Noname manuscript No.
(will be inserted by the editor)

Comparing Six Shrinkage Estimators With Large Sample Theory and Asymptotically Optimal Prediction Intervals

Lasanthi C.R. Pelawa Watagoda · David J. Olive

Received: date / Accepted: date

Abstract 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. This paper compares the six shrinkage estimators: forward selection, lasso, partial least squares, principal components regression, lasso variable selection, and ridge regression, with large sample theory and two new prediction intervals that are asymptotically optimal if the estimator $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. Few prediction intervals have been developed for $p > n$, and they are not asymptotically optimal.

For p fixed, the large sample theory for variable selection estimators like forward selection is new, and the theory shows that lasso variable selection is $\sqrt{n}$ consistent under much milder conditions than lasso. This paper also simplifies the proofs of the large sample theory for lasso, ridge regression, and elastic net.

Keywords Forward Selection · Lasso · Partial Least Squares · Principal Components Regression · Ridge Regression

1 Introduction

This section first reviews six multiple linear regression (MLR) estimators, and then reviews asymptotically optimal prediction intervals. Suppose that the response variable Y_i and at least one predictor variable $x_{i,j}$ are quantitative

Lasanthi C.R. Pelawa Watagoda
Department of Mathematical Sciences, Appalachian State University, Boone, NC 28608-2092, USA.
E-mail: lasanthi@appstate.edu

David J. Olive
Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA.
E-mail: dolive@siu.edu

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 model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1}$$

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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2}$$

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 many regression estimators, a method is needed so that everyone who uses the same units of measurements for the predictors and Y gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. A common method is to 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

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{3}$$

where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$, and $\hat{\boldsymbol{\beta}}$ is found from $\hat{\boldsymbol{\eta}}$.

There are many methods for estimating $\boldsymbol{\beta}$, including forward selection with OLS, principal components regression (PCR), partial least squares (PLS) due to Wold (1975), lasso due to Tibshirani (1996), and ridge regression (RR): see Hoerl and Kennard (1970). Some shrinkage methods do variable selection: apply OLS to the predictors that had nonzero coefficients. These methods include least angle regression, lasso, relaxed lasso, and elastic net. See, for example, Fan and Li (2001), Hastie, Tibshirani, and Wainwright (2015, ch. 5), Sun and Zhang (2012), Tibshirani (1996), and Zou and Hastie (2005). The Meinshausen (2007) relaxed lasso estimator for multiple linear regression fits lasso with penalty λ_n to get a subset of variables with nonzero coefficients, and then fits lasso with a smaller penalty ϕ_n where n is the sample size. A three stage procedure uses this relaxed lasso estimator for variable selection.

These variable selection estimators have had several names in the literature. Least angle regression variable selection is the LARS-OLS hybrid estimator of Efron et al. (2004, p. 421), lasso variable selection is called relaxed lasso by Hastie, Tibshirani, and Wainwright (2015, p. 12), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007, p. 376).

Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \lambda_{1,n}\sum_{i=1}^{p-1} |\eta_i|^j \tag{4}$$

where $\lambda_{1,n} \geq 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression, and $j = 1$ corresponds to lasso. In the literature, $Q(\boldsymbol{\eta})/c$ is often used, where $c = 2, n$, or $2n$ are common. 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 $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$.

These six methods produce M models and use a criterion to select the final model (e.g., C_p or 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. (2013, ch. 6) and Hastie, Tibshirani, and Wainwright (2015, p. 24).

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 (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{5}$$

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 (5) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{I/S} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{6}$$

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 $S \subseteq I$.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to β_1 is always in the model, and $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, 0)^T$. Then the $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that contain 1 are $I_1 = \{1\}$, $S = I_2 = \{1, 2\}$, $I_3 = \{1, 3\}$, $I_4 = \{1, 4\}$, $I_5 = \{1, 2, 3\}$, $I_6 = \{1, 2, 4\}$, $I_7 = \{1, 3, 4\}$, and $I_8 = \{1, 2, 3, 4\}$. There are $2^{p-a_S} = 4$ subsets I_2, I_5, I_6, and I_8 such that $S \subseteq I_j$. Also, $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$ is obtained by regressing Y on $\boldsymbol{x}_{I_7} = (x_1, x_3, x_4)^T$.

Forward selection forms a sequence of submodels $I_1, ..., I_M$ where I_j uses j predictors $x_1^*, ..., x_{j-1}^*, x_j^*$ including the constant $x_1^* = x_1$. For $j > 1$, the variable x_j^* is the variable not in I_{j-1} that reduces the residual sum of squares the most. Often $M = \min(\lceil n/J \rceil, p)$ for some integer J such as $J = 5, 10$, or 20. Here $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$.

Consider the six methods forward selection with OLS, PCR, PLS, lasso, lasso variable selection, and ridge regression. When there is a sequence of M submodels, the final submodel I_d needs to be selected. Let the candidate model I contain a terms, including a constant. For example, let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors for the methods excluding PCR and PLS. Then there are many

criteria used to select the final submodel I_d with a_d terms. For a given data set, the quantities p, n, and $\hat{\sigma}^2$ act as constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset I_{min} that minimizes the criterion.

Let criteria $C_S(I)$ have the form

$$C_S(I) = SSE(I) + aK_n\hat{\sigma}^2.$$

These criteria need a good estimator of σ^2 and n/p large. The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. Typically $\hat{\sigma}^2$ is the OLS full model

$$MSE = \sum_{i=1}^{n} \frac{r_i^2}{n-p}$$

when n/p is large. See Jones (1946) and Mallows (1973) for C_p.

The following criteria also need n/p large. AIC is due to Akaike (1973) and BIC to Schwarz (1978).

$$AIC(I) = n\log\left(\frac{SSE(I)}{n}\right) + 2a, \quad \text{and}$$

$$BIC(I) = n\log\left(\frac{SSE(I)}{n}\right) + a\log(n).$$

Forward selection with C_p and AIC often gives useful results if $n \geq 5p$ and if the final model has $n \geq 10a_d$. For $p < n < 5p$, forward selection with C_p and AIC tends to pick the full model (which overfits since $n < 5p$) too often, especially if $\hat{\sigma}^2 = MSE$. The Hurvich and Tsai (1989) AIC_C criterion can be useful if $n \geq \max(2p, 10a_d)$.

