Adv. & Appl. in Stat. **4**(3) (2004), 265-286

ON SUBSET SELECTION AND BEYOND

ERHARD RESCHENHOFER

Department of Statistics and Decision Support Systems
University of Vienna, Universitätsstr. 5, A-1010 Vienna, Austria

Abstract

This paper considers methods of variable selection that are based on common statistics like the adjusted R-squared statistic, the t-statistic, or the F-statistic and proposes various modifications for the case of non-nested models. The resulting model selection criteria are somehow related to the risk inflation criteria proposed by Foster and George [Ann. Statist. 22 (1994), 1947-1975] and George and Foster [Biometrika 87 (2000), 731-747]. Next, the final prediction error criterion is modified in a similar way so that it can also be used for subset selection. Finally, a universal modeling procedure is discussed that can be used for the simultaneous selection of the model class, the criterion for variable selection, and the method for the estimation of the model parameters.

1. Introduction

According to Kempthorne [6] there is no objectively optimal variable-selection procedure. Suppose the data $y = (y_1, ..., y_n)^T$ follow a normal linear regression

$$y = X\beta + \varepsilon, \tag{1}$$

where the columns $x(1), ..., x(k)$ of the nonstochastic $(n \times k)$-matrix X are linearly independent, $\beta \in \Re^k$ is the vector of regression coefficients, and $\varepsilon_1, ..., \varepsilon_n$ are independent and identically distributed (i.i.d.) normal

2000 Mathematics Subject Classification: 62F07, 62J05.

Key words and phrases: AIC, model selection criteria, risk inflation criteria.

Received August 26, 2004

random variables with mean zero and variance σ^2. Then Kempthorne's [6] Theorem 1 states that for any two-stage estimator

$$y \to S \subseteq \{x(1), ..., x(n)\} \to \hat{\mu}(S) = P_{[S]}(y)$$

of $\mu = E(y) = X\beta$ that (i) selects a subset of regressors and (ii) estimates μ by projecting y onto the subspace $[S]$ spanned by the selected regressors there does not exist another two-stage estimator $\hat{\mu}(S^*)$ such that

$$E\| \hat{\mu}(S^*) - \mu \|^2 \leq E\| \hat{\mu}(S) - \mu \|^2 \text{ for all } \beta, \sigma^2$$

and

$$E\| \hat{\mu}(S^*) - \mu \|^2 < E\| \hat{\mu}(S) - \mu \|^2 \text{ for some } \beta, \sigma^2.$$

Hence all variable-selection procedures are admissible, no procedure can be uniformly better than any other procedure. The admissibility of variable-selection procedures was first investigated by Stone [14], who proved the admissibility of the trivial procedure which always selects all regressors. Kabaila [4] extended Kempthorne's [6] result to the case of the misspecified normal regression model

$$y = X\beta + \eta + \varepsilon,$$

where η is an unknown vector which is orthogonal to the subspace spanned by the columns of X.

In a certain contrast to Kempthorne's [6] and Kabaila's [4] finite-sample results, Shibata's [13] Theorem 2.2 states that variable selection with Akaike's [1] information criterion (AIC) is asymptotically optimal. To understand the exact meaning of this result we consider the trigonometric regression

$$y_t = \mu_t + \varepsilon_t = \sum_{j=1}^{\infty} \beta_j \cos(\omega_t(j-1))/j + \varepsilon_t, \ \omega_t = 2\pi(t-1)/n, \ \sum \beta_j^2 < \infty.$$

Given observations $y = (y_1, ..., y_n)^T$ we estimate $\mu = (\mu_1, ..., \mu_n)^T$ by selecting $k \in \{1, ..., K\}$ and projecting y onto the subspace spanned by the first k regressors. If there are infinitely many nonzero regression

coefficients β_j and $K \le n$, $K \to \infty$, $K = o(n)$, the assumptions of Shibata's Theorem 2.2 are satisfied and we can then conclude that the value of k determined by AIC is asymptotically optimal in the sense that

$$\operatorname*{p\,lim}_{n\to\infty} \| \hat{\mu}(k_{\mathrm{AIC}}) - \mu \|^2 / E\| \hat{\mu}(k^*) - \mu \|^2 = 1 \text{ for all } \beta, \sigma^2,$$

where k^* is an element of $\{1, ..., K\}$ that minimizes $E\| \hat{\mu}(k) - \mu \|^2$. Determination of the number of regressors by minimization of

$$n \log \hat{\sigma}^2(k) + 2k \qquad \text{(AIC)}$$

is not the only possibility to achieve asymptotic optimality, we may as well minimize closely related criteria like

$$n\hat{\sigma}^2(k)\left(1 + \frac{2k}{n}\right) = \hat{\sigma}^2(k)(n + 2k).$$

In each case, the residual variance

$$\hat{\sigma}^2(k) = \frac{1}{n} \| y - \hat{\mu}(k) \|^2$$

measures the goodness of fit and the term $2k$ serves to penalize overfitting. It is a quite remarkable fact that a penalty term of the form $(2 + \delta)k$ implies asymptotic optimality if and only if $\delta = 0$ (Shibata [12, 13]). An immediate consequence of this fact is that the Bayesian information criterion (BIC; Schwarz [11])

$$n \log \hat{\sigma}^2(k) + (\log n)k \qquad \text{(BIC)}$$

is not asymptotically optimal.

By imposing extremely restrictive conditions on the number of models under consideration Shibata [13] was able to prove the asymptotic optimality of AIC even in the case of non-nested models, where there is no natural ordering of regressors and therefore the selection of arbitrary subsets is possible. From an intuitive point of view it is hard to understand why one and the same variable-selection method, namely AIC, should be optimal both in the case of nested models and in the case of subset selection. Consider, for example, the case of two regressors x

and z. In the case of nested models, AIC prefers the full model to a model with only one regressor if

$$n \log \hat{\sigma}^2(x, z) + 4 < n \log \hat{\sigma}^2(x) + 2,$$

whereas in the case of subset selection AIC prefers the full model to a model with only one regressor if

$$n \log \hat{\sigma}^2(x, z) + 4 < n \log \min\{\hat{\sigma}^2(x), \hat{\sigma}^2(z)\} + 2.$$

Using the same penalty term for the smaller model in both cases just does not seem to be appropriate.

