Far East Journal of Theoretical Statistics
Volume 31, Number 2, 2010, Pages 117-133
Published Online: June 5, 2010
This paper is available online at http://pphmj.com/journals/fjts.htm

DISCRIMINATING BETWEEN NONNESTED MODELS

ERHARD RESCHENHOFER

Department of Statistics and Decision Support Systems
University of Vienna
A-1010 Vienna, Austria

Abstract

We propose various model selection methods that have been designed for the case of nonnested models. In contrast to conventional model selection criteria like *AIC* and *BIC*, which penalize only the number of explanatory variables actually included in the model, our methods take also the total number of available variables into account. We compare the performance of the different methods through simulation studies.

1. Introduction

Our methods for discriminating between nonnested models, which will be introduced in the next section, are modifications of conventional methods designed for nested models. It is, therefore, necessary to first give a short review of these conventional methods. For the discussion of popular model selection criteria like Mallows' C_p [8, 15], the final prediction error (*FPE*; [1, 20]), and Akaike's information criterion (*AIC*, [2]), which are closely related to each other, we assume that all competing models are submodels of a normal regression model

$$y = X\beta + u$$

satisfying $\mu = Ey = X\beta$ and $Var(y) = \sigma^2 I$. In the case of nested models, each submodel is characterized by an $n \times k$ submatrix X_k containing the first $k \leq K$

2010 Mathematics Subject Classification: 62F07, 62J05.

Keywords and phrases: model selection criteria, subset selection, mean squared error, mean squared prediction error.

Received March 12, 2010

columns $x_1, ..., x_k$ of the $n \times K$ matrix X. The submatrix X_k is used to estimate the mean μ by projecting y onto the subspace of $\mathbb{R}^n$ spanned by the columns of X_k. To assess the quality of this estimator, $P_k y$, we use the mean squared error

$$\begin{aligned} MSE(X_k) &= E\| P_k y - \mu \|^2 \\ &= E\| P_k (y - \mu) \|^2 + \| P_k y - \mu \|^2 \\ &= k\sigma^2 + \| \mu - P_k \mu \|^2. \end{aligned} \tag{1}$$

We want to select that model dimension k which gives the smallest mean squared error. The residual sum of squares is a biased estimator of the mean squared error, because

$$\begin{aligned} E\ RSS(X_k) &= E\| y - P_k y \|^2 \\ &= E\| (I - P_k)\mu \|^2 + \| (I - P_k)\mu \|^2 \\ &= (n - k)\sigma^2 + \| \mu - P_k \mu \|^2. \end{aligned} \tag{2}$$

Model selection by minimization of the unbiased estimator

$$C_p^*(X_k) = RSS(X_k) - (n - 2k)\hat{\sigma}_K^2 \tag{3}$$

of the mean squared error is equivalent to model selection by minimization of either

$$C_p(X_k) = \frac{RSS(X_k)}{\hat{\sigma}_K^2} - (n - 2k) \tag{4}$$

[8, 15] or

$$FPE^*(X_k) = RSS(X_k) + 2k\hat{\sigma}_K^2, \tag{5}$$

where $\hat{\sigma}_K^2 = RSS(X_k)/(n - K)$ is an unbiased estimator of σ^2. The statistic FPE^* is an unbiased estimator of the mean squared prediction error

$$\begin{aligned} MSPE(X_k) &= E\| z - P_k y \|^2 \\ &= E\| z - \mu \|^2 + \| \mu - P_k \mu \|^2 + E\| P_k (y - \mu) \|^2 \\ &= (n + k)\sigma^2 + \| \mu - P_k \mu \|^2, \end{aligned} \tag{6}$$

where z is an independent sample from the same distribution as y.

If X_k is correctly specified, i.e., if $E\,P_k y = \mu$, then not only $\hat{\sigma}_K^2$ but also $\hat{\sigma}_k^2 = RSS(X_k)/(n-k)$ will be an unbiased estimator of σ^2 and

$$FPE(X_k) = RSS(X_k) + 2k\hat{\sigma}_k^2 = RSS(X_k)\left(1 + \frac{2k}{n-k}\right) \tag{7}$$

[1, 20] will be another unbiased estimator of the mean squared prediction error. For large *n*, model selection by minimization of $FPE(X_k)$ is practically equivalent to model selection by minimization of

$$AIC(X_k) = -2\log L(y; P_k y, \hat{\sigma}_k^2) + 2(k+1), \tag{8}$$

[2], because

$$n\log\left(RSS(X_k)\left(1 + \frac{2k}{n-k}\right)\right) = n\log(RSS(X_k)) + \left(1 + \frac{2k}{n-k}\right)^n$$

$$\sim n\log(RSS(X_k)) + 2k \tag{9}$$

and $\log\hat{\sigma}_k^2$ is the only term in the maximum log likelihood

$$\log L(y; P_k y, \hat{\sigma}_k^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_k^2 - \frac{\| y - P_k y \|^2}{2\hat{\sigma}_k^2}$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\hat{\sigma}_k^2 - \frac{n}{2} \tag{10}$$

which depends on X_k. If X_k is misspecified, i.e., if $EP_k y \neq \mu$, then

$$AIC^Q(X_k) = -2\log L(y; P_k y, \hat{\sigma}_k^2) + 2(k+2)\hat{Q}_k - 2\hat{Q}_k^2, \tag{11}$$

[21], where $\hat{Q}_k = \hat{\sigma}_K^2/\hat{\sigma}_k^2$, is more appropriate for the estimation of the mean squared prediction error than *AIC*. A small-sample version of AIC^Q is given by

$$AIC_C^Q(X_k) = -2\log L(y; P_k y, \hat{\sigma}_k^2) + 2(k+2)\hat{Q}_k - 2\hat{Q}_k$$

$$+ \frac{2k^2\hat{Q}_k^2 + 14k\hat{Q}_k^2 - 8k\hat{Q}_k^3 + 24\hat{Q}_k^2 - 32\hat{Q}_k^3 + 12\hat{Q}_k^4}{n-k-2}, \tag{12}$$

