{*}^{(k)}\right\|_{2}^{2} \\
\boldsymbol{\xi}^{(k+1)} &= \underset{\boldsymbol{\xi}}{\operatorname{argmin}}\, h\left(\boldsymbol{\xi}\right) + \frac{\tau}{2}\left\|\mathbf{A}\boldsymbol{\theta}^{(k+1)} + \mathbf{B}\boldsymbol{\xi} - \mathbf{c} + \boldsymbol{\gamma}_{*}^{(k)}\right\|_{2}^{2} \\
\boldsymbol{\gamma}_{*}^{(k+1)} &= \boldsymbol{\gamma}_{*}^{(k)} + \mathbf{A}\boldsymbol{\theta}^{(k+1)} + \mathbf{B}\boldsymbol{\xi}^{(k+1)} - \mathbf{c}
\end{aligned}
\tag{2.15.}
$$

where $\tau > 0$ is a fixed penalty parameter (Boyd et al., 2011).

3. PENALIZED REGRESSION

In linear regression setting, the regression coefficients are unknown and must be estimated. The idea in estimating the regression coefficients is to find an estimator of coefficient vector $\boldsymbol{\beta}$ such that the resulting plane is as close as possible to the data points. Although there are many ways of measuring closeness, the OLS criterion is a common approach. The OLS estimation is the oldest method used to estimate the coefficients of a linear regression model. The OLS estimator is obtained by minimizing the objective function

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{3.1.}$$

which has the closed-form solution

$$\hat{\boldsymbol{\beta}}^{ls} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

given that the design matrix $\mathbf{X}$ is full column rank. Gauss-Markov theorem states that the OLS is the best linear unbiased estimator (BLUE) in the sense of the minimum variance among the class of unbiased estimators under the assumptions $\boldsymbol{\varepsilon} \sim \left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$ and $\mathbf{X}$ is full column rank. Although the OLS is BLUE, it has limitations if the multicollinearity or high-dimensionality exists in the data set.

In the case of multicollinearity, the matrix $\mathbf{X}^\top \mathbf{X}$ is almost singular and the OLS coefficients have very large variance which causes the OLS coefficients to be very unstable. As a result of this, the resulting model exhibits poor prediction accuracy.

If the data set is high-dimensional, $\mathbf{X}^\top \mathbf{X}$ is not invertible and the OLS can not

be computed. Indeed, if M is the space spanned by the columns of $\mathbf{X}$ and $M^{\perp}$ is the orthogonal complement of M then for all $\mathbf{m} \in M$

$$\widehat{\boldsymbol{\beta}}^{m} = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{+}\mathbf{X}^{\top}\mathbf{y} + \mathbf{m}$$

is a solution to the minimization problem of the function given by Equation (3.1.) where $\left(\mathbf{X}^{\top}\mathbf{X}\right)^{+}$ denotes Moore-Penrose inverse of $\mathbf{X}^{\top}\mathbf{X}$.

Moreover, the OLS estimates all of the model coefficients as nonzero values. Hence, the OLS does not perform variable selection, hence does not lead to a simple model that identifies the explanatory variables which indeed explain the response variable. Alternative estimators to the OLS are proposed to overcome the deficiencies of the OLS in the cases of multicollinearity or high-dimensionality in the data set and lack of variable selection property.

Among many others, some estimators are based on exact and stochastic restrictions on the regression coefficients. These estimators may outperform the OLS by using additional information about the coefficient vector.

In real-life applications, the analyst may encounter additional information on $\boldsymbol{\beta}$ rather than the model given by Equation (2.1.). In such cases, the analyst has to use additional information in estimating regression coefficients. The additional information may appear in different forms. The additional information can be in the form of

$$\mathbf{h} = \mathbf{H}\boldsymbol{\beta} \tag{3.2.}$$

where $\mathbf{h}$ is a $J \times 1$ known vector, $\mathbf{H}$ is a $J \times p$ known matrix with rank $J < p$. This type of information is known as exact linear restrictions (see Rao and Toutenburg, 1995). However, in some cases, the analyst may be unsure about the exactness of the linear restrictions. If there is uncertainty about the exactness of the linear restrictions, then the restrictions will be in the form

$$\mathbf{h} = \mathbf{H}\boldsymbol{\beta} + \mathbf{u}, \qquad \mathbf{u} \sim \left(\mathbf{0}, \sigma^2 \mathbf{I}_J\right) \tag{3.3.}$$

where $\mathbf{u}$ is a $J \times 1$ random vector and independent from $\boldsymbol{\varepsilon}$. The variance of $\mathbf{u}$ captures uncertainty about the information. The restriction given by Equation (3.3.) is known as stochastic linear restriction. By plugging the stochastic restrictions, extra information can be included in the model, which may come from

(a) experience; that is, previous estimates of similar models or different samples,

(b) the regression coefficients or only a part of regression coefficients where the analyst has uncertainty on the information.

Estimations with the restrictions given by Equations (3.2.) and (3.3.) have been considered by researchers in linear regression model. If the error sum of squares is minimized under the exact linear restrictions in Equation (3.2.), the restricted least squares (RLS) can be obtained as (see Rao and Toutenburg, 1995)

$$\widehat{\boldsymbol{\beta}}^{rls} = \widehat{\boldsymbol{\beta}} + \mathbf{S}^{-1}\mathbf{H}^{\top}\left(\mathbf{H}\mathbf{S}^{-1}\mathbf{H}^{\top}\right)^{-1}\left(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}}\right)$$

where $\mathbf{S} = \mathbf{X}^{\top}\mathbf{X}$ and $\widehat{\boldsymbol{\beta}} = \mathbf{S}^{-1}\mathbf{X}^{\top}\mathbf{y}$ is the OLS estimator.

Theil and Goldberger (1961) are the first to use stochastic linear restrictions. They combine the model given by Equation (2.1.) with $\boldsymbol{\varepsilon} \sim \left(\mathbf{0}, \sigma^2\mathbf{V}\right)$ and information (3.3.) with $\boldsymbol{u} \sim \left(\mathbf{0}, \sigma^2\mathbf{W}\right)$ by writing

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{h} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{H} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon} \\ \mathbf{u} \end{pmatrix}, \qquad \begin{pmatrix} \boldsymbol{\varepsilon} \\ \mathbf{u} \end{pmatrix} \sim \left[\begin{pmatrix} \mathbf{0}_n \\ \mathbf{0}_J \end{pmatrix}, \begin{pmatrix} \sigma^2\mathbf{V} & \mathbf{0}_n \\ \mathbf{0}_J & \sigma^2\mathbf{W} \end{pmatrix} \right]$$

and applying the least squares method and obtain an estimator called mixed estimator (ME) as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^m &= \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X} + \mathbf{H}^\top\mathbf{W}^{-1}\mathbf{H}\right)^{-1}\left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y} + \mathbf{H}^\top\mathbf{W}^{-1}\mathbf{h}\right) \\ &= \widehat{\boldsymbol{\beta}}^g + \mathbf{G}^{-1}\mathbf{H}^\top\left(\mathbf{W} + \mathbf{H}\mathbf{G}^{-1}\mathbf{H}^\top\right)^{-1}\left(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}}^g\right) \end{aligned}$$

where $\mathbf{G} = \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}$ and $\widehat{\boldsymbol{\beta}}^g = \mathbf{G}^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}$ is the generalized least squares estimator. This estimator has the property of being a best linear unbiased estimator if the restrictions hold true (Rao and Toutenburg, 1995).

To improve the OLS, RLS and ME estimators when the multicollinearity or high-dimensionality exists in the data set, the penalized regression methods can be used. Also, the penalized regression methods may perform coefficient estimation and variable selection simultaneously, depending on the type of the penalization.

In Section 3.1, a brief explanation of the penalized regression methods is given. Also, well-known ridge, LASSO, two-parameter and elastic net estimators and their properties are examined in the penalized regression methods scope. Also, the idea of the exact or linear restricted linear regression models can be combined with the

penalized regression methods. The methods based on this combination can be named as restricted-penalized regression methods. In Section 3.2 some restricted-penalized regression methods in the literature are examined.

3.1. Penalized Regression Methods

Penalized regression methods have been widely used when multicollinearity exists in the data or the case of high-dimensional data as alternative methods to the OLS. Unlike traditional variable selection approaches, the automatic model selection is performed simultaneously with the estimation of the parameters in some penalized regression methods.

In penalized regression methods, the regression coefficients are estimated by using some penalties on the coefficients. In these methods, it is generally aimed to decrease the variance of the coefficients by introducing some bias to the regression coefficients. In penalized regression methods, the objective function is defined as

$$Q\left(\boldsymbol{\beta};\lambda\right)^{2} = \frac{1}{2n}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_{2}^{2} + \lambda P\left(\boldsymbol{\beta}\right) \tag{3.4.}$$

where $P\left(\boldsymbol{\beta}\right)$ is the penalization function and $\lambda \geq 0$ is known as the tuning (biasing, regularization, complexity) parameter (Kock et al., 2020).

The penalization function is a positive valued function that penalizes the unrealistic (generally large) values of the coefficients and the tuning parameter controls the trade-off between RSS $= \frac{1}{2n}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_{2}^{2}$ and the penalty part in Equation (3.4.).

[2]In some descriptions, the factor $\frac{1}{2n}$ in Equation (3.4.) can be replaced by $\frac{1}{2}$ or 1. According to Hastie et al., (2015), this case makes no difference in the objective function and corresponds to a reparametrization of λ in Equation (3.4.). The advantage of the form in Equation (3.4.) is stated by Hastie et al., (2015) as that this kind of standardization makes λ values comparable for different sample sizes.

There are various penalization functions that penalize or shrink the coefficients in different manners. The effect can be seen in the coefficient estimates and predictions.

By applying penalized regression methods, one can obtain a lower SMSE or PMSE by properly chosen penalization function. In Subsections 3.1.1-3.1.4, the ridge, LASSO, two-parameter and elastic net estimators in the scope of penalized regression methods are given.

3.1.1. The Ridge Estimator

One way to penalize the large values of the coefficients is to use the penalization function $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$ which leads to the ridge estimator of Hoerl and Kennard (1970). Under the linear model given by Equation (2.1.) the ridge estimator is based on minimizing the objective function

$$Q(\boldsymbol{\beta};\lambda) = \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2 \tag{3.5.}$$

where $\lambda \geq 0$. This type of penalization is particularly attractive since the minimization problem of the function given by Equation (3.5.) has the closed-form solution

$$\widehat{\boldsymbol{\beta}}^r(\lambda) = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{y}. \tag{3.6.}$$

The ridge estimator aims to reduce the condition number (square root of the ratio of the largest and smallest eigenvalues of $\mathbf{X}^\top\mathbf{X}$) by adding λ to the diagonal elements of the matrix $\mathbf{X}^\top\mathbf{X}$. The estimator corresponds to the OLS for $\lambda = 0$. Since the matrix $\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p$ is always non-singular, there exists a unique ridge estimator for $\lambda > 0$.

The expectation of the ridge estimator is

$$\mathrm{E}\left(\widehat{\boldsymbol{\beta}}^{r}(\lambda)\right)=\left(\frac{1}{n}\mathbf{X}^{\top}\mathbf{X}+\lambda\mathbf{I}_{p}\right)^{-1}\frac{1}{n}\mathbf{X}^{\top}\mathbf{X}\boldsymbol{\beta}$$

Therefore, it is obvious that the ridge estimator is biased. The ridge estimator implements more shrinkage to the coefficients as the value of λ becomes large because $\mathrm{E}\left(\widehat{\boldsymbol{\beta}}^{r}(\lambda)\right)\rightarrow\mathbf{0}$ as $\lambda\rightarrow\infty$.

The main result of the ridge estimator is about the SMSE comparison with the OLS. Theobald (1974) shows that there always exists a tuning parameter $\lambda>0$ such that $\mathrm{SMSE}\left(\widehat{\boldsymbol{\beta}}^{r}(\lambda)\right)<\mathrm{SMSE}\left(\widehat{\boldsymbol{\beta}}^{ls}\right)$, that is the ridge estimator attains a lower SMSE than the OLS for some $\lambda>0$. As long as the reduction in the variance is more than the squared increase of bias (bias-variance trade-off), the ridge estimator is better than the OLS in the sense of lower SMSE.

3.1.1.1. Constrained and Bayesian Form of the Ridge Estimator

The minimization problem of the function given by Equation (3.5.) is equivalent to the constrained optimization problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_{2}^{2}\quad\text{s.t.}\quad\|\boldsymbol{\beta}\|_{2}^{2}\leq t$$

where $t>0$ is an upper bound for the sum of the squares of the coefficients. The analysis of constraint form provides insight about the nature of the methods. Figure 3.1(a) picturizes the situation for two explanatory variables case. In Figure 3.1(a), the ridge estimator corresponds to the point at which the contours from the residual sum

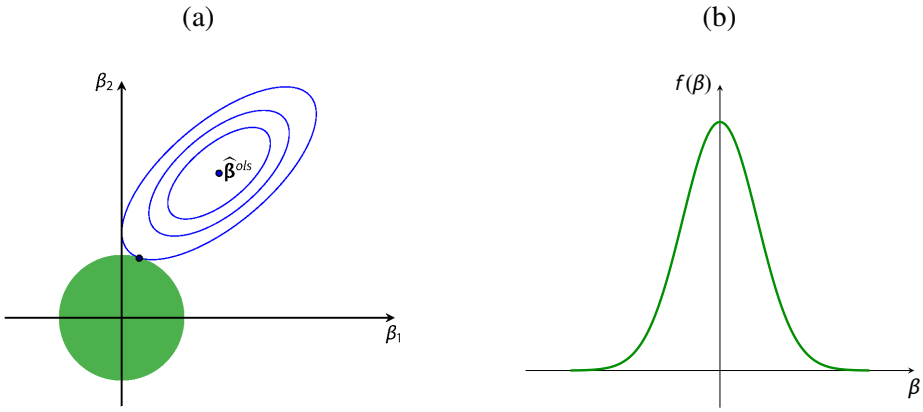

Figure 3.1. (a) Constraint region and contours for the ridge estimator: the circular region corresponds to the constraint $\beta_1^2 + \beta_2^2 \leq t$, the contours correspond to the residual sum of squares (b) Density plot of the normal distribution $\mathcal{N}\left(0, \sigma_r^2\right)$ (adapted from Tibshirani, 1996)

of squares hit the circle representing the constraint $\|\boldsymbol{\beta}\|_2^2 \leq t$ for the first time.

The Bayesian analysis is based on a distributional assumption about the regression coefficients. For the ridge estimator case, let the coefficients follow a normal distribution (Figure 3.1(b)). Let β_j, $j = 1, 2, \ldots, p$ have a prior distribution that is identically and independent normal random variables with mean zero and variance σ_r^2:

$$f\left(\beta_j\right) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \exp\left(-\frac{1}{2}\left(\frac{\beta_j^2}{\sigma_r^2}\right)\right).$$

Then the posterior distribution of $\boldsymbol{\beta}$ given $\sigma^2, \mathbf{y}$ and $\mathbf{X}$ is

$$\begin{aligned} f\left(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}\right) &= f_{\mathbf{y}}\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2\right) f_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right) \\ &\propto \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{2\sigma_r^2}\|\boldsymbol{\beta}\|_2^2\right] \end{aligned}$$

(see van Wieringen, 2015). To obtain the maximum a posteriori (MAP) estimator, $\widehat{\boldsymbol{\beta}}^{map}$, the derivative of this expression with respect to $\boldsymbol{\beta}$ is taken and set equal to zero:

$$\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y} - \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \frac{1}{\sigma_r^2}\widehat{\boldsymbol{\beta}}^{map} = \mathbf{0}$$

hence,

$$\widehat{\boldsymbol{\beta}}^{map} = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{n\sigma_r^2}\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}^\top\mathbf{y}$$

which corresponds to $\widehat{\boldsymbol{\beta}}^{r}(\lambda)$ with $\lambda = \frac{\sigma^2}{n\sigma_r^2}$. The MAP estimator minimizes the risk $\text{MSE} = \text{E}\left\|\widehat{\boldsymbol{\beta}}^{r}(\lambda) - \boldsymbol{\beta}\right\|_2^2$ where the expectation is computed over the joint distribution of $\boldsymbol{\beta}$ and $\mathbf{y}$. As a result, the ridge estimator has the minimal achievable risk under the identically and independent normal prior distribution for $\boldsymbol{\beta}$ (Van Wieringen, 2015).

3.1.1.2. The Ridge Estimator in Orthonormal Design

In the orthonormal design case, the nature of the shrinkage can be seen more clearly. If the matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthonormal, the ridge estimator becomes

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{ro} &= \left(\frac{1}{1+\lambda}\right)\frac{1}{n}\mathbf{X}^\top\mathbf{y} \\ &= \frac{1}{1+\lambda}\widehat{\boldsymbol{\beta}}^{lso}\end{aligned} \tag{3.7.}$$

where

$$\widehat{\boldsymbol{\beta}}^{lso} = \frac{1}{n}\mathbf{X}^{\top}\mathbf{y} \tag{3.8.}$$

is the OLS estimator in the orthonormal design case.

Mayer and Willke (1973) defines the deterministically shrunken estimator as

$$\widehat{\boldsymbol{\beta}}^{se} = d\,\widehat{\boldsymbol{\beta}}^{ls} \tag{3.9.}$$

Thus, in the orthonormal design case, the ridge estimator corresponds to a deterministically shrunken estimator with a factor $d = (1+\lambda)^{-1}$ in Equation (3.9.). This corresponds to a rotation of the OLS vector clockwise by the angle $\tan^{-1}(\lambda/(2+\lambda))$. Since $\widehat{\boldsymbol{\beta}}^{lso}$ is an unbiased estimator for the actual parameter values, a bias occurs for each $\lambda > 0$ in the estimator given by Equation (3.7.); but the variances of the coefficients decrease.

3.1.1.3. Degrees of Freedom of the Ridge Estimator

For the ridge estimator, the fitted values have the form of $g(\mathbf{y}) = \mathbf{H}_{\lambda}\mathbf{y}$ where $\mathbf{H}_{\lambda} = \frac{1}{n}\mathbf{X}\left(\frac{1}{n}\mathbf{X}^{\top}\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^{\top}$. By using Equations (2.7.) and (2.8.), the degrees of freedom of the ridge fit can be obtained as

$$\mathrm{df}\left(\widehat{\mathbf{y}}^{r}\right) = \mathrm{tr}\left(\frac{1}{n}\mathbf{X}\left(\frac{1}{n}\mathbf{X}^{\top}\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^{\top}\right) \tag{3.10.}$$

where $\hat{\mathbf{y}}^r = \mathbf{X}\hat{\boldsymbol{\beta}}^r$ is the ridge fit.

The singular value decomposition of $\mathbf{X}$ has the form $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Here, $\mathbf{U}$ and $\mathbf{V}$ are $n \times p$ and $p \times p$ orthogonal matrices. $\mathbf{D}$ is a $p \times p$ diagonal matrix with nonnegative diagonal entries in decreasing order. If the singular value decomposition of $\mathbf{X}$ is applied to Equation (3.10.), it is obtained that

$$\begin{aligned} \mathrm{df}_r = \mathrm{df}\left(\hat{\mathbf{y}}^r\right) &= \mathrm{tr}\left(\mathbf{U}\mathbf{D}\mathbf{V}^\top \left(\mathbf{V}\mathbf{D}^2\mathbf{V}^\top + \lambda n \mathbf{I}_p\right)^{-1} \mathbf{V}\mathbf{D}\mathbf{U}^\top\right) \\ &= \mathrm{tr}\left(\mathbf{D}^2\left(\mathbf{D}^2 + n\lambda \mathbf{I}_p\right)^{-1}\right) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + n\lambda} \end{aligned}$$

where d_j^2 is the j^{th} diagonal element of $\mathbf{D}^2$ (Hastie et al., 2009). The degrees of freedom of the ridge fit is a decreasing function of λ which implies that the degrees of freedom decreases as the shrinkage on the coefficients increases.

3.1.1.4. Choosing the Tuning Parameter of the Ridge Estimator

Optimal tuning parameter selection is crucial for penalized regression methods. For this purpose, Hoerl and Kennard (1970) propose a graphical method called ridge trace. The plot of the ridge trace shows the coefficient paths as a function of the tuning parameter. According to this method, the optimal value of the tuning parameter is obtained when the coefficients are stabilized. Hoerl et al. (1975) propose an analytical method to select the tuning parameter:

$$\hat{\lambda} = \frac{p\hat{\sigma}^2}{\left(\hat{\boldsymbol{\beta}}^{ls}\right)^\top \hat{\boldsymbol{\beta}}^{ls}} \tag{3.11.}$$

where $\widehat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$ is obtained from the OLS estimation. Many methods are proposed as in Equation (3.11.) to choose the tuning parameter properly, see, for example, Lawless and Wang (1976), Kibria (2003), Kibria and Banik (2016).

The information criteria can be used to choose the tuning parameter (Hastie et al., 2009). Since the degrees of freedom of the ridge fit is known (df_r), the AIC and BIC for the ridge fit can be computed as

$$\mathrm{AIC}_r = n\log\mathrm{RSS} + 2\mathrm{df}_r$$

and

$$\mathrm{BIC}_r = n\log\mathrm{RSS} + \log(n)\cdot\mathrm{df}_r.$$

Then, the tuning parameter is selected as the one corresponding to the smallest AIC or BIC.

The resampling methods explained in Section 2.3 can also be used to choose the tuning parameter.

3.1.2. A Two-Parameter Estimator

A two-parameter estimator proposed by Özkale and Kaçıranlar (2007) is mentioned in this subsection. As λ goes to infinity from zero, $\widehat{\boldsymbol{\beta}}^r(\lambda)$ goes from $\mathbf{0}$ to $\widehat{\boldsymbol{\beta}}^{ls}$ where $\mathbf{0}$ is biased stable estimator of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}^{ls}$ is unbiased unstable estimator of $\boldsymbol{\beta}$. Then as λ goes to infinity from zero, the expected distance between $\widehat{\boldsymbol{\beta}}^r(\lambda)$ and $\boldsymbol{\beta}$ decreases. However, there is not an optimum λ value that produces such a point

estimate. These ideas let Özkale and Kaçıranlar (2007) to define a new estimator which minimizes the objective function given by Equation (3.4.) with the penalization function $P(\boldsymbol{\beta}) = \frac{1}{2}\left\|\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2$, that is

$$Q(\boldsymbol{\beta};\lambda,d) = \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\left\|\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2. \tag{3.12.}$$

$d\widehat{\boldsymbol{\beta}}^{ls}$ shows the prior information on $\boldsymbol{\beta}$ which is chosen as the deterministically shrunken estimator given by Equation (3.9.). Equation (3.12.) has a closed-form solution as

$$\widehat{\boldsymbol{\beta}}(\lambda,d) = \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{y} + \lambda d\widehat{\boldsymbol{\beta}}^{ls}\right). \tag{3.13.}$$

From now on, $\widehat{\boldsymbol{\beta}}(\lambda,d)$ is called as the OK[3] estimator. Özkale and Kaçıranlar (2007) state that according to the objective function given by Equation (3.12.), the OK estimator given by Equation (3.13.) has the minimum RSS in an equivalence class of estimators which are equal distance from $d\widehat{\boldsymbol{\beta}}^{ls}$. Then $\widehat{\boldsymbol{\beta}}(\lambda,d)$ goes from $d\widehat{\boldsymbol{\beta}}^{ls}$ to $\widehat{\boldsymbol{\beta}}^{ls}$ as λ goes infinity from zero. Here, $d\widehat{\boldsymbol{\beta}}^{ls}$ is chosen as prior information because it is proposed in the case of multicollinearity which shrinks each component of the OLS by d; $\left\|d\widehat{\boldsymbol{\beta}}^{ls}\right\|_2 \leq \left\|\widehat{\boldsymbol{\beta}}^{ls}\right\|_2$ for $0 \leq d < 1$, and calculations of $d\widehat{\boldsymbol{\beta}}^{ls}$ are simple because of being linear function of d. Therefore, $d\widehat{\boldsymbol{\beta}}^{ls}$ confines the estimator to a suitable parameter space when compared to $\widehat{\boldsymbol{\beta}}^{r}(\lambda)$.

[3]The OK estimator denotes the first letters of the names of the authors.

3.1.3. The LASSO Estimator

The LASSO (least absolute shrinkage and selection operator) estimator proposed by Tibshirani (1996) is a modern regression method performing both shrinkage on the coefficients and variable selection simultaneously. It is obtained by minimizing the objective function

$$Q(\boldsymbol{\beta};\lambda) = \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 . \tag{3.14.}$$

The LASSO estimator utilizes the penalization function $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ to shrink the coefficients where $\|\boldsymbol{\beta}\|_1$ denotes the ℓ_1-norm of $\boldsymbol{\beta}$. The ridge estimator shrinks the coefficients towards zero but does not set any coefficient to zero because of the properties of ℓ_2-norm. On the contrary, the LASSO estimator both shrinks the coefficients and sets some coefficient estimates to zero due to properties of the ℓ_1-norm. Therefore, the LASSO estimator can reduce the number of the selected variables in the model. This is a very important feature especially for the big data sets and makes the LASSO estimator an attractive method in statistics, machine learning, finance and engineering.

For small values of λ, the LASSO estimates are close to the OLS and for large values of λ, they are close to zero. A proper choice of the tuning parameter value provides a compromise between these two extremes. The analyst needs to choose the optimal value of the tuning parameter from the parameter space $\mathbb{R}^+$ to obtain the best coefficient estimates where $\mathbb{R}^+$ is the set of positive real numbers.

The minimization problem of the function given by Equation (3.14.) is convex, however, it is not strictly convex. Therefore the minimizer of the problem may not be unique. However, the minimizer of the problem is unique for $\lambda > 0$ when the columns

of $\mathbf{X}$ are in general position[4]. Tibshirani (2013) gives a summary of the forms in which the problem has unique or non-unique solutions.

Tibshirani (2013) shows that the KKT conditions to the minimization problem of the function given by Equation (3.14.) are

$$-\frac{1}{n}\mathbf{X}^{\top}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}^{l}(\lambda)\right)+\lambda\mathbf{s}=0, \tag{3.15.}$$

where $\widehat{\boldsymbol{\beta}}^{l}(\lambda)$ is the LASSO estimator and

$$s_j \in \begin{cases} \{-1\}, & x<0 \\ [-1,1], & x=0 \\ \{1\}, & x>0 \end{cases} \tag{3.16.}$$

for $j=1,2,\ldots,p$, that is, $\mathbf{s}\in\mathbb{R}^p$ is the subdifferential of the function $f(\mathbf{x})=\|\mathbf{x}\|_1$ given by Definition 2.3 evaluated at $\widehat{\boldsymbol{\beta}}^{l}(\lambda)$. Hence the necessary and sufficient condition for $\widehat{\boldsymbol{\beta}}^{l}(\lambda)$ to be a solution to Equation (3.14.) is that $\widehat{\boldsymbol{\beta}}^{l}(\lambda)$ must satisfy Equation (3.15.) and Condition (3.16.) for some $\mathbf{s}$.

No expression in explicit form for the minimization problem of the LASSO objective function given by Equation (3.14.) exists because of the nature of the ℓ_1-norm. The objective function in Equation (3.14.) is non-differentiable, except in some situations such as the orthonormal design matrix and the design matrix with single and two explanatory variables. Therefore, one needs to use iterative optimization

[4]The columns of $\mathbf{X}$ are in general position if any affine subspace $\mathrm{L}\subset\mathbb{R}^n$ of dimension $k<n$ contains at most $k+1$ elements of the set $\{\pm\mathbf{x}_1,\ldots,\pm\mathbf{x}_p\}$ other than the points differing only in sign (for details, see Hastie et al., 2015).

methods to solve the minimization problem of the LASSO objective function such as the methods given in Section 2.4.

3.1.3.1. Constrained and Bayesian Form of the LASSO Estimator

The minimization problem of the LASSO objective function given by Equation (3.14.) can be written in the equivalent constrained form as

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq t$$

where $t > 0$ is the constraint parameter. Figure 3.2(a) shows the structure of the LASSO estimator for two explanatory variables case. The LASSO estimator corresponds to the point at which the contours of residual sum of squares touch the rotated square standing for the constraint $\|\boldsymbol{\beta}\|_1 \leq t$ for the first place. Unlike the circular constraint region of the ridge estimator, the constraint region corresponding to the LASSO estimator has corners. If the matching point coincides with the corners of the square, then the relevant coefficient is estimated as zero and the related variable is discarded from the model effectively. This explains visually, why the LASSO estimator performs variable selection while the ridge estimator does not.