Kabaila [5] stressed that Shibata's [12, 13] asymptotic results are pointwise in the space of data generating mechanisms, which are typically specified by β, σ^2, and may therefore be misleading. He considered the class of trigonometric regressions and compared the performance of AIC and BIC for each fixed n and each pair (β, σ^2). He found that AIC is not better than BIC no matter how large n is. However, Kabaila [5] did not rule out that it might be possible to prove the superiority of AIC over BIC in large samples by imposing further restrictions on the rate of decline of the regression coefficients (see also Ploberger and Phillips [7]). But since it is always very difficult to verify assumptions about the rate of decline, practitioners should not expect too much from such a result.

Thus it seems that, at least for the time being, we must manage to live without an "optimal" variable-selection method and be content with methods that perform well in a broad class of data generating mechanisms. Recent research on subset selection has tended to focus on methods that penalize each regressor in a different, possibly data dependent way. In certain situations these new methods apparently outperform conventional criteria like AIC and BIC, whose penalty terms are just constants multiplied by the number of regressors. In the following sections, we discuss the pros and cons of some of these methods and propose a number of modifications. In Section 2, we discuss simple model selection criteria that are based on common statistics like $\overline{R}^2$, t,

and F and propose modifications for the case of subset selection. In Section 3, we also tune more popular criteria like the final prediction error criterion (FPE) for the case of subset selection. Section 4 concludes. Finally, since the penalty terms of the subset criteria are derived under the assumption that all regression coefficients are zero, we describe in the Appendix a universal, data-dependent variable-selection method that is not based on implausible assumptions. This method can also be used for the simultaneous selection of a suitable model class and an appropriate method for the estimation of the model parameters.

2. Model Selection with Common Statistics

The ordinary R-squared (R^2) statistic is a widely used measure of the success of a regression model $y = X\beta + \varepsilon$. It is defined as the fraction of the variance of the dependent variable y explained by the independent variables $x(1), ..., x(k)$. Assuming that all occurring variables are centered, we can write this statistic as

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y} = \frac{y'y - \hat{\varepsilon}'\hat{\varepsilon}}{y'y} = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y}.$$

Clearly, R^2 can never fall and generally increases, if another regressor is added to the model. However, if we modify this statistic by replacing the sum of squared errors (SSE) $\hat{\varepsilon}'\hat{\varepsilon}$, which is a biased estimator of $n\sigma^2$, by the unbiased estimator $\frac{n}{n-k}\hat{\varepsilon}'\hat{\varepsilon}$, this is no longer the case. The adjusted R-squared statistic

$$\overline{R}^2(k) = 1 - \frac{n}{n-k}\frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y} = 1 - \frac{n}{n-k}(1 - R^2) = R^2 - (1 - R^2)\frac{k}{n-k}$$

can decrease as regressors are added. The selection of the dimension k of a model by maximization of $\overline{R}^2(k)$ is equivalent to the selection of k by minimization of

$$\frac{n}{n-k}\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} = \left(1 + \frac{k}{n-k}\right)\hat{\sigma}^2(k)$$

or, equivalently, by minimization of

$$n \log(\hat{\sigma}^2(k)) + n \log\left(1 + \frac{k}{n-k}\right). \qquad (\overline{R}^2 - C)$$

The addition of a new variable $x(j)$ to a set of $j-1$ regressors $x(1), \ldots, x(j-1)$ will cause the adjusted R-squared statistic to rise if and only if

$$\overline{R}^2(j) = 1 - \frac{n}{n-j}\frac{\mathrm{SSE}(j)}{y'y} > \overline{R}^2(j-1) = 1 - \frac{n}{n-(j-1)}\frac{\mathrm{SSE}(j-1)}{y'y}$$

$$\Leftrightarrow \mathrm{SSE}(j-1) > \left(1 + \frac{1}{n-j}\right)\mathrm{SSE}(j)$$

$$\Leftrightarrow t^2 = (n-j)\frac{\mathrm{SSE}(j-1) - \mathrm{SSE}(j)}{\mathrm{SSE}(j)} > 1,$$

where t is just the common t-statistic used for testing the marginal contribution of a single variable. If the regression coefficient under consideration is zero and n is large, then the mean of this statistic will be close to 1. Hence, the condition $\overline{R}^2(j) > \overline{R}^2(j-1)$ for the inclusion of the regressor $x(j)$ is approximately equivalent to the condition that its squared t-statistic exceeds the expected value $Et^2(n-j)$.

Analogously, for the addition of k new variables $x(j-k+1), \ldots, x(j)$ to a set of $j-k$ regressors $x(1), \ldots, x(j-k)$ it could be required that the condition

$$F = \frac{\mathrm{SSE}(j-k) - \mathrm{SSE}(j)}{\mathrm{SSE}(j)}\frac{n-j}{k} > EF(k, n-j) = \frac{n-j}{n-j-2}$$

is satisfied. This condition is equivalent to

$$\hat{\sigma}^2(j-k) > \hat{\sigma}^2(j)\left(1 + \frac{k}{n-j-2}\right).$$

At first glance it seems that this condition might not always imply an "optimal" number of regressors. But this is not the case. For example, suppose that

$$i < j < k,\ \hat{\sigma}^2(i) > \hat{\sigma}^2(j)\left(1 + \frac{j-i}{n-j-2}\right),\ \text{and}\ \hat{\sigma}^2(j) > \hat{\sigma}^2(k)\left(1 + \frac{k-j}{n-k-2}\right).$$