[17]. For $n \to \infty$ this statistic reduces to AIC^Q, for $\hat{Q}_k = 1$ to the corrected *AIC*,

$$AIC_C(X_k) = -2\log L(y; P_k y, \hat{\sigma}_k^2) + 2(k+1) + \frac{2k^2 + 6k + 4}{n-k-2}, \tag{13}$$

[26], and for $n \to \infty$, $\hat{Q}_k = 1$ to *AIC*. In case of correct specification, $AIC_C(X_k)$ is an unbiased estimator of $E(-2\log L(z; P_k y, \hat{\sigma}_k^2))$, because

$$\begin{aligned}
&E(-2\log L(y; P_k y, \hat{\sigma}_k^2)) - E(-2\log L(z; P_k y, \hat{\sigma}_k^2)) \\
&= n - E\frac{\| z - P_k y \|^2}{\hat{\sigma}_k^2} \\
&= n - E\| z - P_k y \|^2 \frac{n}{\sigma^2} E\left(\frac{n\hat{\sigma}_k^2}{\sigma^2}\right)^{-1} \\
&= n - (n+k)\sigma^2 \frac{n}{\sigma^2}\frac{1}{n-k-2} \\
&= -2(k+1) - \frac{2k^2 + 6k + 4}{n-k-2}.
\end{aligned}$$

In the case of nested models, all K potential regressors are arranged in some natural order, hence there is only one model for each model dimension $k \le K$. A comparison of two different models is therefore equivalent to a comparison of two different model dimensions. For example, *AIC* prefers the unique model of dimension k_1 over the unique model of dimension $k_2 > k_1$ when the likelihood terms differ by less than $2(k_2 - k_1)$. In the case of nonnested models, the best among all models of dimension k_1 is compared to the best among all models of dimension k_2. It seems quite obvious that now possible differences between the sizes of the two model classes must also be taken into account. Clearly, the number of models of a certain dimension depends on the total number K of potential regressors. In the next section, we study model selection criteria whose penalty terms depend not only on the number of regressors actually included in the model but also on the total number of available regressors. Because it is easier to deal with misspecification when we use the residual sum of squared errors rather than the maximum log likelihood, we will focus on extensions of FPE^* and *FPE*, respectively. The performance of these extensions is examined in Section 3. Section 4 concludes.

2. Comparing Nonnested Models

In the case of nonnested models, there is more than one model of dimension k if $k < K$. Let $\boldsymbol{X}_k$ denote an $n \times k$ submatrix containing those k columns of X which minimize the residual sum of squares and let $\boldsymbol{\beta}_k$ denote the corresponding subvector of β. Under the additional assumption that the columns of X are orthonormal, i.e., $X'X = I$, the residual sum of squares, the expected value of the residual sum of squares, the mean squared error, and the mean squared prediction error of the apparently best model of dimension k are given by

$$\begin{aligned} RSS(\boldsymbol{X}_k) &= y'(I - \boldsymbol{X}_k(\boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}\boldsymbol{X}_k')\,y \\ &= y'y - Ey'\boldsymbol{X}_k\boldsymbol{X}_k'y \\ &= y'y - E\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k, \end{aligned} \tag{14}$$

$$E\,RSS(\boldsymbol{X}_k) = \mu'\mu + n\sigma^2 - E\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k, \tag{15}$$

$$\begin{aligned} MSE(\boldsymbol{X}_k) &= E(\mu - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k)'(\mu - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k) \\ &= \mu'\mu - 2E\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k + E\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k \\ &= \mu'\mu + 2(E(\hat{\boldsymbol{\beta}}_k' - \boldsymbol{\beta}_k')\hat{\boldsymbol{\beta}}_k) - E\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k \end{aligned} \tag{16}$$

and

$$\begin{aligned} MSPE(\boldsymbol{X}_k) &= E(z - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k)'(z - \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k) \\ &= \mu'\mu + n\sigma^2 + 2E(\hat{\boldsymbol{\beta}}_k' - \boldsymbol{\beta}_k')\hat{\boldsymbol{\beta}}_k - E\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k, \end{aligned} \tag{17}$$

respectively. For the estimation of

$$V(\boldsymbol{X}_k) = E(\hat{\boldsymbol{\beta}}_k' - \boldsymbol{\beta}_k')\hat{\boldsymbol{\beta}}_k \tag{18}$$

we have to impose further restrictions. Perhaps the simplest possibility is to assume that $\beta = 0$. Then $\hat{\beta} = X'y = X'u$ has a K-dimensional multivariate normal distribution with $E\hat{\beta} = 0$ and

$$Var(\hat{\beta}) = EX'uu'X = \sigma^2 I. \tag{19}$$

Hence,

$$V(\boldsymbol{X_k}) = E\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}'\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} = \sigma^2\zeta_1(k, K), \tag{20}$$

where $\zeta_1(k, K)$ is the expected value of the sum of the k largest of K independent $\chi^2(1)$-variables. Now the statistic

$$\hat{\sigma}_{\boldsymbol{k}}^2 = \frac{RSS(\boldsymbol{X_k})}{n - \zeta_1(k, K)} \tag{21}$$

is an unbiased estimator of σ^2 (because $\mu = 0$ if $\beta = 0$) and the statistics

$$\begin{aligned} FPE_{sub}(\boldsymbol{X_k}) &= RSS(\boldsymbol{X_k}) + 2\hat{\sigma}_{\boldsymbol{k}}^2\zeta_1(k, K) \\ &= RSS(\boldsymbol{X_k})\left(1 + \frac{2\zeta_1(k, K)}{n - \zeta_1(k, K)}\right) \end{aligned} \tag{22}$$

