BIAS AND VARIANCE REDUCTION IN ESTIMATION OF MODEL DIMENSION

WEI-YIN LOH AND XIAODONG ZHENG

(Communicated by James Glimm)

ABSTRACT. The problem of estimating the number of regressors to include in a linear regression model is considered. Estimators based on the final prediction error and Akaike's criterion frequently have large positive bias. Shrinkage correction factors and bootstrapping are used to produce new estimators with reduced bias. The asymptotic bias and mean-squared errors of these estimators are derived analytically. Finite-sample estimates are obtained by simulation.

1. INTRODUCTION

Consider the following estimation problem. Data are observed from the linear regression model

$$(1) \quad y_i = x_{i1}\beta_1 + \cdots + x_{iK}\beta_K + \varepsilon_i, \quad \beta_{k_0+1} = \beta_{k_0+2} = \cdots = \beta_K = 0, \quad i = 1, \dots, n,$$

where $0 \leq k_0 \leq K$ and the ε_i , $i = 1, 2, \dots, n$, are independent errors from a normal distribution with zero mean and known standard deviation σ^2 . Assume without loss of generality that $\sigma^2 = 1$. The problem is to estimate k_0 , the model dimension. It is well known that if k_0 is underestimated (the so-called "underfitting problem"), then least-squares estimates of the regression coefficients are biased. On the other hand, if k_0 is overestimated (the "overfitting problem"), then coefficient estimates remain unbiased, but their variance estimates are inflated (see, e.g., [3 §6.1]). Overfitting is often thought to be the lesser of the two evils.

A standard approach to estimation of k_0 employs the final prediction error (FPE) criterion

$$C(k, \lambda) = \text{RSS}(k) + \lambda k, \quad 0 \leq k \leq K,$$

where $\text{RSS}(k)$ is the residual sum of squares under (1) when only the first k covariates are used and where λ is a penalty parameter for overfitting. The choice $\lambda = 2$ is equivalent to Mallows's C_p and Akaike's AIC criteria, and the choice $\lambda = \log n$ to the BIC criterion. The FPE estimator is defined as the

Received by the editors March 12, 1993.

1991 *Mathematics Subject Classification.* Primary 62G09, 62J05; Secondary 62G20.

Key words and phrases. Akaike criterion, cross-validation, final prediction error, model selection.

Research partially supported by ARO grant DAAL03-91-G-0111.

minimizer of $C(k, \lambda)$ as

$$\tilde{k} = \arg \min_{0 \leq k \leq K} C(k, \lambda).$$

(See, e.g., [1, 4] for methods based on cross-validation.) Let $\mathbf{X}_k$ denote the design matrix for the model with k covariates, let $P_k = \mathbf{X}_k(\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k'$, and let $f = \mathbf{X}_{k_0}(\beta_1, \dots, \beta_{k_0})^t$ be the true regression mean function (the superscript t denotes transpose). Suppose the following two conditions hold:

C1. $K < n$ and $(\mathbf{X}_k' \mathbf{X}_k)^{-1}$ exists for all $k \leq K$.

C2. $\lambda > 1$ and there exists a $\delta > 0$ such that $f^t(I - P_k)f \geq \delta n$ for all $k < k_0$.

It follows from the results in [5, Theorem 5.2] and [6] that under these conditions, $\tilde{k} \rightarrow_p T$ as $K \rightarrow \infty$, where T is an integer-valued variable with probability-generating function given by

$$E z^{T-k_0} = \exp \left\{ \sum_{j=1}^{\infty} j^{-1} a_j (z^j - 1) \right\}, \quad 0 < z \leq 1.$$

Here $a_j = P(\chi_j^2 > \lambda j)$, $c = \sum_{j=1}^{\infty} a_j$,

$$(3) \quad E(T) = k_0 + c,$$

and χ_j^2 is a chi-square random variable with j degrees of freedom. It is clear that $T \geq k_0$ a.s. and hence $ET \geq k_0$, since the right-hand side (rhs) of (2) is expressible as an infinite series in nonnegative powers of z . The estimator $\tilde{k}$ is therefore usually positively biased, at least with large samples. Our goal is to find an estimator of k_0 with smaller asymptotic bias and smaller mean-square error than $\tilde{k}$. Although the estimator $(\tilde{k} - c)$ has zero asymptotic bias, it takes negative values with positive probability. This problem is avoided by the truncated estimator $\bar{k} = \max\{(\tilde{k} - c), 0\}$, but (see Table 2 in §4) it is frequently negatively biased.