The EBIC criterion given in Luo and Chen (2013) may be useful when n/p is not large. Let $0 \leq \gamma \leq 1$ and $|I| = a \leq \min(n, p)$ if $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$. We may use $a \leq \min(n/5, p)$. Then

$$EBIC(I) = n\log\left(\frac{SSE(I)}{n}\right) + a\log(n) + 2\gamma\log\left[\binom{p}{a}\right] = BIC(I) + 2\gamma\log\left[\binom{p}{a}\right].$$

This criterion can give good results if $p = p_n = O(n^k)$ and $\gamma > 1 - 1/(2k)$. Hence we will use $\gamma = 1$.

The above criteria can be applied to forward selection and lasso variable selection. The C_p criterion can also be applied to lasso. See Efron and Hastie (2016, pp. 221, 231).

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

the mass. (A highest density region is a union of intervals such that the sum of the lengths is minimized given at least $100(1-\delta)\%$ coverage. For a unimodal error distribution with a probability density function, the population shorth is the population highest density region. See Hyndman (1996) for more about highest density regions.)

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 the shorth(c) estimator. Frey (2013) showed that for large $n\delta$ and iid data, the shorth($k_n = \lceil n(1-\delta) \rceil$) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the large sample $100(1-\delta)\%$ PI shorth(c) =

$$[Z_{(s)}, Z_{(s+c-1)}] \text{ with } \mathrm{c} = \min(\mathrm{n}, \lceil \mathrm{n}[1 - \delta + 1.12\sqrt{\delta/\mathrm{n}}\ ]\ \rceil). \tag{7}$$

Section 2 will develop two prediction intervals that are useful after model or variable selection where n/p need not be large. Section 3 reviews large sample theory for the estimators with some new results, and Section 4 gives a simulation.

2 Prediction Intervals After Model Selection

This section derives asymptotically optimal prediction intervals for the additive error regression model, $Y = m(\boldsymbol{x}) + e$, that can be useful after model selection. Here $m(\boldsymbol{x})$ is a real valued function and the e_i are iid, often with zero mean and constant variance $V(e) = \sigma^2$. The large sample theory for prediction intervals is simple for this model. Emphasis will be on the multiple linear regression model (1) which is a special case with $m(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. Cai et al. (2008) 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 $P(Y_f \in [\hat{Y}_f + \hat{L}_n(r), \hat{Y}_f + \hat{U}_n(r)] \to P(Y_f \in [m(\boldsymbol{x}_f) + L, m(\boldsymbol{x}_f) + U]) = P(e \in [L, U]) = 1 - \delta$ as $n \to \infty$. Three common choices are a) $P(e \le U) = 1 - \delta/2$ and $P(e \le L) = \delta/2$, b) $P(e^2 \le U^2) = P(|e| \le U) = P(-U \le e \le U) = 1 - \delta$ with $L = -U$, and c) the population shorth is the shortest interval (with length $U - L$) 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. The split conformal prediction interval (13), described below, estimates $[-U, U]$ in b).

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. With the exception of ridge regression, let d be the number of "variables" used by the method. Forward selection, lasso, and lasso variable selection use variables $x_1^*, ..., x_d^*$ while PCR and PLS use variables that are

linear combinations of the predictors $V_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ for $j = 1, ..., d$. For PCR, the variables are the principal components of the (standardized) predictors. (We could let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used. See Jansen, Fithian, and Hastie (2015).) See Hong et al. (2018) for why classical prediction intervals after variable selection fail to work.

For n/p large and $d = p$, Olive (2013) developed prediction intervals for models of the form $Y_i = m(\boldsymbol{x}_i) + e_i$, and variable selection models for (1) have this form, as noted by Olive (2018). The first new PI, that can be useful even if n/p is not large, is defined below. This PI modifies the Olive (2013) PI that can only be computed if $n > p$. Olive (2007, 2017a, 2017b, 2018) 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 (2012). For a wide range of regression models, extrapolation occurs if the leverage $h_f = \boldsymbol{x}_{I,f}^T (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{x}_{I,f} > 2d/n$: 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 (9) is changed to $q_n = 1 - \delta$.

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

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.} \tag{8}$$

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

$$c = \lceil n q_n \rceil, \tag{9}$$

and let

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \tag{10}$$

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 new 100 $(1-\delta)$% large sample PI for Y_f is

$$[\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}]. \quad (11)$$

The second new 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 (7) by $n_V = n - n_H$). Then the second new $100(1-\delta)$% large sample PI for Y_f is

$$[\hat{m}_H(\boldsymbol{x}_f) + v_{(s)}, \hat{m}_H(\boldsymbol{x}_f) + v_{(s+c-1)}]. \quad (12)$$

We can also motivate PI (12) by modifying the justification for the Lei et al. (2018) split conformal prediction interval

$$[\hat{m}_H(\boldsymbol{x}_f) - a_q, \hat{m}_H(\boldsymbol{x}_f) + a_q] \quad (13)$$

where a_q is the $100(1-\delta)$th quantile of the absolute validation residuals. Also see Lei (2019). PI (12) 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 (11) to (13) 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 new PIs (11) and (12) are asymptotically optimal for a large class of error distributions while the split

conformal PI (13) 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. For example EBIC and lasso can have $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then forward selection with EBIC and lasso variable selection can produce consistent estimators. PLS can be $\sqrt{n}$ consistent. See the second to last paragraph of Section 3 for references.

None of the three prediction intervals (11), (12), and (13) dominates the other two. If a good fitting method, such as forward selection with EBIC or lasso, is used, and $1.5a_S \leq n \leq 5a_S$, then PI (11) can be much shorter than PIs (12) and (13). For n/d large, PIs (11) and (12) can be shorter than PI (13) if the error distribution is not unimodal and symmetric; however, PI (13) is often shorter if n/d is not large since the sample shorth converges to the population shorth rather slowly. Grübel (1988) shows that for iid data, the length and center of the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval. For a unimodal and symmetric error distribution, the three PIs are asymptotically equivalent, but PI (13) can be the shortest PI due to different correction factors.

If the estimator is poor, the split conformal PI (13) and PI (12) can have coverage closer to the nominal coverage than PI (11). 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 (12) will be long compared to PI (11). For a good fitting model, residuals r_i tend to be smaller in magnitude than errors e_i. Hence PI (11) needs a complicated correction factor. The validation residuals v_j tend to be larger in magnitude than the e_i, and thus the Frey correction factor can be used for PI (12).