([18]; for related statistics see [7, 27]) and

$$FPE_{sub}^*(\boldsymbol{X_k}) = RSS(\boldsymbol{X_k}) + 2\hat{\sigma}_K^2\zeta_1(k, K) \tag{23}$$

are unbiased estimators of the mean squared prediction error. When we use FPE_{sub} or FPE_{sub}^* for the selection of a subset of regressors, we compare the jth largest squared estimate, $\hat{\beta}_{(j)}^2$, to its expected value, $E\hat{\beta}_{(j)}^2$. But when there are $k-1$ dominant regressors which are certain to be included, we actually compare the largest of the remaining $K-k+1$ values, $\hat{\beta}_{(k)}^2$, to the expected value of the kth largest of K independent $\chi^2(1)$-variables, which does not seem right. In such a case, it might be more appropriate to compare $\hat{\beta}_{(k)}^2$ to $\zeta_1(1, K-k+1)$ and consequently prefer model dimension k over model dimension $k-1$ only if the difference between the residual sums of squares is greater than $2\hat{\sigma}_K^2\zeta_1(1, K-k+1)$. Assuming that all parameters except the $k-1$ very large ones vanish we might estimate $V(\boldsymbol{X_k})$ by $\hat{\sigma}_K^2(k-1+\zeta_1(1, K-k+1))$. But the associated criterion

$$FPE_1^*(\boldsymbol{X_k}) = RSS(\boldsymbol{X_k}) + 2\hat{\sigma}_K^2(k-1+\zeta_1(1, K-k+1)) \tag{24}$$

is likely to select too large models, because it allows only for a fair comparison

between the model dimensions $k-1$ and k but not between $k-1$ and $k+1$, $k+2$, etc. However, in a stepwise approach, where for each k, the decision between $k-1$ and k is based on the assumption that $k-1$ regressors are certain to be included, the criterion

$$FPE_0^*(\boldsymbol{X}_{\boldsymbol{k}}) = RSS(\boldsymbol{X}_{\boldsymbol{k}}) + 2\hat{\sigma}_K^2 \sum_{j=1}^{k} \zeta_1(1, K-j+1) \tag{25}$$

(for related criteria see [6, 19]) would certainly make more sense (although there is hardly any situation where it is a meaningful estimator of the mean squared prediction error). For the case, where there are not only some large coefficients, but possibly also a number of nonvanishing coefficients of minor importance, we propose a two-step procedure (*FPE*2). In the first step, the first $k_0 \geq 0$ predictors are selected by minimization of $FPE_0^*(\boldsymbol{X}_{\boldsymbol{k}})$. In the second step, the final model dimension $k \geq k_0$ is selected by minimizing

$$RSS(\boldsymbol{X}_{\boldsymbol{k}}) + 2\hat{\sigma}_K^2 \zeta_1(k-k_0, K-k_0)$$

over $k = k_0, ..., K$, where $\zeta_1(0, K-k_0) = 0$.

The subset selection criteria discussed so far have two shortcomings. The first is that they have been derived under the assumption of orthogonality, which is rarely satisfied in practice. However, Reschenhofer et al. [19] tailored specific subset-selection criteria for a typical base set of non-orthogonal macroeconomic variables and found that these criteria are practically equivalent to analogous criteria derived under the assumption of orthogonality. The second shortcoming is that they involve moments of order statistics from a $\chi^2(1)$-distribution. Although explicit closed-form expressions for such moments have recently been derived by Nadarajah [16], the calculation of the required moments is still no easy task, because the closed-form expressions are sums of special functions (namely, Lauricella functions) that are not contained in standard packages. For convenience, we therefore provide a table of values of $\zeta_1(k, K)$ for $K = 1, ..., 50$ and $k = 1, ..., \min(K, 15)$ (see Table 1). Each value in this table is based on 100 million random samples of size K from a $\chi^2(1)$-distribution. Next, we propose a reverse subset selection approach which does not involve moments of order statistics. Instead of increasing the current model dimension k when $\hat{\beta}^2_{(k+1)}, \hat{\beta}^2_{(k+2)}, ...$ are too large, we decrease the dimension when

$\hat{\beta}^2_{(K)}, \hat{\beta}^2_{(K-1)}, \ldots$ are too small. If, in addition to k regressors with very large parameters, $K - k$ regressors with vanishing parameters are included in the model, the mean squared prediction error will increase by approximately $(K - k)\sigma^2$ and the expected value of the residual sum of squares will decrease by the same value. This change can be estimated either by the decrease in the residual sum of squares or by $(K-k)\hat{\sigma}^2_K$. A smaller value of the first estimator is an indication of overfitting. According to this reverse approach (*RA*), the largest k satisfying

$$RSS(\boldsymbol{X}_{\boldsymbol{k}-1}) - RSS(X) > (K-k+1)\hat{\sigma}^2_K$$

is selected.