2. SHRINKAGE ESTIMATOR

A natural way to reduce the bias of $\tilde{k}$ is to “shrink” it toward 0 as follows. Let $\gamma = E(T)/k_0 = 1 + ck_0^{-1}$ denote the number of times the asymptotic mean of $\tilde{k}$ is larger than k_0 . Then $\tilde{k}\gamma^{-1}$ has zero asymptotic bias for k_0 . Since we do not know the value of γ , we may estimate it with $(1 + c\tilde{k}^{-1})$. This yields the shrinkage estimator (truncated to be integer-valued)

$$\hat{k} = \lfloor \tilde{k}/(1 + c\tilde{k}^{-1}) \rfloor + 1 = \lfloor \tilde{k}^2/(\tilde{k} + c) \rfloor + 1,$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x . To analyze its properties, define the sets

$$A_m = \begin{cases} \{j : \lfloor (m-1)c(1+c-m)^{-1} \rfloor + 1 \\ \leq j \leq \lfloor mc(c-m)^{-1} \rfloor\}, & m = 1, \dots, \lfloor c \rfloor, \\ \{j : j \geq \lfloor c \rfloor(c - \lfloor c \rfloor)^{-1} \rfloor\}, & m = \lfloor c \rfloor + 1, \end{cases}$$

and use the convention that $x/0 = \infty$ if $x > 0$. Further define

$$S = T^2(T+c)^{-1}, \quad W = \lfloor S \rfloor + 1, \quad V = \sum_{m=1}^{\lfloor c \rfloor + 1} mI(T \in A_m).$$

Since $T \geq k_0$ a.s. and $\bigcup_{m=1}^{\lfloor c \rfloor + 1} A_m = \{j : j \geq 1\}$, we have $1 \leq V \leq \lfloor c \rfloor + 1$ with probability one if $k_0 \geq 1$. The following lemma gives the relation between S and V .

Lemma 1.

$$(4) \quad \lfloor S \rfloor = T - V.$$

Proof. Since $\lfloor S \rfloor = \lfloor T - cT(T+c)^{-1} \rfloor$ and T is a nonnegative integer, we have

$$\begin{aligned} \lfloor S \rfloor &= \sum_{m=1}^{\infty} (T-m)I(m-1 < cT(T+c)^{-1} \leq m) \\ (5) \quad &= T - \sum_{m=1}^{\infty} mI(m-1 < cT(T+c)^{-1} \leq m) \\ &\equiv T - \sum_{m=1}^{\infty} mI_{B_m}, \quad \text{say.} \end{aligned}$$

We consider two cases.

Case 1. c is not an integer. Then

$$B_m = \begin{cases} \{(m-1)c(c+1-m)^{-1} < T \leq mc(c-m)^{-1}\}, & 1 \leq m < c, \\ \{T > (m-1)c(c+1-m)^{-1}\}, & m \in (c, c+1), \\ \emptyset, & m > c+1. \end{cases}$$

This yields (5) and hence (4).

Case 2. c is a positive integer. Then

$$B_m = \begin{cases} \{(m-1)c(c+1-m)^{-1} < T \leq mc(c-m)^{-1}\}, & 1 \leq m < c, \\ \{T > c(c-1)\} & m = c, \\ \emptyset, & m \geq c+1. \end{cases}$$

Because $A_{\lfloor c \rfloor + 1} = \{T \geq \infty\} = \emptyset$, this also yields (4). $\square$

We now show that W is the limit of $\hat{k}$.

Theorem 1. Suppose conditions C1 and C2 are satisfied. Then $\hat{k} \rightarrow_P W$.

