Positive-Shrinkage and Pretest Estimation in Multiple Regression: A Monte Carlo Study with Applications

SM Enayetur Raheem, S. Ejaz Ahmed

University of Windsor, Windsor, ON, Canada.

Abstract. Consider a problem of predicting a response variable using a set of covariates in a linear regression model. If it is *a priori* known or suspected that a subset of the covariates do not significantly contribute to the overall fit of the model, a restricted model that excludes these covariates, may be sufficient. If, on the other hand, the subset provides useful information, shrinkage method combines restricted and unrestricted estimators to obtain the parameter estimates. Such an estimator outperforms the classical maximum likelihood estimators. Any *prior* information may be validated through preliminary test (or pretest), and depending on the validity, may be incorporated in the model as a parametric restriction. Thus, pretest estimator chooses between the restricted and unrestricted estimators depending on the outcome of the preliminary test. Examples using three real life data sets are provided to illustrate the application of shrinkage and pretest estimation. Performance of positive-shrinkage and pretest estimators are compared with unrestricted estimator under varying degree of uncertainty of the prior information. Monte Carlo study reconfirms the asymptotic properties of the estimators available in the literature.

Keywords. Data analysis, James-Stein estimation, lasso, Monte Carlo simulation, multiple regression, pretest estimation, quadratic risk, RMSE, shrinkage estimation.

MSC: 62F12, 62F40, 62J05, 62J07.

1   Introduction

Regression analysis is one of the most mature and widely applied branch 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 model fits the data well or whether the estimated parameters are of much use depends on the validity of certain assumptions. In this setup, the estimates are obtained to have insights about the parameters. However, in many practical situations, it is the researchers who provide the estimation of the parameters utilizing the information contained in the sample and other relevant information. The "other" information may be considered as *non-sample information* (NSI). This is also known as *uncertain prior information* (UPI), or simply *prior* information. The non-sample information 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.

The quality of the fit and of the estimated parameters depend largely on the quality of the data used to obtain them. Only reliable information leads to useful results. However, in many practical situations, uncertainty arises as to whether the available information is of much use. 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.

With this keeping in mind, it is however, important to note that the consequences of incorporating non-sample information depend on the quality or usefulness of the information being added in the estimation process. Any uncertain prior information may be tested before they are incorporated in the model. Based on the idea of Bancroft (1944), uncertain prior information may be validated through preliminary test, and depending on the validity, may be incorporated in as a parametric restriction, and choose between the restricted or unrestricted estimation procedure depending on the outcome of the preliminary test.

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 non-sample information if it proves to be useful.

1.1 Review of Literature

Since the beginning, shrinkage estimation have received considerable attention from the researchers. Since 1987, Ahmed and his co-researchers are among others who have analytically demonstrated that shrinkage estimators outshine the classical maximum likelihood estimator. Asymptotic properties of shrinkage and preliminary test estimators using quadratic loss function have been studied, and their dominance over the usual maximum likelihood estimators demonstrated in numerous studies in the literature. Ahmed (1997) gave a detailed description of shrinkage estimation, and discussed large sample estimation techniques in a regression model with non-normal errors.

Khan and Ahmed (2003) considered the problem of estimating the coefficient vector of a classical regression model, and 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.

Estimation of the mean vector of a multivariate normal distribution, under the uncertain prior information that component means are equal but unknown, was studied by Khan and Ahmed (2006). Ahmed and Nicol (2010) among others, considered various large sample estimation techniques in a nonlinear regression model. Nonparametric estimation of the location parameter vector when uncertain prior information about the regression parameters is available was considered by Ahmed and Saleh (1999).

In this paper, we review positive shrinkage, and pretest estimators to compare their performance when certain information about a subset of the covariates are available *a priori.* In particular, we apply shrinkage estimation on three real life data sets to show the usability of positive-shrinkage and pretest estimators for practical purposes.

2 Statement of the Problem

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 parameter vector and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ 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 for the errors, only that εs have a cumulative distribution function $F(\varepsilon)$ with $E(\varepsilon) = \mathbf{0}$, and $E(\varepsilon\varepsilon') = \sigma^2 \mathbf{I}$, where σ^2 is finite. We make the following two assumptions, also called the regularity conditions

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 ith row of $\boldsymbol{X}$

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

In our case, suppose that $\boldsymbol{\beta}$ may be 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 $i = 1, 2$. Here, $\boldsymbol{\beta_1}$ is the coefficient vector for main effects, and $\boldsymbol{\beta_2}$ is a vector for "nuisance" effects. We are essentially interested in the estimation of $\boldsymbol{\beta}_1$ when it is plausible that $\boldsymbol{\beta}_2$ do not contribute significantly in predicting the response. Such a situation may arise when there is over-modeling and one wishes to cut down the irrelevant part from the model (2.1). For example, in studying the relationship between the level of prostate specific antigen (PSA) and some clinical measures, the log cancer volume and log prostate weight can be considered as the main effects while age, log of benign prostate hyperplasia amount, seminal vesicle invasion and others can be regarded as nuisance variables. In this situation, inference about $\boldsymbol{\beta_1}$ may benefit from shrinking the regression coefficients of the full model towards the restricted space while utilizing the available information contained in the nuisance covariates. 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 prior information about the subset of $\boldsymbol{\beta}$ can be written in terms of a restriction, $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$. Here, $\boldsymbol{H}$ is a known $p_2 \times p$ matrix and $\boldsymbol{h}$ is $p_2 \times 1$ vector of known constants.