3. Simulation Studies

Here we evaluate the finite-sample performance of various model selection criteria through simulation studies. We generate the data from the orthonormal regression model

$$\begin{aligned} y &= \beta_1 x_1 + \cdots + \beta_K x_K + u_t \\ &= C\rho \begin{pmatrix} \sqrt{\frac{2}{n}}\cos\left(\frac{2\pi}{n}1\right) \\ \vdots \\ \sqrt{\frac{2}{n}}\cos\left(\frac{2\pi}{n}n\right) \end{pmatrix} + \cdots + C\rho^K \begin{pmatrix} \sqrt{\frac{2}{n}}\cos\left(\frac{2\pi K}{n}1\right) \\ \vdots \\ \sqrt{\frac{2}{n}}\cos\left(\frac{2\pi K}{n}n\right) \end{pmatrix} + u_t, \end{aligned}$$

where $u_1, \ldots, u_n$ are i.i.d. $N(0, 1)$. We use the mean squared error $E\| \hat{\mu}(\hat{k}) - \mu \|^2$ to measure the performance of an estimator

$$\hat{\mu}(\hat{k}) = \hat{\beta}_{(1)} x_{(1)} + \cdots + \hat{\beta}_{(\hat{k})} x_{(\hat{k})} + u_t$$

of $\mu = Ey_t$, where $\hat{\beta}_{(1)}, \ldots, \hat{\beta}_{(\hat{k})}$ are the $\hat{k}$ apparently most significant coefficients, i.e.,

$$\hat{\beta}^2_{(1)} > \cdots > \hat{\beta}^2_{(\hat{k})} > \cdots > \hat{\beta}^2_{(K)},$$

and $\hat{k}$ is obtained with the help of a model selection criterion. For $K = 10$,

$\rho = 0.7, 0.9,$ and $C = 0, 0.5, 1, ..., 6, 7, ..., 10, 15, ..., 30$ we generated 100,000 samples y of size $n = 50, 500$ and calculated the value

$$\mu'\mu - 2\sum_{j=1}^{\hat{k}} \beta_{(j)}\hat{\beta}_{(j)} + \sum_{j=1}^{\hat{k}} \hat{\beta}_{(j)}^2$$

for each sample. The average over all 100,000 values is practically identical to the mean squared error of $\hat{\mu}(\hat{k})$. For the determination of $\hat{k}$ we used some of the model selection methods discussed in the previous sections, namely, *AIC*, FPE^*_{sub}, FPE^*_0, *FPE*2; and *RA*, as well as

$$BIC(X_k) = -2\log L(y; P_k y, \hat{\sigma}^2_K) + (k+1)\log(n)$$

[22]. Methods that are closely related to one of these six methods were not included in our simulation study. For example, *FPE*, C_p, AIC_C, AIC^Q and AIC^Q_C often select the same model as *AIC*, particularly when *n* is large. For the determination of the best subset of regressors, we always used the apparently best subset of each model dimension *k* and not just the first *k* regressors, even in case of *AIC* and *BIC*. Figure 1 (medium sample size: $n = 50,$ quickly decaying coefficients: $\rho = 0.7$), Figure 2 (large sample size: $n = 500,$ quickly decaying coefficients: $\rho = 0.7$), Figure 3 (medium sample size: $n = 50,$ slowly decaying coefficients: $\rho = 0.9$), and Figure 4 (large sample size: $n = 500,$ slowly decaying coefficients: $\rho = 0.9$) show the mean squared errors implied by the six model selection methods. Each figure contains two subfigures which display the results in absolute terms (upper subfigure) and relative to *BIC* (lower subfigure), respectively. Not surprisingly, *AIC* outperforms *BIC* in the case of large coefficients and *BIC* outperforms *AIC* in the case of small coefficients. *RA* is somewhere between *AIC* and *BIC*, but closer to *AIC* in the case of large coefficients. But what is much more interesting is the behavior of the three criteria involving moments of order statistics. Their performance relative to *BIC* is not just a monotonic function of the size of the coefficients. They can outperform *BIC* both in the case of small and large coefficients. A nice example is FPE^*_{sub} which is always among the best criteria except for a "relatively small" region of medium-sized coefficients. This does, of course, not mean that this criterion is optimal in some sense (see also the discussion in the next section). Our goal is only to add additional methods to the base set of model selection criteria, which contains just *AIC* and *BIC* in most applications.

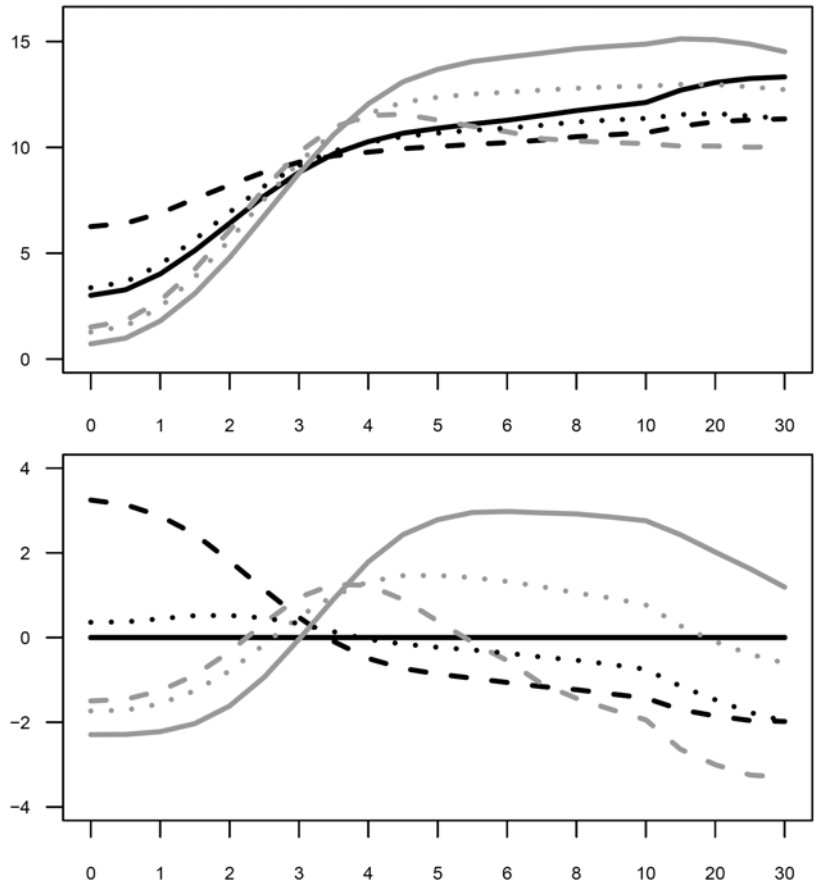

Figure 1. Mean squared errors of the methods *AIC* (dashed black line), FPE^*_{sub} (dashed gray line), FPE^*_0 (solid gray line), *FPE*2 (dotted gray line), *RA* (dotted black line), and *BIC* (solid black line) used for the selection of the best subset of a base set of $K = 10$ orthonormal regressors in a normal regression setting with $n = 50$, $\sigma^2 = 1$, $\beta_1 = C\rho, ..., \beta_K = C\rho^K$, $\rho = 0.7$ and $C = 0, 0.5, 1, ..., 6, 7, ..., 10, 15, ..., 30$. The upper figure displays the results in absolute terms and the lower figure relative to *BIC*.