Proof. An analogous proof to that in the previous theorem shows that

$$\hat{k} = \lfloor \tilde{k}^2/(\tilde{k} + c) \rfloor + 1 = \tilde{k} + 1 - \sum_{m=1}^{\lfloor c \rfloor + 1} mI(\tilde{k} \in A_m).$$

Therefore,

$$\begin{aligned} |\hat{k} - W| &\leq |\tilde{k} - T| + \sum_{m=1}^{\lfloor c \rfloor + 1} m|I(\tilde{k} \in A_m) - I(T \in A_m)| \\ &\leq |\tilde{k} - T| + \sum_{m=1}^{\lfloor c \rfloor + 1} mI(|\tilde{k} - T| \geq 1) \\ &= |\tilde{k} - T| + 2^{-1}(\lfloor c \rfloor + 1)(\lfloor c \rfloor + 2)I(|\tilde{k} - T| \geq 1) \\ &\leq |\tilde{k} - T| + 2^{-1}(\lfloor c \rfloor + 1)(\lfloor c \rfloor + 2)|\tilde{k} - T| \rightarrow_P 0, \end{aligned}$$

since $\tilde{k} \rightarrow_P T$. $\square$

The next two theorems give the mean and variance of S . It is a positively biased estimate of k_0 .

Theorem 2. $ES = k_0 + c^2Q > k_0$, where

$$Q = \int_0^1 z^{k_0+c-1} \exp \left\{ \sum_{j=1}^{\infty} j^{-1} P(\chi_j^2 > \lambda j) (z^j - 1) \right\} dz.$$

Furthermore, a sharp upper bound for ES is $k_0 + c - \Lambda$ where

$$\Lambda = c \left(1 + \sum_{j=1}^{\infty} j^{-1} P[\chi_j^2 > \lambda j] \right)^{-1} \left\{ \sum_{j=1}^{\infty} (j+c)^{-1} P[\chi_j^2 > \lambda j] \right\} > 0$$

is independent of k_0 . Hence S has a smaller bias than T by an amount at least Λ .

Proof. Multiplying z^{k_0+c-1} to both sides of (2) and integrating over z gives

$$E(T+c)^{-1} = \int_0^1 z^{k_0+c-1} \exp \left\{ \sum_{j=1}^{\infty} j^{-1} a_j (z^j - 1) \right\} dz = Q.$$

Because $S = T - c + c^2/(T+c)$ and $ET = k_0 + c$, the first assertion follows.

To derive the upper bound, we use the inequality $\exp(z) < (1-z)^{-1}$, $0 < z < 1$, to obtain

$$(6) \quad Q \leq \int_0^1 z^{k_0+c-1} \left[1 + \sum_{j=1}^{\infty} a_j (1 - z^j) / j \right]^{-1} dz.$$

Write $a = \sum_{j=1}^{\infty} a_j j^{-1}$, and note that $(1+x)^{-1} \leq 1-bx$ for all $0 \leq x \leq b^{-1}-1$, $0 < b < 1$. Therefore, by setting $b = 1/(1+a)$, we see that the rhs of (6) is less than or equal to

$$\int_0^1 z^{k_0+c-1} \left[1 - (1+a)^{-1} \sum_{j=1}^{\infty} a_j (1 - z^j) j^{-1} \right] dz.$$

Thus

$$\begin{aligned} ES &= k_0 + c^2Q \\ &\leq k_0 + c^2 \int_0^1 z^{k_0+c-1} \left[1 - (1+a)^{-1} \sum_{j=1}^{\infty} a_j (1 - z^j) j^{-1} \right] dz \\ (7) \quad &= k_0 + c^2(k_0+c)^{-1} - c^2(1+a)^{-1} \int_0^1 z^{k_0+c-1} \sum_{j=1}^{\infty} a_j (1 - z^j) j^{-1} dz \\ &= k_0 + c^2(k_0+c)^{-1} - c^2(1+a)^{-1} (k_0+c)^{-1} \sum_{j=1}^{\infty} a_j (j+k_0+c)^{-1} \\ &= k_0 + c - \Lambda(k_0), \end{aligned}$$

where $\Lambda(k_0) = ck_0(k_0 + c)^{-1} + c^2(k_0 + c)^{-1}(1 + a)^{-1} \sum_{j=1}^{\infty} a_j(j + k_0 + c)^{-1}$. Treating $k_0 \geq 0$ as a continuous variable, it can be verified that the derivative

$$\Lambda'(k_0) = c^2(1 + a)^{-1} \left\{ (k_0 + c)^{-2} + \sum_{j=1}^{\infty} a_j(k_0^2 + c^2 + 2ck_0) \times (k_0 + c)^{-2}(j + k_0 + c)^{-2}j^{-1} \right\} > 0.$$