2.1   Organization of the Paper

The paper is organized as follows. The statistical model is introduced in section 3. Shrinkage, positive-shrinkage, and pretest estimators are defined in this section. Examples using three real life data sets are presented in section 4. Positive-shrinkage and pretest estimators are obtained, and their performance are compared using cross-validation. Monte Carlo simulation study is described in section 5. Asymptotic bias and risk expressions for the shrinkage estimators are presented in

section 6. Finally, conclusions and future directions are presented in section 7.

3   The Model and Estimation Strategies

The unrestricted least-squares estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}^{\text{UR}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{C}^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

where $\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})$. Under the restriction $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$, the restricted estimator is given by

$$\hat{\boldsymbol{\beta}}^{\text{R}} = \hat{\boldsymbol{\beta}}^{\text{UR}} - \boldsymbol{C}^{-1}\boldsymbol{H}'(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}')^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}^{\text{UR}} - \boldsymbol{h}),$$

which is a linear function of the unrestricted estimator. Let us define the estimator of σ^2 by

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

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 defined by

$$\psi_n = \frac{(\boldsymbol{H}\hat{\boldsymbol{\beta}}^{\text{UR}} - \boldsymbol{h})'(\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}')^{-1}(\boldsymbol{H}\hat{\boldsymbol{\beta}}^{\text{UR}} - \boldsymbol{h})}{s_e^2}, \qquad (3.1)$$

which, under H_0, follows a chi-square distribution with p_2 degrees of freedom.

3.1   Shrinkage Estimator

A Stein-type estimator (STE) $\hat{\boldsymbol{\beta}}_1^{\text{S}}$ of $\boldsymbol{\beta}_1$ can be defined as

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

where ψ_n is defined in (3.1).

One problem with STE is that its components may have a different sign from the coordinates of $\hat{\boldsymbol{\beta}}_1^{\text{UR}}$. This could happen if $(p_2 - 1)\psi_n^{-1}$ is larger than unity. One possibility is when $p_2 = 2$ 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 STE. To overcome the sign problem, we define a positive-rule Stein-type semiparametric estimator (PSTE) by retaining the positive-part of the STE. A PSTE has the form

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

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

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

Ahmed (2001) and others studied the asymptotic properties of Stein-type estimators in various contexts.

3.2 Preliminary Test Estimator

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

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{UR}} - (\hat{\boldsymbol{\beta}}_1^{\mathrm{UR}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{R}})I(\psi_n < c_{n,\alpha}), \tag{3.2}$$

where $I(\cdot)$ is an indicator function, and $c_{n,\alpha}$ is the upper $100(1-\alpha)$ percentage point of the test statistic ψ_n.

In a pretest estimation problem, the *prior* information is tested before choosing the estimator for practical purposes, while shrinkage and positive-shrinkage estimator incorporates in the estimation process whatever *prior* information is available.

Pretest estimator either accepts of rejects the restricted estimator $(\hat{\boldsymbol{\beta}}_1^{\mathrm{R}})$ based on whether $\psi_n < c_{n,\alpha}$, while shrinkage estimator is a smoothed version of the pretest estimator.

4 Examples

In the following, we study three real life examples. For each data set, we fit linear regression models to predict the variable of interest form 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.

4.1 Assessment Criteria

In shrinkage and pretest estimation, we utilize the full-model and sub-model estimates, and combine them in a way that shrinks the least-squares estimates towards the sub-model estimates. In this framework, we utilize, if available, the information contained in the restricted subspace if they contribute significantly in predicting the response. However, in the absence of prior information about the nuisance subset, one might do usual variable selection to filter the nuisance subset out of the covariates. In that, one initiates the process with the model having all the covariates. Then the best subset may be selected based on AIC, BIC or other model selection criteria. Separate estimates from full- and restricted models are then combined to obtain shrinkage estimates. Finally, a model with shrunken coefficients is obtained, which reduces overall prediction error.

We obtain pretest and positive-shrinkage estimates using different sub-models. Performance of each pair of full- and sub-models was evaluated by estimating 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 data, 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.

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 5000 times, and estimate the average prediction errors along with their standard errors. The number of repetitions was initially varied, and settled with this as no noticeable variations in the standard errors were observed for higher values.

4.2 Prostate Data

Hastie *et al.* (2009) demonstrated various model selection techniques by

fitting linear regression model to the prostate data. Specifically, the log of prostate-specific antigen (`lpsa`) was modeled by the log cancer volume (`lcavol`), log prostate weight (`lweight`), age (`age`), log benign prostatic hyperplasia amount (`lbph`), seminal vesicle invasion (`svi`), log capsular penetration (`lcp`), Gleason score (`gleason`), and percentage Gleason scores 4 or 5 (`pgg45`). The idea is to predict `lpsa` from the measured variables.