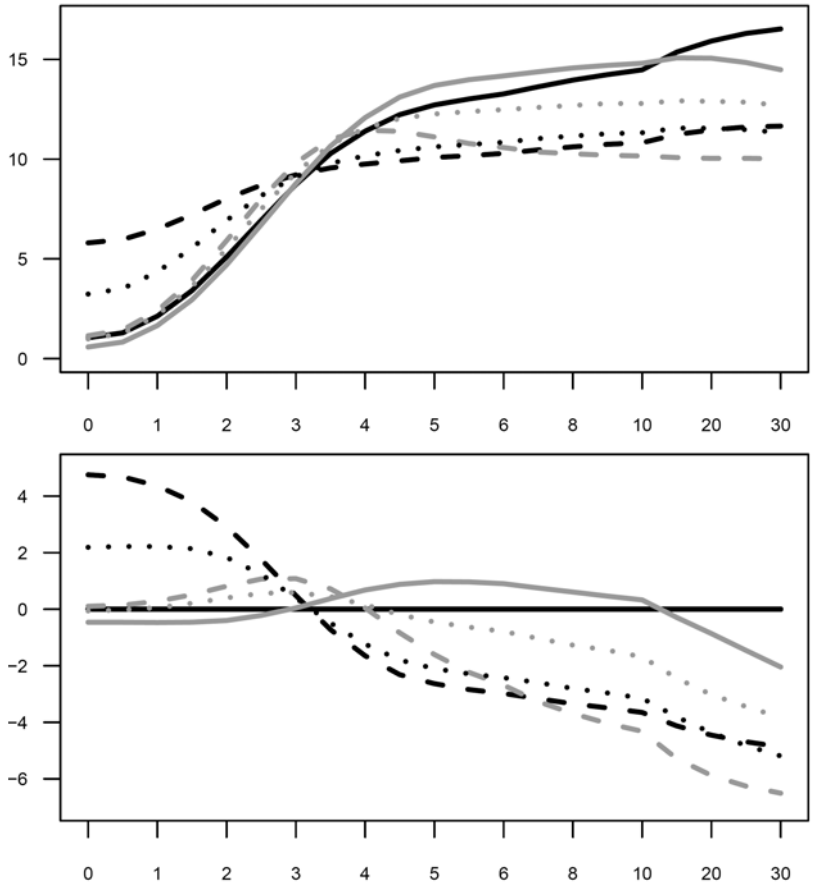

Figure 2. Mean squared errors of the methods *AIC* (dashed black line), FPE^*_{sub} (dashed gray line), FPE^*_0 (solid gray line), *FPE*2 (dotted gray line), *RA* (dotted black line), and *BIC* (solid black line) used for the selection of the best subset of a base set of $K = 10$ orthonormal regressors in a normal regression setting with $n = 500$, $\sigma^2 = 1$, $\beta_1 = C\rho, ..., \beta_K = C\rho^K$, $\rho = 0.7$ and $C = 0, 0.5, 1, ..., 6, 7, ..., 10, 15, ..., 30$. The upper figure displays the results in absolute terms and the lower figure relative to *BIC*.