Then

$$\hat{\sigma}^2(i) > \hat{\sigma}^2(j)\left(1 + \frac{j-i}{n-j-2}\right) > \hat{\sigma}^2(k)\left(1 + \frac{k-j}{n-k-2}\right)\left(1 + \frac{j-i}{n-j-2}\right)$$

$$= \hat{\sigma}^2(k)\left(1 + \frac{k-i}{n-k-2}\right).$$

Indeed, it follows from

$$\hat{\sigma}^2(i) > \hat{\sigma}^2(j)\left(1 + \frac{j-i}{n-j-2}\right) \Leftrightarrow \log(\hat{\sigma}^2(i)) > \log(\hat{\sigma}^2(j)) + \log\left(\frac{n-2-i}{n-2-j}\right)$$

$$\Leftrightarrow \log(\hat{\sigma}^2(i)) - \log(n-2-i)$$

$$> \log(\hat{\sigma}^2(j)) - \log(n-2-j)$$

that the number of regressors is optimal according to the criterion based on the F-statistic, which minimizes

$$\log(\hat{\sigma}^2(k)) - \log(n-2-k)$$

or, equivalently,

$$n[\log(\hat{\sigma}^2(k)) + \log(n-2) - \log(n-2-k)]$$

$$= n\log(\hat{\sigma}^2(k)) + n\log\left(1 + \frac{k}{n-2-k}\right), \qquad (F^* - C)$$

where $k/(n-2-k)$ is just the expected value of the statistic

$$F^* = \frac{k}{n-k}F = \frac{\mathrm{SSE}(0) - \mathrm{SSE}(k)}{\mathrm{SSE}(k)}.$$

If k is small compared to n, then the criteria $\overline{R}^2 - C$ and $F^* - C$ are approximately equivalent to the criterion

$$n\log(\hat{\sigma}^2(k)) + k, \qquad (C1)$$

which assigns the fixed penalty 1 to each regressor. Of course, these three criteria will differ dramatically if k is large. Although there is no justification whatsoever for the use of fixed penalty criteria like $C1$ or AIC in the case of large values of k (see also Section 3), these criteria are generally used as benchmarks in comparisons of different model selection

criteria. This is the case even when the number of regressors is close to the number of observations (see, e.g., George and Foster [3]).

Now suppose that the k explanatory variables to be possibly included in a linear regression model are not given but can be selected from a set $\{x(1), ..., x(k), ..., x(K)\}$ of size $K \geq k$. If $k = 1$, then the obvious candidate for inclusion is the regressor with the largest t-value. Analogously, if $k > 1$, then we might wish to select that k-dimensional subset implying the highest F-value. Unfortunately, the problem of finding this set quickly becomes computationally intractable as k and K increase. In the following, we will therefore assume that the K regressors are orthogonal. In this setup, the subset implying the highest F-value just consists of those regressors with the highest t-values, hence there is no need to examine all possible subsets of size k. However, there is still the problem of determining k. For this purpose, we may use for each k the expected value $e^*(k, K, n)$ (under the assumption that all regression coefficients are zero) of the highest F^*-value as a benchmark and accordingly formulate a subset version of $F^* - C$ as

$$n \log(\hat{\sigma}^2(k)) + n \log(1 + e^*(k, K, n)). \qquad (F^*_{\text{sub}} - C)$$

Since $e^*(k, K, n)$ coincides with the expected value of the sum of the k largest t-statistics divided by $n - k$, the criterion $F_{\text{sub}} - C$ is somehow related to the risk inflation criterion (RIC), which penalizes the inclusion of k regressors with k times $2 \log(K)$. The latter criterion has been motivated initially by minimaxity considerations (see Foster and George [2]) and later by considering the term $2 \log(K)$ as an approximation of the expected value of the maximum of K squared t-statistics (see George and Foster [3]). For $K = 10, 20, ..., 250$ and $n = 1.1K, 2K$, Figure 1 compares the actual expected value of the maximum of K t-statistics (with $n - 1$ degrees of freedom) obtained from K orthogonal regressors with the approximation $2 \log(K)$. In general, this approximation is not very good. For large values of n, the expected value of the maximum t-statistic is close to the expected value of the maximum of K independent

$\chi^2(1)$-statistics. The latter quantity can be approximated reasonably well by $2\log(K) - \log(\log(K))$. We might therefore find it appealing to replace the term $2\log(k)$ occurring in RIC by $2\log(K) - \log(\log(K))$. An obvious disadvantage of any version of RIC is that it compares not only the largest squared t-statistic with an approximation of the expected value of the maximum but all the other t-statistics as well. Therefore, George and Foster [3] proposed to replace the penalty term $2k\log(K)$ for a model with $k \le K$ regressors by

$$2\sum_{j=1}^{k}\log(K/j).$$

The corresponding criterion, the modified risk inflation criterion (MRIC), compares the jth largest squared t-statistic with the approximation $2\log(K/j)$ of its expected value. For $K = 250$, $k = 1, 2, \ldots, K$, and $n = 1.1K, 2K$, Figure 2 compares the expected value of the sum of the k largest t-statistics (with $n - k$ degrees of freedom) with the approximation $2\sum_{j=1..k}\log(K/j)$. Again, the quality of this approximation strongly depends on the sample size n. As n increases the expected value of the sum of the k largest t-statistics approaches the expected value of the sum of the k largest of K independent $\chi^2(1)$-statistics. In the following, we denote the latter expected value by $\varsigma(k, K)$. The discrepancy between the approximation $2\sum_{j=1..k}\log(K/j)$ and $\varsigma(k, K)$ increases as k increases. In the extreme case, where $k = K = n$, $\varsigma(k, K)$ is just the expected value of the sum of n independent $\chi^2(1)$-statistics and is therefore approximately only half as large as

$$2\sum_{j=1..k}\log(K/j) = -2\log\left(\frac{n!}{n^n}\right) = -2\log\left(\frac{n!}{n^n e^{-n}\sqrt{n}} - n + \frac{1}{2}\log(n)\right)$$

$$\approx 2n - \log(n) - 2\sqrt{2\pi}.$$

Anyhow, when n is much larger than K, we might wish to approximate the term $e^*(k, K, n)$ occurring in $F^*_{\text{sub}} - C$ by $\varsigma(k, K)/(n - k)$. Of course,

we cannot expect that the resulting criterion

$$n \log(\hat{\sigma}^2(k)) + n \log\left(1 + \frac{\varsigma(k,\ K)}{n - k}\right) \qquad (\text{Chi}^*_{\text{sub}} - C)$$

or $F^*_{\text{sub}} - C$ itself are generally better than MRIC (see the discussion in Section 1), but once we believe in the rationale behind a certain approach, we should try to translate it into a concrete criterion as accurately as possible.

