2012

# Absolute Penalty and Shrinkage Estimation Strategies in Linear and Partially Linear Models

S.M. Enayetur Raheem
*University of Windsor*

# Absolute Penalty and Shrinkage Estimation Strategies in Linear and Partially Linear Models

by

S.M. Enayetur Raheem

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada

# Absolute Penalty and Shrinkage Estimation Strategies in Linear and Partially Linear Models

by

S.M. Enayetur Raheem

APPROVED BY

---

Dr. Peter XK Song, External Examiner
University of Michigan

---

Dr. A. Ngom
School of Computer Science

---

Dr. M. Hlynka
Department of Mathematics and Statistics

---

Dr. A. A. Hussein
Department of Mathematics and Statistics

---

Dr. S. E. Ahmed, Advisor
Department of Mathematics and Statistics

---

Dr. S. Johnson, Chair of Defense
Faculty of Graduate Studies

16 March 2012

# Declaration of Co-Authorship/ Previous Publication

### I. Co-Authorship Declaration

I hereby declare that this thesis incorporates the outcome of a joint research undertaken in collaboration with my supervisor, Professor S. Ejaz Ahmed. In all cases, the key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author, and the contribution of co-author was primarily through the provision of some theoretical results.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author to include in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

### II. Declaration of Previous Publication

This thesis includes two original papers that have been previously published, and another received invitation for submission.

| Thesis Chapter | Publication title/ full citation | Publication Status |
|---|---|---|
| Chapter 2 | Positive-shrinkage and pretest estimation in multiple regression: A Monte Carlo study with applications. *Journal of the Iranian Statistical Society*, 10(2):267-289, 2011 | Published |
| Chapter 3 | Absolute penalty and shrinkage estimation in partially linear models, *Computational Statistics & Data Analysis*, 56(4): 874-891, 2012 | Published |
| Chapter 2 | Shrinkage and Absolute Penalty Estimation in Linear Models. *WIREs Computational Statistics* | Preprint |

I certify that I have the rights to include the above published materials in my thesis. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained written permission from the copyright owner to include such material in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other university or institution.

# Abstract

In this dissertation we studied asymptotic properties of shrinkage estimators, and compared their performance with absolute penalty estimators (APE) in linear and partially linear models (PLM). A robust shrinkage M-estimator is proposed for PLM, and asymptotic properties are investigated, both analytically and through simulation studies.

In Chapter 2, we compared the performance of shrinkage and some APEs through prediction error criterion in a multiple linear regression setup. In particular, we compared shrinkage estimators with lasso, adaptive lasso and SCAD estimators. Monte Carlo studies were conducted to compare the estimators in two situations: when $p << n$, and when $p$ is large yet $p < n$. Examples using some real data sets are presented to illustrate the usefulness of the suggested methods.

In Chapter 3, we developed shrinkage estimators for a PLM. Efficient procedures for simultaneous sub-model selection and shrinkage estimation have been developed and implemented to obtain the parameter estimates where the nonparametric component is estimated using B-spline basis expansion. The proposed shrinkage estimator performed similarly to adaptive lasso estimators. In overall comparison, shrinkage estimators based on B-splines outperformed the lasso for moderate sample sizes and when the nuisance parameter space is large.

In Chapter 4, we proposed robust shrinkage M-estimators in a PLM with scaled residuals. Ahmed et al. (2006) considered such an M-estimator in a linear regression setup. We extended their work to a PLM.

*Dedicated to my parents.*

# Acknowledgements

All praises are for the Almighty who has given me the strength and ability to pursue for knowledge.

My sincere gratitude goes to my advisor Prof S. Ejaz Ahmed for his guidance which has lead to the completion of this dissertation. I am thankful to him for his support during my doctoral studies and for his mentorship without which it would not have been possible to complete the work in time.

Thanks are due to the external examiner Dr. Peter Song, and to the advisory committee members– Dr. Myron Hlynka, Dr. Abdul Hussein and Dr. Alioune Ngom for reviewing the dissertation and providing with valuable suggestions which have improved it greatly.

With this, I would like to extend my thanks to Dr. Sévérien Nkurunziza for his advices during my doctoral studies. Thanks are also due to Tanvir Quadir, Saber Fallahpour and Shabnam Chitsaz–for their excellent friendship during my studies at this university.

My parents and their expectations have been a constant source of inspirations throughout my life. No words of gratitude would be enough to acknowledge their contributions–I thank you for all your patience, support and prayers. Achievement comes with sacrifice–it is my family who has sacrificed the most. Despite many limitations, hardship, many tears and, at times, frustrations during the past several years, the love and encouragements from my wife Rifat Ara Jahan and my dear ones kept me on track. Special love and *adore* to my pearls–Tasfia and Eiliyah for giving me joyous company in my otherwise busy graduate-student-life.

S.M. Enayetur Raheem
May 15, 2012
Windsor, Ontario, Canada

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| ADB | asymptotic distributional bias |
| ADMSE | asymptotic distributional mean squared error |
| alasso | adaptive lasso |
| AIC | Akaike information criterion |
| APE | absolute penalty estimator/estimation |
| APEs | absolute penalty estimators |
| AQDB | asymptotic quadratic distributional bias |
| AQDR | asymptotic quadratic distributional risk |
| BIC | Bayesian information criterion |
| BSS | best subset selection |
| LAR | least angle regression |
| Lasso | least absolute shrinkage and selection operator |
| MSE | mean squared error |
| NSI | non-sample information |
| PT | pretest estimator |
| PLM | partially linear model |
| PLS | penalized least squares |
| PSE | positive shrinkage estimator |
| PSSE | positive-shrinkage semiparametric estimator |
| RE | restricted estimator |
| RM | restricted M-estimator |
| RMSE | relative mean squared error |
| SCAD | smoothly clipped absolute deviation |
| SE | shrinkage estimator |
| SM | shrinkage M-estimator |
| SM+ | positive-shrinkage M-estimator |
| SRE | semiparametric restricted estimator |
| SSE | semiparametric shrinkage estimator |
| UE | unrestricted estimator |
| UM | unrestricted M-estimator |
| UPI | uncertain prior information |

# List of Symbols

| | |
|---|---|
| $\boldsymbol{\beta}$ | regression parameter vector |
| $p$ | the number of regression parameters |
| $n$ | sample size |
| $H_0$ | null hypothesis |
| $\psi_n$ | test statistic |
| $\lambda$ | tuning parameter |
| $\hat{\boldsymbol{\beta}}^{\mathrm{UE}}$ | unrestricted estimator |
| $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}$ | restricted estimator |
| $\hat{\boldsymbol{\beta}}^{\mathrm{S}}$ | shrinkage estimator |
| $\hat{\boldsymbol{\beta}}^{\mathrm{S+}}$ | positive shrinkage estimator |
| $\hat{\boldsymbol{\beta}}^{\mathrm{PT}}$ | pretest estimator |
| $\hat{\boldsymbol{\beta}}^{\mathrm{UM}}$ | unrestricted M-estimator |
| $\hat{\boldsymbol{\beta}}^{\mathrm{RM}}$ | restricted M-estimator |
| $\hat{\boldsymbol{\beta}}^{\mathrm{SM}}$ | shrinkage M-estimator |
| $\hat{\boldsymbol{\beta}}^{\mathrm{SM+}}$ | positive-shrinkage M-estimator |
| $I(A)$ | indicator function |
| $\boldsymbol{W}$ | positive semi-definite weight matrix in the quadratic loss function |
| $\boldsymbol{\Gamma}$ | asymptotic distributional mean square error |
| $R(\cdot)$ | asymptotic distributional quadratic risk of an estimator |

| | |
|---|---|
| $K_n$ | local alternative hypothesis |
| $\boldsymbol{\omega}$ | a fixed real valued vector in $K_n$ |
| $\Delta$ | non-centrality parameter |
| $\Delta^*$ | a measure of the degree of deviation from the true model |
| $\boldsymbol{G}(y)$ | non-degenerate distribution function of $\boldsymbol{y}$ |

# Chapter 1

# Background

## 1.1 Introduction

Regression analysis is one of the most mature and widely applied branches in statistics. Least squares estimation and related procedures, mostly having a parametric flavor, have received considerable attention from theoretical as well as application perspectives. Statistical models, both linear and non-linear, are used to obtain information about unknown parameters. Whether such models fit the data well or whether the estimated parameters are of much use depends on the validity of certain assumptions. In practical situations, parameters are estimated based on sample information and, if available, other relevant information. The "other" information may be considered as *non-sample information* (NSI) (Ahmed, 2001). This is also known as *uncertain prior information* (UPI). The NSI may or may not positively contribute in the estimation procedure. Nevertheless, it may be advantageous to use the NSI in the estimation process when sample-information may be rather limited and may not

be completely trustworthy.

It is widely accepted that, in applied science, an experiment is often performed with some prior knowledge of the outcomes, or to confirm a hypothetical result, or to re-establish existing results. Suppose in a biological experiment, a researcher is focusing on estimating the growth rate parameter $\eta$ of a certain bacterium after applying some catalyst when it is suspected *a priori* that $\eta = \eta_0$, where $\eta_0$ is a specified value. In a controlled experiment, the ambient condition may not contribute to varying growth rate. Therefore, the biologist may have good reason to suspect that $\eta_0$ is the true growth rate parameter for her experiment, albeit unsure. This suspicion may come from previous studies or experience, and the researcher may utilize previously obtained information i.e., NSI, in the estimation of growth rate parameter.

It is however, important to note that the consequences of incorporating NSI depend on the quality or usefulness of the information being added in the estimation process. Based on the idea of Bancroft (1944), NSI may be validated through preliminary test, and depending on the validity, may be incorporated in the estimation process.

Later, Stein (1956) introduced shrinkage estimation. In this framework, the shrinkage estimator or Stein-type estimator takes a hybrid approach by shrinking the base estimator to a plausible alternative estimator utilizing the NSI.

Apart from Stein-type estimators, there are absolute penalty-type estimators, which are a class of estimators in the penalized least squares family. Such an estimator is commonly known as absolute penalty estimator (APE) since the absolute value of the penalty term is considered in the estimation process. These estimators provide simultaneous variable selection and shrinkage of the coefficients towards zero. Frank and Friedman (1993) introduced bridge regression, a generalized version of

APEs that includes ridge regression as a special case. An important member of the penalized least squares (PLS) family is the $L_1$ penalized least squares estimator or the *lasso* (Least Absolute Shrinkage and Selection Operator) which is due to Tibshirani (1996). Two other related APEs are adaptive lasso (alasso) which is due to Zou (2006), and the smoothly clipped absolute penalty (SCAD), due to Fan and Li (2001). APEs are frequently being used in variable selection and feature extraction problems, and problems involving low- and high-dimensional data. We define low- and high-dimensional later later in this chapter.

## 1.2 Statement of the Problem in this Study

Consider a scenario as follows. We have a set of covariates to fit a regression model to predict a response variable. If it is *a priori* known or suspected that a subset of the covariates do not significantly contribute in the overall prediction of the response variable, they may be left aside and a model without these covariates may be sufficient. In some situations, a subset of the covariates may be considered nuisance such that they are not of main interest, but they must be taken into account in estimating the coefficients of the remaining parameters. A candidate model for the data that involves only the important covariates in predicting the response is called the restricted model or sub-model, whereas the model that includes all the covariates is called the unrestricted model or simply the candidate full model.

To formulate the problem, consider the regression model of the form

$$y = f(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{E}, \tag{1.1}$$

where $y$ is the vector of responses, $\boldsymbol{X}$ is a fixed design matrix, $\boldsymbol{\theta}$ is an unknown vector of parameters, and $\boldsymbol{E}$ is the vector of unobservable random errors.

The shrinkage estimation method combines estimates from the candidate full model and a sub-model. Such an estimator outperforms the classical maximum likelihood estimator in terms of a quadratic risk function. In this framework, the estimates are essentially being shrunken towards the restricted estimators. A schematic flowchart of shrinkage estimation is presented in Figure 1.1.

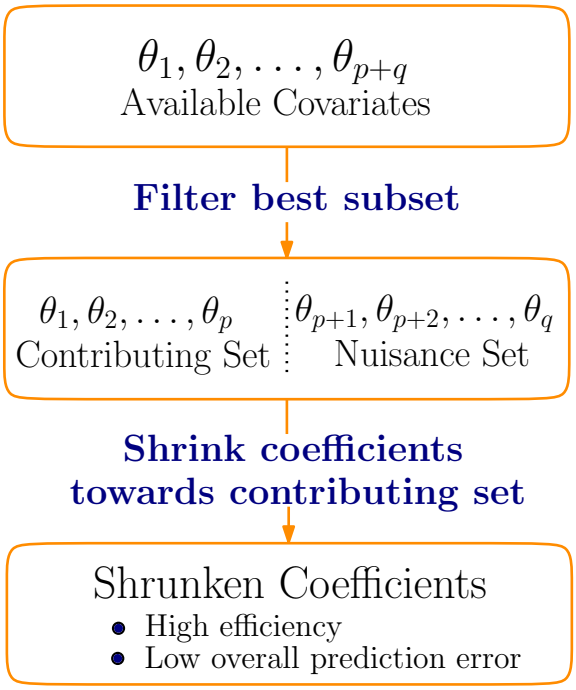


Figure 1.1: Flowchart of shrinkage estimation in multiple regression.

Suppose the dimension of $\boldsymbol{\theta}$ is $(p+q)$. Also suppose that $q$ of the covariates are considered as nuisance. Therefore, the parameter space and the design matrix $\boldsymbol{X}$ may be partitioned, and the model in (1.1) may be written as

$$y_i = f(\boldsymbol{X}_{il}, \boldsymbol{\theta}_p) + f(\boldsymbol{X}_{im}, \boldsymbol{\theta}_q) + E_i, \quad i = 1, 2, \ldots, n \tag{1.2}$$

where $\boldsymbol{X}_{il}$ includes the first $l$ column vectors of $\boldsymbol{X}_i$, and $\boldsymbol{X}_{im}$ contains the last $m$ column vectors of $\boldsymbol{X}_i$ with $l = 1, 2, \ldots, p$, $m = p+1, p+2, \ldots, p+q$.

Let us denote the full model estimators or unrestricted estimators (UE) by $\hat{\boldsymbol{\theta}}^{\mathrm{UE}}$, and the restricted estimators (RE) by $\hat{\boldsymbol{\theta}}^{\mathrm{RE}}$. The nuisance subset may be tested in the form of testing the hypothesis

$$H_0 : \boldsymbol{H}\boldsymbol{\theta} = \boldsymbol{h}$$

with an appropriate test statistic, $T_n$.

Now, the general forms of the UE and RE are, respectively,

$$\hat{\boldsymbol{\theta}}^{\mathrm{UE}} = g(\boldsymbol{X}, y)$$

and

$$\hat{\boldsymbol{\theta}}^{\mathrm{RE}} = \hat{\boldsymbol{\theta}}^{\mathrm{UE}} - g(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{h}, \boldsymbol{H}).$$

In the above expressions, $g(\boldsymbol{X}, y)$ denotes a function of the data and the response vector, and $g(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{h}, \boldsymbol{H})$ is a function of the data, the response vector, $\boldsymbol{H}$ and $\boldsymbol{h}$. Notice that the RE is a linear function of the UE. Now, we are in a position to introduce the shrinkage estimators. In the following, we define shrinkage, positive shrinkage, and pretest estimators.

Shrinkage estimator (SE), $\hat{\boldsymbol{\theta}}^{\mathrm{S}}$, is defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{S}} = \hat{\boldsymbol{\theta}}^{\mathrm{RE}} + (\hat{\boldsymbol{\theta}}^{\mathrm{UE}} - \hat{\boldsymbol{\theta}}^{\mathrm{RE}})(1 - cT_n^{-1}),$$

where $c$ is an optimum constant that minimizes the risk.

Positive shrinkage estimator (PSE), $\hat{\boldsymbol{\theta}}^{\mathrm{S+}}$, is defined by

$$\hat{\boldsymbol{\theta}}^{\mathrm{S+}} = \hat{\boldsymbol{\theta}}^{\mathrm{RE}} + (\hat{\boldsymbol{\theta}}^{\mathrm{UE}} - \hat{\boldsymbol{\theta}}^{\mathrm{RE}})(1 - cT_n^{-1})^+, \text{ where } s^+ = \max(0, s).$$

The pretest estimator (PT), $\hat{\boldsymbol{\theta}}^{\mathrm{PT}}$, is defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{PT}} = \hat{\boldsymbol{\theta}}^{\mathrm{UE}} - (\hat{\boldsymbol{\theta}}^{\mathrm{UE}} - \hat{\boldsymbol{\theta}}^{\mathrm{RE}})I(T_n < d_\alpha),$$

where $d_\alpha$ is the $100(1-\alpha)$ percentage point of the test statistic $T_n$.

Traditionally, statistical methods have been limited to low-dimensional settings, that is, when the number of experimental units ($n$) is much larger than the number of covariates ($p$). However, with the advent of high-performance computing and related technological advancements, new areas of applications have emerged where the number of experimental units is small compared to a high- or ultra high-dimensional parameter space. In this dissertation, we call a data set

(a) classical or low-dimensional, if $p$ is much smaller than $n$, e.g., $n = 100$, $p = 10$,

(b) high-dimensional, if $p$ is large yet smaller than $n$, e.g., $n = 100$, $p = 70$, and

(c) ultra high-dimensional, if $p$ is larger than $n$, e.g., $n = 100$ and $p = 150$.

Technically, as per the above definitions, shrinkage estimation works for low and high-dimension ($p < n$), but does not work for ultra high-dimensional data. On the other hand, APEs such as lasso, alasso and SCAD are suited for problems with low- or high- or ultra high-dimensional data.

It is worth exploring how shrinkage estimators perform in low and high-dimensional context compared to the APEs considered in this dissertation. We are interested to

study shrinkage, pretest and absolute penalty estimation in various contexts. We first consider a multiple linear regression model to compare shrinkage and APE. Then we extend shrinkage estimation to partially linear model (PLM). Finally, a new robust shrinkage M-estimator is proposed in PLM setup in the presence of a scaled error in the model.

Following section summarizes the relevant works available in the reviewed literature.

## 1.3 Review of Literature

In this section, we summarize some of the important works found in the reviewed literature pertaining to this dissertation. We group them into three subsections as follows.

### 1.3.1 Shrinkage, Pretest, and APE in Multiple Regression Models

Since the beginning, pretest and shrinkage estimation techniques have received considerable attention from the researchers. Asymptotic properties of shrinkage and preliminary test estimators using quadratic loss function, and their dominance over the usual maximum likelihood estimators have been demonstrated in numerous studies in the literature. Since 1987, Ahmed and his co-researchers are among others who have analytically demonstrated that shrinkage estimators outshine the classical estimator.

Ahmed (1997) gave a detailed description of shrinkage estimation, and discussed large sample estimation in a regression model with non-normal error terms. A review of shrinkage and some penalized estimators can be found in the work of van Houwelingen (2001). In their work, the James-Stein estimator, pretest estimator, ridge regression, lasso and the Garotte estimators are discussed in an attempt to put them in a semi-Bayesian framework.

An application of empirical Bayes shrinkage estimation can be found in Castner and Schirm (2003). They estimated the state wise numbers of people eligible for food stamps in the Food Stamp Program[1] in the U.S. "The shrinkage estimates derived are substantially more precise than direct sample estimates."(Castner and Schirm, 2003, page ix).

Khan and Ahmed (2003) considered the problem of estimating the coefficient vector of a classical regression model when it is *a priori* suspected that the parameter vector may belong to a subspace. They demonstrated analytically and numerically that the positive-part of Stein-type estimator and the improved preliminary test estimator dominate the usual Stein-type and pretest estimators, respectively. They also showed that positive-part of Stein-type estimator uniformly dominates the unrestricted estimator.

A review of APEs and their application in PLM can be found in Ahmed et al. (2010). There has been no study in the reviewed literature to compare the risk properties of shrinkage and APEs in the context of multiple linear regression models. In this dissertation, we compare shrinkage, pretest, and APEs in multiple linear regression models.

[1]The Food Stamp Program is the largest food and nutrition assistance program administered by the Food and Nutrition Service of Department of Agriculture, U.S.

### 1.3.2 Shrinkage Estimation in Partially Linear Models

A partially linear regression model (PLM) can be written as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + g(t_i) + \varepsilon_i, i = 1, \ldots, n, \tag{1.3}$$

where $y_i$'s are responses, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ and $t_i \in [0, 1]$ are design points, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown parameter vector and $g(\cdot)$ is an unknown bounded real-valued function defined on $[0, 1]$, $\varepsilon_i$'s are unobservable random errors.

When $\varepsilon_i$ are independent and identically distributed (i.i.d.) random variables, Heckman (1986), Rice (1986), Chen (1988), Speckman (1988), Robinson (1988), Chen and Shiau (1991), Chen and Shiau (1994), Donald and Newey (1994), Eubank and Speckman (1990), and Hamilton and Truong (1997) used various estimation methods, such as kernel method, spline method, series estimation, local linear estimation, two-stage estimation, to obtain estimators of the unknown quantities in (1.3). When errors are AR(1) process, Schick (1994) discussed the estimation of the autocorrelation coefficient. Schick (1994, 1996, 1998) further constructed efficient estimators for the regression coefficients and autocorrelation coefficients, respectively. Bhattacharya and Zhao (1997) proposed asymptotically efficient estimator of the regression parameters in a PLM under mild smoothness assumptions on the nonparametric component. A survey of the estimation and application of PLM in (1.3) can be found in the monograph of Härdle et al. (2000). For more recent work on the subject we refer to Wang et al. (2004), Xue et al. (2004), Liang et al. (2004), and Bunea (2004).

information (UPI) about the regression parameters is available have been considered by Ahmed and Saleh (1999).

Ahmed et al. (2007) considered a profile least squares approach based on using kernel estimates of $g(\cdot)$ to construct absolute penalty, shrinkage and pretest estimators of $\boldsymbol{\beta}$ in the case where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2)'$, where $\boldsymbol{\beta}_1$ is a vector of principle parameters and $\boldsymbol{\beta}_2$ is a vector of nuisance parameters. We extend their work to estimate $g(\cdot)$ using B-spline basis expansion, and obtain restricted, shrinkage and positive shrinkage estimators (Raheem et al., 2012).

### 1.3.3 Shrinkage M-estimation in Partially Linear Models

For partially linear regression models, Bianco and Boente (2004) proposed a family of robust estimates for the regression parameters. They studied their asymptotic properties and compared their performance with the classical estimators through simulation.

Ma and Kosorok (2005) proposed weighted M-estimators for semiparametric models in a situation when some of the parameters cannot be estimated at the root-n convergence rate. In the context of a generalized PLM, Boente et al. (2006) proposed a family of robust estimates for the parametric and nonparametric components. They showed that the regression parameters are root-n consistent and asymptotically normal. In a PLM with serially correlated errors, Bianco and Boente (2007) proposed a family of robust estimators for the autoregression parameter and the autoregression function.

For longitudinal data, He et al. (2002) studied robust M-estimation in a PLM. They considered a regression spline to approximate the nonparametric component. He et al. (2005) considered robust generalized estimating equations for generalized partially linear models for longitudinal data. They argued that the regression spline

approach overcomes some of the intricacies associated with the profile-kernel method, and used regression spline to estimate the nonparametric component.

Yu and Ruppert (2002) proposed penalized spline estimation in partially linear single-index models. Root-n consistency and asymptotic normality of the estimators of all parameters have been discussed assuming a fixed number of knots.

Cheng and Huang (2010) provided theoretical justifications for the use of bootstrap as a semiparametric inferential tool. In particular, they considered M-estimation in a semiparametric model that is characterized by a Euclidean parameter of interest and an infinite-dimensional nuisance parameter.

Ahmed et al. (2006) considered robust shrinkage-estimation of the slope parameters in the presence of nuisance scale parameter in linear regression setup. They studied asymptotic properties of variants of Stein-type M-estimators (including the positive-part shrinkage M-estimators). In this dissertation, we extend their work to obtain shrinkage M-estimators in a PLM.

## 1.4 Objective of the Study

There are three objectives of this dissertation.

Since the introduction of lasso in 1996 by Tibshirani, there has been a tremendous amount of development in lasso and related absolute penalty estimation techniques and their applications.

As a tool for simultaneous variable selection and parameter estimation, lasso shrinks the regression coefficients toward zero. Although shrinkage and APEs have been

around for quite some time, little work has been done to compare their relative performance. Ahmed et al. (2007) are the first to compare shrinkage and lasso estimates in the context of a PLM. We did not find any study in the reviewed literature that compares shrinkage and lasso in linear regression models. Therefore, in this study, we compare the performance of positive-shrinkage and lasso estimators in multiple linear regression setup based on the quadratic risk function. We show a real data example and compare their performance by calculating average prediction error through cross validation. Monte Carlo study will be conducted with low- and high-dimensional data to compare the predictive performance of shrinkage estimators with those of lasso, alasso, and SCAD estimators.

Secondly, we intend to develop shrinkage and positive-shrinkage estimators for a PLM. In particular, we wish to study the the suitability of B-spline basis function in estimating the nonparametric component in a PLM to obtain shrinkage estimates. Efficient procedure for simultaneous sub-model selection via shrinkage will be developed and implemented to obtain the parameter estimates after incorporating the B-Spline bases in the model. We will also compare the shrinkage estimation with some APEs.

Thirdly, robust shrinkage estimation will be studied in a PLM. We will consider a PLM with scaled residuals. Such a study have been considered by Ahmed et al. (2006) for linear regression only. We will extend their work to a PLM.

### 1.4.1 Organization of the Study

The dissertation is divided into five chapters. Chapter 1 introduces various shrinkage and pretest estimators. Three absolute penalty estimators, namely, the least absolute

penalty and shrinkage operator (lasso), adaptive lasso (alasso), and the smoothly clipped absolute deviation (SCAD) have been defined. The objective of the study and the highlights of important findings are summarized in this chapter.

We divided Chapter 2 into two parts. In the first part, application of shrinkage and pretest estimation have been demonstrated with three real data examples. The asymptotic properties of the estimators, which are well developed in the literature, have been studied through Monte Carlo experiment.

In the second part of Chapter 2, we compared shrinkage estimators with three absolute penalty estimators, such as, lasso, alasso, and SCAD in linear regression model. Both low-dimensional ($p << n$) and high-dimensional ($p < n$) cases have been considered.

In Chapter 3, we studied asymptotic properties of shrinkage and positive shrinkage estimators in a partially linear model when the nonparametric component is estimated by B-spline basis expansion. The asymptotic bias and risk expressions of the estimators have been derived. Algorithm for simultaneous sub-model selection and shrinkage estimation is presented. Performance of shrinkage estimators have been compared with lasso and alasso estimators using a popular econometric data set.

In Chapter 4 we proposed a robust shrinkage M-estimator of slope parameters in a partially linear regression model. Asymptotic bias and risk properties of the estimators have been studied–both analytically and numerically.

Finally, conclusions and an outline for future research are presented in Chapter 5.

## 1.5 Highlights of Contributions

In this dissertation, we consider shrinkage, pretest and absolute penalty estimation (APE) in linear and partially linear regression regression models (PLM). We also considered robust shrinkage M-estimation in a PLM. We demonstrate, with examples, the application of shrinkage, pretest, and absolute penalty estimation. Asymptotic risk properties of shrinkage and shrinkage-M-estimators have been studied, and analytic expressions for their bias and risks derived.

The highlights of our contributions in this dissertation are summarized below.

1. In Chapter 1 we layout the rationale for shrinkage estimation. The statement of the problem and the techniques of shrinkage estimation have been illustrated through a general regression framework. An up to date review of literature on the topics covered in this dissertation is presented.

2. In Chapter 2, we show application of shrinkage and pretest estimation in multiple regression. Three real data examples are given. Monte Carlo study show that restricted and pretest estimators have superior risk performance compared to the unrestricted and positive-shrinkage estimators when the underlying model is assumed to be correctly specified. However, under model misspecification, positive-shrinkage estimators show superior performance in terms of quadratic risk.

   In comparing shrinkage and APE, we developed and implemented the algorithm for simultaneous sub-model selection using AIC and BIC to obtain shrinkage estimates of the regression coefficients. Estimates based on lasso, adaptive lasso, and SCAD have been obtained and their performance compared

with shrinkage estimators by calculating average prediction errors based on cross validation. We demonstrated using a medical data example that positive-shrinkage estimator outperforms APE over a range of alternative parameter space. In general, positive shrinkage estimators maintain its superiority over all other estimators for moderate sample sizes and when there are large number of nuisance covariates present in the model.

Further, through a Monte Carlo experiment, shrinkage estimators have been compared with APE when the parameter space is high-dimensional, i.e., when $p$ is large. We considered only classical and high-dimensional cases with $p < n$ since shrinkage estimators do not exist for $p > n$.

3. As an extension of Ahmed et al. (2007), in Chapter 3, we considered shrinkage estimation of slope parameters in a PLM. We explored the suitability of using B-spline basis expansion to estimate the nonparametric component, $g(\cdot)$. The dominance of shrinkage and positive-shrinkage estimators over classical ML estimators have been shown using asymptotic quadratic distributional risk functions. We have found that B-spline is very flexible to be incorporated in a regression model when one considers to use uniform knots. If uniform knots are not preferable, B-splines are still attractive albeit the number of knots and their placements need to be obtained first. Since the nonparametric part can also be estimated using kernel-based methods, we compared the bias of the estimators based on when $g(\cdot)$ was approximated by both B-spline bases and kernel-based methods. For the nonparametric component, a flat function, such as $g(t) = 4\sin(\pi t)$, and a highly oscillating non-flat function, such as

$$g(t) = \sin\left(-\frac{2\pi(0.35 \times 10 + 1)}{0.35t + 1}\right), \quad t \in [0, 10],$$

have been considered. Simulation results showed that B-spline-based estimators have less bias than the kernel-based estimators.

4. In Chapter 4, shrinkage M-estimation (SM) is considered in the context of a PLM. We developed shrinkage and positive-shrinkage M-estimators (SM+) when we have prior information about a subset of the covariates. Based on a quadratic risk function, we computed relative risk of SM estimators with respect to the unrestricted M-estimator (UM). We analytically demonstrated that shrinkage estimators outperform classical full model M-estimators throughout the entire parameter space. In simulation experiments, four error distributions have been considered to explore the performance of the proposed estimators.

   We have found that restricted M-estimator (RM) outperforms all other estimators when the nuisance subspace is a zero vector. Overall, SM+ dominates both SM and RM for a wider range of the alternative parameter space.

# Chapter 2

# Absolute Penalty and Shrinkage Estimation in Multiple Regression Models

## 2.1 Introduction

Shrinkage estimators combine sample and non-sample information in a way that shrinks the regression coefficients towards a plausible alternative subspace. In this sense, shrinkage estimates resemble lasso for lasso penalizes the least squares estimates on their sizes and shrinks them towards zero. In shrinkage estimation, however, most of the coefficients shrink, while some of them are eliminated by shrinking to exactly zero.

In this chapter, we demonstrate application of shrinkage estimation and compare the performance of positive-shrinkage and absolute penalty estimators (APE). We

divide the chapter into two halves. In the first half, we illustrate, with examples, the application of shrinkage estimators in a multiple regression setup. We study different shrinkage and APE for low-dimensional ($p << n$) as well as high-dimensional data with $p < n$. In the second half of this chapter, we compare prediction errors of shrinkage estimators with lasso, alasso and SCAD estimators through Monte Carlo simulation. A real data example is given to illustrate the methods.

### 2.1.1 Organization of the Chapter

In Section 2.2 we present estimation strategies using the idea of shrinkage and APE. Shrinkage, positive-shrinkage, pretest, and absolute penalty estimators are defined in this section. Asymptotic properties of shrinkage and pretest estimators are presented in Section 2.3. Application of shrinkage and pretest estimation is illustrated with three real data examples in Section 2.4. In Section 2.5, performance of shrinkage and APEs is compared using prostate data and through a Monte Carlo experiment.

## 2.2 Model and Estimation Strategies

Consider a regression model of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)'$ is a vector of responses, $\boldsymbol{X}$ is an $n \times p$ fixed design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown vector of parameters, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ is the vector of unobservable random errors, and the superscript ($'$) denotes the transpose

of a vector or matrix.

We do not make any distributional assumption about the errors except that $\boldsymbol{\varepsilon}$ has a cumulative distribution function $F(\boldsymbol{\varepsilon})$ with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{I}$, where $\sigma^2$ is finite. We make the following two assumptions, also called the regularity conditions, which are needed to derive the asymptotics of the estimators:

i) $\max\limits_{1\leq i\leq n} \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i \longrightarrow 0$ as $n \longrightarrow \infty$, where $\boldsymbol{x}_i'$ is the $i$th row of $\boldsymbol{X}$

ii) $\lim\limits_{n\to\infty}\left(\dfrac{\boldsymbol{X}'\boldsymbol{X}}{n}\right) = \boldsymbol{C}$, where $\boldsymbol{C}$ is a finite positive-definite matrix.

Suppose that $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. The sub-vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are assumed to have dimensions $p_1$ and $p_2$ respectively, and $p_1 + p_2 = p$, $p_i \geq 0$ for $i = 1, 2$. We are essentially interested in the estimation of $\boldsymbol{\beta}_1$ when it is plausible that $\boldsymbol{\beta}_2$ is a set of nuisance covariates. This situation may arise when there is over-modeling and one wishes to cut down the irrelevant part from the model (2.1). Thus, the parameter space can be partitioned and it is plausible that $\boldsymbol{\beta}_2$ is near some specified $\boldsymbol{\beta}_2^o$, which, without loss of generality, may be set to a null vector.

The above situation may be mathematically written in terms of a restriction on $\boldsymbol{\beta}$ as $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$. Here, $\boldsymbol{H}$ is a known $p_2 \times p$ matrix and $\boldsymbol{h}$ is $p_2 \times 1$ vector of known constants.

The unrestricted estimator (UE) of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{UE}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Under the restriction $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$, the restricted estimator (RE) is given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{RE}} = \hat{\boldsymbol{\beta}}^{\mathrm{UE}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{H}'(\boldsymbol{H}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{H}')^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}^{\mathrm{UE}} - \boldsymbol{h}),$$

which is a linear function of the unrestricted estimator.

We may consider testing the restriction in the form of testing the null hypothesis

$$H_0 : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}.$$

The test statistic is given by

$$\psi_n = \frac{(\boldsymbol{H}\hat{\boldsymbol{\beta}}^{\mathrm{UE}} - \boldsymbol{h})'(\boldsymbol{H}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{H}')^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}^{\mathrm{UE}} - \boldsymbol{h})}{s_e^2}, \tag{2.2}$$

where

$$s_e^2 = \frac{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UE}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UE}})}{n - p}$$

is an estimator of $\sigma^2$. Under $H_0$, $\psi_n$ follows a chi-square distribution with $p_2$ degrees of freedom.

Now, we outline the estimation strategies in the following section.

### 2.2.1 Shrinkage Estimators

#### Shrinkage and Positive-shrinkage Estimators

A Stein-type or shrinkage estimator (SE) $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ of $\boldsymbol{\beta}_1$ can be defined as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{S}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{RE}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RE}})\left\{1 - \kappa\psi_n^{-1}\right\}, \text{ where } \kappa = p_2 - 2, \quad p_2 \geq 3.$$

Here, $\psi_n$ is defined in (2.2).

One problem with SE is that its components may have a different sign from the coordinates of $\hat{\boldsymbol{\beta}}_1^{\text{UE}}$. This could happen if $\kappa\psi_n^{-1}$ is larger than unity. One possibility is when $p_2 \geq 3$ and $\psi_n < 1$. From the practical point of view, the change of sign would affect its interpretability. However, this behavior does not adversely affect the risk performance of SE. To overcome the sign problem, a positive-rule Stein-type estimator (PSE) has been defined by retaining the positive-part of the SE. A PSE has the form

$$\hat{\boldsymbol{\beta}}_1^{\text{S+}} = \hat{\boldsymbol{\beta}}_1^{\text{RE}} + (\hat{\boldsymbol{\beta}}_1^{\text{UE}} - \hat{\boldsymbol{\beta}}_1^{\text{RE}}) \left\{1 - \kappa\psi_n^{-1}\right\}^+, \quad p_2 \geq 3$$

where $z^+ = \max(0, z)$. Alternatively, the PSE can be written as

$$\hat{\boldsymbol{\beta}}_1^{\text{S+}} = \hat{\boldsymbol{\beta}}_1^{\text{RE}} + (\hat{\boldsymbol{\beta}}_1^{\text{UE}} - \hat{\boldsymbol{\beta}}_1^{\text{RE}}) \left\{1 - \kappa\psi_n^{-1}\right\} I(\psi_n < \kappa), \quad p_2 \geq 3,$$

where $I(\cdot)$ is an indicator function.

Throughout they study, we call a PSE as positive-shrinkage estimator. Ahmed (2001) and others studied the asymptotic properties of Stein-type estimators in various contexts.

**Preliminary Test Estimator**

The preliminary test estimator or pretest estimator (PT) for the regression parameter $\boldsymbol{\beta}_1$ is obtained as

$$\hat{\boldsymbol{\beta}}_1^{\text{PT}} = \hat{\boldsymbol{\beta}}_1^{\text{UE}} - (\hat{\boldsymbol{\beta}}_1^{\text{UE}} - \hat{\boldsymbol{\beta}}_1^{\text{RE}}) I(\psi_n < c_{n,\alpha}), \tag{2.3}$$

where $I(\cdot)$ is an indicator function, and $c_{n,\alpha}$ is the upper $100\alpha$ percentage point of the test statistic $\psi_n$.

In pretest estimation, the *prior* information is tested before choosing the estimator for practical purposes whereas shrinkage and positive-shrinkage estimators use the value of the test statistic to obtain the estimates.

Pretest estimator either rejects or fails to reject the restricted estimator ($\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$) based on whether $\psi_n < c_{n,\alpha}$, while shrinkage estimator is a smoothed version of the pretest estimator. For this reason, pretesting is sometimes called "hard thresholding", while shrinkage estimation is called "soft thresholding".

### 2.2.2 Absolute Penalty Estimators

In this section, we define some absolute penalty estimators (APEs). These estimators are members of the penalized least squares family, and are suitable for both high-dimensional and low-dimensional data. Ahmed et al. (2010) mentioned that penalized least squares (PLS) estimation provides a generalization of both nonparametric least squares and weighted projection estimators, and a popular version of the PLS is given by Tikhonov regularization (Tikhonov, 1963).

Frank and Friedman (1993) introduced bridge regression, a generalized version of penalty (or absolute penalty type) estimators. For a given penalty function $\pi(\cdot)$ and regularization parameter $\lambda$, the general form of the objective function can be written as

$$\phi(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\pi(\beta),$$

where the penalty function is of the form

$$\pi(\beta) = \sum_{j=1}^{m} |\beta_j|^{\gamma},\ \gamma > 0. \tag{2.4}$$

The penalty function in (2.4) bounds the $L_\gamma$ norm of the parameters in the given model as $\sum_{j=1}^{m} |\beta_j|^{\gamma} \leq t$, where $t$ is the tuning parameter that controls the amount of shrinkage. Notice that for $\gamma = 2$, we have ridge estimates which are obtained by minimizing the penalized residual sum of squares

$$\tilde{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},\ p = p_1 + p_2, \tag{2.5}$$

where, $\lambda$ is the tuning parameter which controls the amount of shrinkage.