The predictors were first standardized to have zero mean and unit standard deviation before fitting the model. Several model selection criteria and shrinkage methods were tried–details of which may be found in (Hastie *et al.*, 2009, Table 3.3, page 63). We consider the models obtained by AIC, BIC, and best subset selection (BSS) criteria, and consider them as our sub-models. They are listed in Table 1.

Table 1: 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`

Average prediction errors, and their standard deviations for pretest and shrinkage estimators for various sub-models are shown in Table 2. Prediction errors are based on five- and ten-fold cross validation. Average and standard errors are obtained after repeating the process 5000 times.

Looking at the bias corrected cross validation estimate of the prediction errors, on an average, restricted and the pretest estimators based on AIC have the smallest prediction errors. This is followed by pretest and the restricted estimators based on BIC. Interestingly, average prediction errors based on the sub-model given by BSS is much higher than those obtained from the models based on AIC or BIC. For instance, restricted model based on BSS has average prediction error 0.576, and the same for pretest estimator is 0.605. For the same sub-model, positive-shrinkage estimator has average prediction error 0.536, which is much less than R(BSS), and PT(BSS). Clearly, positive shrinkage estimator is beating the restricted and pretest estimators for this sub-model. This is a classic example where utility of positive-shrinkage estimator is practically

Table 2: Average prediction errors for various estimators based on K-fold cross validation repeated 5000 times for prostate data. Numbers in smaller font are the corresponding standard errors.

Estimator	Raw CVE		Bias Corrected CVE	
	$K = 5$	$K = 10$	$K = 5$	$K = 10$
UR	$.556_{.030}$	$.548_{.018}$	$.543_{.026}$	$.542_{.017}$
R(AIC)	$.535_{.023}$	$.529_{.014}$	$.525_{.020}$	$.523_{.013}$
R(BIC)	$.537_{.020}$	$.533_{.012}$	$.529_{.018}$	$.529_{.011}$
R(BSS)	$.582_{.017}$	$.578_{.010}$	$.576_{.015}$	$.576_{.009}$
PS(AIC)	$.554_{.029}$	$.547_{.018}$	$.540_{.025}$	$.541_{.017}$
PS(BIC)	$.546_{.026}$	$.541_{.016}$	$.533_{.023}$	$.535_{.015}$
PS(BSS)	$.549_{.026}$	$.542_{.016}$	$.536_{.023}$	$.536_{.015}$
PT(AIC)	$.536_{.024}$	$.529_{.014}$	$.526_{.021}$	$.525_{.014}$
PT(BIC)	$.538_{.021}$	$.533_{.012}$	$.529_{.019}$	$.529_{.011}$
PT(BSS)	$.599_{.030}$	$.601_{.024}$	$.602_{.036}$	$.605_{.029}$

realized. Restricted and/or pretest estimation may perform better under correct specification of the model (e.g., the models given by AIC and BIC for this data set), whereas, positive-shrinkage estimator is less sensitive to model misspecification.

Apparently, in the presence of imprecise subspace information, restricted and pretest estimators fail to produce the best estimates that reduce average prediction errors. On the other hand, positive-shrinkage estimator maintains a steady risk-superiority under model misspecification. This behaviour is illustrated in more detail through a Monte Carlo study in section 5.

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; per capita income (1974); illiteracy (1970, percent of population); life expectancy in years (1969-71); murder and non-negligent manslaughter rate per 100,000 population (1976); percent high-school graduates (1970); mean number of days with minimum temperature 32 degrees (1931-1960) in capital or large city;

and land area in square miles. We consider life expectancy as the response. 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 3.

Table 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 4: Average prediction errors (thousands) for various estimators based on K-fold cross validation, repeated 5000 times for state data. Numbers in smaller font are the corresponding standard errors.

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

When the models are correctly specified, it is obvious that restricted estimator will perform the best. Such is the scenario for the state data, where the model given by AIC and BIC are the same, and the restricted estimator has the smallest prediction error. Under model uncertainty, however, the scenario will change completely as restricted estimator becomes unbounded when the sub-model deviates from the true structure. This is explored in the simulation study presented in section 5. For the

correctly specified models, such as in Table 4, we see that restricted and pretest estimators have the smallest average prediction errors for both five-fold and ten-fold cross validation. The bias corrected version of the cross validation errors are exactly the same for the restricted and pretest estimators.

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 represents 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), 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 5.

Table 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 of the Galapagos data. Average prediction errors along with their standard errors for unrestricted (UR), restricted (R), positive-shrinkage (PS), and preliminary test or pretest (PT) estimators are presented in Table 6. Prediction errors and the standard errors are shown in thousands. PS(AIC) represents positive shrinkage estimates based on sub-model given by AIC, and PS(BIC) represents the same based on BIC. PT(AIC) and PT(BIC) are similarly defined for pretest

estimators.