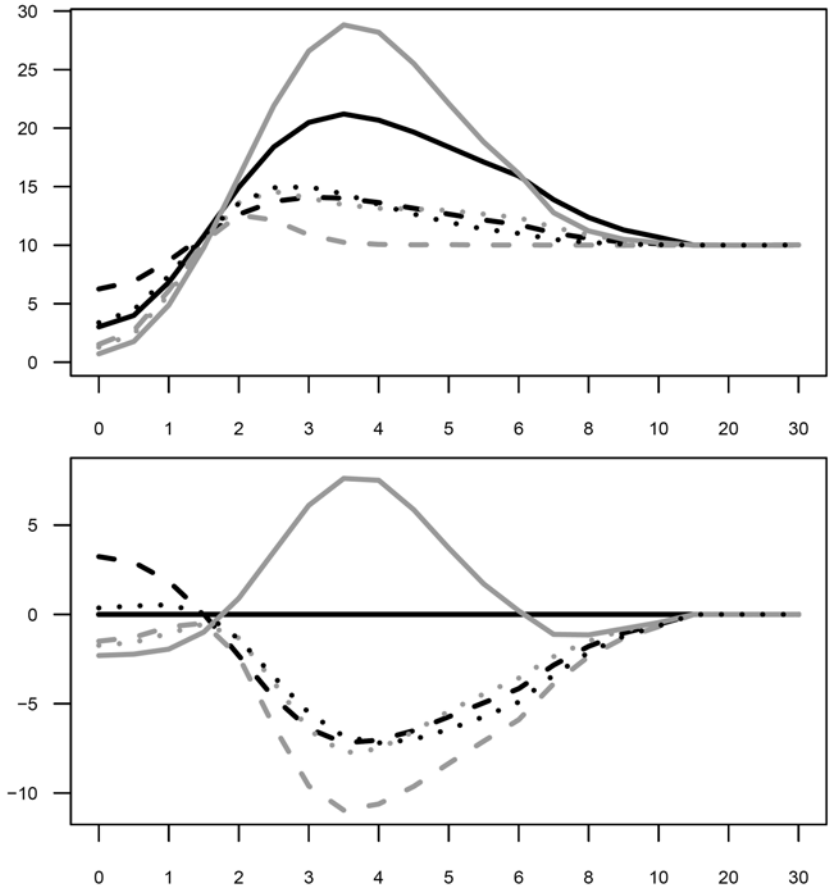

Figure 3. Mean squared errors of the methods *AIC* (dashed black line), FPE^*_{sub} (dashed gray line), FPE^*_0 (solid gray line), *FPE*2 (dotted gray line), *RA* (dotted black line), and *BIC* (solid black line) used for the selection of the best subset of a base set of $K = 10$ orthonormal regressors in a normal regression setting with $n = 50$, $\sigma^2 = 1$, $\beta_1 = C\rho, ..., \beta_K = C\rho^K$, $\rho = 0.9$ and $C = 0, 0.5, 1, ..., 6, 7, ..., 10, 15, ..., 30$. The upper figure displays the results in absolute terms and the lower figure relative to *BIC*.

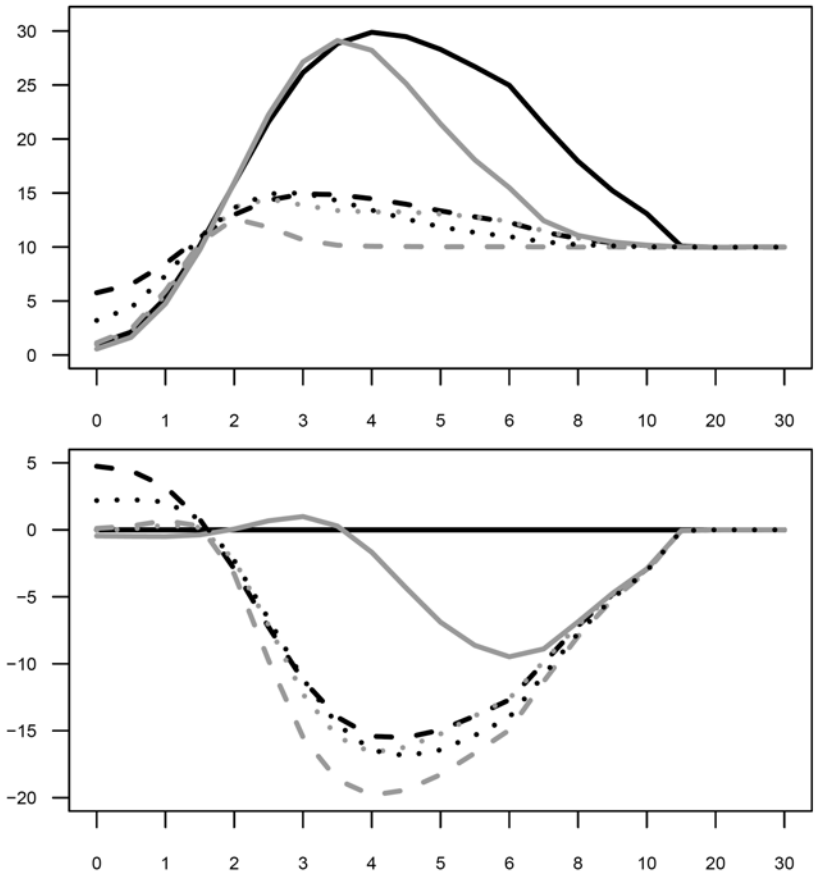

Figure 4. Mean squared errors of the methods *AIC* (dashed black line), FPE^*_{sub} (dashed gray line), FPE^*_0 (solid gray line), *FPE*2 (dotted gray line), *RA* (dotted black line), and *BIC* (solid black line) used for the selection of the best subset of a base set of $K = 10$ orthonormal regressors in a normal regression setting with $n = 500$, $\sigma^2 = 1$, $\beta_1 = C\rho, ..., \beta_K = C\rho^K$, $\rho = 0.9$ and $C = 0, 0.5, 1, ..., 6, 7, ..., 10, 15, ..., 30$. The upper figure displays the results in absolute terms and the lower figure relative to *BIC*.

4. Conclusion

Since Kempthorne [11] has shown that all post-model-selection estimators, which first select a model based on a given data set and then fit the selected model to

the same data set, are admissible for choosing among least-squares fits of a normal linear regression model, there cannot be anything like a universally best model selection method. However, there may be methods that are optimal in special situations. Shibata [24, 25] and Shao [23] argued that model selection criteria like *AIC* may be particularly useful when there exists no finite-dimensional true model. However, Kabaila [10] warned not to overinterpret Shibata's [24, 25] asymptotic results, because they hold only pointwise and may therefore be misleading. Moreover, Leeb [12] showed that generalized cross-validation

$$GCV(X_k) = \frac{1}{n} RSS(X_k)\left(1+\frac{k}{n-k}\right)^2 \tag{26}$$

(see [5]) and Tukey's

$$S_p(X_k) = \frac{1}{n} RSS(X_k)\left(1+\frac{k}{n-k}\right)\left(1+\frac{k}{n-1-k}\right) \tag{27}$$

(see [4, 9, 28]), which are practically equivalent to *AIC* if k is of smaller order than n, outperform *AIC* if the number of predictors is large. Yang [29, 30] questioned the simple story that *AIC* is good for infinite-dimensional models and *BIC* is good for finite-dimensional models and showed that no model selection criterion can share the main strengths of *AIC* and *BIC* simultaneously (see, e.g., also [3, 14]). In the light of the above discussion (for a more detailed account of the current state of the model selection problem, see [13]) it seems naïve to expect that any criterion could on purely theoretical grounds outperform all others. Of course, also simulation studies cannot prove the superiority of one criterion over another. Their outcomes depend largely on the respective model specifications (small vs. large sample size, nested vs. nonnested models, small vs. large dimension, slow vs. fast decay rate, deterministic vs. stochastic regressors, etc.) and can therefore easily be manipulated in any desired direction. However, in actual applications it is often quite easy to determine which criterion out of a base set of criteria is most appropriate for the problems at hand, particularly when n is large. So we only have to make sure that the base set contains a variety of criteria with diverse statistical properties. The model selection methods FPE^*_{sub}, FPE^*_0, *FPE*2; and *RA* proposed in this paper are certainly worthy candidates for inclusion in such a base set.