Frank and Friedman (1993) did not solve for the bridge regression estimators for any $\gamma > 0$. Interestingly, for $\gamma < 2$, it shrinks the coefficient towards zero, and depending on the value of $\lambda$, it sets some of them to exactly zero. Thus, the procedure combines variable selection and shrinking of the coefficients of penalized regression.

An important member of the penalized least squares family is the $L_1$ penalized least squares estimator, which is obtained when $\gamma = 1$. This is known as the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani (1996)). Although LASSO is an acronym, it is now commonly written as "lasso". In this dissertation, we use "lasso" and "LASSO" interchangeably.

A variable or feature selection procedure is said to have the oracle property if it identifies the right subset of variables and has the optimal estimation rate. So, it is desirable for an estimator to have oracle property. Zou (2006) argued that variable selection via the lasso could be inconsistent, and proposed the adaptive lasso, which

has the oracle property. Fan and Li (2001) introduced a non-convex penalty called the smoothly clipped absolute deviation (SCAD) and showed that their estimator satisfies the oracle property. For a comprehensive review of variable selection, please see Hesterberg et al. (2008) and Bühlmann and van de Geer (2011).

In the following, the APEs are defined.

### LASSO

Proposed by Tibshirani (1996), lasso is a member of the penalized least squares family, which performs variable selection and parameter estimation simultaneously. Lasso is closely related to ridge regression.

Lasso solutions are similarly defined by replacing the squared penalty $\sum_{j=1}^{p} \beta_j^2$ in the ridge solution (2.5) with the absolute penalty $\sum_{j=1}^{p} |\beta_j|$ in the lasso,

$$\tilde{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2.6}$$

Although the change apparently looks minor, the absolute penalty term makes it impossible to have an analytic solution for the lasso. Originally, lasso solutions were obtained via quadratic programming. Later, Efron et al. (2004) proposed Least Angle Regression (LAR), a type of stepwise regression with which the lasso estimates can be obtained at the same computational cost as that of ordinary least squares estimation. Further, the lasso estimator remains numerically feasible for dimensions $p$ that are much higher than the sample size $n$. Zou and Hastie (2005) introduced a hybrid

penalized least squares regression with the so called *elastic net penalty*

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|).$$

Here, the penalty function is a linear combination of the ridge regression penalty function and lasso penalty function. Here, $\alpha$ controls the amount of weight given to ridge and lasso penalty. A different type of penalized least square, called *garotte* is due to Breiman (1993).

Ahmed et al. (2007) proposed an APE for partially linear models. Further, they reappraised the properties of shrinkage estimators based on Stein-rule estimation.

Recently, Friedman et al. (2010) have developed fast algorithms for lasso and related methods for generalized linear models. A complete regularization path can be obtained via the cyclical coordinate descent algorithm. A generalization of the lasso, called the randomized lasso, is discussed in Meinshausen and Bühlmann (2010).

### Adaptive LASSO

The adaptive lasso estimator $\hat{\beta}^{\text{alasso}}$ is obtained by

$$\hat{\boldsymbol{\beta}}^{\text{alasso}} = \arg\min_{\beta} \left|\left| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|, \tag{2.7}$$

where the weight function is given by

$$\hat{\boldsymbol{w}} = \frac{1}{|\hat{\boldsymbol{\beta}}^*|^\gamma}; \quad \gamma > 0.$$

Equation (2.7) is a "convex optimization problem and its global minimizer can be efficiently solved" (Zou, 2006). The $\hat{\boldsymbol{\beta}}^*$ is a root-n consistent estimator of $\boldsymbol{\beta}$. For example, $\hat{\boldsymbol{\beta}}^*$ can be $\hat{\boldsymbol{\beta}}_{\text{ols}}$. Once we have the $\hat{\boldsymbol{\beta}}_{\text{ols}}$, we need to select a $\gamma > 0$ and calculate the weights. Finally, the adaptive lasso estimates are obtained from (2.7). The LARS algorithm (Efron et al., 2004) can be used to obtain adaptive lasso estimates. The steps are given below.

Step 1. Reweight the data by defining $\boldsymbol{x}_j^{\text{new}} = \boldsymbol{x}_j^{\text{old}}/\hat{w}_j, \quad j = 1, 2, \ldots, p$

Step 2. Solve the lasso problem as

$$\hat{\boldsymbol{\beta}}^{**} = \arg\min_{\beta} \left|\left| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j^{\text{new}} \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Step 3. Return $\hat{\beta}_j^{\text{alasso}} = \hat{\beta}_j^{**}/\hat{w}_j$

For a detailed discussion on the computation of adaptive lasso, we refere to Zou (2006).

### SCAD

While lasso uses $L_1$ penalty function, the amount of penalty increases linearly in the magnitude of its argument (Fan et al., 2009). As such, lasso produces biased estimates for large regression coefficients. This issue was addressed by Fan and Li (2001) who proposed the smoothly clipped absolute deviation or SCAD. This method not only retains the good features of both subset selection and ridge regression but also produces sparse solutions, ensures continuity of the selected models (for the stability of model selection) and give unbiased estimates for large coefficients. The

estimates are obtained as

$$\hat{\boldsymbol{\beta}}^{\text{SCAD}} = \arg\min_{\beta} \left|\left| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} S_{\alpha,\lambda} ||\beta_j||_1.$$

Here $S_{\alpha,\lambda}(\cdot)$ is the smoothly clipped absolute deviation penalty, and $||\cdot||_1$ denotes $L_1$ norm. SCAD penalty is a symmetric and a quadratic spline on $[0, \infty)$ with knots at $\lambda$ and $\alpha\lambda$, whose first order derivative is given by

$$S_{\alpha,\lambda}(x) = \lambda \left\{ I(|x| \leq \lambda) + \frac{(\alpha\lambda - |x|)_+}{(\alpha - 1)\lambda} I(|x| > \lambda) \right\}, \quad x \geq 0. \tag{2.8}$$

Here $\lambda > 0$ and $\alpha > 2$ are the tuning parameters. For $\alpha = \infty$, the expression (2.8) is equivalent to the $L_1$ penalty in LASSO. The solution of SCAD penalty is originally due to Fan (1997).

## 2.3 Asymptotic Properties of Shrinkage Estimators

In this section, we present the asymptotic distributions of unrestricted (UE), pretest (PT), shrinkage (SE), and positive-shrinkage estimators (PSE), and the test statistic $\psi_n$. This facilitates finding the asymptotic distributional bias (ADB), asymptotic distributional quadratic bias (ADQB), and asymptotic distributional quadratic risk (ADQR) of the estimator of $\boldsymbol{\beta}$.

Ahmed (1997) noted that since the test statistic $\psi_n$ is consistent against fixed $\boldsymbol{\beta}$ such that $\boldsymbol{H}\boldsymbol{\beta} \neq \boldsymbol{h}$, the PT, SE, and PSE will be asymptotically equivalent to UE for fixed alternative up to the order $\mathcal{O}(n^{-\frac{1}{2}})$. Therefore, for the large-sample situation

there is not much to investigate. This means that, for an estimator $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$ under fixed alternative, the asymptotic distribution of $\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})$ is equivalent to $\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathrm{UE}} - \boldsymbol{\beta})$. In this case, to obtain meaningful asymptotics, a class of local alternatives, $\{K_n\}$, is considered, which is given by

$$K_n : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h} + \frac{\boldsymbol{\omega}}{\sqrt{n}}, \tag{2.9}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_{p_2})' \in \Re^{p_2}$ is a fixed vector. We notice that $\boldsymbol{\omega} = \mathbf{0}$ implies $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$, i.e., the fixed alternative is a particular case of (2.9). In the following, we evaluate the performance of each estimator under local alternative.

For an estimator $\boldsymbol{\beta}^*$ and a positive-definite matrix $\boldsymbol{W}$, we define the loss function of the form

$$L(\boldsymbol{\beta}^*; \boldsymbol{\beta}) = n(\boldsymbol{\beta}^* - \boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{\beta}^* - \boldsymbol{\beta}).$$

These loss functions are generally known as weighted quadratic loss functions, where $\boldsymbol{W}$ is the weighting matrix. For $\boldsymbol{W} = \boldsymbol{I}$, we get squared error loss functions.

The expectation of the loss function

$$E[L(\boldsymbol{\beta}^*, \boldsymbol{\beta}); \boldsymbol{W}] = R[(\boldsymbol{\beta}^*, \boldsymbol{\beta}); \boldsymbol{W}],$$

is called the risk function, which can be written as

$$\begin{aligned} R[(\boldsymbol{\beta}^*, \boldsymbol{\beta}); \boldsymbol{W}] &= nE[(\boldsymbol{\beta}^* - \boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{\beta}^* - \boldsymbol{\beta})] \\ &= n\,\mathrm{tr}[\boldsymbol{W}\{E(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})'\}] \\ &= \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Gamma}^*), \end{aligned} \tag{2.10}$$

where $\mathbf{\Gamma}^*$ is the covariance matrix of $\boldsymbol{\beta}^*$.

The performance of the estimators can be evaluated by comparing the risk functions with a suitable matrix $\boldsymbol{W}$. An estimator with a smaller risk is preferred. The estimator $\boldsymbol{\beta}^*$ will be called inadmissible if there exists another estimator $\boldsymbol{\beta}^0$ such that

$$R(\boldsymbol{\beta}^0, \boldsymbol{\beta}) \leq R(\boldsymbol{\beta}^*, \boldsymbol{\beta}) \quad \forall(\boldsymbol{\beta}, \boldsymbol{W}) \tag{2.11}$$

with strict inequality for some $\boldsymbol{\beta}$. In such cases, we say that the estimator $\boldsymbol{\beta}^0$ dominates $\boldsymbol{\beta}^*$. If, however, instead of (2.11) holding for every $n$, we have

$$\lim_{n\to\infty} R(\boldsymbol{\beta}^0, \boldsymbol{\beta}) \leq \lim_{n\to\infty} R(\boldsymbol{\beta}^*, \boldsymbol{\beta}) \quad \forall\boldsymbol{\beta}, \tag{2.12}$$

with strict inequality for some $\boldsymbol{\beta}$, then $\boldsymbol{\beta}^*$ is termed as an asymptotically inadmissible estimator of $\boldsymbol{\beta}$. The expression in (2.11) is not easy to prove. An alternative is to consider the asymptotic distributional quadratic risk (ADQR) for the sequence of local alternatives $\{K_n\}$.

Consider the asymptotic cumulative distribution function (cdf) of $\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})/s_e$ under $\{K_n\}$ exists, and defined as

$$G(\boldsymbol{y}) = P[\lim_{n\to\infty} \sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})/s_e \leq \boldsymbol{y}].$$

This is known as the asymptotic distribution function (ADF) of $\boldsymbol{\beta}^*$. Further let

$$\Gamma = \int\int\cdots\int \boldsymbol{y}\boldsymbol{y}' dG(\boldsymbol{y})$$

be the dispersion matrix which is obtained from ADF. Then the ADQR may be

defined as

$$R(\boldsymbol{\beta}^*;\boldsymbol{\beta}) = \text{tr}(\boldsymbol{W}\boldsymbol{\Gamma}). \tag{2.13}$$

Further, $\boldsymbol{\beta}^*$ strictly dominates $\boldsymbol{\beta}^0$ if $R(\boldsymbol{\beta}^*;\boldsymbol{\beta}) < R(\boldsymbol{\beta}^0;\boldsymbol{\beta})$ for some $(\boldsymbol{\beta}, \boldsymbol{W})$. The asymptotic risk may be obtained by replacing $\boldsymbol{\Gamma}$ with the limit of the actual dispersion matrix of $\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})$ in the ADQR function. However, this may require some extra regularity conditions. Ahmed (2001), Sen (1986), and Saleh and Sen (1985) among others, have explained this point in various other contexts.

To obtain the asymptotic distribution of the proposed estimators, and the test statistic $\psi_n$, we consider the following theorems.

**Theorem 2.3.1.** Under the regularity conditions, and if $\sigma^2 < \infty$, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{\text{UE}} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{C}^{-1}),$$

where $\xrightarrow{d}$ implies convergence in distribution.

The proof of the theorem can be found in Sen and Singer (1993).

**Theorem 2.3.2.** If $\boldsymbol{V} = (V_1, V_2, \cdots, V_p)'$ is a $p$-dimensional normal vector distributed as $N_p(\boldsymbol{\zeta}, \boldsymbol{I}_p)$, then for a measurable function $\varphi$, we have

$$E\left[\boldsymbol{V}\varphi(\boldsymbol{V}'\boldsymbol{V})\right] = \boldsymbol{\zeta}E\left[\varphi(\chi^2_{p+2}(\Delta))\right] \tag{2.14}$$

Here, $\chi^2_\nu(\Delta)$ is a non-central chi-square distribution with $\nu$ degrees of freedom and noncentrality parameter $\Delta$.

**Theorem 2.3.3.** If $\boldsymbol{V} = (V_1, V_2, \cdots, V_p)'$ is a $p$-dimensional normal vector distributed

as $N_p(\boldsymbol{\zeta}, \boldsymbol{I}_p)$, then for a measurable function $\varphi$, we have

$$E\left[\boldsymbol{V}\boldsymbol{V}'\varphi(\boldsymbol{V}'\boldsymbol{V})\right] = \boldsymbol{I}_p E\left[\varphi(\chi^2_{p+2}(\Delta))\right] + \boldsymbol{\zeta}\boldsymbol{\zeta}' E\left[\varphi(\chi^2_{p+4}(\Delta))\right] \tag{2.15}$$

The proof of Theorems 2.3.2 and 2.3.3 can be found in the appendix of Judge and Bock (1978).

**Theorem 2.3.4.** Let $\boldsymbol{X} = (\boldsymbol{X}_1' \vdots \boldsymbol{X}_2')'$ be distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \Sigma_{12} \\ \boldsymbol{\Sigma}_{21} & \Sigma_{22} \end{pmatrix}.$$

Then the conditional distribution of $\boldsymbol{X}_1$, given $\boldsymbol{X}_2 = \boldsymbol{x}_2$, is normal with

$$\mu_{11.2} = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$$

and

$$\Sigma_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

*Proof.* The proof can be found in Johnson and Wichern (2001).

### 2.3.1 Bias Performance

The asymptotic distributional bias (ADB) of an estimator $\boldsymbol{\beta}_1^*$ is defined as

$$\text{ADB}(\boldsymbol{\beta}_1^*) = E\left\{\lim_{n\to\infty} \sqrt{n}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)\right\}.$$

**Theorem 2.3.5.** Under the assumed regularity conditions and the definition above,

and under $\{K_n\}$, the ADB of the estimators are as follows:

$$\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{UE}}) = \mathbf{0} \tag{2.16}$$

$$\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{RE}}) = -\boldsymbol{\xi}, \text{ where } \boldsymbol{\xi} = \boldsymbol{C}^{-1}\boldsymbol{H}'(\boldsymbol{H}C^{-1}\boldsymbol{H}')^{-1}\boldsymbol{\omega} \tag{2.17}$$

$$\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{PT}}) = -\boldsymbol{\xi}H_{p_2+2}(\chi^2_{p_2,\alpha};\Delta), \text{ with } \Delta = \boldsymbol{\xi}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}, \text{ where } \boldsymbol{\Sigma} = \sigma^2\boldsymbol{C}^{-1} \tag{2.18}$$

$$\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{S}}) = -(p_2-2)\boldsymbol{\xi}E\left\{\chi^{-2}_{p_2};\Delta\right\} \tag{2.19}$$

$$\begin{aligned}\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{S+}}) &= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{S}}) \\ &\quad - \boldsymbol{\xi}E\left\{(\,1-(p_2-2)\chi^{-2}_{p_2+2}(\Delta))I(\chi^2_{p_2+2}(\Delta)<(p_2-2)\right\}.\end{aligned} \tag{2.20}$$

The proof of Theorem 2.3.5 can be found in Ahmed (1997) and Saleh (2006).

Here,

$$E\left\{(\chi^2_\nu(\Delta))^{-m}\right\} \tag{2.21}$$

is the expected value of the inverse of a non-central chi-square distribution with $\nu$ degrees of freedom and noncentrality parameter $\Delta$. For nonnegative interger-valued $\nu$ and $m$, and for $\nu > 2m$, the expectations in (2.21) can be obtained using a theorem in Bock et al. (1983, page 7). $H_\nu(\cdot,\Delta)$ is the cdf of a non-central chi-square distribution with $\nu$ degrees of freedom with non-centrality parameter $\Delta = \boldsymbol{\xi}'\Sigma^{-1}\boldsymbol{\xi}$ with $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{C}^{-1}$.

The bias expressions for all the estimators are not in the scalar form. We therefore take recourse by converting them into the quadratic form. Let us define the asymptotic distributional quadratic bias (ADQB) of the estimator $\boldsymbol{\beta}_1^*$ by

$$ADQB(\boldsymbol{\beta}_1^*) = [ADB(\boldsymbol{\beta}_1^*)]'\boldsymbol{\Sigma}^{-1}[ADB(\boldsymbol{\beta}_1^*)] \tag{2.22}$$

where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{C}^{-1}$ is the dispersion matrix of $\hat{\boldsymbol{\beta}}^{\mathrm{UE}}$ as $n \to \infty$.

Using the definition in (2.22), the asymptotic distributional quadratic bias of the estimators are presented below.

$$\mathrm{ADQB}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}) = \mathbf{0}, \tag{2.23}$$

$$\mathrm{ADQB}(\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}) = \Delta \tag{2.24}$$

$$\mathrm{ADQB}(\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}}) = \Delta \left\{ H_{p_2+2}(\chi^2_{p_2,\alpha}; \Delta) \right\}^2 \tag{2.25}$$

$$\begin{aligned} \mathrm{AQDB}(\hat{\boldsymbol{\beta}}^{\mathrm{S}}) = & \left[ -(p_2-2)\boldsymbol{\xi} E\left\{ \chi^{-2}_{p_2+2}(\Delta^2) \right\} \right]' \sigma^{-2} \boldsymbol{C} \left[ -(p_2-2)\boldsymbol{\xi} E\left\{ \chi^{-2}_{p_2+2}(\Delta^2) \right\} \right] \\ & (p_2-2)^2 \Delta \left[ E\left\{ \chi^{-2}_{p_2+2}(\Delta) \right\} \right]^2 \end{aligned} \tag{2.26}$$

$$\begin{aligned} \mathrm{ADQB}(\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}) = & \Delta \left[ H_{p_2+2}(p_2-2;\Delta) + (p_2-2) E\left\{ \chi^{-2}_{p_2+2}(\Delta) \right\} \right. \\ & \left. + E\left\{ \chi^{-2}_{p_2+2}(\Delta) I(\chi^2_{p_2+2}(\Delta) > p_2-2) \right\} \right]. \end{aligned} \tag{2.27}$$

### 2.3.2 Risk Performance

Following (2.13), under the assumed regularity conditions, and local alternative $\{K_n\}$, the ADQR expressions are as follows. The proof can be found in Ahmed (1997) and Saleh (2006).

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}};\boldsymbol{W}) = \sigma^2\mathrm{tr}(\boldsymbol{W}\boldsymbol{C}^{-1}) \tag{2.28}$$

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}};\boldsymbol{W}) = \sigma^2\mathrm{tr}(\boldsymbol{W}\boldsymbol{C}^{-1}) - \sigma^2\mathrm{tr}(\boldsymbol{Q}) + \boldsymbol{\omega}'\boldsymbol{B}^{-1}\boldsymbol{Q}\boldsymbol{\omega} \tag{2.29}$$

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}};\boldsymbol{W}) &= \sigma^2\mathrm{tr}(\boldsymbol{W}\boldsymbol{C}^{-1}) - \sigma^2\mathrm{tr}(\boldsymbol{Q})H_{p_2+2}(\chi^2_{p_2,\alpha};\Delta) \\ &\quad + \boldsymbol{\omega}'\boldsymbol{B}^{-1}\boldsymbol{\omega}\left\{2H_{p_2+2}(\chi^2_{p_2,\alpha};\Delta) - H_{p_2+4}(\chi^2_{p_2,\alpha};\Delta)\right\} \end{aligned} \tag{2.30}$$

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}};\boldsymbol{W}) &= \sigma^2\mathrm{tr}(\boldsymbol{W}\boldsymbol{C}^{-1}) - (p_2-2)\sigma^2\mathrm{tr}(\boldsymbol{Q}_{11})\left\{2E[\chi^{-4}_{p_2+4}(\Delta)]\right. \\ &\quad \left.-(p_2-2)E[\chi^{-4}_{p_2+4}(\Delta)]\right\} \\ &\quad + (p_2-2)(p_2+6)(\gamma_1'\boldsymbol{Q}_{11}\gamma_1)E[\chi^{-4}_{p_2+4}(\Delta)] \end{aligned} \tag{2.31}$$

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}};\boldsymbol{W}) &= R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}};\boldsymbol{W}) + (p_2-2)\sigma^2\mathrm{tr}(\boldsymbol{Q})\left[E\left\{\chi^{-2}_{p_2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le p_2-2)\right\}\right. \\ &\quad \left.-\ (p_2-2)E\left\{\chi^{-4}_{p+2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le p_2-2)\right\}\right] \\ &\quad - \sigma^2\mathrm{tr}(\boldsymbol{Q})H_{p_2+2}(p_2-2;\Delta) + \boldsymbol{\omega}'\boldsymbol{B}^{-1}\boldsymbol{Q}\boldsymbol{\omega}\left\{2H_{p_2+4}(p_2-2;\Delta)\right\} \\ &\quad - (p_2-2)\boldsymbol{\omega}'\boldsymbol{B}^{-1}\boldsymbol{Q}\boldsymbol{\omega}\left[2E\left\{\chi^{-2}_{p_2+2}(\Delta)I(\chi^2_{p_2+2}(\Delta)\le p_2-2)\right\}\right. \\ &\quad -2E\left\{\chi^{-2}_{p_2+4}(\Delta)I(\chi^2_{p_2+4}(\Delta)\le p_2-2)\right\} \\ &\quad \left.+\ (p_2-2)E\left\{\chi^{-4}_{p_2+4}(\Delta)I(\chi^{-4}_{p_2+4}(\Delta)\le p_2-2)\right\}\right], \end{aligned} \tag{2.32}$$

where $\boldsymbol{Q} = \boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{H}'\boldsymbol{B}^{-1}$, and $\boldsymbol{B} = \boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}'$.

Ahmed (1997) have studied the statistical properties of various shrinkage and pretest estimators. It was remarked that none of the unrestricted, restricted, and pretest estimators is inadmissible with respect to any of the others. However, at $\Delta = 0$,

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}};\boldsymbol{W}) \le R(\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}};\boldsymbol{W}) \le R(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}};\boldsymbol{W}).$$

Therefore, for all $(\Delta; \boldsymbol{W})$ and $p_2 \geq 3$,

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}+}; \boldsymbol{W}) \leq R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}; \boldsymbol{W}) \leq R(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}; \boldsymbol{W})$$

is satisfied. Thus, we conclude that $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}+}$ performs better than $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ in the entire parameter space induced by $\Delta$. The gain in risk over $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ is substantial when $\Delta = 0$ or near.

## 2.4 Application of Shrinkage and Pretest Estimation

In the following, we study three real data examples. For each data set, we fit multiple regression models to predict the variable of interest from the available regressors. Shrinkage and pretest estimates are then obtained for the regression parameters. Performance of shrinkage and pretest estimators are assessed as per the criteria outlined in the following section. A Monte Carlo study is carried out afterwards.

### 2.4.1 Assessment Criteria

#### Low-dimensional Scenario ($p << n$)

The shrinkage estimation method utilizes the full- and sub-model estimates, and combines them in a way that shrinks the least-squares estimates towards the sub-model estimates. In this framework, if prior information about a subset of the covariates is available, then the estimates can be obtained by incorporating the available in-

formation in the estimation process. However, in the absence of prior information, one might do usual variable selection to select the best subsets. The best subset may be selected based on AIC, BIC or other model selection criteria. In the end, we have a full model with all the covariates, and a sub-model with the best subset of the covariates. Shrinkage estimates are then obtained from the full-model and the sub-model.

Let us consider AIC and BIC as sub-model selection criteria for a given data. We may choose to use the AIC sub-model or the BIC sub-model to build shrinkage estimators. Suppose, AIC retains three covariates, while BIC drops all but two covariates. So, it is possible to use different sub-models to obtain pretest and shrinkage estimates. In this study, performance of each pair of full- and sub-models were evaluated by calculating the prediction error based on $K$-fold cross validation. In a cross validation, the data set is randomly divided into $K$ subsets of roughly equal size. One subset is left aside, and termed as test set, while the remaining $K-1$ subsets, called training set, are used to fit the model. The fitted model is then used to predict the responses of the test data set. Finally, prediction errors are obtained by taking the squared deviation of the observed and predicted values in the test set. The process is repeated for all $K$s and the prediction errors are combined.

We consider $K = 5, 10$. Both raw cross validation estimate (CVE), and bias corrected cross validation estimate of prediction errors are obtained for each configuration. The bias corrected cross validation estimate is the adjusted cross-validation estimate designed to compensate for the bias introduced by not using leave-one-out cross-validation (Tibshirani and Tibshirani, 2009).

Since cross validation is a random process, the estimated prediction error varies

across runs and for different values of $K$. To account for the random variation, we repeat the cross validation process 2000 times, and estimate the average prediction errors along with their standard errors. The number of repetitions was initially varied, and we settled with this number as no noticeable variations in the standard errors were observed for higher values.

**High- and Ultra High-dimensional Scenario (large $p$, $p < n$, and $p >> n$)**

Shrinkage estimation works when $p$ is high-dimensional provided $p < n$. However, for ultra high dimensional cases ($p >> n$), shrinkage estimators do not exist since maximum likelihood estimators do not exist when the number of unknowns is larger than the number of sample observations.

In the following, we present three real data examples to illustrate shrinkage and pretest estimation.

### 2.4.2 $NO_2$ Data

The data came from a subsample of 60 observations from a large data set from a study where air pollution at a road was related to traffic volume and meteorological variables. Data were collected at Alnabru in Oslo, Norway, between October 2001 and August 2003 by the Norwegian Public Roads Administration. The data are freely available from `http://lib.stat.cmu.edu/datasets/`.

The idea is to predict the logarithm of the concentration of $NO_2$ particles (`conc`) from the following covariates: logarithm of the number of cars per hour (`cars`), temperature (degree C) two-meter above ground (`tmp2m`), wind speed (meters/second)

(`windsp`), temperature difference (degree C) between 25- and 2-meters above ground (`tmpdiff`), wind direction (degrees between 0 and 360) (`winddir`), and hour of day (`hour`).

Candidate full- and sub-models based for the data are listed in Table 2.1. The sub-models are obtained based on AIC and BIC.

Table 2.1: Candidate full- and sub-models for $NO_2$ data.

| Selection Criterion | Model: Response ~ Covariates |
|---|---|
| Full Model | `conc~ cars + windsp + tmp2m + tmpdiff + winddir + hour` |
| BIC | `conc~cars + windsp` |
| AIC | `conc~cars + windsp + tmp2m` |

Table 2.2 summarizes the average prediction errors with their standard deviations for UE, RE, PSE and PTE. The terms listed in the first column of Table 2.2 are defined as follows: UE represents the full-model, RE(AIC) and RE(BIC) denotes the restricted estimators with sub-models obtained by AIC and BIC. PSE(AIC), PSE(BIC) represents positive-shrinkage estimators with AIC and BIC sub-models. PTE(AIC) and PTE(BIC) are similarly denoted to represent pretest estimators.

Comparing the bias corrected estimate of the cross validation error for 10-fold cross validation, PSE(BIC) has the smallest average prediction error of 0.265 with standard error .011. For this data set, RE and PTE are performing very close to PSE, mainly because the sub-models based on AIC and BIC produce the best model to predict concentration of $NO_2$. Recall that, RE and PTE works best when the nuisance set is nearly zero. This data set is an example of such a scenario. However, this may not be the case for every data set, or prior information may not be trustworthy in every situation. Since PSE takes into account both full and sub-model, it is less sensitive

Table 2.2: Average prediction errors based on $K$-fold cross validation repeated 2000 times for $NO_2$ data. Numbers in smaller font are the corresponding standard errors of the prediction errors.

| | Raw CVE | | Bias Corrected CVE | |
|---|---|---|---|---|
| Estimator | $K = 5$ | $K = 10$ | $K = 5$ | $K = 10$ |
| UE | $.299_{.020}$ | $.298_{.019}$ | $.283_{.012}$ | $.281_{.011}$ |
| RE(AIC) | $.267_{.021}$ | $.267_{.019}$ | $.268_{.014}$ | $.266_{.013}$ |
| RE(BIC) | $.266_{.020}$ | $.265_{.012}$ | $.265_{.013}$ | $.265_{.012}$ |
| PSE(AIC) | $.273_{.014}$ | $.275_{.013}$ | $.271_{.012}$ | $.266_{.012}$ |
| PSE(BIC) | $.272_{.013}$ | $.273_{.012}$ | $.270_{.019}$ | $.265_{.011}$ |
| PTE(AIC) | $.271_{.021}$ | $.270_{.015}$ | $.268_{.021}$ | $.266_{.014}$ |
| PTE(BIC) | $.276_{.021}$ | $.271_{.016}$ | $.267_{.019}$ | $.267_{.011}$ |

to model misspecification. We will explore this further through Monte Carlo study in subsection 2.4.5.

### 2.4.3 State Data

Faraway (2002) illustrated variable selection methods on a data set called `state`. There are 97 observations (cases) on 9 variables. The variables are: population estimate as of July 1, 1975 (`Population`); per capita income (1974) (`Income`); illiteracy (1970, percent of population) (`Illiteracy`); life expectancy in years (1969-71) (`Life.exp`); murder and non-negligent manslaughter rate per 100,000 population (1976) (`Murder`); percent high-school graduates (1970) (`Hs.grad`); mean number of days with minimum temperature 32 degrees (1931-1960) in capital or large city (`Frost`); and land area in square miles (`Area`).

We consider predicting life expectancy from the available covariates. It was found that population, murder, high school graduates, and temperature produce the best

model based on AIC or BIC. A model based on CP statistic that includes population, high school graduates, and temperature showed the largest adjusted $R^2$. All the models are listed in Table 2.3.

Table 2.3: Full and candidate sub-models for state data.

| Selection Criterion | Model: Response ~ Covariates |
|---|---|
| Full | `Life.exp~ Population + Murder + Hs.grad + Frost + Income + Illiteracy + Area` |
| AIC/BIC | `Life.exp~ Population + Murder + Hs.grad + Frost` |
| CP | `Life.exp~ Murder + Hs.grad + Frost` |

Table 2.4: Average prediction errors (thousands) based on $K$-fold cross validation, repeated 2000 times for state data. Numbers in smaller font are the corresponding standard errors of the prediction errors.

| | Raw CVE | | Bias Corrected CVE | |
|---|---|---|---|---|
| Estimator | $K = 5$ | $K = 10$ | $K = 5$ | $K = 10$ |
| UE | $.879_{.144}$ | $.847_{.086}$ | $.819_{.119}$ | $.820_{.079}$ |
| RE(AIC) | $.637_{.063}$ | $.614_{.036}$ | $.599_{.052}$ | $.597_{.033}$ |
| RE(CP) | $.639_{.058}$ | $.639_{.033}$ | $.626_{.048}$ | $.626_{.031}$ |
| PSE(AIC) | $.740_{.124}$ | $.690_{.074}$ | $.696_{.104}$ | $.671_{.068}$ |
| PSE(CP) | $.768_{.106}$ | $.746_{.063}$ | $.727_{.090}$ | $.727_{.058}$ |
| PTE(AIC) | $.637_{.066}$ | $.614_{.036}$ | $.599_{.054}$ | $.597_{.033}$ |
| PTE(CP) | $.662_{.069}$ | $.639_{.035}$ | $.629_{.059}$ | $.626_{.032}$ |

For this data set, the models given by AIC and BIC are the same. When the models are correctly specified, restricted estimator performs best. Here, the RE has the smallest prediction error. Under model uncertainty, however, the scenario may completely change completely as the risk of RE theoretically higher than than of the UE when the sub-model deviates from the true underlying model. This is explored in the simulation study presented in section 2.4.5.

### 2.4.4 Galapagos Data

Faraway (2002) analyzed the data about species diversity on the Galapagos islands. The Galapagos data contains 30 rows and seven variables. Each row represents an island, and the covariates represent various geographic measurements. The relationship between the number of species of tortoise and several geographic variables is of interest. The data set has the following covariates: `Species`–the number of species of tortoise found on the island, `Endemics` represents the number of endemic species, `Area` represents the area of the island ($km^2$), `Elevation` measures the highest elevation of the island (m), `Nearest` is the distance from the nearest island (km), `Scruz` measures the distance from Santa Cruz island (km), and `Adjacent` measures the area of the adjacent island ($km^2$). The original data set contained missing values for some of the covariates, which have been imputed by Faraway (2002) for convenience.

The full model and the sub-models based on AIC and BIC are shown in Table 2.5.

Table 2.5: Full and candidate sub-models for Galapagos data.

| Selection Criterion | Model: Response ~ Covariates |
|---|---|
| Full | `Species~ Endemics + Area + Elevation + Nearest + Scruz + Adjacent` |
| AIC | `Species~ Endemics + Area + Elevation` |
| BIC | `Species~ Endemics` |

We obtain restricted, pretest, and positive-shrinkage estimates of the regression parameters for the Galapagos data. Average prediction errors along with their standard errors for UE, RE, PSE, and PTE are presented in Table 2.6. Prediction errors and the standard errors are shown in thousands. PSE(AIC) represents positive shrinkage estimates based on sub-model given by AIC, and PSE(BIC) represents the same based on BIC. PTE(AIC) and PTE(BIC) are similarly defined for pretest estimators.

Table 2.6: Average prediction errors (thousands) based on $K$-fold cross validation, repeated 2000 times for Galapagos data. Numbers in smaller font are the corresponding standard errors of the predictor errors.

| | Raw CVE | | Bias Corrected CVE | |
|---|---|---|---|---|
| Estimator | $K = 5$ | $K = 10$ | $K = 5$ | $K = 10$ |
| UE | $13.87_{8.36}$ | $12.63_{4.36}$ | $11.31_{6.70}$ | $11.48_{3.93}$ |
| RE(AIC) | $12.45_{6.96}$ | $11.62_{4.28}$ | $10.10_{5.57}$ | $10.53_{3.85}$ |
| RE(BIC) | $1.78_{0.59}$ | $1.65_{0.24}$ | $1.46_{0.43}$ | $1.51_{0.29}$ |
| PSE(AIC) | $13.19_{7.82}$ | $11.98_{4.29}$ | $10.75_{6.27}$ | $10.88_{3.87}$ |
| PSE(BIC) | $9.07_{6.53}$ | $7.96_{3.75}$ | $7.54_{5.24}$ | $7.32_{3.38}$ |
| PTE(AIC) | $12.50_{6.98}$ | $11.63_{4.29}$ | $10.14_{5.58}$ | $10.54_{3.86}$ |
| PTE(BIC) | $5.39_{7.56}$ | $3.90_{6.16}$ | $4.40_{6.08}$ | $3.55_{5.56}$ |

For this data set, the RE and PTE have the smallest average prediction errors. We notice that, models based on BIC are smaller in size, and their average prediction errors are smaller than those of the AIC models. The difference in average prediction errors for the two sub-models is noticeably large. Such a large difference between the competing sub-models hints about possible error in model specification, and the consequences that it may cause. A Monte Carlo study conducted later in section 2.4.5 reveals the sensitivity of RE, PSE, and PTE when the hypothesized model deviates considerably from the true one.

It is noted here that the prediction errors are unusually large for this data set. This indicates that the predictors are not quite capturing the variability in the response.

### 2.4.5 Simulation Study: Comparing PSE with UE, RE, PTE

Based on the bias and risk expressions of PSE and PTE in section 2.3, we conduct Monte Carlo simulation experiments to examine the quadratic risk performance of the

estimators. We generate the response and the predictors from the following model:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots, +x_{pi}\beta_p + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.33}$$

where $x_{1i}$ and $x_{2i} \sim N(1, 2)$, and the $x_{si}$ are i.i.d. $N(0, 1)$ for $s = 3, \ldots, p$ and $i = 1, \ldots, n$. Moreover, $\varepsilon_i$ are i.i.d. $N(0, 1)$.

We are interested in testing the hypothesis $H_0 : \boldsymbol{\beta}_j = \mathbf{0}$, for $j = p_1 + 1, p_1 + 2, \ldots, p_1 + p_2$, with $p = p_1 + p_2$. Accordingly, we partition the regression coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$.

The number of simulations is initially varied. Finally, each realization is repeated 2000 times to obtain stable results. For each realization, we calculated bias of the estimators. We define $\Delta^* = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1, \mathbf{0})$, and $||\cdot||$ is the Euclidean norm. To determine the behavior of the estimators for $\Delta^* > 0$, further data sets are generated from those distributions under local alternative hypotheses. Various $\Delta^*$ values between [0,1] are considered.

Our objective is to study the behavior of PSE and PTE under varying degrees of model misspecification, i.e., when $\Delta^* > 0$. RE performs best if the nuisance subset is a zero vector ($\Delta^* = 0$). However, the risk of RE goes higher than the UE when it deviates substantially from $\Delta^* = 0$.

The risk performance of an estimator of $\boldsymbol{\beta}_1$ is measured by comparing its MSE with that of UE as defined below:

$$\text{RMSE}(\hat{\boldsymbol{\beta}}_1^{\text{UE}} : \hat{\boldsymbol{\beta}}_1^*) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{\text{UE}})}{\text{MSE}(\hat{\boldsymbol{\beta}}_1^*)}, \tag{2.34}$$

where $\hat{\boldsymbol{\beta}}_1^*$ is either the RE, PSE or PTE. The amount by which an RMSE is larger

than unity indicates the degree of superiority of the estimator $\hat{\boldsymbol{\beta}}_1^*$ over $\hat{\boldsymbol{\beta}}_1^{\text{UE}}$.

RMSEs for the RE, PSE and PTE are computed for $n = 30, 50, 100$, $p_1 = 3, 6, 9$, and $p_2 = 5, 7, 10$. Since the results are similar for all the configurations, we list the RMSEs in Table 2.7 for $n = 50$ and $(p_1, p_2) = (6, 10)$ only. Comparative RMSEs for RE, PSE and PTE for the configurations $(p_1, p_2) = (6, 5)$, $(6, 10)$, $(9, 5)$, and $(9, 10)$ are illustrated in Figure 2.1.