Table 6: Average prediction errors (thousands) for various estimators based on K-fold cross validation, repeated 5000 times for Galapagos data. Numbers in smaller font are the corresponding standard errors.

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

For this example as well, since we have selected our sub-models based on AIC or BIC, they are likely to be true, which results in restricted and pretest estimators being the best estimators in terms of 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 shows us about the uncertainty in model specification, and the consequences that it cause. Monte Carlo study conducted later in the paper (section 5) reveals the sensitivity of restricted and pretest estimators, and shows that pretest and restricted estimators are outperformed by positive-shrinkage estimators when the underlying model is misspecified.

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.

5 Simulation Studies

Monte Carlo simulation experiments have been conducted to examine the quadratic risk performance of positive-shrinkage and pretest estima-

tors. We simulate the response from the following model:

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

where $x_{1i} = (\zeta_{1i}^{(1)})^2 + \zeta_i^{(1)} + \xi_{1i}$, $x_{2i} = (\zeta_{2i}^{(1)})^2 + \zeta_i^{(1)} + 2\xi_{2i}$, $x_{si} = (\zeta_{si}^{(1)})^2 + \zeta_i^{(1)}$ 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, ε_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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{\beta}_1, \mathbf{0})$. We show results for $\boldsymbol{\beta}_1 = (1,1,1)$, and $\boldsymbol{\beta}_1 = (1,1,1,1)$ only.

The number of simulations were initially varied. Finally, each realization was repeated 2000 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 data sets were generated from those distributions under local alternative hypothesis. Various Δ values between [0,1] have been considered.

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

$$\text{RMSE}(\hat{\boldsymbol{\beta}}_1^{\text{UR}} : \hat{\boldsymbol{\beta}}_1^*) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{\text{UR}})}{\text{MSE}(\hat{\boldsymbol{\beta}}_1^*)}, \tag{5.1}$$

where $\hat{\boldsymbol{\beta}}_1^*$ is one of the estimators considered in this study. The amount by which an RMSE is larger than unity indicates the degree of superiority of the estimator $\hat{\boldsymbol{\beta}}_1^*$ over $\hat{\boldsymbol{\beta}}_1^{\text{UR}}$.

RMSEs for the positive-shrinkage and pretest estimators were computed for $n = 30, 50, 100$, $p_1 = 3, 6, 9$, and $p_2 = 4, 6, 9$. Since the results are similar for all the configurations, we list the RMSEs in Table 7 for $n = 50$. Comparative RMSEs for positive-shrinkage and pretest estimators for $(p_1, p_2) = (3, 3)$, $(3, 6)$, $(4, 3)$, and $(4, 6)$ are illustrated in Figure 1.

5.1 Case 1: $\Delta = 0$

Clearly, for $\Delta = 0$, the restricted estimator outperforms all other estimators for all the cases considered in the simulation study. As the restriction moves away from $\Delta = 0$, the restricted estimator becomes unbounded (see the sharply decaying curve that goes below the horizontal line at $\hat{\boldsymbol{\beta}}_1^{\text{UR}}/\hat{\boldsymbol{\beta}}_1^* = 1$ for $\Delta > 0$). The positive-shrinkage estimator
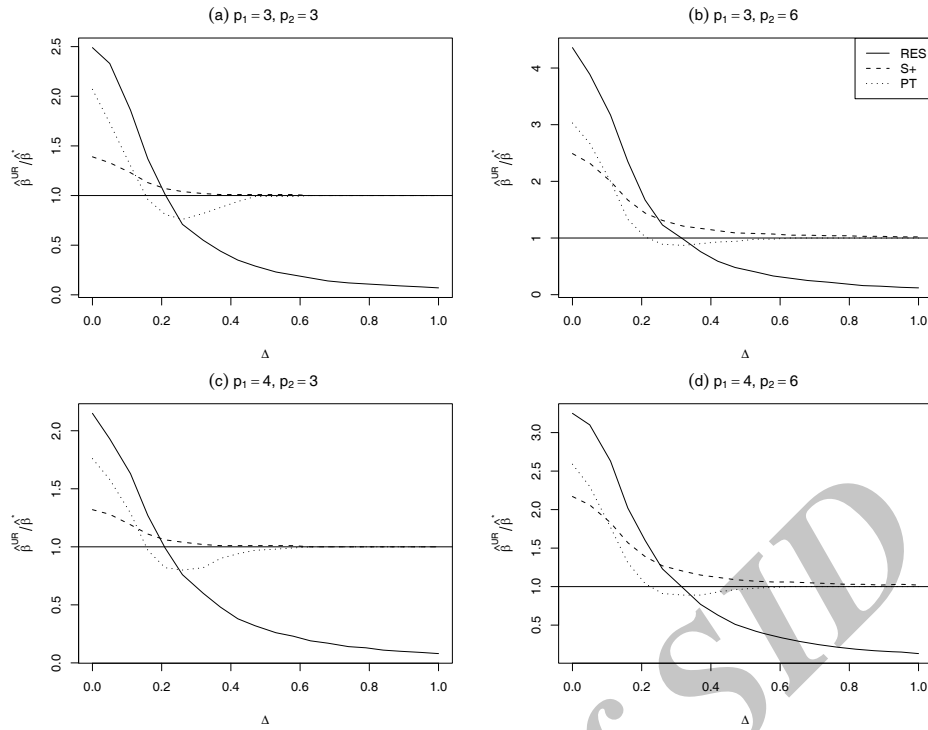
Figure 1: Relative mean squared error for restricted, positive-shrinkage, and pretest estimators for $n = 50$, and $(p_1, p_2) = (3, 3)$, $(3, 6)$, $(4, 3)$, and $(4, 5)$

