Ann. Inst. Statist. Math.
Vol. 45, No. 1, 35–54 (1993)

MODEL SELECTION AND PREDICTION: NORMAL REGRESSION*

T. P. SPEED[1] AND BIN YU[2]

[1] *Department of Statistics, University of California at Berkeley, CA 94720, U.S.A.*
[2] *Department of Statistics, University of Wisconsin-Madison, WI 53706, U.S.A.*

(Received September 27, 1991; revised April 27, 1992)

Abstract. This paper discusses the topic of model selection for finite-dimensional normal regression models. We compare model selection criteria according to prediction errors based upon prediction with refitting, and prediction without refitting. We provide a new lower bound for prediction without refitting, while a lower bound for prediction with refitting was given by Rissanen. Moreover, we specify a set of sufficient conditions for a model selection criterion to achieve these bounds. Then the achievability of the two bounds by the following selection rules are addressed: Rissanen's accumulated prediction error criterion (APE), his stochastic complexity criterion, AIC, BIC and the FPE criteria. In particular, we provide upper bounds on overfitting and underfitting probabilities needed for the achievability. Finally, we offer a brief discussion on the issue of finite-dimensional vs. infinite-dimensional model assumptions.

Key words and phrases: Model selection, prediction lower bound, accumulated prediction error (APE), AIC, BIC, FPE, stochastic complexity, overfit and underfit probability.

1. Introduction

This paper discusses the topic of model selection for prediction in regression analysis. We compare model selection criteria according to the quality of the predictions they give. Two types of prediction errors, prediction with and without refitting, will be considered. A lower bound on the former type of error was given by Rissanen (1986*a*), and in this paper (Section 2) we provide a lower bound for the latter. Moreover, also in Section 2 we specify a set of sufficient conditions for a model selection criterion to achieve these bounds. Roughly speaking, to achieve these bounds, a model selection criterion has to be consistent and satisfy some underfitting and overfitting probability constraints. Section 3 concerns the following model selection criteria: Rissanan's predictive "minimum description length"

* Support from the National Science Foundation, grant DMS 8802378 and support from ARO, grant DAAL03-91-G-007 to B. Yu during the revision are gratefully acknowledged.

(accumulated prediction error, or predictive least squares), stochastic complexity, AIC, BIC and FPE. We consider bounds on their overfitting and underfitting probabilities, and therefore their achievability of the prediction lower bounds. In particular, the selection rule based on the accumulated prediction error and BIC achieve the two prediction lower bounds, but AIC does not unless the largest model considered is the true model.

Detailed proofs are relegated to the last section 5. All of our results are obtained under the assumption that a finite dimensional normal model generates the data under discussion. This contrasts greatly with most previous discussions, notably Shibata (1983*a*, 1983*b*) and Breiman and Freedman (1983), where the "true" model is infinite-dimensional. More discussion on finite-dimensional models vs. infinite-dimensional models can be found in Section 4.

2. Model selection and prediction in regression

In order to compare model selection procedures a number of choices need to be made; these can be critical. Two objectives of regression analysis are data description and prediction. The focus will be on the second, prediction.

Write $y = (y_1, \ldots, y_n)'$ for the n-dimensional column vector of observations, and $X = (x_{ij})$ for the $n \times K$ matrix of covariates or regressors. Inner products and squared norms are denoted by $\langle y, z\rangle = \sum y_t z_t$ and $|y|^2 = \langle y, y\rangle$, respectively. For $1 \le t \le n$, $1 \le k \le K$, denote by $y(t)$ and $X_k(t)$ that $t \times 1$ and $t \times k$ subvector and submatrix of y and X respectively, consisting of the first t rows and, in the case of X, of the first k columns. The subscript k or the parenthetical t will be omitted when they are clear from the context, or when $k = K$ or $t = n$. The t-th row of X is denoted by x_t' and the j-th column by ξ_j, whilst $x_t'(k)$ denotes the t-th row of X_k, with an analogous convention regarding the dropping of t or k. Parameter vectors are denoted by $\beta = (\beta_1, \ldots, \beta_k)'$, written $\beta(k)$ when necessary.

The class of models to be discussed will be denoted by $\{M_k : 1 \le k \le K\}$, where M_k is the model prescribing that y is $N(X_k\beta, \sigma^2 I)$ for some $\beta \in \boldsymbol{R}^k$ and $\sigma^2 > 0$. The number K of models is supposed known, and for the present discussion is held fixed as the sample size $n \to \infty$.

One framework for prediction involving regression is the following: (y_1, x_1), $(y_2, x_2), \ldots, (y_t, x_t)$ are given. The object is to predict y_{t+1} from x_{t+1}. An obvious approach is to select a model on the basis of the data available at time t, and predict y_{t+1} from this model with $t + 1$ replacing t. The response y_t at time t is known before predicting y_{t+1}, so this framework is called *prediction with repeated refitting* because it allows model selection at each time.