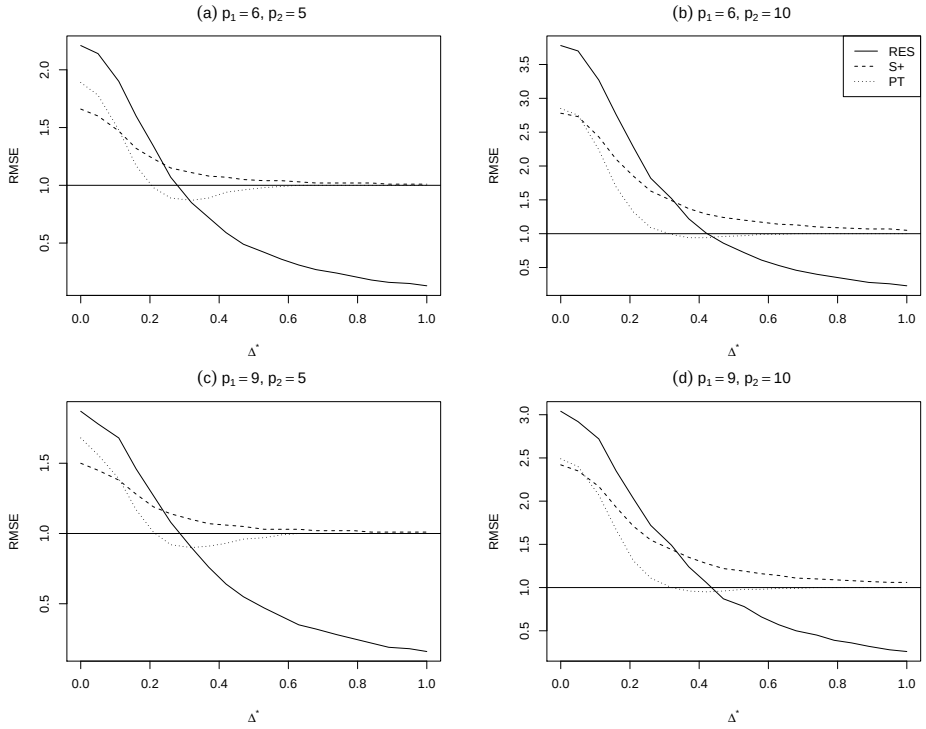


Figure 2.1: Relative mean squared error for restricted, positive-shrinkage, and pretest estimators for $n = 50$, and $(p_1, p_2) = (6, 5)$, $(6, 10)$, $(9, 5)$, $(9, 10)$.

**Case 1:** $\Delta^* = 0$

Clearly, for $\Delta^* = 0$, the RE outperforms all other estimators for all the cases considered in the simulation study.

Table 2.7: Simulated relative mean squared error for restricted, positive-shrinkage, and pretest estimators with respect to unrestricted estimator for $p_1 = 6$, and $p_2 = 10$ for different $\Delta^*$ when $n = 50$.

| $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}}$ |
|---|---|---|---|
| 0.00 | 3.78 | 2.78 | 2.85 |
| 0.05 | 3.70 | 2.73 | 2.75 |
| 0.11 | 3.27 | 2.43 | 2.24 |
| 0.16 | 2.76 | 2.10 | 1.69 |
| 0.21 | 2.28 | 1.85 | 1.32 |
| 0.26 | 1.82 | 1.63 | 1.09 |
| 0.32 | 1.52 | 1.49 | 0.99 |
| 0.37 | 1.22 | 1.37 | 0.94 |
| 0.42 | 1.01 | 1.29 | 0.94 |
| 0.47 | 0.86 | 1.24 | 0.96 |
| 0.53 | 0.72 | 1.20 | 0.97 |
| 0.58 | 0.61 | 1.17 | 0.99 |
| 0.63 | 0.53 | 1.14 | 0.99 |
| 0.68 | 0.46 | 1.13 | 1.00 |
| 0.74 | 0.40 | 1.10 | 1.00 |
| 0.79 | 0.36 | 1.09 | 1.00 |
| 0.84 | 0.32 | 1.08 | 1.00 |
| 0.89 | 0.28 | 1.07 | 1.00 |
| 0.95 | 0.26 | 1.07 | 1.00 |
| 1.00 | 0.23 | 1.05 | 1.00 |

**Case 2:** $\Delta^* > 0$

As the restriction moves away from $\Delta^* = 0$, the RMSE of RE (solid line in Figure 2.1) decays sharply and goes below the horizontal line at RMSE=1. The RMSE of PSE, represented by dashed line in Figure 2.1, approaches 1 at the slowest rate (for a range of $\Delta^*$) as we move away from $\Delta^* = 0$. This indicates that in the event of imprecise subspace information (i.e., even if $\boldsymbol{\beta}_2 \neq \mathbf{0}$), PSE has the smallest quadratic risk among all other estimators for a range of $\Delta^*$. PTE (dotted line in Figure 2.1) outshines PSE when $\Delta^*$ is in the neighborhood of zero. For $\Delta^* > 0$, the RMSE of

PTE becomes inferior to the UE at a faster rate than that of the RE. However, with the increase of $\Delta^*$, at some point, RMSE of PTE approaches 1 from below. This phenomenon suggests that neither PTE nor RE is uniformly better than the other when $\Delta^* > 0$. This is consistent with the theoretical results available in the literature for these estimators (Ahmed, 2012).

Figure 2.1 suggests that PSE maintains its superiority over the RE and PTE for a wide range of $\Delta^*$. For example, when $(p_1, p_2) = (6, 5)$, the risk performance of RE and PTE are superior to that of PSE when $\Delta^* = 0$. However, as $\Delta^*$ moves slightly away from 0, PTE becomes inferior to PSE. PSE dominates RE for $\Delta^*$ around 0.25. After this point RE is not useful while the RMSE of PSE continues to remain over 1 [panels a) and c) in Figure 2.1]. Therefore, PSE might be a preferred estimator since there always remains uncertainty in model-specification. Here $\Delta^*$ measures the degree of deviation from the underlying hypothesis regarding the sub-model, and we see that one cannot go wrong with the PSE even if the assumed model is wrong. In that case, the estimates are as good as the UE in terms of risk.

## 2.5 Comparing Shrinkage and APEs

Our motivation to compare shrinkage and APE came from the work of Hastie et al. (2009) who analyzed the prostate data (Stamey et al., 1989) to evaluate the performance of lasso and some other model selection criteria.

Shrinkage estimators shrink the full model estimates towards the restricted model estimates. In contrast, APEs, for example, the lasso, shrink the ordinary least squares (OLS) estimator towards zero and depending on the value of the tuning parameter $\lambda$,

it sets some coefficients to exactly zero. The output of the lasso resembles shrinkage and pretest methods as it shrinks and selects the variables simultaneously. However, lasso does variable selection automatically by treating all the variables equally. Lasso does not single out the nuisance covariates, or for that matter, the NSI, for special scrutiny as to their usefulness in estimating the main coefficients.

When we have prior information about certain covariates, shrinkage estimators are directly obtained by combining the full and sub-model estimates. On the other hand, if *a priori* information is not available, shrinkage estimation takes a two-step approach in obtaining the estimates. In the first step, a set of covariates are selected based on a suitable model selection criterion such as AIC, BIC or best subset selection. Consequently, the remaining covariates become nuisance, causing a parametric restriction on the full model. In the second step, full and sub-model estimates are combined in a way that minimizes the quadratic risk.

Therefore, it is worth exploring the performance of APEs and shrinkage estimators when it is suspected *a priori* that the parameters may be reduced to a subspace. The following section is dedicated to comparative study of shrinkage and APEs.

In the following, we use the `prostate` data that was used by Hastie et al. (2009) and obtain the shrinkage estimators, and compare their performance with the absolute penalty estimators.

First, we present the prostate data.

### 2.5.1 Prostate Data

Stamey et al. (1989) studied the correlation between the level of prostate specific

antigen (PSA), and a number of clinical measures in men who were about to receive radical prostatectomy. The data consist of 97 measurements on the following variables: log cancer volume (`lcavol`), log prostate weight (`lweight`), age (`age`), log of benign prostatic hyperplasia amount (`lbph`), log of capsular penetration (`lcp`), seminal vesicle invasion (`svi`), Gleason score (`gleason`), and percent of Gleason scores 4 or 5 (`pgg45`).

The idea is to predict log of PSA (`lpsa`) from these measured variables.

### 2.5.2 Predictive Models for Prostate Data

Hastie et al. (2009) demonstrated various model selection techniques by fitting linear regression models to the prostate data. We fit linear regression model to this data, and apply the shrinkage estimation method to obtain positive shrinkage estimates of the regression parameters. We then obtain prediction accuracy of the model by computing cross validation errors, and compare the same with those obtained by the lasso. The predictors were first standardized to have zero mean and unit standard deviation before fitting the model. The correlation table, and the estimated coefficients of linear regression model are available in Hastie et al. (2009, page 50). In their analysis, data were randomly divided into a training and a test part. Several model selection and shrinkage methods such as, OLS, best subset selection (BSS), ridge regression, principal component regression (PCR), partial least squares (PLS), and the lasso, were employed on the training data, and the resulting models were used to predict the outcomes in the test data to obtain prediction errors. Results can be found in Hastie et al. (2009, Table 3.3, page 63). Of the six methods that were used, only best subset selection and lasso methods set some of the coefficients to zero. Best subset

selection gives a model with only `lcavol` and `lweight`, while lasso returns `lcavol`, `lweight`, `lbph`, and `svi` as the best covariates to be included in the model. Since the variables that were dropped were not significantly contributing to the overall fit of the model, we take them as our *p*rior information, and incorporate them in the shrinkage estimation by setting them as restrictions on the full model. In addition to the best subset selection method and lasso, we obtain sub-models based on AIC and BIC for the same data set. The sub-models along with the full-model are listed in Table 2.8. Subsequent calculation of shrinkage and positive-shrinkage estimates uses these four sub-models.

Table 2.8: Full and candidate sub-models for prostate data.

| Selection Criterion | Model: Response ~ Covariates |
|---|---|
| Full Model | `lpsa~ lcavol + lweight + svi + lbph + age + lcp + gleason + pgg45` |
| AIC | `lpsa~ lcavol + lweight + svi + lbph + age` |
| BIC | `lpsa~ lcavol + lweight + svi` |
| BSS | `lpsa~ lcavol + lweight` |
| lasso | `lpsa~ lcavol + lweight + svi + lbph` |

We compute several sets of positive-shrinkage estimates using the sub-models listed in Table 2.8. The model performance is evaluated by computing the prediction error based on $K$-fold cross validation. We consider $K = 5, 10$. In a similar fashion, separate lasso estimates are obtained. For lasso, the tuning parameter is chosen to minimize an estimate of the prediction error based on five- and ten-fold cross validation. Both raw and bias corrected cross validation estimate of prediction error are considered.

We compute adaptive lasso estimates for the prostate data. The advantage of adaptive lasso over the lasso is that it has the oracle property–"it performs as well as if the

true underlying model were given in advance"(Zou, 2006). We use `parcor` R-package (Kraemer and Schaefer, 2010) to obtain adaptive lasso estimates. The software calculates the weights for adaptive lasso by fitting a lasso, where the optimal value of the penalty term is selected via $K$-fold cross-validation. This is a computationally intensive method in which the lasso solutions are computed $K * K$ times.

We also estimate the regression parameters using SCAD penalty. Breheny and Huang (2011) have implemented the SCAD algorithm in their R-package `ncvreg`. In our analysis, we use this package to obtain the SCAD estimates.

Table 2.9 shows average prediction errors and their standard deviations for different shrinkage and absolute penalty estimators based on $K$-fold cross validation repeated 2000 times. We compute four positive shrinkage estimators based on sub-models returned by BSS, AIC, BIC, and lasso. For the purpose of comparison, we first obtain lasso, adaptive lasso and SCAD estimators. Then shrinkage estimators are obtained based on the sub-models given by AIC, BIC, BSS, and lasso. Prediction errors are obtained for each of the cases using 10-fold cross validation.

Table 2.9: Average prediction errors for various models based on $K$-fold cross validation repeated 2000 times for prostate data. Numbers in smaller font are the corresponding standard errors of the prediction errors.

| | Raw CVE | | Bias Corrected CVE | |
|---|---|---|---|---|
| Estimator | $K = 5$ | $K = 10$ | $K = 5$ | $K = 10$ |
| Lasso | $.571_{.030}$ | $.569_{.021}$ | $.565_{.027}$ | $.564_{.018}$ |
| alasso | $.562_{.029}$ | $.557_{.022}$ | $.559_{.026}$ | $.552_{.016}$ |
| SCAD | $.588_{.044}$ | $.563_{.031}$ | $.584_{.043}$ | $.560_{.026}$ |
| PSE(AIC) | $.553_{.029}$ | $.546_{.020}$ | $.549_{.027}$ | $.541_{.016}$ |
| PSE(BIC) | $.559_{.031}$ | $.553_{.021}$ | $.555_{.028}$ | $.547_{.016}$ |
| PSE(BSS) | $.551_{.028}$ | $.549_{.019}$ | $.547_{.025}$ | $.544_{.017}$ |
| PSE(Lasso) | $.546_{025}$ | $.548_{019}$ | $.546_{025}$ | $.543_{016}$ |

The procedure for shrinkage estimation is as follows:

Step 1 Select a candidate sub-model using a suitable model selection criterion, or use prior information to make up the sub-model.

Step 2 Obtain full and sub-model estimates

Step 3 Combine the full and sub-model estimates to obtain shrinkage estimates.

From Table 2.9 we see that the shrinkage estimators have the smallest average prediction error compared to each of the absolute penalty estimators. Recall that the best subset selection produced the smallest sub-model with two covariates followed by BIC, which produced a sub-model with three covariates.

In an attempt to visually display the comparison between shrinkage and APEs, we plot the first fifty values of the average prediction errors for the selected models in Figure 2.2. Since plotting the values for 2000 repetitions would make the comparison difficult to visualize, we plot only the first 50 values.

Figure 2.2 a) shows no striking difference in average prediction errors when prediction errors based on PSE(AIC) and PSE(BIC) models are compared. Figure 2.2 b) shows comparison between PSE(BIC) and the lasso models. Similarly, PSE(BIC) is compared with adaptive lasso and shown in panel c), and with SCAD in panel d) of Figure 2.2. Clearly, shrinkage estimators have smaller overall prediction errors compared to the APEs. It is to be noted that AIC model is larger than the lasso model. The analyses demonstrate that the positive shrinkage estimators minimize the overall risk when we have some prior information about some of the covariates.

In the following section, we conduct Monte Carlo simulation for further investigation.

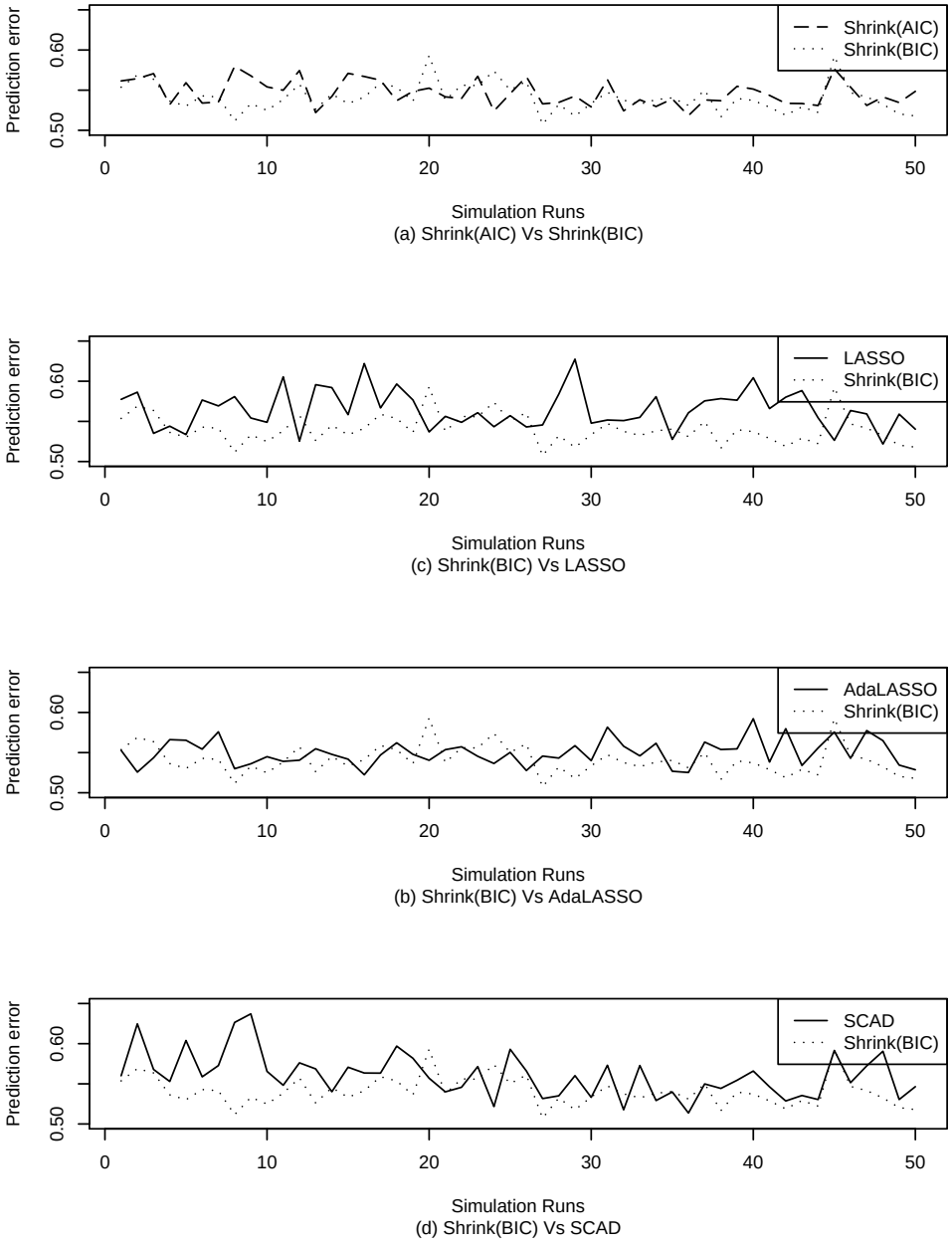


Figure 2.2: Comparison of average prediction error using 10-fold cross validation (first 50 values only) for some positive-shrinkage, lasso, adaptive lasso, and SCAD estimators.

### 2.5.3 Simulation Study: Shrinkage Vs APEs

We perform Monte Carlo simulation experiments to examine the quadratic risk performance of shrinkage estimators with those of APEs. We simulate data from model (2.33) that was previously used in this chapter.

We partition the regression coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$, and consider $\boldsymbol{\beta}_1 = (1, 1, 1, 1)$.

The risk performance of an estimator of $\boldsymbol{\beta}_1$ is measured by calculating its mean squared error (MSE). After calculating the MSEs, we compute efficiencies of the estimators $\hat{\boldsymbol{\beta}}_1^{\text{RE}}$, $\hat{\boldsymbol{\beta}}_1^{\text{S}}$, $\hat{\boldsymbol{\beta}}_1^{\text{S+}}$, $\hat{\boldsymbol{\beta}}^{\text{lasso}}$, $\hat{\boldsymbol{\beta}}^{\text{alasso}}$, and $\hat{\boldsymbol{\beta}}^{\text{SCAD}}$ relative to the unrestricted estimator $\hat{\boldsymbol{\beta}}_1^{\text{UE}}$ using the relative mean squared error (RMSE) criterion, given by

$$\text{RMSE}(\hat{\boldsymbol{\beta}}_1^{\text{UE}} : \hat{\boldsymbol{\beta}}_1^{*}) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{\text{UE}})}{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{*})}. \tag{2.35}$$

Here, $\hat{\boldsymbol{\beta}}_1^{*}$ is one of the shrinkage and APEs. The amount by which an RMSE is larger than unity indicates the degree of superiority of the estimator $\hat{\boldsymbol{\beta}}_1^{*}$ over $\hat{\boldsymbol{\beta}}_1^{\text{UE}}$.

We simulate for $n = 30, 50, 100, 125$, $p_1 = 4, 6, 10$, and $p_2 = 5,\ 9,\ 15$. RMSEs are calculated and are presented in Tables 2.12-2.22 for different values of $\Delta^*$. Table 2.10 summarizes the RMSEs of the estimators when $\Delta^* = 0$.

The tuning parameters for the APE are obtained via cross validation. Ahmed et al. (2007) was the first to compare the shrinkage estimators with an APE (the lasso) in a partially linear regression setup. They compared when $\Delta^* = 0$ only, arguing that APE does not take into consideration that the regression coefficient $\boldsymbol{\beta}$ is partitioned into main and nuisance parts. However, comparison in the classical linear model is not

available in the reviewed literature. Further, in this study, we extend the comparison by adding adaptive lasso and the SCAD penalty estimators in the picture.

**Discussion: Shrinkage Vs APEs**

We compare RMSE of shrinkage and APE for both $\Delta^* = 0$ and $\Delta^* > 0$. Let us compare their performance separately.

**Case 1:** $\Delta^* = 0$

Figure 2.3 shows relative efficiencies of the PSE, $\hat{\boldsymbol{\beta}}_1^{S+}$, and the APE with respect to the UE. Clearly, for $\Delta^* = 0$, the restricted estimator outperforms all other estimators for all the cases considered in this study. Under this condition, $\hat{\boldsymbol{\beta}}_1^{S+}$ outperforms all the APEs. Table 2.10 lists the RMSE of the estimators for $p_1 = 4$, $p_2 = 5, 9, 15$, and $n = 30$, 50, 100, and 125.

**Case 2:** $\Delta^* > 0$

As the restriction moves away from $\Delta^* = 0$, RMSE of the restricted estimator sharply goes below 1. The RMSE of PSE approaches 1 at the slowest rate (for a range of $\Delta^*$) as we move away from $\Delta^* = 0$. This indicates that in the event of imprecise subspace information (i.e., even if $\boldsymbol{\beta}_2 \neq \mathbf{0}$), $\hat{\boldsymbol{\beta}}_1^{S+}$ has the smallest quadratic risk among all other estimators for a range of $\Delta^*$.

Our simulation results suggest that shrinkage and positive shrinkage estimators maintain their superiority over the restricted estimators for a wide range of $\Delta^*$. However, when compared to lasso, alasso and SCAD estimators, the scenario changes when

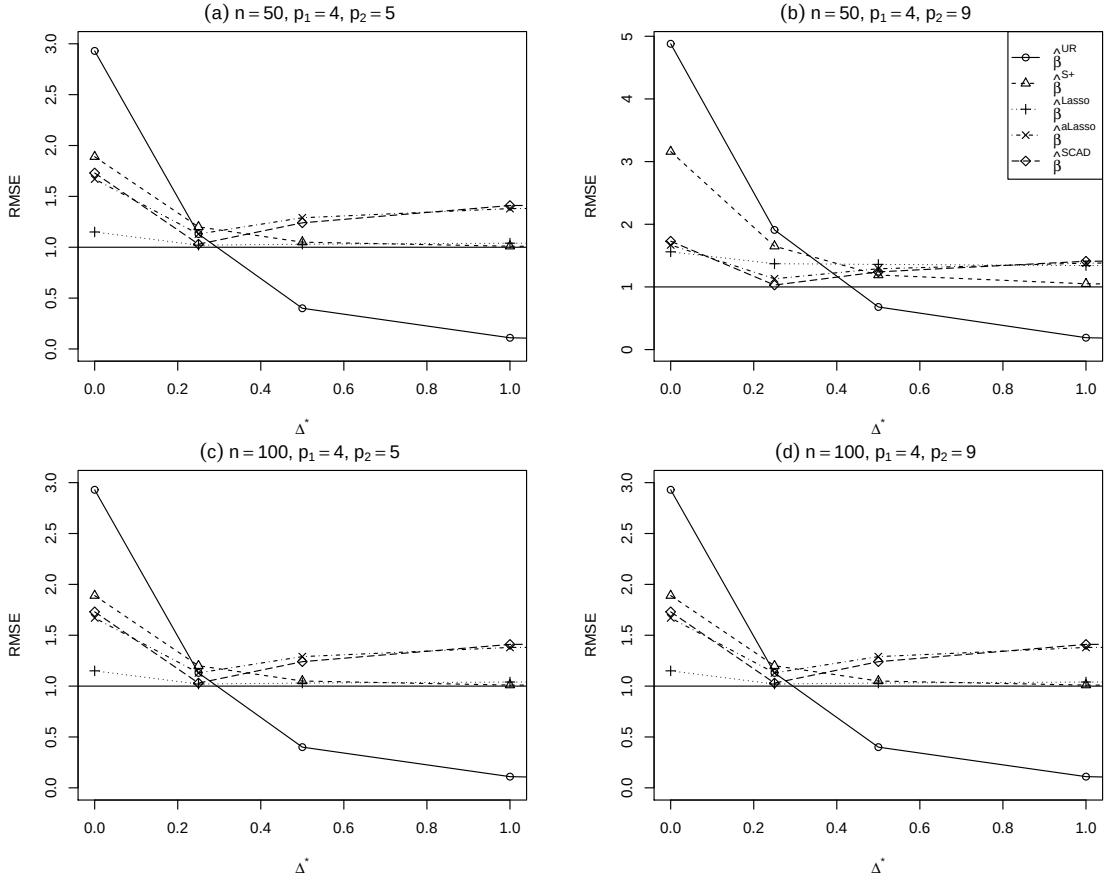


Figure 2.3: Relative efficiency as measured by RMSE criterion for positive shrinkage, lasso, adaptive lasso, and SCAD estimators for different $\Delta^*$, $n$, $p_1$, and $p_2$. A value larger than unity (the horizontal line on the $y$-axis) indicates superiority of the estimator compared to the unrestricted estimator.

we deviate considerably from $\Delta^* = 0$. At some point in the range of $\Delta^*$, adaptive lasso and the SCAD estimators show improved RMSE compared to the unrestricted estimators. Notice the upward-going curves for adaptive lasso and SCAD estimators when $\Delta^*$ is around 0.3 in Figure 2.3.

Based on the $\Delta^*$ values that we consider in our study, performance of shrinkage and positive shrinkage estimators are superior for $\Delta^* \leq 0.25$. However, as $\Delta^*$ increases, RMSE of adaptive lasso gets better. The reason for adaptive lasso to perform better than the rest of the estimators under the alternative hypothesis is that an increase in the $\Delta^*$ makes a previously insignificant covariate possibly significant, which was not being considered under the simulation setup. Note that, in our setup we computed the MSE under fixed $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ even though we let $\Delta^*$ vary considerably.

Table 2.10: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\text{UE}}$ for $p_1 = 4$, $\Delta^* = 0$.

| $n$ | $p_2$ | $\hat{\boldsymbol{\beta}}_1^{\text{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\text{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\text{S+}}$ | $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\text{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\text{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 30 | 5 | 3.30 | 1.71 | **1.98** | 1.17 | 1.66 | 1.64 |
| | 9 | 6.20 | 2.85 | **3.49** | 1.66 | 2.71 | 2.75 |
| | 15 | 14.88 | 5.35 | **6.98** | 3.05 | 5.72 | 5.47 |
| 50 | 5 | 2.93 | 1.67 | **1.89** | 1.15 | 1.67 | 1.73 |
| | 9 | 4.88 | 2.61 | **3.16** | 1.56 | 2.57 | 2.81 |
| | 15 | 14.88 | 5.35 | **6.98** | 3.05 | 4.65 | 4.70 |
| 100 | 5 | 2.76 | 1.62 | **1.84** | 1.15 | 1.79 | 1.73 |
| | 9 | 4.27 | 2.52 | **2.93** | 1.51 | 2.62 | 2.77 |
| | 15 | 6.90 | 3.83 | **4.63** | 2.12 | 4.08 | 4.35 |
| 125 | 5 | 2.69 | 1.62 | **1.83** | 1.14 | 1.87 | 1.71 |
| | 9 | 4.14 | 2.44 | **2.88** | 1.50 | 2.70 | 4.30 |
| | 15 | 6.66 | 3.79 | **4.63** | 2.09 | 4.07 | 4.45 |

### 2.5.4 High-dimensional Scenario

We further compare the RMSE of shrinkage and APEs in a high-dimensional scenario when the number of nuisance parameter is very high compared to the number of main parameters. We keep the number of parameters less than the number of observations in order to be able to compute the shrinkage estimators. Simulated RMSE for fixed $n = 110$ and fixed $p_1 = 5$ are presented in Table 2.11 for varying $p_2$. RMSE are also graphically compared in Figure 2.4. This time, only cases with $\Delta^* = 0$ were considered.

We observe that PSE and SCAD estimators perform closely in terms of RMSE for $p_2$ up to 40. As $p_2$ gets larger, adaptive lasso and SCAD outshines PSE. However, PSE continues to dominate lasso for all $p_2$ considered in this study. See Table 2.11.

Table 2.11: Simulated RMSE when $p_2$ is high-dimensional for fixed $n = 110$ and $p_1 = 5$.

| $p_2$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|
| 20 | 7.17 | 4.33 | 5.25 | 2.31 | 4.43 | 5.05 |
| 40 | 17.50 | 9.94 | 12.21 | 4.55 | 10.61 | 12.73 |
| 60 | 37.82 | 17.11 | 22.78 | 8.56 | 21.10 | 26.39 |
| 80 | 92.85 | 27.48 | 41.32 | 18.63 | 51.06 | 66.04 |
| 90 | 163.22 | 37.32 | 58.24 | 33.9 | 89.07 | 104.72 |

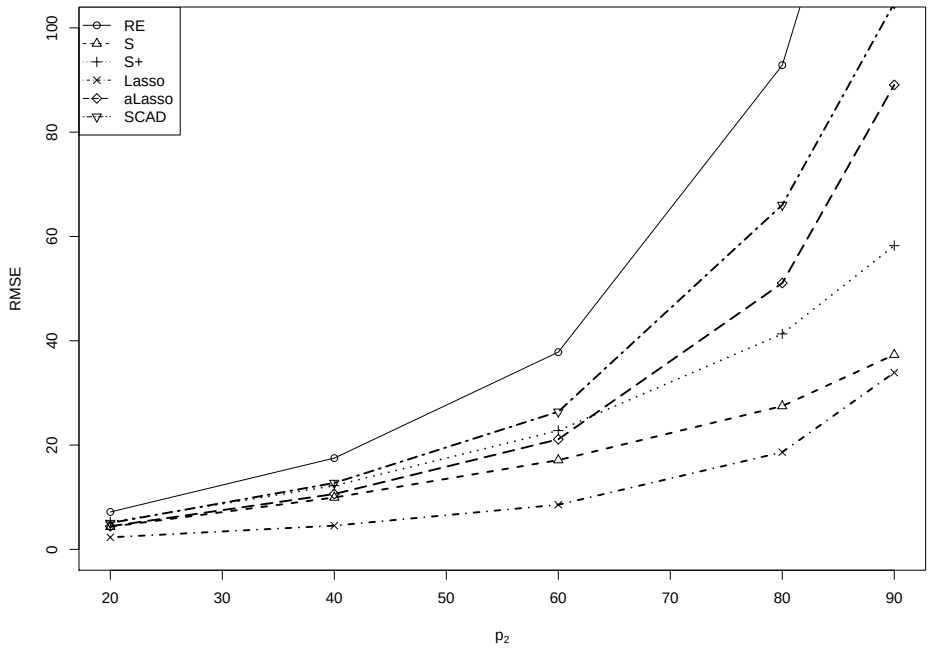


Figure 2.4: Graphical comparison of simulated RMSE for fixed $p_1 = 5$, $n = 110$ when $\Delta^* = 0$.

Table 2.12: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 4$ when $n = 30$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 3.30 | 1.71 | 1.98 | 1.17 | 1.66 | 1.64 |
| | 0.25 | 1.69 | 1.34 | 1.38 | 1.05 | 1.17 | 1.11 |
| | 0.50 | 0.70 | 1.11 | 1.11 | 1.02 | 1.19 | 1.10 |
| | 1.00 | 0.20 | 1.03 | 1.03 | 1.02 | 1.31 | 1.34 |
| | 2.00 | 0.05 | 1.01 | 1.01 | 1.03 | 1.35 | 1.37 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.02 | 1.40 | 1.37 |
| 9 | 0.00 | 6.20 | 2.85 | 3.49 | 1.66 | 2.71 | 2.75 |
| | 0.25 | 3.17 | 2.04 | 2.13 | 1.46 | 1.85 | 1.82 |
| | 0.50 | 1.29 | 1.40 | 1.40 | 1.42 | 1.78 | 1.70 |
| | 1.00 | 0.38 | 1.11 | 1.11 | 1.40 | 2.19 | 2.18 |
| | 2.00 | 0.10 | 1.03 | 1.03 | 1.42 | 2.20 | 2.29 |
| | 4.00 | 0.03 | 1.01 | 1.01 | 1.44 | 2.24 | 2.34 |
| 15 | 0.00 | 14.88 | 5.35 | 6.98 | 3.05 | 5.72 | 5.47 |
| | 0.25 | 7.77 | 3.64 | 3.87 | 2.73 | 3.85 | 3.75 |
| | 0.50 | 3.09 | 2.03 | 2.04 | 2.55 | 3.67 | 3.48 |
| | 1.00 | 0.93 | 1.29 | 1.29 | 2.48 | 4.32 | 4.50 |
| | 2.00 | 0.24 | 1.07 | 1.07 | 2.49 | 4.26 | 4.61 |
| | 4.00 | 0.06 | 1.02 | 1.02 | 2.46 | 4.65 | 4.33 |

Table 2.13: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 4$ when $n = 50$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 2.93 | 1.67 | 1.89 | 1.15 | 1.67 | 1.73 |
| | 0.25 | 1.13 | 1.18 | 1.20 | 1.02 | 1.13 | 1.03 |
| | 0.50 | 0.40 | 1.05 | 1.05 | 1.03 | 1.29 | 1.24 |
| | 1.00 | 0.11 | 1.01 | 1.01 | 1.04 | 1.38 | 1.41 |
| | 2.00 | 0.03 | 1.00 | 1.00 | 1.04 | 1.40 | 1.43 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.04 | 1.43 | 1.43 |
| 9 | 0.00 | 4.88 | 2.61 | 3.16 | 1.56 | 2.57 | 2.81 |
| | 0.25 | 1.91 | 1.62 | 1.65 | 1.37 | 1.72 | 1.59 |
| | 0.50 | 0.68 | 1.19 | 1.19 | 1.36 | 1.91 | 1.84 |
| | 1.00 | 0.19 | 1.05 | 1.05 | 1.34 | 2.11 | 2.25 |
| | 2.00 | 0.05 | 1.01 | 1.01 | 1.35 | 2.10 | 2.31 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.37 | 2.27 | 2.35 |
| 15 | 0.00 | 8.59 | 4.29 | 5.38 | 2.36 | 4.65 | 4.70 |
| | 0.25 | 3.44 | 2.43 | 2.49 | 2.00 | 2.77 | 2.67 |
| | 0.50 | 1.23 | 1.47 | 1.47 | 1.98 | 3.16 | 3.09 |
| | 1.00 | 0.34 | 1.12 | 1.12 | 1.96 | 3.48 | 4.14 |
| | 2.00 | 0.09 | 1.03 | 1.03 | 2.00 | 3.48 | 4.03 |
| | 4.00 | 0.02 | 1.01 | 1.01 | 1.97 | 3.81 | 3.90 |

Table 2.14: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 4$ when $n = 100$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 2.76 | 1.62 | 1.84 | 1.15 | 1.79 | 1.73 |
| | 0.25 | 0.66 | 1.09 | 1.09 | 1.03 | 1.28 | 1.09 |
| | 0.50 | 0.20 | 1.02 | 1.02 | 1.03 | 1.44 | 1.40 |
| | 1.00 | 0.05 | 1.01 | 1.01 | 1.04 | 1.48 | 1.43 |
| | 2.00 | 0.01 | 1.00 | 1.00 | 1.04 | 1.56 | 1.42 |
| | 4.00 | 0.00 | 1.00 | 1.00 | 1.05 | 1.63 | 1.40 |
| 9 | 0.00 | 4.27 | 2.52 | 2.93 | 1.51 | 2.62 | 2.77 |
| | 0.25 | 1.04 | 1.31 | 1.31 | 1.31 | 1.82 | 1.52 |
| | 0.50 | 0.32 | 1.08 | 1.08 | 1.33 | 2.08 | 2.12 |
| | 1.00 | 0.08 | 1.02 | 1.02 | 1.33 | 2.23 | 2.18 |
| | 2.00 | 0.02 | 1.01 | 1.01 | 1.33 | 2.17 | 2.19 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.32 | 2.38 | 2.19 |
| 15 | 0.00 | 6.90 | 3.83 | 4.63 | 2.12 | 4.08 | 4.35 |
| | 0.25 | 1.65 | 1.70 | 1.71 | 1.80 | 2.69 | 2.39 |
| | 0.50 | 0.50 | 1.20 | 1.20 | 1.81 | 3.25 | 3.50 |
| | 1.00 | 0.13 | 1.05 | 1.05 | 1.82 | 3.17 | 3.66 |
| | 2.00 | 0.03 | 1.01 | 1.01 | 1.81 | 3.47 | 3.67 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.82 | 3.75 | 3.54 |

Table 2.15: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\text{UE}}$ for $p_1 = 4$ when $n = 125$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\text{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\text{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\text{S+}}$ | $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\text{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\text{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 2.69 | 1.62 | 1.83 | 1.14 | 1.87 | 1.71 |
| | 0.25 | 0.55 | 1.07 | 1.07 | 1.04 | 1.33 | 1.14 |
| | 0.50 | 0.16 | 1.02 | 1.02 | 1.03 | 1.47 | 1.40 |
| | 1.00 | 0.04 | 1.00 | 1.00 | 1.03 | 1.50 | 1.42 |
| | 2.00 | 0.01 | 1.00 | 1.00 | 1.03 | 1.53 | 1.42 |
| | 4.00 | 0.00 | 1.00 | 1.00 | 1.06 | 1.61 | 1.41 |
| 9 | 0.00 | 4.14 | 2.44 | 2.88 | 1.50 | 2.70 | 4.30 |
| | 0.25 | 0.84 | 1.24 | 1.24 | 1.30 | 1.81 | 2.21 |
| | 0.50 | 0.25 | 1.06 | 1.06 | 1.31 | 2.11 | 3.19 |
| | 1.00 | 0.06 | 1.02 | 1.02 | 1.32 | 2.25 | 3.69 |
| | 2.00 | 0.02 | 1.00 | 1.00 | 1.31 | 2.30 | 3.53 |
| | 4.00 | 0.00 | 1.00 | 1.00 | 1.33 | 2.40 | 3.54 |
| 15 | 0.00 | 6.66 | 3.79 | 4.63 | 2.09 | 4.07 | 4.45 |
| | 0.25 | 1.33 | 1.56 | 1.56 | 1.79 | 2.68 | 2.48 |
| | 0.50 | 0.39 | 1.16 | 1.16 | 1.78 | 3.14 | 3.41 |
| | 1.00 | 0.10 | 1.04 | 1.04 | 1.79 | 3.44 | 3.54 |
| | 2.00 | 0.03 | 1.01 | 1.01 | 1.78 | 3.50 | 3.61 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.79 | 3.66 | 3.49 |