approaches 1 at the slowest rate (for a range of Δ) as we move away from $\Delta = 0$. This indicates that in the event of imprecise subspace information (i.e., even if $\beta_2 \neq \mathbf{0}$), it has the smallest quadratic risk among all other estimators for a range of Δ. Pretest estimator outshines shrinkage estimators when Δ is in the neighbourhood of zero. Otherwise, it becomes unbounded at a faster rate than the restricted estimator. However, with the increase of Δ, at some point, RMSE of pretest estimator approaches 1 from below. This phenomenon suggests that neither pretest nor restricted estimator is uniformly better than the other when $\Delta > 0$.

5.2   Case 2: $\Delta > 0$

Simulation results suggest that positive shrinkage estimator maintains its superiority over the restricted and pretest estimators for a wide range

Table 7: Simulated relative mean squared error for restricted, positive-shrinkage, and pretest estimators with respect to unrestricted estimator for $p_1 = 4$, and $p_2 = 6$ for different Δ when $n = 50$.

Δ^*	$\hat{\boldsymbol{\beta}}_1^{\mathrm{R}}$	$\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}$	$\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}}$
0.00	3.25	2.17	2.59
0.05	3.10	2.06	2.30
0.11	2.63	1.83	1.77
0.16	2.02	1.57	1.31
0.21	1.60	1.39	1.04
0.26	1.23	1.27	0.91
0.32	0.98	1.20	0.89
0.37	0.77	1.15	0.89
0.42	0.63	1.12	0.93
0.47	0.51	1.09	0.96
0.53	0.42	1.07	0.98
0.58	0.36	1.06	0.99
0.63	0.31	1.06	1.00
0.68	0.27	1.05	1.00
0.74	0.23	1.04	1.00
0.79	0.20	1.03	1.00
0.84	0.18	1.03	1.00
0.89	0.16	1.02	1.00
0.95	0.15	1.03	1.00
1.00	0.13	1.02	1.00

of Δ. In particular, when $p_2 = 3$, the performance of positive-shrinkage estimator is superior for Δ up to around 0.35, after which point it is as good as the unrestricted estimator (panels a) and c) in Figure 1). However, when $p_2 = 6$, positive-shrinkage estimator maintains its risk-superiority over all other estimators for a wider range of Δ (see panels b) and d) in Figure 1). This clearly suggests that a positive-shrinkage estimator is preferred as there always remains uncertainty in specifying statistical models correctly. Moreover, one cannot go wrong with the positive-shrinkage estimators even if the assumed model is grossly wrong. In such cases, the estimates are as good or equal to the unrestricted (i.e., full model) estimates.

In the following sections, we review the asymptotic properties of the

estimators, and analytically present their bias and risk expressions.

6 Asymptotic Distribution of the Estimators

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

Under fixed alternative, the asymptotic distribution of $\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta})/s_e$ is equivalent to $\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathrm{UR}} - \boldsymbol{\beta})/s_e$. This suggest that in asymptotic setup, there is not much to investigate under a fixed alternative such as $\boldsymbol{H}\boldsymbol{\beta} \neq \boldsymbol{h}$. Therefore, to obtain meaningful asymptotics, a class of local alternatives, $\{K_n\}$, is considered, which is given by

$$K_n : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h} + \frac{\boldsymbol{\omega}}{\sqrt{n}}, \tag{6.1}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_{p_2})' \in \mathfrak{R}^{p_2}$ is a fixed vector. We notice that $\boldsymbol{\omega} = \boldsymbol{0}$ implies $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$, i.e., the fixed alternative is a particular case of (6.1). In the following, we evaluate the performance of each estimators 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 $\boldsymbol{W} = \boldsymbol{I}$, it is the simple squared error loss function.