If the parameter vector $\boldsymbol{\beta}$ has identical and independent Laplace distributions (see Figure 3.2(b)) as priors, the MAP estimator of the posterior distribution becomes the LASSO estimator. To see this, let β_j, $j = 1, 2, \ldots, p$ follow identically and independent Laplace distributions

$$f\left(\beta_j\right) = \frac{1}{2\sigma_\ell} \exp\left(-\frac{|\beta_j|}{\sigma_\ell}\right).$$

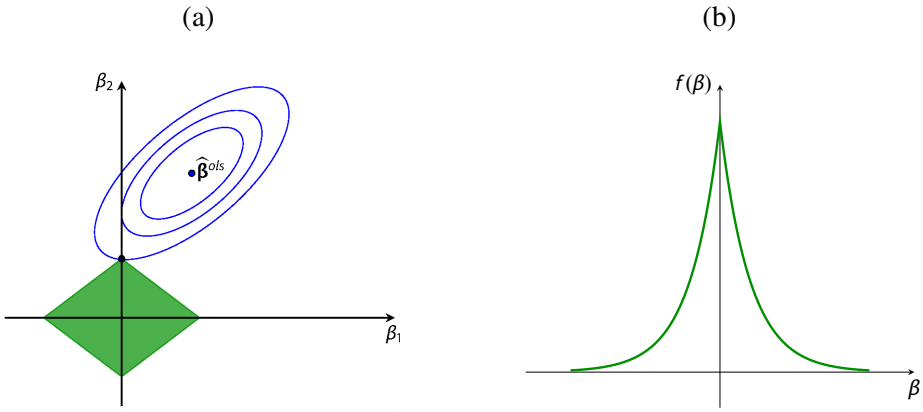

Figure 3.2. (a) Constraint region and contours for the LASSO estimator: the rotated square region corresponds to the constraint $|\beta_1| + |\beta_2| \leq t$, the contours correspond to the residual sum of squares (b) Density plot of the Laplace distribution with location parameter zero and scale parameter σ_ℓ (adapted from Tibshirani, 1996)

The posterior distribution of $\boldsymbol{\beta}$ given $\sigma^2, \mathbf{y}$ and $\mathbf{X}$ is

$$f\left(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X}\right) = f_{\mathbf{y}}\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2\right) f_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right)$$
$$\propto \exp\left[-\frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 - \frac{\sigma^2}{n\sigma_\ell}\left\|\boldsymbol{\beta}\right\|_1\right]. \tag{3.17.}$$

Therefore the MAP estimator obtained from Equation (3.17.) is the LASSO estimator with $\lambda = \sigma^2/(n\sigma_\ell)$.

3.1.3.2. The LASSO Estimator in Orthonormal Design

Consider the orthonormal design case, that is the matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthonormal and $\mathbf{X}^\top\mathbf{X} = n\mathbf{I}_p$. In this case, the minimization problem of the function given by

Equation (3.14.) is separable and the problem has a closed-form solution as

$$\widehat{\beta}_j^{lo}(\lambda) = S\left(\widehat{\beta}_j^{lso}, \lambda\right),\ j = 1, 2, \ldots, p$$

where $\widehat{\beta}^{lso}$ is defined by Equation (3.8.) and $S(\cdot)$ is the soft-thresholding operator (Donoho and Johnstone, 1994) defined as

$$S(z, \lambda) = \text{sign}(z)(|z| - \lambda)_+ \tag{3.18.}$$

where z and λ are scalars. In Equation (3.18.), $\text{sign}(z)$ stands for the sign of z:

$$\text{sign}(z) = \begin{cases} 1, & z > 0 \\ 0, & z = 0 \\ -1, & z < 0 \end{cases} \tag{3.19.}$$

and $(z)_+$ operator stands for the positive part of a:

$$(z)_+ = \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases}. \tag{3.20.}$$

3.1.3.3. Degrees of Freedom of the LASSO Estimator

Since the LASSO estimator is not a linear estimator, the degrees of freedom can not be computed by the method in Equation (2.8.). Therefore, many authors study the degrees of freedom of the LASSO fit.

Let the response variable $\mathbf{y}$ follows a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\mathbf{I}_n$. Efron et al. (2004) prove that when the active set size changes along the path of λ from ∞ to zero, the degrees of freedom of the LASSO fit $\widehat{\mathbf{y}}^l$ is equal to the active set size under the assumptions that $\mathbf{X}$ is a full column rank matrix and explanatory variables entering the model do not leave the model as λ decreases. Zou et al. (2007) show that the degrees of freedom of the LASSO fit is the expected active set size for a fixed value of λ by using the properties of the solution path of the LASSO estimator. Tibshirani and Taylor (2012) extend this result to $\mathbf{X}$ matrix with no assumption. More explicitly, for any design matrix $\mathbf{X}$ and $\lambda \geq 0$, the degrees of freedom of the LASSO fit is

$$\mathrm{df}\left(\widehat{\mathbf{y}}^l\right) = \mathrm{E}\left(\mathrm{rank}\left(\mathbf{X}_{\mathcal{A}}\right)\right) \tag{3.21.}$$

where $\mathcal{A}$ is the active set and $\mathbf{X}_{\mathcal{A}}$ is the columns of $\mathbf{X}$ corresponding to the active set. As a special case of Equation (3.21.), when $\mathrm{rank}\left(\mathbf{X}\right) = p$, $\mathrm{df}\left(\widehat{\mathbf{y}}^l\right) = \mathrm{E}\left(|\mathcal{A}|\right)$ where $|\mathcal{A}|$ is the cardinality[5] of $\mathcal{A}$.

3.1.3.4. Choosing the Tuning Parameter of the LASSO Estimator

Since the number of non zero coefficients of the model is an unbiased estimator of the degrees of freedom of the LASSO fit, this quantity can be used with the

[5]The cardinality of a finite set stands for the number of elements of the set.

information criteria to choose the tuning parameter. Therefore the AIC and BIC for the LASSO fit are computed as

$$\mathrm{AIC}_l = n \log \mathrm{RSS} + 2\,|\mathcal{A}|$$

and

$$\mathrm{BIC}_l = n \log \mathrm{RSS} + \log(n) \cdot |\mathcal{A}|\,.$$

The resampling methods in Section 2.3 can also be used to determine the tuning parameter of the LASSO estimator.

3.1.4. The Elastic Net Estimator

The LASSO estimator is useful for high-dimensional data owing to its model selection and coefficient shrinkage properties. However, the LASSO estimator has some deficiencies in solving some problems depending on the structure of the data. Some examples are listed below.

- The ridge estimator yields better results than the LASSO estimator if there is a high level of multicollinearity in a classical data set (Tibshirani, 1996).
- When the number of the explanatory variables (p) is more than the number of the observations (n), the LASSO estimator retains at most n variables in the model (Tibshirani, 1996). This is a limiting feature for variable selection.
- Highly correlated variables form a group of variables. The LASSO estimator

retains only one variable form the group in the presence of a group of variables in the data set and it does not model other variables (Zou and Hastie, 2005). For example, one expects that a high correlation between genes with similar functions exists in gene expression data sets. Therefore, highly correlated genes in these data sets should be modeled as a group.

- Another point is that the penalization function of the LASSO estimator is not strictly convex, which may cause the solution to the problem given by Equation (3.14.) not to be unique. This results in getting different estimations in a different order of explanatory variables when fitting the model (Zou and Hastie, 2005).

In such cases, alternative estimators based on the generalization of the LASSO estimator have been proposed and the elastic net estimator is one of these estimators. Zou and Hastie (2005) proposed the elastic net estimator to deal with the disadvantages of the LASSO estimator. The elastic net estimator is obtained by minimizing the following function with respect to $\boldsymbol{\beta}$:

$$Q\left(\boldsymbol{\beta};\lambda_1,\lambda_2\right)=\frac{1}{2n}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\lambda_1\left\|\boldsymbol{\beta}\right\|_1+\frac{\lambda_2}{2}\left\|\boldsymbol{\beta}\right\|_2^2. \tag{3.22.}$$

The penalization function of the elastic net estimator is $\lambda_1\left\|\boldsymbol{\beta}\right\|_1+\lambda_2\left\|\boldsymbol{\beta}\right\|_2^2$, that is a linear combination of ℓ_1 and ℓ_2-norm penalization terms. It can be shown that the KKT conditions corresponding to Equation (3.22.) are

$$-\frac{1}{n}\mathbf{X}^\top\mathbf{y}+\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}+\lambda_2\mathbf{I}_p\right)\widehat{\boldsymbol{\beta}}^{en}\left(\lambda_1,\lambda_2\right)+\lambda_1\mathbf{s}=\mathbf{0} \tag{3.23.}$$

where $\mathbf{s}$ is defined by Equation (3.16.) and $\widehat{\boldsymbol{\beta}}^{en}\left(\lambda_1,\lambda_2\right)$ stands for the elastic net esti-

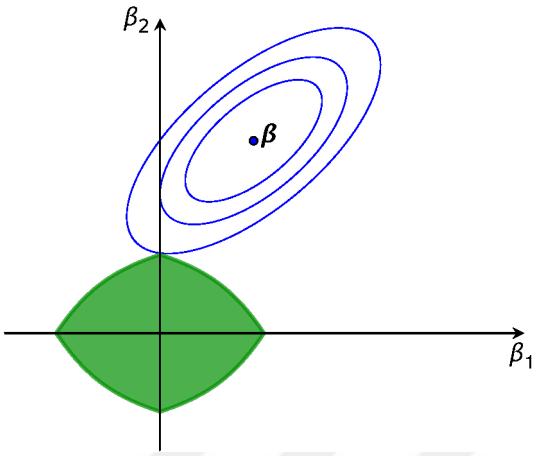

Figure 3.3. The contours of residuals and region of constraints of the elastic net estimator (adapted from Zou and Hastie, 2005)

mator with tuning parameters λ_1 and λ_2 (Tibshirani, 2013).

3.1.4.1. Constrained Form of the Elastic Net Estimator

Under the model given by Equation (2.1.), the elastic net estimator is the solution to the problem in a constrained form given by

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2 \leq t.$$

The plot of the penalty function of the elastic net estimator for $p = 2$ is shown in Figure 3.3. The elastic net estimator corresponds to the point where the contours of residual sum of squares touch the region of constraints for the first time in Figure 3.3. It has a similar pattern to the LASSO estimator at the corners (pointy) and the ridge estimator at the edges. Figure 3.3 shows visually that the properties of the penalization function of the elastic net estimator as a combination of the ridge and LASSO estimators.

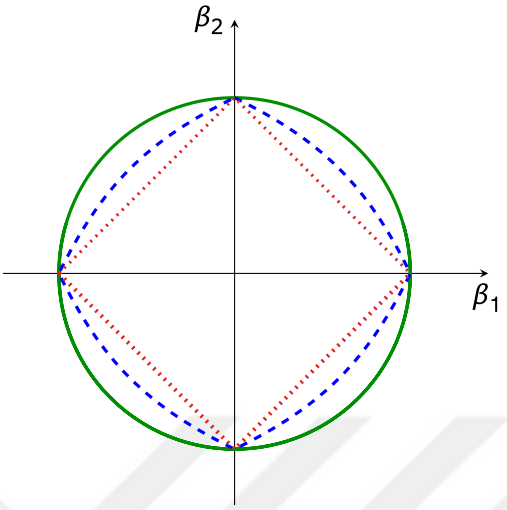

Figure 3.4. Two dimensional contour plots of the ridge (———), LASSO (·····) and elastic net (- - - - - -) penalization functions (adapted from Zou and Hastie, 2005)

Figure 3.4 shows a different point of view which is useful for comparison of the methods. As seen from Figure 3.4 that the penalization functions of the LASSO and elastic net estimators have corners, whereas the penalization function of the ridge estimator has no corner. Figure 3.4 graphically explains why the LASSO and elastic net estimators potentially provide sparse coefficient estimate vectors. The LASSO estimator has linear edges while the elastic net estimator has curved edges because of the ℓ_2-norm term in its penalization function. Because of the curvature introduced by the penalization function, the elastic net estimator tends to yield less sparse estimates than the LASSO estimator.

3.1.4.2. The Elastic Net Estimator in Orthonormal Design

When the matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthonormal, the minimization problem of Equation (3.22.) has a closed-form solution in the form of

$$\beta_j^{eno}(\lambda_1,\lambda_2) = \frac{1}{1+\lambda_2} S\left(\widehat{\beta}_j^{lso},\lambda_1\right),\ j = 1,2,\ldots,p$$

where S is the soft-thresholding operator defined by Equation (3.18.) and $\widehat{\boldsymbol{\beta}}^{lso} = \mathbf{X}^\top \mathbf{y}/n$.

According to this result, the addition of the ℓ_2-norm to the LASSO estimator has a similar effect as the addition of the ℓ_2-norm to the OLS in the orthonormal design. In other words, coefficients are shrunken. As in the ridge estimator, shrinking the coefficients towards zero increases bias and decreases variance. It is aimed to obtain a small variance and high accuracy by trading-off a proper amount of increase in bias.

3.1.4.3. Relation of the Elastic Net Estimator with the LASSO Estimator

The elastic net estimator can be formulated as a special type of LASSO estimator. Let

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{X} \\ \sqrt{n\lambda_2/2}\mathbf{I} \end{pmatrix}$$

be $(n+p)\times p$ augmented matrix on $\mathbf{X}$ and

$$\mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$$

be $(n+p)\times 1$ augmented vector on $\mathbf{y}$. Under these definitions, the function given by Equation (3.22.) becomes a LASSO objective function in the form of

$$Q(\boldsymbol{\beta};\lambda_1,\lambda_2) = \frac{1}{2n}\left\|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}\right\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 . \tag{3.24.}$$

This form of the LASSO objective function implies that all the explanatory variables can be included in the model. Equation (3.24.) also indicates that the elastic net estimator has the computational advantages of the LASSO estimator.

3.1.4.4. Reparametrization of the Tuning Parameters of the Elastic Net Estimator

The tuning parameters of the elastic net estimator can be reparametrized in a more effective form. In Equation (3.22.), let $\lambda_1 = \lambda\alpha$, $\lambda_2 = \lambda\left(\frac{1-\alpha}{2}\right)$ where $0 \le \alpha \le 1$. Then the function given by Equation (3.22.) can be written as

$$Q(\boldsymbol{\beta};\lambda,\alpha) = \left\{\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2}\|\boldsymbol{\beta}\|_2^2\right)\right\}. \tag{3.25.}$$

By using the form given by Equation (3.25.), one can select a few α values in the interval $[0,1]$ and then compute the coefficient estimates of the elastic net estimator for

a grid of λ values. The elastic net estimator corresponds to the ridge estimator when $\alpha = 0$ and the LASSO estimator when $\alpha = 1$. Therefore, the elastic net estimator is based on mutually tuning the ridge and LASSO estimators. The penalization function of the elastic net estimator is non-differentiable for $\alpha \in (0, 1]$ which indicates that the analyst needs to use an iterative method to solve the minimization problem of the function given by Equation (3.25.). The penalization function in Equation (3.25.) is strictly convex for $\alpha \in (0, 1)$ which means that the problem has unique elastic net solution.

3.1.4.5. Grouping Effect of the Elastic Net Estimator

In linear regression analysis, it is expected that the explanatory variables with high correlation have close coefficient estimates. This phenomenon leads to defining the grouping effect.

Definition 3.1. (Zou and Hastie, 2005) The grouping effect represents the situation in which the model coefficients corresponding to the highly correlated explanatory variables are approximately equal in absolute value.

The extreme case in the grouping effect arises when some of the explanatory variables have the same observed values. In this case, the regression method should assign the same coefficients to the same variables and the elastic net estimator satisfies this property. Given standardized $\mathbf{X}$ matrix and centered $\mathbf{y}$ vector, Zou and Hastie (2005) show that if $\widehat{\beta}_i^{en} \cdot \widehat{\beta}_j^{en} > 0$, then the inequality

$$\frac{1}{\|\mathbf{y}\|_1} \left| \widehat{\beta}_i^{en} - \widehat{\beta}_j^{en} \right| \leq \frac{2}{\lambda (1-\alpha)} \sqrt{2\left(1-\rho_{ij}\right)} \tag{3.26.}$$

holds where $\rho_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$ is the sample correlation. Inequality (3.26.) implies to the grouping property of the elastic net estimator. The left hand-side of Inequality (3.26.) is a unitless quantity and describes the coefficient paths of the explanatory variables i and j. For the large values of ρ_{ij}, the difference between the coefficient paths of the explanatory variables i and j almost equals to zero. Hence, if the explanatory variables have the grouping effect, the elastic net estimator extracts that effect by assigning the same coefficient to the corresponding coefficients.

3.2. Restricted-Penalized Regression Methods

The exact and stochastic restricted least square methods can be improved by combining these methods with penalized regression methods. Here, some estimators that combine exact and stochastic restricted least squares and penalized regression methods in the literature are mentioned.

Groß (2003) and Kaçıranlar et al. (2011) define the restricted ridge estimator as

$$\widehat{\boldsymbol{\beta}}^{rr}(\lambda) = \widehat{\boldsymbol{\beta}}^{r}(\lambda) + \left(\mathbf{S} + \lambda \mathbf{I}_p\right)^{-1} \mathbf{H}^\top \left(\mathbf{H}\left(\mathbf{S} + \lambda \mathbf{I}_p\right)^{-1} \mathbf{H}^\top\right)^{-1} \left(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}}^{r}(\lambda)\right)$$

and establish the theoretical properties of the estimator.

Norouzirad and Arashi (2018), Norouzirad et al. (2015) and Tuaç and Arslan (2017) study the restricted LASSO estimator under the restriction given by Equation (3.2.). By the analogy of the RLS, the restricted LASSO estimator that Norouzirad

and Arashi (2018) and Norouzirad et al. (2015) give has the form

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{rl}(\lambda) = \widehat{\boldsymbol{\beta}}^{l}(\lambda) - \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{H}^{\top}\left(\mathbf{W} + \mathbf{H}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{H}^{\top}\right)^{-1} \\ \times\left(\mathbf{H}\widehat{\boldsymbol{\beta}}^{l}(\lambda) - \mathbf{h}\right)\end{aligned} \tag{3.27.}$$

where $\widehat{\boldsymbol{\beta}}^{l}(\lambda)$ is the LASSO estimator obtained by any algorithm.

Tuaç and Arslan (2017) propose the restricted LASSO estimator as the solution of the problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \;\text{ subject to }\; \mathbf{H}\boldsymbol{\beta} = \mathbf{h} \;\text{ and }\; \sum_{j=1}^{p} |\beta_j| \leq t \tag{3.28.}$$

for some $t \geq 0$. The objective function of the problem given by Equation (3.28.) is non-differentiable at origin because of the penalty function. Therefore, Tuaç and Arslan (2017) use a local quadratic approximation (Fan and Li, 2001) to the penalty function and give the restricted LASSO estimator as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{rlt}(\lambda) = \widehat{\boldsymbol{\beta}}^{l}(\lambda) - \left(\mathbf{X}^{\top}\mathbf{X} + \boldsymbol{\Sigma}_{\lambda}\boldsymbol{\beta}^{(0)}\right)^{-1}\mathbf{H}^{\top} \\ \times\left(\mathbf{W} + \mathbf{H}\left(\mathbf{X}^{\top}\mathbf{X} + \boldsymbol{\Sigma}_{\lambda}\boldsymbol{\beta}^{(0)}\right)^{-1}\mathbf{H}^{\top}\right)^{-1}\left(\mathbf{H}\widehat{\boldsymbol{\beta}}^{l}(\lambda) - \mathbf{h}\right)\end{aligned}$$

where $\boldsymbol{\Sigma}_{\lambda}\boldsymbol{\beta}^{(0)} = \lambda \operatorname{diag}\left(\frac{1}{\left|\beta_1^{(0)}\right|}, \frac{1}{\left|\beta_2^{(0)}\right|}, \ldots, \frac{1}{\left|\beta_p^{(0)}\right|}\right)$.

Güler and Güler (2019) state that the estimator given by Equation (3.27.) does not yield sparse coefficient estimates. Therefore, the authors propose the sparsely

restricted penalized estimators named as sparsely restricted LASSO (SRL) and sparsely restricted bridge (SRB) estimators to overcome this deficiency. They propose a two-step algorithm to obtain the SRL and SRB estimates. More explicitly, in the first step of the algorithm, the model is fitted by the LASSO estimator and the coefficients estimated as zero are determined. These zero-estimated coefficients are used for augmenting the matrix $\mathbf{H}$ and the vector $\mathbf{h}$ yielding $\mathbf{H}_*$ matrix and $\mathbf{h}_*$ vector. In the second step, the model coefficients are estimated by the estimator given by

$$\widehat{\boldsymbol{\beta}}^{rl(s)}(\lambda) = \widehat{\boldsymbol{\beta}}^{l}(\lambda) - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{H}_*^\top\left(\mathbf{H}_*\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{H}_*^\top\right)^{-1} \times\left(\mathbf{H}_*\widehat{\boldsymbol{\beta}}^{l}(\lambda) - \mathbf{h}_*\right)$$

for the SRL. To obtain the SRB estimates, the bridge estimates are used in place of the LASSO estimates in the steps of the algorithm.

Özkale (2009) obtains the stochastic restricted ridge estimator by combining the idea of the mixed and ridge estimators as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{m}(\lambda) &= \left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X} + \mathbf{H}^\top\mathbf{W}^{-1}\mathbf{H} + \lambda\mathbf{I}_p\right)^{-1}\left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y} + \mathbf{H}^\top\mathbf{W}^{-1}\mathbf{h}\right)\\ &= \widehat{\boldsymbol{\beta}}^{g}(\lambda) + \mathbf{G}_\lambda^{-1}\mathbf{H}^\top\left(\mathbf{W} + \mathbf{H}\mathbf{G}_\lambda^{-1}\mathbf{H}^\top\right)^{-1}\left(\mathbf{h} - \mathbf{H}\widehat{\boldsymbol{\beta}}^{g}(\lambda)\right)\end{aligned}$$

where $\mathbf{G}_\lambda = \mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X} + \lambda\mathbf{I}_p$ and $\widehat{\boldsymbol{\beta}}^{g}(\lambda) = \mathbf{G}_\lambda^{-1}\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{y}$ is the ridge estimator for the model given by Equation (2.1.) with $\boldsymbol{\varepsilon} \sim \left(\mathbf{0}, \sigma^2\mathbf{V}\right)$.

4. NOVEL PENALIZED REGRESSION METHODS

In this chapter, we introduce novel estimators in the penalized regression methods and examine their characteristics in the scope of this thesis. The estimators are called as the GO[6], shift-ridge, shift-LASSO and stochastic restricted LASSO estimators.

4.1. The GO Estimator

The GO estimator is based on triple-shrinkage of the coefficients. The estimator and its properties are also studied by Genç and Özkale (2020a). It enjoys model selection property of the LASSO estimator like the elastic net estimator; however, it can produce estimates which are close to that of the shrunken estimator rather than the zero estimates as a result of the nature of the third term in the penalization function. It has also a type of grouping property as we will explain the details in Subsection 4.1.3.

In penalized regression methods, it is aimed to decrease the variance by introducing some bias in the coefficient estimates, which is called bias-variance trade-off. In this way, good prediction performance is aimed to be achieved. The ridge, LASSO and elastic net estimators yield biased coefficient estimates. Although the ridge and LASSO estimators are imposed by a single shrinkage, the elastic net estimator is incurred a double shrinkage. This double shrinkage is applied by first finding the ridge estimates for each fixed λ_2 and then solving the LASSO along the λ_1 values. The double shrinkage does not provide significant reduction in the variance of the elastic net estimates. Besides, the bias in the elastic net estimates is more pronounced due to the double shrinkage when compared with the LASSO and ridge estimators (Zou and

[6]The word GO both shows the abbreviations of the authors' names who proposed the estimator and also has the meaning of continuing to be in a particular state which carries the idea of ridge, LASSO and elastic net in a further way.

Hastie, 2005).

Our aim is now to find an estimator which combines the advantages of the elastic net and shrunken estimator given by Equation (3.9.). This new estimator will be proposed in such a way that its length is closer to the true coefficient vector than the OLS estimator and carries the model selection and grouping properties of the elastic net estimator.

Let $0 \leq d < 1$, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ be fixed parameters. The proposed estimator is based on minimizing the objective function

$$Q(\boldsymbol{\beta}; d, \lambda_1, \lambda_2) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \left\| \boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}^{ls} \right\|_2^2. \tag{4.1.}$$

For known values of d, the function

$$\lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \left\| \boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}^{ls} \right\|_2^2 \tag{4.2.}$$

is the penalization function of the new estimator. For usual $n > p$ cases, if there is multicollinearity in the data set, it has been empirically observed that the prediction performance of the LASSO estimator is dominated by the ridge estimator (Tibshirani, 1996). Therefore, it is possible to reinforce further the prediction performance of the LASSO estimator.

Defining $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \lambda_1 / (\lambda_1 + \lambda_2)$, then the function given by Equation (4.1.) becomes

$$Q(\boldsymbol{\beta}; d, \lambda, \alpha) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left\{ \alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \left\| \boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}^{ls} \right\|_2^2 \right\}, \tag{4.3.}$$

where $\lambda \geq 0$, $0 \leq \alpha \leq 1$ and $0 \leq d < 1$. We aim to minimize the function in the form of Equation (4.3.) for convenience as the case of the elastic net estimator.

We also note that the minimization problem of the function given by Equation (4.1.) is equivalent to the problem in constrained form

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \left\| \boldsymbol{\beta} - d\,\hat{\boldsymbol{\beta}}^{ls} \right\|_2^2 \leq t \quad (4.4.)$$

for some t with $0 \leq \alpha \leq 1$.

This new objective function leads to the well-known methods for special cases of the tuning parameters:

- the LASSO estimator for $\alpha = 1$,
- the OK estimator for $\alpha = 0$,
- the elastic net estimator for $d = 0$,
- the ridge estimator for $\alpha = d = 0$
- the OLS estimator for $\lambda = 0$.

As the consequence of using the penalization function given by Equation (4.2.), we obtain an estimator which is a generalization of the OLS, ridge, OK, LASSO and elastic net estimators. The GO estimator has the following properties:

(i) For fixed λ and d values, the GO and OK solutions are the same when $\alpha = 0$ and the GO and LASSO solutions are the same when $\alpha = 1$. Therefore, it traces

a curve path through the parameter space from the OK estimator to the LASSO estimator as α goes from 0 to 1. That is, the GO solutions oscillate between the OK and LASSO solutions for $0 \leq \alpha \leq 1$.

(ii) It is able to select sub-models as in the LASSO estimator case due to the ℓ_1-norm term in the objective function. Hence, it does variable selection as the LASSO and elastic net estimators.

(iii) For any $\alpha < 1$, $\lambda > 0$ and $0 \leq d < 1$, the problem in Equation (4.3.) is strictly convex. Therefore, the GO estimator is the unique solution irrespective of the correlations and duplications in $\mathbf{x}_j$.

(iv) The ℓ_2-norm term included in the objective function has a potential to maintain coefficient estimates with a length close to $\boldsymbol{\beta}$ than the OLS estimates depending on the parameter d in the case of the multicollinearity.

(v) When α and λ^* are fixed, the elastic net estimates vector has a shorter length than the LASSO estimates vector along the λ values which are longer than or equal to a λ^* value which is a specific value of the tuning parameter values used in computing the solutions. However, when α, d and λ^* are fixed, the GO solution is closer to $d\widehat{\boldsymbol{\beta}}^{ls}$ than the LASSO solutions along the λ values which are longer than or equal to a λ^* value which is a specific value of the tuning parameter values used in computing the solutions. This case supports efficiency over the elastic net estimator when the length of the estimator will be far away from the true parameter in the presence of the multicollinearity.

The objective function given by Equation (4.1.) depends on the parameters d, λ_1, λ_2 while the reparametrized form of the objective function given by Equation (4.3.) depends on the parameters d, α, λ. While both forms can be used, we fit the models

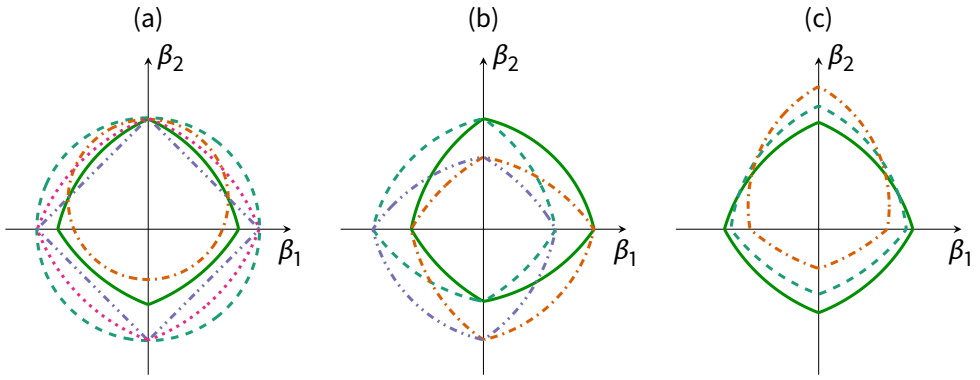

Figure 4.1. (a) Two dimensional contour plots of the ridge penalty (- - - - - -), OK penalty (- · - · - ·), LASSO penalty (- · · - · · -), elastic net penalty (·····) and GO penalty (———) with $\alpha = 0.5$, $d = 0.9$, $\beta_1^{ls} = 0$, $\beta_2^{ls} = 0.5$ (b) Two dimensional contour plots of the GO penalty for $\beta_1^{ls} = 1$, $\beta_2^{ls} = 1$ (———), $\beta_1^{ls} = 1$, $\beta_2^{ls} = -1$ (- - - - - -), $\beta_1^{ls} = -1$, $\beta_2^{ls} = 1$ (- · - · - ·), $\beta_1^{ls} = -1$, $\beta_2^{ls} = -1$ (- · · - · · -) with $\alpha = d = 0.5$ (c) Two dimensional contour plots of the GO penalty for $d = 0.2$ (———), $d = 0.5$ (- - - - - -), $d = 0.9$ (- · - · - ·) with $\alpha = 0.5$, $\beta_1^{ls} = 0$, $\beta_2^{ls} = 1.5$

by using the reparametrized form because α is bounded in the interval $[0,1]$ and d is bounded in the interval $[0,1)$ which is in the similar form of the elastic net estimator described in Hastie et al. (2015). In our case, one more step is needed for the parameter d. Therefore we select a grid of d values in the interval $[0,1)$ and, for each d value, select a grid of α values. Then, for each pair of (d,α), we select a grid of λ values and find the best value of λ by the cross-validation method. Finally, we fit the model by using the optimal values of the tuning parameters d, α and λ. This tuning parameter selection method can be seen as an extension of the method described in Hastie et al. (2015) to the new estimator by the new tuning parameter d.