Hence $\inf_{k_0 \geq 0} \Lambda(k_0) = \Lambda(0) = c(1 + a)^{-1} \sum_{j=1}^{\infty} a_j(j + c)^{-1}$. This completes the proof. The bound is attained in the limit as $c \rightarrow 0$. $\square$

Theorem 3. $\text{Var}(S) = \text{Var}(T) + J$, where $J = 2c^2 + c^4L - 2c^2(k_0 + 2c)Q - c^4Q^2$ and

$$L = \int_0^1 z^{k_0+c-1} (-\log z) \exp \left\{ \sum_{j=1}^{\infty} j^{-1} P(\chi_j^2 > \lambda j) (z^j - 1) \right\} dz.$$

Proof. Since $S = T + c + c^2(T + c)^{-1} - 2c$, we have

$$\begin{aligned} \text{Var}(S) &= \text{Var}(T) + 2c^2 \text{Cov}(T + c, [T + c]^{-1}) + c^4 \text{Var}([T + c]^{-1}) \\ &= \text{Var}(T) + 2c^2\{1 - (k_0 + 2c)Q\} + c^4\{E(T + c)^{-2} - Q^2\}. \end{aligned}$$

Hence it suffices to show that $E(T + c)^{-2} = L$. Again multiplying z^{k_0+c-1} to each side of (2) and integrating over $z \in (0, w]$, where $0 < w < 1$, gives

$$(8) \quad Ew^{T+c}(T + c)^{-1} = \int_0^w z^{k_0+c-1} \exp \left\{ \sum_{j=1}^{\infty} j^{-1} a_j(z^j - 1) \right\} dz.$$

Now multiply w^{-1} to both sides of (8) and integrate over $w \in (0, 1)$ to obtain

$$E(T + c)^{-2} = \int_0^1 \int_0^w w^{-1} z^{k_0+c-1} \exp \left\{ \sum_{j=1}^{\infty} j^{-1} a_j(z^j - 1) \right\} dz dw.$$

This last double integral is verified to be equal to L by reversing the order of the integrals. $\square$

The next result gives conditions under which the value of W lies between those of S and T . It also shows that the limiting bias of $\hat{k}$ is nonnegative and that it is not greater than that of $\tilde{k}$.

Theorem 4. Suppose $k_0 \geq 1$ and $\lambda > 1$. Then $S \leq W \leq T$ a.s. Furthermore $W = T$ a.s. if and only if $0 < c \leq 1$. Thus, with regard to bias, W is no better than S , but it is at least as good as T ; W has a strictly smaller bias than T if and only if $c > 1$.

Proof. That $S \leq W$ is trivial. Now according to (4), $W = T - V + 1$. Because $1 \leq V \leq [c] + 1$ if $k_0 \geq 1$, we have $W \leq T$. Furthermore, $P(W = T) = 1$ if and only if $P(V = 1) = 1$ or $P(A_1) = 1$. Now if $0 < c \leq 1$ then $P(A_1) = P(T \geq 1) = 1$ because $T \geq k_0 \geq 1$. Otherwise if $c > 1$, then $P(A_1) = P(1 \leq T \leq [c(c-1)^{-1}]) < 1$ since T has unbounded support. $\square$

The following lemma from [2, p. 104] is needed for the proofs of the next two theorems.

Lemma 2. Suppose Z is an integrable random variable, b is a finite constant, and g is a convex function on the real line $\mathcal{R}$. If $\alpha(t)$ and $t - \alpha(t)$ are nondecreasing functions on $\mathcal{R}$, then $\alpha(Z)$ is integrable and

$$Eg(Z - EZ + b) \geq Eg(\alpha(Z) - E\alpha(Z) + b).$$

We are now ready to show that if $\lambda > 1$, the limiting variance of $\hat{k}$ is not greater than that of $\tilde{k}$.