Table 1. The value in row *K* and column *k* is the expected value of the sum of the *k* largest of *K* independent and identically $\chi^2(1)$ -distributed random variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1														
2	1.64	2													
3	2.1	2.81	3												
4	2.47	3.47	3.88	4											
5	2.77	4.03	4.65	4.92	5										
6	3.03	4.51	5.32	5.75	5.94	6									
7	3.26	4.94	5.92	6.49	6.81	6.95	7								
8	3.46	5.32	6.46	7.17	7.61	7.85	7.96	8							
9	3.64	5.66	6.95	7.8	8.34	8.69	8.88	8.97	9						
10	3.8	5.97	7.4	8.37	9.03	9.47	9.74	9.9	9.98	10					
11	3.95	6.26	7.81	8.9	9.67	10.2	10.56	10.78	10.92	10.98	11				
12	4.09	6.52	8.19	9.39	10.26	10.88	11.33	11.62	11.82	11.93	11.98	12			
13	4.21	6.77	8.55	9.85	10.82	11.53	12.05	12.42	12.68	12.84	12.94	12.98	13		
14	4.33	7	8.89	10.29	11.34	12.14	12.74	13.18	13.5	13.72	13.86	13.95	13.99	14	
15	4.44	7.22	9.2	10.69	11.84	12.72	13.39	13.91	14.29	14.56	14.76	14.88	14.95	14.99	15
16	4.55	7.42	9.5	11.08	12.31	13.27	14.02	14.6	15.04	15.37	15.62	15.78	15.89	15.96	15.99
17	4.65	7.61	9.78	11.45	12.75	13.79	14.61	15.26	15.76	16.15	16.45	16.66	16.81	16.9	16.96
18	4.74	7.79	10.05	11.79	13.18	14.29	15.18	15.89	16.45	16.9	17.24	17.5	17.69	17.83	17.91
19	4.83	7.97	10.3	12.12	13.58	14.76	15.72	16.49	17.12	17.62	18.01	18.32	18.55	18.73	18.84
20	4.92	8.13	10.54	12.44	13.97	15.22	16.24	17.07	17.76	18.31	18.76	19.11	19.39	19.6	19.75
21	5	8.29	10.78	12.74	14.34	15.65	16.73	17.63	18.37	18.98	19.48	19.88	20.2	20.44	20.63
22	5.08	8.44	11	13.03	14.7	16.07	17.21	18.17	18.96	19.62	20.17	20.62	20.98	21.27	21.49
23	5.15	8.59	11.21	13.31	15.04	16.47	17.67	18.68	19.53	20.24	20.84	21.33	21.74	22.07	22.33
24	5.22	8.73	11.42	13.58	15.37	16.86	18.12	19.18	20.08	20.85	21.49	22.03	22.47	22.84	23.14
25	5.29	8.86	11.61	13.84	15.68	17.23	18.55	19.66	20.62	21.43	22.12	22.7	23.19	23.6	23.94
26	5.36	8.99	11.8	14.09	15.99	17.6	18.96	20.13	21.13	21.99	22.72	23.35	23.88	24.33	24.71
27	5.42	9.12	11.99	14.33	16.29	17.95	19.36	20.58	21.63	22.54	23.32	23.99	24.56	25.05	25.46
28	5.48	9.24	12.17	14.56	16.57	18.28	19.75	21.02	22.12	23.07	23.89	24.6	25.21	25.74	26.19
29	5.54	9.35	12.34	14.79	16.85	18.61	20.13	21.44	22.59	23.58	24.45	25.2	25.85	26.42	26.9
30	5.6	9.47	12.51	15.01	17.12	18.93	20.5	21.86	23.04	24.08	24.99	25.78	26.47	27.08	27.6
31	5.66	9.58	12.67	15.22	17.38	19.24	20.85	22.26	23.49	24.57	25.52	26.35	27.08	27.72	28.28
32	5.71	9.68	12.82	15.43	17.63	19.54	21.2	22.65	23.92	25.04	26.03	26.9	27.67	28.35	28.94
33	5.76	9.79	12.98	15.63	17.88	19.83	21.53	23.02	24.34	25.5	26.53	27.44	28.25	28.96	29.58
34	5.81	9.89	13.12	15.82	18.12	20.11	21.86	23.39	24.75	25.95	27.02	27.97	28.81	29.55	30.22
35	5.86	9.98	13.27	16.01	18.35	20.39	22.18	23.75	25.15	26.39	27.5	28.48	29.36	30.14	30.83
36	5.91	10.08	13.41	16.19	18.58	20.66	22.49	24.1	25.54	26.82	27.96	28.98	29.89	30.71	31.44
37	5.96	10.17	13.55	16.37	18.8	20.92	22.79	24.44	25.92	27.24	28.41	29.47	30.42	31.27	32.03
38	6	10.26	13.68	16.55	19.02	21.18	23.09	24.78	26.29	27.64	28.86	29.95	30.93	31.81	32.6
39	6.05	10.35	13.81	16.72	19.23	21.43	23.37	25.1	26.65	28.04	29.29	30.42	31.43	32.34	33.17
40	6.09	10.44	13.94	16.89	19.44	21.67	23.66	25.42	27.01	28.43	29.71	30.87	31.92	32.87	33.72
41	6.14	10.52	14.06	17.05	19.64	21.91	23.93	25.74	27.35	28.81	30.13	31.32	32.4	33.38	34.26
42	6.18	10.6	14.18	17.21	19.84	22.15	24.2	26.04	27.69	29.19	30.54	31.76	32.87	33.88	34.8
43	6.22	10.68	14.3	17.37	20.03	22.38	24.47	26.34	28.03	29.55	30.93	32.19	33.33	34.37	35.32
44	6.26	10.76	14.42	17.52	20.22	22.6	24.72	26.63	28.35	29.91	31.32	32.61	33.78	34.85	35.83
45	6.3	10.84	14.53	17.67	20.4	22.82	24.98	26.92	28.67	30.26	31.71	33.02	34.23	35.33	36.33
46	6.34	10.91	14.64	17.81	20.58	23.03	25.23	27.2	28.98	30.6	32.08	33.43	34.66	35.79	36.82
47	6.37	10.99	14.75	17.96	20.76	23.25	25.47	27.47	29.29	30.94	32.45	33.83	35.09	36.25	37.31
48	6.41	11.06	14.86	18.1	20.94	23.45	25.71	27.74	29.59	31.27	32.81	34.22	35.51	36.69	37.78
49	6.45	11.13	14.96	18.24	21.11	23.65	25.94	28.01	29.89	31.6	33.17	34.6	35.92	37.13	38.25
50	6.48	11.2	15.06	18.37	21.27	23.85	26.17	28.27	30.18	31.92	33.51	34.98	36.32	37.56	38.7