A graphical comparison of the mentioned methods for different parameter values is given in Figure 4.1. As can be seen from Figure 4.1, the curve of the GO estimator has corner points as the same with the case of the LASSO and elastic

net estimators. These points are the singularity points (i.e. a point without first derivative) of the penalization function of the GO estimator. These points indicate that the method has variable selection feature. Also, the strictly convex structure of the penalization function of the GO estimator is seen in Figure 4.1, explicitly. Furthermore, the penalization function of the GO estimator can be symmetric (Figure 4.1(a) and (c)) or non-symmetric (Figure 4.1(b)) depending on the coefficient estimates of the shrunken estimator. Besides, the parameter d brings extra flexibility as displayed in Figure 4.1(c). Figure 4.1(c) also depicts the parameter space explained in Subsection 3.1.2.

4.1.1. ℓ_1 Adjustment for the GO Estimator

The relationship between the GO and LASSO estimators can be shown by making use of augmented matrices. We assume that λ and α are known. Given the design matrix $\mathbf{X}$ and response vector $\mathbf{y}$, we define $(n+p)\times p$ matrix $\mathbf{X}^*$ and $(n+p)\times 1$ vector $\mathbf{y}^*$ as follows:

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{X} \\ \sqrt{n\lambda(1-\alpha)}\mathbf{I} \end{pmatrix} \text{ and } \mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ \sqrt{n\lambda(1-\alpha)}d\widehat{\boldsymbol{\beta}}^{ls} \end{pmatrix}. \tag{4.5.}$$

Then, it is straightforward to show that

$$\begin{aligned}\frac{1}{2n}\left\|\mathbf{y}^*-\mathbf{X}^*\boldsymbol{\beta}\right\|_2^2 &= \frac{1}{2n}\left(\begin{pmatrix}\mathbf{y}\\ \sqrt{n\lambda(1-\alpha)}d\widehat{\boldsymbol{\beta}}^{ls}\end{pmatrix}-\begin{pmatrix}\mathbf{X}\\ \sqrt{n\lambda(1-\alpha)}\mathbf{I}\end{pmatrix}\boldsymbol{\beta}\right)^{\top}\times\\ &\qquad\left(\begin{pmatrix}\mathbf{y}\\ \sqrt{n\lambda(1-\alpha)}d\widehat{\boldsymbol{\beta}}^{ls}\end{pmatrix}-\begin{pmatrix}\mathbf{X}\\ \sqrt{n\lambda(1-\alpha)}\mathbf{I}\end{pmatrix}\boldsymbol{\beta}\right)\\ &= \frac{1}{2n}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+n\lambda(1-\alpha)\left(\boldsymbol{\beta}-d\widehat{\boldsymbol{\beta}}^{ls}\right)^{\top}\left(\boldsymbol{\beta}-d\widehat{\boldsymbol{\beta}}^{ls}\right)\right]\\ &= \frac{1}{2n}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\frac{\lambda(1-\alpha)}{2}\left\|\boldsymbol{\beta}-d\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2.\end{aligned}$$

Therefore, under Equalities (4.5.), we obtain

$$\frac{1}{2n}\left\|\mathbf{y}^*-\mathbf{X}^*\boldsymbol{\beta}\right\|_2^2+\lambda\alpha\left\|\boldsymbol{\beta}\right\|_1=\frac{1}{2n}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\lambda\alpha\left\|\boldsymbol{\beta}\right\|_1+\frac{\lambda(1-\alpha)}{2}\left\|\boldsymbol{\beta}-d\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2$$

which indicates that the GO estimator is equivalent to the LASSO estimator based on the augmented matrix $\mathbf{X}^*$ and vector $\mathbf{y}^*$. Hence, the GO estimator enjoys the computational advantages of the LASSO estimator.

4.1.2. The Coordinate Descent Algorithm for the GO Estimator

The penalization function in Equation (4.2.) is separable in its components. Accordingly, the coordinate descent algorithm can be used to compute the GO estimates. When other parameters are fixed, we write the objective function in terms of $\beta_j,\ j=1,2,\ldots,,p$ as

$$\begin{aligned}f\left(\beta_j;d,\lambda,\alpha\right) &= \frac{1}{2n}\sum_{i=1}^{n}\left(r_{ij}-x_{ij}\beta_j\right)^2+\lambda\alpha\sum_{j=1}^{p}\left|\beta_j\right|\\ &\quad+\frac{\lambda(1-\alpha)}{2}\sum_{j=1}^{p}\left(\beta_j-d\widehat{\beta}_j^{ls}\right)^2\end{aligned}$$

where $r_{ij} = y_i - \sum_{k \neq j} x_{ik}\beta_k$ is called the partial residual of the variable $\mathbf{x}_j$. The subdifferential of the function $f\left(\beta_j; d, \lambda, \alpha\right)$ is

$$\partial f\left(\beta_j; d, \lambda, \alpha\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 + \lambda\left(1-\alpha\right)\right)\beta_j - \frac{1}{n}\sum_{i=1}^{n} x_{ij} r_{ij} - \lambda\left(1-\alpha\right) d\, \widehat{\beta}_j^{ls} + \lambda\alpha s_j$$

for $j = 1, 2, \ldots, p$ where s_j is defined by Equation (3.16.). Hence, by equating $\partial f\left(\beta_j; d, \lambda, \alpha\right)$ to zero, we get the estimates of the new estimator in an iterative manner depending on the soft-thresholding operator defined by Equation (3.18.) as

$$\widehat{\beta}_j^{go}\left(d, \lambda, \alpha\right) \leftarrow \frac{S\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\widehat{r}_{ij} + \lambda\left(1-\alpha\right) d\, \widehat{\beta}_j^{ls}, \lambda\alpha\right)}{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 + \lambda\left(1-\alpha\right)}, \; j = 1, 2, \ldots, p.$$

where $\beta_j\left(d, \lambda, \alpha\right)$ term in the iterations represents the GO estimator and $S\left(\cdot\right)$ is soft-thresholding operator given by Equation (3.18.). When the explanatory variables have mean zero and standard deviation 1, as a special case, the iterations for the solution are

$$\widehat{\beta}_j^{go}\left(d, \lambda, \alpha\right) \leftarrow \frac{S\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\widehat{r}_{ij} + \lambda\left(1-\alpha\right) d\, \widehat{\beta}_j^{ls}, \lambda\alpha\right)}{1 + \lambda\left(1-\alpha\right)}.$$

We, now, examine the GO estimator in the standardized form when the matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthonormal. In this case, we have a closed-form solution for the coefficients

as

$$\widehat{\beta}_j^{go}(d,\lambda,\alpha) = \frac{S\left((1+\lambda(1-\alpha)d)\,\widehat{\beta}_j^{ls},\lambda\alpha\right)}{1+\lambda(1-\alpha)}$$
$$= \operatorname{sign}\left((1+\lambda(1-\alpha)d)\,\widehat{\beta}_j^{ls}\right)\frac{\left((1+\lambda(1-\alpha)d)\left|\widehat{\beta}_j^{ls}\right|-\lambda\alpha\right)_+}{1+\lambda(1-\alpha)}$$

where $\operatorname{sign}(\cdot)$ is the signum function defined by Equation (3.19.) and $(z)_+$ is the positive part function defined by Equation (3.20.).

Figure 4.2 shows the operational characteristics of the ridge, LASSO, elastic net and GO estimators when $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthonormal. The lines in Figure 4.2 represent the mentioned estimators as different functions of the OLS solution. The elastic net and GO estimators are two-stage procedures though ridge and LASSO are one-stage procedures as mentioned in Zou and Hastie (2005). It is seen from Figure 4.2 that the interval in which coefficient estimates are zero for the LASSO and elastic net estimators are the same, while it is narrow for the GO estimator depending on the parameter d.

In order to gain a visual point of view, the coefficient plots of the LASSO, elastic net and GO estimators are shown on the `mpg` data set from Henderson and Velleman (1981). The data set was extracted from the 1974 Motor Trend magazine, and consists of miles per gallon (`mpg`) and ten features of automobile design and performance of 32 automobiles which are 1973–74 models. The explanatory variables are the number of cylinders (`cyl`), displacement (`disp`), gross horsepower (`hp`), rear axle ratio (`drat`), weight in 1000 lbs (`wt`), 1/4 mile time (`qsec`), V/S (`vs`), transmission (`am`), number of forward gears (`gear`), number of carburetors (`carb`) and the response is miles per gallon (`mpg`). We fit the LASSO, elastic net and GO estimators when

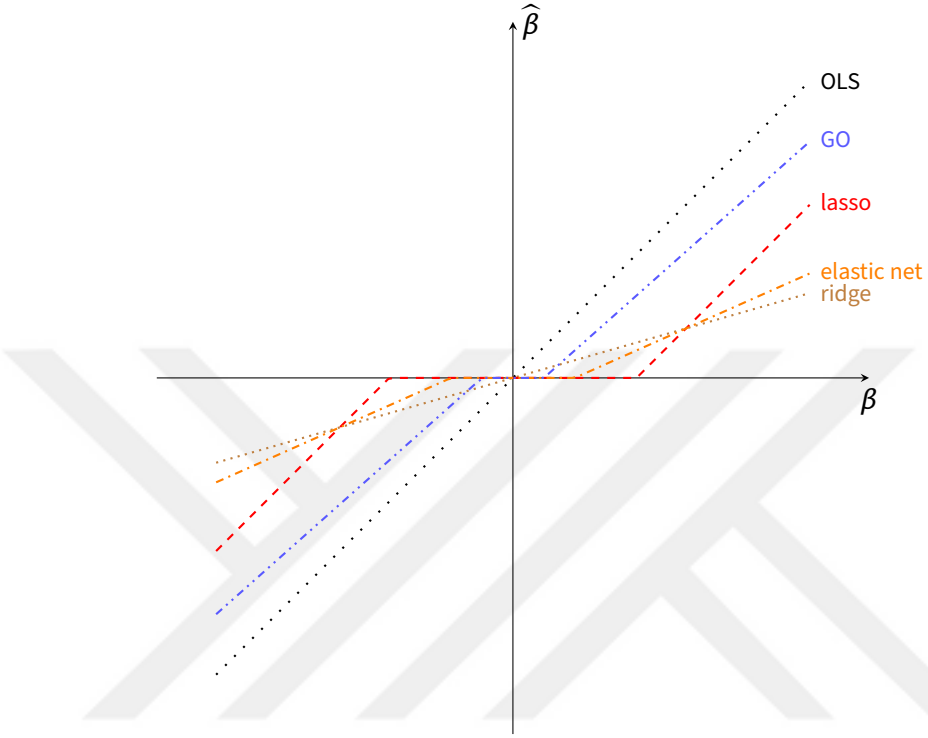

Figure 4.2. Exact solutions for the ridge (- - - - - -), LASSO (- · - · - ·), elastic net (- · · - · · -) and GO (———) estimators in the orthonormal design where the shrinkage parameters are $\lambda = 2.5$, $\alpha = 0.5$, $d = 0.8$ and (· · · · · ·) denotes the OLS estimator

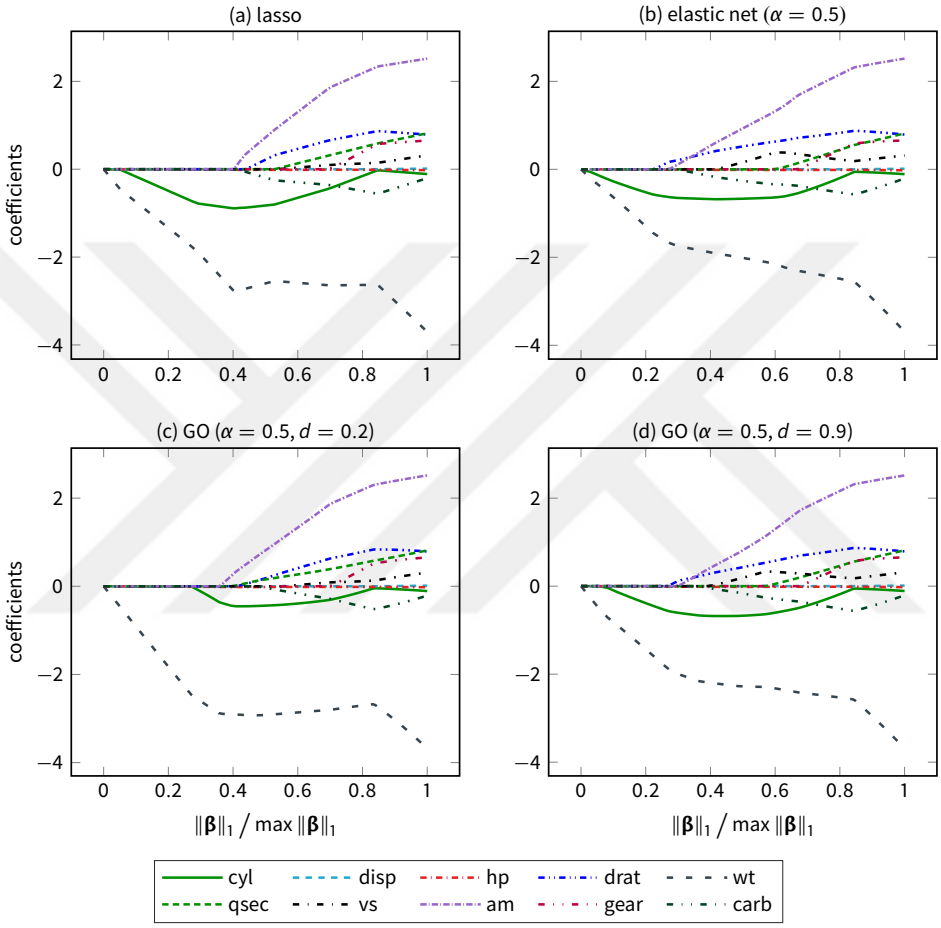

Figure 4.3. The LASSO (a), elastic net (b) and GO (c, d) coefficient estimates as a function of $\|\boldsymbol{\beta}\|_1 \,/\, \max \|\boldsymbol{\beta}\|_1$ for the `mpg` data set when $\alpha = 0.5$, $d = 0.2$ and $d = 0.9$ for 100 different values of λ

$\alpha = 0.5$ and $d = 0.2$ or $d = 0.9$ for 100 different λ values. The coefficient paths of the fitted models are shown in Figure 4.3. The coefficient paths are plotted as a function of the parameter $\|\boldsymbol{\beta}\|_1 \,/\, \max \|\boldsymbol{\beta}\|_1$ for convenience in comparison of different models. The coefficient paths obtained with the LASSO estimator show instability while this instability has improved somewhat with the elastic net estimator. When the coefficient paths obtained by the GO estimator are examined, it can be seen that the knot points introduced by the LASSO and elastic net estimators exist in the coefficient paths of the GO estimator. However, the coefficient paths follow a more stable course. In addition, the points at which the variables enter into the model change depending on the values of d. Moreover, as d increases, the coefficient paths become more pronounced for the more significant variables such as `wt` and `am`. Thus the GO estimator gives a better result in terms of reducing the instability in the coefficient paths and emphasizing the more significant coefficients.

4.1.3. A New Grouping Property for the GO Estimator

Zou and Hastie (2005) have shown that the elastic net estimates have the grouping effect. A different type of grouping effect can be defined for the GO estimator. To clarify this property for the GO estimator, we first give Theorem 4.1.

Theorem 4.1. Given the response vector $\mathbf{y}$ is centered to have mean zero and the columns of design matrix $\mathbf{X}$ is centered to have mean zero and standard deviation 1. Let λ, α and d parameters are fixed. We assume $\widehat{\beta}_i^{go}\widehat{\beta}_j^{go} > 0$ where $\widehat{\beta}_i^{go} = \widehat{\beta}_i^{go}(d,\lambda,\alpha)$ and $\widehat{\beta}_j^{go} = \widehat{\beta}_j^{go}(d,\lambda,\alpha)$ represent the i^{th} and j^{th} GO estimates, respectively. Define the distance

$$U(i,j) = \frac{1}{\sqrt{\|\mathbf{y}\|_1^2 + n\lambda(1-\alpha)d^2\left\|\widehat{\beta}^{ls}\right\|_2^2}}\left|\left(\widehat{\beta}_i^{go} - d\,\widehat{\beta}_i^{ls}\right) - \left(\widehat{\beta}_j^{go} - d\,\widehat{\beta}_j^{ls}\right)\right|, \quad (4.6.)$$

then the inequality

$$U(i,j) \leq \frac{1}{n\lambda(1-\alpha)}\sqrt{2(1-\rho)}$$

holds true where ρ is the sample correlation coefficient between the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.
Proof: We start the proof with the function given by Equation (4.1.). We first take the derivatives of the function with respect to β_i and β_j and set the results to zero:

$$\left.\frac{\partial Q}{\partial \beta_i}\right|_{\beta_i=\widehat{\beta}_i^{go}} = -\frac{1}{n}\mathbf{x}_i^{\top}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}^{go}\right)+\lambda\alpha\widehat{s}_i+\lambda(1-\alpha)\left(\widehat{\beta}_i^{go}-d\,\widehat{\beta}_i^{ls}\right)=0 \qquad (4.7.)$$

$$\left.\frac{\partial Q}{\partial \beta_j}\right|_{\beta_j=\widehat{\beta}_j^{go}} = -\frac{1}{n}\mathbf{x}_j^{\top}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}^{go}\right)+\lambda\alpha\widehat{s}_j+\lambda(1-\alpha)\left(\widehat{\beta}_j^{go}-d\,\widehat{\beta}_j^{ls}\right)=0 \qquad (4.8.)$$

where $\widehat{s}_i$ and $\widehat{s}_j$ subdifferentials have the same sign. Then we subtract Equation (4.7.) from Equation (4.8.) and we get

$$\widehat{\beta}_j^{go}-\widehat{\beta}_i^{go}-d\left(\widehat{\beta}_j^{ls}-\widehat{\beta}_i^{ls}\right)=\frac{1}{n\lambda(1-\alpha)}\left(\mathbf{x}_j^{\top}-\mathbf{x}_i^{\top}\right)\widehat{\mathbf{r}} \qquad (4.9.)$$

where $\widehat{\mathbf{r}}=\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}^{go}$. By Cauchy Schwarz inequality for Equation (4.9.), we can write

$$\left|\widehat{\beta}_j^{go}-\widehat{\beta}_i^{go}-d\left(\widehat{\beta}_j^{ls}-\widehat{\beta}_i^{ls}\right)\right| \leq \frac{1}{n\lambda(1-\alpha)}\sqrt{\left\|\mathbf{x}_i-\mathbf{x}_j\right\|_2^2\left\|\widehat{\mathbf{r}}\right\|_2^2}. \qquad (4.10.)$$

We know that the equality $\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 = 2n\,(1-\rho)$ holds true because $\mathbf{x}_i$ and $\mathbf{x}_j$ are standardized. Applying this result to Inequality (4.10.) , we obtain

$$\left|\widehat{\beta}_j^{go} - \widehat{\beta}_i^{go} - d\left(\widehat{\beta}_j^{ls} - \widehat{\beta}_i^{ls}\right)\right| \leq \frac{1}{\sqrt{n}\lambda\,(1-\alpha)}\sqrt{2\,(1-\rho)\,\|\widehat{\mathbf{r}}\|_2^2}. \tag{4.11.}$$

We also know the inequality $Q\left(\widehat{\boldsymbol{\beta}}^{go};d,\lambda,\alpha\right) \leq Q\,(\mathbf{0};d,\lambda,\alpha)$ is valid because $\widehat{\boldsymbol{\beta}}^{go}$ minimizes the function $Q\,(\boldsymbol{\beta};d,\lambda,\alpha)$. Therefore, we get

$$\frac{1}{2n}\|\widehat{\mathbf{r}}\|_2^2 + \lambda\alpha\left\|\widehat{\boldsymbol{\beta}}^{go}\right\|_1 + \frac{\lambda\,(1-\alpha)}{2}\left\|\widehat{\boldsymbol{\beta}}^{go} - d\,\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2 \leq \frac{1}{2n}\|\mathbf{y}\|_2^2 + \frac{\lambda\,(1-\alpha)\,d^2}{2}\left\|\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2.$$

Considering that the norms are non-negative, we write

$$\|\widehat{\mathbf{r}}\|_2^2 \leq \|\mathbf{y}\|_2^2 + n\lambda\,(1-\alpha)\,d^2\left\|\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2. \tag{4.12.}$$

If we take account of Equations (4.12.) and (4.11.) together, we obtain

$$\begin{aligned}\left|\widehat{\beta}_j^{go} - \widehat{\beta}_i^{go} - d\left(\widehat{\beta}_j^{ls} - \widehat{\beta}_i^{ls}\right)\right| &\leq \frac{1}{\sqrt{n}\lambda\,(1-\alpha)}\sqrt{2\,(1-\rho)\left(\|\mathbf{y}\|_2^2 + n\lambda\,(1-\alpha)\,d^2\left\|\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2\right)} \\ &\leq \frac{1}{\sqrt{n}\lambda\,(1-\alpha)}\sqrt{2\,(1-\rho)}\sqrt{\|\mathbf{y}\|_1^2 + n\lambda\,(1-\alpha)\,d^2\left\|\widehat{\boldsymbol{\beta}}^{ls}\right\|_2^2}.\end{aligned}$$

As a result, the inequality holds and the proof is completed.

The unit-less quantity $U\,(i\,,j\,)$ in Equation (4.6.) allows to reach some conclusions regarding the differences between the coefficients of the explanatory variables i and j for the GO estimator with the shrunken estimates:

- The differences $\widehat{\beta}_j^{go} - d\,\widehat{\beta}_j^{ls}$ and $\widehat{\beta}_i^{go} - d\,\widehat{\beta}_i^{ls}$ are approximately equal. Therefore, the approximation speed of the j^{th} GO estimate to the corresponding shrunken

estimate is almost equal to the approximation speed of the i^{th} GO estimate to the corresponding shrunken estimate.

- Taking into account that the differences $\widehat{\beta}_j^{go} - \widehat{\beta}_i^{go}$ and $d\left(\widehat{\beta}_j^{ls} - \widehat{\beta}_i^{ls}\right)$ are approximately equal; that is, the difference between the j^{th} and i^{th} GO estimates is almost equal to the difference between the corresponding shrunken estimates.

- If x_i and x_j are highly correlated, i.e. $\rho = 1$ (if $\rho = -1$ then consider x_i and $-x_j$) between the coefficient paths of x_i and x_j is almost $d\left(\widehat{\beta}_j^{ls} - \widehat{\beta}_i^{ls}\right)$. This property also presents that as $d \to 0$, the grouping property of the GO estimator approaches to that of the elastic net estimator.

These conclusions can be seen from Figure 4.4 where the data were simulated for the sample size of $n = 100$ by following Hastie et al. (2015):

$$
\begin{aligned}
Y &= 3Z_1 - 1.5Z_2 + 2\varepsilon, \text{ with } \varepsilon \sim \mathcal{N}(0,1), \\
X_j &= Z_1 + \xi_j/5, \text{ with } \xi_j \sim \mathcal{N}(0,1), \text{ for } j = 1,2,3, \text{ and} \\
X_j &= Z_2 + \xi_j/5, \text{ with } \xi_j \sim \mathcal{N}(0,1), \text{ for } j = 4,5,6
\end{aligned}
$$

where $Z_1 \sim \mathcal{N}(0,1)$ and $Z_2 \sim \mathcal{N}(0,1)$ are independent variables. In this setting, there are two sets of three variables, with pairwise correlation about 0.97 in each group. We give the coefficient paths of the methods in Figure 4.4. The horizontal lines in Figure 4.4(a) and (b) are the OLS estimates while these lines represent the shrunken estimator in Figure 4.4(c) and (d). All the methods coincidence with the OLS solution for $\lambda = 0$ (when $\|\boldsymbol{\beta}\|_1$ has its largest value). According to Figure 4.4, the coefficients of the LASSO estimator do not show good performance since the coefficient paths of the

LASSO estimator tend to be erratic and it does not capture the relative importance of the variables. Unlike the LASSO estimator, the elastic net estimator tends to combine the variables in each group together. Therefore it reflects the relative prominence of the individual variables in the groups. The GO estimator shows a grouping effect in a different manner. As can be seen from Figure 4.4, the GO estimator drags the coefficient estimates towards the shrunken estimates individually. For small value of d, the pattern is similar to the paths of the elastic net estimator while for the large value of d, the individual estimates approach their $d\,\widehat{\boldsymbol{\beta}}^{ls}$ counterparts more quickly but the groups can still be seen easily.

4.2. The Shift-ridge and Shift-LASSO Estimators

In this section, we study new methods to deal with both multicollinearity and high-dimensionality in linear regression models. The methods are based on the combination of the ideas of the elastic net and GO estimators. The new methods satisfy the properties of the continuous shrinkage, automatic variable selection and selecting group of correlated explanatory variables of the elastic net. In addition to these properties, the new methods shrink the regression coefficients to a vector which represents the prior information. Also, the new methods can deal with the high-dimensional data which the GO estimator can not do.

4.2.1. The New Methods

Consider the linear regression setup given by Equation (2.1.) with n observations and p explanatory variables. The GO estimator introduced in Section 4.1 is a good alternative to the penalization methods ridge, OK, LASSO and elastic net and has promising properties given in Section 4.1. However, the GO estimator can not be applied to the high-dimensional data because it contains the coefficient estimates of

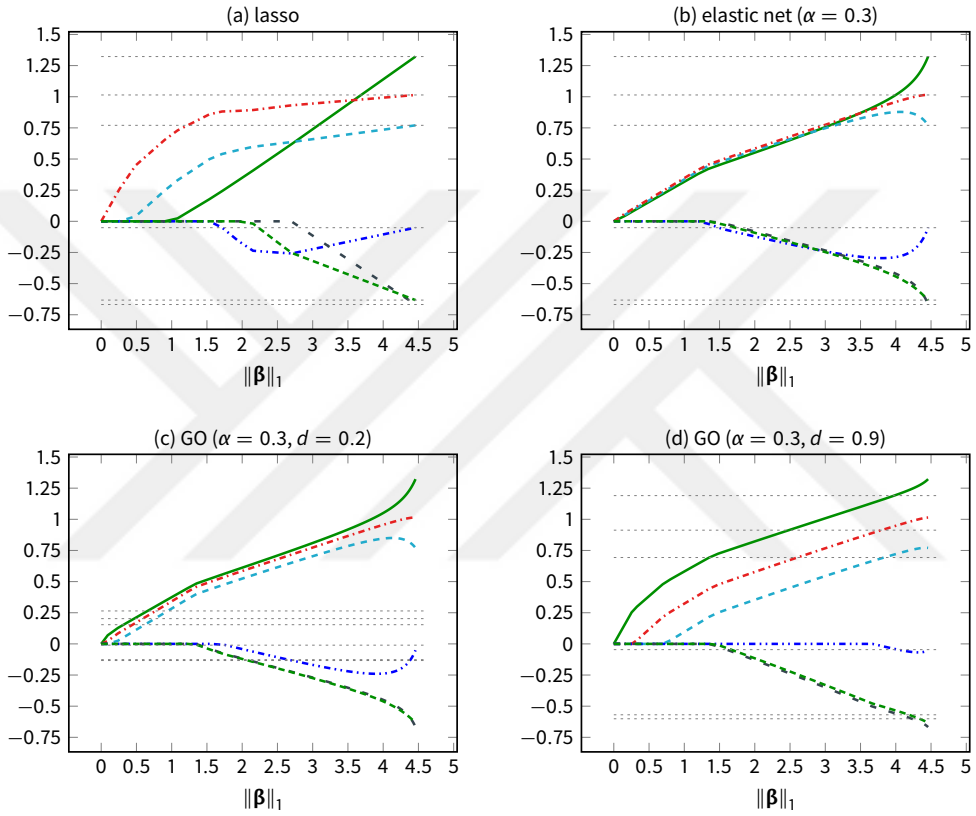

Figure 4.4. The LASSO (a), elastic net (b) and GO (c, d) coefficient estimates as a function of $\|\boldsymbol{\beta}\|_1$ for the simulated data set when $\alpha = 0.3$, $d = 0.2$ and $d = 0.9$ for 100 different values of λ

the shrunken estimator $d\widehat{\boldsymbol{\beta}}^{ls}$ which can not be computed for the high-dimensional data. A choice other than $d\widehat{\boldsymbol{\beta}}^{ls}$ can be applied in Equation (4.1.) so that the estimates from Equation (4.1.) become appropriate for the high-dimensional data. Genç and Özkale (2020b) propose to apply $\mathbf{b} = \widehat{\boldsymbol{\beta}}^{r}(\lambda)$ or $\mathbf{b} = \widehat{\boldsymbol{\beta}}^{l}(\lambda)$ instead of $d\widehat{\boldsymbol{\beta}}^{ls}$ where $\mathbf{b}$ represents a prior information available. Thus the objective function of the new methods is

$$Q(\boldsymbol{\beta};\lambda,\alpha) = \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2}\|\boldsymbol{\beta}-\mathbf{b}\|_2^2\right). \tag{4.13.}$$

Based on the choice of $\mathbf{b}$, the method is named as the shift-ridge ($\widehat{\boldsymbol{\beta}}^{sr}(\lambda,\alpha)$) or shift-LASSO ($\widehat{\boldsymbol{\beta}}^{sl}(\lambda,\alpha)$), respectively. Therefore, the new method satisfies similar properties as the GO estimator; additionally, it is appropriate for the high-dimensional data. Moreover, the new method includes the ridge, LASSO, elastic net, GO and two-parameter estimators depending on the value of the tuning parameter:

(1) the ridge estimator when $\alpha = 0$ and no prior information; that is, $\mathbf{b} = 0$,

(2) the LASSO estimator when $\alpha = 1$,

(3) the elastic net estimator when no prior information,

(4) the GO estimator when $\mathbf{b} = d\widehat{\boldsymbol{\beta}}^{ls}$ and

(5) the OK estimator when $\alpha = 0$ and $\mathbf{b} = d\widehat{\boldsymbol{\beta}}^{ls}$.