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

$$\begin{aligned} R(\boldsymbol{\beta}^*, \boldsymbol{\beta}); \boldsymbol{W}) &= nE[(\boldsymbol{\beta}^* - \boldsymbol{\beta})' \boldsymbol{W} (\boldsymbol{\beta}^* - \boldsymbol{\beta})] \\ &= n \operatorname{tr}[\boldsymbol{W} \{E(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})'\}] \\ &= \operatorname{tr}(\boldsymbol{W}\boldsymbol{\Gamma}^*), \end{aligned} \tag{6.2}$$

where $\boldsymbol{\Gamma}^*$ 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

$$R(\boldsymbol{\beta}^0, \boldsymbol{\beta}) \leq R(\boldsymbol{\beta}^*, \boldsymbol{\beta}) \quad \forall(\boldsymbol{\beta}, \boldsymbol{W}) \tag{6.3}$$

with strict inequality holds for some $\boldsymbol{\beta}$. In such case, we say that the estimator $\boldsymbol{\beta}^0$ dominates $\boldsymbol{\beta}^*$. If, however, instead of (6.3) holding for every n, we have

$$\lim_{n \to \infty} R(\boldsymbol{\beta}^0, \boldsymbol{\beta}) \leq \lim_{n \to \infty} R(\boldsymbol{\beta}^*, \boldsymbol{\beta}) \quad \forall \boldsymbol{\beta}, \tag{6.4}$$

with strict inequality for some $\boldsymbol{\beta}$, then $\boldsymbol{\beta}^*$ is termed as asymptotically inadmissible estimator of $\boldsymbol{\beta}$. The expression in (6.3) is not easy to prove. An alternative is to consider the asymptotic distributional quadratic risk (ADQR) for the sequence of local alternative $\{K_n\}$.

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

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

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

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

be the dispersion matrix which is obtained from ADF, the ADQR may be defined as

$$R(\boldsymbol{\beta}^*; \boldsymbol{\beta}) = \operatorname{tr}(\boldsymbol{W}\boldsymbol{\Gamma}). \tag{6.5}$$

An estimator $\boldsymbol{\beta}^*$ is said to dominate an estimator $\boldsymbol{\beta}^0$ asymptotically if $R(\boldsymbol{\beta}^*; \boldsymbol{\beta}) \leq R(\boldsymbol{\beta}^0; \boldsymbol{\beta})$. 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. Sen (1986), and Saleh and Sen (1985) among others, have explained this point in various other contexts.

6.1 Asymptotic Bias and Risk Performance

To obtain the asymptotic distribution of the proposed estimators, and the test statistic ψ_n, we consider the following theorem.

Theorem 6.1. *Under the regularity conditions, and if $\sigma^2 < \infty$, as $n \to \infty$,*

$$\sqrt{n} \, s_e^{-1}(\hat{\boldsymbol{\beta}}^{UR} - \boldsymbol{\beta}) \stackrel{d}{\sim} N_p(\boldsymbol{0}, \boldsymbol{C}^{-1}).$$

6.1.1 Bias Performance

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

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

Theorem 6.2. *Under the assumed regularity conditions and theorem above, and under $\{K_n\}$, the ADB of the estimators are as follows:*

$$ADB(\hat{\boldsymbol{\beta}}_1^{UR}) = \mathbf{0} \tag{6.6}$$

$$ADB(\hat{\boldsymbol{\beta}}_1^{R}) = -\boldsymbol{C}_n^{-1}\boldsymbol{H}\boldsymbol{B}^{-1}\boldsymbol{\omega} \tag{6.7}$$

$$ADB(\hat{\boldsymbol{\beta}}_1^{PT}) = -\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{B}^{-1}\delta\boldsymbol{H}_{p_2+2}(\chi_{p_2,\alpha}^2;\Delta) \tag{6.8}$$

$$ADB(\hat{\boldsymbol{\beta}}_1^{S+}) = -\boldsymbol{C}^{-1}\boldsymbol{H}\boldsymbol{B}^{-1}\boldsymbol{\omega}\left[\boldsymbol{H}_{p_2+2}(p_2-2;\Delta) + (p_2-2)E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\}\right.$$
$$\left. + E\left\{\chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^2(\Delta) > p_2-2)\right\}\right] \tag{6.9}$$

where

$$E(\chi_p^{-2j}(\Delta)) = \int_0^\infty x^{-2j}d\Phi_p(x;\Delta)$$

and $\Phi_p(x;\Delta)$ is the cdf of a p-variate normal distribution with mean vector $\mathbf{0}$, and covariance matrix, Δ.

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 quadratic distributional bias (AQDB) of an estimator $\boldsymbol{\delta}$ of $\boldsymbol{\beta}_1$ by

$$AQDB(\boldsymbol{\delta}) = [ADB(\boldsymbol{\delta})]'\boldsymbol{\Sigma}[ADB(\boldsymbol{\delta})]$$

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

Using the definition, and following Ahmed (1997), the asymptotic quadratic distributional bias of the various estimators are presented below.

$$\mathrm{AQDB}(\hat{\boldsymbol{\beta}}_1^{\mathrm{UR}}) = \mathbf{0}, \tag{6.10}$$

$$\mathrm{AQDB}(\hat{\boldsymbol{\beta}}_1^{\mathrm{R}}) = \frac{\boldsymbol{\xi}'\boldsymbol{\xi}}{\sigma^2\boldsymbol{C}^{-1}} = \Delta \tag{6.11}$$

$$\mathrm{AQDB}(\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}}) = \Delta\left\{\boldsymbol{H}_{p_2+2}(\chi_{p_2,\alpha}^2;\Delta)\right\}^2 \tag{6.12}$$

$$ADQB(\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}) = \Delta\left[\boldsymbol{H}_{p_2+2}(p_2-2;\Delta) + (p_2-2)E\left\{\chi_{p_2+2}^{-2}(\Delta)\right\}\right.$$
$$\left. + E\left\{\chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^2(\Delta) > p_2-2)\right\}\right]. \tag{6.13}$$

6.1.2   Risk Performance

Following Ahmed (1997), we present the risk expressions of the estimators.

