Summer 7-1-2019

PREDICTION INTERVALS FOR SCALED SHRINKAGE ESTIMATORS

Lingling Zhang
lingling.zhang@siu.edu

PREDICTION INTERVALS FOR SCALED SHRINKAGE ESTIMATORS

by

Lingling Zhang

B.S., Shandong Normal University, 2017

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
30th May, 2019

RESEARCH PAPER APPROVAL

PREDICTION INTERVALS FOR SCALED SHRINKAGE ESTIMATORS

by

Lingling Zhang

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Mathematics

Approved by:

David J. Olive

H.R.Hughes

Michael Sullivan

Graduate School
Southern Illinois University Carbondale
30th May, 2019

AN ABSTRACT OF THE RESEARCH PAPER OF

Lingling Zhang, for the Master of Science degree in MATHEMATICS, presented May, 2019, at Southern Illinois University Carbondale.

TITLE: PREDICTION INTERVALS FOR SCALED SHRINKAGE ESTIMATORS

MAJOR PROFESSOR: Dr. David J. Olive

Consider the multiple linear regression model $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$ with sample size n. Let $\hat{\boldsymbol{\beta}}$ be a shrinkage estimator of $\boldsymbol{\beta}$ such as elastic net, lasso, or ridge regression. These estimators often shrink the slope estimators $\hat{\beta}_i$ too much. Then the intercept estimator $\hat{\beta}_1$ is also poor. As a remedy, do a simple linear regression of Y on $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ to get the scaled shrinkage estimator $\hat{\boldsymbol{\beta}}_{SA}$ where $\hat{\beta}_{iSA} = \hat{b}\hat{\beta}_i$ for $i = 2, ..., p$ and $\hat{\beta}_{1SA} = \hat{a} + \hat{b}\hat{\beta}_1$. Two prediction intervals are used to compare the shrinkage estimators with the scaled shrinkage estimators.

KEY WORDS: Elastic Net, Lasso, Ridge Regression, Prediction Interval.

ACKNOWLEDGMENTS

I would like to take this opportunity to thank my research advisor, Dr. David Olive for overseeing my Master's project and Dr. H.R.Hughes and Dr. Michael Sullivan for sitting on my committee. I would also like to thank the professors of Southern Illinois University for their instruction and care over the past one years. Finally I want to say to my family (especially my mother), thank you for all of your support and encouragement. You have pushed me to succeed and I could not have done it without you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Suppose that the response variable Y_i and at least one predictor variable $x_{i,j}$ are quantitative with $x_{i,1} \equiv 1$. Let $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p}) = (1 \;\; \boldsymbol{u}_i^T)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ where β_1 corresponds to the intercept. Then the multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1.1}$$

for $i = 1, ..., n$. This model is also called the full model. Here n is the sample size, and assume that the zero mean random variables e_i are independent and identically distributed (iid) with variance $V(e_i) = \sigma^2$. In matrix notation, these n equations become

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

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. The ith fitted value $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ and the ith residual $r_i = Y_i - \hat{Y}_i$ where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. Ordinary least squares (OLS) is often used for inference if n/p is large.

For some shrinkage estimators, such as lasso, $\hat{Y}_i$ depends on the scale of the predictors. Algorithms for such estimators often use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W} = (W_{ij})$ where $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Note that the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is $\boldsymbol{R}_{\boldsymbol{u}} = \boldsymbol{W}^T\boldsymbol{W}/n$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{1.3}$$

where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Three important shrinkage estimators of $\boldsymbol{\beta}$ are the elastic net due to Zou and Hastie [27] , lasso due to Tibshirani [26], and ridge regression (RR): see Hoerl and Kennard [8]. Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j \tag{1.4}$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression, $j = 1$ corresponds to lasso, and $a = 1, 2, n$, and $2n$ are common. A fourth estimator, relaxed lasso, applies OLS to a constant and the predictors that had nonzero lasso coefficients. See Efron et al. [3] and Meinshausen [13]. The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Following Hastie, Tibshirani, and Wainwright [7], the elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{1.5}$$

where $\lambda_1 = (1 - \alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ with $0 \leq \alpha \leq 1$.

The elastic net, lasso, relaxed lasso, and ridge regression estimators produce M models and use a criterion to select the final model (e.g., 10-fold cross validation (CV)). The number of models M depends on the method. Lasso and ridge regression have a parameter λ. When $\lambda = 0$, the OLS full model is used. These methods also use a maximum value λ_M of λ and a grid of M λ values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{M-1} < \lambda_M$. For lasso, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ for $i < M$. See James et al. [9] and Hastie, Tibshirani, and Wainwright [7].

Variable selection is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large, and so that the model with the remaining predictors is useful for prediction. Following Olive and Hawkins [20], a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{1.6}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (1.6) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{1.7}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$.

Consider regressing Y on $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ to get $\tilde{\boldsymbol{Y}} = \hat{a} + \hat{b}\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_A$ be a shrinkage estimator. The Olive [19] and Pelawa Watagoda and Olive [21] scaled shrinkage estimator $\hat{\boldsymbol{\beta}}_{SA}$ has $\hat{\beta}_{iSA} = \hat{b}\hat{\beta}_i$ for $i = 2, ..., p$ and $\hat{\beta}_{1SA} = \hat{a} + \hat{b}\hat{\beta}_1$. Shrinkage estimators often shrink the slope estimators $\hat{\beta}_i$ too much. Relaxed lasso is a remedy if the model is sparse: a_S is small. A fitted model is sparse if the number d of nonzero coefficients in $\hat{\boldsymbol{\beta}}$ is small. We want $n \geq 10d$ to avoid overfitting. Relaxed lasso is useful if the population model and fitted model are both sparse. The scaled shrinkage estimator may be useful if the population model or fitted model is not sparse. Ridge regression has $d = p$, and hence is not a sparse fitted model. For ridge regression, we could let d be a plug in degrees of freedom: compute the degrees of freedom as if the model was selected in advance rather than after model selection with 10-fold CV. Thus a plug in degrees of freedom is not the actual degrees of freedom, which tends to be hard to compute when model or variable selection is used.

Response plots of the fitted values $\hat{Y}$ versus the response Y are useful for checking linearity of the MLR model and for detecting outliers. If the error distribution is unimodal and not highly skewed, if $n \geq 10d$, and if the MLR model (1.1) is good, then the plotted points in the response plot should scatter in a roughly even band about the identity line with zero intercept and unit slope. Residual plots should also be made. We call $\boldsymbol{x}^T\boldsymbol{\beta}$ a sufficient predictor and $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ an estimated sufficient predictor (ESP).