In the rest of this section, the vector $\mathbf{b}$ represents both the shift-ridge and the shift-LASSO estimators depending on the selection of the vector $\mathbf{b}$.

Because of the term involving ℓ_1-norm in the penalization function, the shift-ridge and shift-LASSO estimators perform automatic variable selection. The second term of the penalization function implies that the coefficient estimates approach to $\mathbf{b}$.

Therefore, the estimates of the new method shift towards an estimator whose length is closer to the true parameter vector than that of the OLS estimator.

The relationship between the shift-ridge and shift-LASSO estimators with the LASSO estimator can be shown by using augmented matrices. Given the design matrix $\mathbf{X}$, response vector $\mathbf{y}$, tuning parameters (λ, α), we define the augmented matrices

$$\mathbf{X}_A = \begin{pmatrix} \mathbf{X} \\ \sqrt{n\lambda\,(1-\alpha)}\mathbf{I} \end{pmatrix} \text{ and } \mathbf{y}_A = \begin{pmatrix} \mathbf{y} \\ \sqrt{n\lambda\,(1-\alpha)}\mathbf{b} \end{pmatrix}$$

with dimensions $(n+p)\times p$ and $(n+p)\times 1$, respectively. Then we can verify the arguments in Subsection 4.1.1 as

$$\frac{1}{2n}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\lambda\alpha\left\|\boldsymbol{\beta}\right\|_1+\frac{\lambda\,(1-\alpha)}{2}\left\|\boldsymbol{\beta}-\mathbf{b}\right\|_2^2=\frac{1}{2n}\left\|\mathbf{y}_A-\mathbf{X}_A\boldsymbol{\beta}\right\|_2^2+\lambda\alpha\left\|\boldsymbol{\beta}\right\|_1. \quad (4.14.)$$

According to Equation (4.14.), the shift-ridge and shift-LASSO estimators are in the form of the LASSO estimator based on the augmented matrix $\mathbf{X}_A$ and response vector $\mathbf{y}_A$. Since the augmented matrix $\mathbf{X}_A$ is $(n+p)\times p$ with rank p, the shift-ridge and shift-LASSO estimators can select all the explanatory variables in the design matrix as the elastic net estimator does. This property overcomes the deficiency of selecting at most n explanatory variables of the LASSO estimator in the high-dimensional data. Equation (4.14.) also indicates that the shift-ridge and shift-LASSO estimators can perform automatic variable selection in a similar way to the LASSO estimator.

The objective function given by Equation (4.13.) depends on the parameters λ and α. One way of choosing the tuning parameters is described as follows: First, pick a grid of α values. Then pick a grid of λ values corresponding to each of the α values and apply the cross-validation to the training data to select the optimal λ. Choose α as

the one giving the smallest cross-validation error.

4.2.2. Algorithms for Fitting the Shift-ridge and Shift-LASSO Estimators

Because of the nature of ℓ_1-norm, no closed-form solution of the shift-ridge and shift-LASSO estimators exists. Therefore, we need to implement iterative methods to solve the problem of the shift-ridge and shift-LASSO estimators. We use the ADMM and coordinate descent algorithms given in Section 2.4 to solve the problems.

The ADMM and coordinate descent algorithms are useful for this problem because it involves ℓ_1-norm penalty which is non-differentiable at zero (when $\widehat{\beta}_j = b_j$). The coordinate descent is an iterative algorithm that updates coefficients by choosing a single coordinate to update, and performing a univariate minimization over this coordinate. The algorithm is particularly useful when the problem satisfies the separability condition as mentioned in Subsection 2.4.1.1 and the problem that we introduce holds this condition. A full cycle of the coordinate descent through all the variables costs $O(pn)$. The ADMM is a flexible algorithm that can handle a convex optimization problem with non-differentiable constraints. Furthermore, it has some attractive features for large-scale applications (Hastie et al., 2015). This algorithm has cost $O(pn)$ as is the coordinate descent for our problem except a singular value decomposition of $\mathbf{X}$ in the first step which can be done off-line.

If the vector $\mathbf{b}$ is known, the cost of the ADMM or coordinate descent for our problem is $O(pn)$, the same cost with the classical LASSO problem. The cost will increase if $\mathbf{b}$ is estimated by a different method. In this case, the increase in the cost is completely determined by the procedure of estimating $\mathbf{b}$. Once $\mathbf{b}$ is estimated, we fix the estimate and use it during the analysis. There is no additional cost for cross-validation other than the estimation of $\mathbf{b}$ once.

4.2.2.1. The ADMM Algorithm for the Shift-ridge and Shift-LASSO Estimators

After some algebraic computations, we can give the scaled form of the ADMM iterations in Equation (2.15.) for the shift-ridge and shift-LASSO estimators by

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)}(\lambda,\alpha) &= \left(\frac{1}{n}\mathbf{X}^{\top}\mathbf{X}+\tau\mathbf{I}_p\right)^{-1}\left(\frac{1}{n}\mathbf{X}^{\top}\mathbf{y}+\tau\left(\boldsymbol{\theta}^{(k)}(\lambda,\alpha)-\mathbf{u}^{(k)}\right)\right)\\
\boldsymbol{\theta}^{(k+1)}(\lambda,\alpha) &= \frac{S\left(\lambda(1-\alpha)\mathbf{b}+\tau\left(\boldsymbol{\beta}^{(k+1)}(\lambda,\alpha)+\mathbf{u}^{(k)}\right),\lambda\alpha\right)}{\lambda(1-\alpha)+\tau}\\
\mathbf{u}^{(k+1)} &= \mathbf{u}^{(k)}+\boldsymbol{\beta}^{(k+1)}(\lambda,\alpha)-\boldsymbol{\theta}^{(k+1)}(\lambda,\alpha)
\end{aligned} \tag{4.15.}$$

where the $\boldsymbol{\beta}$ vector in the iterations represents the shift-ridge or shift-LASSO estimator. Therefore, we can get the solution vector by updating $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\mathbf{u}$ in Equation (4.15.) until a convergence criterion is met.

4.2.2.2. The Coordinate Descent Algorithm for the Shift-ridge and Shift-LASSO Estimators

To see the minimization problem of the function given by Equation (4.13.) is separable in components, we rewrite the problem as a function of β_j by holding the other coefficients and the tuning parameters fixed.

$$\begin{aligned}
\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, f\left(\beta_j;\lambda,\alpha\right) = \frac{1}{2n}\sum_{i=1}^{n}\left(r_{ij}-x_{ij}\beta_j\right)^2+\lambda\alpha\sum_{j=1}^{p}\left|\beta_j\right|\\
+\frac{\lambda(1-\alpha)}{2}\sum_{j=1}^{p}\left(\beta_j-b_j\right)^2
\end{aligned}$$

where $r_{ij} = y_i - \sum_{k \neq j} x_{ik}\beta_k$ is the j^{th} partial residual. Here, we assume that the explanatory variables have mean zero and standard deviation 1. The subdifferential of $f\left(\beta_j;\lambda,\alpha\right)$ is

$$\begin{aligned}\partial f\left(\beta_j;\lambda,\alpha\right) = &\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 + \lambda\left(1-\alpha\right)\right)\beta_j - \frac{1}{n}\sum_{i=1}^{n} x_{ij} r_{ij} \\ &-\lambda\left(1-\alpha\right)b_j + \lambda\alpha s_j \end{aligned} \tag{4.16.}$$

for $j = 1,2,\ldots,p$ where s_j is defined by Equation (3.16.). By setting Equation (4.16.) to zero, we get the coordinate descent updates

$$\widehat{\beta}_j^{sh}\left(\lambda,\alpha\right) \leftarrow \frac{S\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\widehat{r}_{ij} + \lambda\left(1-\alpha\right)b_j, \lambda\alpha\right)}{1+\lambda\left(1-\alpha\right)},\ j = 1,2,\ldots,p$$

where $\widehat{\beta}_j^{sh}\left(\lambda,\alpha\right)$ represents the j^{th} coefficient estimate of the shift-ridge or shift-LASSO estimator and $S\left(\cdot\right)$ is soft-thresholding operator given by Equation (3.18.).

As a special case, if the matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthonormal, the problem has the closed-form solution

$$\widehat{\beta}_j^{sh}\left(\lambda,\alpha\right) = \frac{S\left(\frac{1}{n}\mathbf{x}_j^{\top}\mathbf{y} + \lambda\left(1-\alpha\right)b_j, \lambda\alpha\right)}{1+\lambda\left(1-\alpha\right)}.$$

Figure 4.5 shows the characteristics of the ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators when $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthonormal. The shift-ridge and shift-LASSO estimators use $\widehat{\boldsymbol{\beta}}^{r}\left(\lambda\right)$ or $\widehat{\boldsymbol{\beta}}^{l}\left(\lambda\right)$ rather than $d\widehat{\boldsymbol{\beta}}^{ls}$. The solutions

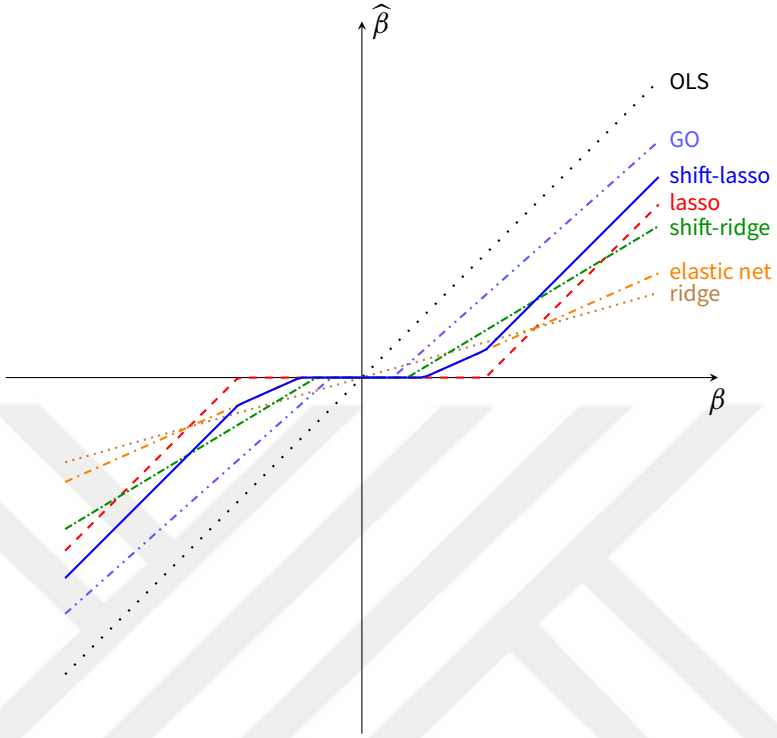

Figure 4.5. Solutions to the ridge (·····), LASSO (- - - - -), elastic net (-·-·-·-), GO (-·-·-·-), shift-ridge (-·-·-·-), shift-LASSO (———) estimators in the orthonormal design where the tuning parameters are $\lambda = 2.5$, $\alpha = 0.5$, $d = 0.8$ and (· · · · ·) denotes the OLS solution

are displayed as functions of the OLS estimates. The zero interval of the coefficient estimates is narrow for the shift-ridge and the shift-LASSO estimators when compared to the LASSO estimator. However, the shift-ridge and shift-LASSO estimators have wider zero intervals compared to the GO estimator (see also Figure 4.2). The shift-LASSO shows a different pattern from other methods based on soft-thresholding. It has two extra break points as a result of shifting towards LASSO, which causes more sparsity on the coefficients compared to the shift-ridge estimator.

4.2.3. A New Grouping Property for the Shift-ridge and Shift-LASSO Estimators

One of the main advantages of the elastic net estimator over the LASSO estimator is that the elastic net estimator groups the highly correlated variables. Genç and Özkale (2020a) demonstrated that the GO estimation method provides a different grouping variables property. According to this property, the differences $\widehat{\beta}_i^{go} - d\,\widehat{\beta}_i^{ls}$ and $\widehat{\beta}_j^{go} - d\,\widehat{\beta}_j^{ls}$ are close to each other if the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is high, where $d\,\widehat{\beta}^{ls}$ represents the shrunken estimates vector and $\widehat{\beta}_i^{go}$ and $\widehat{\beta}_j^{go}$ represent the i^{th} and j^{th} GO estimates, respectively. This property can be generalized by predetermined b_i and b_j instead of $d\,\widehat{\beta}_i^{ls}$ and $d\,\widehat{\beta}_j^{ls}$ as given by the following theorem.

Theorem 4.2. Let the design matrix $\mathbf{X}$ be in the standardized form, and the response vector $\mathbf{y}$ be centered. Assume that λ and α are fixed and signs of $\widehat{\beta}_i^{sh}$ and $\widehat{\beta}_j^{sh}$ are the same where $\widehat{\beta}_i^{sh} = \widehat{\beta}_i^{sh}(\lambda,\alpha)$ and $\widehat{\beta}_j^{sh} = \widehat{\beta}_j^{sh}(\lambda,\alpha)$ represent the i^{th} and j^{th} shift-ridge or shift-LASSO estimates, respectively. Then we have the inequality

$$U(i,j) \leq \frac{1}{n\lambda(1-\alpha)}\sqrt{2(1-\rho)}$$

where

$$U(i,j) = \frac{1}{\sqrt{\|\mathbf{y}\|_1^2 + n\lambda(1-\alpha)\|\mathbf{b}\|_2^2}}\left|\left(\widehat{\beta}_i^{sh} - b_i\right) - \left(\widehat{\beta}_j^{sh} - b_j\right)\right| \tag{4.17.}$$

and ρ is the sample correlation coefficient between the variables $\mathbf{x}_i$ and $\mathbf{x}_j$.

Proof: We follow the adapted version of the proof of Theorem 4.1 to prove this theorem.

Let

$$Q(\boldsymbol{\beta};\lambda,\alpha) = \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2}\|\boldsymbol{\beta}-\mathbf{b}\|_2^2\right).$$

We write the subdifferrentials of this function with respect to β_i, β_j and set them equal to zero:

$$\left.\frac{\partial Q}{\partial \beta_i}\right|_{\beta_i=\widehat{\beta}_i^{sh}} = -\frac{1}{n}\mathbf{x}_i^\top\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}^{sh}\right) + \lambda\alpha\widehat{s}_i + \lambda(1-\alpha)\left(\widehat{\beta}_i^{sh}-b_i\right) = 0 \quad (4.18.)$$

$$\left.\frac{\partial Q}{\partial \beta_j}\right|_{\beta_j=\widehat{\beta}_j^{sh}} = -\frac{1}{n}\mathbf{x}_j^\top\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}^{sh}\right) + \lambda\alpha\widehat{s}_j + \lambda(1-\alpha)\left(\widehat{\beta}_j^{sh}-b_j\right) = 0 \quad (4.19.)$$

where $\widehat{s}_i$ and $\widehat{s}_j$ have the same sign.

Subtracting Equation (4.18.) from Equation (4.19.) and applying Cauchy Schwarz inequality, we get

$$\left|\widehat{\beta}_j^{sh}-\widehat{\beta}_i^{sh}-(b_j-b_i)\right| \leq \frac{1}{n\lambda(1-\alpha)}\sqrt{\|\mathbf{x}_i-\mathbf{x}_j\|_2^2\,\|\widehat{\mathbf{r}}\|_2^2}. \quad (4.20.)$$

where $\widehat{\mathbf{r}} = \mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}$. Since $\|\mathbf{x}_i-\mathbf{x}_j\|_2^2 = 2(1-\rho)$, we obtain

$$\left|\widehat{\beta}_j^{sh}-\widehat{\beta}_i^{sh}-(b_j-b_i)\right| \leq \frac{1}{n\lambda(1-\alpha)}\sqrt{2(1-\rho)\,\|\widehat{\mathbf{r}}\|_2^2}. \quad (4.21.)$$

Furthermore, $Q\left(\widehat{\boldsymbol{\beta}}^{sh};\lambda,\alpha\right) \leq Q(\mathbf{0};\lambda,\alpha)$ holds because $\widehat{\boldsymbol{\beta}}^{sh}$ is the minimizer of Q.

Hence, we write

$$\frac{1}{2n}\left\|\hat{\mathbf{r}}\right\|_2^2+\lambda\alpha\left\|\widehat{\boldsymbol{\beta}}^{sh}\right\|_1+\frac{\lambda\left(1-\alpha\right)}{2}\left\|\widehat{\boldsymbol{\beta}}^{sh}-\mathbf{b}\right\|_2^2\leq\frac{1}{2n}\left\|\mathbf{y}\right\|_2^2+\frac{\lambda\left(1-\alpha\right)}{2}\left\|\mathbf{b}\right\|_2^2$$

which implies that

$$\left\|\hat{\mathbf{r}}\right\|_2^2\leq\left\|\mathbf{y}\right\|_2^2+n\lambda\left(1-\alpha\right)\left\|\mathbf{b}\right\|_2^2. \tag{4.22.}$$

If we consider Equations (4.21.) and (4.22.) together, then

$$\left|\widehat{\beta}_j^{sh}-\widehat{\beta}_i^{sh}-\left(b_j-b_i\right)\right|\leq\frac{1}{n\lambda\left(1-\alpha\right)}\sqrt{2\left(1-\rho\right)}\sqrt{\left\|\mathbf{y}\right\|_1^2+n\lambda\left(1-\alpha\right)\left\|\mathbf{b}\right\|_2^2}$$

which completes the proof.

We comment Theorem 4.2 in line with Theorem 4.1. $U\left(i,j\right)$ is a unitless quantity. $\left|\left(\widehat{\beta}_i^{sh}-b_i\right)-\left(\widehat{\beta}_j^{sh}-b_j\right)\right|$ is about zero if the correlation coefficient between $\mathbf{x}_i$ and $\mathbf{x}_j$ is large. Therefore, $\widehat{\beta}_j^{sh}-\widehat{\beta}_i^{sh}$ and b_j-b_i are approximately equal, which means the difference between j^{th} and i^{th} estimates of the shift-ridge or shift-LASSO estimator is nearly equal to the corresponding coefficient estimates of the prior information. Also, as the j^{th} estimate of the shift-ridge or shift-LASSO estimator goes to b_j, the i^{th} estimate of the shift-ridge or shift-LASSO estimator goes to b_i with approximately the same speed. Moreover, the grouping property of the shift-ridge and shift-LASSO estimators approaches to that of the elastic net estimator if the vector of the prior information gets close to zero.

4.3. The Stochastic Restricted LASSO Estimator

In this section, we introduce the stochastic restricted LASSO estimator in linear regression, which is a generalization of the mixed estimator with ℓ_1- type penalization. The new estimator enjoys the automatic variable selection property of the LASSO estimator while has better performance in the sense of prediction mean squared error. The stochastic restricted LASSO estimator is also studied by Genç and Özkale (2020c).

The idea of stochastic linear restrictions mentioned in Chapter 3 can be applied with the LASSO idea. We combine the linear regression model given by Equation (2.1.) with the stochastic linear restrictions given by Equation (3.3.) as

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*, \qquad \left(\mathbf{0}, \sigma^2\mathbf{I}_{n+J}\right) \tag{4.23.}$$

where $\mathbf{y}^* = \begin{pmatrix}\mathbf{y}\\ \mathbf{h}\end{pmatrix}$, $\mathbf{X}^* = \begin{pmatrix}\mathbf{X}\\ \mathbf{H}\end{pmatrix}$, $\boldsymbol{\varepsilon}^* = \begin{pmatrix}\boldsymbol{\varepsilon}\\ \mathbf{u}\end{pmatrix}$, $\mathbf{0}$ is the $(n+J)$-vector or of zeros.

The stochastic restricted LASSO model leads to the minimization of the function in the Lagrange form

$$\begin{aligned} Q(\boldsymbol{\beta};\lambda) &= \left(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}\right)^\top \left(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{h} - \mathbf{H}\boldsymbol{\beta})^\top (\mathbf{h} - \mathbf{H}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|. \end{aligned} \tag{4.24.}$$

One can easily see that the function in Equation (4.24.) is the penalized form of a special mixed estimator of Theil and Goldberger (1961) by the ℓ_1-norm $\sum_{j=1}^{p}\left|\beta_j\right|$ of the coefficients. We represent the solution of the minimization problem of the function

given by Equation (4.24.) by $\widehat{\boldsymbol{\beta}}^{srl}(\lambda)$.

KKT conditions given in Section 2.4 provide information about the solution of the stochastic restricted LASSO problem. The KKT conditions of the convex problem given by Equation (4.24.) are

$$\left(\mathbf{X}^{\top}\mathbf{X}+\mathbf{H}^{\top}\mathbf{H}\right)\widehat{\boldsymbol{\beta}}^{srl}(\lambda)+\left(\mathbf{X}^{\top}\mathbf{y}+\mathbf{H}^{\top}\mathbf{h}\right)=\frac{\lambda}{2}\mathbf{s} \tag{4.25.}$$

where $\widehat{\boldsymbol{\beta}}^{srl}(\lambda)$ is the stochastic restricted LASSO solution at λ and $\mathbf{s}\in\partial\left(\sum_{j=1}^{p}\left|\beta_j\right|\right)$ with elements s_j defined by Equation (3.16.) for $j=1,2,\ldots,p$. Equation (4.25.) implies that even though the solution $\widehat{\boldsymbol{\beta}}^{srl}(\lambda)$ to the problem given by Equation (4.24.) may not be unique, the optimal subdifferential $\mathbf{s}$ is unique because it is a function of unique fitted value $\mathbf{X}\widehat{\boldsymbol{\beta}}^{srl}(\lambda)$ for $\lambda>0$ as the case of the LASSO estimator.

4.3.1. The Coordinate Descent Algorithm for the Stochastic Restricted LASSO Estimator

The coordinate descent algorithm given in Subsection 2.4.1.1 can be extended to the stochastic restricted LASSO estimator. To see this, $Q(\boldsymbol{\beta};\lambda)$ given by Equation (4.24.) can be written in the form of

$$Q(\beta_j;\lambda)=\sum_{i=1}^{n}\left(v_i^{(j)}-x_{ij}\beta_j\right)^2+\sum_{i=1}^{J}\left(z_i^{(j)}-h_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{4.26.}$$

for $j=1,2,\ldots,p$, which is a function of β_j when other parameters are fixed. In Equation (4.26.), $v_i^{(j)}=y_i-\sum_{k\neq j}x_{ik}\beta_k$ and $z_i^{(j)}=h_i-\sum_{k\neq j}h_{ik}\beta_k$ are the partial residuals respectively corresponding to the linear models given by Equations

(2.1.) and (3.3.). The subdifferential of the function given by Equation (4.26.) with respect to β_j is

$$\begin{aligned}\partial Q\left(\beta_j;\lambda\right) &= -2\sum_{i=1}^{n} x_{ij}\left(v_i^{(j)} - x_{ij}\beta_j\right) - 2\sum_{i=1}^{J} h_{ij}\left(z_i^{(j)} - h_{ij}\beta_j\right) + \lambda s_j \\ &= -2\left[\sum_{i=1}^{n} x_{ij}v_i^{(j)} + \sum_{i=1}^{J} h_{ij}z_i^{(j)}\right] + 2\beta_j\left[\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{J} h_{ij}^2\right] + \lambda s_j.\end{aligned}$$

Solving $\partial Q\left(\beta_j;\lambda\right) = 0$, we obtain the coordinate descent algorithm iterations for β_j at λ as

$$\widehat{\beta}_j^{srl}(\lambda) \leftarrow \frac{S\left(\sum_{i=1}^{n} x_{ij}v_i^{(j)} + \sum_{i=1}^{J} h_{ij}z_i^{(j)}, \lambda/2\right)}{\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{J} h_{ij}^2}, \; j = 1,2,\ldots,p.$$

When the explanatory variables are centered and standardized with respect to unit-length scaling, the iteration steps can be written as

$$\widehat{\beta}_j^{srl}(\lambda) \leftarrow \frac{S\left(\sum_{i=1}^{n} x_{ij}v_i^{(j)} + \sum_{i=1}^{J} h_{ij}z_i^{(j)}, \lambda/2\right)}{1 + \sum_{i=1}^{J} h_{ij}^2}, \; j = 1,2,\ldots,p.$$

The update rule of the coordinate descent algorithm implies that the stochastic restricted LASSO estimator can perform variable selection for large enough values of λ even if there are not restrictions in the form of $\beta_j + u = 0, u \sim \left(0,\sigma^2\right)$ that is the analyst believes that the j^{th} variable is not significant. In other words, the variable selection property of the stochastic restricted LASSO estimator is restriction-independent.

The computational cost of a complete cycle of the coordinate descent for the stochastic restricted LASSO estimator through all p variables is $O(p(n+J))$ operations. If the number of the linear restrictions J is very small compared to n then the computational cost is approximately $O(pn)$, that is, it is equal to the cost of the method when there are not restrictions. This implies that if $n \gg J$, the stochastic restricted LASSO estimator has no extra cost over the LASSO estimator (on the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$).

5. REAL DATA ANALYSES

We introduced the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators in Chapter 4. In this chapter, we investigate the performance of the estimators with several real-life data sets. The `R` and `C++` programs are used for the coefficient estimation of the estimators and other computations.

5.1. Data Sets

In this section, we describe the real data sets used to compare the methods. The data sets we used for this goal are the prostate (Stamey et al., 1989), diabetes (Efron et al., 2004) and riboflavin (Bühlmann, 2014) data sets.

5.1.1. The Prostate Data

The prostate data come from a study examining the correlation between the level of prostate specific antigen and a number of clinical measures from men who were about to receive a radial prostatectomy. The data consist of 97 observations and 8 explanatory variables. The response is the log of prostate specific antigen (`lpsa`) and the explanatory variables are the logarithm of cancer volume (`lcavol`), the logarithm of prostate weight (`lweight`), age in years (`age`), the natural logarithm of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), the logarithm of capsule penetration (`lcp`), Gleason score (`gleason`) and percent of Gleason score 4 or 5 (`pgg45`). The data were studied in some articles, for instance, Tibshirani (1996) and Zou and Hastie (2005). The correlation matrix of explanatory variables of the prostate data is given by Table 5.1. It is seen from Table 5.1 that the correlation between `gleason` and `pgg45` is 0.75, which is a moderate correlation. Also, the condition number corresponding to the explanatory variables of the data set is 244.19, which

Table 5.1. Correlation table for prostate data

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
lcavol	1.00	0.28	0.22	0.03	0.54	0.68	0.43	0.43	0.73
lweight		1.00	0.35	0.44	0.16	0.16	0.06	0.11	0.43
age			1.00	0.35	0.12	0.13	0.27	0.28	0.17
lbph				1.00	-0.09	-0.01	0.08	0.08	0.18
svi					1.00	0.67	0.32	0.46	0.57
lcp						1.00	0.51	0.63	0.55
gleason							1.00	0.75	0.37
pgg45								1.00	0.42
lpsa									1.00

implies the existence of the multicollinearity among the explanatory variables.

5.1.2. The Diabetes Data

The diabetes data contain 442 observations with 10 explanatory variables obtained from diabetes patients. The explanatory variables are age, sex, body mass index (`BMI`), average blood pressure (`BP`) and six blood serum measurements (`S1` through `S6`) and the response variable is a quantitative measure of disease progression one year after baseline. The data show that some moderate and high correlations among the explanatory variables, specifically, high correlations are 0.9 between `S1`, `S2` and -0.74 between `S3, S4` as shown in Table 5.2.

5.1.3. The Riboflavin Data

The riboflavin data is a genomic data set about riboflavin (vitamin B2) production with Bacillus subtilis. The data were first studied by Bühlmann (2014) and can be available from the `hdi` package of software `R`. The data were also analyzed by Javanmard and Montanari (2014), Zhang et al., (2019), Wang et al., (2019), etc. The

Table 5.2. Correlation table for diabetes data

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
AGE	1	0.17	0.19	0.34	0.26	0.22	-0.08	0.2	0.27	0.3	0.19
SEX		1	0.09	0.24	0.04	0.14	-0.38	0.33	0.15	0.21	0.04
BMI			1	0.4	0.25	0.26	-0.37	0.41	0.45	0.39	0.59
BP				1	0.24	0.19	-0.18	0.26	0.39	0.39	0.44
S1					1	0.9	0.05	0.54	0.52	0.33	0.21
S2						1	-0.2	0.66	0.32	0.29	0.17
S3							1	-0.74	-0.4	-0.27	-0.39
S4								1	0.62	0.42	0.43
S5									1	0.46	0.57
S6										1	0.38
Y											1

main purpose of analyzing the data was to identify the genes that have a positive effect on the production of riboflavin and then to build genetically the bacteria that produce higher yields. The data consist of 71 observations and 4088 explanatory variables corresponding to genes which describe the expression level of each gene of the bacteria. The explanatory variables are handled as the logarithm of the expression level of the genes. The response variable is the logarithm of the riboflavin production rate of the bacteria.