Theorem 6.3. *Under the assumed regularity conditions, and local alternative $\{K_n\}$, the ADQR expressions are as follows:*

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{UR}}; \boldsymbol{W}) = \sigma^2 \mathrm{tr}(\boldsymbol{WC}^{-1}) \tag{6.14}$$

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{R}}; \boldsymbol{W}) = \sigma^2 \mathrm{tr}(\boldsymbol{WC}^{-1}) - \sigma^2 \mathrm{tr}(\boldsymbol{Q}) + \boldsymbol{\omega}'\boldsymbol{B}^{-1}\boldsymbol{Q}\boldsymbol{\omega} \tag{6.15}$$

$$\begin{aligned}
R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}; \boldsymbol{W}) = {}& \sigma^2 \mathrm{tr}(\boldsymbol{WC}^{-1}) - (p_2 - 2)\sigma^2 \mathrm{tr}(\boldsymbol{Q}_{11}) \\
& \times \left\{ 2E[\chi_{p_2+4}^{-4}(\Delta)] - (p_2 - 2)E[\chi_{p_2+4}^{-4}(\Delta)] \right\} \\
& + (p_2 - 2)(p_2 + 6)(\gamma_1'\boldsymbol{Q}_{11}\gamma_1)E[\chi_{p_2+4}^{-4}(\Delta)] \tag{6.16}
\end{aligned}$$

$$\begin{aligned}
R(\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}}; \boldsymbol{W}) = {}& \sigma^2 \mathrm{tr}(\boldsymbol{WC}^{-1}) - \sigma^2 \mathrm{tr}(\boldsymbol{Q})\boldsymbol{H}_{p_2+2}(\chi_{p_2,\alpha}^2; \Delta) \\
& + \boldsymbol{\omega}'\boldsymbol{B}^{-1}\boldsymbol{\omega}\left\{ 2\boldsymbol{H}_{p_2+2}(\chi_{p_2,\alpha}^2; \Delta) - \boldsymbol{H}_{p_2+4}(\chi_{p_2,\alpha}^2; \Delta) \right\} \tag{6.17}
\end{aligned}$$

$$\begin{aligned}
R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S+}}; \boldsymbol{W}) = {}& R(\hat{\boldsymbol{\beta}}_1^{\mathrm{S}}; \boldsymbol{W}) + (p_2 - 2)\sigma^2 \mathrm{tr}(\boldsymbol{Q}) \\
& \times \Big[E\left\{ \chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^2(\Delta) \leq p_2 - 2) \right\} \\
& \quad - (p_2 - 2)E\left\{ \chi_{p+2+2}^{-4}(\Delta)I(\chi_{p_2+2}^2(\Delta) \leq p_2 - 2) \right\} \Big] \\
& - \sigma^2 \mathrm{tr}(\boldsymbol{Q})\boldsymbol{H}_{p_2+2}(p_2 - 2; \Delta) \\
& + \boldsymbol{\omega}'\boldsymbol{B}^{-1}\boldsymbol{Q}\boldsymbol{\omega}\left\{ 2\boldsymbol{H}_{p_2+4}(p_2 - 2; \Delta) \right\} \\
& - (p_2 - 2)\boldsymbol{\omega}'\boldsymbol{B}^{-1}\boldsymbol{Q}\boldsymbol{\omega}\Big[2E\left\{ \chi_{p_2+2}^{-2}(\Delta)I(\chi_{p_2+2}^2(\Delta) \leq p_2 - 2) \right\} \\
& \quad -2E\left\{ \chi_{p_2+4}^{-2}(\Delta)I(\chi_{p_2+4}^2(\Delta) \leq p_2 - 2) \right\} \\
& \quad + (p_2 - 2)E\left\{ \chi_{p_2+4}^{-4}(\Delta)I(\chi_{p_2+4}^{-4}(\Delta) \leq p_2 - 2) \right\} \Big], \tag{6.18}
\end{aligned}$$

where $\boldsymbol{Q} = \boldsymbol{HC}^{-1}\boldsymbol{WC}^{-1}\boldsymbol{H}'\boldsymbol{B}^{-1}$.

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 $\Delta = 0$,

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{R}} \succ \hat{\boldsymbol{\beta}}_1^{\mathrm{P}} \succ \hat{\boldsymbol{\beta}}_1^{\mathrm{UR}}.$$

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

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

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

7 Discussion

In this paper, we reviewed positive-shrinkage and pretest estimation in the context of a multiple linear regression model. In our study, we presented asymptotic bias and the risk expressions for the estimators.