Example 1. Suppose $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_{101} x_{101} + e = x_2 + e$ with $n = 100$ and $p = 101$. This model is sparse and lasso performs well. Ridge regression shrinks too much and $\hat{\beta}_1$ is poor, but the correlation $cor(\hat{\boldsymbol{Y}}_{RR}, \boldsymbol{Y}) = 0.91$. See the response plots in Figure 1.1 which has the 90% pointwise prediction interval (PI) (2.8) bands added to the plot as two lines parallel to the identity line. See Section 2. The response plot in Figure 1.2 shows the scaled ridge regression estimator fits the data much better than the ridge regression estimator in Figure 1.1. Some R code is below.

```
library(glmnet)
set.seed(13)
par(mfrow=c(2,1))
x <- matrix(rnorm(10000),nrow=100,ncol=100)
```

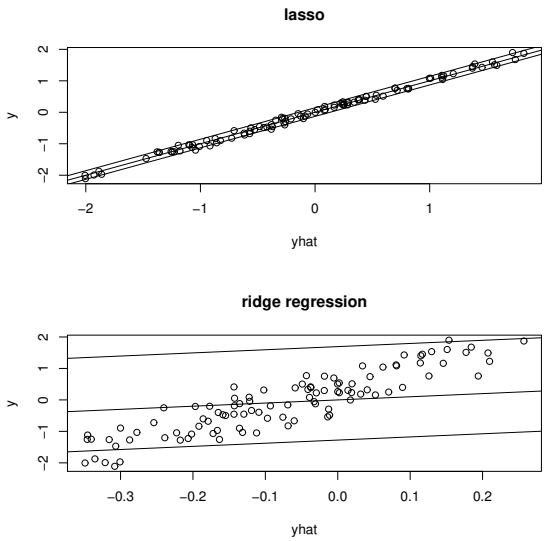

Figure 1.1. Response Plots for Example 1.

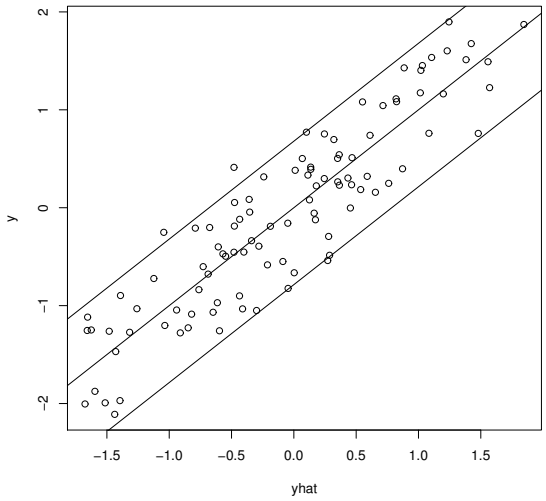

Figure 1.2. Response Plot for Scaled Ridge Regression Estimator.

```
Y <- x[,1] + rnorm(100,sd=0.1)
#sparse model, iid predictors
out <- cv.glmnet(x,Y,alpha=1) #lasso
lam <- out$lambda.min
fit <- predict(out,s=lam,newx=x)
res<- Y-fit
#PI bands used d = 1
AERplot2(yhat=fit,y=Y,res=res)
title("lasso")
cor(fit,Y) #about 0.997
tem <- lsfit(fit,Y)
tem$coef   #changes even if set.seed is used
#   Intercept            1
#0.0009741988 1.0132965955
out <- cv.glmnet(x,Y,alpha=0) #ridge regression
lam <- out$lambda.min
fit <- predict(out,s=lam,newx=x)
res<- Y-fit
#PI bands used d = 1
AERplot2(yhat=fit,y=Y,res=res)
#$respi
#[1] -1.276461  1.693856  #PI length about  2.97
title("ridge regression")
par(mfrow=c(1,1))
#ridge regression shrank betahat and ESP too much
cor(fit,Y) #about 0.91
tem <- lsfit(fit,Y)
tem$coef
```

```
# Intercept          1
#0.3523725   5.8094443   #Fig. 1.1 has -0.7008187  5.7954084
fit2 <- Y-tem$resid   #Y = yhat + r, fit2 = yhat for scaled RR estimator
plot(fit2,Y)  #response plot is much better
abline(0,1)
rrcoef <- predict(out,type="coefficients",s=lam)
plot(rrcoef)
bhat <- tem$coef[2]*rrcoef
bhat[1] <- bhat[1] + tem$coef[1]
#bhat is the betahat for the new ESP fit2
fit3 <- x%*%bhat[-1] + bhat[1]
plot(fit2,fit3)
max(abs(fit2-fit3))
#[1] 1.110223e-15
plot(rrcoef)
plot(bhat)
res2 <- Y - fit2
AERplot2(yhat=fit2,y=Y,res=res2)
$respi
[1] -0.7857706  0.6794579  #PI length about 1.47
title("Response Plot for Scaled Ridge Regression Estimator")
```

Section 2 gives the two prediction intervals used in the simulation study, and Section 3 gives some large sample theory for shrinkage estimators. Section 4 gives a simulation for the prediction intervals to compare lasso and ridge regression with scaled lasso and scaled ridge regression. Sections 2 and 3 follow Pelawa Watagoda and Olive [21] closely.

CHAPTER 2

PREDICTION INTERVALS AFTER MODEL SELECTION

Consider predicting a future test response variable Y_f given a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ and training data $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$. A large sample $100(1-\delta)\%$ prediction interval (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1-\delta$ as the sample size $n \to \infty$. A PI is asymptotically optimal if $[\hat{L}_n, \hat{U}_n] \to [L_s, U_s]$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering $100(1-\delta)\%$ of the mass.