Anyhow, this seems to be an appropriate moment for a reality check. We have started this section with comparing common statistics to their expected values under the null hypothesis and are now fiddling around with odd criteria involving ordered $\chi^2(1)$-statistics. What exactly are we trying to estimate? In the next section, we will look at the problem of subset selection from a different angle. Taking a more orthodox and less heuristic approach, we will focus on the unbiased estimation of the squared prediction error. Eventually, we will end up with a similar criterion, which again involves $\varsigma(k,\ K)$.

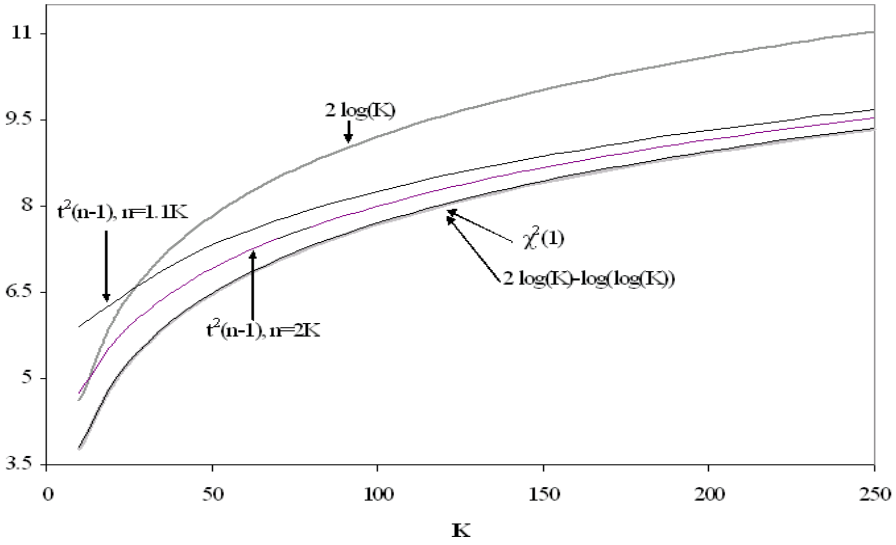

Figure 1. Expected values of the maximum of K squared $t(n-1)$-statistics (where $n = 1.1K,\ 2K$) and K independent $\chi^2(1)$-statistics, respectively, together with two approximations

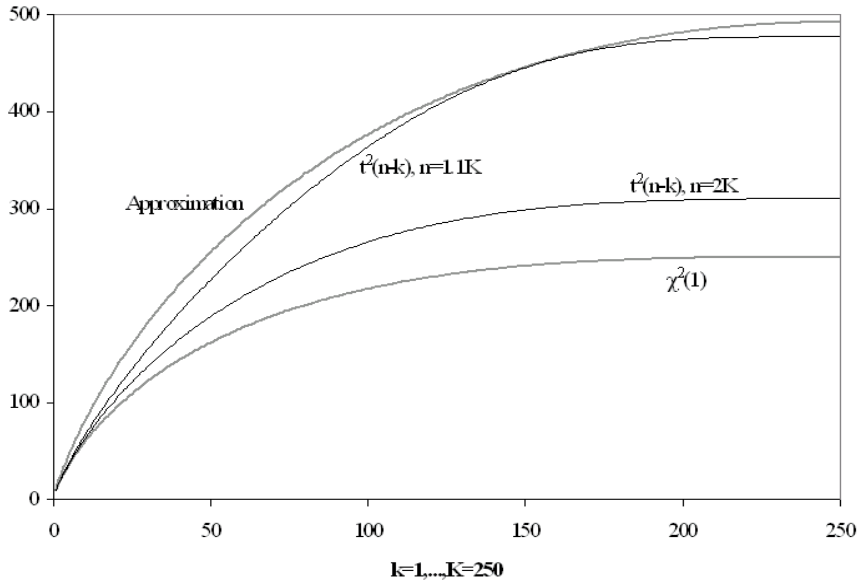

Figure 2. Expected values of the sum of the k largest of $K = 250$ squared $t(n-k)$-statistics (where $n = 1.1K, 2K$) and of the sum of the k largest of K independent $\chi^2(1)$-statistics, respectively, together with the approximation $2\log(K/1) + \cdots + 2\log(K/k)$

3. Model Selection Based on the Prediction Error

In practice, criteria like the adjusted R-squared statistic are hardly ever used for the determination of the number of regressors in the case of nested models. Applied workers prefer more conservative criteria like FPE, AIC, or even BIC. Since the criteria for subset selection discussed in the previous section may be considered as modifications of the adjusted R-squared statistic, we may suspect that they will also not become very popular. In this section, we will therefore try to tune a suitable version of FPE for the case of subset selection. We start with a short discussion of AIC and related criteria (including FPE).