Asymptotically optimal PIs estimate the population shorth of the zero mean error distribution. Hence PIs that use the shorth of the residuals, such as PIs (11) and (12), are the only easily computed asymptotically optimal PIs for a wide range of consistent estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for the multiple linear regression model (1). Other asymptotically optimal PIs need p fixed or only estimate the population shorth for unimodal symmetric zero mean error distributions. For example, the Lei et al. (2018) split conformal prediction interval (13) needs the latter distributions for asymptotic optimality. If the error distribution is $e \sim EXP(1) - 1$, then the asymptotic length of the 95% PI (11) or (12) is 2.966 while that of the split conformal PI is $2(1.966) = 3.992$. The Olive (2007, 2013) asymptotically optimal PIs need n/p large.

3 Large Sample Theory

The prediction intervals (11) and (12) are asymptotically optimal if $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. The six estimators forward selection with OLS, principal components regression, partial least squares, lasso, ridge regression, and lasso variable selection have R programs and large sample theory related

to that of OLS. This section gives new theory for variable selection estimators such as forward selection and lasso variable selection, and simplifies the proofs for lasso and ridge regression. First we will let p be fixed.

Next, we review large sample theory for lasso, ridge regression, and the elastic net, defined below. Knight and Fu (2000) and Slawski, zu Castell, and Tutz (2010) proved that lasso, ridge regression, and the elastic net are asymptotically equivalent to the OLS full model if $\lambda_{1,n}/\sqrt{n} \xrightarrow{P} 0$. Knight and Fu (2000) 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. Pilz (2020) also has some theory for the elastic net.

The following results are used to give simpler proofs, and to make comparisons of the three estimators and the full model OLS estimator simpler. Since model selection with $\lambda_1, ..., \lambda_M$ is used, we need $\hat{\lambda}_{1,n}$ to be well behaved.

Assume that the sample correlation matrix

$$\boldsymbol{R_u} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{14}$$

Note that $\boldsymbol{V}^{-1} = \boldsymbol{\rho_u}$, the population correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$, if the $\boldsymbol{u}_i$ are a random sample from a population. Under (14), if $\lambda_{1,n}/n \to 0$ then

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

Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,...,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Then from Sen and Singer (1993, p. 280), the OLS estimator satisfies

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{15}$$

The following identity from Gunst and Mason (1980, p. 342) 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 (2016, p. 308), 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 the ith element of s_n is $s_{in} \in [-1, 1]$ 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 Hastie, Tibshirani, and Wainwright (2015, p. 57), the elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{16}$$

where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ with $0 \leq \alpha \leq 1$. Also see Zou and Hastie (2005).

Following Jia and Yu (2010), by standard Karush-Kuhn-Tucker conditions for convex optimality for Equation (16), $\hat{\boldsymbol{\eta}}_{EN}$ is optimal if

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

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

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_R - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{17}$$

Hence

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

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

Thus elastic net is consistent if $\lambda_{1,n}/n \to 0$ as $n \to \infty$. 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 $\hat{\lambda}_2/\sqrt{n} \xrightarrow{P} 2\psi\tau$. Also note that

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

The following theorem, summarizing results from Knight and Fu (2000) and Slawski, zu Castell, and Tutz (2010), shows that elastic net, lasso, and ridge regression are asymptotically equivalent to the OLS full model if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$. 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 (15) hold for the model* $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

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

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 six estimators asymptotically equivalent to the OLS full model: take, for example, $\lambda_{1n} = \sqrt{n}/\log(n)$ for lasso and lasso variable selection, and make $P(d = p) \to 1$ for forward selection, PCR, PLS, and lasso variable selection. Lasso and elastic net do variable selection better if $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$ as $n \to \infty$ than if $\lambda_n/\sqrt{n} \to 0$, as noted by Fan and Li (2001). PCR tends to be inconsistent if $P(d = p)$ does not go to one.

Usually $\hat{\lambda}_{1,n}$ is selected using a criterion such as k–fold CV or GCV. 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 $\hat{\boldsymbol{\eta}} = \boldsymbol{0}$ 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.

Next, we give large sample theory for OLS variable selection estimators such as forward selection and lasso variable selection. Suppose that model (5) holds. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume the maximum leverage

$$\max_{i=1,...,n} \boldsymbol{x}_{iI_j}^T(\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{x}_{iI_j} \to 0$$

in probability as $n \to \infty$ for each I_j with $S \subseteq I_j$ where the dimension of I_j is a_j. For the OLS model with $S \subseteq I_j$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \sigma^2\boldsymbol{V}_j)$ where $(\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})/n \xrightarrow{P} \boldsymbol{V}_j^{-1}$. See, for example, Sen and Singer (1993, p. 280). Then

$$\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}_{j,0}) \tag{18}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. Also use zero padding for the model I_{min}. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets. For example, if each subset contains at least one variable, then there are $J = 2^p - 1$ subsets.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $W = W_{VS} = k$ if $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ where $P(W_{VS} = k) = \pi_{kn}$ for $k = 1, ..., J$. Then $(\hat{\boldsymbol{\beta}}_{VS:n}, W_{VS:n}) = (\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$ has a joint distribution where the sample size n is usually suppressed. Note that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. Define $P(A|B_k)P(B_k) = 0$ if $P(B_k) = 0$. Let $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(W_{VS} = k)$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(W_{VS} = k) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \le t_1, ..., z_p \le t_p)$ by $P(\boldsymbol{z} \le \boldsymbol{t})$. Then

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]\pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}]\pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities π_{kn}, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}.

Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the MLE with AIC. Here $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\boldsymbol{0}$. Hence $E(\boldsymbol{w}_j) = 0$, and $\text{Cov}(\boldsymbol{w}_j) = \boldsymbol{\Sigma}_j$ exits. Pelawa Watagoda and Olive (2019) defined $\hat{\boldsymbol{\beta}}_{MIX}$ to be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities equal to π_{kn}. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with same probabilities π_{kn} of the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the I_k are randomly selected. Note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$ of Equation (18). The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$ does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ to that of $\boldsymbol{w}_{jn}$. Similarly, selection bias does change the distribution of the selected $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_j$. Let $W = W_{VS,\infty}$ where $P(W = k) = \pi_k$.

The first assumption in Theorem 2 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the variable selection estimator corresponding to I_{min} underfits with probability going to zero, and the assumption holds under regularity conditions if BIC or AIC is used. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). For multiple linear regression with Mallows (1973) C_p or AIC, see Li (1987), Nishii (1984), and Shao (1993). Let $\hat{\boldsymbol{\beta}}_{I_{min}}$ be the OLS estimator applied to a constant and the variables with nonzero

shrinkage estimator coefficients. If the shrinkage estimator is a consistent estimator of $\boldsymbol{\beta}$, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. See Zhao and Yu (2006, p. 2554). The reasonable Theorem 2 assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild. Rathnayake and Olive (2020) extend this theory to many other variable selection estimators such as generalized linear models. Charkhi and Claeskens (2018) have a related result for forward selection with AIC when the iid errors are $N(0, \sigma^2)$.