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals if the data are iid. Let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics of $Z_1, ..., Z_n$. Then let the shortest closed interval containing at least c of the Z_i be

$$\text{shorth(c)} = [\text{Z}_{(\text{s})}, \text{Z}_{(\text{s+c}-1)}]. \tag{2.1}$$

Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{2.2}$$

Frey [5] showed that for large $n\delta$ and iid data, the shorth(k_n) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth$(c)$ estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(n, \lceil n[1-\delta+1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{2.3}$$

Example 2. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76
order data: 76 78 89 111 778
                  13 = 89 - 76
                      33 = 111 - 78
                         689 = 778 - 89
shorth(3) = [76,89]
```

The additive error regression model is $Y = m(\boldsymbol{x}) + e$ where $m(\boldsymbol{x})$ is a real valued function and the e_i are iid, often with zero mean and constant variance $V(e) = \sigma^2$. Model (1.1) is a special case with $m(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. The large sample theory for prediction intervals is simple for this model. Cai et al. [1] proved that the shorth PI works for multiple linear regression. Let the residuals $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$ for $i = 1, ..., n$. Assume $\hat{m}(\boldsymbol{x})$ is a consistent estimator of $m(\boldsymbol{x})$ such that the sample percentiles $[\hat{L}_n(r), \hat{U}_n(r)]$ of the residuals are consistent estimators of the population percentiles $[L, U]$ of the error distribution where $P(e \in [L, U]) = 1 - \delta$. Let $\hat{Y}_f = \hat{m}(\boldsymbol{x}_f)$. Then $P(Y_f \in [\hat{Y}_f + \hat{L}_n(r), \hat{Y}_f + \hat{U}_n(r)] \rightarrow P(Y_f \in [m(\boldsymbol{x}_f) + L, m(\boldsymbol{x}_f) + U]) = P(e \in [L, U]) = 1 - \delta$ as $n \rightarrow \infty$. Three common choices are a) $P(e \leq U) = 1 - \delta/2$ and $P(e \leq L) = \delta/2$, b) $P(e^2 \leq U^2) = P(|e| \leq U) = P(-U \leq e \leq U) = 1 - \delta$ with $L = -U$, and c) the population shorth is the shortest interval $U - L$ such that $P[e \in [L, U]) = 1 - \delta$. The PI c) is asymptotically optimal while a) and b) are asymptotically optimal on the class of symmetric zero mean unimodal error distributions.

Prediction intervals based on the shorth of the residuals need a correction factor for good coverage since the residuals tend to underestimate the errors in magnitude. For lasso, let d be the number of variables used by the method: the number of nonzero $\hat{\boldsymbol{\beta}}_i$, including $\hat{\boldsymbol{\beta}}_1$. We could also let d be equal to a plug in estimate of model degrees of freedom.

For n/p large and $d = p$, Olive [15] developed prediction intervals for models of the form $Y_i = m(\boldsymbol{x}_i) + e_i$. The first Pelawa Watagoda and Olive [21] PI, that can be useful even if n/p is not large, is defined below. This PI modifies the Olive [15] PI that can only be computed if $n > p$. Olive [14][16][17][18] used similar correction factors for several prediction intervals and prediction regions with $d = p$. We want $n \geq 10d$ so that the model does not overfit.

If the OLS model I has d predictors, and $S \subseteq I$, then

$$E(MSE(I)) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-d}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right)$$

and $MSE(I)$ is a $\sqrt{n}$ consistent estimator of σ^2 for many error distributions by Su and Cook [25]. For a wide range of regression models, extrapolation occurs if the leverage $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{x}_f > 2d/n$: if $\boldsymbol{x}_{I,f}$ is too far from the data $\boldsymbol{x}_{I,1}, ..., \boldsymbol{x}_{I,n}$, then the model may not hold and prediction

can be arbitrarily bad. These results suggests that

$$\sqrt{\frac{n}{n-d}}\sqrt{(1+h_f)} \;\; r_i \approx \sqrt{\frac{n+2d}{n-d}} \;\; r_i \approx e_i.$$

In simulations for prediction intervals and prediction regions with $n = 20d$, the maximum simulated undercoverage was near 5% if q_n in (2.5) is changed to $q_n = 1 - \delta$.

Next we give the correction factor and the first prediction interval. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.} \tag{2.4}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let

$$c = \lceil nq_n \rceil, \tag{2.5}$$

and let

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \tag{2.6}$$

if $d \leq 8n/9$, and

$$b_n = 5\left(1 + \frac{15}{n}\right),$$

otherwise. As d gets close to n, the model overfits and the coverage will be less than the nominal. The piecewise formula for b_n allows the prediction interval to be computed even if $d \geq n$. Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then the first 100 $(1-\delta)$% large sample PI for Y_f is

$$[\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}]. \tag{2.7}$$

The second PI randomly divides the data into two half sets H and V where H has $n_H = \lceil n/2 \rceil$ of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $\hat{m}_H(\boldsymbol{x})$ is computed using the training data set H. Then the validation residuals $v_j = Y_{i_j} - \hat{m}_H(\boldsymbol{x}_{i_j})$ are computed for the $j = 1, ..., n_V$ cases in the validation set V. Find the Frey PI $[v_{(s)}, v_{(s+c-1)}]$ of the validation residuals (replacing n in (2.3) by $n_V = n - n_H$). Then the second new $100(1 - \delta)$% large sample PI for Y_f is

$$[\hat{m}_H(\boldsymbol{x}_f) + v_{(s)}, \hat{m}_H(\boldsymbol{x}_f) + v_{(s+c-1)}]. \tag{2.8}$$

We can also motivate PI (2.8) by modifying the justification for the Lei et al. [12] split conformal prediction interval $[\hat{m}_H(\boldsymbol{x}_f) - a_q, \hat{m}_H(\boldsymbol{x}_f) + a_q]$ where a_q is an appropriate quantile of the absolute validation residuals. PI (2.8) is a modification of the split conformal PI that is asymptotically optimal. Suppose $(Y_i, \boldsymbol{x}_i)$ are iid for $i = 1, ..., n, n+1$ where $(Y_f, \boldsymbol{x}_f) = (Y_{n+1}, \boldsymbol{x}_{n+1})$. Compute $\hat{m}_H(\boldsymbol{x})$ from the cases in H. For example, get $\hat{\boldsymbol{\beta}}_H$ from the cases in H. Consider the validation residuals v_i for $i = 1, ..., n_V$ and the validation residual v_{n_V+1} for case $(Y_f, \boldsymbol{x}_f)$. Since these $n_V + 1$ cases are iid, the probability that v_t has rank j for $j = 1, ..., n_V + 1$ is $1/(n_V + 1)$ for each t, i.e., the ranks follow the discrete uniform distribution. Let $t = n_V + 1$ and let the $v_{(j)}$ be the ordered residuals using $j = 1, ..., n_V$. That is, get the order statistics without using the unknown validation residual v_{n_V+1}. Then $v_{(i)}$ has rank i if $v_{(i)} < v_{n_V+1}$ but rank $i + 1$ if $v_{(i)} > v_{n_V+1}$. Thus