Table 2.16: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\text{UE}}$ for $p_1 = 6$ when $n = 30$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\text{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\text{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\text{S+}}$ | $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\text{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\text{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 2.26 | 1.51 | 1.66 | 1.07 | 1.64 | 1.41 |
| | 0.25 | 1.13 | 1.16 | 1.17 | 1.00 | 0.98 | 1.05 |
| | 0.50 | 0.46 | 1.05 | 1.05 | 1.00 | 1.16 | 1.05 |
| | 1.00 | 0.13 | 1.01 | 1.01 | 1.00 | 1.48 | 1.23 |
| | 2.00 | 0.04 | 1.00 | 1.00 | 1.00 | 1.28 | 1.28 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.01 | 1.27 | 1.26 |
| 9 | 0.00 | 3.52 | 2.26 | 2.60 | 1.36 | 2.06 | 2.20 |
| | 0.25 | 1.74 | 1.55 | 1.57 | 1.23 | 1.70 | 1.64 |
| | 0.50 | 0.71 | 1.18 | 1.18 | 1.22 | 1.72 | 1.51 |
| | 1.00 | 0.21 | 1.05 | 1.05 | 1.22 | 1.79 | 1.83 |
| | 2.00 | 0.05 | 1.01 | 1.01 | 1.23 | 1.88 | 1.91 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.24 | 1.88 | 1.99 |
| 15 | 0.00 | 6.11 | 3.67 | 4.35 | 1.94 | 4.32 | 4.41 |
| | 0.25 | 3.05 | 2.29 | 2.34 | 1.72 | 3.48 | 3.21 |
| | 0.50 | 1.23 | 1.45 | 1.45 | 1.69 | 3.24 | 2.90 |
| | 1.00 | 0.36 | 1.12 | 1.12 | 1.71 | 3.66 | 3.35 |
| | 2.00 | 0.09 | 1.03 | 1.03 | 1.70 | 3.78 | 3.53 |
| | 4.00 | 0.02 | 1.01 | 1.01 | 1.71 | 3.80 | 3.69 |

Table 2.17: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 6$ when $n = 50$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 2.56 | 1.59 | 1.79 | 1.07 | 1.41 | 1.53 |
| | 0.25 | 1.59 | 1.30 | 1.33 | 0.99 | 1.09 | 1.03 |
| | 0.50 | 0.74 | 1.10 | 1.11 | 0.97 | 1.21 | 1.20 |
| | 1.00 | 0.24 | 1.03 | 1.03 | 0.98 | 1.23 | 1.30 |
| | 2.00 | 0.06 | 1.01 | 1.01 | 0.98 | 1.26 | 1.32 |
| | 4.00 | 0.02 | 1.00 | 1.00 | 1.00 | 1.33 | 1.34 |
| 9 | 0.00 | 4.61 | 2.54 | 3.02 | 1.43 | 2.18 | 2.33 |
| | 0.25 | 2.87 | 1.95 | 2.04 | 1.32 | 1.57 | 1.47 |
| | 0.50 | 1.31 | 1.39 | 1.40 | 1.27 | 1.72 | 1.73 |
| | 1.00 | 0.43 | 1.11 | 1.11 | 1.28 | 1.80 | 1.98 |
| | 2.00 | 0.11 | 1.03 | 1.03 | 1.30 | 1.85 | 2.01 |
| | 4.00 | 0.03 | 1.01 | 1.01 | 1.29 | 1.82 | 2.01 |
| 15 | 0.00 | 11.51 | 4.88 | 6.26 | 2.51 | 3.42 | 3.97 |
| | 0.25 | 7.04 | 3.49 | 3.77 | 2.30 | 2.39 | 2.48 |
| | 0.50 | 3.37 | 2.10 | 2.11 | 2.23 | 2.75 | 2.72 |
| | 1.00 | 1.07 | 1.31 | 1.31 | 2.21 | 2.91 | 3.47 |
| | 2.00 | 0.29 | 1.08 | 1.08 | 2.17 | 3.01 | 3.48 |
| | 4.00 | 0.07 | 1.02 | 1.02 | 2.24 | 3.05 | 3.39 |

Table 2.18: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 6$ when $n = 100$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 2.10 | 1.46 | 1.60 | 1.07 | 1.53 | 1.52 |
| | 0.25 | 0.70 | 1.07 | 1.07 | 1.00 | 1.19 | 1.06 |
| | 0.50 | 0.24 | 1.02 | 1.02 | 1.01 | 1.28 | 1.30 |
| | 1.00 | 0.06 | 1.00 | 1.00 | 1.01 | 1.33 | 1.32 |
| | 2.00 | 0.02 | 1.00 | 1.00 | 1.01 | 1.34 | 1.32 |
| | 4.00 | 0.00 | 1.00 | 1.00 | 1.01 | 1.39 | 1.30 |
| 9 | 0.00 | 3.10 | 2.14 | 2.38 | 1.31 | 2.07 | 2.22 |
| | 0.25 | 1.03 | 1.26 | 1.26 | 1.20 | 1.61 | 1.40 |
| | 0.50 | 0.34 | 1.07 | 1.07 | 1.20 | 1.81 | 1.86 |
| | 1.00 | 0.09 | 1.02 | 1.02 | 1.20 | 1.85 | 1.93 |
| | 2.00 | 0.02 | 1.00 | 1.00 | 1.21 | 1.89 | 1.93 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.21 | 1.95 | 1.87 |
| 15 | 0.00 | 4.80 | 3.19 | 3.66 | 1.75 | 3.14 | 3.45 |
| | 0.25 | 1.59 | 1.63 | 1.64 | 1.55 | 2.25 | 2.13 |
| | 0.50 | 0.53 | 1.18 | 1.18 | 1.57 | 2.59 | 2.83 |
| | 1.00 | 0.14 | 1.05 | 1.05 | 1.56 | 2.70 | 2.98 |
| | 2.00 | 0.04 | 1.01 | 1.01 | 1.57 | 2.79 | 2.96 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.57 | 2.91 | 2.96 |

Table 2.19: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 6$ when $n = 125$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 2.08 | 1.45 | 1.60 | 1.07 | 1.58 | 1.50 |
| | 0.25 | 0.60 | 1.06 | 1.06 | 1.01 | 1.22 | 1.11 |
| | 0.50 | 0.19 | 1.02 | 1.02 | 1.00 | 1.32 | 1.29 |
| | 1.00 | 0.05 | 1.00 | 1.00 | 1.00 | 1.36 | 1.32 |
| | 2.00 | 0.01 | 1.00 | 1.00 | 1.01 | 1.39 | 1.31 |
| | 4.00 | 0.00 | 1.00 | 1.00 | 1.02 | 1.41 | 1.30 |
| 9 | 0.00 | 2.99 | 2.08 | 2.33 | 1.31 | 2.17 | 2.17 |
| | 0.25 | 0.86 | 1.21 | 1.21 | 1.20 | 1.64 | 1.48 |
| | 0.50 | 0.28 | 1.06 | 1.06 | 1.20 | 1.83 | 1.88 |
| | 1.00 | 0.08 | 1.01 | 1.01 | 1.19 | 1.84 | 1.91 |
| | 2.00 | 0.02 | 1.00 | 1.00 | 1.20 | 1.90 | 1.89 |
| | 4.00 | 0.00 | 1.00 | 1.00 | 1.22 | 1.97 | 1.91 |
| 15 | 0.00 | 4.55 | 3.10 | 3.54 | 1.71 | 3.17 | 3.31 |
| | 0.25 | 1.31 | 1.50 | 1.50 | 1.54 | 2.31 | 2.17 |
| | 0.50 | 0.42 | 1.14 | 1.14 | 1.54 | 2.57 | 2.84 |
| | 1.00 | 0.11 | 1.04 | 1.04 | 1.54 | 2.63 | 2.90 |
| | 2.00 | 0.03 | 1.01 | 1.01 | 1.54 | 2.75 | 2.95 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.56 | 2.96 | 2.84 |

Table 2.20: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 10$ when $n = 50$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 1.78 | 1.37 | 1.46 | 1.01 | 1.24 | 1.35 |
| | 0.25 | 1.12 | 1.14 | 1.15 | 0.97 | 1.06 | 1.03 |
| | 0.50 | 0.53 | 1.04 | 1.04 | 0.97 | 1.12 | 1.13 |
| | 1.00 | 0.17 | 1.01 | 1.01 | 0.96 | 1.15 | 1.21 |
| | 2.00 | 0.05 | 1.00 | 1.00 | 0.97 | 1.13 | 1.22 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 0.97 | 1.16 | 1.22 |
| 9 | 0.00 | 2.61 | 1.91 | 2.12 | 1.18 | 1.69 | 1.92 |
| | 0.25 | 1.63 | 1.46 | 1.48 | 1.11 | 1.41 | 1.37 |
| | 0.50 | 0.77 | 1.17 | 1.17 | 1.11 | 1.45 | 1.54 |
| | 1.00 | 0.25 | 1.05 | 1.05 | 1.11 | 1.50 | 1.74 |
| | 2.00 | 0.07 | 1.01 | 1.01 | 1.11 | 1.53 | 1.70 |
| | 4.00 | 0.02 | 1.00 | 1.00 | 1.11 | 1.54 | 1.71 |
| 15 | 0.00 | 4.37 | 3.05 | 3.45 | 1.56 | 2.65 | 3.05 |
| | 0.25 | 2.74 | 2.15 | 2.20 | 1.47 | 2.12 | 2.25 |
| | 0.50 | 1.28 | 1.45 | 1.45 | 1.45 | 2.10 | 2.44 |
| | 1.00 | 0.41 | 1.13 | 1.13 | 1.45 | 2.40 | 2.94 |
| | 2.00 | 0.11 | 1.03 | 1.03 | 1.45 | 2.42 | 2.84 |
| | 4.00 | 0.03 | 1.01 | 1.01 | 1.45 | 2.36 | 2.90 |

Table 2.21: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\text{UE}}$ for $p_1 = 10$ when $n = 100$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\text{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\text{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\text{S+}}$ | $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\text{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\text{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 1.65 | 1.32 | 1.39 | 1.01 | 1.31 | 1.32 |
| | 0.25 | 0.77 | 1.06 | 1.06 | 0.98 | 1.12 | 1.02 |
| | 0.50 | 0.30 | 1.01 | 1.01 | 0.98 | 1.18 | 1.18 |
| | 1.00 | 0.09 | 1.00 | 1.00 | 0.98 | 1.20 | 1.20 |
| | 2.00 | 0.02 | 1.00 | 1.00 | 0.98 | 1.21 | 1.20 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 0.99 | 1.21 | 1.19 |
| 9 | 0.00 | 2.23 | 1.75 | 1.90 | 1.16 | 1.69 | 1.78 |
| | 0.25 | 1.04 | 1.21 | 1.21 | 1.10 | 1.41 | 1.28 |
| | 0.50 | 0.40 | 1.06 | 1.06 | 1.10 | 1.50 | 1.59 |
| | 1.00 | 0.12 | 1.02 | 1.02 | 1.10 | 1.53 | 1.59 |
| | 2.00 | 0.03 | 1.00 | 1.00 | 1.10 | 1.54 | 1.62 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.10 | 1.56 | 1.60 |
| 15 | 0.00 | 3.20 | 2.48 | 2.73 | 1.42 | 2.33 | 2.59 |
| | 0.25 | 1.50 | 1.53 | 1.53 | 1.33 | 1.88 | 1.83 |
| | 0.50 | 0.58 | 1.16 | 1.16 | 1.33 | 2.07 | 2.25 |
| | 1.00 | 0.17 | 1.04 | 1.04 | 1.34 | 2.12 | 2.37 |
| | 2.00 | 0.04 | 1.01 | 1.01 | 1.33 | 2.17 | 2.36 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.33 | 2.21 | 2.36 |

Table 2.22: Simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 10$ when $n = 125$.

| $p_2$ | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{alasso}}$ | $\hat{\boldsymbol{\beta}}^{\mathrm{SCAD}}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.00 | 1.63 | 1.31 | 1.39 | 1.02 | 1.31 | 1.30 |
| | 0.25 | 0.68 | 1.04 | 1.04 | 0.99 | 1.14 | 1.07 |
| | 0.50 | 0.24 | 1.01 | 1.01 | 0.99 | 1.20 | 1.20 |
| | 1.00 | 0.07 | 1.00 | 1.00 | 0.98 | 1.21 | 1.21 |
| | 2.00 | 0.02 | 1.00 | 1.00 | 0.99 | 1.23 | 1.20 |
| | 4.00 | 0.00 | 1.00 | 1.00 | 0.99 | 1.22 | 1.21 |
| 9 | 0.00 | 2.19 | 1.73 | 1.87 | 1.16 | 1.70 | 1.78 |
| | 0.25 | 0.90 | 1.17 | 1.17 | 1.10 | 1.48 | 1.34 |
| | 0.50 | 0.33 | 1.05 | 1.05 | 1.10 | 1.52 | 1.62 |
| | 1.00 | 0.09 | 1.01 | 1.01 | 1.10 | 1.53 | 1.59 |
| | 2.00 | 0.02 | 1.00 | 1.00 | 1.10 | 1.57 | 1.63 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.11 | 1.60 | 1.58 |
| 15 | 0.00 | 3.05 | 2.40 | 2.64 | 1.41 | 2.30 | 2.50 |
| | 0.25 | 1.28 | 1.42 | 1.42 | 1.32 | 1.99 | 1.81 |
| | 0.50 | 0.47 | 1.13 | 1.13 | 1.33 | 2.09 | 2.28 |
| | 1.00 | 0.13 | 1.03 | 1.03 | 1.32 | 2.10 | 2.30 |
| | 2.00 | 0.03 | 1.01 | 1.01 | 1.33 | 2.12 | 2.34 |
| | 4.00 | 0.01 | 1.00 | 1.00 | 1.32 | 2.17 | 2.29 |

## 2.6 Conclusion

In this chapter, we demonstrated application of shrinkage and pretest estimation in linear models using three real life data sets. We also compared the performance of PSE and APE numerically using the prostate data. Monte Carlo experiments were conducted to study the behavior of shrinkage and APE under various conditions.

In the first half of this chapter (up to Section 2.4), we presented shrinkage and pretest estimation in the context of a multiple linear regression model. To illustrate the methods, three different data sets have been considered to obtain restricted, positive shrinkage, and pretest estimators. Average prediction errors based on repeated cross validation estimate of the error rates show that pretest and restricted estimators have superior risk performance compared to the unrestricted and positive-shrinkage estimators when the underlying model is correctly specified. This is not unusual since the restricted estimator dominates all other estimators when the prior information is correct. Since the data considered in this study have been interactively analyzed using various model selection criteria, it is expected that the sub-models consist of the best subsets of the available covariates for the respective data sets. Theoretically, this is equivalent to the case where $\Delta^* = 0$, or very close to zero. The real data examples considered here, however, do not tell us how sensitive are the prediction errors under model misspecification. Therefore, we conduct Monte Carlo simulation to study the behaviour of PSE and PT estimators under varying $\Delta^*$ and different sizes of the nuisance subsets.

Our study re-established the fact that the restricted estimator outperforms the unrestricted estimator at or near the pivot ($\Delta^* = 0$). However, as we deviate from the pivot ($\Delta^* > 0$), relative risk of the RE becomes higher than that of the UE.

RMSE of PSE decays at the slowest rate with the increase of $\Delta^*$, and performs steadily throughout a wider range of the alternative parameter subspace. When the nuisance subset is large, PSE outperforms all other estimators.

In the second part of this chapter (Section 2.5), we compared shrinkage and APE in the context of a multiple linear regression model. In our study, we developed and implemented a procedure for simultaneous sub-model selection using AIC or BIC to obtain shrinkage estimates of the regression coefficients. Based on a quadratic risk function, we computed RMSE of SE, PSE and several APE such as lasso, adaptive lasso, and SCAD with respect to UE. Asymptotic risk properties of the proposed estimators have been reappraised and their dominance over classical estimators demonstrated. All the computations were programmed with R (R Development Core Team, 2010).

For high-dimensional data, our simulation study reconfirms the dominance of PSE over APEs for moderate to large number of nuisance covariates (Table 2.11). However, SCAD and adaptive lasso outperform PSE when the number of nuisance covariates gets extremely large relative to the sample size.

# Chapter 3

# Shrinkage Estimation in Partially Linear Models

## 3.1 Introduction

A semiparametric model is one which includes both parametric and nonparametric components. Several semiparametric models have been proposed in the literature. Partially linear model, semiparametric single index models, and varying coefficient models are among the popular ones.

A partially linear model (PLM) is a semiparametric regression model of the form

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + g(t_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{3.1}$$

where $y_i$'s are responses, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ and $t_i \in [0, 1]$ are design points, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown parameter vector, $g(\cdot)$ is an unknown bounded real-valued

function defined on $[0, 1]$, and $\varepsilon_i$'s are unobservable random errors.

Some earlier surveys of the estimation and application of model (3.1) can be found in the monograph of Härdle et al. (2000). Bunea (2004) suggested a consistent covariate selection technique in a semiparametric regression model through penalized least squared criterion for selection. They showed that the selected estimator of the linear part is asymptotically normal. In finding the strategy of bandwidth selection in kernel-based estimation in semiparametric models, Liang (2006) numerically compared the performance of profile-kernel, penalized spline method, and back-fitting methods for a partially linear model. Real data from a study of relation between log-earnings of an individual, personal characteristics (such as gender), and measures of a person's human capital (such as year of schooling or job experience) were used to compare the three methods. The nonparametric component was estimated as a function of the "local unemployment rate" by smoothing-splines. Sun et al. (2008) considered polynomial spline estimation of partially linear single-index model in the context of proportional hazards regression.

For (3.1), Ahmed et al. (2007) considered a profile least squares approach based on using kernel estimates of $g(\cdot)$ to construct absolute penalty, shrinkage, and pretest estimators of the regression parameter $\boldsymbol{\beta}$. They also studied APE with shrinkage and positive-shrinkage estimators through Monte Carlo simulation.

In this study, we explore the suitability of using B-spline basis function in approximating the nonparametric component since B-splines are easy to compute and they can be incorporated in a regression model. Our motivation came from the work of Engle et al. (1986) who used a PLM in studying the relationship between electricity demand and temperature. We propose shrinkage estimators in a PLM and illustrate

and compare our estimators with lasso and adaptive lasso (alasso) estimators using econometric data. We present the details in Section 3.2.

In this chapter, we consider a PLM (3.1) where the vector of coefficients $\boldsymbol{\beta}$ in the linear part can be partitioned as $(\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$ where $\boldsymbol{\beta_1}$ is the coefficient vector for main effects, and $\boldsymbol{\beta_2}$ is the vector for nuisance effects. For example, in modeling family income, age and education of the wage-earner may be the main effects while the number of kids they have, or parents' years of schooling can be regarded as nuisance variables. In this situation, inference about $\boldsymbol{\beta_1}$ may benefit from moving the least squares estimate for the full model to the direction of the least squares estimate without the nuisance variables (Steinian shrinkage), or from dropping the nuisance variables if there is evidence that they do not provide useful information (through pretesting). In this framework, the shrinkage estimator takes a hybrid approach by shrinking the base estimator to a plausible alternative estimator.

In our case, the sub-vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are assumed to have dimensions $p_1$ and $p_2$ respectively, and $p_1 + p_2 = p$, $p_i \geq 0$ for $i = 1, 2$. We are essentially interested in the estimation of $\boldsymbol{\beta}_1$ when it is plausible that $\boldsymbol{\beta}_2$ is close to zero. This situation may arise when there is over-modeling and one wishes to cut down the irrelevant part (uncertain factors) from the model. Thus, the parameter space can be partitioned and it is plausible that $\boldsymbol{\beta}_2$ is near some specified $\boldsymbol{\beta}_2^o$ which, without loss of generality, may be set to a null vector. In our framework, the nonparametric component is estimated by B-spline basis function.

### 3.1.1 Organization of the Chapter

In the following, we begin with a motivating example using econometric data. A brief description of the data is given. Analysis of data and development of the full model and sub-model is illustrated in detail in Section 3.3. We present the proposed estimators in Section 3.4. Prediction errors and log-likelihood values of the proposed estimators are compared with lasso and adaptive lasso estimators through bootstrap resampling in Section 3.5. In Section 3.6, we design and conduct a Monte Carlo experiment to study the performance of the proposed estimators and compare them with an APE. Asymptotic bias and risk performance of the estimators are presented in Section 3.8. Finally, we make some conclusions and recommendations.

## 3.2 Motivating Example

Engle et al. (1986) studied the relationship between demand for electricity and temperature, and found that electricity demand and temperature are nonlinearly related. Numerous authors have shown applications of PLM in many areas. Most of the applications are, however, in the areas of econometrics or on demographic and socio-economic data. Härdle et al. (2000) mentioned that "well-known applications in econometrics literature that can be put in the form of PLM are the *human capital* earnings function and the wage curve," where log-earnings of an individual were related to sex, marital status, schooling, and labour market experience. It is also suggested in economic theory that log-earnings and labour market experience are nonlinearly related.

In the following, we fit a PLM to an econometric data while estimating the non-

parametric component $g(\cdot)$ using B-spline basis function.

### 3.2.1 Data and Variables

Mroz (1987) used a sample of 1975 Panel Study on Income Dynamics (PSID) labour supply data to systematically study several theoretic and statistical assumptions used in many empirical models of female labour supply. PSID data is freely available from `http://ideas.repec.org/s/boc/bocins.html`

Fox (2002) used these data for a semiparametric logistic regression. Fox (2005) commented that a semiparametric model may be used wherever there is a reason to believe that one or more covariates enter the regression linearly. This could be known from prior studies or there are prior reasons to believe so (although rare), or examination of the data might suggest a linear relationship for some covariates. A more general scenario is when some of the covariates are categorical and they enter in the model as dummy variables.

The female labour supply data consist of 753 observations on 19 variables. Data were collected from married white women between the ages 30 and 60 in 1975. Of them, 428 were working at some time during the year 1975.

Depending on whether the woman was in the labour force during 1975 (`inlf=1`), average and standard deviation (in parenthesis) of the variables are presented in Table 3.1. The variable `nwifeinc` represents nonwife income (in thousands) for the household and is defined as the household's total money (`faminc`) minus the wife's labour income. The rest of the variables are self explanatory. For a detail description of each of the variables, please see Mroz (1987, pages 771, 796).

Table 3.1: Description of variables, and summary of PSID 1975 female labour supply data.

| Variables | Description | Summary: Average (SD) | |
|---|---|---|---|
| | | All women | Working women |
| `inlf` | = 1 if in labour force in 1975 | – | – |
| `hours` | Hours worked in 1975 | 740.60 | 1302.93 |
| | | (871.3) | (776.27) |
| `k5` | Kids less than 6 years | 0.24 | 0.14 |
| | | (0.52) | (0.39) |
| `k618` | Kids 6-18 years | 1.35 | 1.35 |
| | | (1.31) | (1.31) |
| `age` | Woman's age in years | 42.53 | 41.97 |
| | | (8.07) | (7.72) |
| `educ` | Years of schooling | 12.28 | 12.65 |
| | | (2.28) | (2.28) |
| `wage` | Estimated hourly wage from earnings | – | 185.26 |
| | | | (107.26) |
| `repwage` | Reported wage at interview in 1976 | 1.84 | 3.18 |
| | | (2.42) | (2.44) |
| `hushrs` | Hours worked by husband in 1975 | 2267.27 | 2233.46 |
| | | (595.56) | (582.91) |
| `husage` | Husband's age | 45.12 | 44.61 |
| | | (8.06) | (7.95) |
| `huseduc` | Husband's years of schooling | 12.49 | 12.61 |
| | | (3.02) | 3.03 |
| `huswage` | Husband's hourly wage in 1975 | 7.48 | 7.22 |
| | | (4.23) | (3.57) |
| `faminc` | Family income in 1975 (000s) | 23.08 | 24.13 |
| | | (12.19) | (11.67) |
| `mtr` | Federal marginal tax rate facing woman | 0.68 | 0.66 |
| | | (0.08) | (0.07) |
| `motheduc` | Mother's years of schooling | 9.25 | 9.51 |
| | | (3.36) | (3.31) |
| `fatheduc` | Father's years of schooling | 8.80 | 8.99 |
| | | (3.57) | (3.52) |
| `unem` | Unemployment rate in county of residence | 8.62 | 8.54 |
| | | (3.11) | (3.03) |
| `city` | = 1 if living in SMSA | – | – |
| `exper` | Actual labour market experience | 10.63 | 13.03 |
| | | (8.07) | (8.05) |
| `nwifeinc` | (`faminc` - `wage` × `hours`)/1000 | 20.12 | 18.93 |
| | | (11.63) | (10.59) |

In the next section we define the statistical model and analyze the data by fitting a semiparametric regression model. Unlike Fox (2005), where smoothing spline was used for fitting nonparametric part, we use B-spline basis function. Since we are estimating the nonparametric part through a different method, we briefly present the results of our analysis in the following section. We used `gam` function in `mgcv` package in R (R Development Core Team, 2010) for model fitting.

## 3.3 Statistical Model

We assume that $\mathbf{1}_n = (1, \ldots, 1)'$ is not in the space spanned by the column vectors of $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$. As a result, according to Chen (1988), model (3.1) is identifiable. In addition, we assume the design points $\boldsymbol{x}_i$ and $t_i$ as fixed for $i = 1, \ldots, n$. The design space of $t$ is $[0, 1]$ and it is assumed that the sequence of designs (we drop the dependence on $n$) forms an asymptotically regular sequence (Sacks and Ylvisaker, 1970) in the sense that

$$\max_{i=1,\ldots,n} \left| \int_0^{t_i} p(t)dt - \frac{i-1}{n-1} \right| = o(n^{-\frac{3}{2}}).$$

Here $p(\cdot)$ denotes a positive density function on the interval $[0, 1]$ which is Lipschitz continuous of order one. Let us introduce a restriction on the parameters in model (3.1) as

$$y_i = \boldsymbol{X}\boldsymbol{\beta} + g(t_i) + \varepsilon_i \quad \text{subject to } \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}, \tag{3.2}$$

where $\boldsymbol{H}$ is a $p_2 \times p$ restriction matrix, and $\boldsymbol{h}$ is a $p_2 \times 1$ vector of constants. In this paper, we consider $\boldsymbol{H} = [\mathbf{0}, \boldsymbol{I}]$, and $\boldsymbol{h} = \mathbf{0}$.

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')'$ be the semiparametric least squares estimator of $\boldsymbol{\beta}$ for the model

(3.1). Here, $\hat{\boldsymbol{\beta}}$ is a column vectors. Then we call $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ the semiparametric unrestricted least squares estimator of $\boldsymbol{\beta}_1$. If $\boldsymbol{\beta}_2 = \mathbf{0}$, then the model in (3.1) reduces to

$$y_i = x_{i1}\beta_1^{(*)} + \ldots + x_{ip_1}\beta_{p_1}^{(*)} + g^{(*)}(t_i) + \varepsilon_i^{(*)}, \quad i = 1, 2, \ldots, n. \tag{3.3}$$

Here $(*)$ is used to differentiate the slope parameters in (3.3) from those in (3.1). The reduced model in (3.3) gives restricted estimator of $\boldsymbol{\beta}_1$. Let us denote the semiparametric restricted least squares estimator by $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$.

We develop shrinkage and PSE of $\boldsymbol{\beta}_1$, and denote them by $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ and $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$, respectively. Our main objective is efficient estimation of $\boldsymbol{\beta}_1$ when it is suspected that $\boldsymbol{\beta}_2 = \mathbf{0}$ or close to zero.

### 3.3.1 Model Building Strategy: Candidate Full and Submodels

Similar to Mroz (1987), we consider `hours`– woman's hours of work in 1975, as our response variable. Because of the nature of our response variable, we only used the portion of the data when the women were in labour force. Thus, we had 428 cases (rows) in our working data. Our candidate full model consists of age (`age`), non-wife income (`nwifeinc`), children aged five and younger (`k5`), number of children between ages six and eighteen (`k618`), wife's college attendance (`wc`), husband's college attendance (`hc`), unemployment rate in the county of residence (`unem`), actual labour force experience (`exper`), and marginal tax rate (`mtr`). A brief summary of the variables in our model is given in Table 3.2.

After applying stepwise variable selection procedure based on AIC, BIC, and ab-

Table 3.2: Description of Variables in the Model for Working Women.

| Covariates | Description | Remarks |
|---|---|---|
| `hours` | Hours worked in 1975 | Min=12, max=4950, median=1303 |
| `age` | Age (in years) of woman | Min=30, max=60, median=42 |
| `nwifeinc` | Non-wife income | Income in thousands |
| `k5` | Number of kids five and younger | 0-1, a few 2's and 3's, factor variable |
| `k618` | Number of kids six to 18 years | 0-4, few >4, factor variable |
| `wc` | Whether wife attended college | 1 (if `educ` > 12), else 0 |
| `hc` | Whether husband attended college | 1 (if `huseduc` > 12), else 0 |
| `unem` | Unemployment rate | Min=3, max=14, median=7.5 |
| `mtr` | Marginal tax rate facing women | Min=0.44, max=0.94, median=0.69 |
| `exper` | Actual labour market experience | Min=0, max=38, median=12 |

solute penalty (lasso), we obtained three candidate sub-models. AIC selection procedure picked `wc`, `nwifeinc`, `mtr`, `exper`, `unem`, `k5`, and `age`. BIC picked `wc`, `nwifeinc`, `mtr`, and `exper` (see Table 3.3). To apply lasso on a model with categorical predictors (such as `k5` and `k618`), we need to make dummy variables for each levels of the categorical covariate. For instance, `k618` had four levels, so, four dummy variables were created, namely, `k6180` (when `k6180=0`), `k6181` (when `k618=1`), `k6182` (when `k618=2`), and `k6183` (when `k618` $\geq$`3`). When applied to data, lasso method retained all the variables of the full model except `k6183`. We used `glmnet` package in R to implement lasso on our data.

Table 3.3: Selection of covariates by AIC, BIC.

| Models | Selected / Chosen Variables |
|---|---|
| Full Model | `wc`, `nwifeinc`, `mtr`, `exper`, `unem`, `k5`, `age`, `k618`, `hc` |
| AIC | `wc`, `nwifeinc`, `mtr`, `exper`, `unem`, `k5`, `age` |
| BIC | `wc`, `nwifeinc`, `mtr`, `exper` |

We begin with the model given by AIC and investigate if any of the covariates can be modelled nonparametrically. For this, we fit several models to test for nonlinearity of each of the covariates. The selected covariates, deviance and residual degrees of

freedom of these models are listed in Table 3.4. The codes in Table 3.4 are as follows: F = factor or dummy variable, L = linear term, S = a smoothed term estimated by B-spline basis expansion. Of all the models, we have found that model 2 has the smallest deviance in which `nwifeinc` was estimated nonparametrically using B-spline approximation. Analysis of deviance (Table 3.5) confirms that `nwifeinc` has a significant nonlinear relationship with the response `hours` (p-value 0.0065) making model 2 as the preferred one.

Table 3.4: Deviance table for various models fitted with `mroz` data.

| | Predictors | | | | | | | Deviance | |
|---|---|---|---|---|---|---|---|---|---|
| Model | `k5` | `wc` | `age` | `unem` | `exper` | `nwifeinc` | `mtr` | (000,000) | df (res) |
| 0 | F | F | L | L | L | L | L | 196.26 | 419 |
| 1 | F | F | L | L | L | L | S | 191.19 | 412 |
| 2 | F | F | L | L | L | S | L | 186.62 | 411 |
| 3 | F | F | L | L | S | L | L | 191.21 | 411 |
| 4 | F | F | L | S | L | L | L | 195.08 | 414 |
| 5 | F | F | S | L | L | L | L | 192.33 | 411 |

Code: L = linear term, S = smoothed term.

Table 3.5: Analysis of deviance table for tests of nonlinearity of `age`, `unem`, `exper`, `nwifeinc` and `mtr`.

| Model contrasted | Predictor | Difference in deviance | Difference in df (res) | $p$-value |
|---|---|---|---|---|
| 1–0 | `mtr` | 5.15 | 7 | 0.1337 |
| 2–0 | `nwifeinc` | 9.65 | 8 | 0.0065 |
| 3–0 | `exper` | 5.05 | 8 | 0.2097 |
| 4–0 | `unem` | 1.19 | 5 | 0.7731 |
| 5–0 | `age` | 3.94 | 8 | 0.3941 |

Keeping model 2 in mind, we test for significance of each of the predictors by dropping them one at a time. For this, additional models (Table 3.6) were fitted and contrasted with model 2. Results are reported in Table 3.7. Analysis of deviance confirms that there is strong evidence of partial relationship of woman's hours of

work to wife's college attendance, labour force experience, marginal tax rate, and non-wife income of the family but not to children five and younger, age of woman, and unemployment rate. Interestingly, the significant covariates found through deviance analysis are also the ones that were picked by the BIC.

Table 3.6: Deviance table for additional models to test for significance of each of the predictors.

| | Predictors | | | | | | | Deviance | |
|---|---|---|---|---|---|---|---|---|---|
| Model | `k5` | `wc` | `age` | `unem` | `exper` | `nwifeinc` | `mtr` | | df (res) |
| 2 (Ref) | F | F | L | L | L | S | L | 186.62 | 411 |
| 6 | - | F | L | L | L | S | L | 187.89 | 413 |
| 7 | F | - | L | L | L | S | L | 189.19 | 412 |
| 8 | F | F | - | L | L | S | L | 187.68 | 412 |
| 9 | F | F | L | - | L | S | L | 187.60 | 412 |
| 10 | F | F | L | L | - | S | L | 191.42 | 414 |
| 11 | F | F | L | L | L | - | L | 191.59 | 412 |
| 12 | F | F | L | L | L | S | - | 221.78 | 412 |

Code: F= Factor or dummy, L = linear term, S = smoothed term.

Table 3.7: Analysis of deviance table for additional models when contrasted with model 2.

| Model contrasted | Predictor | Difference in deviance | Difference in df (res) | $p$-value |
|---|---|---|---|---|
| 6–2 | `k5` | 1.27 | 2 | 0.2454 |
| 7–2 | `wc` | 2.57 | 1 | 0.0173 |
| 8–2 | `age` | 1.06 | 1 | 0.1266 |
| 9–2 | `unem` | 9.86 | 1 | 0.1405 |
| 10–2 | `exper` | 4.81 | 3 | 0.0142 |
| 11–2 | `nwifeinc` | 4.98 | 1 | 0.0009 |
| 12–2 | `mtr` | 3.51 | 1 | $\ll$0.0001 |

For the sake of visualizing nonlinearity, we jointly plotted `mtr` and `nwifeinc` in a three dimensional space in Figure 3.1 a), after holding other predictors fixed. The two-dimensional plot in panel d) of Figure 3.1 visibly shows a nonlinear relationship between non-wife income and woman's hours of work. We notice that the confidence

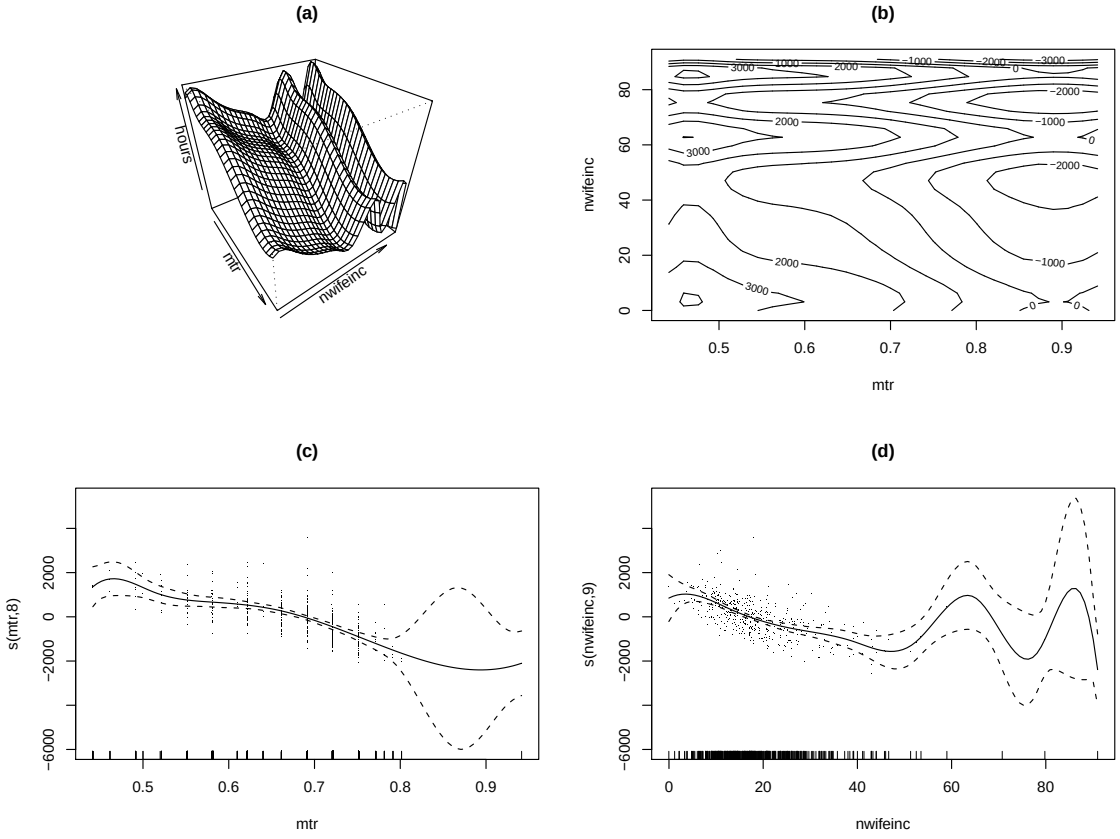


Figure 3.1: (a) Visualizing nonlinear relationship of `mtr` and `nwifeinc` with woman's hours of work. (b) contour plot, (c) 2-D plot of `mtr` shows the smoothed curve estimated by B-spline basis function, and (d) shows the smoothed curve for `nwifeinc` estimated by B-spline basis function with uniform knots. Dashed lines in c) and d) are 95% confidence envelopes of the smoothed curves.

envelopes in panels c) and d) get wider near the edge of the curves. The reason behind the large scale for the confidence envelope is due to the small number of sample points as `mtr` and `nwifeinc` increase. This causes high variability in the fitted values causing the confidence envelopes to explode.

Finally, with the inclusion of a nonparametric part, our candidate full- and sub-models are listed below. Since the model produced by lasso did not eliminate any

covariate completely, we are not considering it as a sub-model.

Full-Model: `hours` = `wc` + $g$(`nwifeinc`) + `mtr` + `exper` + `unem` + `k5` + `age` + `k618` + `hc`

Sub-Model: `hours` = `wc` + $g$(`nwifeinc`) + `mtr` + `exper`

Here `g(nwifeinc)` denotes a component estimated by B-spline basis function.

It is to be mentioned here that, although we have found that the covariates `unem`, `k5`, `age`, `k618`, and `hc` do not contribute significantly in predicting `hours`, and subsequently being dropped from the model, the shrinkage estimates based on full- and sub-models above may result in a model with all the variables of the full model depending on the quantity $1 - (p_2 - 2)\psi_n^{-1}$ defined in Section 3.4.2. However, the coefficients will be shrunken, and some of them might be zero.

## 3.4 Estimation Strategies

We first define a semiparametric least square estimator for the parameter vector $\boldsymbol{\beta}$ based on $g(\cdot)$ approximated by a B-spline series. The book by de Boor (2001) is an excellent source for various properties of splines as well as many computer algorithms. Let $k$ be an integer larger than or equal to $\nu$ where $\nu$ will be defined in Assumption 3.7.2. Further, let $S_{m_n,k}$ be the class of functions $s(\cdot)$ on $[0,1]$ with the following properties:

(i) $s(\cdot)$ is a polynomial of degree $k$ on each of the sub-intervals $[(i-1)/m_n, i/m_n], i = 1, \ldots, m_n$, where $m_n$ is a positive integer which depends on $n$.

(ii) $s(\cdot)$ is $(k-1)$ times differentiable.