Theorem 2 *Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then*

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{19}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$. Thus $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

PROOF. a) Since $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{w}_n$ is $F_{\boldsymbol{w}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{z})$ at continuity points of the $F_{\boldsymbol{w}_j}(\boldsymbol{t})$ as $n \to \infty$. □

If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959). By both this result and Theorem 2, the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent.

We expect that prediction intervals will often work better for smaller sample sizes for estimators with better large sample theory. For fixed p, the forward selection with C_p and lasso variable selection estimators $\hat{\boldsymbol{\beta}}_{I_{min},0}$ are $\sqrt{n}$ consistent by Theorem 2, and PLS is $\sqrt{n}$ consistent by Cook and Forzani (2018, 2019). Note that $\hat{\boldsymbol{\beta}}_{I_{min}}$ and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ give the same prediction intervals (11) to (13). Lasso and ridge regression have the next best large sample theory, and PCR has the worst. PCR may give good prediction if the column space of $\boldsymbol{X}$ is approximately the column space of $\boldsymbol{W}_d$, where the columns of $\boldsymbol{W}_d$ are the d PCR components used by the PCR estimator. The two column spaces are equal if $d = p$. Even if $p > n$, lasso variable selection may outperform lasso if $\boldsymbol{X}^T_{I_{min}} \boldsymbol{X}_{I_{min}}$ is not ill conditioned, where I_{min} corresponds to the variables with nonzero lasso coefficients, including a constant. Lasso may outperform lasso variable selection if $\boldsymbol{X}^T_{I_{min}} \boldsymbol{X}_{I_{min}}$ is ill conditioned. Belloni and Chernozhukov (2013) suggest lasso variable selection can outperform lasso.

If $p > n$, the regularity conditions for $\hat{\boldsymbol{\beta}}$ to be a consistent estimator of $\boldsymbol{\beta}$ are much stronger, and a simplifying structure is usually needed. One widely used structure is that the model is sparse: the subset S in (5) has a_S small. Results from Hastie, Tibshirani, and Wainwright (2015, pp. 20, 296, ch. 6, ch. 11) and Luo and Chen (2013) suggest that lasso, lasso variable selection, and forward selection with EBIC can perform well for sparse models, especially if the nontrivial predictors are nearly orthogonal or nearly uncorrelated. A

second widely used structure is $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$. Then $\boldsymbol{Y} = \boldsymbol{X}_I \boldsymbol{\beta}_I + \boldsymbol{e}_I$ follows a multiple linear regression model for every subset I. Some models have smaller σ_I^2, and $n(\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}$ should not be ill conditioned for forward selection and lasso variable selection. If $p > n$, under two sets of strong regularity conditions, PLS can be $\sqrt{n}$ consistent or inconsistent. See Chun and Keleş (2010), Cook (2018), Cook and Forzani (2018, 2019), and Cook, Helland, and Su (2013).

The simulations have parameters ψ and k, where ψ determines the correlation of the k nontrivial predictors and $a_S = k+1$. For n/p not large, the model will be sparse if $k = 1$ or $k = 19$. The nontrivial predictors are independent and hence uncorrelated if $\psi = 0$. In the simulation, $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$ when the errors are iid $N(0, 1)$. A third simplifying structure occurs when $\psi = 0.9$. Then all of the nontrivial predictors are very highly correlated with the line in the direction of $(1, ..., 1)^T$, and the first PCR component or a small group of predictors may be useful for prediction.

4 Examples and Simulations

Suppose that $n \geq 10d$ where model I has d predictors, including a constant. Response plots of the fitted values $\hat{Y}$ versus the response Y are useful for checking linearity, checking whether the error distribution is skewed, and for detecting outliers. See Brillinger (1977, 1983) and Cook and Weisberg (1999, pp. 417, 425, 432). Suppose the plotted points in the response plot and residual plot of $\hat{Y}$ vs. r scatter about the identity line with zero intercept and unit slope and $r = 0$ line in roughly even bands. For OLS forward selection with C_p, we suggest $n \geq 5p$ and $n \geq 10d_{I_{min}}$. A model with $n < 5d$ overfits. Much larger values of n may be needed if the error distribution is skewed or multimodal. In the forward selection simulations, PI (11) often had good coverage but was rather long if $n \approx 5p$. See the following example. The Hurvich and Tsai (1989) AIC_C criterion can be useful when $n \geq \max(2p, 10d_{I_{min}})$.

Example 1. The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. See (http://parker.ad.siu.edu/Olive/sldata.txt). We will study the relationship between $Y =$ the *number of women married to civilians* in the district with the predictors $x_1 =$ constant, $x_2 = pop =$ the *population of the district in 1843*, $x_3 = mmen =$ the *number of married civilian men* in the district, $x_4 = mmilmen =$ the *number of married men in the military* in the district, and $x_5 = milwmn =$ the *number of women married to husbands in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence Y and x_3 are highly correlated but not equal. Similarly, x_4 and x_5 are highly correlated but not equal. $Y = x_3 + e$ is a good model.

Consider PI (11). Forward selection selected the model with the minimum C_p while the other methods used 10-fold CV. PLS and PCR used (p components) the OLS full model with 90% PI length 2395.74, forward selection used a constant and *mmen* with PI length 2114.72, ridge regression had PI length

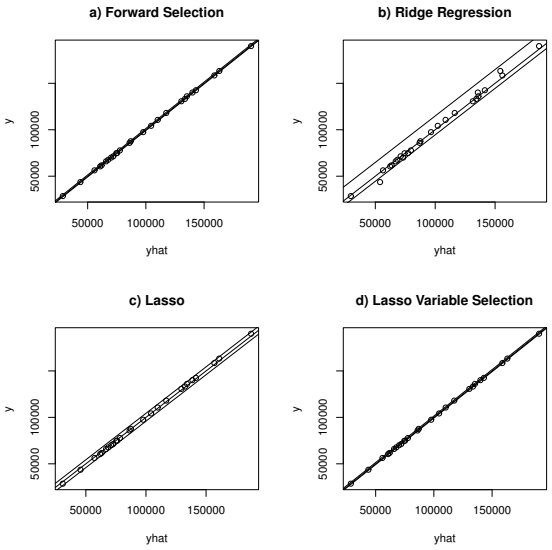

Fig. 1 Marry Data Response Plots.