$$P(Y_f \in [\hat{m}_H(\boldsymbol{x}_f) + v_{(k)}, \hat{m}_H(\boldsymbol{x}_f) + v_{(k+b-1)}]) = P(v_{(k)} \leq v_{n_V+1} \leq v_{(k+b-1)}) \geq$$

$P(v_{n_V+1}$ has rank between $k+1$ and $k+b-1$ and there are no tied ranks$) \geq (b-1)/(n_V+1) \approx 1-\delta$ if $b = \lceil (n_V + 1)(1 - \delta) \rceil + 1$ and $k + b - 1 \leq n_V$. This probability statement holds for a fixed k such as $k = \lceil n_V \ \delta/2 \rceil$. The statement is not true when the shorth(b) estimator is used since the shortest interval using $k = s$ can have s change with the data set. That is, s is not fixed. Hence if PI's were made from J independent data sets, the PI's with fixed k would contain Y_f about $J(1 - \delta)$ times, but this value would be smaller for the shorth(b) prediction intervals where s can change with the data set. The above argument works if the estimator $\hat{m}(\boldsymbol{x})$ is "symmetric in the data," which is satisfied for multiple linear regression estimators.

The PIs (2.7) and (2.8) can be used with $\hat{m}(\boldsymbol{x}) = \hat{Y}_f = \boldsymbol{x}_{I_d}^T \hat{\boldsymbol{\beta}}_{I_d}$ where I_d denotes the index of predictors selected from the model or variable selection method. If $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, the Pelawa and Watagoda and Olive [21] PIs (2.7) and (2.8) are asymptotically optimal for a large class of error distributions while the split conformal PI needs the error distribution to be unimodal and symmetric for asymptotic optimality. Since $\hat{m}_H$ uses $n/2$ cases, $\hat{m}_H$ has about half the efficiency of $\hat{m}$. When $p \geq n$, the regularity conditions for consistent estimators are strong. See the last paragraph of Section 3 for references. If the estimator is not consistent, the split conformal PI and PI (2.8) can have coverage closer to the nominal coverage than PI (2.7). For example, if $\hat{m}$ interpolates the data and $\hat{m}_H$ interpolates the training data from H, then the validation

residuals will be huge. Hence PI (2.8) will be long compared to PI (2.7). For a good fitting model, residuals r_i tend to be smaller in magnitude than errors e_i. Hence complicated correction factors are needed. The validation residuals v_j tend to be larger in magnitude than the e_i, and thus the Frey correction factor can be used.

CHAPTER 3

LARGE SAMPLE THEORY

The estimators elastic net, lasso, and ridge regression have R programs and large sample theory related to that of OLS. First we will let p be fixed.

Assume that the sample correlation matrix

$$\boldsymbol{R_u} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1} \tag{3.1}$$

where $\boldsymbol{V}^{-1} = \boldsymbol{\rho_u}$, the population correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$, if the $\boldsymbol{u}_i$ are a random sample from a population. Under (3.1), if $\lambda_{1,n}/n \to 0$ then

$$\frac{\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}, \text{ and } \mathrm{n}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \lambda_{1,\mathrm{n}}\boldsymbol{I}_{\mathrm{p-1}})^{-1} \xrightarrow{\mathrm{P}} \boldsymbol{V}.$$

Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,\ldots,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Then from Sen and Singer [23], the OLS estimator satisfies

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{3.2}$$

The following identity from Gunst and Mason [6] is useful for ridge regression inference: $\hat{\boldsymbol{\eta}}_R =$

$$(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z} = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

$$[\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{B}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

$$\hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1n}}{n}n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

The following identity from Efron and Hastie [2], for example, is useful for inference for the lasso estimator $\hat{\boldsymbol{\eta}}_L$:

$$\frac{-1}{n}\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_L) + \frac{\lambda_{1,n}}{2n}\boldsymbol{s}_n = \boldsymbol{0} \text{ or } -\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_{\mathrm{L}}) + \frac{\lambda_{1,\mathrm{n}}}{2}\boldsymbol{s}_{\mathrm{n}} = \boldsymbol{0}$$

where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_{i,L})$ if $\hat{\eta}_{i,L} \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 1$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 1$. Note that $\boldsymbol{s}_n = \boldsymbol{s}_{n,\hat{\boldsymbol{\eta}}_L}$ depends on $\hat{\boldsymbol{\eta}}_L$. Thus $\hat{\boldsymbol{\eta}}_L$

$$= (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n.$$

Following Jia and Yu [10], by standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for Equation (1.5), $\hat{\boldsymbol{\eta}}_{EN}$ is optimal if

$$2\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{EN} - 2\boldsymbol{W}^T\boldsymbol{Z} + 2\lambda_1\hat{\boldsymbol{\eta}}_{EN} + \lambda_2\boldsymbol{s}_n = 0, \quad \text{or}$$

$$(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})\hat{\boldsymbol{\eta}}_{EN} = \boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \quad \text{or}$$

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_R - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{3.3}$$

Hence

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_1}{n}\, n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_2}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, \boldsymbol{s}_n$$

$$= \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, [\frac{\lambda_1}{n}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n].$$

Note that if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then $\hat{\lambda}_1/\sqrt{n} \xrightarrow{P} (1-\psi)\tau$ and $\hat{\lambda}_2/\sqrt{n} \xrightarrow{P} 2\psi\tau$. Under these conditions,

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - n(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_1\boldsymbol{I}_{p-1})^{-1}\, [\frac{\hat{\lambda}_1}{\sqrt{n}}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\hat{\lambda}_2}{2\sqrt{n}}\boldsymbol{s}_n].$$

The following theorem shows the elastic net, lasso, and ridge regression are asymptotically equivalent to the OLS full model if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$. The theorem follows from results in Knight and Fu [11] and Slawski, zu Castell, and Tutz [24]. Knight and Fu [11] proved that lasso and ridge regression are consistent estimators of $\boldsymbol{\beta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, and $\sqrt{n}$ consistent if $\lambda_{1,n} = O(\sqrt{n})$ so $\lambda_{1,n}/\sqrt{n}$ is bounded. Let $\hat{\boldsymbol{\eta}}_A$ be $\hat{\boldsymbol{\eta}}_{EN}$, $\hat{\boldsymbol{\eta}}_L$, or $\hat{\boldsymbol{\eta}}_R$. Note that c) follows from b) if $\psi = 0$, and d) follows from b) (using $2\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 2\tau$) if $\psi = 1$. Recall that we are assuming that p is fixed.