A quite different framework assumes the existence of an initial data set $\{(y_1, x_1), \ldots, (y_n, x_n)\}$, often called a training sample, and the regressors $\tilde{x}_1, \ldots, \tilde{x}_m$ associated with a number of other units, the requirement being to predict the corresponding responses $\tilde{y}_1, \ldots, \tilde{y}_m$. A familiar variant on this would be when the "prediction" is in fact the *allocation* of units into predetermined groups. The standard solution to this problem is to select a model on the basis of the initial data set, and then predict or allocate using the model selected. This framework will be called *prediction without refitting*.

In this section, the above two frameworks for prediction will be discussed in detail: lower bounds are given in each case, and sufficient conditions for a model selection procedure to achieve them are obtained. However, we leave to Section 3 the achievability of these lower bounds by common selection procedures.

2.1 *Prediction with repeated refitting*

A natural measure of the quality of a sequence of predictions in the repeated refitting framework is the sum

$$\text{APE}_n = \sum_{t=1}^{n} (y_t - \hat{y}_{t|t-1})^2 \tag{2.1}$$

where $\hat{y}_{t|t-1}$ denotes a predictor of y_t made on the basis of data up to and including time $t-1$, and any covariates available at time t. Model selection is thus permitted at every stage. The predictors which we consider below are $\hat{y}_{t|t-1} = x_t' \hat{\beta}_{t-1}(\hat{k}_{t-1})$, where $\hat{\beta}_{t-1}(\hat{k}_{t-1})$ is the least squares estimator based on model $M_{\hat{k}_{t-1}}$ at time t, and we will compare selection procedures leading to different $\hat{k}_s$ according to the average size of APE which is achieved for large n. For the purposes of our asymptotic analysis, it is not necessary to specify how we define $\hat{k}_t$ for $t \leq K$. In practice a number of reasonable approaches exist.

Our comparison is based upon a general inequality derived by Rissanen ((1986*a*), p. 1087). As in Sections 3 and 4 we denote by k^* the dimension associated with the true model, and $\hat{y}_{t|t-1}$ is *any* predictor of y_t which is a measurable function of $y_1, \ldots, y_{t-1}$, and $x_1, \ldots, x_t$. Although all our discussions so far have supposed that the error variance σ^2 is known and equal to unity, we will state the inequality for an arbitrary unknown σ^2. It asserts that for all k^* there is a Lebesgue null subset $A(k^*)$ of $\boldsymbol{R}^{k^*}$ such that for $\beta^* \notin A(k^*)$:

$$\liminf_{n \to \infty} \frac{\boldsymbol{E}_{\beta^*} \{\sum_1^n (y_t - \hat{y}_{t|t-1})^2 - n\sigma^2\}}{k^* \log n} \geq \sigma^2. \tag{2.2}$$

We say that the lower bound (2.2) is achieved by a model selection criterion if it is achieved by the corresponding predictor $y_{t|t-1}$.

We need some assumptions before we can state our results on the achievability of the prediction lower bound (2.2).

Assume (cf. Lai *et al.* (1979)) that there exists a positive definite $K \times K$ matrix $C = C_K$ such that

$$\lim_{N \to \infty} N^{-1} \sum_{t=M+1}^{M+N} x_t x_t' = C \tag{2.3}$$

uniformly in $M \geq 0$. If $M = 0$, the left-hand side is just $\lim_N N^{-1} X(N)' X(N)$. A further specialization gives $\lim_N N^{-1} X_k(N)' X_k(N) = C_k$, where C_k denotes the principal $k \times k$ submatrix of C. Assume also that

(2.4) $M_{k^*} \subseteq M_K$ is the smallest true model, and $\beta(k^*)$ the true parameter.

With this background we can now state the following result, proved in Section 5 below.

THEOREM 2.1. *Suppose that* (2.3) *and* (2.4) *hold and that* $\hat{k}_n$, *the dimension defined by a model selection procedure, satisfies*:

(i) $\text{pr}(\hat{k}_n < k^*) = O(n^{-2}(\log n)^{-c})$ *as* $n \to \infty$, *for some* $c > 1$, *and*

(ii) $\text{pr}(\hat{k}_n > k^*) \leq O((\log n)^{-\alpha})$ *as* $n \to \infty$, *for some* $\alpha > 2$.

Then the predictor $\hat{y}_{t|t-1} = x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1})$ *achieves the lower bound* (2.2).

2.2 *Prediction without refitting*

Now let us suppose that we have observed $(y_1, x_1), \ldots, (y_n, x_n)$ and are required to predict the responses $\tilde{y}_1, \ldots, \tilde{y}_m$ corresponding to units with covariate vectors $\tilde{x}_1, \ldots, \tilde{x}_m$. In most discussions of this aspect of model selection, see e.g. Nishi (1984) and Shibata (1986*a*), $m = n$ and $x_i = \tilde{x}_i$, $1 \leq i \leq n$. Our framework is more realistic and although the general conclusions do not seem to be different from Shibata's, this was not obvious a priori.