5.2. Real Data Analysis for the GO Estimator

In this section, we investigate the OLS, ridge, OK, LASSO, elastic net and GO estimators discussed in Section 4.1 with prostate and diabetes data sets. The explanatory variables were centered and standardized such that $\sum_{i=1}^{n} x_{ij}^2/n = 1$ for $j = 1, 2, \ldots, p$ (unit-normal scaling). The estimators were applied to the standardized data. The data are divided into 70% training and 30% test data sets. Ten-fold cross-validation is applied to the training data for tuning parameter selection. The standard error estimates of the model coefficients are obtained by the bootstrap method with

Table 5.3. Comparison of the OLS, ridge, OK, LASSO, elastic net and GO estimators on prostate data set with parameter values and selected variables

Method	Parameter values	Variables selected
OLS	-	All the variables
Ridge	$\lambda = 0.088$	All the variables
OK	$\lambda = 1.702$, $d = 0.60$	All the variables
LASSO	$\lambda = 0.125$	1,2,4,5,8
Elastic net	$\alpha = 0.64$, $\lambda = 0.147$	1,2,4,5,8
GO	$\alpha = 0.33$, $\lambda = 0.577$, $d = 0.83$	1,2,4,5,8

1000 repetitions. We compare the model performances PMSE, which is defined by Equation (2.4.), over the response values in the test set. We also report the selected variables for each data set.

5.2.1. The Prostate Data Analysis for the GO Estimator

Table 5.3 shows the resulted values of the optimal tuning parameters of each method in the analysis result and the variables that are selected according to each method. The tuning parameter estimates of the elastic net and GO estimators are obtained by minimizing mean cross-validation error while that of the LASSO is obtained by one standard error rule. The variables selected by the GO estimator according to the results in Table 5.3 are `lcavol`, `lweight`, `lbph`, `svi`, `pgg45`. Elastic net and LASSO estimators also select the same variables.

The model coefficient estimates and bootstrap standard error estimates obtained from the analysis are given in Table 5.4. The OLS, ridge and OK estimates do not have zero components. On the contrary, the LASSO, elastic net and GO estimators yield coefficient estimates with zeros. The problem of the incorrect sign in the estimates of these variables is sorted out by the ridge and OK estimators. The LASSO, elastic net and GO estimators discard both of these variables from the model. Three of the LASSO, elastic net and GO estimates are zero. While the signs of the elastic net and

GO estimators are consistent, the GO estimator shrinks the coefficients more than the elastic net estimator, which satisfies our expectations. The length of the estimates of the coefficient vectors and PMSE values obtained from the analysis are given at the bottom row of Table 5.4. The length of the GO estimates vector is the smallest which is a result of the number of variables determined to be zero and shrinking. The length of the GO estimates vector is 0.38% smaller than the length of the elastic net estimates vector and 0.44% than that of LASSO estimates. The GO has the best result in terms of PMSE which is followed by the LASSO and elastic net estimators, respectively. Figure 5.1 shows the box plots of the 1000 bootstrap replications of the coefficient estimates of the prostate data obtained by each estimation method. The bootstrap distributions of the variables `age`, `lcp` and `gleason` are symmetric bootstrap distributed in which `gleason` exhibits skewed bootstrap distribution for the LASSO, elastic net and GO estimators.

The results in Table 5.3 correspond to tuning parameters selected via ten-fold cross-validation. To gain a more general perspective, we give the plot of PMSE values versus λ values for the Prostate data set in Figure 5.2 where α and d are fixed at their optimum values given in Table 5.3. In this plot, the ranges for λ values are determined such that they include the optimal λ values obtained by cross-validation. The optimal λ values are slightly different from Table 5.3 since we fit the models on the training set without cross-validation and compute the PMSE values to obtain Figure 5.2 for illustrative purposes. According to the plot, the LASSO and elastic net estimators are superior to the GO estimator for very small values of λ. For λ values approximately between 0.4 and 1.4, the GO estimator outperforms the other methods. The OK and ridge estimators dominate the GO estimator for large λ values. Figure 5.2 shows that none of the methods is superior to the others uniformly for all the values of λ, superiority of the one estimator over others depends on the chosen λ value. However,

Table 5.4. Coefficient estimates, length of the coefficient vectors and PMSE values for the OLS, ridge, OK, LASSO, elastic net and GO estimators in prostate data

	OLS	Ridge	OK	LASSO	Elastic net	GO
Intercept	2.4523	2.4523	2.4523	2.4523	2.4523	2.4523
	(0.0904)	(0.0896)	(0.0905)	(0.0934)	(0.0923)	(0.0951)
lcavol	0.711	0.6075	0.5252	0.5678	0.5401	0.5315
	(0.1326)	(0.1075)	(0.0907)	(0.1149)	(0.1059)	(0.1069)
lweight	0.2905	0.2843	0.2415	0.2193	0.2286	0.1893
	(0.1207)	(0.1013)	(0.0845)	(0.0995)	(0.0955)	(0.0952)
age	-0.1415	-0.1103	-0.0786	0	0	0
	(0.087)	(0.0788)	(0.0659)	(0.0156)	(0.0262)	(0.0111)
lbph	0.2104	0.2005	0.1679	0.0819	0.1097	0.0655
	(0.1185)	(0.1019)	(0.0845)	(0.0894)	(0.0912)	(0.0813)
svi	0.3073	0.2833	0.2481	0.1538	0.1817	0.1429
	(0.136)	(0.1109)	(0.0933)	(0.1109)	(0.1077)	(0.1045)
lcp	-0.2868	-0.1621	-0.1391	0	0	0
	(0.138)	(0.0915)	(0.0879)	(0.0077)	(0.0111)	(0.0126)
gleason	-0.0208	0.0121	0.0093	0	0	0
	(0.1505)	(0.1082)	(0.0997)	(0.0345)	(0.0405)	(0.0293)
pgg45	0.2753	0.2058	0.2035	0.0504	0.0774	0.0626
	(0.146)	(0.0931)	(0.0944)	(0.0724)	(0.0746)	(0.0731)
$\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}$	2.6308	2.5817	2.5505	2.5332	2.5316	2.5221
PMSE	0.5522	0.5209	0.5092	0.4475	0.4593	0.4433

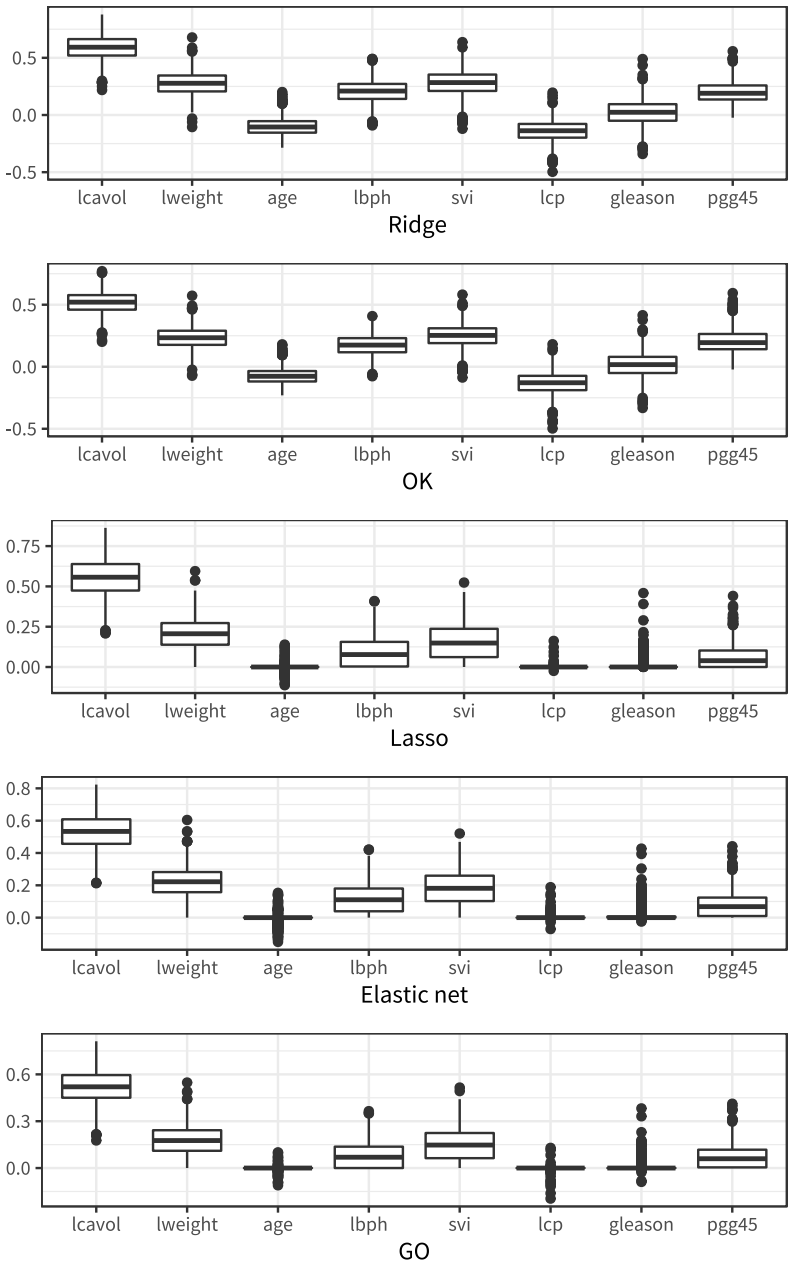

Figure 5.1. Box plots of bootstrap coefficient estimates of the prostate data for the ridge, OK, LASSO, elastic net and GO estimators

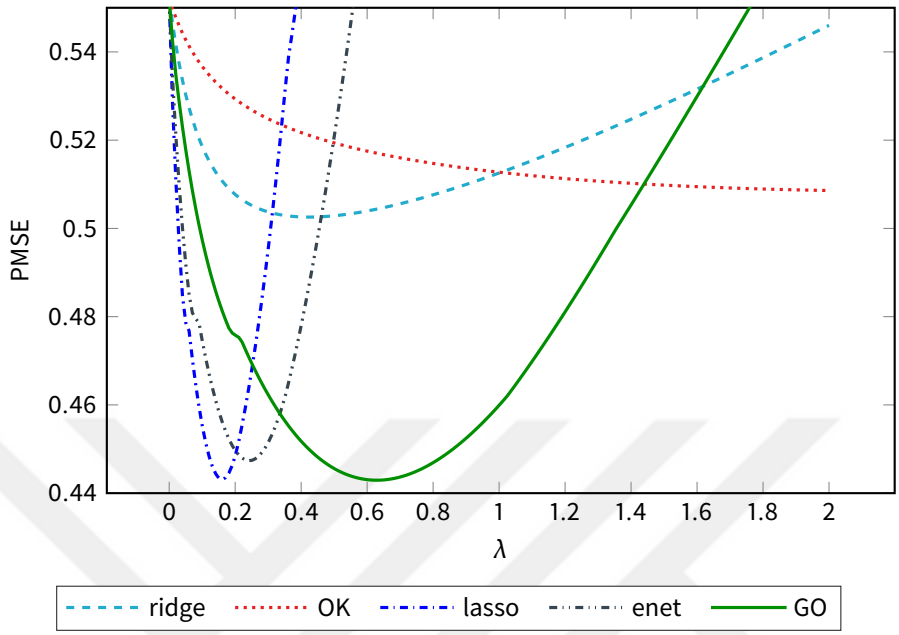

Figure 5.2. PMSE values of the prostate data set for the ridge, OK, LASSO, elastic net and GO estimators

in general, the GO estimator at its optimal tuning parameter values outperforms the other estimators at their corresponding optimal tuning parameters.

5.2.2. The Diabetes Data Analysis for the GO Estimator

The optimal tuning parameters and selected variables by the methods for the diabetes data are shown in Table 5.5. The estimates of the tuning parameters are obtained by minimizing the mean cross-validation error. The LASSO estimator selects five variables, `SEX`, `BMI`, `BP`, `S3`, `S5` while the GO estimator selects eight variables including the variables selected by LASSO, `S1`, `S2`, `S4`, additionally, and the elastic net estimator does not do variable selection. Therefore, the LASSO estimator discards positively highly correlated variables `S1` and `S2` together while selects one of the negatively highly correlated variables `S3` and `S4`. The GO estimator discards `AGE` and

Table 5.5. Comparison of the OLS, ridge, OK, LASSO, elastic net and GO estimators on diabetes data with parameter values and selected variables

Method	Parameter values	Variables selected
OLS	-	All the variables
Ridge	$\lambda = 5.745$	All the variables
OK	$\lambda = 4.692$, $d = 0.24$	All the variables
LASSO	$\lambda = 2.017$	2,3,4,7,9
Elastic net	$\alpha = 0.54$, $\lambda = 0.131$	All the variables
GO	$\alpha = 0.24$, $\lambda = 4.762$, $d = 0.99$	2,3,4,5,6,7,8,9

S6 which are also discarded by the LASSO estimator.

The model coefficient estimates and bootstrap standard error estimates of diabetes data are given in Table 5.6. The problem of the incorrect sign in the estimates of S1 and S2 is sorted out by ridge regression as in the prostate data. We give norms of the coefficient vectors and PMSE values in the bottom two rows of Table 5.6. The GO estimator has the smallest PMSE value among all the methods while the LASSO estimator has the largest one. This shows that the GO estimator reduces the PMSE for the sake of a little decrease in the variable selection property. The norm of the GO estimator is smaller than that of the elastic net estimator. In this data set, the GO estimator dominates the elastic net estimator in terms of variable selection, norm of coefficient vector and PMSE values. Also, Figure 5.3 shows the box plots of the bootstrap coefficient estimates of the diabetes data obtained by the estimation methods after 1000 bootstrap iterations. The estimated coefficient of AGE displays symmetric bootstrap distribution while the estimated coefficient of S1, S2, S4 and S6 display skewed bootstrap distributions in view of the LASSO estimator. On the other hand, both the estimated coefficients of AGE and S6 show symmetric bootstrap distributions.

The PMSE values of the estimators versus λ when α and d are fixed at their corresponding optimum values are given by Figure 5.4. It shows a similar pattern to Figure 5.2. That is, there is not one optimum λ that one estimator outperforms the

Table 5.6. Coefficient estimates, length of the coefficient vectors and PMSE values for the OLS, ridge, OK, LASSO, elastic net and GO estimators in diabetes data

	OLS	Ridge	OK	LASSO	Elastic net	GO
Intercept	150.9159	150.9159	150.9159	150.9159	150.9159	150.9159
	(3.1626)	(3.163)	(3.1592)	(3.174)	(3.1598)	(3.1626)
AGE	-0.997	-0.6364	-0.7478	0	-0.8353	0
	(3.2605)	(2.9521)	(3.0605)	(1.6769)	(3.1862)	(2.3359)
SEX	-13.1097	-11.599	-12.1438	-9.2736	-12.9168	-10.9202
	(3.4059)	(3.0604)	(3.1817)	(3.3944)	(3.3934)	(3.3647)
BMI	22.3371	21.6054	21.9366	22.171	22.4908	22.4623
	(3.8076)	(3.3324)	(3.4971)	(3.7863)	(3.7978)	(3.8028)
BP	16.1427	15.1594	15.5082	13.5725	15.9869	14.6981
	(3.6001)	(3.1219)	(3.2847)	(3.5157)	(3.5736)	(3.5323)
S1	-37.0154	-3.5769	-12.1239	0	-25.6381	-17.3065
	(24.7961)	(2.924)	(8.1700)	(2.4766)	(18.1816)	(15.6423)
S2	25.712	-1.5712	5.34	0	16.3201	9.349
	(20.4986)	(3.1489)	(7.0291)	(1.322)	(15.2081)	(13.3582)
S3	3.6807	-9.719	-6.407	-12.4039	-1.1508	-2.7727
	(12.191)	(3.0815)	(5.0367)	(3.6598)	(9.2833)	(6.6328)
S4	8.9369	6.6412	7.1637	0	7.8256	7.0682
	(9.3458)	(4.9118)	(6.2358)	(3.792)	(8.6078)	(6.2669)
S5	36.4086	22.4843	26.2281	23.5105	32.2762	29.4738
	(9.4536)	(3.7269)	(4.5873)	(4.2761)	(7.5218)	(7.0676)
S6	0.1857	1.6659	1.1786	0	0.286	0
	(3.5786)	(3.1892)	(3.3255)	(1.9728)	(3.484)	(2.4932)
$\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}$	164.7976	155.7860	156.8519	155.705	160.4139	157.8883
PMSE	3078.3164	3074.4676	3067.7078	3111.6118	3067.4051	3058.2837

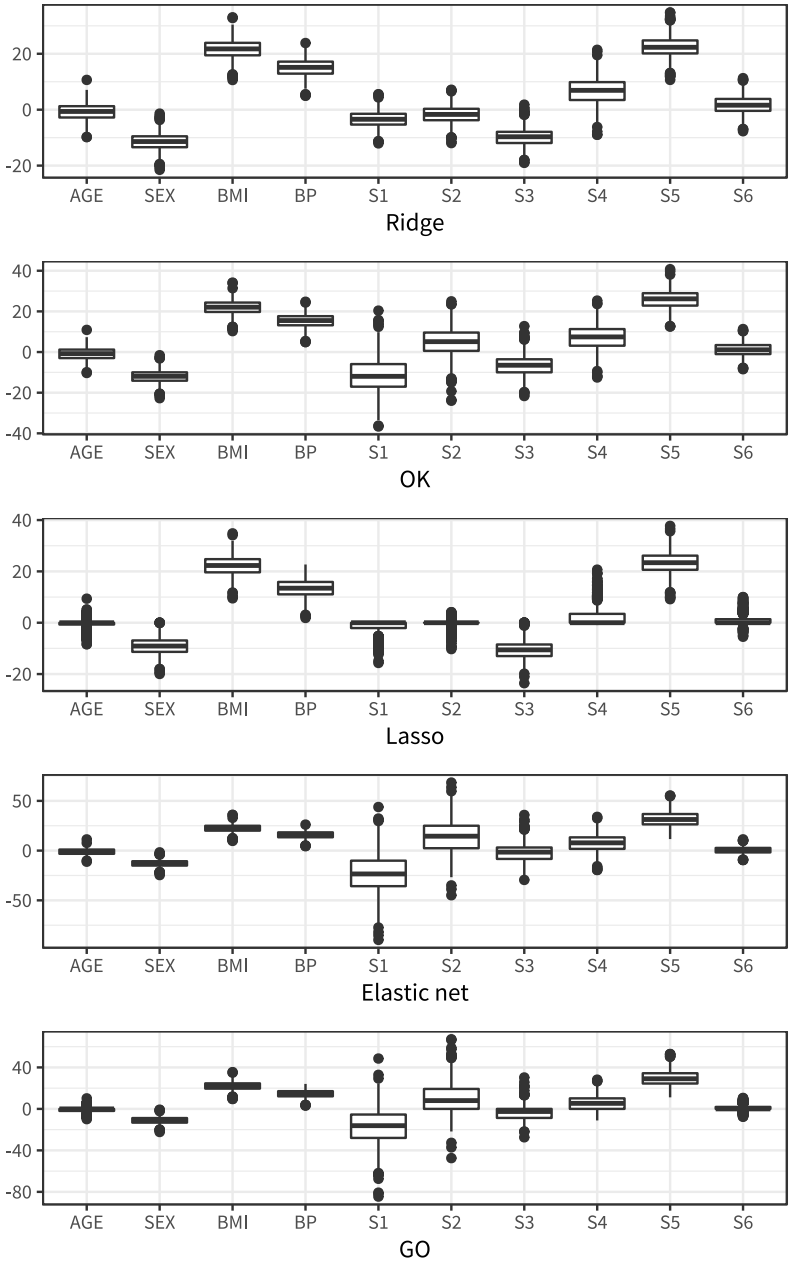

Figure 5.3. Box plots of bootstrap coefficient estimates of the diabetes data for the ridge, OK, LASSO, elastic net and GO estimators

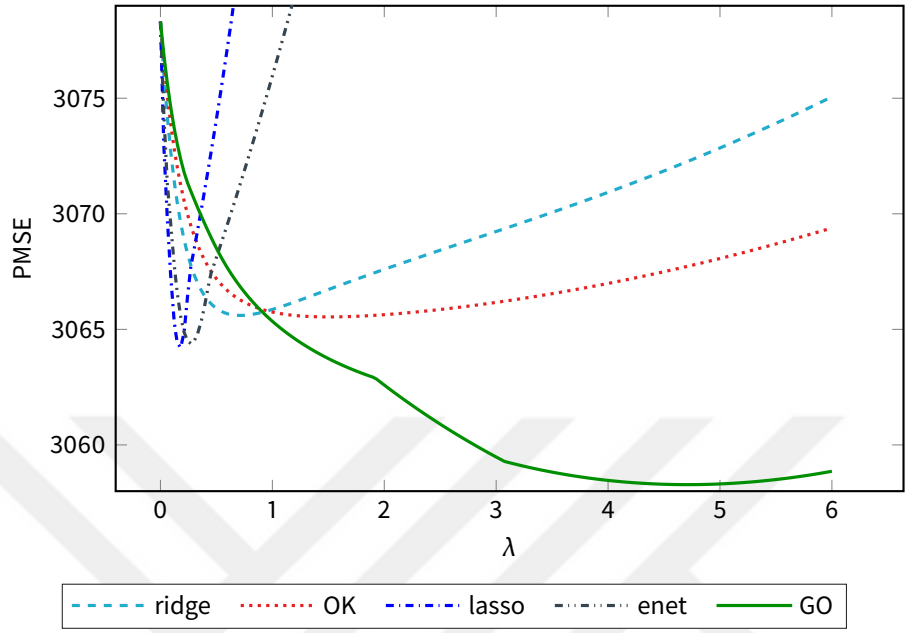

Figure 5.4. PMSE values of the diabetes data set for the ridge, OK, LASSO, elastic net and GO estimators

others. The superiority of one estimator depends on the selected λ value. The GO estimator in this data set dominates the others for λ values larger than 1.

The examination of the data analysis presents that the GO estimator is the best for the prostate and diabetes data sets when the tuning parameters are optimized.

5.3. Real Data Analysis for the Shift-ridge and Shift-LASSO Estimators

We investigate the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators discussed in Section 4.2 with riboflavin data in this section. The explanatory variables were centered and standardized such that $\sum_{i=1}^{n} x_{ij}^2/n = 1$ for $j = 1, 2, \ldots, p$ (unit-normal scaling) and the response variable is centered. We start analyzing the data set by splitting it into training and test sets with proportion 80%/20%. Five-fold cross-validation is applied to the training data for tuning parameter selection.

Table 5.7. Comparison of the ridge, LASSO, elastic net shift-ridge and shift-LASSO estimators on riboflavin data with tuning parameters, PMSE, active set sizes and reduce rates

					Reduce rate (%)	
	λ	α	PMSE	Active set size	shift-ridge	shift-LASSO
Ridge	0.224	-	0.3015	4088	45.85	57.27
LASSO	0.031	-	0.1894	42	13.82	32.00
Elastic net	0.389	0.05	0.1619	219	-0.81	20.46
Shift-ridge	0.191	0.15	0.1632	109	-	21.10
Shift-LASSO	0.865	0.25	0.1288	18	-26.74	-

The estimated tuning parameters, PMSE values and active set sizes of the analysis are given in Table 5.7. For the purpose of comparing the performances of the methods, Table 5.7 also presents reduce rate (RR) with respect to the shift-ridge and shift-LASSO, which is computed as

$$RR = \frac{\text{PMSE}\left(\widehat{\boldsymbol{\beta}}\right) - \text{PMSE}\left(\widehat{\boldsymbol{\beta}}^{sh}\right)}{\text{PMSE}\left(\widehat{\boldsymbol{\beta}}\right)} \cdot 100$$

where $\widehat{\boldsymbol{\beta}}$ is any estimator to be compared with the shift-ridge or shift-LASSO, $\widehat{\boldsymbol{\beta}}^{sh}$ represents the shift-ridge or shift-LASSO and $\text{PMSE}(\cdot)$ is the computed PMSE of the relevant estimator.

According to Table 5.7, the shift-LASSO estimator has the best PMSE in the sense that it has the smallest PMSE. The shift-LASSO estimator is followed by the elastic net and shift-ridge estimators, respectively. The ridge, LASSO, elastic net and shift-ridge estimators are 57.27%, 32.00%, 20.46% and 21.10% larger than the shift-

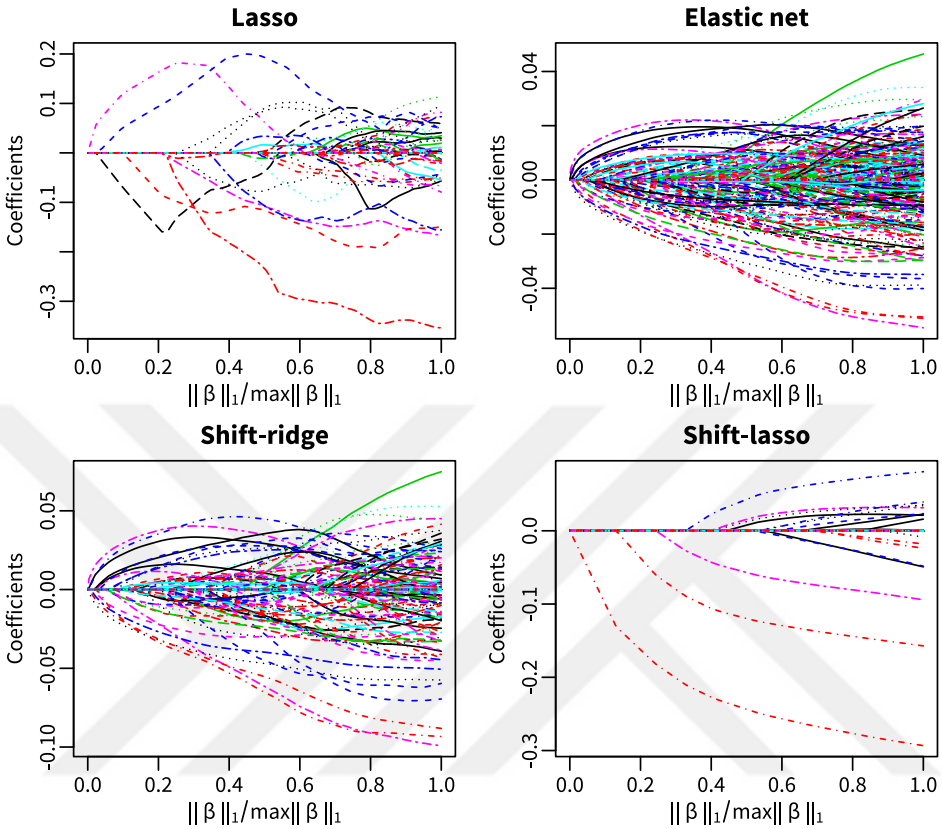

Figure 5.5. Coefficient paths of the ridge, LASSO, elastic net shift-ridge and shift-LASSO estimators for the riboflavin data

LASSO estimator with respect to the PMSE criterion. The elastic net estimator does not improve much (only %0.81 smaller than shift-ridge estimator) and has comparable results with the shift-ridge estimator.

Although the elastic net and shift-ridge estimators have close PMSE values, the shift-ridge estimator selects less variables than the elastic net estimator. Also, the shift-LASSO estimator yields the most parsimonious model for this data set. The resulting model contains 18 genes corresponding to the nonzero parameters of the shift-LASSO estimator.

Figure 5.5 shows the estimates as a function of $\|\boldsymbol{\beta}\|_1 \,/\, \max \|\boldsymbol{\beta}\|_1$ for the riboflavin data for 100 different values of λ when α is set to its optimal value for the LASSO, elastic net, shift-ridge and shift-LASSO estimators. The elastic net and shift-ridge estimators have very similar paths and a smoother path than the LASSO estimator. Also, considering the top right and bottom left of Figure 5.5, it can be seen that the shrinkage structures of the elastic net and shift-ridge estimators are different although they look very similar. The shift-LASSO estimator also has a smoother path than the LASSO estimator and sets most of the coefficients to zero faster than the LASSO estimator, which is coherent with the optimal estimates that the methods give.

The real data analysis shows that the shift-LASSO estimator dominates the other methods in the sense of having smaller PMSE. When the results are evaluated, it is observed that the shift-ridge and shift-LASSO estimators are alternative methods to the well-known methods in the literature in the high-dimensional setting.

5.4. Real Data Analysis for the Stochastic Restricted LASSO Estimator

In this section, we analyze the prostate and riboflavin data sets to compare the performance of the stochastic restricted LASSO estimator with the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators. The stochastic restricted LASSO estimator is defined in Section 4.3. For each data sets, we center the response variable, $n^{-1}\sum_{i=1}^{n} y_i = 0$, and standardize the explanatory variables by unit-length scaling, $n^{-1}\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, 2, \ldots, p$ before the analysis.

5.4.1. The Prostate Data Analysis for the Stochastic Restricted LASSO Estimator

The prostate data are divided into 70% training and 30% test data sets for the analysis. We compare the model performances with the PMSE as well as the SMSE. For any estimator $\widehat{\boldsymbol{\beta}}$, the PMSE is defined by Equation (2.4.) and the SMSE is defined by Equation (2.2.). The LASSO and stochastic restricted LASSO estimators do not have closed-form variance and bias expressions, hence, we apply the bootstrap method with 1000 repetitions to estimate these quantities and obtain the SMSE and PMSE values of the estimators. To estimate the tuning parameter, we select the grid of λ values $\lambda = 2^{k-14}$, $k = 0, 1, \ldots, 28$ in the same line with Park and Yoon (2011) and follow two methods: Ten-fold cross-validation and minimizing BIC. First, we fit the OLS, ridge and LASSO estimators to the data. For the stochastic restricted ridge and stochastic restricted LASSO estimators, we consider three cases of restrictions:

Case 1: For this case, we utilize from the results given by Table 5.8. After examining the coefficient estimates, we see that the sum of the coefficient estimates of `lcavol` and `lweight` is approximately same for different estimators. However, there is uncertainty about the total. Therefore, by Chebyshev inequality, we consider the inequality $\mathrm{P}(|\beta_1 + \beta_2 - 7| < 3\sigma) \geq 0.888$ which implies $\beta_1 + \beta_2 \in [5.8, 8.2]$ with probability 0.888 for $\sigma = 0.4$. This information let us take $\mathbf{H} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$ and $\mathbf{h} = 7$, that is we have a stochastic restriction $7 = \beta_1 + \beta_2 + u$, $u \sim (0, 0.4^2)$.