Then $S_{m_n,k}$ is called the class of all splines of degree $k$ with $m_n$-equispaced knots. Note that $S_{m_n,k}$ has a basis of $m_n + k$ normalized B-spline $\{B_{m_nj}(\cdot) : j = 1, \ldots, m_n + k\}$, and $g(\cdot)$ can be approximated by a linear combination $\boldsymbol{\theta}'\boldsymbol{B}_{m_n}(\cdot)$ of the bases, where $\boldsymbol{\theta} \in \mathcal{R}^{m_n+k}$ and $\boldsymbol{B}_{m_n}(\cdot) = (B_{m_n,1}(\cdot), \ldots, B_{m_n,m_n+k}(\cdot))'$. With $\boldsymbol{\theta}'\boldsymbol{B}_{m_n}(\cdot)$, model in (3.1) becomes

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\theta}'\boldsymbol{B}_{m_n}(t_i) + \varepsilon_i. \tag{3.4}$$

For $\boldsymbol{\beta} \in \mathcal{R}^p$ and $\boldsymbol{\theta} \in \mathcal{R}^{m_n+k}$, let

$$S_n(\boldsymbol{\beta}, \boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \left[ y_i - \boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{\theta}'\boldsymbol{B}_{m_n}(t_i) \right]^2. \tag{3.5}$$

In the following, we discuss and develop UE, SE, PSE, and an APE as defined in Section 2.2.1.

### 3.4.1 Unrestricted and Restricted Estimators

If $S_n(\cdot, \cdot)$ is minimized at $(\hat{\boldsymbol{\beta}}, \hat{\theta})$, then we have

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{B}_{m_n}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}_{B_{m_n}}\boldsymbol{Y} \quad \text{and} \quad \hat{\boldsymbol{\theta}} = (\boldsymbol{B}_{m_n}'\boldsymbol{B}_{m_n})^{-1}\boldsymbol{B}_{m_n}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\boldsymbol{x}_s = (x_{1s}, \ldots, x_{ns})'$, $s = 1, \ldots, p$, $\boldsymbol{M}_{\boldsymbol{B}_{m_n}} = \boldsymbol{I} - \boldsymbol{B}_{m_n}(\boldsymbol{B}_{m_n}'\boldsymbol{B}_{m_n})^{-1}\boldsymbol{B}_{m_n}'$ and $\boldsymbol{B}_{m_n} = (B_{m_n}(t_1), \ldots, B_{m_n}(t_n))$. The estimator $\hat{\boldsymbol{\beta}}$ is called a semiparametric least squares estimator (SLSE) of $\boldsymbol{\beta}$. The SLSE possess some good statistical properties. With respect to a quadratic risk function, $\hat{\boldsymbol{\beta}}$ can be dominated by a class of shrinkage estimators.

Using the inverse matrix formula, the semiparametric unrestricted least squared

estimator $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ of $\boldsymbol{\beta}_1$ is

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}} = (\boldsymbol{X}_1' \boldsymbol{M}_{\boldsymbol{B}_{m_n}} \boldsymbol{M}_{\boldsymbol{B}_{m_n}\boldsymbol{X}_2} \boldsymbol{M}_{\boldsymbol{B}_{m_n}} \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1' \boldsymbol{M}_{\boldsymbol{B}_{m_n}} \boldsymbol{M}_{\boldsymbol{B}_{m_n}\boldsymbol{X}_2} \boldsymbol{M}_{\boldsymbol{B}_{m_n}} \boldsymbol{Y},$$

where $\boldsymbol{X}_1$ is composed of the first $p_1$ row vectors of $\boldsymbol{X}$, $\boldsymbol{X}_2$ is composed of the last $p_2$ row vectors of $\boldsymbol{X}$, and $\boldsymbol{M}_{\boldsymbol{B}_{m_n}\boldsymbol{X}_2} = \boldsymbol{I} - \boldsymbol{B}_{m_n}\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{B}_{m_n}'\boldsymbol{B}_{m_n}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{B}_{m_n}'$. When $\boldsymbol{\beta}_2 = 0$, we have the restricted partially linear regression (reduced) model which is

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip_1}\beta_{p_1} + g(t_i) + \varepsilon_i,\ i = 1, \ldots, n. \tag{3.6}$$

Using the semiparametric least squares estimation for $\boldsymbol{\beta}$, similar to Ahmed et al. (2007), an estimator of $\boldsymbol{\beta}_1$ can be obtained,which has the form

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}} = (\boldsymbol{X}_1' \boldsymbol{M}_{\boldsymbol{B}_{m_n}} \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1' \boldsymbol{M}_{\boldsymbol{B}_{m_n}} \boldsymbol{Y}.$$

$\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ is called a semiparametric restricted estimator of $\boldsymbol{\beta}_1$.

### 3.4.2 Shrinkage Estimators

A semiparametric shrinkage estimator (SSE) $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ of $\boldsymbol{\beta}_1$ can be defined as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{S}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{RE}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}) \left\{1 - (p_2 - 2)\psi_n^{-1}\right\}, \quad p_2 \geq 3,$$

where

$$\psi_n = \frac{n}{\hat{\sigma}_n^2} \hat{\boldsymbol{\beta}}_2' \boldsymbol{X}_2' \boldsymbol{B}_{m_n}' \boldsymbol{M}_{\boldsymbol{B}_{m_n}\boldsymbol{X}_2} \boldsymbol{B}_{m_n} \boldsymbol{X}_2 \hat{\boldsymbol{\beta}}_2,$$

with

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} - \boldsymbol{B}_{m_n}'(t_i)\hat{\boldsymbol{\theta}})^2.$$

A positive-part shrinkage semiparametric estimator (PSSE) is obtained by retaining the positive-part of the SSE. We denote PSSE by $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$. A PSSE has the form

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{RE}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RE}})\left\{1 - (p_2 - 2)\psi_n^{-1}\right\}^+, \quad p_2 \geq 3$$

where $z^+ = \max(0, z)$.

### 3.4.3 Absolute Penalty Estimators

Absolute penaly estimation (APE) was defined in Section 2.2.2. In this chapter, we use lasso and adaptive lasso estimators for the purpose of comparison with shrinkage estimators. Therefore, we briefly present the definitions of lasso and adaptive lasso (alasso) in the following.

Lasso is a member of the penalized least squares family, which performs simultenaous variable selection and parameter estimation. Lasso was proposed by Tibshirani (1996). Lasso solutions are obtained as

$$\tilde{\boldsymbol{\beta}}_{\mathrm{lasso}} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}, \tag{3.7}$$

where $\lambda$ is the tuning parameter which controls the amount of shrinkage. The tuning parameter is selected via cross-validation.

For a root-n consistent estimator $\hat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}$, let us denote the alasso estimator by $\hat{\beta}^{\mathrm{alasso}}$. We may consider $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$ as an estimator of $\boldsymbol{\beta}^*$. For a chosen value of $\gamma > 0$, we

calculate the weights $\hat{w}_j = 1/|\hat{\beta}_j^*|^\gamma$. Finally, the adaptive lasso estimates are obtained as

$$\hat{\boldsymbol{\beta}}^{\text{alasso}} = \arg\min_{\beta} \left|\left| \boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|. \tag{3.8}$$

The algorithm to obtain the alasso estimates is described in detaile in Section 2.2.2.

## 3.5 Application

In the previous section we analyzed labour supply data and developed a sub-model. In this section we evaluate the performance of shrinkage, positive-shrinkage, lasso, and alasso estimates through prediction errors and log likelihood criteria. For lasso, we used `glmnet` package, and the `adalasso()` function in `parcor` R-package was used to to compute alasso estimates.

Prediction errors were obtained following the discussion on page 18 of Hastie et al. (2009). Our results are based on 9999 case resampled bootstrap samples. Initially, we varied the number of replications and settled with this as no noticeable variation were observed for larger samples. For each bootstrap replicate, average prediction errors were calculated by ten-fold cross validation. Figure 3.2 shows that lasso estimator has the smallest prediction error similar to that of the full model. The rest of estimators perform equally well in terms of prediction errors. On the other hand, all the estimators perform equally in terms of loglikelihood, with the restricted estimator having slightly larger loglikelihood value. Although the alasso estimator has higher prediction error than the lasso, it is interesting to note that our proposed estimators are behaving quite similarly with the alasso. On the other hand, lasso is behaving more like the full model. The reason might be the fact that the lasso model has as

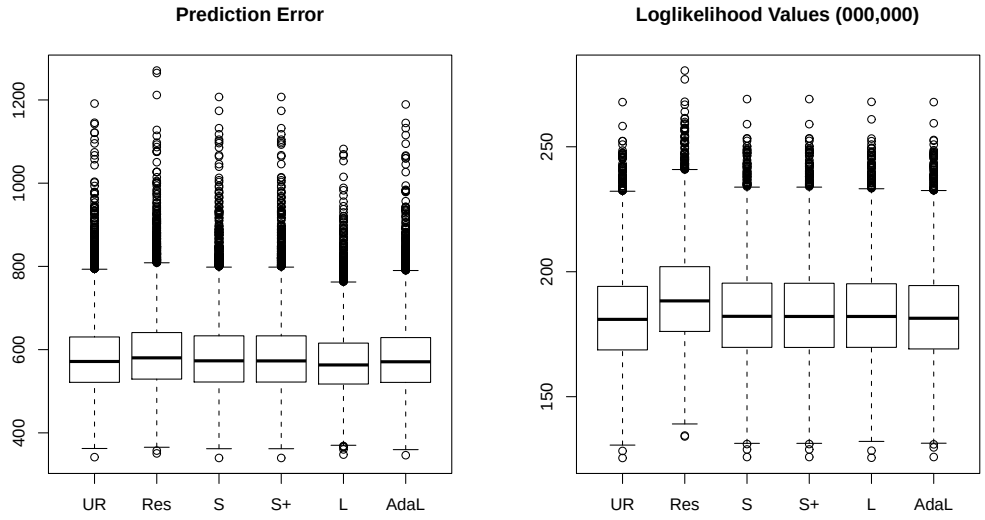


Figure 3.2: Comparison of the estimators through prediction errors and loglikelihood values.

many covariates as there are in the full model. Noticeably, the log-likelihood of the proposed estimators are similar to the log-likelihood of the full model.

## 3.6 Simulation Studies

We perform Monte Carlo simulation experiments to examine the quadratic risk performance of the proposed estimators. We simulate the response from the following model:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{pi}\beta_p + g(t_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $t_i = (i - 0.5)/n$, $x_{1i} = (\zeta_{1i}^{(1)})^2 + \zeta_i^{(1)} + \xi_{1i}$, $x_{2i} = (\zeta_{2i}^{(1)})^2 + \zeta_i^{(1)} + 2\xi_{2i}$, $x_{si} = (\zeta_{si}^{(1)})^2 + \zeta_i^{(1)}$ with $\zeta_{si}^{(1)}$ i.i.d. $\sim N(0,1)$, $\zeta_i^{(1)}$ i.i.d. $\sim N(0,1)$, $\xi_{1i}$ $\sim$Bernoulli(0.45) and $\xi_{2i}$ $\sim$Bernoulli(0.45) for all $s = 3, \ldots, p$, and $i = 1, \ldots, n$. Moreover, $\varepsilon_i$ are i.i.d. $N(0,1)$, $n >> p$, and $g(t) = \sin(4\pi t)$.

We are interested in testing the hypothesis $H_0 : \boldsymbol{\beta}_j = \mathbf{0}$, for $j = p_1 + 1, p_1 + 2, \ldots, p_1 + p_2$, with $p = p_1 + p_2$. Our aim is to estimate $\beta_1$, $\beta_2$, $\beta_3$, and $\beta_4$ when the remaining regression parameters may not be useful. We partition the regression coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$ with $\boldsymbol{\beta}_1 = (2, 1.5, 1, 0.6)$.

The number of simulations was initially varied. Next, each realization was repeated 5000 times to obtain stable results. For each realization, we calculated bias of the estimators. We defined $\Delta^* = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1, \mathbf{0})$, and $||\cdot||$ is the Euclidean norm. To determine the behavior of the estimators for $\Delta^* > 0$, further datasets were generated from those distributions under local alternative hypothesis. We consider $\Delta^* = 0, .1, .2, .3, .4, .5, .8, 1, 2,$ and 4.

The risk performance of an estimator of $\boldsymbol{\beta}_1$ was measured by calculating its mean squared error (MSE). After calculating the MSEs, we numerically calculated the efficiency of the proposed estimators $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$, relative to the unrestricted estimator $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ using the relative mean squared error (RMSE) criterion defined by

$$\mathrm{RMSE}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}} : \hat{\boldsymbol{\beta}}_1^*) = \frac{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}})}{\mathrm{MSE}(\hat{\boldsymbol{\beta}}_1^*)}, \tag{3.9}$$

where $\hat{\boldsymbol{\beta}}_1^*$ is one of the proposed estimators. An RMSE greater than 1 indicates that $\hat{\boldsymbol{\beta}}_1^*$ is superior to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$.

In this study, we used B-spline basis expansion with uniform knots for estimating the nonparametric component. According to He and Shi (1996), uniform knots are usually sufficient when the function $g(\cdot)$ does not exhibit dramatic changes in its derivatives. Thus, we just need to determine the number of knots. The method discussed in He and Shi (1996) is used to determine this number. In a separate simulation study (results not presented here), we found that a degree-three B-spline

with three knots performs best for sample sizes larger than 40, and two knots are sufficient for moderate sample sizes ($n \leq 35$).

To compute RMSEs, we considered $n = 30, 50, 80, 100, 125$, $p_1 = 3, 4$, and $p_2 = 5$, 9, 15. Since the results of our simulation study are similar for all the combinations, we graphically present results in Figure 3.3 for $n = 50, 80$, $p_1 = 4$, and $p_2 = 5, 9, 15$. The horizontal line at RMSE=1 facilitates a comparison among the estimators. Any point above this horizontal line indicates superiority of the proposed estimator compared to the unrestricted one.

In general, restricted estimators ($\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$) have the largest RMSE, which indicates their superiority over other estimators when the null hypothesis is true ($\Delta^* = 0$). Not surprisingly, RMSE of $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ decays quite sharply as we deviate from the null hypothesis ($\Delta^* > 0$), and quickly goes below the horizontal line. On the other hand, shrinkage ($\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$) and positive-shrinkage ($\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$) estimators perform steadily for a range of $\Delta^*$.

The findings of the simulation study may be summarized as follows.

(i) Figure 3.3 shows that the restricted estimator outperforms all other estimators for all the cases considered in this study. However, this is true when the restriction is at or near $\Delta^* = 0$. As the restriction moves away from $\Delta^* = 0$, the restricted estimator becomes inefficient (see the sharply decaying RMSE curve that goes below the horizontal line at RMSE=1 when $\Delta^* > 0$).

(ii) The RMSE of the positive-shrinkage estimator $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ approaches 1 at the slowest rate as we move away from $\Delta^* = 0$. This indicates that in the event of imprecise subspace information (i.e., even if $\boldsymbol{\beta}_2 \neq \mathbf{0}$), it has the smallest quadratic risk

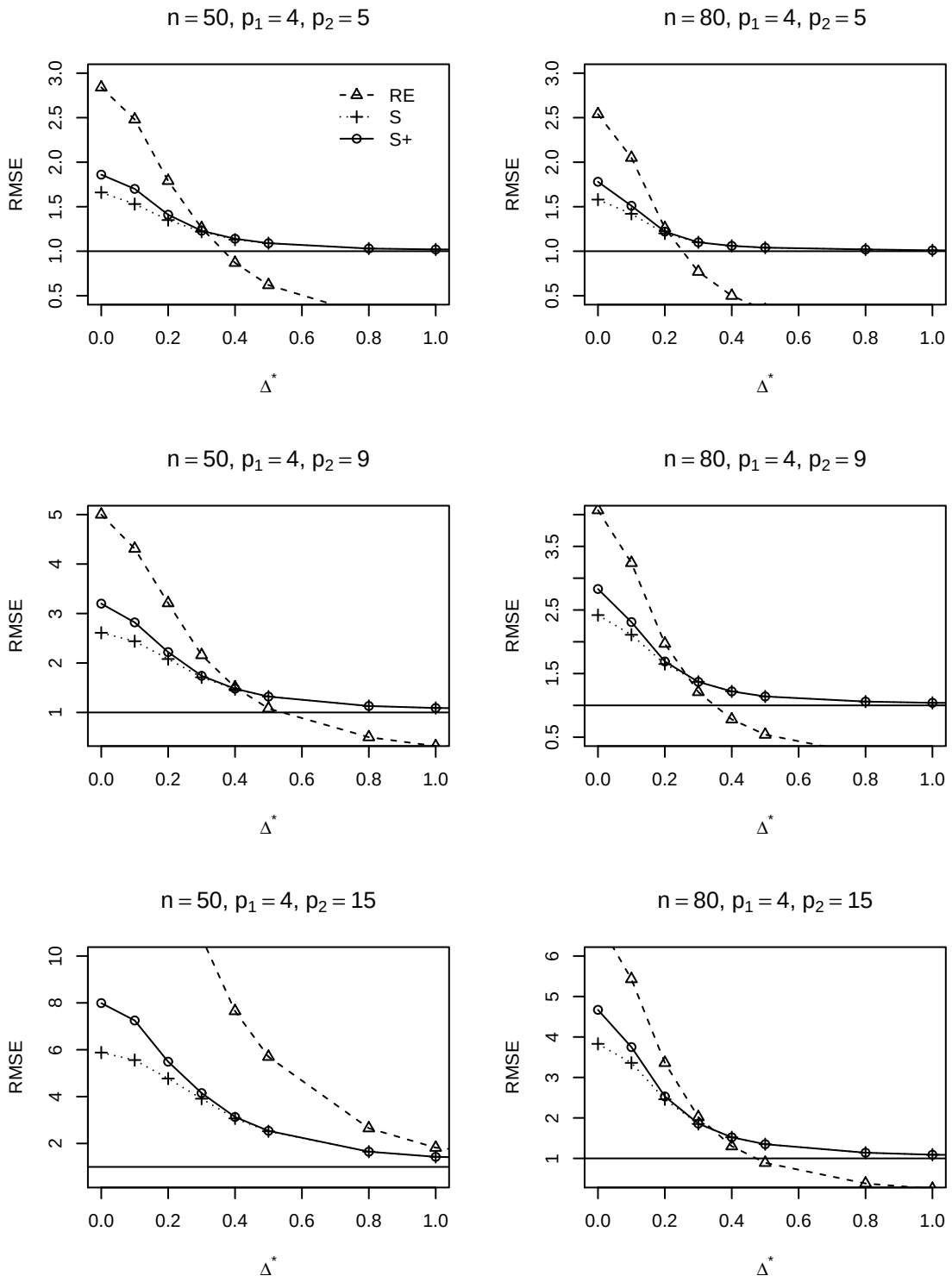


Figure 3.3: Relative mean squared error of the estimators as a function of the non-centrality parameter $\Delta^*$ for sample sizes $n = 50, 80$, $p_1 = 4$, and $p_2 = 5, 9, 15$.

among all other estimators, making it an ideal choice for real-life applications.

In summary, the simulation results are in agreement with our asymptotic results and the general theory of these estimators available in the literature.

### 3.6.1 Comparison with Absolute Penalty Estimator

We compare shrinkage estimators with an APE (lasso only), based on the RMSE criterion. The tuning parameter for the APE was estimated using cross validation (CV) and generalized cross validation (GCV). In our simulation, we considered $p_1 = 3, 4$ and $p_2 = 3, 4, 5, 6, 9, 11, 15$. Only $\Delta^* = 0$ was considered since, according to Ahmed et al. (2007), APE does not take into consideration that the parameter vector $\boldsymbol{\beta}$ is partitioned into main and nuisance parts, and is at a disadvantaged position when $\Delta^* > 0$. Simulated RMSEs are presented in Table 3.8, and 3.9. Figure 3.4 shows RMSEs when $p_1 = 3$, and Figure 3.5 shows the same when $p_1 = 4$. Both figures reveal that shrinkage estimates have a smaller risk than the APE for moderate-sized samples. As the number of nuisance parameters increases, shrinkage estimators perform better than the APE.

For a succinct comparison between positive-shrinkage and APE, we plotted RMSEs in a three-dimensional diagram (see Figures 3.6, 3.7). The horizontal axis represents $n$, the diagonal axis shows $p_2$, while the RMSEs are plotted on the vertical axis. Solid black circles represent positive shrinkage estimates, and hollow circles, represented by APE (CV), indicate APE with cross validation. Clearly, shrinkage estimator is doing better for moderate sample sizes and when $p_2$ is large. On the other hand, APE has higher RMSE than the shrinkage estimators for large sample sizes and when the number of main parameters is large.

Table 3.8: Shrinkage versus APE: simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 3$, $\Delta^* = 0$.

| $n$ | $p_2$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | APE(CV) | APE(GCV) |
|---|---|---|---|---|---|---|
| 30 | 3 | 2.49 | 1.26 | 1.37 | 1.46 | 1.63 |
| | 4 | 3.21 | 1.52 | 1.72 | 1.64 | 1.86 |
| | 5 | 3.94 | 1.82 | 2.07 | 1.92 | 2.18 |
| | 6 | 4.84 | 2.15 | 2.49 | 2.12 | 2.46 |
| | 9 | 8.19 | 3.12 | 3.80 | 2.83 | 3.57 |
| | 11 | 11.89 | 3.98 | 4.87 | 3.13 | 4.39 |
| | 15 | 27.22 | 6.02 | 7.67 | 2.37 | 5.69 |
| 40 | 3 | 2.40 | 1.25 | 1.35 | 1.62 | 1.72 |
| | 4 | 2.92 | 1.49 | 1.68 | 1.89 | 2.08 |
| | 5 | 3.53 | 1.74 | 2.00 | 2.10 | 2.27 |
| | 6 | 4.10 | 1.99 | 2.32 | 2.37 | 2.55 |
| | 9 | 6.51 | 2.90 | 3.43 | 3.08 | 3.46 |
| | 11 | 8.40 | 3.55 | 4.24 | 3.49 | 4.05 |
| | 15 | 13.65 | 4.99 | 6.06 | 4.52 | 5.78 |
| 80 | 3 | 2.23 | 1.26 | 1.35 | 1.82 | 1.85 |
| | 4 | 2.67 | 1.47 | 1.66 | 2.15 | 2.20 |
| | 5 | 3.18 | 1.70 | 1.96 | 2.42 | 2.46 |
| | 6 | 3.63 | 1.91 | 2.26 | 2.65 | 2.72 |
| | 9 | 5.15 | 2.66 | 3.21 | 3.48 | 3.61 |
| | 11 | 6.19 | 3.17 | 3.88 | 4.00 | 4.17 |
| | 15 | 8.33 | 4.25 | 5.23 | 5.03 | 5.30 |
| 100 | 3 | 2.24 | 1.23 | 1.34 | 1.63 | 1.65 |
| | 4 | 2.66 | 1.45 | 1.65 | 1.86 | 1.87 |
| | 5 | 3.07 | 1.69 | 1.94 | 2.47 | 2.50 |
| | 6 | 3.51 | 1.92 | 2.24 | 2.27 | 2.31 |
| | 9 | 4.91 | 2.62 | 3.15 | 2.91 | 2.97 |
| | 11 | 5.90 | 3.13 | 3.81 | 3.31 | 3.40 |
| | 15 | 8.00 | 4.14 | 5.13 | 4.09 | 4.24 |

We recommend a positive-shrinkage estimator for practical purposes when sample size is moderate, and when the number of nuisance parameters is large.

Table 3.9: Shrinkage versus APE: simulated RMSE with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for $p_1 = 4$, $\Delta^* = 0$.

| $n$ | $p_2$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$ | APE(CV) | APE(GCV) |
|---|---|---|---|---|---|---|
| 30 | 3 | 2.18 | 1.23 | 1.31 | 1.27 | 1.41 |
| | 4 | 2.68 | 1.46 | 1.63 | 1.49 | 1.69 |
| | 5 | 3.29 | 1.73 | 1.93 | 1.69 | 1.93 |
| | 6 | 3.96 | 1.98 | 2.28 | 1.87 | 2.18 |
| | 9 | 6.71 | 2.96 | 3.49 | 2.44 | 3.12 |
| | 11 | 9.70 | 3.66 | 4.54 | 2.62 | 3.76 |
| | 15 | 24.68 | 5.99 | 7.64 | 1.64 | 5.07 |
| 40 | 3 | 2.05 | 1.22 | 1.31 | 1.46 | 1.55 |
| | 4 | 2.45 | 1.42 | 1.59 | 1.64 | 1.74 |
| | 5 | 2.90 | 1.66 | 1.85 | 1.83 | 1.96 |
| | 6 | 2.91 | 1.67 | 1.90 | 2.03 | 2.19 |
| | 9 | 4.61 | 2.60 | 3.06 | 2.65 | 2.98 |
| | 11 | 5.11 | 2.68 | 3.20 | 3.04 | 3.50 |
| | 15 | 11.27 | 4.68 | 6.06 | 3.81 | 4.79 |
| 80 | 3 | 1.92 | 1.20 | 1.29 | 1.61 | 1.62 |
| | 4 | 2.23 | 1.40 | 1.54 | 1.85 | 1.89 |
| | 5 | 2.56 | 1.60 | 1.78 | 2.05 | 2.09 |
| | 6 | 2.93 | 1.80 | 2.04 | 2.25 | 2.30 |
| | 9 | 3.99 | 2.36 | 2.78 | 2.90 | 2.99 |
| | 11 | 4.80 | 2.80 | 3.32 | 3.28 | 3.40 |
| | 15 | 6.54 | 3.78 | 4.55 | 4.13 | 4.36 |
| 100 | 3 | 1.87 | 1.20 | 1.28 | 1.87 | 1.89 |
| | 4 | 2.18 | 1.37 | 1.52 | 2.21 | 2.23 |
| | 5 | 2.50 | 1.57 | 1.77 | 2.11 | 2.13 |
| | 6 | 2.85 | 1.76 | 2.01 | 2.77 | 2.81 |
| | 9 | 3.88 | 2.39 | 2.78 | 3.61 | 3.70 |
| | 11 | 4.61 | 2.79 | 3.27 | 4.12 | 4.21 |
| | 15 | 6.20 | 3.66 | 4.35 | 5.12 | 5.33 |

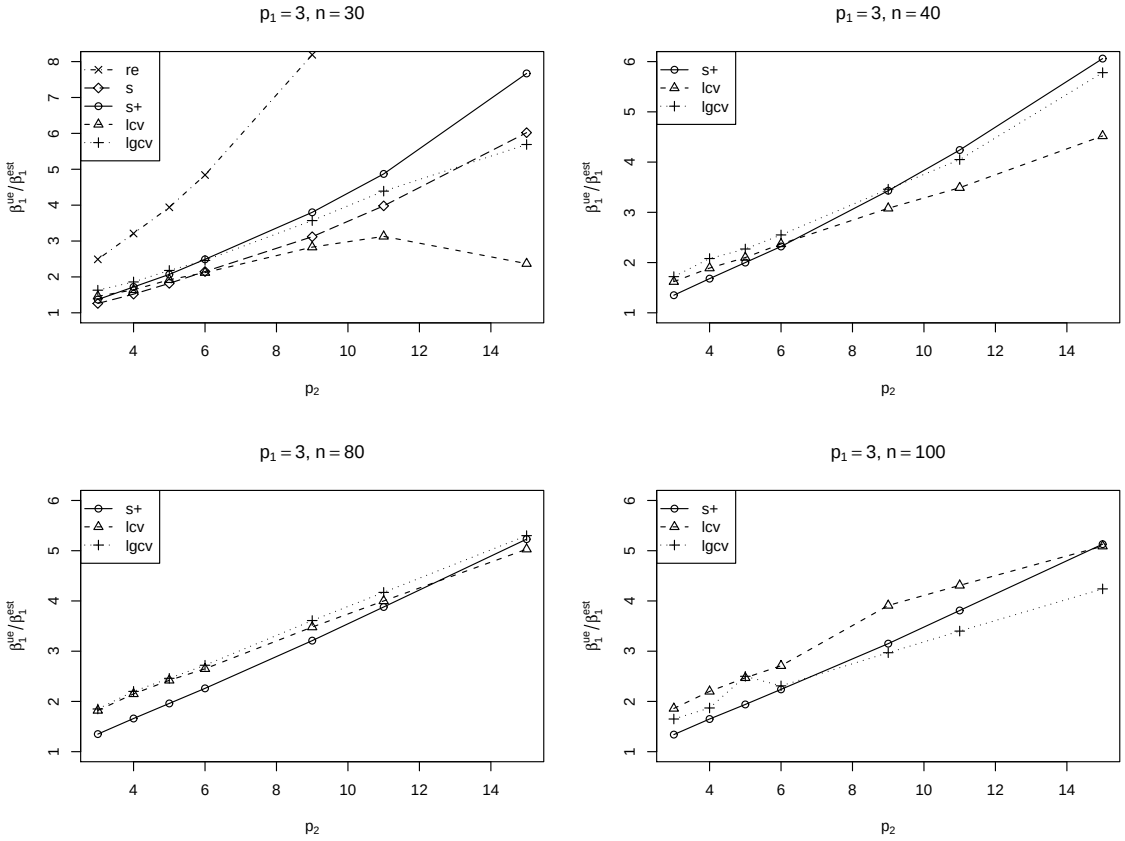


Figure 3.4: Graphical comparison of simulated RMSE plot for shrinkage and lasso when $p_1 = 3$ and $p_2$ varies for different $n$.

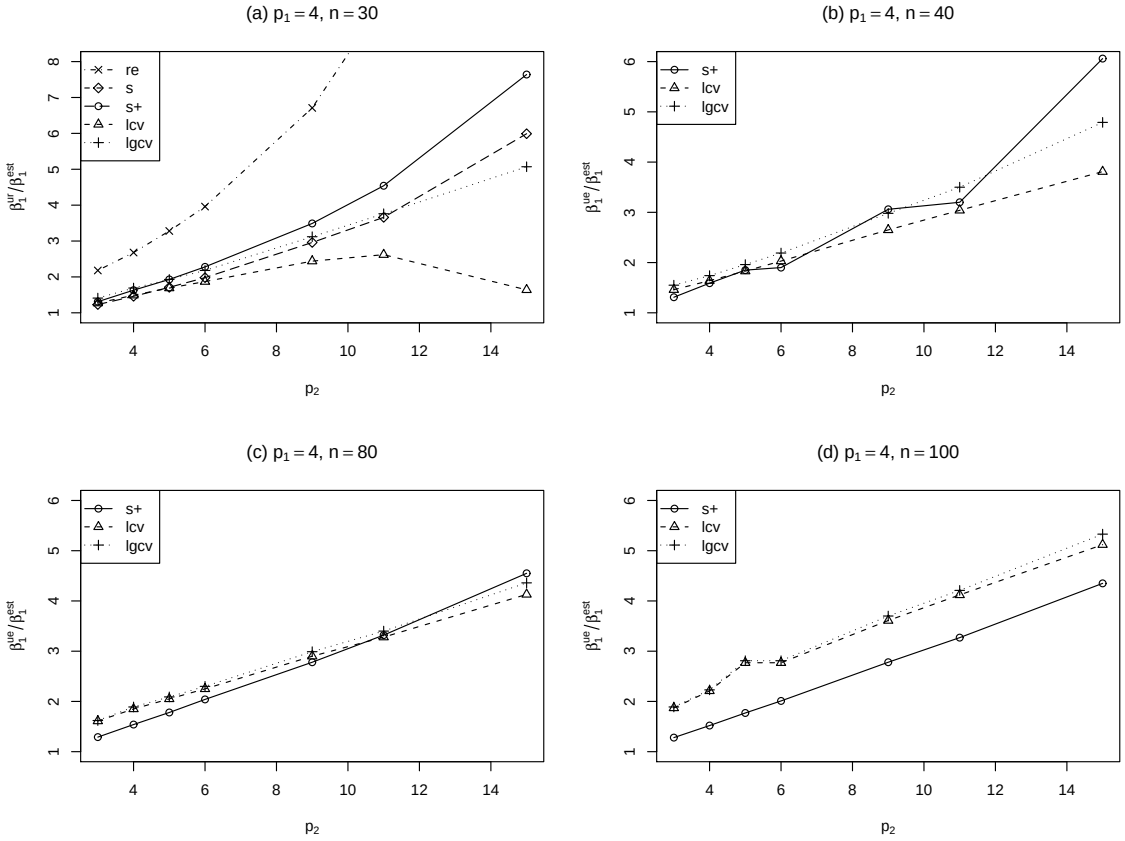


Figure 3.5: Graphical comparison of simulated RMSE plot for shrinkage and lasso when $p_1 = 4$ and $p_2$ varies for different $n$.

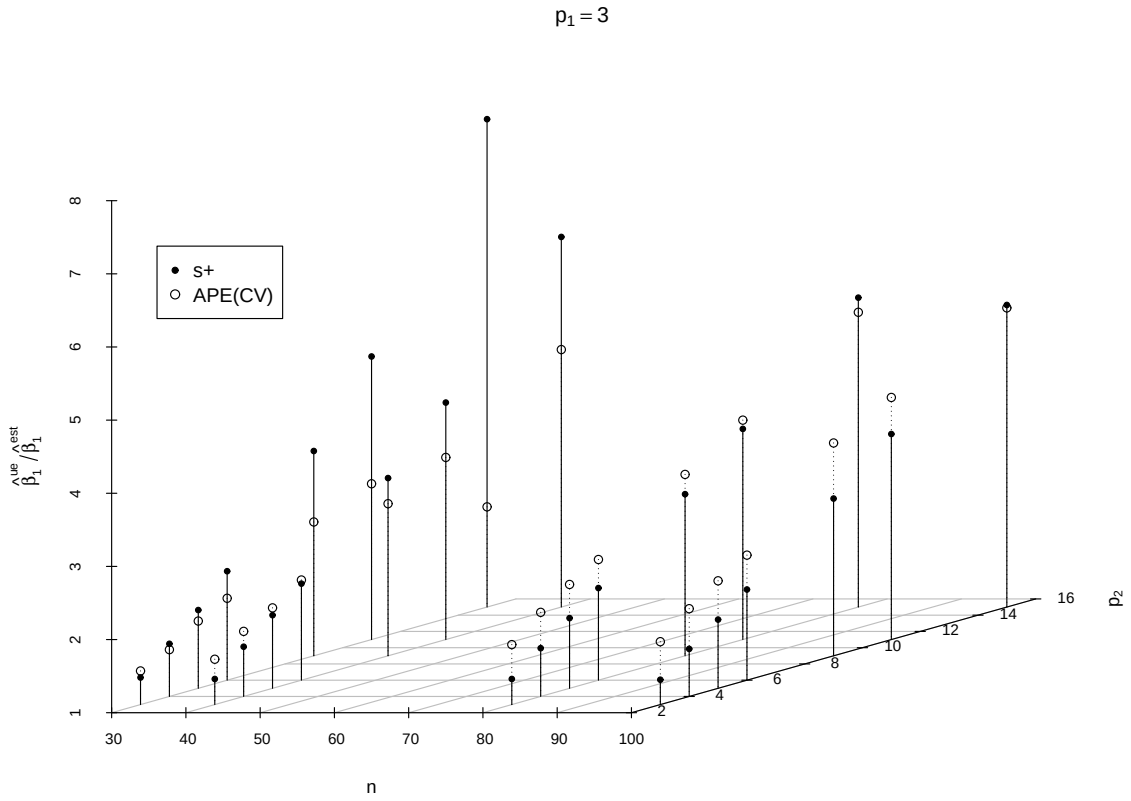


Figure 3.6: Three-dimensional plot of RMSE against $n$ and $p_2$ for $p_1 = 3$ to compare positive shrinkage estimator and APE(CV).

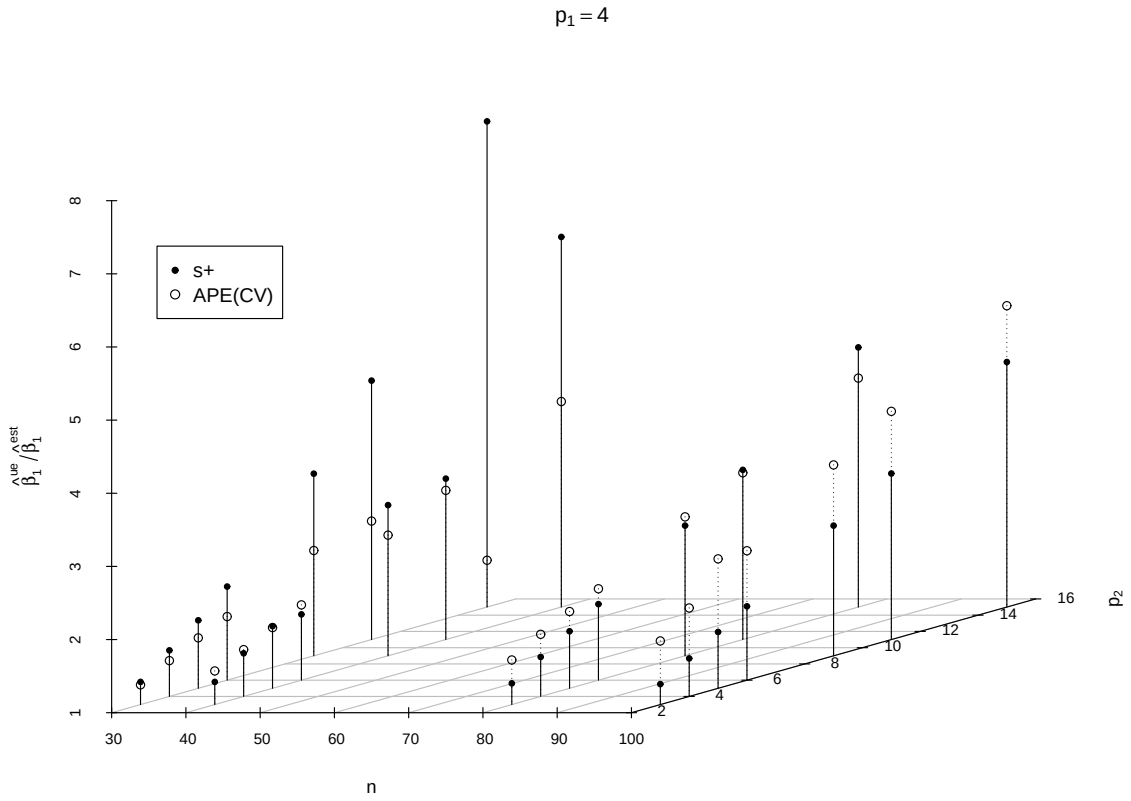


Figure 3.7: Three-dimensional plot of RMSE against $n$ and $p_2$ for $p_1 = 4$ to compare positive shrinkage estimator and APE(CV).

## 3.7 First-Order Asymptotics

The following assumptions are needed to derive the main results. These assumptions, while they look a bit lengthy, are actually quite mild and can be easily satisfied (see remark following the assumptions).

**Assumption 3.7.1.** There exist bounded functions $h_s(\cdot)$ on $[0,1]$, $s = 1, \ldots, p$, such that

$$x_{is} = h_s(t_i) + u_{is},\ i = 1, \ldots, n, s = 1, \ldots, p, \qquad (a)$$

where $u_i = (u_{i1}, \ldots, u_{ip})'$ are real vectors satisfying

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n u_{ik}u_{ij}}{n} = b_{kj}, \quad \text{for } k = 1, \ldots, p,\ j = 1, \ldots, p, \qquad (b)$$

and the matrix $\boldsymbol{B} = (b_{kj})$ is nonsingular. Moreover,

$$\max_{1\le k,j\le p} \|\boldsymbol{A}\boldsymbol{u}_j^*\| = O\Big( [\mathrm{tr}(\boldsymbol{A}'\boldsymbol{A})]^{\frac{1}{2}} \Big), \text{ for any matrix } \boldsymbol{A}, \qquad (c)$$

where $\boldsymbol{u}_k^* = (u_{1k}, \ldots, u_{nk})'$ and $||\cdot||$ denotes the Euclidean norm.