In the classical linear regression model

$$y \sim N(X\beta, \sigma^2 I)$$

the AIC can be interpreted as an estimator of the expected Kullback-

Leibler discrepancy

$$D = -2E \log(f(y^*;\ X\hat{\beta},\ \hat{\sigma}^2 I)),$$

where $\hat{\beta} = (X'X)^{-1}X'y$, $\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta})$, and y^* is a second sample from $N(X\beta,\ \sigma^2 I)$, which is independent of y. Indeed, the AIC-statistic

$$-2 \log(f(y;\ X\hat{\beta},\ \hat{\sigma}^2 I)) + 2(k+1),$$

which differs from $-2n \log(\hat{\sigma}^2) + 2k$ just by a constant not depending on the rank k of X, is an asymptotically unbiased estimator for D provided that the model is correctly specified. Unfortunately, the finite sample-bias of the AIC-statistic will only be small, if k is small in relation to the sample size n. Whenever this is not the case, Sugiura's [15] corrected AIC

$$-2 \log(f(y;\ X\hat{\beta},\ \hat{\sigma}^2(k) I)) + 2(k+1) + \frac{2k^2 + 6k + 4}{n - k - 2}, \qquad (\mathrm{AIC}_C)$$

which is an unbiased estimator of D, should be used instead. Clearly, both the AIC and the AIC_C will be severely biased, if the model is misspecified. Extensions of the AIC and the AIC_C to the case of misspecified models were proposed by Sawa [10] and Reschenhofer [9], respectively. But while these improved criteria in general are superior estimators for the expected Kullback-Leibler discrepancy, their model selection properties are typically almost identical to those of the conventional, severely biased criteria (see Reschenhofer [8]). This is mainly due to the fact that in those cases, where the bias of the conventional criteria is large, the maximum likelihood term is decisive anyhow.

If k is large, then AIC will not only differ significantly from AIC_C but also from the final prediction error criterion

$$\frac{n+k}{n-k}\hat{\sigma}^2(k) = \left(1 + \frac{2k}{n-k}\right)\hat{\sigma}^2(k),$$

which is an unbiased estimator for the mean squared prediction error

$$\sigma_p^2(k) = \frac{1}{n} E(y^* - X\hat{\beta})^T (y^* - X\hat{\beta}),$$

if the regression model is correctly specified. The minimization of the FPE criterion is equivalent to the minimization of

$$n \log(\hat{\sigma}^2(k)) + n \log\left(1 + \frac{2k}{n-k}\right). \qquad \text{(FPE)}$$

For $k = 1, ..., K = 250$, $n = 1.1K, 2K$, Figure 3a compares the AIC-penalty term with the penalty terms of AIC_C and FPE, respectively. For larger values of k, AIC can obviously neither be regarded as a meaningful estimator for the Kullback-Leibler discrepancy nor for the log prediction error. Figure 3a also shows that similarly dramatic differences occur between the different versions of the $\overline{R}^2$-criterion.

For the discussion of the case of subset selection, we consider the simple orthogonal set-up used by George and Foster [3]. There we have $X = I$ so that $y = X\beta + \varepsilon$ reduces to $y = \beta + \varepsilon$. If k of the first K regressors are to be selected, then we choose the j-th regressor if y_j^2 is among the k largest of the squared observations $y_1^2, ..., y_K^2$. The parameter β_j is estimated by $\hat{\beta}_j = y_j$ if the j-th regressor is selected and by $\hat{\beta}_j = 0$ otherwise. The error variance is estimated by

$$\hat{\sigma}^2(k) = \frac{1}{n}\left(\sum_{j=k+1}^{K} y_{(j)}^2 + \sum_{j=K+1}^{n} y_j^2\right),$$

where $y_{(j)}^2$ denotes the j-th largest of the K values $y_1^2, ..., y_K^2$. Under the hypothesis that all β_j are zero, the expected value of this estimator is given by

$$E\hat{\sigma}^2(k) = \frac{\sigma^2}{n} E\left(\sum_{j=k+1}^{n} \frac{y_{(j)}^2}{\sigma^2} + \sum_{j=K+1}^{n} \frac{y_j^2}{\sigma^2}\right)$$

$$= \frac{\sigma^2}{n}\left(\sum_{j=1}^{n} \frac{y_j^2}{\sigma^2} - E\sum_{j=1}^{k} \frac{y_{(j)}^2}{\sigma^2}\right) = \sigma^2 \frac{n - \varsigma(k, K)}{n}.$$

Taking the logarithm of the unbiased estimator

$$\hat{\sigma}^2(k)\frac{n}{n-\varsigma(k,\,K)}$$

and multiplying by n, we obtain the criterion

$$n\log(\hat{\sigma}^2(k)) + n\log\left(1+\frac{\varsigma(k,\,K)}{n-\varsigma(k,\,K)}\right), \qquad (\mathrm{Chi}^{**}_{\mathrm{sub}} - C)$$

which looks very similar to $\mathrm{Chi}^{*}_{\mathrm{sub}} - C$, but is actually much more conservative because $\varsigma(k,\,K)$ is typically much greater than k (see also Figure 3b). Nevertheless, this criterion must still be considered as a subset version of $\overline{R}^2$ or an improved version of MRIC rather than as a subset version of AIC or FPE. However, it is quite clear what we must do in order to obtain a subset version of FPE. We just need to find an unbiased estimator for the mean squared prediction error, which is given by

$$\begin{aligned}
E\,\frac{1}{n}\sum_{j=1}^{n}(y_j^* - \hat{\beta}_j)^2 &= \frac{1}{n}E\left(\sum_{j\ \mathrm{incl.}}(y_j^* - y_j)^2 + \sum_{j\ \mathrm{excl.}} y_j^{*2}\right)\\
&= \frac{1}{n}\left(E\sum_{\mathrm{all}\ j} y_j^{*2} - 2\sum_{j\ \mathrm{incl.}} Ey_j^* Ey_j + E\sum_{j\ \mathrm{incl.}} y_j^2\right)\\
&= \sigma^2\left(1+\frac{\varsigma(k,\,n)}{n}\right)\\
&= \sigma^2\,\frac{n+\varsigma(k,\,n)}{n},
\end{aligned}$$