20336.58, lasso and lasso variable selection used a constant, *mmen*, and *pop* with lengths 8482.62 and 2226.53, respectively. Figure 1 shows the response plots for forward selection, ridge regression, lasso, and lasso variable selection. The response plots for PLS, PCR, and the OLS full model were identical and similar to those of forward selection and lasso variable selection. The plots suggest that the MLR model is appropriate since the plotted points scatter about the identity line. The 90% pointwise PI (11) bands are also shown, and consist of two lines parallel to the identity line. These bands are very narrow in Figure 1 a) and d).

We used 5-fold CV with coverage and average 95% PI length to compare the forward selection models. All 4 models had 100% coverage with 5-fold CV (the 26 large sample 95% PIs formed using 5-fold CV each contained the corresponding Y_i). Hence the model with the shortest average PI length should be selected. The average PI lengths were 2591.243, 2741.154, 2902.628, and 2972.963 for the models with 2 to 5 predictors.

Example 1 illustrates a useful diagnostic that would be slow to simulate. Modify k–fold cross validation to compute the PI coverage and average PI length on all M models. Then n PIs are made for Y_i using $\boldsymbol{x}_f = \boldsymbol{x}_i$ for $i = 1, ..., n$. The coverage is the proportion of times the n PIs contained Y_i. When n is small, 2% undercoverage might be acceptable, but as n increases the amount of undercoverage should decrease to zero so that models that produce undercoverage are not favored. For example, choose the model I_d with the shortest average PI length given that the nominal large sample $100(1 - \delta)\%$

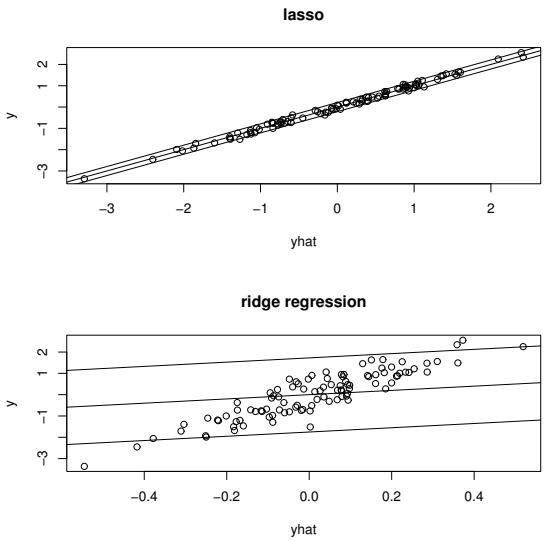

Fig. 2 Response Plots for Example 2.

PI had coverage

$$\geq c_n = \max(1 - \delta - \frac{1}{3\sqrt{n}}, 1 - \delta - 0.02).$$

If no model I_i had coverage $\geq c_n$, pick the model with the largest coverage.

Example 2. Suppose $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_{101} x_{101} + e = x_2 + e$ is simulated 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 Figure 2 which has the 90% pointwise PI (11) bands. A scaled shrinkage estimator is obtained by regressing $\boldsymbol{Y}$ on $\hat{\boldsymbol{Y}}$ to get $\hat{\boldsymbol{\beta}}_s$ where $\hat{\beta}_{is} = \hat{b}\hat{\beta}_i$ for $i = 2, ..., p$ and $\hat{\beta}_{1s} = \hat{a} + \hat{b}\hat{\beta}_1$ and $\hat{\boldsymbol{\beta}}$ is the estimator such as ridge regression. See Olive (2020, ch. 8) for R code. In the simulations for sparse models, lasso sometimes shrinks too much but lasso variable selection helps. Scaled shrinkage estimators may be useful if the population model or fitted model is not sparse.

Simulation

For the simulation, we used several R functions including forward selection (FS) as computed with the `regsubsets` function from the `leaps` library, principal components regression (PCR) with the `pcr` function and partial least squares (PLS) with the `plsr` function from the `pls` library, and ridge regression (RR) and lasso with the `cv.glmnet` function from the `glmnet` library. Lasso variable selection (LVS) was applied to the selected lasso model.

The simulation generated $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n, n+1$ where $(Y_f, \boldsymbol{x}_f) = (Y_{n+1}, \boldsymbol{x}_{n+1})$. Then a 95% prediction interval from a regression method was

Table 1 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n	p	ψ	k		FS	lasso	LVS	RR	PLS	PCR
100	20	0	1	cov	0.9644	0.9750	0.9666	0.9560	0.9438	0.9772
				len	4.4490	4.8245	4.6873	4.5723	4.4149	5.5647
100	40	0	1	cov	0.9654	0.9774	0.9588	0.9274	0.8810	0.9882
				len	4.4294	4.8889	4.6226	4.4291	4.0202	7.3393
100	200	0	1	cov	0.9648	0.9764	0.9268	0.9584	0.6616	0.9922
				len	4.4268	4.9762	4.2748	6.1612	2.7695	12.412
100	50	0	49	cov	0.8996	0.9719	0.9736	0.9820	0.8448	1.0000
				len	22.067	6.8345	6.8092	7.7234	4.2141	38.904
200	20	0	19	cov	0.9788	0.9766	0.9788	0.9792	0.9550	0.9786
				len	4.9613	4.9636	4.9613	5.0458	4.3211	4.9610
200	40	0	19	cov	0.9742	0.9762	0.9740	0.9738	0.9324	0.9792
				len	4.9285	5.2205	5.1146	5.2103	4.2152	5.3616
200	200	0	19	cov	0.9728	0.9778	0.9098	0.9956	0.3500	1.0000
				len	4.8835	5.7714	4.5465	22.351	2.1451	51.896
400	20	0.9	19	cov	0.9664	0.9748	0.9604	0.9726	0.9554	0.9536
				len	4.5121	10.609	4.5619	10.663	4.0017	3.9771
400	40	0.9	19	cov	0.9674	0.9608	0.9518	0.9578	0.9482	0.9646
				len	4.5682	14.670	4.8656	14.481	4.0070	4.3797
400	400	0.9	19	cov	0.9348	0.9636	0.9556	0.9632	0.9462	0.9478
				len	4.3687	47.361	4.8530	48.021	4.2914	4.4764
400	400	0	399	cov	0.9486	0.8508	0.5704	1.0000	0.0948	1.0000
				len	78.411	37.541	20.408	244.28	1.1749	305.93
400	800	0.9	19	cov	0.9268	0.9652	0.9542	0.9672	0.9438	0.9554
				len	4.3427	67.294	4.7803	66.577	4.2965	4.6533

made. This process was repeated for 5000 runs and the proportion of runs (cov) where the PI contained Y_f was recorded along with the average length (len) of the 5000 PIs. 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 $\text{Cov}(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ 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). Normal distributions usually appear in simulations, and 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 population shorth 95% PI lengths estimated by 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 $p = 2n$, where regularity conditions for consistent estimators are strong. The values $k = 1$ and $k = 19$ correspond to sparse models since a model is sparse if $a_S = k + 1$ is small in Equation (5). Lasso, lasso variable selection, and forward selection with EBIC can perform well for sparse models 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, a setting favorable for PLS and PCR. The simulated data sets are rather small since the some of the R estimators are rather slow.