**Assumption 3.7.2.** The functions $g(\cdot)$ and $h_j(\cdot)$ satisfy the Lipschitz condition of order $\nu$, i.e., there exists a constant $c$ such that

$$|f_j^{(\nu)}(s) - f_j^{(\nu)}(t)| \le c|s-t|^\nu, \quad \text{for any } s,t \in [0,1], \quad j = 0, 1, \ldots, p,$$

where $f_0(\cdot) = g(\cdot)$ and $f_j(\cdot) = h_j(\cdot),\ j = 1, \ldots, p$.

*Remark* 1. In many applications, the above $u_{ij}$ behave like zero-mean random variables and $h_j(t_i)$ are the regression of $x_{ij}$ on $t_i$. Specifically, suppose that the design points $(\boldsymbol{x}_i, t_i)$ are independent and identically distributed (i.i.d.) random variables,

and let $h_s(t_i) = E(x_{ij}|t_i)$ and $u_{is} = x_{is} - h_s(t_i)$ with $E\boldsymbol{u}_i\boldsymbol{u}_i' = \boldsymbol{B}$. Then by the law of large numbers (b) holds with probability 1. Assumption 3.7.1 (a) and (b) are used in Gao (1995a, 1995b, 1997), Liang (2000), Härdle, Liang and Gao (2000) among others. According to Moyeed and Diggle (1994), (c) holds when $u_{ij}$ behave like zero-mean uncorrelated random variables.

**Lemma 3.7.1.** If $f(\cdot)$ satisfies assumption 3.7.2, then we have

$$\sup_{t\in[0,1]} |f(t) - \boldsymbol{B}_{m_n}(t)(\boldsymbol{B}_{m_n}'\boldsymbol{B}_{m_n})^{-1}\boldsymbol{B}_{m_n}'\boldsymbol{f}| = O(n^{-1}m_n) + O(m_n^{-\nu}),$$

where $\boldsymbol{f} = (f(t_1), \ldots, f(t_n))'$.

**Lemma 3.7.2.** For the B-spline basis we have i) $\sum_{i=1}^{m_n+k} B_{m_n}^2(t) \leq 1$ for all $t$. ii) all the eigenvalues of $n^{-1}\boldsymbol{B}_{m_n}'\boldsymbol{B}_{m_n}$ are between $c_1 m_n^{-1}$ and $c_2 m_n^{-1}$, where $0 < c_1 < c_2 < \infty$.

*Proof.* The proof of Lemmas 3.7.1 and 3.7.2 can be found in Burman (1991).

**Lemma 3.7.3.** Suppose that assumptions 3.7.1 and 3.7.2 hold, and $\varepsilon_i$ are independent with mean zero, equal variance $\sigma^2$ and $\mu_{3i} = E\varepsilon_i^3$ being uniformly bounded. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_D N(0, \sigma^2\boldsymbol{B}^{-1}) \text{ and } \boldsymbol{B}_{m_n}'(t)\hat{\boldsymbol{\theta}} - g(t) = O_p(n^{-\frac{\nu}{2\nu+1}}),$$

where "$\to_D$" denotes convergence in distribution and $\boldsymbol{B}$ is defined in assumption 3.7.1.

*Proof.* The proof of the asymptotic normality of $\hat{\boldsymbol{\beta}}$ is similar to that of Theorem 3 in

Gao (1995). Following the proof of Lemma 3 in Ahmed et al. (2007) we write

$$\begin{aligned}\hat{\boldsymbol{\theta}}'\boldsymbol{B}_{m_n}(t) - g(t) &= \boldsymbol{B}'_{m_n}(t)(\boldsymbol{B}'_{m_n}\boldsymbol{B}_{m_n})^{-1}\boldsymbol{B}_{m_n}\boldsymbol{g} - g(t) + \boldsymbol{B}'_{m_n}(t)(\boldsymbol{B}'_{m_n}\boldsymbol{B}_{m_n})^{-1}\boldsymbol{B}_{m_n}\boldsymbol{\varepsilon} \\ &+ \boldsymbol{B}'_{m_n}(t)(\boldsymbol{B}'_{m_n}\boldsymbol{B}_{m_n})^{-1}\sum_{i=1}^{n}\boldsymbol{B}_{m_n}(t_i)\boldsymbol{x}'_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= I_1 + I_2 + I_3, \text{ say,}\end{aligned}$$

where $\boldsymbol{g} = (g(t_1), \ldots, g(t_n))'$. By Lemma 3.7.1, $I_1 = O(n^{-1}m_n) + O(m_n^{-\nu})$. Moreover,

$$\text{Cov}(I_2) = \sigma^2\boldsymbol{B}'_{m_n}(t)(\boldsymbol{B}'_{m_n}\boldsymbol{B}_{m_n})^{-1}\boldsymbol{B}_{m_n}(t) = O(n^{-1}m_n).$$

This implies $I_2 = O_p(n^{-\frac{\nu}{2\nu+1}})$. Combining the root-n consistency of $\hat{\boldsymbol{\beta}}$ by the same argument we can show $I_3 = O_p(n^{-\frac{\nu}{2\nu+1}})$ and the proof follows.

**Lemma 3.7.4.** Suppose that assumptions 3.7.1 and 3.7.2 hold. Then $n^{-1}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{B}_{m_n}}\boldsymbol{X} = \boldsymbol{B} + O(n^{-\frac{2\nu}{2\nu+1}})$.

*Proof.* The proof of Lemma 3.7.4 is similar to that of Lemma 2 in Gao (1995).

**Lemma 3.7.5.** Suppose that assumptions 3.7.1 and 3.7.2 hold. Then we have

$$\hat{\sigma}_n^2 = \sigma^2 + O_p\left(n^{-\frac{1}{2}}\right), \ \hat{\boldsymbol{\beta}}_1^R = (\boldsymbol{I}, \boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12})\hat{\boldsymbol{\beta}} + o_p\left(n^{-\frac{1}{2}}\right), \text{ and } T_n = n\sigma^2\hat{\boldsymbol{\beta}}'_2\boldsymbol{B}_{22.1}\hat{\boldsymbol{\beta}}_2 + o_p(1),$$

where $\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix}$ and $\boldsymbol{B}_{22.1} = \boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}$.

*Proof.* By definition of $\hat{\sigma}_n^2$, we have

$$\begin{aligned}\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{x}_i'(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]^2 + \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{B}_{m_n}'(t_i)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right]^2 + \frac{1}{n}\sum_{i=1}^{n}(g(t_i) \\ &\quad -\boldsymbol{B}_{m_n}'(t_i)\boldsymbol{\theta})^2 \\ &\quad +\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{x}_i'(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{B}_{m_n}'(t_i)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(g(t_i)-\boldsymbol{B}_{m_n}'(t_i)\boldsymbol{\theta}) \\ &\quad +\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{x}_i'(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\boldsymbol{B}_{m_n}'(t_i)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \frac{2}{n}\sum_{i=1}^{n}\boldsymbol{x}_i'(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})(g(t_i)-\boldsymbol{B}_{m_n}'(t_i)\boldsymbol{\theta}) \\ &\quad +\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{B}_{m_n}'(t_i)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(g(t_i)-\boldsymbol{B}_{m_n}'(t_i)\boldsymbol{\theta}) = I_1+\ldots+I_{10}, \text{ say.}\end{aligned}$$

It is seen that $I_1 = \sigma^2 + O_p(n^{-\frac{1}{2}})$. Based on Lemmas 3.7.2 and 3.7.3 we can show $I_2 = o_p(n^{-\frac{1}{2}})$ and $I_3 = o_p(n^{-\frac{1}{2}})$. By Lemma 3.7.1, $I_4 = o(n^{-\frac{1}{2}})$. In addition, by the Cauchy-Schwartz inequality we can prove that the other terms are $o_p(n^{-\frac{1}{2}})$. These lemmas provide the basis for deriving the asymptotic distributional bias (ADB) and asymptotic distributional risk (ADR) for the proposed semiparametric estimators.

## 3.8 Asymptotic Properties of Shrinkage Estimators

We investigate the performance of the proposed estimators when, without loss of generality, $\boldsymbol{\beta}_2$ is close to $\mathbf{0}$. Thus we consider the sequence $\{K_n\}$ given by

$$K_n : \boldsymbol{\beta}_{2(n)} = n^{-\frac{1}{2}}\boldsymbol{\omega}, \quad \boldsymbol{\omega} \neq \mathbf{0} \text{ fixed vector.}$$

### 3.8.1 Bias Performance of the Estimators

The asymptotic distributional bias (ADB) of an estimator $\boldsymbol{\beta}_1^*$ is defined as

$$\text{ADB}(\boldsymbol{\beta_1^*}) = E\left\{\lim_{n\to\infty} n^{\frac{1}{2}}(\boldsymbol{\delta} - \boldsymbol{\beta}_1)\right\}.$$

**Theorem 3.8.1.** Suppose that assumptions 3.7.1 and 3.7.2 hold. Under $\{K_n\}$, the ADB of the estimators are as follows:

$$\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{UE}}) = \mathbf{0}, \tag{3.10}$$

$$\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{RE}}) = -\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\omega, \tag{3.11}$$

$$\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{S}}) = -(p_2-2)\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}E(\chi_{p_2,\alpha}^{-2};\Delta), \tag{3.12}$$

and

$$\begin{aligned}\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{S+}}) = &-\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}H_{p_2+2}(p_2-2;\Delta)\\ &-(p_2-2)\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}\left\{E\left[\chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^2(\Delta)>p_2-2)\right]\right\},\end{aligned} \tag{3.13}$$

where $\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix}$ with $\boldsymbol{B}$ defined in assumption 3.7.1, $\Delta = (\boldsymbol{\omega}'\boldsymbol{B}_{22.1}\boldsymbol{\omega})\sigma^{-2}$, $\boldsymbol{B}_{22.1} = \boldsymbol{B}_{22} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}$, $H_v(x;\Delta)$ denotes the noncentral chi-square distribution function with noncentrality parameter $\Delta$ and $v$ degrees of freedom. Here $E\left\{(\chi_\nu^2(\Delta))^{-m}\right\}$ is the expected value of the inverse of a non-central chi-square distribution with $\nu$ degrees of freedom and noncentrality parameter $\Delta$. For nonnegative integer-valued $\nu$ and $m$, and for $\nu > 2m$, the expectations can be obtained using the Theorem in Bock et al. (1983, page 7).

*Proof.* It is easy to prove this theorem using Theorem 4.1 in Ahmed et al. (2007).

We omit the details.

The bias expressions for all the estimators are not in scalar form. We therefore take recourse by converting them into quadratic form. Let us define the asymptotic distributional quadratic bias (ADQB) of an estimator $\boldsymbol{\beta}_1^*$ of $\boldsymbol{\beta}_1$ by

$$\text{ADQB}(\boldsymbol{\beta}_1^*) = [\text{ADB}(\boldsymbol{\beta}_1^*)]' \boldsymbol{B}_{11.2} [\text{ADB}(\boldsymbol{\beta}_1^*)].$$

**Theorem 3.8.2.** Suppose that conditions in Theorem 4.5.2 hold. Then the ADQB of the estimators under consideration are given by

$$\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{UE}}) = \mathbf{0}, \tag{3.14}$$

$$\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{RE}}) = \boldsymbol{\omega}' \boldsymbol{B}_{21} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{11.2} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{12} \boldsymbol{\omega}, \tag{3.15}$$

$$\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{S}}) = (p_2 - 2)^2 \omega' \boldsymbol{B}_{21} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{11.2} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{12} \boldsymbol{\omega} \left[ E(\chi_{p_2,\alpha}^{-2}; \Delta) \right]^2, \tag{3.16}$$

and

$$\begin{aligned} \text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{S}+}) = {} & \boldsymbol{\omega}' \boldsymbol{B}_{21} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{11.2} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{12} \boldsymbol{\omega} \\ & \cdot \left\{ H_{p_2+2}(p_2 - 2; \Delta) - (p_2 - 2) E \left[ \chi_{p_2+2}^{-2}(\Delta) I(\chi_{p_2+2}^2(\Delta) > p_2 - 2) \right] \right\}^2. \end{aligned} \tag{3.17}$$

For $\boldsymbol{B}_{12} = \mathbf{0}$, we have $\boldsymbol{B}_{21} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{12} = 0$ and $\boldsymbol{B}_{11.2} = \boldsymbol{B}_{11}$ and hence all the ADQB reduce to common value zero for all $\boldsymbol{\omega}$. All these variations, thus, become ADQB-equivalent. Hence, in the sequel we assume that $\boldsymbol{B}_{12} \neq \mathbf{0}$ and the remaining discussions follow.

The ADQB of $\hat{\boldsymbol{\beta}}_1^{\text{RE}}$ is an unbounded function of $\boldsymbol{\omega}' \boldsymbol{B}_{21} \boldsymbol{B}_{11}^{-1} \boldsymbol{B}_{11.2} \boldsymbol{B}_{11}^{-1} \ \boldsymbol{B}_{12} \boldsymbol{\omega}$.

In order to investigate $\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{S}})$ and $\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{S}+})$, we use the following result

from matrix algebra:

$$\begin{aligned} \mathrm{ch}_{\min}(\sigma^2 \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{11.2}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1}) &\leq \frac{\sigma^2 \boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{11.2}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}}{\boldsymbol{\omega}'\boldsymbol{B}_{22.1}\boldsymbol{\omega}} \\ &\leq \mathrm{ch}_{\max}(\sigma^2 \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{11.2}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1}). \end{aligned}$$

Therefore, ADQB($\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$) starts from zero at $\boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{11.2}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega} = 0$, increases to a point then decreases towards zero due to $E(\chi^{-2}_{p_2+2}(\Delta))$ being a decreasing log-convex function of $\Delta$. The behavior of $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}+}$ is similar to $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$, however, the quadratic bias curve of $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}+}$ remains below the curve of $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ for all values of $\Delta$.

**Simulation Study for Bias**

Simulated biases for the slope parameters are shown in Table 3.8.1. Here we considered $p_1 = 3, p_2 = 4$ with true parameter vector $\beta = (1, 1, 1, 0, 0, 0, 0)'$. We also tested a highly oscillating non-flat function to compare bias of the slope parameters for B-spline and kernel-based estimators. The B-spline performed better than the kernel for this function. Zheng et al. (2006) used a highly oscillating non-flat function that is identical to the one used in our paper. They considered

$$g(t) = \sin\left(-\frac{2\pi(0.35 \times 10 + 1)}{0.35t + 1}\right), \quad t \in [0, 10]. \tag{3.18}$$

Simulated bias of the slope parameters using this function are given in Table 3.8.1.

Table 3.10: Simulated bias of the slope parameters when the true parameter vector was $\beta = (1, 1, 1, 0, 0, 0, 0)'$. Here, $p_1 = 3$, $p_2 = 4$, and the results are based on 5000 Monte Carlo runs, when $g(t)$ is a flat function.

| | | B-spline | | | Kernel | | |
|---|---|---|---|---|---|---|---|
| $\Delta$ | $\beta$ | RE | S | S+ | RW | S | S+ |
| 0 | $\beta_1$ | -0.0013 | -0.0011 | -0.0009 | -0.0013 | -0.0014 | -0.0013 |
| | $\beta_2$ | 0.0002 | 0.0005 | 0.0001 | -0.0003 | -0.0006 | -0.0004 |
| | $\beta_3$ | 0.0012 | 0.0010 | 0.0011 | 0.0017 | 0.0013 | 0.0017 |
| | $\beta_4$ | 0.0000 | -0.0005 | 0.0003 | 0.0000 | 0.0018 | 0.0011 |
| | $\beta_5$ | 0.0000 | 0.0003 | 0.0002 | 0.0000 | 0.0003 | -0.0004 |
| | $\beta_6$ | 0.0000 | -0.0011 | -0.0009 | 0.0000 | -0.0013 | -0.0016 |
| | $\beta_7$ | 0.0000 | 0.0002 | 0.0006 | 0.0000 | 0.0006 | -0.0001 |
| 2 | $\beta_1$ | 0.0323 | -0.0001 | -0.0001 | 0.0167 | -0.0029 | -0.0029 |
| | $\beta_2$ | 0.0290 | 0.0019 | 0.0019 | 0.0152 | -0.0004 | -0.0004 |
| | $\beta_3$ | 0.0239 | 0.0011 | 0.0011 | 0.0156 | -0.0010 | -0.0010 |
| | $\beta_4$ | 0.0000 | 1.9902 | 1.9902 | 0.0000 | 1.9903 | 1.9903 |
| | $\beta_5$ | 0.0000 | -0.0022 | -0.0022 | 0.0000 | -0.0029 | -0.0029 |
| | $\beta_6$ | 0.0000 | 0.0000 | 0.0000 | 0.0000 | -0.0007 | -0.0007 |
| | $\beta_7$ | 0.0000 | -0.0005 | -0.0005 | 0.0000 | 0.0013 | 0.0013 |

Table 3.11: Simulated bias of the slope parameters when the true parameter vector was $\beta = (1, 1, 1, 0, 0, 0, 0)'$. Here, $p_1 = 3$, $p_2 = 4$, and the results are based on 5000 Monte Carlo runs, when $g(t)$ is a highly oscillating non-flat function.

| | | B-spline | | | Kernel | | |
|---|---|---|---|---|---|---|---|
| $\Delta$ | $\beta$ | RE | S | S+ | RE | S | S+ |
| 0 | $\beta_1$ | 0.0009 | 0.0008 | 0.0008 | 0.0096 | 0.0099 | 0.0100 |
| | $\beta_2$ | 0.0025 | 0.0019 | 0.0022 | 0.0079 | 0.0083 | 0.0080 |
| | $\beta_3$ | -0.0013 | -0.0015 | -0.0013 | 0.0077 | 0.0072 | 0.0072 |
| | $\beta_4$ | 0.0000 | -0.0003 | 0.0003 | 0.0000 | 0.0027 | 0.0030 |
| | $\beta_5$ | 0.0000 | 0.0005 | 0.0004 | 0.0000 | 0.0032 | 0.0043 |
| | $\beta_6$ | 0.0000 | 0.0005 | 0.0009 | 0.0000 | 0.0048 | 0.0057 |
| | $\beta_7$ | 0.0000 | 0.0031 | 0.0014 | 0.0000 | 0.0032 | 0.0037 |
| 2 | $\beta_1$ | 0.0309 | 0.0031 | 0.0031 | 0.0314 | 0.0124 | 0.0124 |
| | $\beta_2$ | 0.0362 | 0.0016 | 0.0016 | 0.0369 | 0.0087 | 0.0087 |
| | $\beta_3$ | 0.0186 | 0.0016 | 0.0016 | 0.0174 | 0.0050 | 0.0050 |
| | $\beta_4$ | 0.0000 | 1.9896 | 1.9896 | 0.0000 | 1.9997 | 1.9997 |
| | $\beta_5$ | 0.0000 | 0.0011 | 0.0011 | 0.0000 | 0.0091 | 0.0091 |
| | $\beta_6$ | 0.0000 | 0.0034 | 0.0034 | 0.0000 | 0.0061 | 0.0061 |
| | $\beta_7$ | 0.0000 | -0.0005 | -0.0005 | 0.0000 | 0.0072 | 0.0072 |

### 3.8.2 Risk Performance of the Estimators

To study the asymptotic distributional quadratic risk (ADQR) of an estimator, we define a quadratic loss function using a positive definite matrix (p.d.m.) $\boldsymbol{M}$, as follows

$$\mathcal{L}(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1) = n(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)'\boldsymbol{M}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1),$$

where $\boldsymbol{\beta}_1^*$ is any one of the proposed estimators. If $\boldsymbol{V}_n$ is the asymptotic dispersion matrix of $\boldsymbol{\beta}_1^*$, the ADQR of $\sqrt{n}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)$ is given by $\mathrm{tr}(\boldsymbol{M}\boldsymbol{V}_n)$. Now we assume that for the estimator $\boldsymbol{\beta}_1^*$ of $\boldsymbol{\beta}_1$, the cumulative distribution function of $\boldsymbol{\beta}_1^*$ under $\{K_n\}$ exists, and is

$$F(\boldsymbol{x}) = P\left\{\lim_{n\to\infty} \sqrt{n}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1) \le \boldsymbol{x} | K_n\right\},$$

where $F(\boldsymbol{x})$ is non degenerate. Then, the ADQR of $\boldsymbol{\beta}_1^*$ is defined as

$$R(\boldsymbol{\beta}_1^*, \boldsymbol{M}) = \mathrm{tr}\left\{\boldsymbol{M}\int_{\mathcal{R}^{p_1}}\int \boldsymbol{x}\boldsymbol{x}' dF(\boldsymbol{x})\right\} = \mathrm{tr}(\boldsymbol{M}\boldsymbol{V}),$$

where $\boldsymbol{V}$ is the dispersion matrix of the asymptotic distribution $F(\boldsymbol{x})$.

**Theorem 3.8.3.** Suppose that assumptions 3.7.1 and 3.7.2 hold. Under $\{K_n\}$, the asymptotic covariance matrix of the estimators are:

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}) = \sigma^2\boldsymbol{B}_{11.2}^{-1} \quad \text{where} \quad \boldsymbol{B}_{11.2} = \boldsymbol{B}_{11} - \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21}, \tag{3.19}$$

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}}) = \sigma^2\boldsymbol{B}_{11}^{-1} + \boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}, \tag{3.20}$$

$$\begin{aligned}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}^{\mathrm{S}}) = {} & \sigma^2\boldsymbol{B}_{11.2}^{-1} - \\ & (p_2-2)\sigma^3\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1}\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\left\{2E(\chi^2_{p_2,\alpha}(\Delta)) - E(\chi^2_{p_2+2,\alpha}(\Delta))\right\} \\ & + (p_2^2-4)\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}E(\chi^{-4}_{p_2+4}(\Delta)),\end{aligned} \tag{3.21}$$

and

$$
\begin{aligned}
& \boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}^{\mathrm{S}+}) \\
= \quad & \boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}^{S}) + (p_2-2)\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1}\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\left\{2E\left[\chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^{2}(\Delta)\leq p_2-2)\right]\right. \\
- \quad & \left.(p_2-2)E\left[\chi_{p_2+2}^{-4}(\Delta)I(\chi_{p_2+2}^{2}(\Delta)\leq p_2-2)\right]\right\} - \sigma^2\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1}\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1} \\
& \cdot H_{p_2+2}(p_2-2;\Delta) + \boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\left[2H_{p_2+2}(p_2-2;\Delta)-\right. \\
& \left.H_{p_2+4}(p_2-2;\Delta)\right] \\
- \quad & (p_2-2)\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\left\{2E\left[\chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^{2}(\Delta)\leq p_2-2)\right]\right. \\
- \quad & \left.2E\left[\chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^{2}(\Delta)\leq p_2-2)\right] + \right. \\
& \left.(p_2-2)E\left[\chi_{p_2+2}^{-4}(\Delta)I(\chi_{p_2+2}^{2}(\Delta)\leq p_2-2)\right]\right\}
\end{aligned}
$$

**Theorem 3.8.4.** Suppose that assumptions 3.7.1 and 3.7.2 hold, then under $\{K_n\}$ the ADQR of the estimators are:

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}};\boldsymbol{M}) = \sigma^2\mathrm{tr}(\boldsymbol{M}\boldsymbol{B}_{11.2}^{-1}), \tag{3.22}$$

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{RE}};\boldsymbol{M}) = \sigma^2\mathrm{tr}(\boldsymbol{M}\boldsymbol{B}_{11}^{-1}) + \boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}, \tag{3.23}$$

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}};\boldsymbol{M}) = \sigma^2 & \left[\mathrm{tr}(\boldsymbol{M}\boldsymbol{B}_{11.2}^{-1}) - (p_2-2)\mathrm{tr}(\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1})\right. \\
& \left.\cdot\left\{2E(\chi_{p_2,\alpha}^{2}(\Delta)) - (p_2-2)E(\chi_{p_2+2}^{-4}(\Delta))\right\}\right] \\
& + (p_2^2-4)\boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}E(\chi_{p_2+4}^{-4}(\Delta)),
\end{aligned} \tag{3.24}
$$

and

$$
\begin{aligned}
& R(\hat{\boldsymbol{\beta}}^{\mathrm{S}+}; \boldsymbol{M}) \\
= \; & R(\hat{\boldsymbol{\beta}}^{\mathrm{S}}; \boldsymbol{M}) \\
& +(p_2-2)\mathrm{tr}(\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1}) \left\{2E\left[\chi_{p_2+2}^{-2}(\Delta) I(\chi_{p_2+2}^2(\Delta) \le p_2-2)\right]\right. \\
& \left.-(p_2-2)E\left[\chi_{p_2+2}^{-4}(\Delta) I(\chi_{p_2+2}^2(\Delta) \le p_2-2)\right]\right\} - \sigma^2\mathrm{tr}(\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1}) \\
& \cdot H_{p_2+2}(p_2-2;\Delta) + \boldsymbol{\omega}'\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}\left[2H_{p_2+2}(p_2-2;\Delta)\right. \\
& \left.-H_{p_2+4}(p_2-2;\Delta)\right] \\
& -(p_2-2)\boldsymbol{\omega}\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{\omega}' \left\{2E\left[\chi_{p_2+2}^{-2}(\Delta) I(\chi_{p_2+2}^2(\Delta) \le p_2-2)\right]\right. \\
& -2E\left[\chi_{p_2+2}^{-2}(\Delta) I(\chi_{p_2+2}^2(\Delta) \le p_2-2)\right] \\
& \left.+(p_2-2)E\left[\chi_{p_2+2}^{-4}(\Delta) I(\chi_{p_2+2}^2(\Delta) \le p_2-2)\right]\right\}, \qquad (3.25)
\end{aligned}
$$

Proof of these theorems are similar to the proofs in Ahmed et al. (2007, p. 452). According to Theorem 3.8.4, for $\boldsymbol{B}_{12} = \boldsymbol{0}$, we have $\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12} = 0$ and $\boldsymbol{B}_{11.2} = \boldsymbol{B}_{11}$ and hence all the ADQR reduce to common value $\sigma^2\mathrm{tr}(\boldsymbol{M}\boldsymbol{B}_{11}^{-1})$ for all $\boldsymbol{\omega}$. Hence, all these variations become ADQR-equivalent. In the sequel we therefore assume that $\boldsymbol{B}_{12} \neq 0$ and the remaining discussion follows. Moreover, if we consider the Mahalanobis distance (loss) then $\boldsymbol{M} = \sigma^{-2}\boldsymbol{B}_{11.2}$ and $\mathrm{R}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}})$ reduces to $p_1$.

By comparing $\mathrm{R}(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}})$ and $\mathrm{R}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}})$, the following dominance condition holds. If $\boldsymbol{M} \in \boldsymbol{M}^{\boldsymbol{D}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ dominates $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ for any $\boldsymbol{\omega}$ in the sense of ADQR, where

$$
\boldsymbol{M}^{\boldsymbol{D}} = \left\{ \boldsymbol{M} : \frac{\mathrm{tr}(\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1})}{\mathrm{ch}_{\max}(\boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{M}\boldsymbol{B}_{11}^{-1}\boldsymbol{B}_{12}\boldsymbol{B}_{22.1}^{-1})} \ge \frac{p_2+2}{2} \right\}.
$$

Comparing $\mathrm{R}(\hat{\boldsymbol{\beta}}^{\mathrm{S}+})$ with $\mathrm{R}(\hat{\boldsymbol{\beta}}^{\mathrm{S}})$, we observe that $\hat{\boldsymbol{\beta}}^{\mathrm{S}+}$ dominates $\hat{\boldsymbol{\beta}}^{\mathrm{S}}$ for all the

values of $\boldsymbol{\omega}$, with strict inequality holding for some $\boldsymbol{\omega}$. Therefore, the risk of $\hat{\boldsymbol{\beta}}^{\mathrm{S}+}$ is also smaller than the risk of $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ in the entire parameter space and the upper limit is attained when $\Delta$ approaches $\infty$. $\mathrm{ADQR}(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}+})$ increases monotonically towards $\mathrm{R}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}})$ from below, as $\Delta$ moves away from 0. This implies that

$$\mathrm{R}(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}+}) \leq \mathrm{R}(\hat{\boldsymbol{\beta}}_1^{S}) \leq \mathrm{R}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}), \text{ for any } \boldsymbol{M} \in \boldsymbol{M}^{\boldsymbol{D}} \text{ and } \boldsymbol{\omega},$$

with strict inequality holding for some $\boldsymbol{\omega}$. Thus, we conclude that $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}$ and $\hat{\boldsymbol{\beta}}_1^{\mathrm{S}+}$ perform better than $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ in the entire parameter space induced by $\Delta$. The gain in risk over $\hat{\boldsymbol{\beta}}_1^{\mathrm{UE}}$ is substantial when $\Delta = 0$ or near.

## 3.9 Conclusion

In this chapter, we considered shrinkage estimation in the context of a partially linear model where the nonparametric component was approximated by a B-spline basis function. In our study, we developed and implemented a procedure for simultaneous sub-model selection to obtain shrinkage estimators, and compared their performance with lasso and adaptive lasso estimators. All the computations were performed using R (R Development Core Team, 2010). As an example, we analyzed a popular econometric data set, and used prediction errors and loglikelihood criteria to compare shrinkage and positive shrinkage estimators with those obtained from the models produced by lasso and adaptive lasso estimators. Our analyses showed that positive shrinkage and shrinkage estimators perform equally compared to the adaptive lasso estimators. In our analysis the candidate sub-models were obtained through stepwise variable selection procedure based on AIC and BIC. For the data example, the initial

selection via AIC, interactive model selection procedure based on deviance analysis was performed to obtain the final sub-model. As noted earlier, the final sub-model was the same as the one given by BIC.

We used the RMSE criterion to compare shrinkage estimators with an APE (lasso only) through Monte Carlo simulation. We found that positive-shrinkage estimators have smaller relative risk when the number of nuisance parameters in the model was large. Not surprisingly, APE performed better when the nuisance subset was small. We suspected that alongside the number of nuisance parameters, performance of the shrinkage and absolute penalty estimators may vary depending on the number of main parameters ($p_1$). In our study, we got an indication that, for an increased number of main parameters, APE may outperform shrinkage estimators. Behavior of the proposed estimators compared to lasso and adaptive lasso estimators for different sizes of main parameters is currently under investigation. In our setup, APE tend to outperform shrinkage estimators when the sample size gets larger.

Monte Carlo simulation study re-established the fact that the restricted estimators outperform the usual unrestricted estimators at or near the pivot ($\Delta^* = 0$). However, as we deviate away from the pivot, risk of the restricted estimator becomes large. On the other hand, shrinkage and positive-shrinkage estimators perform steadily throughout the alternative parameter sub-space.

To summarize, we observed that the positive part shrinkage estimator may be considered in a PLM when the nonparametric component is estimated by B-spline basis function. B-splines are easier to implement and to incorporate in a regression model when one considers uniform knots. In many practical situations, this might be a good alternative to the kernel-based estimators for the nonparametric part.

# Chapter 4

# Robust Shrinkage M-Estimation in Partially Linear Models

## 4.1 Introduction

In this chapter, we consider robust shrinkage M-estimation in a partially linear model (PLM). Ahmed et al. (2006) considered robust shrinkage-estimation of regression parameters in the presence of nuisance scale parameters when it is *a priori* suspected that the regression could be restricted to a linear subspace. They studied the asymptotic properties of variants of Stein-rule M-estimators (including the positive-part shrinkage M-estimators). We extend their work to a PLM and develop shrinkage M-estimators. In our work, we obtained risk-based robust shrinkage estimators of regression parameters.

PLM is more flexible than a linear model since it includes a nonlinear component along with the linear components. A PLM may provide a better alternative to the

classical linear regression in a situation where one or more covariates have nonlinear relationship with the response variable.

Robust regression models are designed to overcome some of the limitations of classical linear regression. For example, least squares regression is highly non-robust to outliers, and is subject to underlying assumptions. Violation of these assumptions will have serious impact on the validity of the fitted model. Similar to a classical linear regression model, a PLM may be affected due to the presence of outliers or violation of underlying assumptions.

In this chapter, we focus on robust shrinkage M-estimation of regression coefficients in a PLM where the nonparametric part is estimated by the kernel-based method. We consider four different error distributions, namely, standard normal, scaled normal, standard Laplace, and logistic distribution. We consider a PLM with scaled error term which is defined in (4.1).

### 4.1.1 Organization of the Chapter

In Section 4.2, we discuss robust M-estimation in a seminarametric model. The steps to estimate the nonparametric function in a PLM are discussed in detail. We reviewed the consistency and asymptotic normality of the estimators. Shrinkage M-estimators are defined in Section 4.3. Asymptotic bias and risk expressions for the M-estimators are derived in Section 4.5. Finally, Monte Carlo simulation results are presented in Section 4.6.

## 4.2 Semiparametric M-Estimation

Consider a PLM of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + g(T) + \sigma\boldsymbol{e}, \tag{4.1}$$

where $\boldsymbol{X}' = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_n')$, $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)'$, and $\boldsymbol{x}_i'$s are known row vectors of length $p$, $e_i$s are independent and identically distributed (iid) random variables having a continuous distribution, $F$, free from an unknown scale parameter $\sigma > 0$, $g(T)$ is an unknown real-valued function, and $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the vector of regression parameters. The $'$ denotes transpose of a vector or a matrix.

After estimating $g(\cdot)$ using kernel smoothing, we confine ourselves to the estimation of $\boldsymbol{\beta}$ based on the partial residuals which attains the usual parametric convergence rate $n^{-1/2}$ without under-smoothing the nonparametric component $g(\cdot)$ (Speckman, 1988).

Let us assume that $\{\boldsymbol{x}_i', t_i, y_i; i = 1, 2, \ldots, n\}$ satisfy model (4.1). If $\boldsymbol{\beta}$ is known to be the true parameter, then by $E(\epsilon) = 0$, we have

$$g(t_i) = E(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \quad i = 1, 2, \ldots, n.$$

A natural nonparametric estimator of $g(\cdot)$ given $\boldsymbol{\beta}$ is

$$g^*(t, \boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}).$$

Here, $W_{ni}$ are defined as

$$W_{ni} = \frac{K((t_i - t)/h)}{\sum_{j=1}^{n} K((t_j - t)/h)}, \tag{4.2}$$

with $K(\cdot)$ being a kernel function–a non-negative function integrable on $\mathfrak{R}$, and $h$ is a bandwidth parameter. We need to make the following assumptions on $W_{ni}$.

**Assumption 4.2.1.** The function $g(\cdot)$ satisfies the Lipschitz condition of order 1 on $[0, 1]$.

**Assumption 4.2.2.** The probability weight functions $W_{ni}(\cdot)$ satisfy

a) $\max_{1\le i\le n} \sum_{j=1}^{n} W_{ni}(t_j) = \mathcal{O}(1)$

b) $\max_{1\le i,\, j\le n} \sum_{j=1}^{n} W_{ni}(t_j) = \mathcal{O}(n^{-2/3})$,

c) $\max_{1\le j\le n} \sum_{i=1}^{n} W_{ni}(t_j) I(|t_i - t_j| > c_n) = \mathcal{O}(d_n)$, where $I$ is the indicator function, $c_n$ satisfies $\limsup_{n\to\infty} n c_n^3$, and $d_n$ satisfies $\limsup_{n\to\infty} n d_n^3 < \infty$.

*Remark* 2. The usual polynomial and trigonometric functions satisfy Assumption 4.2.1

*Remark* 3. Under regular conditions, the Nadaraya-Watson kernel weights, Priestley and Chao kernel weights, locally linear weights and Gasser-Müller kernel weights satisfy Assumption 4.2.2. If we consider the pdf of $U[-1, 1]$ as the kernel function as

$$K(t) = \frac{1}{2} I_{[-1,1]}(t),$$

with $t_i = \frac{i}{n}$, and the bandwidth $cn^{-1/3}$ where $c$ is constant, then the Priestley and

Chao kernel weights satisfy Assumption 4.2.2, and the weights are

$$W_{ni}(t) = \frac{1}{2cn^{\frac{2}{3}}} \left( \left| t - \frac{i}{n} \right| \leq cn^{-\frac{1}{3}} \right)^{(t)}.$$

For a detailed discussion on the assumptions above, see Ahmed et al. (2007).

Now, we define $\gamma_0(t) = E(y|T = t)$ and $\boldsymbol{\gamma}(t) = (\gamma_1(t), \gamma_2(t), \ldots, \gamma_n(t))'$, where $\gamma_j(t) = E(x_j|T = t)$. We estimate $\boldsymbol{\beta}$ using

$$\hat{\boldsymbol{\beta}} = \arg\min SS(\boldsymbol{\beta}) = (\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}, \tag{4.3}$$

with

$$SS(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - x_i'\boldsymbol{\beta} - g^*(t_i, \boldsymbol{\beta})\right)^2 = \sum_{i=1}^{n} (\tilde{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\beta})^2. \tag{4.4}$$

Here, $\widetilde{\boldsymbol{Y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)'$, $\widetilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n)'$, $\tilde{y}_i = y_i - \gamma_0(t)$, and $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{\gamma}(t)$ for $i = 1, 2, \ldots, n$. The conditional expectations $\gamma_0(t)$ and $\boldsymbol{\gamma}(t)$ can be obtained using classical nonparametric approach through

$$\hat{\gamma}_0(t) = \sum_{i=1}^{n} W_{ni}(t) y_i, \text{ and}$$
$$\hat{\gamma}_j(t) = \sum_{i=1}^{n} W_{ni}(t) x_{ij}$$

where $W_{ni}(t)$ is defined in (4.2). Clearly, once we obtain the estimates $\hat{\gamma}_0(t)$ and $\hat{\gamma}_j(t)$, they can be plugged into (4.4) prior to the estimation of regression parameters $\boldsymbol{\beta}$.

The above procedure was independently proposed by Speckman (1988) and Denby (1986). Chen and Shiau (1994) mentioned that the procedure can be related to the

partial regression procedure in linear regression. A similar approach was taken by Ahmed et al. (2007) in estimating the nonparametric component in a partially linear model.

We obtain the robust M-estimators of the parameters of a PLM using a two-step procedure as follows:

Step 1 We first estimate $\gamma_0(t)$ and $\gamma_j(t)$ through kernel smoothing as described above. We denote the estimates by $\hat{\gamma}_0(t)$ and $\hat{\gamma}_j(t)$, respectively.

Step 2 The estimates in Step 1 are then plugged into the model. Then the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by regressing the residuals $y_i - \hat{\gamma}_0(t)$ and $\boldsymbol{x}_i - \hat{\boldsymbol{\gamma}}(t)$ using a robust procedure. We denote the residuals as $\hat{r}_i = y_i - \hat{\gamma}_0(t)$ and $\boldsymbol{u}_i = \boldsymbol{x}_i - \hat{\boldsymbol{\gamma}}(t_i)$.