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, which forms 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.

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 shows 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, however, do not tell us how sensitive are the prediction errors under model misspecification. Therefore, we conduct Monte Carlo simulation to study such characteristics for positive-shrinkage and pretest estimators under varying Δ, and different sizes of the nuisance subsets.

In Monte Carlo study, we numerically computed relative mean squared errors for the restricted, positive-shrinkage, and pretest estimators with

respect to the unrestricted estimator. 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 ($\Delta > 0$), risk of the restricted estimator becomes unbounded. Pretest estimator becomes unbounded even faster than the restricted estimator for the cases considered in the simulation. However, as the Δ increases, pretest estimator performs better for some Δ, and approaches from below to merge with the line where RMSE is unity. On the other hand, positive-shrinkage estimator decays at the slowest rate with the increase of Δ, and perform steadily throughout a wider range of the alternative parameter subspace. In particular, when the nuisance subset is large, positive-shrinkage estimators outperforms all other estimators, which can be seen in panels b) and d) in Figure 1.

7.1 Future Directions

Pretest estimator either selects restricted or unrestricted estimator depending on the significance based on a test statistic, while positive-shrinkage estimator shrinks the covariates towards the restricted subspace. The nuisance subset is ideally a null space when they do not contribute significantly in the estimation process. In this sense, shrinkage estimators resemble penalized estimators such as the least absolute penalty and selection operator, *lasso*. Proposed by Tibshirani (1996), lasso is a member of the penalized least squares (PLS) family, which performs variable selection and parameter estimation simultaneously. Lasso is closely related with ridge regression, which is obtained by minimizing the penalized residual sum of squares,

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

Here, γ is the tuning parameter which controls the amount of shrinkage. Lasso solutions are similarly defined by replacing the squared penalty $\sum_{j=1}^{p_1+p_2} \beta_j^2$ in the ridge solution with the absolute penalty $\sum_{j=1}^{p_1+p_2} |\beta_j|$ in the lasso,

$$\tilde{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p_1+p_2} x_{ij}\beta_j)^2 + \gamma \sum_{j=1}^{p_1+p_2} |\beta_j| \right\}.$$

Although the change apparently looks subtle, the absolute penalty term made it impossible to have analytic solution for the lasso. Originally,

lasso solutions were obtained via quadratic programming. Nowadays, faster algorithm, such as cyclical coordinate descent, is used to obtain lasso estimates.

Shrinkage estimation does variable selection by shrinking the coefficients towards the restricted sub-space. In doing so, some of the coefficients shrink towards zero, while some over-shrinks–producing a negative sign for the coefficient. The change of sign may be uncomfortable for practitioners, although it does not affect the risk performance. The positive-part shrinkage estimator takes care of the negative sign by setting the coefficient to exactly zero. In the process, most of the coefficients are shrunk while some of them are eliminated by shrinking to zero.

Since the introduction of lasso, there has been a tremendous amount of development in lasso and related absolute penalty estimation (APE) during the past one and a half decade. Although the lasso and shrinkage methods have been around for quite some time, little work has been done to compare their relative performance. Recently, Ahmed *et al.* (2007) compared positive shrinkage and lasso in a partially linear regression setup. However, no comparative study for shrinkage and absolute penalty estimators in multiple linear regression model has been found in the reviewed literature. We are currently working on this front, and the findings will be disseminated through future communications.

References

Ahmed, S. E. (1997), Asymptotic shrinkage estimation: the regression case. Applied Statistical Science II, 113–139

Ahmed, S. E. (2001), Shrinkage estimation of regression coefficients from censored data with multiple observations. In Empirical Bayes and Likelihood Inference, Lecture Notes in Statistics, ed. S. E. Ahmed and N. Reid, volume 148, 103-120, New York: Springer-Verlag.

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 and Nicol, C. J. (2010), An application of shrinkage estimation to the nonlinear regression model. Computational Statistics & Data Analysis, In Press.

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.

Faraway, J. J. (2002), Practical Regression and Anova using R.

Hastia, T., Tibshirani, R., and Friedman, J. (2009), The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer.

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.

Khan, B. U. and Ahmed, S. E. (2006), Comparisons of improved risk estimators of the multivariate mean vector. Computational Statistics & Data Analysis, **50**(2), 402-421.

Saleh, A. K. Md. E. and Sen, P. K. (1985), On shrinkage M-estimator of location parameters. Communications in Statistics–Theory & Methods, **14**, 2313–2329.

Sen, P. K. (1986), On the asymptotic distributional risk shrinkage and preliminary test vaersion of the mean of a multivariate normal distribution. Sankhya, **48**, 354–371.

Stein, C. (1956), The admissibility of Hotelling's T^2-test. Mathematical Statistics, **27**, 616–623.

Tibshirani, R. (1996), Regression shrinkage and selection via the LASSO. Journal of 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.