if all parameters β_j vanish. Hence

$$\hat{\sigma}^2(k)\frac{n+\varsigma(k,\,K)}{n-\varsigma(k,\,K)}$$

is an unbiased estimator for the mean squared prediction error and the criterion

$$n\log(\hat{\sigma}^2(k)) + n\log\left(1+\frac{2\varsigma(k,\,K)}{n-\varsigma(k,\,K)}\right), \qquad (\mathrm{FPE}_{\mathrm{sub}})$$

may be considered as a genuine subset version of FPE. Of course, this criterion penalizes additional regressors much heavier than $\mathrm{Chi}^{**}_{\mathrm{sub}} - C$ (see also Figure 3b). At this point, we want to emphasize that, quite apart from their statistical properties, the derivations of $\mathrm{Chi}^{**}_{\mathrm{sub}} - C$ and $\mathrm{FPE}_{\mathrm{sub}}$ appear to be much sounder than those of $\mathrm{Chi}^{*}_{\mathrm{sub}} - C$ and MRIC.

Criteria for subset selection often become increasingly milder as the number of already included regressors increases. Keeping this very property in mind, we could be tempted to design simulation experiments, which bluntly favor criteria of this type. Unfortunately, the results of such simulation studies are sometimes interpreted as evidence of some kind of overall optimality of MRIC and the likes. We will discuss this issue in more detail in the next section.

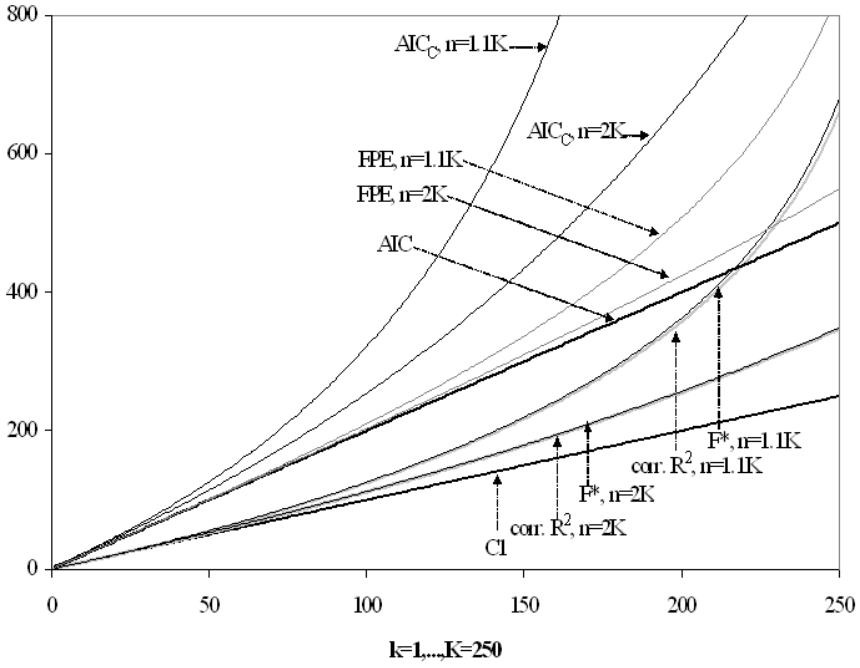

Figure 3a. Comparison of different penalty terms for nested models, where $k = 1, \ldots, K = 250$ and $n = 1.1K, 2K$

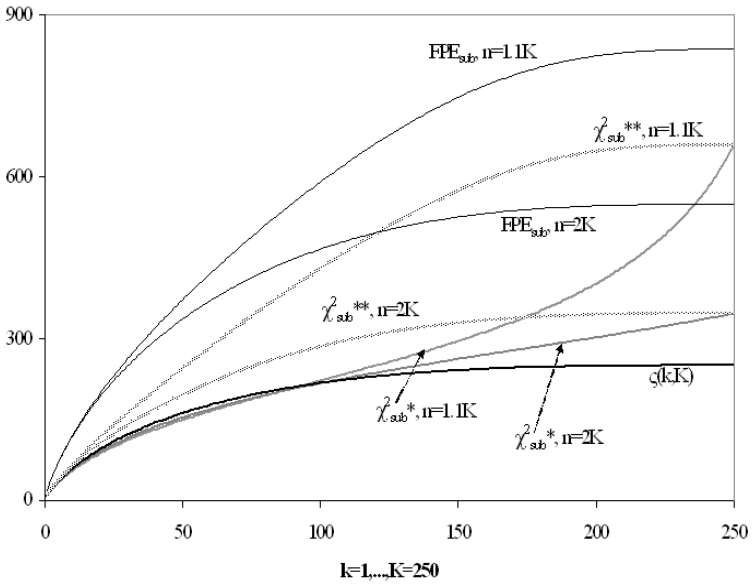

Figure 3b. Comparison of different penalty terms for subset models, where $k = 1, ..., K = 250$ and $n = 1.1K, 2K$

4. Discussion

The emergence of AIC as a standard criterion in the case of nested models is mainly due to three facts. Firstly, the rationale behind AIC is straightforward and easy to understand. Secondly, various asymptotic results have been obtained that suggest that AIC could be optimal in some sense. Thirdly, practical experience shows that AIC very rarely selects an absurd model. In most cases the models selected by AIC are quite reasonable. Despite the fact that numerous simulation studies have been carried out to examine the performance of AIC, these studies are hardly ever cited as arguments for or against AIC. In the case of subset models, simulation studies are more important because there are hardly any asymptotic results and we also have much less experience with these models. However, in the light of our discussion in Section 1 we would expect that in a fairly designed simulation study each criterion comes off well in certain situations and badly in other situations. But how we can then explain the overall good performance of MRIC and the likes demonstrated in several Monte Carlo studies? Choosing an appropriate

version of MRIC for a particular setting and looking at the problem from a particular viewpoint will do the trick.