The simulations were done in *R*. See R Core Team (2016). The results were similar for all five error distributions, and we show some results for the normal and shifted exponential distributions. A much larger simulation study is in Pelawa Watagoda (2017). Tables 1 and 2 show some simulation results for PI (11) where forward selection used C_p for $n \geq 10p$ and EBIC for $n < 10p$. The other methods minimized 10-fold CV. For forward selection, the maximum number of variables used was approximately $\min(\lceil n/5 \rceil, p)$. Ridge regression used the same d that was used for lasso.

For $n \geq 5p$, coverages tended to be near or higher than the nominal value of 0.95. The average PI length was often near 1.3 times the asymptotically optimal length for $n = 10p$ and close to the optimal length for $n = 100p$. C_p and EBIC produced good PIs for forward selection, and 10-fold CV produced good PIs for PCR and PLS. 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. Lasso did appear to select $S \subseteq I_{min}$ since lasso variable selection was good.

For n/p not large, good performance needed stronger regularity conditions, and all six methods can have problems. PLS tended to have severe undercoverage with small average length, but sometimes performed well for $\psi = 0.9$. The PCR length was often too long for $\psi = 0$. If there was $k = 1$ active population predictor, then forward selection with EBIC, lasso, and lasso variable selection often performed well. For $k = 19$, forward selection with EBIC often performed well, as did lasso and lasso variable selection for $\psi = 0$. For dense models with $k = p - 1$ and n/p not large, there was often undercoverage. Here forward selection with EBIC would use about the maximum of $n/5$ variables. Then coverage was low in Table 1 for $n = 100, p = 50$ and $k = 49$. Let $d - 1$ be the number of active nontrivial 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. EBIC occasionally had undercoverage, especially for $k = 19$ or $p - 1$, which was usually more severe for $\psi = 0.9$ or $1/\sqrt{p}$.

Table 2 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1) - 1$

n	p	ψ	k		FS	lasso	LVS	RR	PLS	PCR
100	20	0	1	cov	0.9622	0.9728	0.9648	0.9544	0.9460	0.9724
				len	3.7909	4.4344	4.3865	4.4375	4.2818	5.5065
2000	20	0	1	cov	0.9506	0.9502	0.9500	0.9488	0.9486	0.9542
				len	3.1631	3.1199	3.1444	3.2380	3.1960	3.3220
200	20	0.9	1	cov	0.9588	0.9666	0.9664	0.9666	0.9556	0.9612
				len	3.7985	3.6785	3.7002	3.7491	3.5049	3.7844
200	20	0.9	19	cov	0.9704	0.9760	0.9706	0.9784	0.9578	0.9592
				len	4.6128	12.1188	4.8732	12.0363	3.3929	3.7374
200	200	0.9	19	cov	0.9338	0.9750	0.9564	0.9740	0.9440	0.9596
				len	4.6271	37.3888	5.1167	56.2609	4.0550	4.6994
400	40	0.9	19	cov	0.9678	0.9654	0.9492	0.9624	0.9426	0.9574
				len	4.3433	14.7390	4.7625	14.6602	3.6229	4.1045

Table 3 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n,p,ψ,k		FS	CFS	LVS	CLVS	Lasso	CL	RR	CRR
200,20, 0,19	cov	0.9574	0.9446	0.9522	0.9420	0.9538	0.9382	0.9542	0.9430
	len	4.6519	4.3003	4.6375	4.2888	4.6547	4.2964	4.7215	4.3569
200,40,0,19	cov	0.9564	0.9412	0.9524	0.9440	0.9550	0.9406	0.9548	0.9404
	len	4.9188	4.5426	5.2665	4.8637	5.1073	4.7193	5.3481	4.9348
200,200, 0,19	cov	0.9488	0.9320	0.9548	0.9392	0.9480	0.9380	0.9536	0.9394
	len	7.0096	6.4739	5.1671	4.7698	31.1417	28.7921	47.9315	44.3321
400,20,0.9,19	cov	0.9498	0.9406	0.9488	0.9438	0.9524	0.9426	0.9550	0.9426
	len	4.4153	4.1981	4.5849	4.3591	9.4405	8.9728	9.2546	8.8054
400,40,0.9,19	cov	0.9504	0.9404	0.9476	0.9388	0.9496	0.9400	0.9470	0.9410
	len	4.7796	4.5423	4.9704	4.7292	13.3756	12.7209	12.9560	12.3118
400,400,0.9,19	cov	0.9480	0.9398	0.9554	0.9444	0.9506	0.9422	0.9506	0.9408
	len	5.2736	5.0131	4.9764	4.7296	43.5032	41.3620	42.6686	40.5578
400,800,0.9,19	cov	0.9550	0.9474	0.9522	0.9412	0.9550	0.9450	0.9550	0.9446
	len	5.3626	5.0943	4.9382	4.6904	60.9247	57.8783	60.3589	57.3323

Tables 3 and 4 show some results for PIs (12) and (13). Here forward selection using the minimum C_p model if $n_H \geq 10p$ and EBIC otherwise. The coverage was very good. Labels such as CFS and CLVS used PI (13). For lasso variable selection , the program sometimes failed to run for 5000 runs, e.g., if the number of variables selected $d = n_H$. In Table 3, PIs (12) and (13) are asymptotically equivalent, but PI (13) had shorter lengths for moderate n. In Table 4, PI (12) is shorter than PI (13) asymptotically, but for moderate n, PI (13) was often shorter.

Table 5 shows some results for PIs (11) and (12) for lasso and ridge regression. The header lasso indicates PI (11) was used while vlasso indicates that PI (12) was used. PI (12) tended to work better when the fit was poor while PI (11) was better for $n = 2p$ and $k = p - 1$. The PIs are asymptotically equivalent for consistent estimators.