References

[1] H. Akaike, Fitting autoregressive models for prediction, Ann. Inst. Statist. Math. 21 (1969), 243-247.

[2] H. Akaike, Information theory and an extension of the maximum likelihood principle, B. N. Petrov and F. Csaki, eds., Second International Symposium on Information Theory, Akademia Kiado, Budapest, 1973, pp. 267-281.

[3] R. Beran, Comment on an asymptotic theory for linear model selection by J. Shao. Statist. Sinica 7 (1997), 243-248.

[4] L. Breiman and D. Freedman, How many variables should be entered in a regression equation? J. Amer. Statist. Assoc. 78 (1983), 131-136.

[5] P. Craven and G. Wahba, Smoothing noisy data with spline functions, Numer. Math. 31 (1978), 377-403.

[6] D. P. Foster and E. I. George, The risk inflation criterion for multiple regression, Ann. Statist. 22 (1994), 1947-1975.

[7] E. I. George and D. P. Foster, Calibration and empirical Bayes variable selection, Biometrika 87 (2000), 731-747.

[8] J. W. Gorman and R. J. Toman, Selection of variables for fitting equations to data, Technometrics 8 (1966), 27-51.

[9] R. R. Hocking, The analysis and selection of variables in linear regression, Biometrics 32 (1976), 1-49.

[10] P. Kabaila, On variable selection in linear regression, Econometric Theory 18 (2002), 913-925.

[11] P. J. Kempthorne, Admissible variable-selection procedures when fitting regression models by least squares for prediction, Biometrika 71 (1984), 593-597.

[12] H. Leeb, Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process, Bernoulli 14 (2008), 661-690.

[13] H. Leeb and B. M. Poetscher, Model selection, Handbook of Financial Time Series, 2008, pp. 888-925.

[14] H. Leeb and B. M. Poetscher, Sparse estimators and the oracle property, or the return of Hodges' estimator, J. Econometrics 142 (2008), 201-211.

[15] C. L. Mallows, Choosing variables in a linear regression: a graphical aid, Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 1964.

[16] S. Nadarajah, Explicit expressions for moments of $\chi 2$ order statistics, Bull. Inst. Math. Acad. Sin. (N.S.) 3 (2008), 433-444.

[17] E. Reschenhofer, Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification, Econometric Theory 15 (1999), 377-387.

[18] E. Reschenhofer, On subset selection and beyond, Adv. Appl. Stat. 4 (2004), 265-286.

[19] E. Reschenhofer, M. Schilde, E. Oberecker, E. Payr, H. Tandogan and L. Wakolbinger, Identifying the determinants of foreign direct investment: a data-specific model selection approach, 2009, submitted.

[20] D. Rothman, Letter to the editor, Technometrics 10 (1968), 432.

[21] T. Sawa, Information criteria for discriminating among alternative regression models, Econometrica 46 (1978), 1273-1291.

[22] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978), 461-464.

[23] J. Shao, An asymptotic theory for linear model selection, Statist. Sinica 7 (1997), 221-242.

[24] R. Shibata, Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, Ann. Statist. 8 (1980), 147-164.

[25] R. Shibata, An optimal selection of regression variables, Biometrika 68 (1981), 45-54.

[26] N. Sugiura, Further analysts of the data by Akaike's information criterion and the finite corrections, Comm. Statist. Theory Methods 7 (1978), 13-26.

[27] R. Tibshirani and K. Knight, The covariance inflation criterion for adaptive model selection, J. R. Stat. Soc. Ser. B Stat. Methodol. 61 (1999), 529-546.

[28] J. W. Tukey, Discussion of topics in the investigation of linear relations fitted by the method of least squares by F. J. Anscombe, J. Roy. Statist. Soc. Ser. B 29 (1967), 47-48.

[29] Y. Yang, Can the strengths of *AIC* and *BIC* be shared? A conflict between model identification and regression estimation, Biometrika 92 (2005), 937-950.

[30] Y. Yang, Prediction/estimation with simple linear models: Is it really that simple? Econometric Theory 23 (2006), 1-36.