Theorem 5. For all $k_0 \geq 0$, $\lambda > 1$, and $p \geq 1$,

$$(9) \quad E|S - ES|^p \leq E|T - ET|^p,$$

$$(10) \quad E|W - EW|^p \leq E|T - ET|^p.$$

In particular, we have $\text{Var}(S) \leq \text{Var}(T)$ and $\text{Var}(W) \leq \text{Var}(T)$.

Proof. Set $Z = T$, $b = 0$, and $\alpha(t) = t^2(t+c)^{-1}$, $t \geq 0$, in Lemma 2. It is clear that $\alpha(t)$ and $t - \alpha(t) = ct(t+c)^{-1}$ are both nondecreasing functions of $t \geq 0$. Because $T \geq 0$, we have

$$(11) \quad Eg(T - ET) \geq Eg(S - ES)$$

for any convex g , in particular, for $g(t) = |t|^p$, $p \geq 1$. This proves (9).

To prove (10), choose $\alpha(t) = t + 1 - \sum_{m=1}^{\lfloor ct \rfloor + 1} mI(t \in A_m) = \lfloor t^2(t+c)^{-1} \rfloor + 1$, $t \geq 0$. Since $t^2(t+c)^{-1}$ is nondecreasing in $t \geq 0$, $\alpha(t)$ is also a nondecreasing function of t . On the other hand, $\sum_{m=1}^{\lfloor ct \rfloor + 1} mI(t \in A_m)$ is nondecreasing in t ; therefore, so is $t - \alpha(t)$. The rest of the proof is similar to that leading to (9). $\square$

Let $\langle x \rangle$ be the nearest integer to x , with $\langle m + 1/2 \rangle = m$ for integral m . Another way to round off $\tilde{k}^2/(\tilde{k} + c)$ to be integer-valued is to define the estimator $\hat{k}' = \langle \tilde{k}^2/(\tilde{k} + c) \rangle$. The following theorem shows that it too has no larger limiting variance than $\tilde{k}$.

Theorem 6. For all $\lambda > 1$, $k_0 \geq 0$, $p \geq 1$,

$$(12) \quad E|\langle S \rangle - E\langle S \rangle|^p \leq E|T - ET|^p.$$

Hence $\langle S \rangle$ has smaller variance than T .

Proof. Since S is nondecreasing in T , so is $\langle S \rangle$. Furthermore,

$$\begin{aligned} \langle S \rangle &= \sum_{m=0}^{\infty} (T - m)I(m - 1/2 < cT[T+c]^{-1} \leq m + 1/2) \\ &= T - \sum_{m=0}^{\infty} mI(m - 1/2 < cT[T+c]^{-1} \leq m + 1/2). \end{aligned}$$

Thus

$$T - \langle S \rangle = \sum_{m=0}^{\infty} mI(m - 1/2 < cT[T+c]^{-1} \leq m + 1/2)$$

is also a nondecreasing function of T . Lemma 2 may then be applied as in the proof of Theorem 5. $\square$

3. OTHER ESTIMATORS

3.1. Bootstrap estimate. It is tempting to think that the shrinkage method may be iterated to yield further improvements. Specifically, we can start with $\hat{k}$ instead of $\tilde{k}$ and shrink the former. Let $\gamma_1 = E\hat{k}/k_0$. Then $\hat{k}/\gamma_1$ has zero bias for k_0 . To make this into an estimator, we need to estimate $E(\hat{k})$ and k_0 . The latter can be estimated by $\hat{k}$. We use the bootstrap to estimate $E\hat{k}$ as follows.