### 4.2.1 Consistency and Asymptotic Normality

Let $\tilde{\rho}$ and $\widetilde{W}$ be score and weight functions, respectively. The asymptotic distribution of $\boldsymbol{\beta}$ is defined as a solution of

$$\sum_{i=1}^{n} \tilde{\rho}\left(\frac{\hat{r}_i - \hat{\boldsymbol{u}}_i\hat{\boldsymbol{\beta}}}{s_n}\right)\widetilde{W}(||\hat{\boldsymbol{u}}_i||)\hat{\boldsymbol{u}}_i = 0, \tag{4.5}$$

where $\hat{r}_i = y_i - \hat{\gamma}_0(t)$, $\boldsymbol{u}_i = \boldsymbol{x}_i - \hat{\boldsymbol{\gamma}}(t_i)$, and $s_n$ is an estimate of the residual scale. Also consider that $\hat{\gamma}_0(t)$ and $\hat{\boldsymbol{\gamma}}(t_i)$ are the consistent estimators of $\gamma_0(t)$ and $\boldsymbol{\gamma}(t_i)$, respectively. Now, denote the random vector $(R(T), \boldsymbol{U}(T)')'$ with the same distribution as $(r_i, \boldsymbol{u}_i')'$.

Consistency of the regression parameters in a semiparametric model has been

proved in great detail in Bianco and Boente (2004). We omit the details but present only the set of assumptions, lemma, and theorem which are needed for proving asymptotic normality and consistency of the estimators.

To derive the asymptotic distribution of $\boldsymbol{\beta}$ we must have $t_i$ in a compact set, so without loss of generality, we assume that $t_i \in [0,1]$. We need the following set of assumptions. See Bianco and Boente (2004) for details.

A1 $\tilde{\rho}$ is odd, bounded, continuous, and twice differentiable with bounded derivative $\tilde{\rho}'$ and $\tilde{\rho}''$ such that $\phi_1(t) = t\tilde{\rho}'(t)$ and $\phi_2(t) = t\tilde{\rho}''(t)$ are bounded.

A2 $E(\tilde{W}(||\boldsymbol{U}(T)||)||\boldsymbol{U}(T)'||^2) < \infty$ and the matrix

$$\boldsymbol{A} = E\left(\tilde{\rho}'\left(\frac{R(T) - \boldsymbol{U}(T)'\boldsymbol{\beta}}{\sigma}\right)\widetilde{W}(||\boldsymbol{U}(T)||)\boldsymbol{U}(T)\boldsymbol{U}(T)'\right)$$

is nonzero.

A3 $\widetilde{W}(u) = \tilde{\rho}_1(u)u^{-1} > 0$ is a bounded function which satisfies the Lipschitz condition of order 1. Further, $\tilde{\rho}_1$ is bounded with bounded derivative.

A4 $E(\widetilde{W}(||\boldsymbol{U}(T)||)\boldsymbol{U}(T)|T = t) = 0$ for almost all $t$.

A5 The functions $\boldsymbol{x}_j(t), 0 \leq j \leq p$ are continuous in $[0,1]$ with continuous first derivative.

*Remark* 4. According to Robinson (1988), condition A2 is needed so that no element of $\boldsymbol{X}$ can be predictable by $T$. A2 guarantees that there is no multicollinearity in the columns of $\boldsymbol{X} - \tilde{\boldsymbol{X}}_j(T)$. In other words, $\boldsymbol{X}$ has to be free from multicollinearity.

*Remark* 5. Again, according to Robinson (1988), condition A5 is a standard requirement in kernel estimation in semiparametric models in order to guarantee asymptotic

normality.

**Lemma 4.2.1.** Let $(y_i, \boldsymbol{x}_i', t_i)', 1 \leq i \leq n$ be independent random vectors satisfying (4.1) with $\epsilon_i$ independent of $(\boldsymbol{x}_i', t_i)'$. Assume that $t_i$ are random variable with $t_i \in [0, 1]$. Denote $(R(T), \boldsymbol{U}(T)')'$ a random vector with the same distribution as

$$(r_i, \boldsymbol{u}_i')' = (y_i - \hat{\gamma}_0(t_i), [\boldsymbol{x}_i - \hat{\boldsymbol{\gamma}}(t_i)]')'.$$

Further, let $\hat{\boldsymbol{\gamma}}_j(t_i)$, $0 \leq j \leq p$ be the estimates of $\gamma_j(t_i)$ such that

$$\sum_{t \in [0,1]} |\hat{\gamma}_j(t) - \gamma_j(t)| \xrightarrow{p} 0, \quad 0 \leq j \leq p.$$

If $\tilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ and $s_n \xrightarrow{p} \sigma$, then under the stated assumptions A1-A3, $\boldsymbol{A}_n \xrightarrow{p} \boldsymbol{A}$ where $\boldsymbol{A}$ is defined in A2, and

$$\boldsymbol{A}_n = n^{-1} \sum_{i=1}^{n} \tilde{\rho}' \left( \frac{\hat{r}_i - \hat{\boldsymbol{u}}_i' \widetilde{\boldsymbol{\beta}}}{s_n} \right) \widetilde{W}(||\hat{\boldsymbol{u}}_i||) \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i'.$$

Here, $\xrightarrow{p}$ denotes convergence in probability.

*Proof.* The proof is available in the appendix of Bianco and Boente (2004).

**Theorem 4.2.1.** Let $(y_i, \boldsymbol{x}_i', t_i)', 1 \leq i \leq n$ be independent random vectors satisfying (4.1) with $\epsilon_i$ independent of $(\boldsymbol{x}_i', t_i)'$. Assume that $t_i$ are random variables with $t_{i_n} \in [0, 1]$. Denote $(R(T), \boldsymbol{U}(T)')'$ a random vector with the same distribution as

$$(r_i, \boldsymbol{u}_i')' = (y_i - \gamma_0(t_i), [\boldsymbol{x}_i - \boldsymbol{\gamma}(t_i)]')'.$$

Further, let $\hat{\gamma}_j(t)$, $0 \leq j \leq p$ be estimates of $\gamma_j(t)$ such that first derivative of $\hat{\gamma}_j(t)$

exists and is continuous, and

$$n^{1/4} \sum_{t\in[0,1]} |\hat{\gamma}_j(t) - \gamma_j(t)| \xrightarrow{p} 0, \quad 0 \le j \le p \tag{4.6}$$

$$\sum_{t\in[0,1]} |\hat{\gamma}_j(t) - \gamma_j(t)| \xrightarrow{p} 0, \quad 0 \le j \le p \tag{4.7}$$

Then, if $s_n \xrightarrow{p} \sigma$, under A1-A5,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{Q})$$

with $\boldsymbol{Q} = \boldsymbol{A}^{-1}\boldsymbol{\Sigma}(\boldsymbol{A}^{-1})'$, where $\boldsymbol{A}$ is defined previously in A2 and

$$\boldsymbol{\Sigma} = \sigma^2 E\left(\tilde{\rho}^2\left(\frac{R(T) - \boldsymbol{U}(T)'\boldsymbol{\beta}}{\sigma}\right)\widetilde{W}^2(||\boldsymbol{U}(T)||)\boldsymbol{U}(T)\boldsymbol{U}(T)\right)$$

*Proof.* The proof is available in Bianco and Boente (2004).

## 4.3 Shrinkage M-Estimation

Similar to the previously defined shrinkage estimators, a Stein-type M-estimator (SM) $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$ of $\boldsymbol{\beta}_1$ can be defined as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})\left\{1 - \kappa\psi_n^{-1}\right\}, \text{ where } \kappa = p_2 - 2, \quad p_2 \ge 3.$$

where $\psi_n$ is a test statistic defined in (4.16). And the positive-rule Stein-type M-estimator (SM+) has the form

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})\left\{1 - \kappa\psi_n^{-1}\right\}^+, \quad p_2 \ge 3,$$

where $z^+ = \max(0, z)$. Alternatively, this can be written as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}+} = \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}} + (\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) \left\{1 - \kappa\psi_n^{-1}\right\} I(\psi_n < \kappa), \quad p_2 \geq 3.$$

For a suitable absolutely continuous function $\rho : \mathfrak{R} \to \mathfrak{R}$, with derivative $\phi$, an M-estimator of $\boldsymbol{\beta}$ is defined as a solution of the minimization

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(\tilde{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\beta}). \tag{4.8}$$

Generally, an M-estimator is regression-equivariant, i.e.,

$$\boldsymbol{M}_n(c\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{a}) = c\boldsymbol{M}_n(\boldsymbol{Y}) + c\,\boldsymbol{a}, \quad \text{for } \boldsymbol{a} \in \mathfrak{R}_p$$

and robustness depends on the choice of $\rho(\cdot)$. But it is generally not scale-equivariant. That is, it may not satisfy

$$\boldsymbol{M}_n(c\boldsymbol{Y}) = c\boldsymbol{M}_n(\boldsymbol{Y}) \quad \text{for } c > 0.$$

To have the estimators scale and regression equivariant, we need to studentize them. The studentized M-estimator is defined as as solution of the minimization

$$\min_{\boldsymbol{b} \in \mathfrak{R}^p} \sum_{i=1}^{n} \rho\left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\beta}}{S_n}\right) \tag{4.9}$$

where $S_n = S_n(\boldsymbol{Y}) \geq 0$ is an appropriate scale statistic that is regression equivariant and scale equivariant, i.e.,

$$S_n(c(\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{a})) = cS_n(\boldsymbol{Y}) \quad \text{for } \boldsymbol{a} \in \mathfrak{R}_p \text{ and } c > 0$$

According to Jurečková and Sen (1996), the minimization in (4.9) should be supplemented by a rule how to define $\boldsymbol{M}_n$ in the case when $S_n(\boldsymbol{Y}) = 0$. However, in general, this happens with probability zero, and the specific rule does not affect the asymptotic properties of $\boldsymbol{M}_n$. There are additional regularity conditions needed with (4.9), which we present in the following. Details may be found in Jurečková and Sen (1996, page 217)].

### 4.3.1 Regularity Conditions

Here, we list the regularity conditions needed for the minimization problem in (4.9). Detailed discussions about these conditions can be found in Jurečková and Sen (1996, p. 217-218).

For the studentized M-estimators, consider that $\phi = \rho'$ can be decomposed as

$$\phi = \phi_1 + \phi_2 + \phi_3, \tag{4.10}$$

where $\phi_1$ is an absolutely continuous function with absolutely continuous derivative, $\phi_2$ is a continuous piecewise linear function that is constant in a neighbourhood of $\pm\infty$, and $\phi_3$ is a nondecreasing step function.

The following conditions are imposed on (4.9).

RC.1. $S_n(Y)$ is regression invariant and scale equivariant, $S_n > 0$ a.s., and

$$\sqrt{n}(S_n - S) = O_p(1)$$

for some functional $S = S(F) > 0$.

RC.2. The function $h(t) = \int \rho((z-t)/S)dF(z)$ has the unique minimum at $t = 0$.

RC.3. For some $\delta > 0$ and $\eta > 1$,

$$\int_{\infty}^{-\infty} \left\{ |z| \sup_{|u|\leq\delta} \sup_{|v|\leq\delta} \left| \phi_1'' \left( \frac{e^{-v}(z+u)}{S} \right) \right| \right\}^{\eta} dF(z) < \infty$$

and

$$\int_{\infty}^{-\infty} \left\{ |z|^2 \sup_{|u|\leq\delta} \left| \frac{\phi_1''(z+u)}{S} \right| \right\}^{\eta} dF(z) < \infty$$

where $\phi_1'(z) = \frac{d}{dz}\phi_1(z)$, and $\phi_1''(z) = \frac{d^2}{dz^2}\phi_1(z)$.

RC.4. $\phi_3$ is a continuous, piecewise linear function with knots at $\mu_1, \ldots, \mu_k$, which is constant in a neighborhood of $\pm\infty$. Hence the derivative $\phi_3'$ of $\phi_3$ is a step function

$$\phi_3'(z) = \alpha_\nu \quad \text{for } \mu_\nu < z < \mu_{\nu+1}, \nu = 0, 1, \ldots, k,$$

where $\alpha_0, \alpha_1, \ldots, \alpha_k \in \mathfrak{R}_1$, $\alpha_0 = \alpha_k = 0$ and $\infty = \mu_0 < \mu_1 < \cdots < \mu_k < \mu_{k+1} = \infty$. Further, we assume that $f(z) = \frac{dF(z)}{dz}$ is bounded in neighbourhood of $S_{\mu_j}, j = 1, 2, \ldots, k$.

RC.5. $\phi_3(z) = \lambda_\nu$ for $q_\nu < z \leq q_{\nu+1}$, $\nu = 1, 2, \ldots, m$ where $-\infty = q_0 < q_1 < \cdots < q_m < q_{m+1} = \infty$, $-\infty < \lambda_0 < \lambda_1 < \cdots < \lambda_m < \infty$. We further assume that $f'(z)$ and $f''(z)$ are bounded in the neighbourhood of $S_{q_j}, j = 1, 2, \ldots, m$.

Now, to define the shrinkage M-estimators, we redefine the matrix $\boldsymbol{A}_n$ as

$$\boldsymbol{C} = \boldsymbol{A}_n \boldsymbol{A}_n' = \begin{pmatrix} \boldsymbol{A}_{n11}'\boldsymbol{A}_{n11} & \boldsymbol{A}_{n21}'\boldsymbol{A}_{n12} \\ \boldsymbol{A}_{n21}'\boldsymbol{A}_{n21} & \boldsymbol{A}_{n22}'\boldsymbol{A}_{n22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix}$$

Also, we define

$$\boldsymbol{C}_{22.1} = \boldsymbol{C}_{22} - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12},$$

which we shall require later. Notice that, if $\boldsymbol{C}_{21} = \boldsymbol{0}$, $\boldsymbol{C}_{22.1} = \boldsymbol{C}_{22}$. Otherwise, $\boldsymbol{C}_{22} - \boldsymbol{C}_{22.1}$ is positive semi-definite, as we shall require.

A studentized unrestricted M-estimator (UME) of $\boldsymbol{\beta}$ is defined as a solution of (4.9). Let us denote it by

$$\hat{\boldsymbol{\beta}}^{\mathrm{UM}} = \left(\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}\right), \left(\hat{\boldsymbol{\beta}}_2^{\mathrm{UM}}\right)\right).$$

A studentized restricted M-estimator of $\boldsymbol{\beta}_1$ is obtained by minimizing

$$\min_{\boldsymbol{b}\in\Re^{p_1}} \sum_{i=1}^{n} \rho\left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_{i1}'\boldsymbol{b}}{S_n}\right), \tag{4.11}$$

and denote it by $\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}$. Here, $S_n$ is regression-invariant, and so is not affected by the restricted environment. Since $\rho(\cdot)$ is assumed to have derivative $\phi(\cdot)$, we rewrite $\hat{\boldsymbol{\beta}}^{\mathrm{UM}}$ as a solution of

$$\boldsymbol{M}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i\, \phi\left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\theta}}{S_n}\right) = \boldsymbol{0}. \tag{4.12}$$

In other words,

$$\boldsymbol{M}_n(\hat{\boldsymbol{\beta}}^{\mathrm{UM}}) = \boldsymbol{0}.$$

Similarly, $\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}$ is a solution of

$$\boldsymbol{M}_{n_1}(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_{i1}\, \phi\left(\frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_{i1}'\boldsymbol{\theta}_1}{S_n}\right) = \boldsymbol{0}. \tag{4.13}$$

Now, let

$$\hat{\boldsymbol{M}}_{n_2}^{\mathrm{RM}} = \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_{i2}\, \phi \left( \frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_{i1}' \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}}{S_n} \right). \tag{4.14}$$

Recall that $\hat{\boldsymbol{M}}_{n_2}^{\mathrm{RM}}$ is a $p_1$-vector and $\boldsymbol{M}_{n_1}$ is a $p_2$-vector. Let us denote also

$$\hat{\sigma}_{\phi\, n\, R}^2 = (n - p_2)^{-1} \sum_{i=1}^{n} \phi^2 \left( \frac{\tilde{y}_i - \tilde{\boldsymbol{x}}_{i1}' \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}}{S_n} \right). \tag{4.15}$$

Now, considering the studentized environment for our problem, a suitable test statistic can be formulated following the procedure discussed in Jurečková and Sen (1996, section 10.2), as

$$\psi_n = \frac{\left[\hat{\boldsymbol{M}}_{n_2}^{\mathrm{RM}}\right]' \boldsymbol{C}_{22.1}^{-1} \left[\hat{\boldsymbol{M}}_{n_2}^{\mathrm{RM}}\right]}{\hat{\sigma}_{\Phi\, n\, R}}. \tag{4.16}$$

Directly applying the Lemma 5.5.1 in Jurečková and Sen (1996, page 220), it can be shown that

$$\psi_n \xrightarrow{d} \chi_{p_2}^2 \quad \text{under } H_0.$$

For details of proof, please see the above reference. However, under (local) alternative hypotheses

$$\psi_n \xrightarrow{d} \chi_{p_2,\,\Delta}^2,$$

where $\Delta$ is the noncentrality parameter.

It is to be mentioned here that unlike least-squares estimators, M-estimators are not linear. Even if the distribution function $F$, is normal, the finite sample distribution theory of M-estimators is not simple. Asymptotic methods [Sen and Saleh (1987) Jurečková and Sen (1996)] have been used to overcome this difficulty. However, these asymptotic methods are related primarily to convergence in distribution, which may not generally guarantee convergence in quadratic risk (Ahmed et al., 2006). This is

taken care of with the introduction of asymptotic distributional risk (ADR) (Sen, 1986), which is based on the concept of a shrinking neighbourhood of the pivot for which the ADR serves a useful and interpretable role in they asymptotic risk analysis.

## 4.4 Asymptotic Properties of the Estimators

In this section, we derive the asymptotic distributions of the estimators and the test statistic $\psi_n$. This facilitates in finding the asymptotic distributional bias (ADB), asymptotic distributional quadratic bias (ADQB), and asymptotic distributional quadratic risk (ADQR) of the estimator of $\boldsymbol{\beta}$.

Under the assumed regularity conditions and as

$$\lim_{n\to\infty} \frac{C_n}{n} = \boldsymbol{Q} \tag{4.17}$$

where

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix},$$

it is known that under fixed alternative, $\boldsymbol{\beta}_2 = \mathbf{0}$,

$$\frac{\psi_n}{n} \to \gamma(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2; \boldsymbol{Q}) > 0 \quad \text{as } n \to \infty$$

such that the shrinkage factor $\kappa\psi_n^{-1} = \mathcal{O}_p(n^{-1})$. This implies, asymptotically, there is no shrinkage effect. Therefore, to obtain meaningful asymptotics, we consider a class

of local alternatives, $\{K_n\}$, given by

$$K_n : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2n} = \frac{\boldsymbol{\omega}}{\sqrt{n}}, \tag{4.18}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_{p_2})' \in \Re^{p_2}$ is a fixed vector, and $||\boldsymbol{\omega}|| < \infty$ so that the null hypothesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ reduces to $H_0 : \boldsymbol{\omega} = \mathbf{0}$.

It is to be reminded that under such local alternatives, the estimators $\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$, and $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}}$ may not be asymptotically unbiased for $\boldsymbol{\beta}_1$. Therefore, we consider a quadratic loss function. For an estimator $\boldsymbol{\beta}_1^*$ and a positive-definite matrix $\boldsymbol{W}$, we define the loss function of the form

$$L(\boldsymbol{\beta}_1^*; \boldsymbol{\beta}_1) = n(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)'\boldsymbol{W}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1).$$

These loss functions are generally known as weighted quadratic loss functions, where $\boldsymbol{W}$ is the weighting matrix. For $\boldsymbol{W} = \boldsymbol{I}$, it is the simple squared error loss function. The expectation of the loss function

$$E[L(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1); \boldsymbol{W}] = R[(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1); \boldsymbol{W}]$$

is called the risk function, which can be written as

$$\begin{aligned} R[(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1); \boldsymbol{W}] &= nE[(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)'\boldsymbol{W}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)] \\ &= n\,\mathrm{tr}[\boldsymbol{W}\{E(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)'\}] \\ &= \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Omega}^*), \end{aligned} \tag{4.19}$$

where $\boldsymbol{\Omega}^*$ is the covariance matrix of $\sqrt{n}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)$. Whenever

$$\lim_{n\to\infty} \hat{\boldsymbol{\Omega}}_n^* = \hat{\boldsymbol{\Omega}}^*$$

exists, the asymptotic risk is defined by

$$R_n(\boldsymbol{\beta}_{1n}^*, \boldsymbol{\beta}_1; \boldsymbol{W}) \to R(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1; \boldsymbol{W}) = \text{tr}(\boldsymbol{W}\hat{\boldsymbol{\Omega}}^*).$$

Consider the asymptotic cumulative distribution function (cdf) of $\sqrt{n}(\boldsymbol{\beta}_{1n}^* - \boldsymbol{\beta}_1)$ under $\{K_n\}$ exists, and is defined as

$$G(\boldsymbol{y}) = P\left[\lim_{n\to\infty} \sqrt{n}(\boldsymbol{\beta}_{1n}^* - \boldsymbol{\beta}_1) \leq \boldsymbol{y}\right].$$

This is known as the asymptotic distribution function (ADF) of $\boldsymbol{\beta}_1^*$. Suppose that $G_n \to G$ at all points of continuity as $n \to \infty$, and let $\hat{\boldsymbol{\Omega}}^*$ be the covariance matrix of $G$. Then the ADR of $\boldsymbol{\beta}_{1n}$ is defined as

$$R(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1; \boldsymbol{W}) = \text{tr}(\boldsymbol{W}\boldsymbol{\Omega}_G^*)$$

As noted in Ahmed et al. (2006), if $G_n \to G$ in second moment, then ADR is the asymptotic risk. However, this is a stronger mode of convergence, and is hard to analytically prove for shrinkage M-estimators. Therefore, they suggested using asymptotic distributional risk.

Now let

$$\Gamma = \int\int \cdots \int \boldsymbol{y}\boldsymbol{y}' dG(\boldsymbol{y})$$

be the dispersion matrix which is obtained from ADF. The asymptotic distributional quadratic risk (ADQR) may be defined as

$$R(\boldsymbol{\beta}_1^*; \boldsymbol{\beta}_1) = \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Gamma}). \tag{4.20}$$

Here $\boldsymbol{\Gamma}$ is the asymptotic distributional mean squared error (ADMSE) of the estimators.

To derive the ADB and ADQB of the estimators, we present two important theorems.

**Theorem 4.4.1.** Consider an absolutely continuous function $f(\cdot)$ with derivative $f'(\cdot)$ which exists everywhere, and finite Fisher information

$$I(f) = \int_R \left(\frac{-f'(x)}{f(x)}\right)^2 dF(x) < \infty$$

Under $\{K_n\}$ and the assumed regularity conditions, $\psi_n$ has asymptotically a non-central chi-square distribution with non-centrality parameter $\Delta = \boldsymbol{\omega}'\boldsymbol{Q}_{22.1}\boldsymbol{\omega}\gamma^{-2}$. Here

$$\gamma^2 = \frac{\int_R \phi^2(y)dF(y)}{\int_R \phi(x)[-f'(x)/f(x)]dF(x)}, \tag{4.21}$$

and $\phi(\cdot)$ is defined in Section 4.3.1.

**Theorem 4.4.2.** Under the assumed regularity conditions, as $n \to \infty$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathrm{UM}} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \gamma^2\boldsymbol{Q}^{-1}). \tag{4.22}$$

Proofs of these theorems are available in Jurečková and Sen (1996).

## 4.5 Asymptotic Bias and Risk

**Theorem 4.5.1.** Under the local alternative $K_n$ and the assumed regularity conditions, we have as $n \to \infty$

(i) $\eta_1 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \boldsymbol{\beta}_1) \xrightarrow{d} N(\mathbf{0}, \gamma^2 \boldsymbol{Q}_{11.2}^{-1})$

(ii) $\eta_2 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) \xrightarrow{d} N(\boldsymbol{\delta}, \boldsymbol{\Sigma}^*), \quad \boldsymbol{\delta} = -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{\omega}$

(iii) $\eta_3 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}} - \boldsymbol{\beta}_1) \xrightarrow{d} N(-\boldsymbol{\delta}, \boldsymbol{\Omega}^*) \quad \boldsymbol{\Omega}^* = \gamma^2 \boldsymbol{Q}_{11}^{-1}$

Also, under $\{K_n\}$

$$\sqrt{n}\left((\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \boldsymbol{\beta}_1)', (\hat{\boldsymbol{\beta}}_2^{\mathrm{UM}} - n^{-\frac{1}{2}}\boldsymbol{\omega})'\right)' \xrightarrow{d} N(\mathbf{0}, \gamma^2 \boldsymbol{Q}^{-1})$$

where

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix}$$

Now, let us denote the joint distributions as follows:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \sim N_{p_1+p_2}\left[\begin{pmatrix} \mathbf{0} \\ \boldsymbol{\delta} \end{pmatrix}, \begin{pmatrix} \gamma^2 \boldsymbol{Q}_{11.2}^{-1} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}^* \end{pmatrix}\right]$$

$$\begin{pmatrix} \eta_2 \\ \eta_3 \end{pmatrix} \sim N_{p_1+p_2}\left[\begin{pmatrix} \boldsymbol{\delta} \\ -\boldsymbol{\delta} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}^* & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}^* \end{pmatrix}\right]$$

Now we derive $\boldsymbol{\Sigma}_{12}$ as

$$\begin{aligned}\boldsymbol{\Sigma}_{12} &= Cov(\eta_1, \eta_2) \\ &= Cov(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) \\ &= Cov(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}) - Cov(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) \\ &= Var(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}) - Cov(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) \\ &= \gamma^2 \boldsymbol{Q}_{11.2}^{-1} - Cov(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})\end{aligned}$$

where

$$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) &= Cov(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} + \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\hat{\boldsymbol{\beta}}_2^{\mathrm{UM}}) \\ &= Var(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}) + Cov(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}, \hat{\boldsymbol{\beta}}_2^{\mathrm{UM}})[\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}]' \\ &= \gamma^2 \boldsymbol{Q}_{11.2}^{-1} + \gamma^2 \boldsymbol{Q}_{12}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\end{aligned}$$

Therefore,

$$\begin{aligned}\boldsymbol{\Sigma}_{12} &= \gamma^2 \boldsymbol{Q}_{11.2}^{-1} - \gamma^2 \boldsymbol{Q}_{11.2}^{-1} - \gamma^2 \boldsymbol{Q}_{12}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} \\ &= -\gamma^2 \boldsymbol{Q}_{12}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}^* &= \boldsymbol{\Omega}^* - \gamma^2 \boldsymbol{Q}_{11.2}^{-1} + \boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}_{21} \\ &= \gamma^2(\boldsymbol{Q}_{11}^{-1} - \boldsymbol{Q}_{11.2}^{-1} - 2\boldsymbol{Q}_{12}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1})\end{aligned}$$

### 4.5.1 Bias Performance

The asymptotic distributional bias (ADB) of an estimator $\boldsymbol{\beta}^*$ is defined as

$$\text{ADB}(\boldsymbol{\beta}^*) = E\left\{\lim_{n\to\infty} n^{\frac{1}{2}}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right\}.$$

**Theorem 4.5.2.** Under the assumed regularity conditions and theorem above, and under $\{K_n\}$, the ADB of the estimators are as follows:

$$\begin{aligned}
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) &= \mathbf{0} \\
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{RM}}) &= -\boldsymbol{\delta} \\
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) &= \kappa\boldsymbol{\delta} E\left\{\chi^{-2}_{p_2+2}(\Delta)\right\} \\
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) &= ADB(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) - \boldsymbol{\delta}\left[H_{p_2+2}(\kappa,\Delta) - E\left\{\kappa\chi^{-2}_{p_2+2}(\Delta) I(\chi^2_{p_2+2}(\Delta) < \kappa)\right\}\right]
\end{aligned}$$

*Proof.* Obviously, $\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) = 0$

$$\begin{aligned}
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{RM}}) &= E\left\{\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{RM}} - \boldsymbol{\beta}_1)\right\} \\
&= E\left\{\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} + \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\hat{\boldsymbol{\beta}}_2^{\text{UM}} - \boldsymbol{\beta}_1)\right\} \\
&= E\left\{\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} - \boldsymbol{\beta}_1)\right\} + E\left\{\lim_{n\to\infty} \sqrt{n}(\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\hat{\boldsymbol{\beta}}_2^{\text{UM}})\right\} \\
&= E\left\{\lim_{n\to\infty} \sqrt{n}\left(\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\hat{\boldsymbol{\beta}}_2^{\text{UM}})\right)\right\} \\
&= \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{\omega} \\
&= -\boldsymbol{\delta}.
\end{aligned}$$

$$\begin{aligned}
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) &= E\left\{\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}-\boldsymbol{\beta}_1)\right\} \\
&= E\left\{\lim_{n\to\infty}\left(\sqrt{n}\hat{\boldsymbol{\beta}}_1^{\text{SM}}-\sqrt{n}\boldsymbol{\beta}_1\right)\right\} \\
&= E\left\{\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}-\hat{\boldsymbol{\beta}}_1^{\text{RM}})(-\kappa\psi_n^{-1})\right\} \\
&= -\kappa E\left\{\eta_2\psi_n^{-1}\right\} \\
&= -\kappa(-\boldsymbol{\delta})E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\} \\
&= \kappa\boldsymbol{\delta}E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\}.
\end{aligned}$$

$$\begin{aligned}
\text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) &= E\left\{\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}-\boldsymbol{\beta}_1)\right\} \\
&= E\left\{\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}-\boldsymbol{\beta}_1)-\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}-\hat{\boldsymbol{\beta}}_1^{\text{RM}})(1-\kappa\psi_n^{-1})I(\psi_n<\kappa)\right\} \\
&= E\left\{\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}-\boldsymbol{\beta}_1)\right\}-E\left\{\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}-\hat{\boldsymbol{\beta}}_1^{\text{RM}})(1-\kappa\psi_n^{-1})I(\psi_n<\kappa)\right\} \\
&= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}})-E\left\{\eta_2(1-\kappa\psi_n^{-1})I(\psi_n<\kappa)\right\} \\
&= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}})-\boldsymbol{\delta}E\left\{\left(1-\kappa\chi_{p_2+2}^{-2}(\Delta^2)\right)I\left(\chi_{p_2+2}^2(\Delta^2)<\kappa\right)\right\} \\
&= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}})-\boldsymbol{\delta}E\left\{I\left(\chi_{p_2+2}^2(\Delta)\right)<\kappa\right\}-\boldsymbol{\delta}E\left\{\kappa\chi_{p_2+2}^{-2}(\Delta)I\left(\chi_{p_2+2}^2(\Delta)<\kappa\right)\right\} \\
&= \text{ADB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}})-\boldsymbol{\delta}\left[H_{p_2+2}(\kappa,\Delta)-E\left\{\kappa\chi_{p_2+2}^{-2}(\Delta)I\left(\chi_{p_2+2}^2(\Delta)<\kappa\right)\right\}\right].
\end{aligned}$$

The bias expressions for all the estimators are not in scalar form. We therefore convert them into quadratic form. Let us define the asymptotic distributional quadratic

bias (ADQB) of an estimator $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}_1$ by

$$ADQB(\boldsymbol{\beta}^*) = [ADB(\boldsymbol{\beta}^*)]'\boldsymbol{\Sigma}[ADB(\boldsymbol{\beta}^*)]$$

where $\boldsymbol{\Sigma}^{-1}$ is the dispersion matrix of $\hat{\boldsymbol{\beta}}_1^{\text{UM}}$ as $n \to \infty$. In our case, the dispersion matrix is $\boldsymbol{Q}_{11}$.

Using the definition, the asymptotic quadratic distributional bias of the various estimators are derived below.

$$\begin{aligned}
\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{UM}}) &= \mathbf{0}, \\
\text{ADQR}(\hat{\boldsymbol{\beta}}_1^{\text{RM}}) &= \boldsymbol{\omega}'\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{\omega} \\
\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{SM}}) &= \kappa^2\boldsymbol{\delta}'\boldsymbol{Q}_{11}^{-1}\boldsymbol{\delta}\left[E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\}\right]^2 \\
\text{ADQB}(\hat{\boldsymbol{\beta}}_1^{\text{SM+}}) &= \boldsymbol{\delta}'\boldsymbol{Q}_{11}\boldsymbol{\delta}\left[H_{p_2+2}(\kappa,\Delta) - E\left\{\kappa\chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^2(\Delta) < \kappa)\right\}\right].
\end{aligned}$$

In the following, we derive the expressions for asymptotic distributional mean square error (ADMSE). Let us denote it by $\boldsymbol{\Gamma}$.

The ADMSE's are listed below

$$
\begin{aligned}
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}) &= \gamma^2 \boldsymbol{Q}_{11.2}^{-1} \\
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) &= \gamma^2 \boldsymbol{Q}_{11}^{-1} + \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\boldsymbol{Q}_{21}Q_{11}^{-1} \\
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) &= \gamma^2 \boldsymbol{Q}_{11.2}^{-1} - 2\kappa \left[ E(\chi_{p_2+2}^{-2}(\Delta))\Sigma_{21} + \boldsymbol{\delta}\boldsymbol{\delta}' E(\chi_{p_2+4}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \right. \\
&\quad \left. -\boldsymbol{\delta}\boldsymbol{\delta}' E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \right] \\
&\quad + \kappa^2 \left[ \boldsymbol{\Sigma}^* E(\chi_{p_2+2}^{-4}(\Delta)) + \boldsymbol{\delta}\boldsymbol{\delta}' E(\chi_{p_2+4}^{-4}(\Delta)) \right]. \\
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}}) &= \boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) - 2\boldsymbol{\Sigma}_{21} E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta)) I(\chi_{p_2+2}^2(\Delta) < \kappa) \\
&\quad - 2\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} E(1 - \kappa\chi_{p_2+4}^{-2}(\Delta)) I(\chi_{p_2+4}^2(\Delta) < \kappa)\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&\quad + 2\boldsymbol{\delta}\boldsymbol{\delta}' E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta)) I(\chi_{p_2+2}^2(\Delta) < \kappa) \\
&\quad + \boldsymbol{\Sigma}^* E(1 - \kappa\chi_{p_2+2}^{-2}(\Delta))^2 I(\chi_{p_2+2}^2(\Delta) < \kappa) \\
&\quad + \boldsymbol{\delta}\boldsymbol{\delta}' E\left\{ (1 - \chi_{p_2+4}^{-2}(\Delta))^2 I(\chi_{p_2+4}^2(\Delta) < \kappa) \right\}.
\end{aligned}
$$

*Proof.*

$$
\begin{aligned}
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}) &= E\left\{ \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \beta_1)\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \beta_1)' \right\} \\
&= E\{\eta_1\eta_1'\} \\
&= \{Cov(\eta_1\eta_1') + E(\eta_1)E(\eta_1)'\} \\
&= Var(\eta_1) \\
&= \gamma^2 \boldsymbol{Q}_{11.2}^{-1}.
\end{aligned}
$$

$$\begin{aligned}
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) &= E\left\{\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}-\beta_1)\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}-\beta_1)'\right\} \\
&= E\{\eta_3\eta_3'\} \\
&= Cov(\eta_3,\eta_3') + E(\eta_3)E(\eta_3)' \\
&= Var(\eta_3) + E(\eta_3)E(\eta_3)' \\
&= \gamma^2\boldsymbol{Q}_{11}^{-1} + \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\boldsymbol{Q}_{21}Q_{11}^{-1}.
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) &= E\left\{\lim_{n\to\infty}\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}-\beta_1)\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}-\beta_1)'\right\} \\
&= E\left\{\lim_{n\to\infty} n\left[(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}-\beta_1)-(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}-\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})\kappa\psi_n^{-1}\right]\right. \\
&\quad \left.\left[(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}-\beta_1)-(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}-\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})\kappa\psi_n^{-1}\right]'\right\} \\
&= E\left\{[\eta_1-\eta_2\kappa\psi_n^{-1}][\eta_1-\eta_2\kappa\psi_n^{-1}]'\right\} \\
&= E\left\{\eta_1\eta_1' - 2\kappa\psi_n^{-1}\eta_2\eta_1' + \kappa^2\psi_n^{-2}\eta_2\eta_2'\right\}. \qquad \text{(A)}
\end{aligned}$$

Now

$$\begin{aligned}
E\left\{\psi_n^{-1}\eta_2\eta_1'\right\} &= E\left\{E(\eta_2\eta_1'\psi_n^{-1}|\eta_2)\right\} \\
&= E\left\{\eta_2 E(\eta_1'\psi_n^{-1}|\eta_2)\right\} \\
&= E\left\{\eta_2\left[0+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{*-1}(\eta_2-\delta)\right]'\psi_n^{-1}\right\} \\
&= E\left\{\eta_2(\eta_2-\boldsymbol{\delta})'\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{12}'\psi_n^{-1}\right\} \\
&= E\left\{\eta_2\eta_2'\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\psi_n^{-1}\right\} - E\left\{\eta_2\boldsymbol{\delta}'\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\psi_n^{-1}\right\} \\
&= \left[Var(\eta_2)E(\chi_{p_2+2}^{-2}(\Delta)) + E(\eta_2)E(\eta_2)'E(\chi_{p_2+4}^{-2}(\Delta))\right]\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&\quad - E(\eta_2)\boldsymbol{\delta}'E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&= \left[\boldsymbol{\Sigma}^* E(\chi_{p_2+2}^{-2}(\Delta)) + \boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+4}^{-2}(\Delta))\right]\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&\quad - \boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&= E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}_{21} + \boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+4}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21} \\
&\quad - \boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}.
\end{aligned}$$

Now, substituting $E\{\psi_n^{-1}\eta_2\eta_2'\}$ in (A), we get

$$\begin{aligned}
\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) &= E\{\eta_1\eta_1'\} - 2\kappa E\left\{\psi_n^{-1}\eta_2\eta_1'\right\} + \kappa E\left\{\psi_n^{-2}\eta_2\eta_2'\right\} \\
&= Var(\eta_1) - 2\kappa\left[E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}_{21} + \boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+4}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right. \\
&\quad \left.-\boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right] \\
&\quad + \kappa^2\left\{Var(\eta_2)E(\chi_{p_2+2}^{-4}(\Delta)) + E(\eta_2)E(\eta_2)')E(\chi_{p_2+4}^{-4}(\Delta))\right\} \\
&= \gamma^2\boldsymbol{Q}_{11.2}^{-1} - 2\kappa\left[E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}_{21} + \boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+4}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right. \\
&\quad \left.-\boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+2}^{-2}(\Delta))\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\right] \\
&\quad + \kappa^2\left[\boldsymbol{\Sigma}^* E(\chi_{p_2+2}^{-4}(\Delta)) + \boldsymbol{\delta}\boldsymbol{\delta}'E(\chi_{p_2+4}^{-4}(\Delta))\right].
\end{aligned}$$

$$
\begin{aligned}
\mathbf{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}}) &= E\left\{\lim_{n\to\infty} n(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}} - \beta_1)(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}} - \beta_1)'\right\} \\
&= \Gamma(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) - 2E\left\{\lim_{n\to\infty} n(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})(\hat{\boldsymbol{\beta}}^{\mathrm{UM}} - \beta_1)'(1-\kappa\psi_n^{-1})I(\psi_n < \kappa)\right\} \\
&\quad + E\left\{\lim_{n\to\infty} n(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})'(1-\kappa\psi_n^{-1})^2 I(\psi_n < \kappa)\right\} \\
&= \Gamma(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) - 2E\left\{\eta_2\eta_1'(1-\kappa\psi_n^{-1})I(\psi_n < \kappa)\right\} \\
&\quad + E\left\{\eta_2\eta_2'(1-\kappa\psi_n^{-1})^2 I(\psi_n < \kappa)\right\}. \qquad \text{(B)}
\end{aligned}
$$

Now, using the rule of conditional expectation,

$$
\begin{aligned}
&E\left\{\eta_2\eta_2'(1-\kappa\psi_n^{-1})I(\psi_n < \kappa)\right\} \\
&= E\left[\eta_2 E\left\{\eta_1'(1-\kappa\psi_n^{-1}I(\psi_n < \kappa)\right\}|\eta_2\right] \\
&= E\left[\eta_2\left\{0 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}^{*-1}(\eta_2 - \boldsymbol{\delta})\right\}'(1-\kappa\psi_n^{-1})I(\psi_n < \kappa)\right] \\
&= E\left\{\eta_2(\eta_2 - \boldsymbol{\delta})'\mathbf{\Sigma}^{*-1}\mathbf{\Sigma}_{21}(1-\kappa\psi_n^{-1})I(\psi_n < \kappa)\right\} \\
&= E\left\{\eta_2\eta_2'\mathbf{\Sigma}^{*-1}\mathbf{\Sigma}_{21}(1-\kappa\psi_n^{-1})I(\psi_n < \kappa)\right\} \\
&\quad - E\left\{\eta_2\boldsymbol{\delta}'\mathbf{\Sigma}^{*-1}\mathbf{\Sigma}_{21}(1-\kappa\psi_n^{-1})I(\psi_n < \kappa)\right\} \\
&= \left\{Var(\eta_2)E(1-\kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)\mathbf{\Sigma}^{*-1}\mathbf{\Sigma}_{21}\right. \\
&\quad \left. +\boldsymbol{\delta}\boldsymbol{\delta}'E(1-\kappa\chi_{p_2+4}^{-2}(\Delta))I(\chi_{p_2+4}^2(\Delta) < \kappa)\mathbf{\Sigma}^{*-1}\mathbf{\Sigma}_{21}\right\} \\
&\quad - \left\{\boldsymbol{\delta}\boldsymbol{\delta}'E(1-\kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^2(\Delta) < \kappa)\right\}.
\end{aligned}
$$

Now, substituting the above in (B), we get

$$\begin{aligned}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}+}) = \boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) &- 2\boldsymbol{\Sigma}_{21}E(1-\kappa\chi^{-2}_{p_2+2}(\Delta))I(\chi^2_{p_2+2}(\Delta)<\kappa)\\ &- 2\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}E(1-\kappa\chi^{-2}_{p_2+4}(\Delta))I(\chi^2_{p_2+4}(\Delta)<\kappa)\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\\ &+ 2\boldsymbol{\delta}\boldsymbol{\delta}'E(1-\kappa\chi^{-2}_{p_2+2}(\Delta))I(\chi^2_{p_2+2}(\Delta)<\kappa)\\ &+ \boldsymbol{\Sigma}^*E(1-\kappa\chi^{-2}_{p_2+2}(\Delta))^2I(\chi^2_{p_2+2}(\Delta)<\kappa)\\ &+ \boldsymbol{\delta}\boldsymbol{\delta}'E\left\{(1-\kappa\chi^{-2}_{p_2+4}(\Delta))^2I(\chi^2_{p_2+4}(\Delta)<\kappa)\right\}.\end{aligned}$$

### 4.5.2 Risk Performance

Using the definition (4.20), ADQR expressions are given below.