Theorem 1. *Assume that the conditions of the OLS theory (3.2) hold for the model* $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\mathbf{0}, \sigma^2 \boldsymbol{V}).$$

b) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$, *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

c) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}).$$

d) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

We can make the three estimators asymptotically equivalent to the OLS full model: take, for example, $\lambda_{1n} = \sqrt{n}/\log(n)$. If $\hat{\lambda}_{1n}/\sqrt{n} \to \tau > 0$, then lasso tends to have at least one $\hat{\beta}_j = 0$ for large n by Ewald and Schneider [4]. Lasso may not be $\sqrt{n}$ consistent if lasso selects S with high probability.

Usually $\hat{\lambda}_{1,n}$ is selected using a criterion such as k–fold CV. It is not clear whether $\hat{\lambda}_{1,n} = o(n)$. For the elastic net and lasso, λ_M/n does not go to zero as $n \to \infty$ since $\hat{\boldsymbol{\eta}} = \mathbf{0}$ is not a consistent estimator. Hence λ_M is likely proportional to n, and using $\lambda_i = i\lambda_M/M$ for $i = 1, ..., M$ will not produce a consistent estimator.

Consider regressing Y on $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ to get $\tilde{\boldsymbol{Y}} = \hat{a} + \hat{b}\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. If $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, then $\hat{a} \xrightarrow{P} 0$ and $\hat{b} \xrightarrow{P} 1$ as $n \to \infty$. Hence the scaled shrinkage estimator is a consistent estimator of $\boldsymbol{\beta}$ if the shrinkage estimator $\hat{\boldsymbol{\beta}}$ is consistent. Note that if $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$, then $\hat{a} = 0$ and $\hat{b} = 1$, since otherwise the scaled shrinkage estimator would have a smaller residual sum of squares than the OLS estimator, which is impossible since OLS minimizes the residual sum of squares. Thus scaling has no effect on relaxed lasso or OLS variable selection.

If $p > n$, the regularity conditions for $\hat{\boldsymbol{\beta}}$ to be a consistent estimator of $\boldsymbol{\beta}$ are much stronger, but results from Hastie, Tibshirani, and Wainwright [7] suggest that lasso can perform well for sparse models: the subset S in (1.6) has a_S small.

CHAPTER 4

SIMULATIONS

For the simulation, ridge regression (RR) and lasso were computed with the `cv.glmnet` function from the `glmnet` library with the *R* software. Let $\boldsymbol{x} = (1\ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the $m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\text{Cov}(\boldsymbol{u}_i) = \boldsymbol{\Sigma_u} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (m-2)\psi^2]$. Hence the correlations are $cor(x_i, x_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k+1$ ones and $p-k-1$ zeros. The zero mean errors e_i were iid from five distributions: i) N(0,1), ii) t_3, iii) EXP(1) - 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). The uniform distribution is the distribution where the shorth undercoverage is maximized by Frey (2013). Distributions ii) and v) have heavy tails, and distribution iii) is not symmetric.

The lengths of the asymptotically optimal 95% PIs are i) 3.92 = 2(1.96), ii) 6.365, iii) 2.996, iv) 1.90 = 2(0.95), and v) 13.490. The simulation used 5000 runs, so an observed coverage in [0.94, 0.96] gives no reason to doubt that the PI has the nominal coverage of 0.95. The simulation used $p = 20, 40, 50, n$, or $2n$; $\psi = 0, 1/\sqrt{p}$, or 0.9; and $k = 1, 19$, or $p - 1$. The OLS full model fails when $p = n$ and $p = 2n$ and regularity conditions for consistent estimators are strong. The values $k = 1$ and $k = 19$ are sparse models where lasso can perform well when n/p is not large. If $k = p - 1$ and $p \geq n$, then the model is dense. When $\psi = 0$, the predictors are uncorrelated, when $\psi = 1/\sqrt{p}$, the correlation goes to 0.5 as p increases and the predictors are moderately correlated. For $\psi = 0.9$, the predictors are highly correlated with 1 dominant principal component.

The simulations were done in *R*. See R Core Team (2016). The results were similar for all five error distributions. Tables 4.1 - 8.1 show some simulation results for PI (2.7) and (2.8) where

lasso and ridge regression minimized 10-fold CV. Ridge regression used the same d that was used for lasso. Table headers lasso is for PI (2.7), vlasso is for PI (2.8), SL is for scaled lasso with PI (2.7), VSL is for scaled lasso with PI (2.8), RR is ridge regression for PI (2.7), VRR is RR for PI (2.8), SRR is for scaled RR with PI (2.7), VSRR is for scaled RR with PI (2.8).

Table 4.1. Simulated Large Sample 95% PI Coverages and Lengths for error type 1 (nruns=5000), $e_i \sim N(0,1)$

n	p	ψ	k		lasso	SL	vlasso	VSL	RR	SRR	VRR	VSRR
100	20	0	1	cov	0.9750	0.9730	0.9632	0.9634	0.9564	0.9512	0.9606	0.9594
				len	4.8245	4.7603	4.7831	4.7423	4.5741	4.4758	5.3277	6.2438
100	40	0	1	cov	0.9774	0.9750	0.9624	0.9596	0.9276	0.8632	0.9614	0.9618
				len	4.8889	4.7876	4.8416	4.7905	4.4260	4.1104	5.7438	9.6068
100	200	0	1	cov	0.9764	0.9722	0.9644	0.9678	0.9578	0.7532	0.9588	0.9592
				len	4.9762	4.7555	4.9884	4.8867	6.1622	3.4142	6.2541	13.9480
100	50	0	49	cov	0.9714	0.9708	0.9606	0.9612	0.9822	0.9770	0.9618	0.9564
				len	6.8345	6.8227	22.3265	22.6899	7.7229	7.2399	27.7275	66.7933
200	20	0	19	cov	0.9766	0.9786	0.9572	0.9574	0.9790	0.9766	0.9548	0.9570
				len	4.9636	4.9612	4.6446	4.6486	5.0454	4.9683	4.7066	4.6495
200	40	0	19	cov	0.9762	0.9780	0.9488	0.9454	0.9742	0.9732	0.9478	0.9516
				len	5.2205	5.1611	5.1065	5.0654	5.2097	5.1209	5.3689	5.3300
200	200	0	19	cov	0.9778	0.9728	0.9534	0.9530	0.9960	0.5440	0.9614	0.9562
				len	5.7714	5.3180	7.0898	6.8564	22.3516	11.8611	16.5520	15.0981
400	20	0.9	19	cov	0.9748	0.9692	0.9584	0.9554	0.9726	0.9646	0.9590	0.9572
				len	10.6086	6.1460	10.1626	5.8390	10.6631	4.3647	9.9861	4.1160
400	40	0.9	19	cov	0.9608	0.9596	0.9530	0.9534	0.9578	0.9640	0.9538	0.9570
				len	14.6702	8.3137	14.5228	7.9291	14.4812	4.4158	14.1356	4.3511
400	400	0.9	19	cov	0.9636	0.9636	0.9546	0.9548	0.9632	0.8786	0.9556	0.9550
				len	47.3608	8.9214	45.4396	8.5698	48.0207	4.6275	44.5228	5.2317
400	400	0	399	cov	0.8508	0.8166	0.9518	0.9536	1.000	0.9988	0.9548	0.9606
				len	37.5418	34.5665	78.0652	81.6053	244.1004	131.3563	69.5812	62.9434
400	800	0.9	19	cov	0.9652	0.9698	0.9584	0.9564	0.9672	0.9274	0.9588	0.9580
				len	67.2939	9.0407	63.7856	8.5959	66.5770	4.6898	63.1034	4.9308

Some R code is below.