Define $x_i = (x_{i1}, \dots, x_{iK})$, and let $(\hat{\beta}_1, \dots, \hat{\beta}_K)$ be the least squares estimates of the coefficients. Let ε_i^* , $i = 1, \dots, n$, be independent standard normal variables. We generate bootstrap samples $\{(x_{i1}, y_1^*), \dots, (x_n, y_n^*)\}$, where

$$y_i^* = x_{i1}\hat{\beta}_1 + \dots + x_{iK}\hat{\beta}_K + \varepsilon_i^*, \quad i = 1, \dots, n.$$

Let E_* denote expectation under bootstrap sampling. This yields the following two bootstrap estimates of k_0 :

$$\hat{k}_1^* = \lfloor \hat{k}^2 / E_* \hat{k} \rfloor + 1, \quad \hat{k}_2^* = \langle \hat{k}^2 / E_* \hat{k} \rangle.$$

3.2. Alternative initial estimate. Instead of using $\tilde{k}$ as the initial estimate of k_0 in §2, we may use $\bar{k} = \max\{\langle \tilde{k} - c \rangle, 0\}$, which is the unbiased estimate truncated at 0. Letting $\bar{\gamma} = 1 + c\bar{k}^{-1}$, this yields the alternative estimator

$$\bar{k}^0 = \lfloor \tilde{k} \bar{\gamma}^{-1} \rfloor + 1 = \lfloor \tilde{k} (1 + c\bar{k}^{-1})^{-1} \rfloor + 1.$$

4. SIMULATION RESULTS

Tables 1 and 2 on the next page give Monte Carlo estimates of the mean-square errors and biases of the estimators. The results are based on samples of size 200 and 10,000 simulation trials. Standard errors are given in parentheses. Four different values of the true model dimension ($k_0 = 0, 2, 5, 10$) and five values of λ are used. The full dimension of each simulated regression model is $K = 10k_0$, except for the case $k_0 = 0$ where we use $K = 5$. In each simulation, an orthogonal design matrix was generated according to the true model, with parameters $\beta_i = i$. (Similar results were obtained with $\beta_i = k_0 - i + 1$.) The estimators included for consideration are $\tilde{k}$, $\bar{k}$, $\hat{k}$, $\hat{k}'$, $\bar{k}^0$, $\hat{k}_1^*$, and $\hat{k}_2^*$.

The six entries from these tables that had the worst bias (large negative bias is considered worse than large positive bias) are graphed in Figure 1 on page 1271. Also shown are the corresponding mean-square errors. Only six of the estimators are represented in the graphs ($\tilde{k}$ is excluded because of its large mean-square errors). Because large negative bias is undesirable, we conclude that the estimators $\bar{k}$, $\hat{k}'$, and $\bar{k}^0$ are inferior to $\hat{k}$, $\hat{k}_1^*$, and $\hat{k}_2^*$. All six estimators have mean-square errors that are substantially smaller than those of $\tilde{k}$ when $\lambda < 2$.

TABLE 1. Estimated mean-square errors with estimated SEs in parentheses; $n = 200$; $K = 10k_0$ for $k_0 \neq 0$; $K = 5$ when $k_0 = 0$