Case 2: Similar to Case 1, we utilize from the inequality $\mathrm{P}\left(\left|\sum_{j=1}^{8} \beta_j - 10\right| < 3\sigma\right) \geq 0.888$ which implies $\sum_{j=1}^{8} \beta_j \in [8.80, 11.20]$ with probability 0.888 for $\sigma = 0.4$. Then, by $\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$ and $\mathbf{h} = 10$, we have the stochastic

restriction $10 = \beta_1 + \beta_2 + \cdots + \beta_8 + u, u \sim \left(0, 0.4^2\right)$.

Case 3: In this case, we combine the restrictions in Case 1 and Case 2 and consider these restrictions simultaneously, which implies

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \mathbf{h} = \begin{bmatrix} 7 \\ 10 \end{bmatrix} \text{ and } \mathbf{u} \sim \left[\mathbf{0}, 0.4^2\mathbf{I}_2\right].$$

Therefore, we expect that any estimate between these intervals from the stochastic restricted estimators should perform better than their unrestricted counterparts with the assumption that the standard deviations of the coefficients are given values.

We present the results in Table 5.8. We used abbreviations for the estimators to gain space for Table 5.8. The abbreviations are RE for ridge estimator, LE for LASSO estimator, SRRE for stochastic restricted ridge estimator, SRLE for stochastic restricted LASSO estimator, CV for ten-fold cross-validation and BIC for minimum BIC value on the test set.

The correlation between the variables `gleason` and `pgg45` is 0.75 which is a moderate correlation. The problem of the incorrect sign of `gleason` is fixed by the penalized estimators with both CV and BIC methods. The other coefficient estimates have the same sign pattern for all the estimators except the zero estimates introduced by the LE and SRLE. As we expect, only the LE and SRLE yield coefficient estimates as zero. According to the coefficient estimates from Table 5.8, the SRLE(CV) retains the same variables with the LE(CV) in Case 1. In Case 2, the SRLE(CV) behaves like the LE(CV) while the SRLE(BIC) behaves like the LE(BIC) in variable selection. In Case 3, the SLR(CV) selects less variables than the LE(CV) while the SRLE(BIC) selects the same variables with the LE(BIC). All the estimators have smaller SMSE

Table 5.8. Coefficient estimates, SMSE and PMSE values for the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators in prostate data

Case	Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	SMSE	PMSE
	OLS	5.820	2.377	-1.158	1.722	2.515	-2.348	-0.170	2.253	8.721	0.908
	RE(CV)	5.074	2.337	-0.936	1.653	2.343	-1.442	0.073	1.743	8.186	0.827
	RE(BIC)	3.891	2.154	-0.514	1.462	2.026	-0.251	0.300	1.205	6.935	0.721
	LE(CV)	5.620	2.346	-1.031	1.652	2.390	-1.944	0.000	1.917	7.719	0.868
	LE(BIC)	4.709	1.993	0.000	1.160	1.623	0.000	0.000	0.728	4.319	0.701
Case 1	SRRE(CV)	4.631	2.323	-0.779	1.563	2.190	-0.919	0.189	1.476	7.167	0.785
	SRRE(BIC)	4.261	2.422	-0.601	1.344	1.898	-0.412	0.292	1.202	6.318	0.747
	SRLE(CV)	5.093	2.086	-0.787	1.711	2.378	-1.340	0.000	1.668	6.318	0.805
	SRLE(BIC)	5.158	2.075	-0.916	1.813	2.559	-1.670	0.000	1.870	7.167	0.830
Case 2	SRRE(CV)	4.565	2.216	-0.820	1.456	2.117	-0.931	0.123	1.448	7.508	0.719
	SRRE(BIC)	3.887	2.128	-0.538	1.412	1.991	-0.269	0.275	1.191	6.673	0.696
	SRLE(CV)	5.457	2.227	-0.967	1.395	2.104	-1.611	0.000	1.614	6.675	0.758
	SRLE(BIC)	4.686	1.921	0.000	0.981	1.490	0.000	0.000	0.613	3.397	0.644
Case 3	SRRE(CV)	4.659	2.277	-0.845	1.420	2.080	-0.979	0.118	1.450	6.811	0.725
	SRRE(BIC)	3.781	2.451	-0.354	1.096	1.586	0.067	0.372	0.972	4.720	0.701
	SRLE(CV)	4.840	2.072	-0.049	1.046	1.484	0.000	0.000	0.671	3.350	0.694
	SRLE(BIC)	4.852	2.036	0.000	0.886	1.365	0.000	0.000	0.566	2.611	0.660

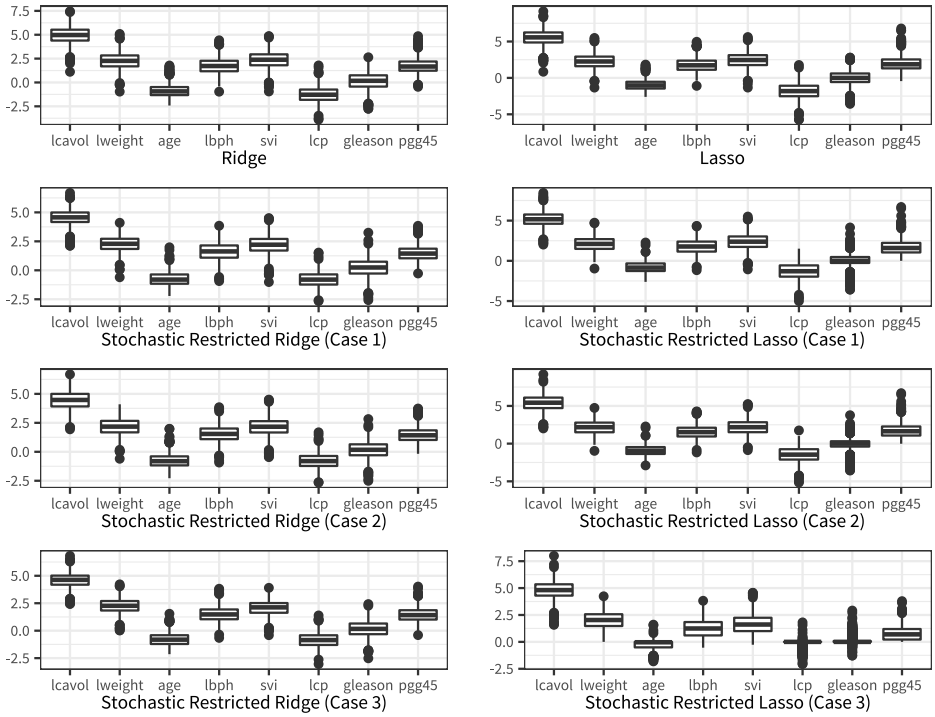

Figure 5.6. Box plots of bootstrap coefficient estimates of the prostate data for the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators obtained by cross-validation

than the OLS estimator. The SRRE outperforms the RE when the tuning parameter is chosen by cross-validation or BIC. Therefore, the SRRE has improvement over the RE. The SRLE outperforms the LE with regard to the SMSE except that the LE(BIC) has better performance than the SRLE(BIC) in Case 1. These comments are also valid for the PMSE criterion.

Figure 5.6 and 5.7 show the box plots of the 1000 bootstrap replications of the coefficient estimates of the prostate data obtained by each estimation method with the criteria cross-validation and BIC, respectively. The coefficient estimates obtained from cross-validation given in Figure 5.6 imply that the SRRE and SRLE concentrate

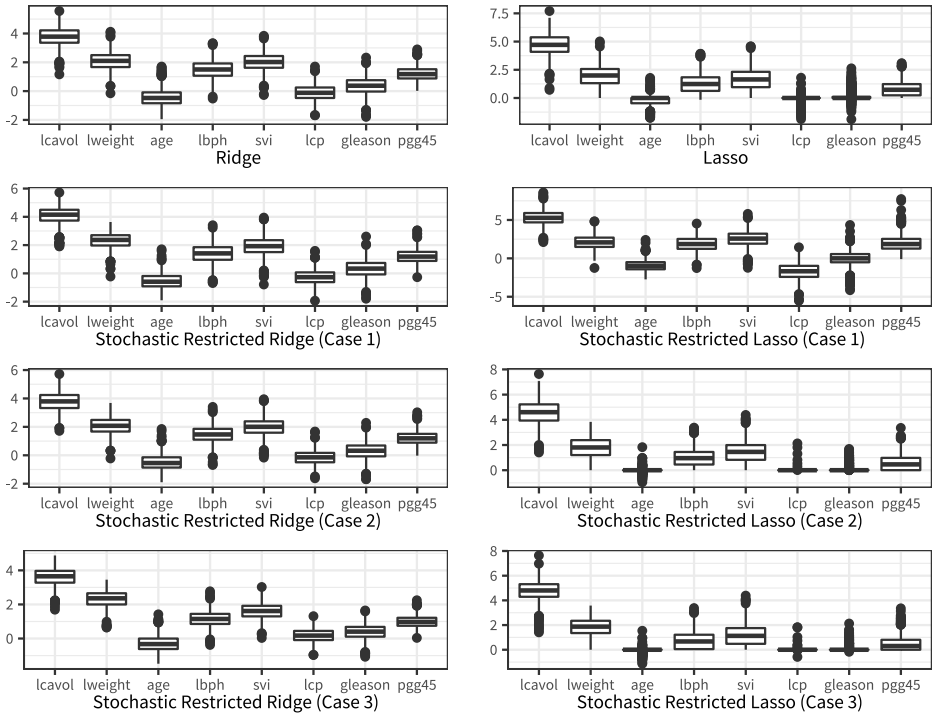

Figure 5.7. Box plots of bootstrap coefficient estimates of the prostate data for the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators obtained by BIC criterion

on the intervals that the restrictions hold for each case. `gleason` is discarded by all the LASSO-type methods while in Case 3, `lcp` is also out of the model. Moreover, the LE and SRLE show smaller standard deviations in this case. The plots corresponding to the coefficient estimates from BIC are given in Figure 5.7. The general pattern is similar to that of Figure 5.6 except that some of the estimates from the LE and SRLE have higher standard deviations compared to the estimates from the cross-validation in Case 2 and Case 3. However, this method discards more variables than the cross-validation in these cases.

For illustrative purposes, we plot the SMSE values obtained from the training set as a function of λ for the penalized regression methods in Figure 5.8. We plot the ridge-stochastic restricted ridge estimators and the LASSO-stochastic restricted LASSO estimators separately to emphasize the differences in the paths they follow.

In all cases, the stochastic restricted ridge and stochastic restricted LASSO estimators have smaller SMSE than their counterparts. This difference is more pronounced in Case 3.

We also plot the PMSE values as a function of λ in Figure 5.9 for the prostate data. Different from the results in Table 5.8, we here obtained the PMSE values from only one validation set (without cross-validation) to illustrate the performances of the methods graphically. For the same reason, the results reported in Table 5.8 can be different from the results of the graphical representation. We observe that each method outperforms the other methods for some λ values.

5.4.2. The Riboflavin Data Analysis for the Stochastic Restricted LASSO Estimator

In this subsection, we aim to compare the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators on the riboflavin data. We start

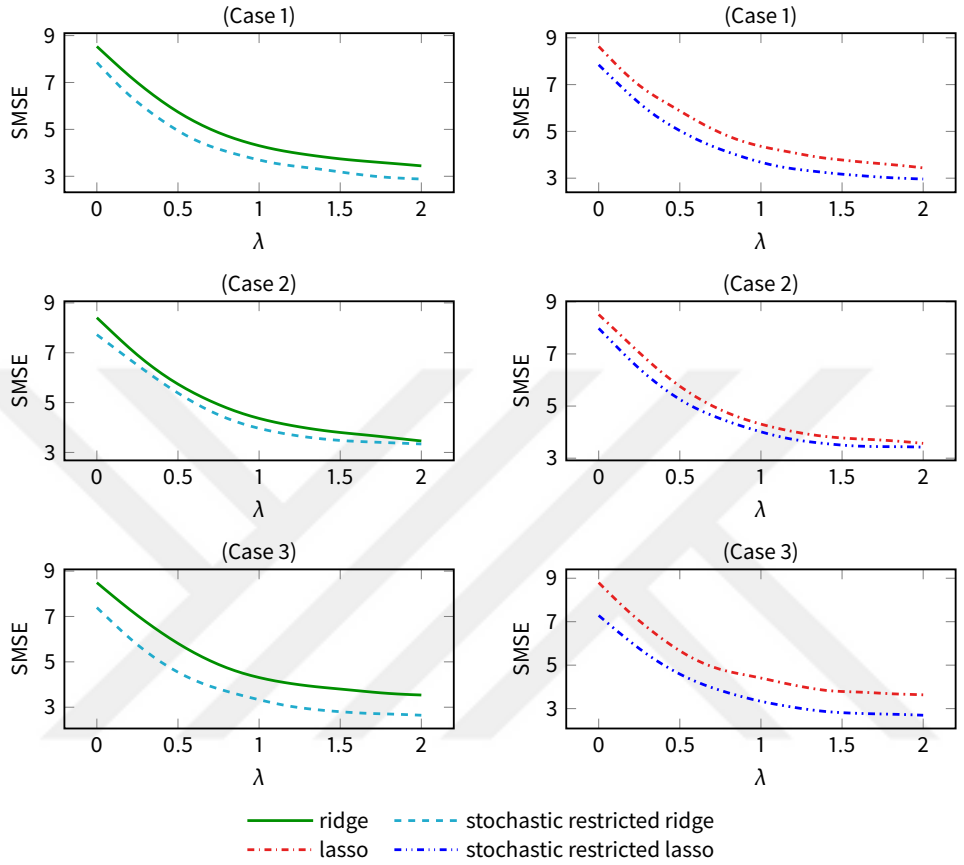

Figure 5.8. SMSE plots as a function of λ for the ridge-stochastic restricted ridge estimators and the LASSO-stochastic restricted LASSO estimators

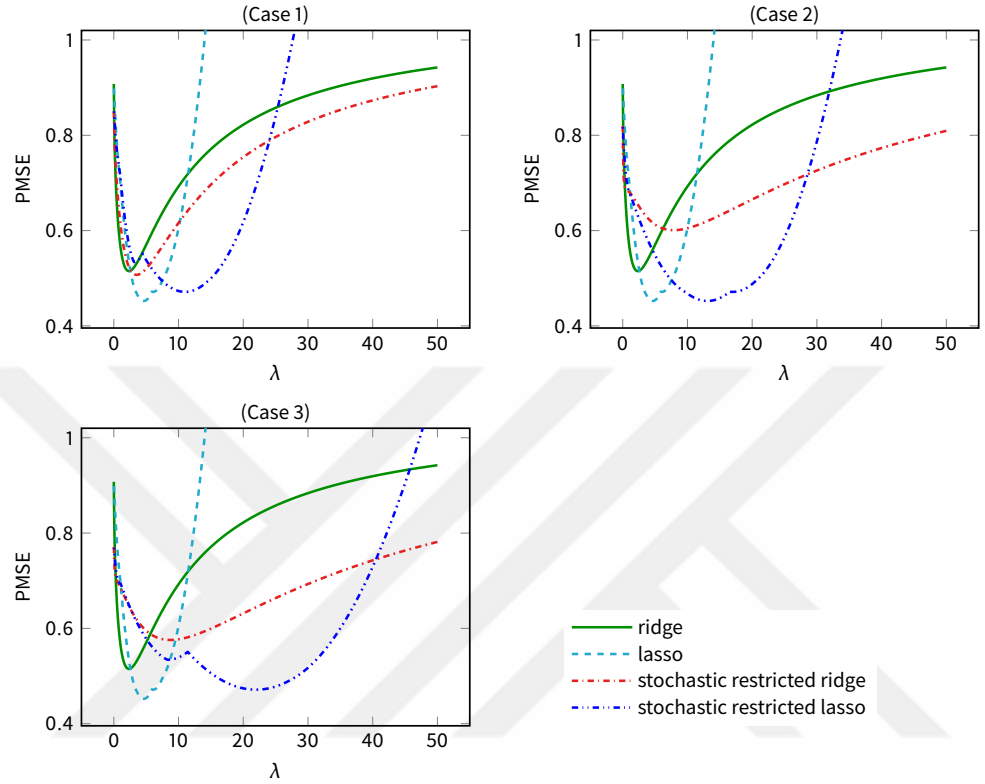

Figure 5.9. PMSE plots as a function of λ for the ridge-stochastic restricted ridge estimators and the LASSO-stochastic restricted LASSO estimators

Table 5.9. Comparison of the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators on riboflavin data where the tuning parameters are selected by ten-fold cross-validation

	RE(CV)	LE(CV)	SRRE(CV)	SRLE(CV)
λ	0.0078	0.0002	0.0001	0.0010
PMSE	2.042	1.796	2.097	1.528
Active set size	4088	59	4088	43

analyzing the data set by splitting it into training and test sets with proportion 80%/20%. We use the same grid of λ values with the prostate data analysis given in Subsection 5.4.1 and estimate the tuning parameter by ten-fold cross-validation and BIC criterion. To compare the performance of the estimators, we used the PMSE for examining the prediction accuracy and active set sizes.

The stochastic linear restriction is taken as follows: When the ridge estimates of the regression coefficients are obtained, it is observed that some of the regression coefficients shrink very close to zero. Therefore, there is uncertainty whether these regression coefficients are really zero or not. Hence, we obtained these regression coefficients by using the criterion $\left|\beta_j\right| < 0.0005$ where 0.0005 is arbitrarily selected such that sufficient regression coefficients are very close to zero. These variables are also discarded from the model by the LASSO. Afterwards h_j is set to 0 and the j-th row of $\mathbf{H}$ matrix is set as the vector whose elements are zero's except its j-th element is 1. Since there is uncertainty whether the determined regression coefficients are really zero, we used the stochastic linear restriction. Uncertainty measure is satisfied as $u_i \sim \left(0, \hat{\sigma}_R^2\right)$ where $\hat{\sigma}_R^2$ is the ridge estimate of the variance of the 71 observations.

The estimated tuning parameters, PMSE values and active set sizes of the analysis are given in Tables 5.9-5.10.

According to Tables 5.9-5.10, the stochastic restricted LASSO estimator has

Table 5.10. Comparison of the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators on riboflavin data where the tuning parameters are selected by BIC

	RE(BIC)	LE(BIC)	SRRE(BIC)	SRLE(BIC)
λ	0.0010	0.0020	0.0078	0.0010
PMSE	2.086	1.555	1.941	1.528
Active set size	4088	43	4088	43

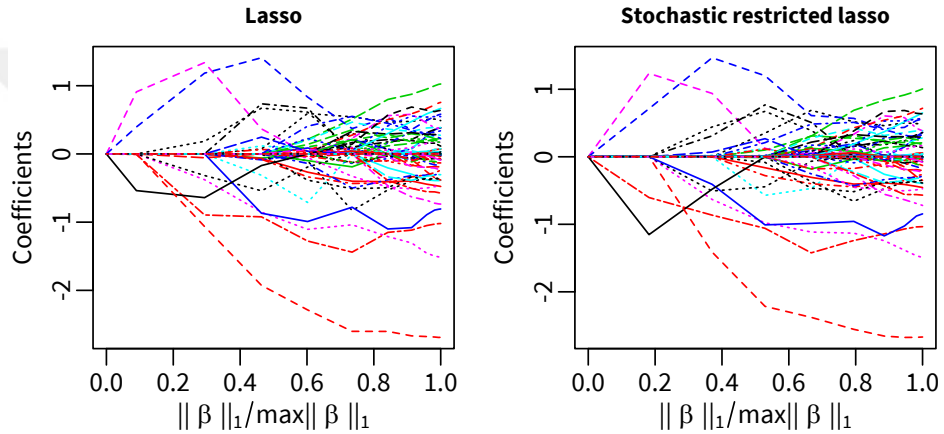

Figure 5.10. Coefficient path of the LASSO and stochastic restricted LASSO estimators in riboflavin data.

the best PMSE in the sense that it has the smallest PMSE. Both the cross-validation and the BIC yield the same PMSE. The stochastic restricted LASSO is followed by the LASSO estimator. The PMSE of the LASSO estimator obtained by cross-validation is 14.92% larger than that of the stochastic restricted LASSO estimator while this value is 1.74% for the PMSE obtained by BIC. The stochastic restricted LASSO estimator selects 43 variables when both cross-validation and BIC are used while the LASSO estimator selects 59 variables when cross-validation is used and 43 variables when BIC is used.

Figure 5.10 shows the LASSO and stochastic restricted LASSO estimates as

a function of $\|\boldsymbol{\beta}\|_1 \,/\, \max \|\boldsymbol{\beta}\|_1$ for the riboflavin data set for the values of λ. It can be seen from Figure 5.10 that the shrinkage structures of the LASSO and stochastic restricted LASSO estimators are different with respect to the path they follow and the points where the estimates become zero.

The riboflavin data set demonstrates that the stochastic restricted LASSO estimator is an alternative to the LASSO estimator in high-dimensional data set in having smaller PMSE and proper variables.

Overall, the stochastic restricted LASSO estimator is a competitive alternative to the ridge, LASSO and stochastic restricted ridge estimators with respect to shrinking the coefficients and variable selection for appropriately selected tuning parameter.

6. SIMULATION STUDIES

We investigated the performances of the GO, shift-ridge, shift-LASSO and stochastic restricted LASSO estimators by real data analysis in Chapter 5. In this chapter, we perform simulation studies to exhibit the performance of these estimators. To achieve this goal, we generate data sets consisting of training, validation and test sets. The training set is used for estimating the coefficients and the validation set is used for estimating the tuning parameters. The model performance is measured by the median of mean squared errors (MSE) where MSE values are computed as

$$\text{MSE} = \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)^{\top} \mathbf{V} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) \tag{6.1.}$$

on the test set, where $\mathbf{V}$ is the covariance matrix of $\mathbf{X}$ and $\widehat{\boldsymbol{\beta}}$ is the concerned estimator. $(\cdot,\cdot,\cdot)$ notation indicates the number of observations in the training, validation and test sets, respectively. Moreover, standard error estimates for each estimated test MSE are obtained by the bootstrap method with 1000 repetitions.

The correct models for the simulation studies are as given by Equation (2.1.). We report the signal-to-noise ratio (SNR) values for all the simulations, which is defined for the linear regression as

$$\text{SNR} = \frac{\text{Var}\left(\mathbf{X}\boldsymbol{\beta}\right)}{\sigma^2}$$

where $\boldsymbol{\beta}$ and σ^2 are estimated by the concerned estimator. Small values of SNR imply that the data set is sparse, while large values of SNR imply that the data set is dense (Hastie et al., 2009).

6.1. Simulation Studies for the GO Estimator

The GO estimator was introduced in Section 4.1. We set simulation studies to compare the OLS, ridge, LASSO, elastic net, OK and GO estimators in this section. We choose 100 λ values on the log scale for a grid of $\{\alpha, d\}$ pairs to fit the models in the training set, and use the validation set to choose the optimal tuning parameters. Then, the MSE values are computed on the test set according to Equation (6.1.) by using the optimal tuning parameters. The first two simulation studies are basically from Tibshirani (1996) and the third one is from Zou and Hastie (2005). We repeat the simulations for four σ values that are 3,5,15,25 and four ρ values that are $0.5, 0.7, 0.9, 0.99$ where ρ is used to evaluate the degree of correlation.

(i) Simulation Study A1

In this simulation study, 100 data sets consisting of $(20, 20, 200)$ observations and 8 explanatory variables are generated. The true coefficient vector is $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$. In this setting, we intend to compare the methods when some of the explanatory variables do not relate to the response. We repeat the simulation studies so that the correlation between the explanatory variables $\mathbf{x}_i$ and $\mathbf{x}_j$ is selected to be $\rho_{ij} = \rho^{|i-j|}$ where $\rho = 0.5, 0.7, 0.9, 0.99$ and $\sigma = 3, 5, 15, 25$.

(ii) Simulation Study A2

In this simulation study, there is no difference in terms of settings with Simulation Study A1 other than the selected true parameter vector $\boldsymbol{\beta}$. This $\boldsymbol{\beta}$ is selected to be as $\boldsymbol{\beta} = (0.85, 0.85, \ldots, 0.85)^\top$. The aim of this setting is to explore if a method has superiority over the other method when all the explanatory variables are related to the response variable.

(iii) Simulation Study A3

In this simulation study, 100 data sets consisting of $(50, 50, 400)$ observations and 40 explanatory variables are generated. $\boldsymbol{\beta}$ and σ parameters are set to

$$\boldsymbol{\beta} = \left(\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25} \right)^{\top}$$

and $\sigma = 3, 5, 15, 25$. The explanatory variables $\mathbf{x}_i$ are formed as follows:

$\mathbf{x}_i = Z_1 + \varepsilon_i,\ Z_1 \sim \mathcal{N}(0,1),\ i = 1, 2, \ldots, 5,$

$\mathbf{x}_i = Z_2 + \varepsilon_i, Z_2 \sim \mathcal{N}(0,1), i = 6, 7, \ldots, 10,$

$\mathbf{x}_i = Z_3 + \varepsilon_i, Z_3 \sim \mathcal{N}(0,1), i = 11, 12, \ldots, 15,$

$\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 16, 17, \ldots, 40$ where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, 0.01)$, $i = 1, 2, \ldots, 15$.

This data set exhibits the grouping effect.

Table 6.1 through Table 6.9 show the simulation results. We report median test MSE values, expected standard error of the median test MSE, reduction rate gained by using the GO estimator and active set size for the estimators in each table. The relative efficiency is computed as

$$RR = \frac{\text{MSE}\left(\widehat{\boldsymbol{\beta}}\right) - \text{MSE}\left(\widehat{\boldsymbol{\beta}}^{GO}\right)}{\text{MSE}\left(\widehat{\boldsymbol{\beta}}\right)}$$

where $\widehat{\boldsymbol{\beta}}$ is any estimator to be compared with the GO estimator and $\text{MSE}(\cdot)$ is the estimated test MSE of the relevant estimator. The aim of using this quantity is to show the relative performance of the GO estimator. We also report active set sizes which represent the number of nonzero coefficient estimates obtained by a method.