We use the simple orthogonal set-up of George and Foster [3]. There we have $X = I$ so that $y = X\beta + \varepsilon$ reduces to $y = \beta + \varepsilon$. Since in this case $K = n$, the orthogonality of the regressors is crucial for the examination of all 2^K sub-models. For each selection criterion, β_j is estimated by $\hat{\beta}_j = y_j$ if the j-th regressor is selected and by $\hat{\beta}_j = 0$ otherwise. The nonzero elements $\beta_1, ..., \beta_q$ of β are generated independently as $N(0, r^2)$. The remaining elements $\beta_{q+1}, ..., \beta_n$ are set zero. The errors $\varepsilon_1, ..., \varepsilon_n$ are generated independently as $N(0, \sigma^2)$, where $\sigma^2 = 1$. Under the null hypothesis that all regression coefficients are zero, the quantities $y_1^2, ..., y_n^2$ will then be i.i.d. $\chi^2(1)$ and, assuming that σ^2 is known (!), we may therefore simply select that k which minimizes the sum of the $n - k$ smallest squared observations plus the MRIC penalty for k regressors.

Following George and Foster [3], we set $r^2 = 5$, $n = 1000$, and $q = 0$, 10, 25, 50, 100, 200, 300, 400, 500, 750, 1000. For each specification the sum of prediction errors

$$\sum_{j=1}^{n} (y_j^* - \hat{y}_j^*)^2, \text{ where } \hat{y}_j^* = \begin{cases} y_j & \text{if } j \text{ is selected,} \\ 0 & \text{if } j \text{ is not selected,} \end{cases}$$

is averaged over 25,000 repetitions. The results obtained for the fixed penalty criteria $C(i)$, $i = 0, 1, 2, 3, 4, 5, 10, 15, 20, 25$ (where $C(i)$ includes the j-th regressor if $y_j^2 > i$) are quite unspectacular, the performance of each of these criteria just keeps deteriorating linearly as q increases (see Figure 4a). In contrast, since MRIC assigns large penalties to the very first regressors and small penalties to the very last regressors, it does not come as a surprise that it performs well in the case of very small or very large values of q. But how is its performance in the case of medium-sized values of q? Figure 4a shows that for each q its performance is always "close" to that of the respective best fixed penalty criterion. But this does not "prove" the superiority of MRIC, because its performance for medium-

sized values of q depends critically on the size of r^2. As r^2 increases, the relative performance of MRIC deteriorates quickly (see Figure 4b). The results of simulation studies like those presented in Figure 4a, which suggest that there might be some kind of "overall optimality", must therefore be interpreted with the utmost care.

As long as we do not have more experience with subset criteria or understand their theoretical properties better, it seems that the best we can do is to use the criterion with the soundest derivation. Among the criteria discussed in this article, this is possibly the subset version of FPE. Unfortunately, we cannot be completely happy with FPE_{sub}, because it has been derived under the assumption that all regression coefficients vanish. In contrast, for the derivation of the original criterion FPE we only had to assume that the models under consideration are correctly specified. In the Appendix, we therefore sketch a universal model selection procedure that is not based on unrealistic assumptions about the model parameters and can be used in a variety of situations.

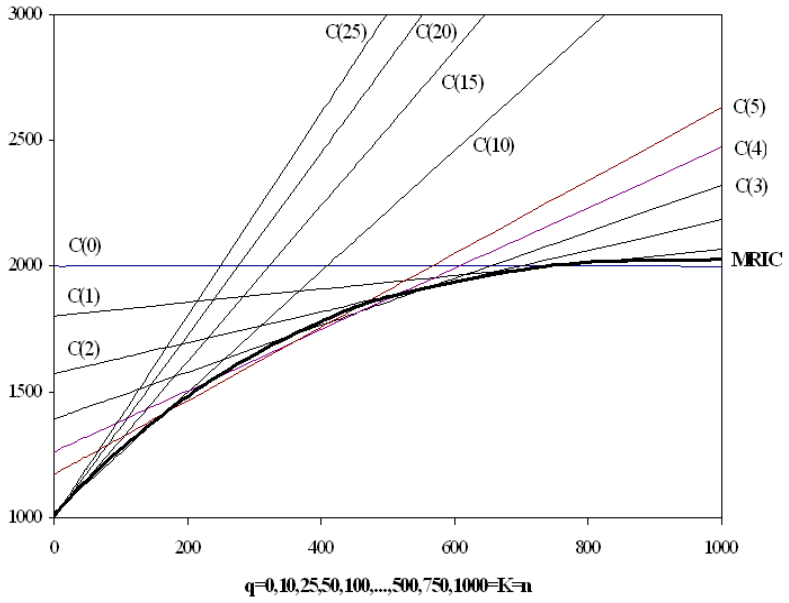

Figure 4a. Comparison of the performance (sum of prediction errors) of MRIC and various fixed penalty criteria (for $r^2 = 5$ and different numbers of nonzero regression coefficients, i.e., $q = 0, 10, 25, 50, 100, 200, 300, 400, 500, 750, 1000 = K = n$)

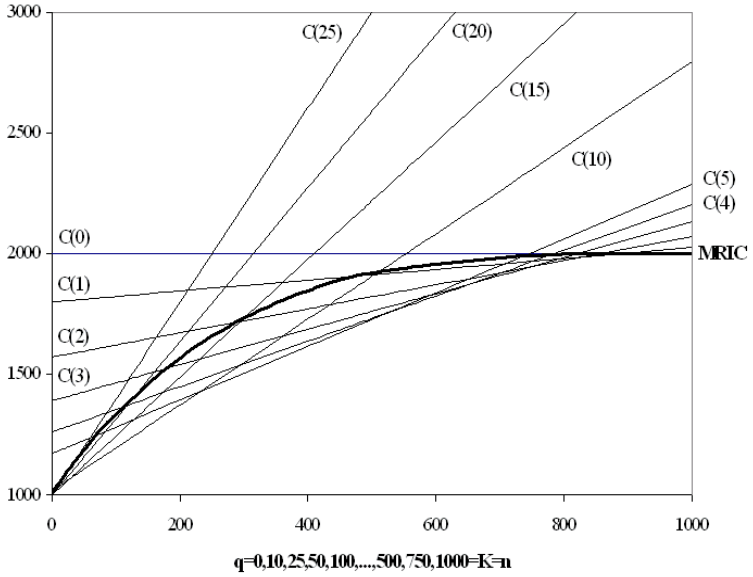

Figure 4b. Comparison of the performance (sum of prediction errors) of MRIC and various fixed penalty criteria (for $r^2 = 100$ and different numbers of nonzero regression coefficients, i.e., $q = 0, 10, 25, 50, 100, 200, 300, 400, 500, 750, 1000 = K = n$)