λ	k_0	$\hat{k}$	$\bar{k}$	$\hat{k}$	$\hat{k}'$	$\bar{k}^0$	$\hat{k}_1^*$	$\hat{k}_2^*$
1.1	0	6.1 (0.2)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	1.0 (0.0)	4.0 (0.0)	1.0 (0.0)
	2	76.9 (2.7)	4.0 (0.0)	1.1 (0.1)	2.5 (0.1)	1.0 (0.0)	1.7 (0.1)	1.3 (0.1)
	5	449 (16.)	25.0 (0.0)	27.6 (1.2)	29.6 (1.2)	16.0 (0.0)	30.0 (1.5)	32.3 (1.5)
	10	1392 (54.)	100 (0.0)	215 (11.)	212 (10.)	81 (0.0)	207.5 (12.0)	209.8 (12.0)
1.5	0	3.4 (0.2)	0.1 (0.0)	2.1 (0.1)	0.9 (0.1)	1.2 (0.0)	3.3 (0.1)	2.7 (0.1)
	2	23.9 (1.5)	13.0 (1.1)	14.4 (1.1)	13.9 (1.2)	13.2 (1.1)	15.1 (1.2)	15.4 (1.2)
	5	61.8 (5.8)	48.6 (5.6)	46.7 (5.3)	44.6 (5.1)	49.9 (5.6)	42.8 (5.1)	44.8 (5.3)
	10	91.6 (12.)	74.6 (11.)	73.0 (11.)	73.1 (11.)	75.8 (11.)	56.4 (8.8)	57.0 (8.7)
2.0	0	1.8 (0.1)	0.8 (0.1)	2.5 (0.1)	1.0 (0.1)	2.5 (0.1)	3.7 (0.2)	3.0 (0.2)
	2	6.2 (0.6)	5.5 (0.6)	5.5 (0.6)	5.3 (0.6)	5.5 (0.6)	7.1 (0.8)	6.4 (0.8)
	5	7.1 (1.0)	6.2 (1.0)	7.1 (1.0)	6.2 (1.0)	7.1 (1.0)	8.6 (1.1)	8.4 (1.2)
	10	5.2 (1.0)	4.5 (1.0)	5.2 (1.0)	4.5 (1.0)	5.2 (1.0)	7.1 (1.0)	7.0 (1.1)
2.5	0	0.9 (0.1)	0.9 (0.1)	1.7 (0.1)	0.9 (0.1)	1.7 (0.1)	2.5 (0.1)	2.1 (0.1)
	2	2.7 (0.3)	2.7 (0.3)	1.7 (0.3)	2.7 (0.3)	1.7 (0.3)	3.1 (0.5)	2.6 (0.5)
	5	1.8 (0.3)	1.8 (0.3)	1.8 (0.3)	1.8 (0.3)	1.8 (0.3)	2.7 (0.4)	2.2 (0.3)
	10	1.3 (0.2)	1.3 (0.2)	1.3 (0.2)	1.3 (0.2)	1.3 (0.2)	2.3 (0.3)	1.7 (0.3)
3.0	0	0.5 (0.1)	0.5 (0.1)	1.4 (0.1)	0.5 (0.1)	1.4 (0.1)	1.9 (0.1)	1.7 (0.1)
	2	2.3 (0.2)	2.3 (0.2)	1.0 (0.2)	2.3 (0.2)	1.0 (0.2)	2.0 (0.2)	1.5 (0.2)
	5	0.4 (0.1)	0.4 (0.1)	0.4 (0.1)	0.4 (0.1)	0.4 (0.1)	1.1 (0.1)	0.7 (0.1)
	10	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)	1.3 (0.3)	0.9 (0.3)

TABLE 2. Estimated bias with estimated SEs in parentheses; $n = 200$; $K = 10k_0$ for $k_0 \neq 0$; $K = 5$ when $k_0 = 0$.