```
srrpisim(n=100,p=20,k=1,nruns=5000,psi=0.0,type=1)
$dlas
[1] 4.947
```

```
$dvlas

[1] 5.163

$laspicov

[1] 0.975

$laspimenlen

[1] 4.824475

$slaspicov

[1] 0.973

$slaspimenlen

[1] 4.760299

$vlaspicov

[1] 0.9632

$vlaspimenlen

[1] 4.783059

$vslaspicov

[1] 0.9634

$vslaspimenlen

[1] 4.742325

$rrpicov

[1] 0.9564

$rrpimenlen

[1] 4.57409

$srrpicov

[1] 0.9512

$srrpimenlen

[1] 4.475827

$vrrpicov

[1] 0.9606
```

```
$vrrpimenlen

[1] 5.327717

$vsrrpicov

[1] 0.9594

$vsrrpimenlen

[1] 6.243801
```

CHAPTER 5

ERROR TYPE 2 EXAMPLES

Table 5.1. Simulated Large Sample 95% PI Coverages and Lengths for error type 2 (nruns=5000), $e_i \sim t_3$

n	p	ψ	k		lasso	SL	vlasso	VSL	RR	SRR	VRR	VSRR
100	20	0	1	cov	0.9632	0.9614	0.9578	0.9576	0.9540	0.9148	0.9574	0.9596
				len	8.3460	8.2156	10.0936	10.1526	7.9940	7.6514	10.3417	30.5300
100	40	0	1	cov	0.9658	0.9628	0.9618	0.9652	0.9506	0.7776	0.9620	0.9640
				len	8.4640	8.2446	10.0878	10.2011	7.9295	7.0406	10.4545	38.0839
100	200	0	1	cov	0.9620	0.9566	0.9552	0.9560	0.9572	0.6936	0.9570	0.9576
				len	8.6988	8.1331	10.3071	10.4749	8.9997	5.3480	10.6292	23.1341
100	50	0	49	cov	0.9696	0.9694	0.9560	0.9572	0.9768	0.9744	0.9596	0.9602
				len	11.5426	11.5344	24.8382	25.3699	12.2149	11.8410	28.6124	73.1487
200	20	0	19	cov	0.9720	0.9712	0.9572	0.9578	0.9740	0.9718	0.9548	0.9584
				len	8.7377	8.7347	8.1649	8.1750	8.7863	8.7404	8.1421	8.1812
200	40	0	19	cov	0.9768	0.9752	0.9554	0.9556	0.9732	0.9734	0.9560	0.9548
				len	9.0936	8.9944	8.8251	8.8504	8.9376	8.8697	9.0151	9.2021
200	200	0	19	cov	0.9758	0.9688	0.9598	0.9586	0.9936	0.5230	0.9554	0.9572
				len	9.8875	9.1091	11.7832	11.7247	23.6466	13.4114	17.5470	16.4468
400	20	0.9	19	cov	0.9624	0.9616	0.9522	0.9516	0.9668	0.9596	0.9524	0.9552
				len	10.7252	8.2400	10.4784	8.0216	10.7629	7.3461	10.3135	7.1320
400	40	0.9	19	cov	0.9616	0.9616	0.9540	0.9570	0.9604	0.9590	0.9540	0.9546
				len	14.8075	9.7732	14.5875	9.4978	14.9966	7.0746	14.6606	7.1897
400	400	0.9	19	cov	0.9594	0.9630	0.9534	0.9574	0.9634	0.9386	0.9548	0.9560
				len	47.8514	10.4872	45.7736	10.1547	48.5160	7.3242	44.9570	7.7311
400	400	0	399	cov	0.8520	0.8154	0.9532	0.9516	1.000	0.9980	0.9524	0.9524
				len	38.1747	35.2417	78.2062	81.7967	243.5236	132.0560	69.8679	63.2659
400	800	0.9	19	cov	0.9654	0.9634	0.9514	0.9536	0.9652	0.9478	0.9536	0.9514
				len	67.7355	10.6404	64.0052	10.1836	66.9951	7.4740	63.3265	7.5353

CHAPTER 6

ERROR TYPE 3 EXAMPLES

Table 6.1. Simulated Large Sample 95% PI Coverages and Lengths for error type 3 (nruns=5000), $e_i \sim EXP(1) - 1$

n	p	ψ	k		lasso	SL	vlasso	VSL	RR	SRR	VRR	VSRR
100	20	0	1	cov	0.9728	0.9706	0.9582	0.9596	0.9546	0.9444	0.9612	0.9550
				len	4.4345	4.3082	5.0089	4.9130	4.4384	4.3619	5.6692	6.9304
100	40	0	1	cov	0.9750	0.9750	0.9586	0.9580	0.9374	0.8664	0.9598	0.9622
				len	4.5535	4.3831	5.0908	4.9986	4.4035	4.1185	6.1098	11.6162
100	200	0	1	cov	0.9736	0.9740	0.9560	0.9582	0.9574	0.7684	0.9594	0.9584
				len	4.7104	4.4060	5.2616	5.1164	6.2218	3.4469	6.6069	13.8421
100	50	0	49	cov	0.9716	0.9706	0.9618	0.9616	0.9814	0.9722	0.9608	0.9646
				len	6.9460	6.9326	22.4097	22.7736	7.8316	7.3600	27.8306	67.1252
200	20	0	19	cov	0.9780	0.9776	0.9592	0.9600	0.9786	0.9776	0.9598	0.9610
				len	4.7186	4.7174	4.6171	4.6211	4.8407	4.7255	4.7052	4.6243
200	40	0	19	cov	0.9784	0.9776	0.9560	0.9560	0.9744	0.9738	0.9582	0.9588
				len	5.0942	5.0210	5.1472	5.1013	5.1455	5.0467	5.4365	5.3922
200	200	0	19	cov	0.9734	0.9726	0.9510	0.9522	0.9930	0.5450	0.9550	0.9574
				len	5.7836	5.2834	7.1394	6.9027	22.3106	11.8392	16.5806	15.0300
400	20	0.9	19	cov	0.9704	0.9658	0.9572	0.9560	0.9694	0.9372	0.9548	0.9606
				len	10.7134	6.1668	10.2824	5.8881	10.7144	3.6054	10.1098	3.5926
400	40	0.9	19	cov	0.9654	0.9630	0.9538	0.9568	0.9622	0.9418	0.9522	0.9540
				len	14.7387	8.3503	14.6056	7.9963	14.6616	4.0356	14.3988	4.1625
400	400	0.9	19	cov	0.9660	0.9632	0.9588	0.9556	0.9658	0.8572	0.9576	0.9592
				len	47.3841	8.9392	45.5246	8.6300	48.0632	4.3641	44.5903	5.1930
400	400	0	399	cov	0.8446	0.8062	0.9586	0.9570	1.000	0.9996	0.9558	0.9560
				len	37.5185	34.5573	78.0564	81.6033	243.7929	131.4923	69.5474	62.8434
400	800	0.9	19	cov	0.9682	0.9674	0.9582	0.9544	0.9656	0.9162	0.9548	0.9580
				len	67.2399	9.0631	63.7545	8.6423	66.4799	4.4452	63.0266	4.8751

CHAPTER 7

ERROR TYPE 4 EXAMPLES