References

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

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

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

[4] P. Kabaila, Admissible variable-selection procedures when fitting misspecified regression models by least squares, Comm. Statist. Theory Methods 26(10) (1997), 2303-2306.

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

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

[7] W. Ploberger and P. C. B. Phillips, An introduction to best empirical models when the parameter space is infinite dimensional, working paper, 2003.

[8] E. Reschenhofer, Some remarks on the entropy maximization principle, Estadística 49-51 (1997-99), 131-165.

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

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

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

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

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

[14] C. J. Stone, Admissible selection of an accurate and parsimonious normal linear regression model, Ann. Statist. 9 (1981), 475-485.

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

Appendix: The Mother of all Criteria

Let $M_1, ..., M_r$ be r competing modeling procedures and $y = (y_1, ..., y_n)$ a sample of n observations. For each $i \in \{1, ..., r\}$, we first apply the modeling procedure M_i to the observations to obtain a completely specified data generating mechanism G_i, which is then used to produce m independent pairs of samples $(y^{(i)}(j), y^{*(i)}(j))$, where

$$y^{(i)}(j) = (y_1^{(i)}(j), ..., y_n^{(i)}(j)),$$

and

$$y^{*(i)}(j) = (y_1^{*(i)}(j), ..., y_n^{*(i)}(j)), \quad j = 1, ..., m.$$

Now M_i is applied to the synthetic samples $y^{(i)}(1), ..., y^{(i)}(m)$ yielding the data generating mechanisms $G^{(i)}(1), ..., G^{(i)}(m)$. Using some suitable goodness-of-fit measure F, we determine the goodness of fit both for the m pairs $(y^{(i)}(1), G^{(i)}(1)), ..., (y^{(i)}(m), G^{(i)}(m))$ and for the m pairs $(y^{*(i)}(1), G^{(i)}(1)), ..., (y^{*(i)}(m), G^{(i)}(m))$ and use the discrepancy between the biased

measures $F(y^{(i)}(j), G^{(i)}(j))$ and the unbiased measures $F(y^{*(i)}(j), G^{(i)}(j))$ to evaluate the bias of M_i as

$$b(i) = \frac{1}{m}\sum_{j=1}^{m} F(y^{(i)}(j), G^{(i)}(j)) - \frac{1}{m}\sum_{j=1}^{m} F(y^{*(i)}(j), G^{(i)}(j)).$$

Finally, the "optimal" model is chosen by minimization of the criterion

$$F(y, G_i) - b(i). \qquad \text{(MOAC)}$$

This universal criterion, which we call the mother of all criteria (MOAC), may not only be used for the selection of a suitable model in a given class of models but could also be used for the selection of the model class itself and for many other selection tasks.

To illustrate the possible applications of this criterion we consider the problem of finding a suitable model for the sequence y of first differences of the logarithms of a macroeconomic time series like the real U.S. GDP. Possible models for such data are autoregressive (AR) models, moving average (MA) models, autoregressive moving average (ARMA) models, and fractionally integrated ARMA (ARFIMA) models. We assume that all models are Gaussian. Each model is completely specified, if the vector θ of model parameters (including the mean, the AR parameters, the MA parameters, the fractionally differencing parameter, and the variance of the innovations) is specified. A suitable goodness-of-fit measure is given by the Gaussian log likelihood, i.e., $F(y, G) = F(y, \theta) = \log f(\theta \mid y)$.

Example 1. In a traditional application, $G_1, \ldots, G_r$ (i.e., $\theta_1, \ldots, \theta_r$) could be obtained by fitting AR models of order $1, \ldots, r$ to the data. Here MOAC only determines the dimension of the model.

Example 2. In a more unorthodox application, G_1 could be obtained by selecting a suitable AR model with AIC and G_2 by selecting another AR model with BIC. Here MOAC assigns different penalties to the two models selected by AIC and BIC, respectively, and uses these penalties to select one of the two models.

Example 3. G_1 and G_2 are obtained by selecting a suitable AR model and a suitable MA model, respectively, with AIC.

Example 4. Both G_1 and G_2 are obtained by selecting a suitable AR model with AIC, but the model parameters are estimated in different ways (e.g., maximum likelihood estimation for the parameters of G_1 and least squares estimation for the parameters of G_2). This example is not very exciting, admittedly, but it serves the purpose of illustrating the range of possible applications.

Example 5. $G_1, ..., G_r$ are obtained by fitting different subset AR models to the data.

Example 6. G_1 and G_2 are the best AR models according to AIC and BIC, respectively, G_3 and G_4 are the best ARMA models according to AIC and BIC, respectively, G_5 and G_6 are the best ARFIMA models according to AIC and BIC, respectively, G_7 is the AR model selected as in Example 1, and G_8 is the subset AR model selected as in Example 5.

Of course, very little can be said about the properties of the models selected in these examples. The best way to assess the usefulness of this approach would be to use it in different concrete situations and to examine whether the selected models are meaningful. We should also examine different versions of MOAC. For example, in a situation, where all competing models are submodels of one large model M_r, all synthetic series could be generated with G_r. Regardless of which version is used, it might well turn out that in all but very simple applications the involved computations are too costly. Large simulation studies would definitely be computationally intractable.

■