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}) &= \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}})) \\
&= \mathrm{tr}(\boldsymbol{W}\gamma^2\boldsymbol{Q}_{11.2}^{-1}) \\
R(\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}) &= \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}})) \\
&= \mathrm{tr}(\boldsymbol{W}\gamma^2\boldsymbol{Q}_{11}^{-1}) + \mathrm{tr}(\boldsymbol{W}\boldsymbol{M}), \quad \text{where } \boldsymbol{M} = \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\boldsymbol{\omega}\boldsymbol{\omega}'\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} \\
R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) &= \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}})) \\
&= R(\hat{\boldsymbol{\beta}}_1^{\mathrm{UM}}) - 2\kappa E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\}\mathrm{tr}(\boldsymbol{W}\boldsymbol{\Sigma}_{21}) \\
&\quad - 2\kappa E\left\{\chi_{p_2+4}^{-2}(\Delta)\right\}\mathrm{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}) \\
&\quad + 2\kappa E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\}\mathrm{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}) \\
&\quad + \kappa^2 E\left\{\chi_{p_2+2}^{-4}(\Delta)\right\}\mathrm{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{*}) \\
&\quad + \kappa^2 E\left\{\chi_{p_2+4}^{-2}(\Delta)\right\}\mathrm{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}') \\
R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}}) &= \mathrm{tr}(\boldsymbol{W}\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}})) \\
&= R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) - 2E(1-\kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+2}^{2}(\Delta)<\kappa)\mathrm{tr}(\boldsymbol{W}\boldsymbol{\Sigma}_{21}) \\
&\quad - 2E(1-\kappa\chi_{p_2+4}^{-2}(\Delta))I(\chi_{p_2+4}^{2}(\Delta)<\kappa)\mathrm{tr}(\boldsymbol{W}\boldsymbol{\delta}'\boldsymbol{\delta}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Sigma}_{21}) \\
&\quad + 2E(1-\kappa\chi_{p_2+2}^{-2}(\Delta))I(\chi_{p_2+4}^{2}(\Delta)<\kappa)\mathrm{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}') \\
&\quad + E\left\{(1-\kappa\chi_{p_2+2}^{-2}(\Delta))^2 I(\chi_{p_2+2}^{2}(\Delta)<\kappa)\right\}\mathrm{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{*}) \\
&\quad + E\left\{(1-\kappa\chi_{p_2+4}^{-2}(\Delta))^2 I(\chi_{p_2+4}^{2}(\Delta)<\kappa)\right\}\mathrm{tr}(\boldsymbol{W}\boldsymbol{\delta}\boldsymbol{\delta}')
\end{aligned}
$$

## 4.6 Simulation Studies

We perform Monte Carlo simulation experiments to examine the quadratic risk performance of the proposed estimators. We simulate the response from the following

model:

$$y_i = \sum_{l=1}^{p_1} x_{il}\beta_l + \sum_{m=p_1+1}^{p} x_{im}\beta_m + \sin(4\pi t_i) + \varepsilon_i \tag{4.23}$$

where $\beta_l$ is a $p_1 \times 1$ vector and $\beta_m$ is $p_2 \times 1$ vector of parameters and $p = p_1 + p_2$.

To simulate the data, we consider, $x_{i1} = (\zeta_{i1}^{(1)})^2 + \zeta_i^{(1)} + \xi_{i1}$, $x_{i2} = (\zeta_{i2}^{(1)})^2 + \zeta_i^{(1)} + 2\xi_{i2}$, $x_{is} = (\zeta_{is}^{(1)})^2 + \zeta_i^{(1)}$ with $\zeta_{is}^{(1)}$ i.i.d. $\sim N(0,1)$, $\zeta_i^{(1)}$ i.i.d. $\sim N(0,1)$, $\xi_{i1}$ ∼Bernoulli(0.35) and $\xi_{i2}$ ∼Bernoulli(0.35) for all $s = 3, \ldots, p$, $p = p_1 + p_2$, and $i = 1, \ldots, n$. Four different error distributions have been considered which are defined later in this chapter.

We are interested in testing the hypothesis $H_0 : (\beta_{p_1+1}, \beta_{p_1+2}, \ldots, \beta_{p_1+p_2}) = \mathbf{0}$. Our aim is to estimate $\boldsymbol{\beta}_1$, when the remaining regression parameters may not be useful. We partition the regression coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$.

The number of simulations was initially varied. Finally, each realization was repeated 5000 times to obtain stable results. For each realization, we calculated bias of the estimators. We defined $\Delta^* = ||\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}||$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1, \mathbf{0})$ and $||\cdot||$ is the Euclidean norm. $\Delta^*$ and $S_n$ were estimated by median absolute deviation (MAD).

To determine the behavior of the estimators for $\Delta^* > 0$, further data sets were generated from those distributions under local alternative hypothesis.

### 4.6.1 Error Distributions

Four different error distributions have been considered. They are outlined briefly below.

**Normal and Contaminated Normal**

$$F(x) = \lambda N(0, \sigma^2) + (1 - \lambda)N(0, 1) \tag{4.24}$$

where $\lambda$ is the parameters indicating whether standard normal or its contaminated version is returned. We consider $\lambda = 0$ and .9. For $\lambda = 0$ we get standard normal errors, while scale contaminated normal errors are obtained for $\lambda = .9$.

**Standard Logistic**

The standard logistic distribution has cdf

$$F(x) = \frac{1}{1 + e^{-x}}, \quad x \in \Re \tag{4.25}$$

**Standard Laplace**

The standard Laplace distribution has cdf

$$F(x) = \frac{1}{2}\left[1 + \text{sign}(x)(1 - e^{-|x|})\right], \quad x \in \Re. \tag{4.26}$$

### 4.6.2 Risk Comparison

The risk performance of an estimator of $\boldsymbol{\beta}_1$ was measured by calculating its MSE. After calculating the MSEs, we numerically calculated the efficiency of the proposed estimators $\hat{\boldsymbol{\beta}}_1^{\text{RM}}$, $\hat{\boldsymbol{\beta}}_1^{\text{SM}}$, $\hat{\boldsymbol{\beta}}_1^{\text{SM+}}$ relative to the unrestricted estimator $\hat{\boldsymbol{\beta}}_1^{\text{UM}}$ using the

relative mean squared error criterion, given as follows:

$$\text{RMSE}(\hat{\boldsymbol{\beta}}_1^{\text{UM}} : \hat{\boldsymbol{\beta}}_1^{*}) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{\text{UM}})}{\text{MSE}(\hat{\boldsymbol{\beta}}_1{}^{*})}, \tag{4.27}$$

where $\hat{\boldsymbol{\beta}}_1^{*}$ is one of the proposed estimators. The amount by which an RMSE is larger than unity indicates the degree of superiority of the estimator $\hat{\boldsymbol{\beta}}_1^{*}$ over $\hat{\boldsymbol{\beta}}_1^{\text{UM}}$.

To compute RMSE we consider $n = 30, 50$ and $(p_1, p_2)$= (3, 5), (3, 9), (5, 9) and (5, 20) based on Huber's $\rho-$function. Results are shown in Tables 4.1–4.4. Since the results of our simulation study are similar for all the combinations, we conduct separate simulations to visually compare the estimators for $n = 50$, and $(p_1, p_2) = (3, 4)$.

Figure 4.1 shows the RMSEs of various M-estimators for Huber's $\rho$-function. Here, $\Delta^*$ indicates the correctness of the sub-model under null hypothesis. $\Delta^* > 0$ indicates the degree of deviation from the hypothesized model. We found that the RM estimators are the best when $\Delta^* = 0$. However, the RM estimators become inefficient and the RMSE goes below 1 very quickly as $\Delta^*$ deviates from zero. The RMSE of restricted estimator is depicted by the dashed line in Figure 4.1. In the simulation study, the RM shows similar behaviour for all the error distributions considered in this study.

Positive shrinkage estimator (SM+) appears to be most stable in terms of RMSE as $\Delta^*$ becomes large. Although the RM estimator outperforms all other estimators for $\Delta^* = 0$, SM+ dominates in terms of RMSE for $\Delta^*$ as small as 0.10 for all the error distributions except standard Laplace. When error distribution is standard Laplace, SM+ dominates RM for $\Delta^* \geq 20$.

An RMSE larger than 1 indicates that the risk of the corresponding estimator is smaller than the risk of unrestricted M-estimator. For example, an RMSE of "$x$" indicates that gain in risk of the estimator is "$x$"-times that of UM. As for example, Table 4.1 presents the RMSEs based on Huber's $\rho$-function for sample size 30, and $(p_1, p_2) = (3, 5)$. For standard normal error, gain in risk for positive-shrinkage M-estimator is 3.161 times that of the ordinary M-estimator provided that the model specification is correct (i.e., $\Delta^* = 0$). For the same configuration, when the error distribution is standard Laplace, the gain is risk for SM+ is 2.273 times that of UM.

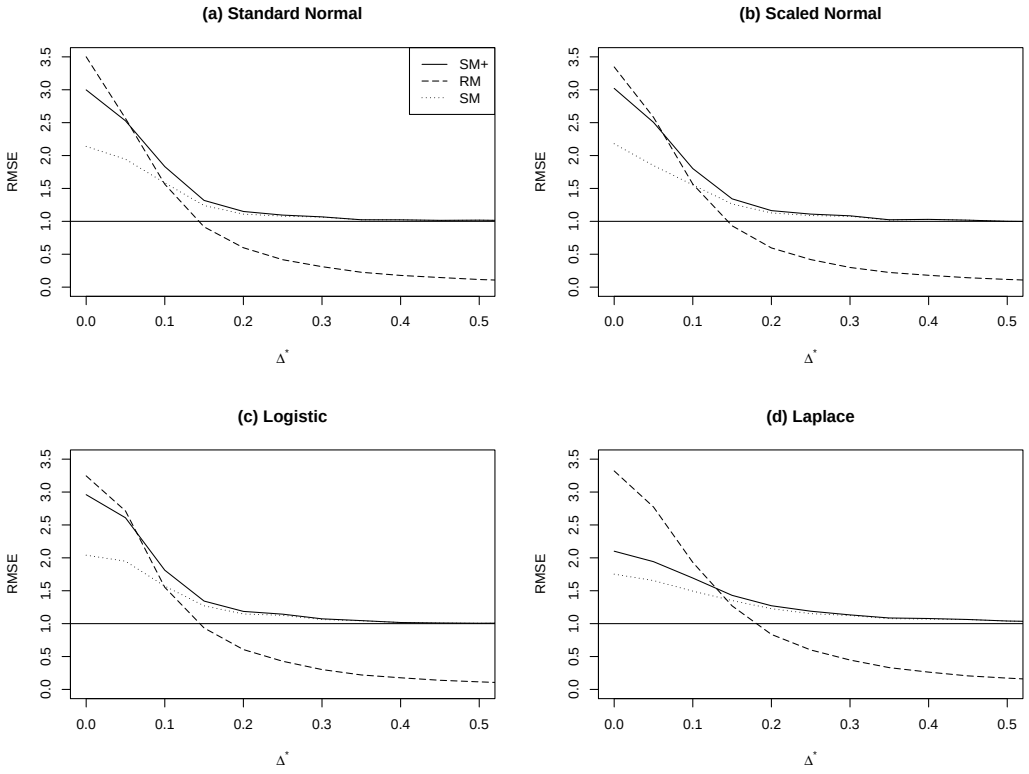


Figure 4.1: Relative mean squared errors for RM, SM, and SM+ estimators with respect to unrestricted M-estimator for $n = 50$, $(p_1, p_2) = (3, 4)$ when Huber's $\rho$-function is considered.

Table 4.1: Relative mean squared errors for restricted, shrinkage, and positive shrinkage M-estimators for $(p_1, p_2) = (3,\ 5)$, $n = 30$, based on Huber's $\rho-$function for different error distributions.

| Error | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}}$ |
|---|---|---|---|---|
| Standard | 0.00 | 3.695 | 2.035 | 3.161 |
| Normal | 0.05 | 3.472 | 2.084 | 3.224 |
| | 0.10 | 2.386 | 1.758 | 2.515 |
| | 0.15 | 1.619 | 1.496 | 1.835 |
| | 0.20 | 1.065 | 1.216 | 1.479 |
| | 0.25 | 0.814 | 1.202 | 1.261 |
| | 1.00 | 0.060 | 0.995 | 0.997 |
| Scaled | 0.00 | 3.867 | 2.163 | 3.195 |
| Normal | 0.05 | 2.842 | 1.733 | 2.621 |
| | 0.10 | 2.250 | 1.707 | 2.352 |
| | 0.15 | 1.476 | 1.492 | 1.839 |
| | 0.20 | 1.100 | 1.288 | 1.473 |
| | 0.25 | 0.737 | 1.082 | 1.163 |
| | 1.00 | 0.060 | 1.017 | 1.017 |
| Standard | 0.00 | 3.532 | 1.921 | 2.991 |
| Logistic | 0.05 | 3.288 | 1.922 | 3.004 |
| | 0.10 | 2.400 | 1.846 | 2.434 |
| | 0.15 | 1.656 | 1.551 | 1.853 |
| | 0.20 | 1.129 | 1.323 | 1.464 |
| | 0.25 | 0.758 | 1.158 | 1.208 |
| | 1.00 | 0.062 | 0.996 | 0.996 |
| Standard | 0.00 | 3.853 | 1.895 | 2.273 |
| Laplace | 0.05 | 3.628 | 1.743 | 2.056 |
| | 0.10 | 2.719 | 1.597 | 1.974 |
| | 0.15 | 2.179 | 1.426 | 1.753 |
| | 0.20 | 1.418 | 1.329 | 1.538 |
| | 0.25 | 1.090 | 1.273 | 1.360 |
| | 1.00 | 0.093 | 1.014 | 1.016 |

Table 4.2: Relative mean squared errors for restricted, shrinkage, and positive shrinkage M-estimators for $(p_1, p_2) = (3,\ 9)$, $n = 50$, based on Huber's $\rho-$function for different error distributions.

| Error | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}}$ |
|---|---|---|---|---|
| Standard | 0.00 | 5.552 | 3.607 | 5.462 |
| Normal | 0.05 | 4.269 | 3.098 | 4.407 |
| | 0.10 | 2.574 | 2.322 | 2.919 |
| | 0.15 | 1.601 | 1.827 | 2.027 |
| | 0.20 | 1.085 | 1.503 | 1.583 |
| | 0.25 | 0.726 | 1.340 | 1.370 |
| | 1.00 | 0.051 | 0.994 | 0.994 |
| Scaled | 0.00 | 5.443 | 3.576 | 5.301 |
| Normal | 0.05 | 4.457 | 3.009 | 4.510 |
| | 0.10 | 2.755 | 2.442 | 3.096 |
| | 0.15 | 1.668 | 1.788 | 2.028 |
| | 0.20 | 1.022 | 1.417 | 1.500 |
| | 0.25 | 0.750 | 1.313 | 1.353 |
| | 1.00 | 0.051 | 1.001 | 1.001 |
| Standard | 0.00 | 5.702 | 3.364 | 5.410 |
| Logistic | 0.05 | 4.422 | 3.316 | 4.459 |
| | 0.10 | 2.641 | 2.364 | 2.927 |
| | 0.15 | 1.666 | 1.941 | 2.139 |
| | 0.20 | 1.040 | 1.475 | 1.582 |
| | 0.25 | 0.710 | 1.338 | 1.374 |
| | 1.00 | 0.049 | 1.009 | 1.009 |
| Standard | 0.00 | 5.827 | 2.734 | 3.256 |
| Laplace | 0.05 | 5.187 | 2.489 | 2.989 |
| | 0.10 | 3.351 | 2.333 | 2.624 |
| | 0.15 | 2.332 | 2.142 | 2.242 |
| | 0.20 | 1.462 | 1.692 | 1.783 |
| | 0.25 | 1.039 | 1.565 | 1.606 |
| | 1.00 | 0.075 | 1.021 | 1.021 |

Table 4.3: Relative mean squared errors for restricted, shrinkage, and positive shrinkage M-estimators for $(p_1, p_2) = (5,\ 9)$, $n = 50$, based on Huber's $\rho-$function for different error distributions.

| Error | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}}$ |
|---|---|---|---|---|
| Standard | 0.00 | 3.838 | 2.772 | 3.705 |
| Normal | 0.05 | 3.202 | 2.438 | 3.179 |
| | 0.10 | 2.494 | 2.168 | 2.641 |
| | 0.15 | 1.638 | 1.708 | 1.923 |
| | 0.20 | 1.097 | 1.453 | 1.531 |
| | 0.25 | 0.778 | 1.201 | 1.260 |
| | 1.00 | 0.062 | 1.016 | 1.016 |
| Scaled | 0.00 | 3.653 | 2.592 | 3.472 |
| Normal | 0.05 | 3.172 | 2.328 | 3.161 |
| | 0.10 | 2.391 | 2.035 | 2.530 |
| | 0.15 | 1.632 | 1.748 | 1.986 |
| | 0.20 | 1.123 | 1.453 | 1.550 |
| | 0.25 | 0.797 | 1.244 | 1.297 |
| | 1.00 | 0.064 | 1.028 | 1.028 |
| Standard | 0.00 | 3.740 | 2.730 | 3.678 |
| Logistic | 0.05 | 3.299 | 2.554 | 3.298 |
| | 0.10 | 2.424 | 2.125 | 2.544 |
| | 0.15 | 1.596 | 1.690 | 1.931 |
| | 0.20 | 1.077 | 1.383 | 1.480 |
| | 0.25 | 0.769 | 1.283 | 1.326 |
| | 1.00 | 0.059 | 1.002 | 1.002 |
| Standard | 0.00 | 3.919 | 2.125 | 2.626 |
| Laplace | 0.05 | 3.872 | 2.158 | 2.461 |
| | 0.10 | 2.613 | 1.923 | 2.207 |
| | 0.15 | 2.015 | 1.798 | 1.895 |
| | 0.20 | 1.519 | 1.637 | 1.703 |
| | 0.25 | 1.157 | 1.464 | 1.528 |
| | 1.00 | 0.099 | 1.022 | 1.022 |

Table 4.4: Relative mean squared errors for restricted, shrinkage, and positive shrinkage M-estimators for $(p_1, p_2) = (5,\ 20)$, $n = 50$, based on Huber's $\rho-$function for different error distributions.

| Error | $\Delta^*$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{RM}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$ | $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM+}}$ |
|---|---|---|---|---|
| Standard | 0.00 | 7.469 | 5.415 | 7.328 |
| Normal | 0.05 | 6.034 | 4.502 | 6.145 |
| | 0.10 | 3.809 | 3.343 | 3.992 |
| | 0.15 | 2.230 | 2.487 | 2.727 |
| | 0.20 | 1.437 | 1.901 | 1.985 |
| | 0.25 | 1.019 | 1.638 | 1.672 |
| | 1.00 | 0.072 | 1.037 | 1.037 |
| Scaled | 0.00 | 7.900 | 5.809 | 7.974 |
| Normal | 0.05 | 6.115 | 4.864 | 6.171 |
| | 0.10 | 3.593 | 3.111 | 3.820 |
| | 0.15 | 2.295 | 2.446 | 2.726 |
| | 0.20 | 1.491 | 1.967 | 2.054 |
| | 0.25 | 1.005 | 1.613 | 1.636 |
| | 1.00 | 0.073 | 1.028 | 1.028 |
| Standard | 0.00 | 7.366 | 5.569 | 7.238 |
| Logistic | 0.05 | 6.085 | 4.701 | 6.042 |
| | 0.10 | 3.767 | 3.391 | 4.100 |
| | 0.15 | 2.296 | 2.494 | 2.714 |
| | 0.20 | 1.482 | 1.920 | 2.004 |
| | 0.25 | 1.018 | 1.589 | 1.624 |
| | 1.00 | 0.072 | 1.025 | 1.025 |
| Standard | 0.00 | 8.550 | 3.841 | 4.200 |
| Laplace | 0.05 | 7.369 | 3.424 | 3.817 |
| | 0.10 | 4.973 | 3.262 | 3.457 |
| | 0.15 | 3.116 | 2.510 | 2.622 |
| | 0.20 | 2.149 | 2.266 | 2.342 |
| | 0.25 | 1.473 | 1.938 | 1.945 |
| | 1.00 | 0.112 | 1.054 | 1.054 |

## 4.7 Conclusion

In this chapter, we considered shrinkage M-estimation in the context of a partially linear regression model. We developed shrinkage and positive-shrinkage M-estimators when we have prior information about a subset of the covariates. Based on a quadratic risk function, we computed relative risk of shrinkage M-estimators with respect to the unrestricted M-estimator. Asymptotic properties of the estimators have been studied and their risk expressions derived.

In the simulation study, we numerically computed relative mean squared errors of the restricted-M, shrinkage-M, and positive-shrinkage M-estimators compared to the unrestricted M-estimator. Four different error distributions have been considered to study the performance of the proposed estimators. A Monte Carlo simulation study provided support for the positive-shrinkage estimators under varying degrees of model misspecification.

Restricted M-estimator (RM) outperforms all other estimators when the nuisance subspace is zero. However, a small departure from this condition makes the RM very inefficient, questioning its applicability for practical purposes.

# Chapter 5

# Conclusions and Future Work

In this dissertation, shrinkage and absolute penalty estimation have been studied in linear and partially linear models. Application of shrinkage and pretest estimators have been demonstrated in fully parametric and semiparametric regression models with real data examples. We compared the performance of shrinkage and absolute penalty estimators and demonstrated the usefulness of shrinkage estimators under some conditions.

We have discussed the following topics in this dissertation

(i) Application of shrinkage and pretest estimation in linear models

(ii) Comparison of positive-shrinkage and absolute penalty estimators (lasso, adaptive lasso, and SCAD) in linear models

(iii) Use of B-spline basis expansion in partially linear models to obtain shrinkage estimators, and their comparison with lasso and adaptive lasso estimators

(iv) Robust shrinkage M-estimation in partially linear models

In the following, we summarize our findings.

In the first part of Chapter 2, we presented shrinkage, positive-shrinkage and pretest estimation in the context of a multiple linear regression model, and demonstrated their applicability using three real data examples. To illustrate the methods, average prediction errors based on repeated cross validation estimate of the error rates were computed. We numerically showed that pretest and restricted estimators have superior risk performance compared to unrestricted and positive-shrinkage estimators when the underlying model is correctly specified. However, under model misspecification, positive-shrinkage estimators showed superior performance in terms of minimizing prediction errors.

In the second part of Chapter 2, we developed and implemented the algorithm for simultaneous subset selection using AIC and BIC to obtain shrinkage estimates of the regression coefficients in a multiple regression context. Several absolute penalty estimators such as lasso, adaptive lasso, and SCAD were studied, and their relative risks compared with those obtained using shrinkage estimators. Through a real data example, we illustrated that the positive-shrinkage estimator outperforms absolute penalty estimators for varying degrees of model misspecification. In general, the positive shrinkage estimator maintains its superiority over all other estimators for moderate sample sizes and when there is a large number of nuisance covariates present in the model.

We compared the performance of shrinkage and APE for both low- and high-dimentional scenarios. In low-dimensional data, positive-shrinkage estimators outperform all other estimators in a given neighbourhood of the pivot ($\Delta^* = 0$). For high-dimensional data with large $p$ and $p < n$, PSE and SCAD estimators perform

equally in terms of RMSE when the number of nuisance parameters is up to 40. As the number goes higher, adaptive lasso and SCAD became dominant over the shrinkage estimators.

In Chapter 3, we introduced shrinkage estimation of parameters of a semiparametric regression model. This work is an extension of Ahmed et al. (2007) in which a kernel-based estimator was considered to estimate the nonlinear component of a PLM. We explored the suitability of using B-spline basis expansion to estimate the nonparametric part. Since B-splines are easier to incorporate in a regression model, one may prefer B-spline over a kernel-based method for the nonparametric component.

Through a simulation study, both flat and a highly oscillating non-flat function for the nonparametric component were studied. We found that the B-spline-based approach shows less bias in the estimates than the estimates obtained by the kernel-based approach. Therefore, in many practical situations, B-spline will be a good alternative to kernel-based estimators for estimating the nonparametric component in a PLM especially when uniform knots are suitable. Further, shrinkage estimators based on B-splines outperformed the absolute penalty estimator (lasso) for moderate sample sizes and when the nuisance parameter space was large.

In Chapter 4 we considered shrinkage M-estimation in the context of a partially linear regression model. We developed shrinkage and positive-shrinkage estimators when we have prior information about a subset of the covariates. Based on a quadratic risk function, we computed relative risk of shrinkage M-estimator with respect to the unrestricted M-estimator. Through a simulation study, we numerically computed relative mean squared errors of the RM, SM, and SM+ compared to the unrestricted M-estimator. Four different error distributions were considered to study the perfor-

mance of the proposed estimators. A Monte Carlo simulation study provided support for the SM+ estimators under varying sizes of the nuisance parameter space.

## Future Work

There are possibilities of extending our works in the following ways.

In Chapter 2, we compared shrinkage estimation strategies with absolute penalty estimators (APE), such as, lasso, adaptive lasso, and smoothly clipped absolute deviation. We found that the shrinkage strategies work better than the APEs in terms of quadratic risk, when the number of restriction on the parameter space is large. We suspected that the comparative performance of shrinkage and APEs may also depend on the number of main parameters. In this respect, a further study may be conducted to explore the effect of the size of $p_1$ for varying $p_2$ on the performance of shrinkage and APE strategies. It will be worth exploring if there exist a ratio of $p_1$ to $p_2$ that guarantees dominance of shrinkage estimators over the APEs.

We observed that adaptive lasso, although computationally very expensive, performs better than the lasso. In adaptive lasso, lasso estimates are used as weights to obtain the final estimates. We are currently considering the use of positive-shrinkage estimates as weights in computing the adaptive lasso estimates. The results will be released through future publications.

As a continuation of the work in Chapter 3, we will introduce penalized spline or P-spline into the picture and compare the performance of kernel, B-spline, and P-spline based shrinkage estimators.

In Chapter 4, we developed robust M-estimators based on the shrinkage principle

in partially linear models. In our study, the nonlinear component was estimated using kernel-smoothing, and the robust M-estimates were obtained for the regression parameters on the linearized model. Considering a robust estimation of the nonparametric part is our immediate research goal. To achieve this, appropriate algorithm and software are to be developed.

At present, we are developing an R-package for shrinkage estimation in the multiple regression model. The project is hosted at `http://shrink.r-forge.r-project.org/`. The final version will be released in the future through Comprehensive R Archive Network (CRAN).

# Bibliography

Ahmed, S. E. (1997). Asymptotic shrinkage estimation: the regression case. *Applied Statistical Science II*, pages 113–139.

Ahmed, S. E. (2001). Shrinkage estimation of regression coefficients from censored data with multiple observations. In Ahmed, S. E. and Reid, N., editors, *Empirical Bayes and Likelihood Inference, Lecture Notes in Statistics*, volume 148, pages 103–120. Springer-Verlag, New York.

Ahmed, S. E. (2012). *Absolute Penalty, Shrinkage and Shrinkage Pretest Strategies: Estimation and Model Selection in High and Low Dimensional Data*. Springer.

Ahmed, S. E., Doksum, K. A., Hossain, S., and You, J. (2007). Shrinkage, pretest and absolute penalty estimators in partially linear models. *Australian & New Zealand Journal of Statistics*, 49:435–454.

Ahmed, S. E., Hussein, A. A., and Sen, P. K. (2006). Risk comparison of some shrinkage M-estimators in linear models. *Nonparametric Statistics*, 18:401–415.

Ahmed, S. E., Raheem, E., and Hossain, M. S. (2010). *International Encyclopedia of Statistical Science*, chapter Absolute Penalty Estimation. Springer.

Ahmed, S. E. and Saleh, A. E. (1999). Improved nonparametric estimation of location vectors in multivariate regression models. *Journal of Nonparametric Statistics*, 11.

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significances. *Annals of Mathematical Statistics*, 15:190–204.

Bhattacharya, P. and Zhao, P.-L. (1997). Semiparametric inference in a partial linear model. *The Annals of Statistics*, 25(1):244–262.

Bianco, A. and Boente, G. (2004). Robust estimators in semiparametric partly linear regression models. *Journal of Statistical Planning and Inference*, 122:229–252.

Bianco, A. and Boente, G. (2007). Robust estimators under semi-parametric partly linear autoregression: Asymptotic behaviour and bandwidth selection. *Journal of Time Series Analysis*, 28(2):274–306.

Bock, M. E., Judge, G., and Yancey, T. (Marc 1983). A simple form for inverse moments of non-central chi-square random variables and the risk of james-stein estimators. Technical report, Department of Statistics, Purdue University.

Boente, G., He, X., and Zhou, J. (2006). Robust estimates in generalized partially linear models. *The Annals of Statistics*, 34(6):2856–2878.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253.

Breiman, L. (1993). Better subset regression using the nonnegative garrote. Technical report, University of California, Berkeley.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics*, 32:898–927.

Burman, P. (1991). Regression function estimation from dependent observations. *Journal of Multivariate Analysis*, 36:263–279.

Castner, L. A. and Schirm, A. L. (2003). Empirical bayes shrinkage estimates of state food stamp participation rates for 1998-2000. Technical report, Mathematica Policy Research, Inc.

Chen, H. (1988). Convergence rates for parametric components in a partially linear model. *Annals of Statistics*, 16:136–147.

Chen, H. and Shiau, J. (1991). A two-stage spline smoothing method for partially linear models. *Journal of Statistics Planning and Inference*, 27:187–202.

Chen, H. and Shiau, J. (1994). Data-driven efficient estimation for a partially linear model. *Annals of Statistics*, 22:211–237.

Cheng, G. and Huang, J. Z. (2010). Bootstrap consistency for general semiparametric M-estimation. *Annals of Statistics*, 38(5):2884–2915.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York.

Denby, L. (1986). Smooth regression functions. Technical report, Statistical Research Report 26, AT & T Bell Laboratories.

Donald, G. and Newey, K. (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis*, 50:30–40.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

Engle, R. F., Granger, W. J., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 80:310–319.

Eubank, R. and Speckman, P. (1990). Trigonometric series regression estimators with an application to partially linear models. *Journal of Multivariate Analysis*, 32:70–85.

Fan, J. (1997). Comments on wavelets in statistics: A review by A. Antoniadis. *Journal of the Italian Statistical Association*, 6:131–138.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541.

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

Faraway, J. J. (2002). *Practical Regression and Anova using R*.

Fox, J. (2002). *An R and S-PLUS Companion to Applied Regression*. Sage, Thousand Oaks.

Fox, J. (2005). Introduction to nonparametric regression. Website. `http://socserv.socsci.mcmaster.ca/jfox/Courses/Oxford-2005/index.html`.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Gao, J. T. (1995). Asymptotic theory for partially linear models. *Communications in Statistics–Theory & Methods*, A 24(8):1985–2009.

Hamilton, A. and Truong, K. (1997). Local linear estimation in partially linear models. *Journal of Multivariate Analysis*, 60:1–19.

Härdle, W., Liang, H., and Gao, J. (2000). *Partially linear models*. Physica-Verlag, Heidelber.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

He, X., Fung, W. K., and Zhu, Z. (2005). Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association*, 100(472):1176–1184.

He, X. and Shi, P. D. (1996). Bivariate tensor-product b-spline in a partly linear model. *Journal of Multivariate Analysis*, 58:162–181.

He, X., Zhu, Z., and Fung, W. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89(3):579–590.

Heckman, N. (1986). Spline smoothing in a partially linear model. *Journal of the Royal Statistical Society, Series B*, 48:244–248.

Hesterberg, T., Choi, N., Meier, L., and Fraley, C. (2008). Least angle and $l_1$ penalized regression: A review. *Statistics Surveys*, 2:61–93.

Johnson, R. A. and Wichern, D. W. (2001). *Applied Multivariate Statistical Analysis, 3rd Ed.* Prentice-Hall.

Judge, G. G. and Bock, M. E. (1978). *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics.* North Holland, Amsterdam.

Jurečková, J. and Sen, P. K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations.* Wiley.

Khan, B. U. and Ahmed, S. E. (2003). Improved estimation of coefficient vector in a regression model. *Communications in Statistics - Simulation and Computation*, 32(3):747–769.

Kraemer, N. and Schaefer, J. (2010). *parcor: Regularized estimation of partial correlation matrices.* R package version 0.2-2.

Liang, H. (2006). Estimation in partially linear models and numerical comparisons. *Computational Statistics and Data Analysis*, 50:675–687.

Liang, H., Wong, S., Robins, J., and Carroll, R. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association*, 99:357–367.

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96(1):190–217.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B*, 72:417–473.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55(4):765–799.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raheem, S. E., Ahmed, S. E., and Doksum, K. A. (2012). Absolute penalty and shrinkage estimation in partially linear models. *Computational Statistics & Data Analysis*, 56(4):874–891.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics & Probability Letters*, 4:203–208.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56:931–954.

Sacks, J. and Ylvisaker, D. (1970). Design for regression problems with correlated errors III. *Annals of Mathematical Statistics*, 41:2057–2074.

Saleh, A. K. M. E. (2006). *Theory of Preliminary Test and Stein-Type Estimation with Applications*. Wiley.

Saleh, A. K. M. E. and Sen, P. K. (1985). On shrinkage M-estimator of location parameters. *Communications in Statistics–Theory & Methods*, 14:2313–2329.

Schick, A. (1994). Estimation of the autocorrelation coefficient in the presence of a regression trend. *Statistics & Probability Letters*, 21:371–380.

Schick, A. (1996). Efficient estimation in a semiparametric additive regression model with autoregressive errors. *Stochastic Processes and their Applications*, 61:339–361.

Schick, A. (1998). An adaptive estimator of the autocorrelation coefficient in regression models with autoregressive errors. *Journal of Time Series Analysis*, 19:575–589.

Sen, P. and Singer, J. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman and Hall.

Sen, P. K. (1986). On the asymptotic distributional risk shrinkage and preliminary test version of the mean of a multivariate normal distribution. *Sankhya*, 48:354–371.

Sen, P. K. and Saleh, A. K. M. E. (1987). On preliminary test and shrinkage M-estimation in linear models. *The Annals of Statistics*, 15:4:1580–1592.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series. B*, 50:413–437.

Stamey, T., Kabalin, J., McNeal, J., Jhonstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii radical prostectomy treated patients. *Journal of Urology*, 16:1076–1083.

Stein, C. (1956). The admissibility of hotelling's $t^2$-test. *Mathematical Statistics*, 27:616–623.

Sun, J., Kopciuk, K. A., and Lu, X. (2008). Polynomial spline estimation of partially linear single-index proportional hazards regression models. *Computational Statistics & Data Analysis*, 53:176–188.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, pages 267–288.

Tibshirani, R. J. and Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics*, 3(2):822–829.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. Soviet Math Dok— *English translation of Dokl Akad Nauk SSSR 151, 1963, 501-504*, 4:1035–1038.

van Houwelingen, J. C. (2001). Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica*, 55(1):17–34.

Wang, Q., Linton, O., and Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, 99:334–345.

Xue, H., Lam, K. F., and Gouying, L. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of American the Statistical Association*, 99:346–356.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054.

Zheng, D., Wang, J., and Zhao, Y. (2006). Non-flat function estimation with a multi-scale support vector regression. *Neurocomputing*, 70:420–429.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(456):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

# Vita Auctoris

Enayetur Raheem was born in 1977 in Bangladesh. He completed his BSc (Honours) and MSc in Applied Statistics from University of Dhaka, Dhaka, Bangladesh. Shortly after graduation, he joined in the University of Dhaka as a lecturer and worked there before coming to McMaster University in Hamilton, Ontario, Canada in the Fall 2003 where he completed MSc in Statistics. He was conferred the degree of Doctor of Philosophy in Statistics in Summer 2012 at University of Windsor, Windsor, Ontario, Canada.