Table 4 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1) - 1$

n,p,ψ,k		FS	CFS	LVS	CLVS	Lasso	CL	RR	CRR
200,20,0,1	cov	0.9596	0.9504	0.9588	0.9374	0.9604	0.9432	0.9574	0.9438
	len	4.6055	4.2617	4.5984	4.2302	4.5899	4.2301	4.6807	4.2863
2000,20,0,1	cov	0.9560	0.9508	0.9530	0.9464	0.9544	0.9462	0.9530	0.9462
	len	3.3469	3.9899	3.3240	3.9849	3.2709	3.9786	3.4307	3.9943
200,20,0.9,1	cov	0.9564	0.9402	0.9584	0.9362	0.9634	0.9412	0.9638	0.9418
	len	3.9184	3.8957	3.8765	3.8660	3.8406	3.8483	3.8467	3.8509
200,20,0.9,19	cov	0.9630	0.9448	0.9510	0.9368	0.9554	0.9430	0.9572	0.9420
	len	5.0543	4.6022	4.8139	4.3841	9.8640	9.0748	9.5218	8.7366
200,200,0.9,19	cov	0.9570	0.9434	0.9588	0.9418	0.9552	0.9392	0.9544	0.9394
	len	5.8095	5.2561	5.2366	4.7292	31.1920	28.8602	47.9229	44.3251
400,40,0.9,19	cov	0.9476	0.9402	0.9494	0.9416	0.9584	0.9496	0.9562	0.9466
	len	4.6992	4.4750	4.9314	4.6703	13.4070	12.7442	13.0579	12.4015

Table 5 PIs (11) and (12): Simulated Large Sample 95% PI Coverages and Lengths

n	p	ψ	k		dist	lasso	vlasso	RR	vRR
100	20	0	1	cov	N(0,1)	0.9750	0.9632	0.9564	0.9606
				len		4.8245	4.7831	4.5741	5.3277
100	20	0	1	cov	EXP(1)−1	0.9728	0.9582	0.9546	0.9612
				len		4.4345	5.0089	4.4384	5.6692
100	50	0	49	cov	N(0,1)	0.9714	0.9606	0.9822	0.9618
				len		6.8345	22.3265	7.7229	27.7275
100	50	0	49	cov	EXP(1)−1	0.9716	0.9618	0.9814	0.9608
				len		6.9460	22.4097	7.8316	27.8306
400	400	0	399	cov	N(0,1)	0.8508	0.9518	1.0000	0.9548
				len		37.5418	78.0652	244.1004	69.5812
400	400	0	399	cov	EXP(1)−1	0.8446	0.9586	1.0000	0.9558
				len		37.5185	78.0564	243.7929	69.5474

The collection of Olive (2020) *R* functions *slpack*, available from (http://parker.ad.siu.edu/Olive/slpack.txt), has some useful functions for the inference. Tables 1 and 2 were made with `mspisim`. For PI (12), the function `valvspisim2` is for forward selection using the minimum C_p model if $n_H \geq 10p$ and EBIC otherwise, and the function also computes the split conformal PI. The function `valrrpisim2` simulates lasso and ridge regression. The function `valrelpisim` simulates the lasso variable selection model corresponding to the lasso model chosen with 10-fold CV. This function sometimes fails with 5000 runs. For example, the function fails if the number of variables selected $d \geq n/2$. Tables 3 and 4 used these three functions. The function `pifold` can be used to do k-fold CV with PI coverage and average length. The function `AERplot2` makes a response plot with pointwise PIs (11) added. The function `srrpisim` can be used to simulate PIs for the scaled lasso and scaled ridge regression estimators of Example 2, and was used for Table 5.

5 Conclusion

Let p be fixed. Then $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is $\sqrt{n}$ consistent for forward selection with C_p and for lasso variable selection by Theorem 2. If $a_S < p$, then lasso tends not to be $\sqrt{n}$ consistent if lasso selects S with high probability by Fan and Li (2001). PLS appears to be $\sqrt{n}$ consistent by results in Cook and Forzani (2018, 2019). PCR tends to be inconsistent unless $P(d = p) \to 1$ as $n \to \infty$. It is not clear if ridge regression, as implemented by the `glmnet` package in R, is consistent. In the simulations with $n \geq 5p$, forward selection with C_p or EBIC, lasso variable selection, and PLS performed best. A recent survey for principal component regression is Artigue and Smith (2019). Using lasso or forward selection on the principal components may improve principal component regression, although this method may use the full OLS model too often.

If n/p is not large, regularity conditions for consistent estimators are strong. PLS is unreliable since sometimes PLS is inconsistent and sometimes $\sqrt{n}$ consistent by Cook and Forzani (2018, 2019). PCR was also unreliable. In the simulations, forward selection with EBIC and lasso variable selection performed best for sparse models. Suppose a researcher picks a method, say PLS. Then fit PLS and several competing methods. Using response plots and checking PIs (12) and (13) for coverage and average lengths on the validation set may suggest that PLS is useful or that an alternative method may be better for prediction.

As noted in Section 2, none of the three prediction intervals (11), (12), and (13) dominates the other two. The new prediction intervals (11) and (12) are asymptotically optimal for a large class of error distributions if $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. Prediction intervals described in Lei et al. (2018) and the new prediction intervals (11) and (12) are among the only prediction intervals that may be useful when the error distribution is unknown and n/p is small. PI (12) modified the split conformal PI (13) to be asymptotically optimal on a much larger class of error distributions. See Wasserman (2014) and Butler and Rothman (1980) for PIs related to PI (13). Denham (1997) gave a PI for PLS when the number of PLS components V_j is selected in advance. Also, see Romera (2010). Mohie El–Din and Shafay (2013) also derived prediction intervals based on order statistics. Lin, Foster, and Ungar (2012) noted that lasso and related methods can perform poorly in the presence of multicollinearity.

Acknowledgements The authors thank the Editor and two referees for their work.