λ	k_0	$\hat{k}$	$\bar{k}$	$\hat{k}$	$\hat{k}'$	$\bar{k}^0$	$\hat{k}_1^*$	$\hat{k}_2^*$
1.1	0	1.6 (0.1)	0.0 (0.0)	1.0 (0.0)	0.0 (0.0)	1.0 (0.0)	2.0 (0.0)	1.0 (0.0)
	2	6.0 (0.2)	-2.0 (0.0)	-0.4 (0.0)	-1.2 (0.0)	-1.0 (0.0)	-0.3 (0.0)	-0.4 (0.0)
	5	14.9 (0.5)	-5.0 (0.0)	0.1 (0.2)	-0.5 (0.2)	-4.0 (0.0)	-0.3 (0.2)	-1.0 (0.2)
	10	25.0 (0.9)	-10.0 (0.0)	2.3 (0.5)	1.8 (0.5)	-9.0 (0.0)	0.4 (0.5)	-0.2 (0.5)
1.5	0	1.0 (0.1)	0.1 (0.0)	1.3 (0.0)	0.4 (0.0)	1.1 (0.0)	1.5 (0.0)	1.4 (0.0)
	2	2.0 (0.1)	-0.4 (0.1)	1.0 (0.1)	0.6 (0.1)	0.6 (0.1)	1.0 (0.1)	0.2 (0.1)
	5	3.7 (0.2)	-0.4 (0.2)	1.5 (0.2)	3.7 (0.2)	0.7 (0.2)	0.3 (0.2)	-0.6 (0.2)
	10	4.1 (0.3)	0.1 (0.3)	1.8 (0.3)	1.1 (0.3)	1.1 (0.3)	-0.6 (0.2)	-1.0 (0.2)
2.0	0	0.6 (0.0)	0.3 (0.0)	1.3 (0.0)	0.4 (0.0)	1.3 (0.0)	1.5 (0.0)	1.4 (0.0)
	2	0.3 (0.1)	-0.5 (0.1)	0.5 (0.1)	-0.4 (0.1)	0.5 (0.1)	0.7 (0.1)	0.4 (0.1)
	5	1.0 (0.1)	0.0 (0.1)	1.0 (0.1)	0 (0.1)	1.0 (0.1)	1.1 (0.1)	0.2 (0.1)
	10	0.9 (0.1)	-0.2 (0.1)	0.9 (0.1)	-0.2 (0.1)	0.9 (0.1)	0.9 (0.1)	0.1 (0.1)
2.5	0	0.4 (0.0)	0.4 (0.0)	1.2 (0.0)	0.4 (0.0)	1.2 (0.0)	1.3 (0.0)	1.2 (0.0)
	2	-0.4 (0.1)	-0.4 (0.1)	-0.1 (0.0)	-0.4 (0.1)	-0.1 (0.0)	0.1 (0.1)	0.0 (0.1)
	5	0.4 (0.0)	0.4 (0.0)	0.4 (0.0)	0.4 (0.0)	0.4 (0.0)	0.5 (0.0)	0.4 (0.0)
	10	0.4 (0.0)	0.4 (0.0)	0.4 (0.0)	0.4 (0.0)	0.4 (0.0)	0.5 (0.0)	0.4 (0.0)
3.0	0	0.2 (0.0)	0.2 (0.0)	1.1 (0.0)	0.2 (0.0)	1.1 (0.0)	1.1 (0.0)	0.1 (0.0)
	2	-0.8 (0.0)	-0.8 (0.0)	-0.4 (0.0)	-0.8 (0.0)	-0.4 (0.0)	0.0 (0.1)	-0.3 (0.0)
	5	0.2 (0.0)	0.2 (0.0)	0.2 (0.0)	0.2 (0.0)	0.2 (0.0)	0.3 (0.0)	0.2 (0.0)
	10	0.2 (0.0)	0.2 (0.0)	0.2 (0.0)	0.2 (0.0)	0.2 (0.0)	0.3 (0.0)	0.2 (0.0)

FIGURE 1. Barplots of MSE and Bias. The numbers 2–7 in each plot refer to the column numbers associated with the estimators in Tables 1 and 2; i.e., “2” denotes $\bar{k}$, “3” denotes $\hat{k}$, and so on

5. CONCLUSION

We proved that two shrinkage estimators, $\hat{k}$ and $\hat{k}'$, possess reduced asymptotic mean-square errors compared to the FPE estimator $\tilde{k}$. Simulations show that in finite-sample situations, the amount of reduction depends on the size of the bias of $\tilde{k}$. Larger amounts of reduction are possible when the bias of $\tilde{k}$ is large. Iterating the shrinkage method by bootstrapping also appears to yield estimators with good finite-sample properties.

REFERENCES

1. P. Burman, *A comparative study of ordinary cross-validation, ν -fold cross-validation and the repeated learning-testing methods*, *Biometrika* **76** (1989), 503–514.
2. Y. S. Chow and H. Teicher, *Probability theory: independence, interchangeability, martingales*, 2nd ed., Springer-Verlag, New York, 1988.
3. G. A. F. Seber, *Linear regression analysis*, Wiley, New York, 1977.
4. J. Shao, *Linear model selection by cross-validation*, *J. Amer. Statist. Assoc.* **88** (1993), 486–494.
5. F. Spitzer, *A combinatorial lemma and its application to probability theory*, *Trans. Amer. Math. Soc.* **82** (1956), 323–339.
6. P. Zhang, *On the distributional properties of model selection criteria*, *J. Amer. Statist. Assoc.* **87** (1992), 732–737.

DEPARTMENT OF STATISTICS, UNIVERSITY OF WISCONSIN, MADISON, WISCONSIN 53706

E-mail address, W.-Y. Loh: loh@stat.wisc.edu

E-mail address, X. Zheng: zheng@stat.wisc.edu