Table 7.1. Simulated Large Sample 95% PI Coverages and Lengths for error type 4 (nruns=5000), $e_i \sim uniform(-1, 1)$

n	p	ψ	k		lasso	SL	vlasso	VSL	RR	SRR	VRR	VSRR
100	20	0	1	cov	0.9916	0.9944	0.9598	0.9616	0.9472	0.9446	0.9610	0.9612
				len	2.3751	2.3152	2.2886	2.1934	2.3774	2.3502	2.9692	3.0508
100	40	0	1	cov	0.9904	0.9926	0.9610	0.9640	0.8934	0.8712	0.9592	0.9578
				len	2.4176	2.3387	2.3673	2.2456	2.3064	2.2337	3.6998	3.9876
100	200	0	1	cov	0.9864	0.9874	0.9604	0.9566	0.9650	0.7702	0.9588	0.9570
				len	2.4945	2.3506	2.5004	2.3102	5.0174	2.5686	4.9370	6.8526
100	50	0	49	cov	0.9786	0.9790	0.9556	0.9574	0.9824	0.9796	0.9536	0.9582
				len	3.8554	3.8403	21.2821	21.6118	5.2428	4.5342	27.3574	64.3802
200	20	0	19	cov	0.9856	0.9870	0.9550	0.9544	0.9832	0.9858	0.9570	0.9544
				len	2.4703	2.4643	2.4170	2.4162	2.6855	2.4826	2.6499	2.4460
200	40	0	19	cov	0.9870	0.9812	0.9528	0.9496	0.9814	0.9798	0.9532	0.9566
				len	2.6805	2.6324	2.7691	2.7224	2.8913	2.6985	3.2480	3.0434
200	200	0	19	cov	0.9806	0.9754	0.9562	0.9530	0.9942	0.5700	0.9548	0.9560
				len	3.1177	2.8209	4.0109	3.8002	21.8417	11.2607	16.1584	14.4926
400	20	0.9	19	cov	0.9668	0.9660	0.9486	0.9538	0.9668	0.9308	0.9490	0.9498
				len	10.8020	5.0925	10.0931	4.8031	10.9788	2.0595	10.2298	1.9778
400	40	0.9	19	cov	0.9672	0.9600	0.9572	0.9522	0.9642	0.9308	0.9532	0.9532
				len	15.1260	7.6269	14.5715	7.2321	15.3211	2.5657	14.8317	2.5335
400	400	0.9	19	cov	0.9616	0.9636	0.9524	0.9540	0.9622	0.7530	0.9512	0.9532
				len	47.1768	8.2101	45.3406	7.9058	47.8504	2.9071	44.4210	3.9334
400	400	0	399	cov	0.8478	0.8114	0.9502	0.9498	1.000	0.9984	0.9532	0.9522
				len	37.2128	34.2539	78.0147	81.4731	244.8390	131.6709	69.6196	62.8650
400	800	0.9	19	cov	0.9608	0.9664	0.9500	0.9554	0.9630	0.8516	0.9480	0.9540
				len	67.0137	8.3055	63.7723	7.9172	66.2645	2.9511	63.0501	3.5062

CHAPTER 8

ERROR TYPE 5 EXAMPLES

Table 8.1. Simulated Large Sample 95% PI Coverages and Lengths for error type 5(nruns=5000), $e_i \sim 0.9N(0,1) + 0.1N(0,100)$

n	p	ψ	k		lasso	SL	vlasso	VSL	RR	SRR	VRR	VSRR
100	20	0	1	cov	0.9560	0.9572	0.9620	0.9612	0.9562	0.7412	0.9592	0.9616
				len	17.3998	16.9038	23.0477	23.2749	17.1965	15.7771	23.0808	70.5580
100	40	0	1	cov	0.9482	0.9482	0.9586	0.9600	0.9478	0.5964	0.9584	0.9608
				len	17.7428	16.8885	23.0184	23.3237	17.4268	14.5801	23.0644	61.8767
100	200	0	1	cov	0.9524	0.9444	0.9562	0.9590	0.9490	0.6018	0.9572	0.9574
				len	17.7586	15.9205	23.2490	23.6644	17.7365	10.9210	23.1955	35.8158
100	50	0	49	cov	0.9658	0.9654	0.9614	0.9604	0.9732	0.9678	0.9622	0.9592
				len	24.3794	23.8435	33.3387	34.3426	26.6367	24.9746	33.7416	79.8000
200	20	0	19	cov	0.9620	0.9622	0.9498	0.9500	0.9612	0.9620	0.9516	0.9496
				len	20.6508	20.6380	18.7335	18.7802	20.7582	20.6607	18.6119	18.8466
200	40	0	19	cov	0.9654	0.9660	0.9570	0.9598	0.9644	0.9622	0.9576	0.9576
				len	21.2357	20.8590	19.6860	19.9495	21.0476	20.6025	19.4958	20.5310
200	200	0	19	cov	0.9694	0.9600	0.9580	0.9582	0.9826	0.5510	0.9556	0.9582
				len	21.9094	19.2936	22.0987	22.7933	30.0395	19.0683	22.9558	25.5007
400	20	0.9	19	cov	0.9556	0.9574	0.9552	0.9540	0.9570	0.9568	0.9550	0.9552
				len	16.3836	16.2787	16.7407	16.5939	16.4058	16.3116	16.4657	16.3143
400	40	0.9	19	cov	0.9532	0.9538	0.9568	0.9546	0.9574	0.9520	0.9570	0.9560
				len	16.4705	15.8309	17.7666	17.0164	16.3447	15.0483	17.5700	16.4268
400	400	0.9	19	cov	0.9638	0.9632	0.9522	0.9588	0.9648	0.9542	0.9520	0.9580
				len	49.4316	16.7885	45.6336	17.3154	48.2310	15.7169	45.7570	16.4345
400	400	0	399	cov	0.8502	0.8140	0.9482	0.9472	1.000	0.9966	0.9554	0.9520
				len	41.2273	38.1620	79.5473	83.3671	238.8424	132.2243	70.9232	64.6938
400	800	0.9	19	cov	0.9646	0.9604	0.9522	0.9604	0.9648	0.9596	0.9510	0.9584
				len	68.5556	17.3883	64.7917	17.4152	68.2060	16.3547	64.4588	16.4773

CHAPTER 9

CONCLUSIONS

Sometimes scaling resulted in PIs that were too short so there was undercoverage. Scaling with validation residuals was a useful technique.

The simulations were done in *R*. See R Core Team [22]. The collection of Olive [19] *R* functions *slpack*, available from `http://lagrange.math.siu.edu/Olive/slpack.txt`, has some useful functions for the inference. The tables were made with the function `srrpisim`.