Table 6.1. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A1 with $\rho = 0.5$ and $\sigma = 3, 5, 15, 25$

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
2.4086	3	Median MSE	6.2397	3.6591	3.2466	3.0897	2.6849	2.4119
		Expected SE	0.4469	0.2397	0.2204	0.2043	0.2374	0.2462
		Reduce Rate (%)	61	34	26	22	10	-
		Active Set Size	8	8	8	5	6	6
0.8671	5	Median MSE	17.3326	7.6809	6.1133	7.3324	6.2771	5.1883
		Expected SE	1.2414	0.9775	0.4202	0.4926	0.5718	0.4646
		Reduce Rate (%)	70	32	15	29	17	-
		Active Set Size	8	8	8	4	6	6
0.0963	15	Median MSE	155.9933	19.2472	15.7913	20.3145	18.3442	13.7951
		Expected SE	11.1726	0.7082	1.3093	0.6718	1.3296	1.0754
		Reduce Rate (%)	91	28	13	32	25	-
		Active Set Size	8	8	8	1	7.5	5
0.0347	25	Median MSE	433.3147	20.8933	19.3861	21.6616	20.0229	16.6611
		Expected SE	31.0350	0.3558	0.8733	0.4987	0.4789	1.0462
		Reduce Rate (%)	96	20	14	23	17	-
		Active Set Size	8	8	8	1	5	4

Table 6.2. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A1 with $\rho = 0.7$ and $\sigma = 3, 5, 15, 25$

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
3.0232	3	Median MSE	6.2397	2.8349	2.6518	2.6451	2.2751	1.9305
		Expected SE	0.4469	0.3008	0.1929	0.1637	0.2553	0.2367
		Reduce Rate (%)	69	32	27	27	15	-
		Active Set Size	8	8	8	5	6	6
1.0883	5	Median MSE	17.3326	5.5443	4.6958	6.3407	4.7104	3.9345
		Expected SE	1.2414	0.9943	0.5308	0.6649	0.687	0.3197
		Reduce Rate (%)	77	29	16	38	16	-
		Active Set Size	8	8	8	4	6	6
0.1209	15	Median MSE	155.9933	21.4887	14.7010	22.6971	15.6023	12.9109
		Expected SE	11.1726	2.2702	1.6676	1.9088	2.2376	1.4152
		Reduce Rate (%)	92	40	12	43	17	-
		Active Set Size	8	8	8	2	5	5
0.0435	25	Median MSE	433.3148	25.3838	22.4764	26.3236	23.9441	15.7888
		Expected SE	31.0350	0.6502	1.4724	0.4781	0.9037	1.4407
		Reduce Rate (%)	96	38	30	40	34	-
		Active Set Size	8	8	8	1	7	4.5

Table 6.3. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A1 with $\rho = 0.9$ and $\sigma = 3, 5, 15, 25$

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
4.0327	3	Median MSE	6.2397	1.7804	1.7218	2.0418	1.3544	1.2371
		Expected SE	0.4469	0.1055	0.1330	0.1556	0.1206	0.1089
		Reduce Rate (%)	80	31	28	39	9	-
		Active Set Size	8	8	8	4	7	6
1.4518	5	Median MSE	17.3326	3.3666	2.9936	3.9421	2.6771	2.3439
		Expected SE	1.2414	0.3862	0.2288	0.4057	0.2873	0.1832
		Reduce Rate (%)	86	30	22	41	12	-
		Active Set Size	8	8	8	3	6	5.5
0.1613	15	Median MSE	155.9933	16.4650	9.8937	17.7329	12.7782	7.2925
		Expected SE	11.1726	1.9642	1.4736	2.1569	2.7218	0.8456
		Reduce Rate (%)	95	56	26	59	43	-
		Active Set Size	8	8	8	2	6	4
0.0581	25	Median MSE	433.3148	32.1639	20.2757	32.6794	26.1888	13.2347
		Expected SE	31.0350	2.4084	2.4768	2.353	4.9652	2.628
		Reduce Rate (%)	97	59	35	60	49	-
		Active Set Size	8	8	8	1	4	4

Table 6.4. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A1 with $\rho = 0.99$ and $\sigma = 3, 5, 15, 25$

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
4.6379	3	Median MSE	6.2397	0.8113	0.6814	1.0152	0.6976	0.5444
		Expected SE	0.4469	0.092	0.0705	0.1044	0.0803	0.0563
		Reduce Rate (%)	91	33	20	46	22	-
		Active Set Size	8	8	8	3	8	5
1.6696	5	Median MSE	17.3326	1.6475	1.1933	2.1464	1.3026	0.8530
		Expected SE	1.2414	0.2011	0.2096	0.1813	0.1975	0.1054
		Reduce Rate (%)	95	48	29	60	35	-
		Active Set Size	8	8	8	2	7	4
0.1855	15	Median MSE	155.9933	10.8158	6.2407	12.1466	7.9841	3.1547
		Expected SE	11.1726	2.2583	1.3769	2.1924	1.7404	0.903
		Reduce Rate (%)	98	71	49	74	60	-
		Active Set Size	8	8	8	1	6	3
0.0668	25	Median MSE	433.3148	28.8275	13.9926	33.2426	22.5054	5.0638
		Expected SE	31.035	4.7474	3.9549	4.4507	4.7887	2.4369
		Reduce Rate (%)	99	82	64	85	77	-
		Active Set Size	8	8	8	1	5.5	3

Table 6.5. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A2 with $\rho = 0.5$ and $\sigma = 3, 5, 15, 25$

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
1.6233	3	Median MSE	6.2397	2.0884	2.0423	3.7518	2.0040	1.8671
		Expected SE	0.4469	0.1809	0.1432	0.2416	0.144	0.1344
		Reduce Rate (%)	70	11	9	50	7	-
		Active Set Size	8	8	8	6	8	8
0.5844	5	Median MSE	17.3326	4.0936	3.8814	7.0729	3.8713	3.5339
		Expected SE	1.2414	0.4165	0.3036	0.5037	0.3285	0.2246
		Reduce Rate (%)	80	14	9	50	9	-
		Active Set Size	8	8	8	4	8	8
0.0649	15	Median MSE	155.9933	13.8994	11.2536	13.8551	13.2417	9.1454
		Expected SE	11.1726	0.4934	1.1358	0.281	0.7261	0.833
		Reduce Rate (%)	94	34	19	34	31	-
		Active Set Size	8	8	8	2	8	6
0.0234	25	Median MSE	433.3147	14.3371	13.9100	14.4432	13.8551	10.7135
		Expected SE	31.0350	0.3148	0.4726	0.3477	0.2855	0.9413
		Reduce Rate (%)	98	25	23	26	23	-
		Active Set Size	8	8	8	1	7	4

Table 6.6. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A2 with $\rho = 0.7$ and $\sigma = 3, 5, 15, 25$

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
2.4761	3	Median MSE	6.2397	1.4869	1.4216	3.2160	1.4162	1.3323
		Expected SE	0.4469	0.1691	0.159	0.1946	0.1705	0.1591
		Reduce Rate (%)	79	10	6	59	6	-
		Active Set Size	8	8	8	6	8	8
0.8914	5	Median MSE	17.3326	3.4158	3.2862	5.6981	3.1250	2.8514
		Expected SE	1.2414	0.3796	0.39	0.668	0.3313	0.3524
		Reduce Rate (%)	84	17	13	50	9	-
		Active Set Size	8	8	8	4	8	8
0.0990	15	Median MSE	155.9933	17.9938	11.1979	19.2029	13.6782	8.4004
		Expected SE	11.1726	1.8989	1.749	1.4821	1.8439	0.9231
		Reduce Rate (%)	95	53	25	56	39	-
		Active Set Size	8	8	8	2	7	6
0.03566	25	Median MSE	433.3148	21.6056	17.2387	21.8910	19.2620	11.0754
		Expected SE	31.035	0.5638	1.9535	0.5583	1.2376	1.3216
		Reduce Rate (%)	97	49	36	49	43	-
		Active Set Size	8	8	8	1	7	4

Table 6.7. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A2 with $\rho = 0.9$ and $\sigma = 3, 5, 15, 25$

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
3.9796	3	Median MSE	6.2397	1.0380	0.9502	2.0092	0.9658	0.8632
		Expected SE	0.4469	0.0957	0.0944	0.2156	0.1048	0.1155
		Reduce Rate (%)	86	17	9	57	11	-
		Active Set Size	8	8	8	5	8	8
1.4326	5	Median MSE	17.3326	2.3148	2.1148	3.9345	2.2952	1.8010
		Expected SE	1.2414	0.2488	0.2472	0.5321	0.2028	0.2709
		Reduce Rate (%)	90	22	15	54	22	-
		Active Set Size	8	8	8	4	8	8
0.1592	15	Median MSE	155.9933	13.6949	8.1950	17.6718	11.2320	6.0673
		Expected SE	11.1726	2.4971	1.435	1.9424	1.6909	0.9902
		Reduce Rate (%)	96	56	26	66	46	-
		Active Set Size	8	8	8	2	7	6
0.0573	25	Median MSE	433.3148	30.7133	17.6941	33.3901	22.8443	10.1068
		Expected SE	31.0350	3.7427	3.3003	2.3025	4.2352	2.5955
		Reduce Rate (%)	98	67	43	70	56	-
		Active Set Size	8	8	8	1	6	5

Table 6.8. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A2 with $\rho = 0.99$ and $\sigma = 3, 5, 15, 25$

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
5.0096	3	Median MSE	6.2397	0.4791	0.3848	1.0229	0.4287	0.3159
		Expected SE	0.4469	0.0748	0.0731	0.1160	0.0697	0.0443
		Reduce Rate (%)	95	34	18	69	26	-
		Active Set Size	8	8	8	3	8	8
1.8035	5	Median MSE	17.3326	1.2443	0.9243	1.9024	0.9599	0.6123
		Expected SE	1.2414	0.2235	0.1656	0.1913	0.1657	0.077
		Reduce Rate (%)	96	51	34	68	36	-
		Active Set Size	8	8	8	2	8	5.5
0.2004	15	Median MSE	155.9933	10.1615	5.2711	13.6084	7.7666	1.9951
		Expected SE	11.1726	2.1075	1.4264	2.5371	1.8601	0.9832
		Reduce Rate (%)	99	80	62	85	74	-
		Active Set Size	8	8	8	1	6	4
0.0721	25	Median MSE	433.3148	27.7260	13.8746	32.0113	22.6510	4.6229
		Expected SE	31.035	5.0445	3.7601	4.5271	4.7876	2.7054
		Reduce Rate (%)	99	83	67	86	80	-
		Active Set Size	8	8	8	1	5	3

Table 6.9. Measurements of the quality for the OLS, ridge, OK, LASSO, elastic net and GO estimators from Simulation Study A3

SNR	σ	**Measurements**	**OLS**	**ridge**	**OK**	**LASSO**	**enet**	**GO**
72.2968	3	Median MSE	33.7870	7.6657	7.5958	3.5398	2.2854	2.2188
		Expected SE	1.7986	0.4642	0.3776	0.2725	0.236	0.2169
		Reduce Rate (%)	93	71	71	37	3	-
		Active Set Size	40	40	40	17	22	22
26.0269	5	Median MSE	93.8526	16.8531	16.4937	7.3685	5.9541	5.0700
		Expected SE	4.9899	0.8381	0.7856	0.7131	0.6297	0.5281
		Reduce Rate (%)	95	70	69	31	15	-
		Active Set Size	40	40	40	15	21	19
2.8919	15	Median MSE	844.6634	64.1817	63.8736	50.8868	45.9231	25.7057
		Expected SE	44.8514	4.1977	3.9976	6.6323	6.1655	2.5761
		Reduce Rate (%)	97	60	60	49	44	-
		Active Set Size	40	40	40	11	16.5	9
1.0411	25	Median MSE	2346.2817	118.7550	116.3064	130.2332	107.9384	50.9381
		Expected SE	124.556	8.5027	8.1876	15.13	9.5843	5.9959
		Reduce Rate (%)	98	57	56	61	53	-
		Active Set Size	40	40	40	10	26.5	9

We report the results of all the simulation studies as follows:

As a general statement, the GO estimator dominates the other estimators with respect to the test MSE median in all of the simulation studies for all the values of ρ and σ. It competes with the other estimators in the sense of expected standard error of test MSE median and active set size (for model selection). We reported the performance results with respect to σ and ρ, separately for Simulation Study A1 and Simulation Study A2. The comments of Simulation Study A3 are given after the comments of Simulation Study A1 and Simulation Study A2.

Performance with respect to σ:

In Simulation Study A1, for fixed and small values of σ, the test MSE values of the estimators generally get smaller as the values of ρ increase. However, the reverse is correct for large σ values, yet the GO estimator is less influenced in this case. For very high values of ρ, the gain from the test MSE becomes very large. For large σ values, the reduction rates by the GO estimator in the test MSE median generally increase, as the values of ρ get large. The reductions are very large when both ρ and σ are large. The expected standard errors show a similar pattern in general. The LASSO estimator is the most parsimonious method in point of the model selection in all cases. The GO estimator usually selects less variables than the elastic net estimator for large values of ρ, this is the case in favor of the elastic net estimator for the small values of ρ.

In Simulation Study A2, where no true coefficient is zero, the elastic net and GO estimators do not perform variable selection for small σ values as it is expected. They exhibit similar behavior to the ridge estimator in this case. However, they eliminate some variables for the large values of σ. The LASSO estimator always carries out variable selection and, for the large values of σ, the LASSO estimator yields very close to the null model. The GO estimator yields more parsimonious models than the elastic net estimator in this simulation study for the large values of σ. The GO

estimator shows relatively better performance in test MSE in general, as ρ increases at large σ. In this study, the elastic net estimator dominates the GO estimator with regard to the variable selection. Besides, the reduction rates obtained by the GO estimator seems to be less compared to Simulation Study A1. Finally, the GO estimator generally produces well estimates of the estimated standard errors of the test MSE median.

Performance with respect to ρ:

In Simulation Study A1, the GO estimator dominates the other estimators in test MSE medians for all values of ρ. There is not a general pattern of the reduction rate; however, it is generally large when σ is large for fixed values of ρ. As the values of σ increase, the active set sizes of the LASSO is far less than the true active set size, yet the elastic net and GO estimators yield close to the true active set size. This indicates that LASSO over-shrinks the coefficients for large values of σ; however, the elastic net and GO estimators are more robust estimators in this case and yield coefficient estimates with active set size very close to the true size.

In Simulation Study A2, the estimated test MSE medians and standard error estimates are smaller than the values obtained in Simulation Study A1 depending on the new true coefficient vector. For large ρ values, the performance of the LASSO estimator gets worse as σ increases. This result is consistent with the claim that the ridge estimator can dominate the LASSO estimator when $n > p$ if the data set has multicollinearity as Tibshirani (1996) mentioned. In general, the comparison results of the estimators are very similar to that of Simulation Study A1. We also give similar comments for the active set size to Simulation Study A1, except that the estimated active set sizes are large for the elastic net and GO estimators because of the different specification of the true coefficient vector $\boldsymbol{\beta}$.

Results of Simulation Study A3

In Simulation Study A3, the GO estimator yields the smallest test MSE medians

and the reduction rate is very large compared to the other estimators. The expected standard errors of the test MSE median are the lowest for most of the cases. Furthermore, the elastic net and GO estimators picturize the true active set size especially for small values of σ. However, the active set size of the GO estimator for the large values of σ is close to the size obtained by the LASSO estimator. This situation can be explained by bias-variance trade-off: the GO estimator produces more biased estimates compared to the elastic net estimator as σ gets large to overcome the high variance.

6.2. Simulation Studies for the Shift-ridge and Shift-LASSO Estimators

The shift-ridge and shift-LASSO estimators were introduced in Section 4.2. In this section, simulation studies are given to examine the performance of the shift-ridge and shift-LASSO estimators compared to the OLS, ridge, LASSO, elastic net and GO estimators in correlated and high-dimensional data which are illustrated in Subsections 6.2.1 and 6.2.2, respectively.

To choose the tuning parameters, we select the grid of λ values as $\lambda = 2^{k-6}$, $k = 1, \ldots, 20$ in the same line with Park and Yoon (2011). Then, we pick a grid of α values and fit the models by the coordinate descent algorithm on the training set for all the values of λ corresponding to each value of α. We determine the optimal α and λ values as the minimum error on the validation set. Finally, we compute the MSE values on the test set according to Equation (6.1.). We repeat this process for each replication of the simulation study and report the median test MSE, expected standard error of median test MSE and the active set sizes.

6.2.1. Simulation Studies for the Shift-ridge and Shift-LASSO Estimators on Correlated Data

In this subsection, we set up four simulation experiments. In all these simulation settings, the explanatory variables are generated by considering a proper correlation. Simulation studies from B1 to B3 are studied by Tibshirani (1996) and Simulation Study B4 is studied by Zou and Hastie (2005). Simulation study 1 represents the correlated data and sparse situation, Simulation Study B2 represents the correlated data and non-sparse situation, Simulation Study B3 is set up to observe the change in sparsity although it carries the same idea with that of Simulation Study B1 and Simulation Study B4 is set up to see the grouping effect.

(i) Simulation Study B1

This simulation setting is the same as Simulation Study A1 in Section 6.1, except that we take $\rho = 0.5$ and the σ values are chosen as $\sigma = 3, 15$ in this setting. This gives SNR of 2.34 for $\sigma = 3$ and 0.09 for $\sigma = 15$.

(ii) Simulation Study B2

We repeat the Simulation Study B1. The only difference is the true coefficient vector:

$$\boldsymbol{\beta} = \left(\underbrace{0.85, 0.85, \ldots, 0.85}_{8} \right)^{\top}.$$

This gives SNR of 1.63 for $\sigma = 3$ and 0.07 for $\sigma = 15$.

(iii) Simulation Study B3

We simulate 100 data sets consisting of $(100, 100, 400)$ observations and 40

explanatory variables. The true coefficient vector is

$$\boldsymbol{\beta} = \left(\underbrace{0,\ldots,0}_{10}, \underbrace{2,\ldots,2}_{10}, \underbrace{0,\ldots,0}_{10}, \underbrace{2,\ldots,2}_{10} \right)^{\top}$$

and the pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is $\rho_{ij} = 0.5$. This gives SNR of 92.86 for $\sigma = 3$ and 3.71 for $\sigma = 15$.

(iv) Simulation Study B4

This simulation setting is the same as Simulation Study A3 in Section 6.1, except that we take the σ values as $\sigma = 3, 15$ in this setting. This gives SNR of 70.47 for $\sigma = 3$ and 2.82 for $\sigma = 15$.

We report the test MSE, expected standard error and active set sizes of the simulation studies in Tables 6.10 and 6.11. We also summarize the results in Figures 6.1 and 6.2.

The key conclusions obtained from Tables 6.10-6.11 and Figures 6.1-6.2 together are as follows:

(a) Considering the test MSE median results from the Simulation Studies B1 through B4, on the whole, the OLS estimator yields the worst results while the GO estimator outperforms the methods in the Simulation Studies B1 and B2 for both $\sigma = 3$ and $\sigma = 15$ and the shift-ridge estimator follows the GO estimator for these two simulation studies. The shift-ridge estimator outperforms the methods in the Simulation Study B3 for $\sigma = 15$ and Simulation Study B4 for $\sigma = 3$.
When correlation is present among the important explanatory variables (see Simulation Studies B1 and B3), the LASSO, elastic net and GO estimators have better prediction performance than the ridge and OLS estimators under the sparse

Table 6.10. Comparisons of the OLS, ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 on the simulated correlated and grouping effected data sets for $\sigma = 3$

$\sigma = 3$		Sim. study 1	Sim. study 2	Sim. study 3	Sim. study 4
OLS	Median MSE	5.7054	6.0411	6.1504	37.2065
	Expected SE	0.3915	0.5053	0.2450	2.2654
	Active Set Size	8	8	40	40
Ridge	Median MSE	3.4342	1.7480	4.3236	6.9909
	Expected SE	0.2217	0.1244	0.1390	0.3586
	Active Set Size	8	8	40	40
LASSO	Median MSE	3.0413	2.9104	3.5557	3.8722
	Expected SE	0.2682	0.1777	0.1853	0.1842
	Active Set Size	5	6	30	17
Elastic net	Median MSE	2.6246	1.6793	3.4421	2.4207
	Expected SE	0.1816	0.1205	0.1160	0.1686
	Active Set Size	6	8	32	23
GO	Median MSE	2.3704	1.5171	3.2958	2.3732
	Expected SE	0.2212	0.0784	0.1316	0.1538
	Active Set Size	6	8	32	22
Shift-ridge	Median MSE	2.5471	1.5880	3.3458	2.0929
	Expected SE	0.2165	0.1073	0.1447	0.1742
	Active Set Size	6.5	8	31	23
Shift-LASSO	Median MSE	2.5253	2.6117	3.4071	3.2104
	Expected SE	0.2156	0.1392	0.1583	0.1402
	Active Set Size	6	7	33	35

Table 6.11. Comparisons of the OLS, ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 on the simulated correlated and grouping effected data sets for $\sigma = 15$

$\sigma = 15$		Sim. study 1	Sim. study 2	Sim. study 3	Sim. study 4
OLS	Median MSE	142.6348	151.0279	153.7631	930.0687
	Expected SE	9.7876	12.6317	6.1251	57.1266
	Active Set Size	8	8	40	40
Ridge	Median MSE	18.9538	12.5863	23.9838	65.8343
	Expected SE	0.7879	0.5322	1.0200	2.4339
	Active Set Size	8	8	40	40
LASSO	Median MSE	21.0448	14.3109	46.0833	52.1598
	Expected SE	0.4925	0.4487	1.1627	3.1299
	Active Set Size	1	1	20	12
Elastic net	Median MSE	17.0069	11.0622	21.6600	47.1859
	Expected SE	0.8043	0.7579	0.7466	3.3418
	Active Set Size	6	7	40	22
GO	Median MSE	13.4039	8.3849	21.2358	26.9547
	Expected SE	0.9919	0.4591	0.7194	2.2215
	Active Set Size	5	6	40	11
Shift-ridge	Median MSE	16.8852	11.0351	20.1282	45.2120
	Expected SE	0.9351	0.7646	0.4852	3.5243
	Active Set Size	6.5	6	40	23
Shift-LASSO	Median MSE	17.4813	11.5643	43.6009	40.3600
	Expected SE	0.7228	0.6984	1.1359	3.5044
	Active Set Size	5	4	24.5	32

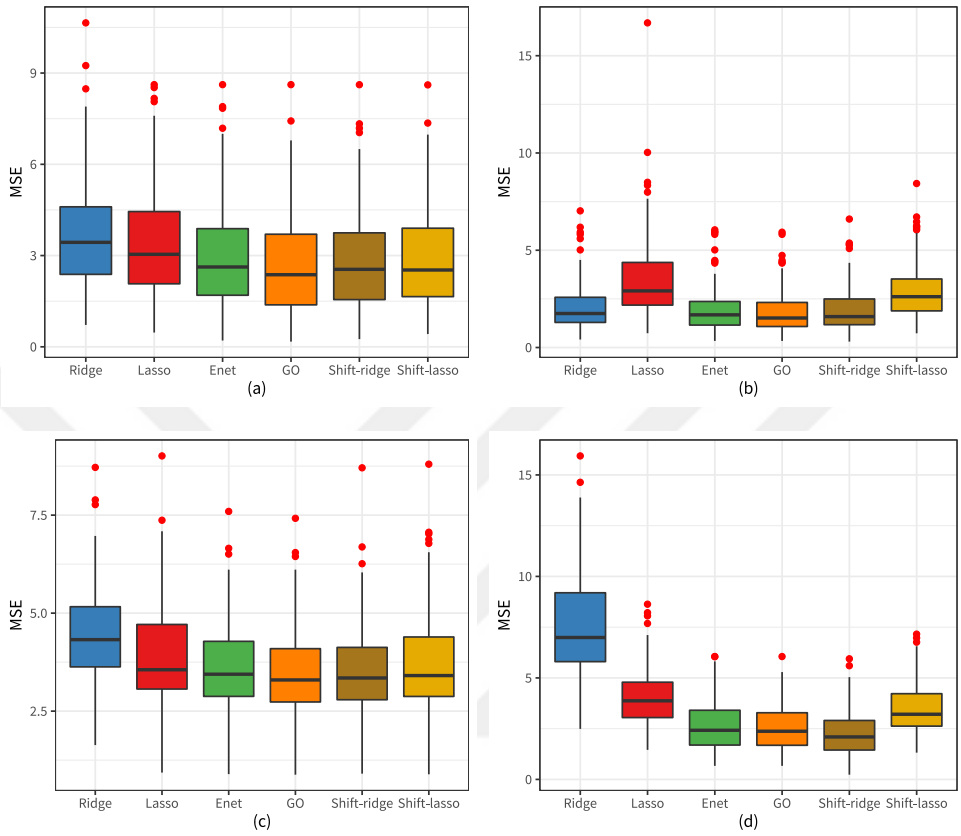

Figure 6.1. Box plots to compare the accuracy of the ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 for $\sigma = 3$

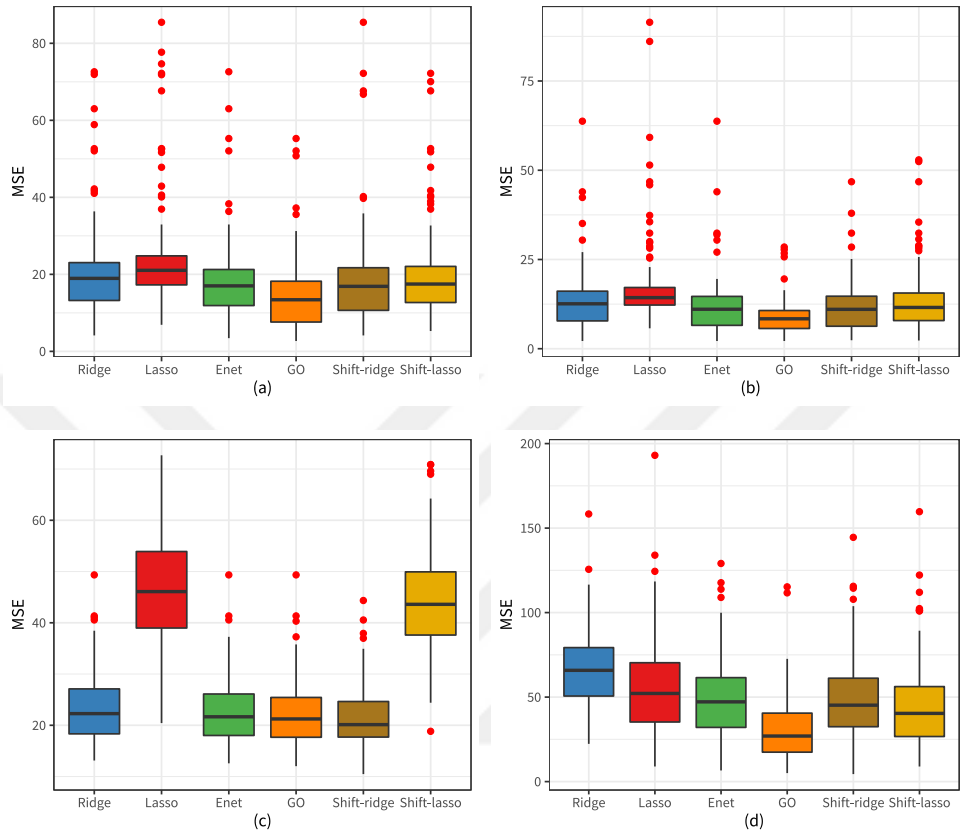

Figure 6.2. Box plots to compare the accuracy of the OLS, ridge, LASSO, elastic net, GO, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 for $\sigma = 15$

situation for $\sigma = 3$. However, the elastic net and GO estimators outperform the ridge estimator under the sparse situation for $\sigma = 15$. In addition to this accepted result, both the shift-ridge and shift-LASSO estimators perform better than the LASSO and elastic net estimators in case of correlated data associated with sparse situation for $\sigma = 3$ and for $\sigma = 15$, either the shift-ridge or shift-LASSO estimator performs better than the LASSO and elastic net estimators in the same case.

The GO estimator does better with correlated data and non-sparse situation (see Simulation Study B2). The shift-ridge estimator behaves almost identically to the GO estimator while the LASSO estimator only outperforms the OLS estimator and shift-LASSO estimator outperforms both the LASSO and OLS estimators. These results obtained for correlated data and non-sparse situation are true for both $\sigma = 3$ and $\sigma = 15$.

The GO, shift-ridge, shift-LASSO and elastic net estimators all have the grouping effect. The reason that the GO, shift-ridge, shift-LASSO and elastic net estimators are better than OLS, the ridge and LASSO estimators is this property. One of our proposed estimators, either the shift-ridge or shift-LASSO estimator dominates the elastic net estimator in case of grouping effect. This result justifies the shift-ridge and shift-LASSO estimators to be better than the existing estimators in the literature when group correlations exist in the data set.

(b) According to the box plots the range of all estimators changes as σ increases. While (a) and (e) are almost of the same structure, there are serious differences in (b) and (f). This shows the effect of σ on correlated data and sparsity situation. (c) and (d) are almost the same, but there are serious changes in the box plots of Simulation Studies B1, B3 and B4. This implies that σ does not have a serious

effect on correlated data and non-sparsity situation. Under the grouping effect, the ridge and LASSO have the worst results, while the other four estimators change considerably as σ increases.

(c) With regard to the active set sizes, the LASSO estimator tends to select less variables than the elastic net, shift-ridge and shift-LASSO estimators. Although the elastic net, GO and shift-ridge estimators give the true number of the active set sizes for $\sigma = 3$, it changes for $\sigma = 15$. σ has a serious effect on the active set size for Simulation Studies B1, B2 and B3; no positive or negative impact can be mentioned on this effect. However, the effect of σ is not visible on the elastic net, shift-ridge and shift-LASSO estimators in Simulation Study B4 which shows the grouping effect.

6.2.2. Simulation Studies for the Shift-ridge and Shift-LASSO Estimators on High-Dimensional Data

In this subsection, we modify the Simulation Studies from B1 to B4 given in Subsection 6.2.1 to the high-dimensional case, respectively. Our setups are as follows.

(i) Simulation Study B5

For this study, we simulate 100 data sets consisting of $(120, 200, 200)$ observations and 400 explanatory variables. We fill the first 49 components of $\boldsymbol{\beta}$ with random numbers from the uniform distribution within the range $[2, 5]$ and set the remaining coefficients to 0. The other settings are the same as Simulation Study B1. This gives SNR of 204.54 for $\sigma = 3$ and 7.85 for $\sigma = 15$.

(ii) Simulation Study B6

We conduct the Simulation Study B2 by establishing the true coefficient vector

$$\boldsymbol{\beta} = \left(\underbrace{0.85, 0.85, \ldots, 0.85}_{400} \right)^{\top}$$

with $(120, 200, 200)$ observations. This gives SNR of 96.82 for $\sigma = 3$ and 3.87 for $\sigma = 15$.

(iii) Simulation Study B7

We modify the Simulation Study B3 by setting the $\boldsymbol{\beta}$ vector to a 400×1 vector which is obtained as

$$\boldsymbol{\beta} = \left(\underbrace{0, \ldots, 0}_{50}, \underbrace{2, \ldots, 2}_{50}, \underbrace{0, \ldots, 0}_{50}, \underbrace{2, \ldots, 2}_{50}, \underbrace{0, \ldots, 0}_{50}, \underbrace{2, \ldots, 2}_{50}, \underbrace{0, \ldots, 0}_{50}, \underbrace{2, \ldots, 2}_{50} \right)^{\top}.$$

We generate the data sets with $(120, 200, 400)$ observations. This gives SNR of 8798.76 for $\sigma = 3$ and 351.95 for $\sigma = 15$.

(iv) Simulation Study B8

We simulate 100 data sets consisting of $(120, 200, 400)$ observations and set $\boldsymbol{\beta}$ to

$$\boldsymbol{\beta} = \left(\underbrace{3, \ldots, 3}_{150}, \underbrace{0, \ldots, 0}_{250} \right)^{\top}.$$

The explanatory variables $\mathbf{x}_i$ are generated in the same line with Simulation

Table 6.12. Comparisons of the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 on the simulated high-dimensional data sets for $\sigma = 3$

$\sigma = 3$		Sim. study 1	Sim. study 2	Sim. study 3	Sim. study 4
Ridge	Median MSE	494.0665	230.7031	615.7105	78.0853
	Expected SE	11.1448	6.5261	11.4179	2.3013
	Active Set Size	400	400	400	400
LASSO	Median MSE	44.1903	562.3882	1194.9797	1385.7103
	Expected SE	3.1540	6.7719	17.4742	48.9770
	Active Set Size	102	116	117	23.5
Elastic net	Median MSE	43.1942	230.7031	615.7105	2.8165
	Expected SE	3.0655	6.5261	11.5808	0.1657
	Active Set Size	104	400	288	247
Shift-ridge	Median MSE	43.1942	227.9636	607.2032	1.9247
	Expected SE	2.9788	6.5403	11.2628	0.1314
	Active Set Size	104	400	288	204
Shift-LASSO	Median MSE	43.9896	542.1704	1155.5241	165.261
	Expected SE	2.9995	5.6567	20.8404	3.3364
	Active Set Size	180	277.5	163	237

Study B4 except each group of $\mathbf{x}_i$ consists of 50 variables in the first three groups and the last group contains 250 variables. This gives SNR of 7919.37 for $\sigma = 3$ and 316.77 for $\sigma = 15$.