References

1. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In Petrov BN, Csakim F (eds) Proceedings, 2nd international symposium on information theory. Akademiai Kiado, Budapest, pp 267-281
2. Artigue H, Smith G (2019) The principal problem with principal components regression. Cogent Math Stat 6, online
3. Belloni A, Chernozhukov V (2013) Least squares after model selection in high-dimensional sparse models. Bernoulli 19:521-547

4. Brillinger DR (1977) The identification of a particular nonlinear time series. Biometrika 64:509-515
5. Brillinger DR (1983) A generalized linear model with "Gaussian" regressor variables. In Bickel PJ, Doksum KA, Hodges JL (eds) A festschrift for Erich L. Lehmann. Wadsworth, Pacific Grove, pp 97-114
6. Butler R, Rothman E (1980) Predictive intervals based on reuse of the sample. J Am Stat Assoc 75:881-889
7. Cai T, Tian L, Solomon SD, Wei LJ (2008) Predicting future responses based on possibly misspecified working models. Biometrika 95:75-92
8. Charkhi A, Claeskens G (2018) Asymptotic post-selection inference for the Akaike information criterion. Biometrika 105:645-664
9. Chun H, Keleş S (2010). Sparse partial least squares regression for simultaneous dimension reduction and predictor selection. J Royal Stat Soc B 72:3-25
10. Claeskens G, Hjort NL (2008) Model Selection and Model Averaging. Cambridge University Press, New York
11. Cook RD (2018) An introduction to envelopes: dimension reduction for efficient estimation in multivariate statistics. Wiley, Hoboken
12. Cook RD, Forzani L (2018) Big data and partial least squares prediction. Can J Stat 46:62-78
13. Cook RD, Forzani L (2019) Partial least squares prediction in high-dimensional regression. Ann Stat 47:884-908
14. Cook RD, Helland IS, Su Z (2013) Envelopes and partial least squares regression. J Royal Stat Soc B 75:851-877
15. Cook RD, Weisberg S (1999) Applied regression including computing and graphics. Wiley, New York
16. Denham MC (1997) Prediction intervals in partial least squares. J Chemometrics 11:39-52
17. Efron B, Hastie T (2016) Computer age statistical inference. Cambridge University Press, New York
18. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression (with discussion). Ann Stat 32:407-451
19. Fan J, Li R (2001) Variable selection via noncave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348-1360.
20. Frey J (2013) Data-driven nonparametric prediction intervals. J Stat Plan Inference 143:1039-1048
21. Grübel R (1988) The length of the shorth. Ann Stat 16:619-628.
22. Gunst RF, Mason RL (1980) Regression analysis and its application. Marcel Dekker, New York
23. Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press Taylor & Francis, Boca Raton
24. Hebbler B (1847) Statistics of Prussia. J Royal Stat Soc A 10:154-186
25. Hoerl AE, Kennard D (1970) Ridge regression: biased estimation for nonorthogonal problems. Technom 12:55-67
26. Hong L, Kuffner TA, Martin R (2018) On overfitting and post-selection uncertainty assessments. Biometrika 105:221-224
27. Hurvich C, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76:297-307
28. Hyndman RJ (1996) Computing and graphing highest density regions. Am Stat 50:120-126
29. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R. Springer, New York
30. Jansen L, Fithian W, Hastie T (2015) Effective degrees of freedom: a flawed metaphor. Biometrika 102:479-485
31. Jia J, Yu B (2010) On model selection consistency of the elastic net when $p >> n$. Stat Sinica 20:595-611
32. Jones HL (1946) Linear regression functions with neglected variables. J Am Stat Assoc 41:356-369
33. Knight K, Fu WJ (2000) Asymptotics for lasso-type estimators. Ann Stat 28:1356-1378

34. Lei J (2019) Fast exact conformalization of lasso using piecewise linear homotopy. Biometrika 106:749-764
35. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018) Distribution-free predictive inference for regression. J Am Stat Assoc 113:1094-1111
36. Li K–C (1987) Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann Stat 15:958-975
37. Lin D, Foster DP, Ungar LH (2012) VIF regression, a fast regression algorithm for large data. J Am Stat Assoc 106:232-247
38. Luo S, Chen Z (2013) Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. J Stat Plan Inference 143:494-504
39. Mallows C (1973) Some comments on C_p. Technom 15:661-676
40. Meinshausen N (2007) Relaxed lasso. Comput Stat Data Anal 52:374-393
41. Mohie El–Din MM, Shafay AR (2013) One– and two–sample Bayesian prediction intervals based on progressively type–II censored data. Stat Pap 54:287-307
42. Nishii R (1984) Asymptotic properties of criteria for selection of variables in multiple regression. Ann Stat 12:758-765
43. Olive DJ (2007) Prediction intervals for regression models. Computat Stat Data Anal 51:3115-3122
44. Olive DJ (2013) Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. Internat J Stat Probab 2:90-100
45. Olive DJ (2017a) Linear Regression. Springer, New York
46. Olive DJ (2017b) Robust Multivariate Analysis. Springer, New York
47. Olive DJ (2018) Applications of hyperellipsoidal prediction regions. Stat Pap 59:913-931
48. Olive DJ (2020) Prediction and statistical learning, unpublished online course notes. (http://parker.ad.siu.edu/Olive/slearnbk.htm)
49. Olive DJ, Hawkins DM (2005) Variable selection for 1D regression models. Technom 47:43-50
50. Pelawa Watagoda LCR (2017) Inference after variable selection, PhD Thesis, Southern Illinois University. (http://parker.ad.siu.edu/Olive/slasanthiphd.pdf)
51. Pelawa Watagoda LCR, Olive DJ (2019) Bootstrapping multiple linear regression after variable selection. Stat Pap to appear
52. Pilz M (2020) Consistency of the elastic net under a finite second moment assumption on the noise. J Stat Plan Inference 204:72-79
53. Pratt JW (1959) On a general concept of "in Probability." Ann Math Statist 30:549-558
54. Pötscher B (1991) Effects of model selection on inference. Econometric Theory 7:163-185
55. R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
56. Rathnayake RC, Olive DJ (2020) Bootstrapping some GLMs and survival regression models after variable selection. Unpublished manuscript at (http://parker.ad.siu.edu/Olive/ppbootglm.pdf).
57. Romera R (2010) Prediction intervals in partial least squares regression via a new local linearization approach. Chemometrics Intelligent Laboratory Systems 103:122–128
58. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461-464
59. Sen PK, Singer JM (1993) Large sample methods in statistics: an introduction with applications. Chapman & Hall, New York
60. Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88:486-494
61. Slawski M, zu Castell W, Tutz G (2010) Feature selection guided by structural information. Ann Applied Stat 4:1056-1080
62. Su Z, Cook RD (2012) Inner envelopes: efficient estimation in multivariate linear regression. Biometrika 99:687-702
63. Sun T, Zhang, C-H (2012) Scaled sparse linear regression. Biometrika 99:879-898.
64. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stat Soc B 58:267-288
65. Wasserman L (2014) Discussion: a significance test for the lasso. Ann Stat 42:501-508
66. Wold H (1975) Soft modelling by latent variables: the nonlinear partial least squares (NIPALS) approach. In Gani J (ed) Perspectives in probability and statistics, papers in honor of M.S. Bartlett. Academic Press, San Diego, pp 117-144
67. Zhao P, Yu B (2006) On model selection consistency of lasso. J Mach Learn Res 72:2541-2563
68. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc B 67:301-320