For lasso and ridge regression, 10-fold CV produced good PIs if $\psi = 0$ or if k was small, but if both $k \geq 19$ and $\psi \geq 0.5$, then 10-fold CV tended to shrink too much and the PI lengths were often too long. Pelawa Watagoda and Olive (2019) noted that lasso did appear to select $S \subseteq I_{min}$ for sparse models since relaxed lasso was good in their simulation.

For n/p not large, good performance needed stronger regularity conditions. If there was $k = 1$ active population predictor, then lasso often performed well. For $k = 19$, lasso often performed well for $\psi = 0$. For dense models with $k = p - 1$ and n/p not large, there was often undercoverage. Let $d - 1$ be the number of active predictors in the selected model. For $N(0, 1)$ errors, $\psi = 0$, and $d < k$, an asymptotic population 95% PI has length $3.92\sqrt{k - d + 1}$. Note that when the $(Y_i, \boldsymbol{u}_i^T)^T$ follow a multivariate normal distribution, every subset follows a multiple linear regression model. PI (2.8) often had good coverage.

From the 5 simulation tables, the results are similar. For the first 7 lines in every table, Scaling did not have much effect. And the lasso often did better than RR.

For the rest data of every table, when n=400, the scaled lasso and scaled RR lengths are much better than lasso and RR, respectively, but the coverage is often too low. PI (2.8) has good coverage, but the PI length was too long if $1.5k \leq n \leq 3k$. PI (2.8) was better for ridge regression than PI (2.7) for $k = 399$ and $n = 400$.

REFERENCES

[1] Cai, T., Tian, L., Solomon, S.D., and Wei, L.J. (2008), "Predicting Future Responses Based on Possibly Misspecified Working Models," *Biometrika*, 95, 75-92.

[2] Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference*,Cambridge University Press, New York, NY.

[3] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," (with discussion), *The Annals of Statistics,* 32, 407-451.

[4] Ewald, K., and Schneider, U. (2018), "Uniformly Valid Confidence Sets Based on the Lasso," *Electronic Journal of Statistics*, 12, 1358-1387.

[5] Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference,* 143, 1039-1048.

[6] Gunst, R.F., and Mason, R.L. (1980), *Regression Analysis and Its Application,* Marcel Dekker, New York, NY.

[7] Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, CRC Press Taylor & Francis, Boca Raton, FL.

[8] Hoerl, A.E., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

[9] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning With Applications in R*, Springer, New York, NY.

[10] Jia, J., and Yu, B. (2010), "On Model Selection Consistency of the Elastic Net When $p >> n$," *Statistica Sinica*, 20, 595-611.

[11] Knight, K., and Fu, W.J. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

[12] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094-1111.

[13] Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics & Data Analysis*, 52,

374-393.

[14] Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics & Data Analysis,* 51, 3115-3122.

[15] Olive, D.J. (2013), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

[16] Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

[17] Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

[18] Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, 59, 913-931.

[19] Olive, D.J. (2019), *Prediction and Statistical Learning*, online course notes, see (http://lagrange.math.siu.edu/Olive/slearnbk.htm).

[20] Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

[21] Pelawa Watagoda, L.C.R., and Olive, D.J. (2019), "Comparing Shrinkage Estimators With Asymptotically Optimal Prediction Intervals," preprint at (http://lagrange.math.siu.edu/Olive/pppicomp.pdf).

[22] R Core Team (2016), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

[23] Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.

[24] Slawski, M., zu Castell, W., and Tutz, G., (2010), "Feature Selection Guided by Structural Information," *The Annals of Applied Statistics,* 4, 1056-1080.

[25] Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

[26] Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

[27] 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

Graduate School
Southern Illinois University

Lingling Zhang

linglingzhang95@outlook.com

Shandong Normal University
Bachelor of Science, Mathematics, July, 2017

Research Paper Title:
Prediction Intervals For Scaled Shrinkage Estimators

Major Professor: Dr. David J. Olive