We report the test MSE, expected standard error and active set sizes of the simulation studies in Tables 6.12 and 6.13. We also summarize the results by Figures 6.3 and 6.4.

The key conclusions obtained from Tables 6.12-6.13 and Figures 6.3-6.4 together are as follows: In Simulation Studies B5 through B8, there are no OLS and GO estimates since the data sets are high-dimensional.

(a) The shift-ridge estimator outperforms the methods in Simulation Studies B5,

Table 6.13. Comparisons of the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators in Simulation Studies B1-B4 on the simulated high-dimensional data sets for $\sigma = 15$

$\sigma = 15$		Sim. study 1	Sim. study 2	Sim. study 3	Sim. study 4
Ridge	Median MSE	619.0709	345.0024	753.0122	198.5243
	Expected SE	15.1278	6.3683	13.6969	4.5519
	Active Set Size	400	400	400	400
LASSO	Median MSE	413.9773	665.1324	1337.7361	1376.7863
	Expected SE	10.2094	9.3768	24.7810	49.5158
	Active Set Size	80	90	113	22
Elastic net	Median MSE	367.2449	345.0024	752.887	42.3865
	Expected SE	14.0804	6.4848	14.0706	3.1564
	Active Set Size	114	400	293	278.5
Shift-ridge	Median MSE	355.2210	341.9129	748.6040	39.6969
	Expected SE	14.0210	6.0769	13.6952	2.8023
	Active Set Size	115.5	400	293	259.5
Shift-LASSO	Median MSE	403.1605	607.8012	1272.3395	209.1053
	Expected SE	13.5947	9.1352	20.4040	4.5607
	Active Set Size	222	317	169	240

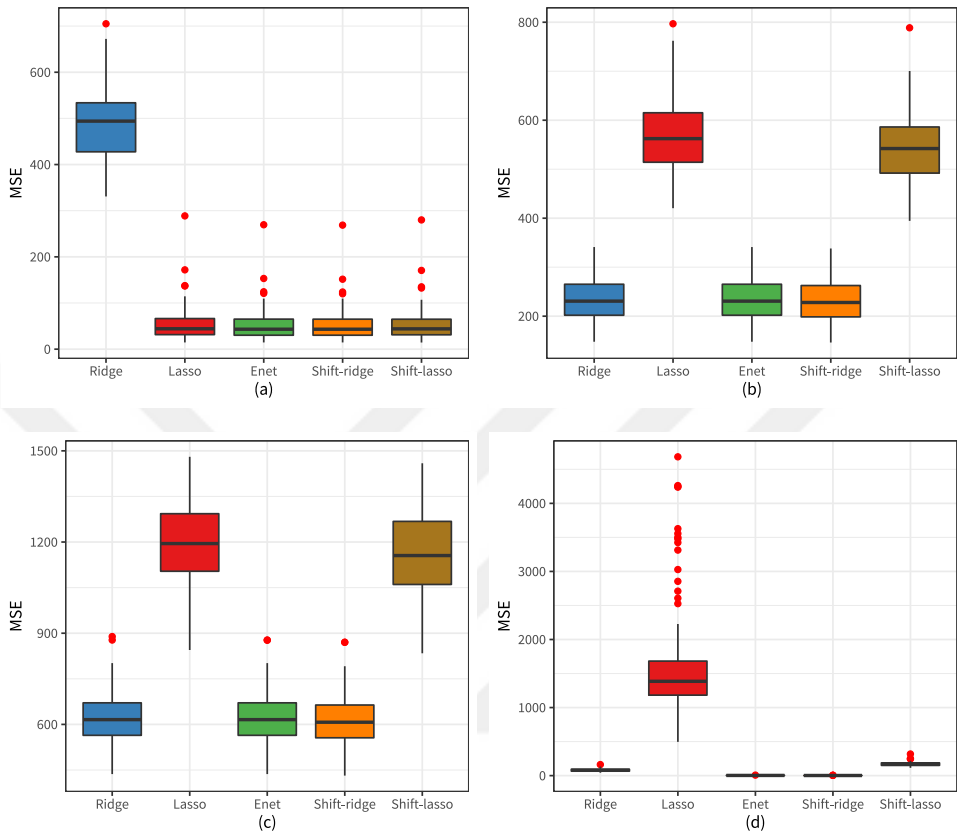

Figure 6.3. Box plots to compare the accuracy of the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators in Simulation Studies B5-B8 for $\sigma = 3$

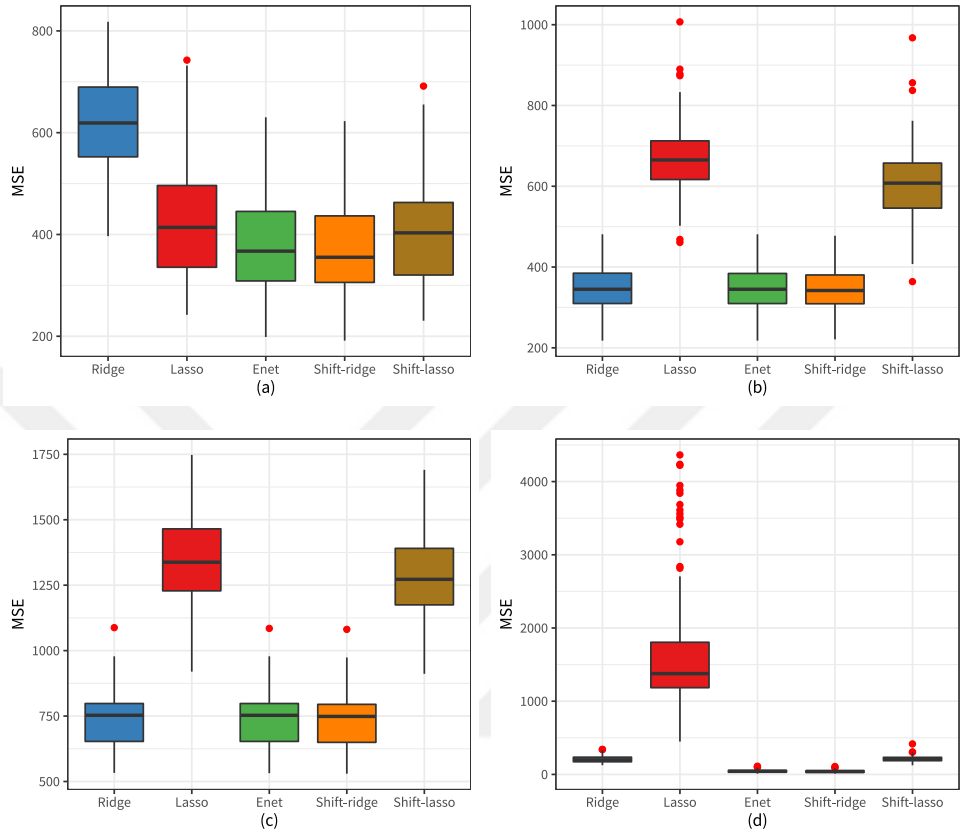

Figure 6.4. Box plots to compare the accuracy of the ridge, LASSO, elastic net, shift-ridge and shift-LASSO estimators in Simulation Studies B5-B8 for $\sigma = 15$

B6, B7 and B8 with regard to the test MSE. In general, the shift-ridge and shift-LASSO estimators have an important improvement over the ridge and LASSO estimators, respectively. The LASSO estimator performs worst in Simulation Studies B6, B7 and B8 while the ridge estimator is the worst in Simulation Study B5 in the sense of test MSE. In high-dimensional setting, the shift-LASSO estimator has a serious improvement over the LASSO estimator in grouped effected data whereas it is slightly better than the LASSO estimator in correlated data under sparse and non-sparse situations.

(b) The box plots show that there is not serious change between the results of $\sigma = 3$ and $\sigma = 15$. Although the most effected one is the Simulation Study B5, there is not a difference in the order of the superiority of the estimators. The shift-ridge estimator dominates the others which is followed by the elastic net estimator.

(c) The LASSO estimator is the most parsimonious among the estimators in Simulation Studies B5-B8. Although the elastic net and shift-ridge estimators do variable selection they give the true active set sizes in Simulation Study B6.

As a general statement, the shift-ridge or shift-LASSO estimators yield good results as compared to the other methods in most of the simulation settings with regard to the test MSE and active set sizes.

6.3. Simulation Studies for the Stochastic Restricted LASSO Estimator

We introduced the stochastic restricted LASSO estimator in Section 4.3. In this section, we compare the estimators OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators with simulation studies.

To choose the tuning parameters, we select the grid of λ values as $\lambda = 2^{k-14}$, $k = 0, \ldots, 28$ in a similar line with Park and Yoon (2011). Then, we fit the models

by the coordinate descent algorithm on the training set for all the values of λ. We determine the optimal λ value as the minimum error on the validation set. Finally, we compute the MSE values on the test set according to Equation (6.1.).

We determine the stochastic restrictions in the form of

$$\mathbf{h} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right).$$

We plug the stochastic restrictions in the training set to obtain the coefficient estimates of the stochastic restricted ridge and stochastic restricted LASSO estimators in each simulation study.

(i) Simulation Study C1

In this study, we generate 100 data sets in the same line with Simulation Study A1 in Section 6.1, except that $\rho = 0.5$ and $\sigma = 3$. For this setting, the restriction matrix and the corresponding restriction vector are

$$\mathbf{H} = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The SNR for this study is 2.34. The value of σ for this restriction is also chosen as 3.

(ii) Simulation Study C2

In this study, there is no difference in terms of settings with Simulation C1 with regard to the data generating process and standard deviation of the error term. The true parameter values are $\beta_j = 0.85,\ j = 1, 2, \ldots, 8$ for this setting. The

restriction matrix and the corresponding restriction vector are

$$\mathbf{H} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

with $\sigma = 2$. The SNR for this study is 1.64.

(iii) Simulation Study C3

There is no difference in terms of settings with Simulation C1 with regard to the data generating process in this study. The true parameter values are $\beta_1 = 5$, $\beta_j = 0$ for $j = 2, 3, \ldots, 8$ and $\sigma = 2$ for this setting. The restriction matrix and the corresponding restriction vector are

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{h} = 5.$$

The SNR for this study is 6.19.

(iv) Simulation Study C4

In this study, we generate 100 data sets consisting of $(50, 50, 400)$ observations and 40 explanatory variables. The true coefficient vector is $\boldsymbol{\beta} = 2 \cdot \begin{bmatrix} \mathbf{0}^\top & \mathbf{1}^\top & \mathbf{0}^\top & \mathbf{1}^\top \end{bmatrix}^\top$ where $\mathbf{0}$ and $\mathbf{1}$ are 10×1 vectors of zeros and ones

and $\sigma = 15$. The explanatory variables $\mathbf{x}_i$ are formed as $x_{ij} = z_{ij} + z_i$ where z_{ij} and z_i are independent standard normal variates. The restriction matrix and the corresponding restriction vector are

$$\mathbf{H} = \begin{bmatrix} \mathbf{0}^\top & \mathbf{1}^\top & \mathbf{0}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{1}^\top \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 20 \\ 20 \end{bmatrix}.$$

The SNR for this study is 7.59.

We report the median of MSE values, active set sizes, as well as the accuracy and F values defined in Equations 2.5.-2.6. for the estimators in Tables 6.14-6.17. However, we do not report F values for the methods which can not perform variable selection in Simulation Study C2 because it always equals to 1 for those methods due to the structure of the true coefficient vector. Also, we use the abbreviations SRR and SRL for stochastic restricted ridge and stochastic restricted LASSO, respectively in the tables and comments.

According to Tables 6.14-6.17, all the methods outperform the OLS estimator in terms of MSE. Also, SRR and SRL always yield better results than their non-stochastic counterparts with respect to the MSE criterion for all the simulation studies.

The results from Table 6.14 implies that the SRL estimator has the best performance in terms of MSE which is followed by the LASSO. The LASSO and SRL estimators have the same active set sizes and accuracy values, however, SRL has a larger F value than the LASSO estimator. In the sense of these measures, The SRL and LASSO estimators outperform the others. As a general statement, the SRL estimator

Table 6.14. Comparisons of the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Study C1

Measurements	OLS	Ridge	LASSO	SRR	SRL
MSE	5.1929	3.5411	3.0428	3.1066	2.8917
Active Set Size	8	8	6	8	6
Accuracy	0.38	0.38	0.62	0.38	0.62
F	0.55	0.55	0.63	0.55	0.67

Table 6.15. Comparisons of the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Study C2

Measurements	OLS	Ridge	LASSO	SRR	SRL
MSE	5.7956	1.9070	3.3797	1.8930	2.8879
Active Set Size	8	8	6	8	7
Accuracy	-	-	0.75	-	0.88
F	-	-	0.86	-	0.93

has the best performance in Simulation Study C1.

Table 6.15 indicates that the ridge type methods show better performance than the LASSO type methods in MSE. This is because of the structure of the true coefficient vector which has no zero. In this simulation study, the SRL estimator has a closer active set size to the true value and has higher accuracy and F values than LASSO.

In Simulation Study C3, only one coefficient is nonzero in the true coefficient vector. Therefore, we expect the methods that can perform variable selection should do better than the other methods. Table 6.16 supports this idea. The SRL estimator has the best performance of MSE and has the same pattern with the LASSO estimator in point of the variable selection measures.

Table 6.16. Comparisons of the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Study C3

Measurements	OLS	Ridge	LASSO	SRR	SRL
MSE	2.1761	1.9835	0.6544	1.8216	0.4591
Active Set Size	8	8	3	8	3
Accuracy	0.12	0.12	0.75	0.12	0.75
F	0.22	0.22	0.5	0.22	0.5

Table 6.17. Comparisons of the OLS, ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Study C4

Measurements	OLS	Ridge	LASSO	SRR	SRL
MSE	153.4421	35.4656	62.9076	31.5145	57.9153
Active Set Size	40	40	23	40	23
Accuracy	0.5	0.5	0.69	0.5	0.72
F	0.67	0.67	0.73	0.67	0.74

Table 6.17 implies that the ridge and SRR estimators have better performance over the other methods. However, the active set sizes of the LASSO and SRL estimators are very close to the true number of the active variables. Moreover, the larger values of accuracy and F measures of the SRL estimator mean that the SRL estimator grabs the true zero-nonzero pattern in the true coefficient vector better than the LASSO estimator.

The box plots of the MSE values for each simulation study are given in Figure 6.5. The box plots support the results in Tables 6.14-6.17. The SRL estimator has improvement over the LASSO estimator in both MSE values and the standard deviation of MSE values. The performances of all the methods differ depending on the simulation settings.

As a general statement, the results in Tables 6.14-6.17 and Figure 6.5 indicate that the SRL estimator outperforms the LASSO estimator and competes with the other methods with regard to the small MSE. Moreover, the estimator has either the same or

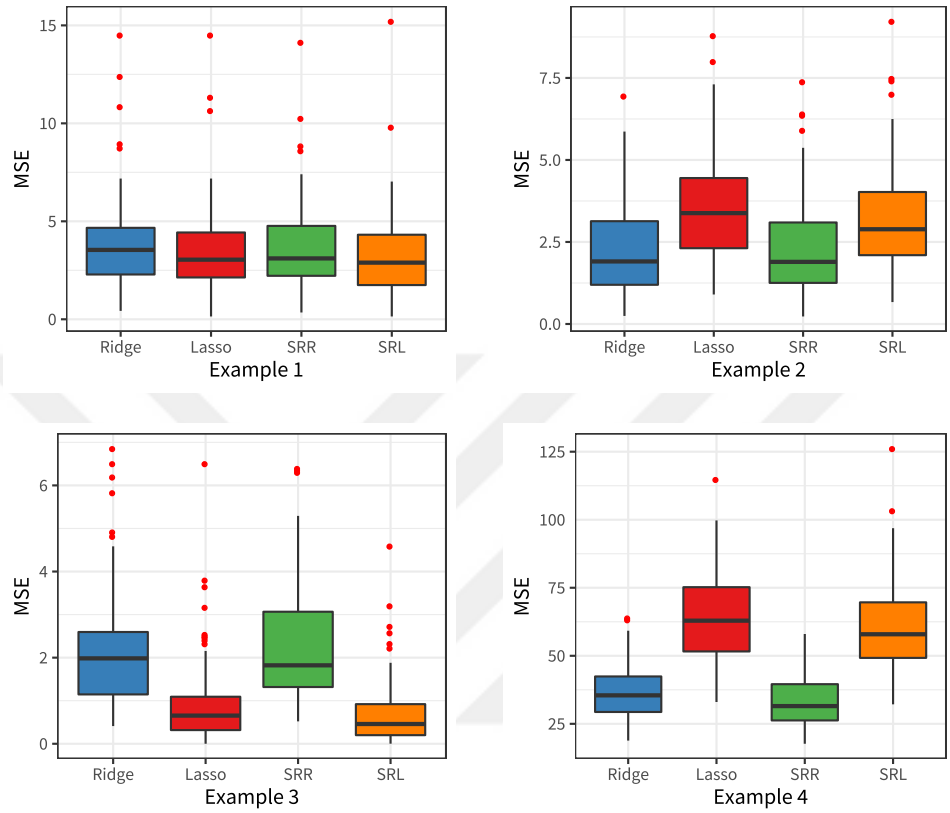

Figure 6.5. Box plots of MSE values of the ridge, LASSO, stochastic restricted ridge and stochastic restricted LASSO estimators for Simulation Studies C1-C4

better results in terms of accuracy and F which are criteria about the variable selection.

7. CONCLUSION

The subject of this study is to study the methods that can perform coefficient estimation and variable selection simultaneously in linear regression analysis.

In this study, we first gave the concept of penalized regression and brief explanations about some well-known penalized regression methods in the literature which are the ridge, LASSO, and elastic net estimators. In this context, we examined the strengths and weaknesses of these estimators and their properties for special data structures.

We proposed a new shrinkage and variable selection method, the GO estimator, in the linear regression, which is based on triple shrinkage on the regression coefficients. The GO estimator includes the ridge, LASSO and elastic net estimators as special cases. The term based on the shrunken estimator in the GO estimator can provide estimates with smaller length depending on the size of a new tuning parameter compared to the elastic net estimator preserving the variable selection feature in the case of multicollinearity. The GO estimator has the property of the grouping effect similar to that of the elastic net estimator. The well-known coordinate descent algorithm is used to compute the coefficient path of the GO estimator, efficiently.

We proposed the shift-ridge and shift-LASSO estimators by shrinking the coefficients towards a predetermined coefficient vector which represents the prior information. The estimators can be seen as the generalized version of the GO estimator to the high-dimensional data. The methods could result in smaller length estimates of the coefficients depending on the prior information compared to the elastic net estimator. We gave the coordinate descent and ADMM algorithms to obtain the coefficient estimates. We showed that the shift-ridge and shift-LASSO estimators have a similar grouping effect to that of the GO estimator when the explanatory variables

are highly correlated.

We also proposed the stochastic restricted LASSO estimator in linear regression model. The estimator is a generalization of the mixed estimator with ℓ_1 type penalization. We gave the coordinate descent algorithm to obtain the estimated coefficient vector of the method.

We conducted real data analysis and simulation studies by using the programs `R` and `C++` to compare the estimators with several methods in the literature for each estimator. The real data analysis and simulation studies showed that the new estimators enjoy the automatic variable selection property while has better performance in the sense of prediction mean squared error. Therefore the new estimators can be seen as alternative penalized regression methods to the methods in the literature depending on the structure of the data sets to deal with multicollinearity and high-dimensional data.

All the theoretical and practical results given in the thesis prove that the novel methods proposed in the thesis are very useful for high-dimensional and ill-conditioned data and deserve to be competitive alternatives to the penalized regression methods in the literature.

REFERENCES

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), Second International Symposium on Information Theory, (pp. 267-281). Academiai Kiado: Budapest.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19 (6): 716-723.

Belsley, D. A., Kuh, E., Welsch R. E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley , New York.

Bertsimas, D., King, A. and Mazumder, R., 2016. Best subset selection via a modern optimization lens. The Annals of Statistics 44(2), 813-852.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine learning, 3(1), 1-122.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge: Cambridge University Press.

Burnham, K. P., Anderson, D. R., 2002. Model Selection and Multimodel Inference: A practical information-theoretic approach (2nd ed.). Springer-Verlag.

Bühlmann, P., Kalisch, M., Meier, L., 2014. High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application, 1:1, 255-278.

Cawley, G. C., Talbot, N. L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul), 2079-2107.

Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., Braun, H.-J., 2010. Prediction of genetic values of quantitative traits in plant

breeding using pedigree and molecular markers. Genetics, 186(2), 713-724.

Donoho, D. L., Johnstone, J. M., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81(3), 425-455.

Edwards, J. B., 1969. The Relation between the F-Test and R^2. The American Statistician, 23(5), 28-28.

Efron, B., 1979. Bootstrap methods: another look at the jackknife, Annals of Statistics 7: 1-26.

Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science, 54-75.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of statistics, 32(2), 407-499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456), 1348-1360.

Fan, Y., Tang, C. Y., 2013. Tuning parameter selection in high dimensional penalized likelihood. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75, 3, pp. 531-552.

Frank, L. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools. Technometrics, 35(2), 109-135.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. The annals of applied statistics, 1(2), 302-332.

Genç, M., Özkale, M. R., 2020a. Regularization and variable selection with triple shrinkage in linear regression: A generalization of lasso. Submitted.

Genç, M., Özkale, M. R., 2020b. Usage of the GO estimator in high dimensional linear models. Computational Statistics, 1-23. doi: 10.1007/s00180-020-01001-2.

Genç, M., Özkale, M. R., 2020c. Stochastic restricted lasso in linear regression model. Submitted.

Giraud, C., 2015. Introduction to High-Dimensional Statistics. New York: Chapman and Hall/CRC.

Greene, W., 2000. Econometric Analysis. 4th edn, Prentice Hall, Upper Saddle River, NJ.

Groß, J., 2003. Restricted ridge estimation. Statistics & probability letters, 65(1), 57-64.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 5439, 531-537.

Gusnanto, A., Pawitan, Y., 2015. Sparse alternatives to ridge regression: a random effects approach. Journal of Applied Statistics, 42(1), 12-26.

Güler, H., Güler, E. O., 2019. Sparsely restricted penalized estimators. Communications in Statistics-Theory and Methods, 1-15.

Hastie, T. and Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.

Hastie T., Tibshirani R., Friedman J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition. New York City, USA: Springer.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC.

Henderson, H. V., Velleman, P. F., 1981. Building multiple regression models interactively. Biometrics, 37(2), 391-411.

Hiriart-Urruty, J. B., Lemaréchal, C., 2012. Fundamentals of convex analysis. Springer Science & Business Media.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 1, pp. 55-67.

Hoerl, A. E., Kannard, R. W., Baldwin, K. F., 1975. Ridge regression: Some simulations. Communications in Statistics-Theory and Methods, 4(2), 105-123.

Jagannathan, R., Ma, T., 2003. Risk reduction in large portfolios: why imposing the wrong constraints helps. Journal of Finance, 58:1651-1683.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning: with Applications in R. (Vol. 112, p. 18). New York: Springer.

Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. The Journal of Machine Learning Research, 15(1), 2869-2909.

Kaçıranlar, S., Sakallıoğlu, S., Özkale, M. R., Güler, H., 2011. More on the restricted ridge regression estimation. Journal of Statistical Computation and Simulation, 81(11), 1433-1448.

Karush, W., 1939. Minima of functions of several variables with inequalities as side constraints. M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago.

Kendall, M. G., 1957. A Course in Multivariate Analysis. Griffin, London, pp. 70-75.

Kibria, B. M. G., 2003. Performance of some new ridge regression estimators. Communications in Statistics-Simulation and Computation, 32(2), 419-435.

Kibria, B. M., Banik, S., 2016. Some ridge regression estimators and their performances. Journal of Modern Applied Statistical Methods, 15(1), 12.

Kock, A. B., Medeiros, M., Vasconcelos, G., 2020. Penalized Time Series Regression. In Macroeconomic Forecasting in the Era of Big Data (pp. 193-228). Springer, Cham.

Kuhn, H., Tucker, A. 1951. Nonlinear programming. Proceedings of 2nd Berkeley symposium (pp. 481-492). Berkeley: University of California Press.

Lawless, J. F., Wang, P., 1976. A simulation study of ridge and other regression estimators. Communications in Statistics-Theory and Methods, 5(4), 307-323.

Mallows, C. L., 1964. Choosing variables in a linear regression: A graphical aid. In Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kans.

Mallows, C. L., 1966. Choosing a subset regression. In the Joint Statistical Meetings, Los Angeles.

Mallows, C. L., 1973. Some comments on C_p. Technometrics, 15(4), 661-675.

Mallows, C. L., 1995. More comments on C_p. Technometrics, 37(4), 362-372.

Manning, C., Raghavan, P., Schütze, H., 2010. Introduction to information retrieval. Natural Language Engineering, 16(1), 100-103.

Marquardt, D. W., 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12(3), 591-612.

Mayer, L. S., Willke, T. A., 1973. On biased estimation in linear models. Technometrics, 15(3), 497-508.

Miller, A., 2002. Subset Selection in Regression. New York: Chapman and Hall/CRC.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2012. Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

Norouzirad, M., Arashi, M., 2018. Preliminary test and Stein-type shrinkage LASSO-based estimators. SORT-Statistics and Operations Research Transactions, 1(1), 45-58.

Norouzirad, M., Arashi, M., and Saleh, A., 2015. Restricted lasso and double shrinking. arXiv preprint arXiv:1505.02913.

Özkale, M. R., 2009. A stochastic restricted ridge regression estimator. Journal of Multivariate Analysis, 100(8):1706-1716.

Özkale, M. R., Kaçıranlar, S., 2007. The restricted and unrestricted two-parameter estimators. Communications in Statistics-Theory and Methods, 36(15), 2707-2725.

Park, C., Yoon, Y. J., 2011. Bridge regression: adaptivity and group selection. Journal of Statistical Planning and Inference, 141(11), 3506-3519.

Rao, C. R., Toutenburg, H., 1995. Linear Models: Least Squares and Alternatives. New York: Springer-Verlag.

Schwartz, G., 1978. Estimating the dimension of a model. Annals of Statistics. 6, 461-464.

Seber, G. A., 1977. Linear Regression Analysis. Wiley, New York.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., Yang, N., 1989. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. The Journal of Urology, 141(5), 1076-1083.

Theil, H. and Goldberger, A. S., 1961. On pure and mixed statistical estimation in economics. International Economic Review, 2(1):65-78.

Theobald, C. M., 1974. Generalizations of mean square error applied to ridge regression. Journal of the Royal Statistical Society: Series B (Methodological), 36(1), 103-106.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Tibshirani, R. J., 2013. The lasso problem and uniqueness. Electronic Journal of statistics, 7, 1456-1490.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 1, pp. 91-108.

Tibshirani, R. J., Taylor, J., 2012. Degrees of freedom in lasso problems. The Annals of Statistics, 40(2), 1198-1232.

Tseng, P., 2001. Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of optimization theory and applications, 109(3),

475-494.

Tuaç, Y. and Arslan, O., 2017. Variable selection in restricted linear regression models. arXiv preprint arXiv:1710.04105.

van der Hilst, R. D., De Hoop, M. V., Wang, P., Shim, S. H., Ma, P., Tenorio, L., 2007. Seismostratigraphy and thermal structure of Earth's core-mantle boundary region. Science, 315(5820), 1813-1817.

van Wieringen, W. N., 2015. Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169.

Wang, Y., Jiang, Y., Zhang, J., Chen, Z., Xie, B., Zhao, C., 2019. Robust variable selection based on the random quantile LASSO. Communications in Statistics-Simulation and Computation, 1-11.

Wright, S. J., 2015. Coordinate descent algorithms. Mathematical Programming, 151(1), 3-34.

Yang, Y., 2005. Can the strengths of AIC and BIC be shared?. Biometrika, 92: 937-950.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 1, pp. 49-67.

Zhang, C., Wu, Y., Zhu, M., 2019. Pruning variable selection ensembles. Statistical Analysis and Data Mining: The ASA Data Science Journal, 12(3), 168-184.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 2, pp. 301-320.

Zou, H., Hastie, T., Tibshirani, R., 2007. On the "degrees of freedom" of the lasso. The Annals of Statistics, 35(5), 2173-2192.

CURRICULUM VITAE

Murat GENÇ was born in Adana in 1984. He completed his elementary, secondary and high school education in Adana. He fulfilled his higher education in Department of Mathematics at Çukurova University as the second top student by June of 2005. He achieved his non-thesis master's degree at mathematics teaching at Çukurova University by August of 2006. He had served for his military service in Tasmus Company Command-Hasdal in İstanbul until September of 2007. After working as a civil servant in Ankara University and Provincal Administration in Adana by the November of 2012, he started to work at the Department of Statistics at Çukurova University as research assistant. He fulfilled his master degree education in Department of Statistics at Çukurova University by the August of 2014 and started his PhD education in the same department. He completed his PhD thesis with title "Ridge, Lasso and Elastic Net Methods in High Dimensional Data" in September of 2020.