Our predictors will all be of the form $\tilde{x}_u'\hat{\beta}(\hat{k})$, $u = 1, \ldots, m$ where $\hat{k}$ corresponds to a model selected on the basis of $\{(y_t, x_t) : t = 1, \ldots, n\}$. Given that $\hat{k} = k$, a natural measure of the quality of our set of m predictions is given by the *prediction error*

$$\text{PE}(k) = \boldsymbol{E}\{|\tilde{y} - \tilde{X}_k\hat{\beta}(k)|^2 \mid y\} = m\sigma^2 + |\tilde{X}_{k^*}\beta(k^*) - \tilde{X}_k\hat{\beta}(k)|^2,$$

which averages over the new observations and conditions on the initial data. Following this line of thought, an equally natural measure of the effectiveness of the model selection procedure leading to $\hat{k}$ is $\boldsymbol{E}\{\text{PE}(\hat{k}) - m\sigma^2\}$, where this time the expectation is over the possible initial data sets. What we now do is give some results on the behaviour of this quantity under a range of assumptions about $\tilde{X}$.

Our results are asymptotic in both n, the size of the initial sample, and m, the number of predictions being made. For this reason we need to supplement assumption (2.3) with an analogous, but weaker hypothesis concerning $\tilde{X}$ namely: that there exists a $K \times K$ positive definite $\tilde{C} = \tilde{C}_K$ such that

$$\lim_{M \to \infty} M^{-1} \sum_{u=1}^{M} \tilde{x}_u\tilde{x}_u' = \tilde{C}. \tag{2.5}$$

In the theorems which follow, $\hat{k} = \{\hat{k}_n\}$ is the index resulting from a procedure selecting from the models $\{M_k : 1 \leq k \leq K\}$.

The components of condition (B) below are defined by the partitioning

$$C_{k+1} = \begin{bmatrix} C_k & D_{k,k+1} \\ D_{k,k+1} & E_{k,k+1} \end{bmatrix},$$

where C_k, $k \leq K$ is defined following (2.3).

THEOREM 2.2. *Assume conditions* (2.3), (2.4) *and* (2.5). *Then under any of the following conditions*:

(A) $\lim_{n\to\infty} \mathrm{pr}(\hat{k}_n < k^*) > 0$;
(B) $C_k^{-1} D_{k,k+1} = \tilde{C}_k^{-1} \tilde{D}_{k,k+1}$, $k^* \le k < K$;
(C) $\hat{k} = \hat{k}_{\mathrm{FPE}_\alpha}$ *for a sequence* $\alpha = (\alpha_n)$ *with* $n^{-1}\alpha_n \to 0$ *where* FPE_α *is the Final Prediction Error criterion defined in Section* 3, *we may conclude*

$$\lim_{m,n\to\infty} nm^{-1} \boldsymbol{E}\{\mathrm{PE}(\hat{k}_n) - m\sigma^2\} \ge \mathrm{tr}\{C_{k^*}^{-1} \tilde{C}_{k^*}\}\sigma^2. \tag{2.6}$$

The proof will be given in Section 5. It can be seen from the proof of this theorem that there will be other "symmetric" selection rules other than FPE_α for which the conclusion holds.

The next question of interest is the following: what kinds of selection rules attain the lower bound (2.6)?

THEOREM 2.3. *The lower bound* (2.6) *is attained for any consistent selection rule whose underfitting probability* $\mathrm{pr}(\hat{k}_n < k^*)$ *is* $o(n^{-2})$ *as* $n \to \infty$.

3. APE, stochastic complexity, and FPE

In this section, we consider the achievability of the two lower bounds in Section 2 of some commonly-used model selection criteria. We derive upper bounds on the underfitting and overfitting probabilities of these criteria and then use Theorem 2.1 or Theorem 2.3.

First, we consider the criterion based upon accumulated (one-step) prediction errors (APE) (or predictive least squares). This criterion is the predictive MDL criterion introduced in Rissanen (1984, 1986*b*). Many authors have discussed this criterion as detailed in the remark after Theorem 3.1.

We now introduce the definition of APE. Only ordinary least squares estimates will be used. For $1 \le k \le K$, $k+1 \le s \le n$, write

$$\hat{\beta}_s(k) = (X_k(s)' X_k(s))^{-1} X_k(s)' y(s)$$

and $\hat{\beta}(k) = \hat{\beta}_n(k)$. All of the matrices $X_k(t)$ will be assumed to have rank k when $t > k$. The *recursive residuals*, also called one-step prediction errors, based on M_k are $e_t(k) = y_t - x_t(k)'\hat{\beta}_{t-1}(k)$. The ordinary residuals are $r_{t,n}(k) = y_t - x_t(k)'\hat{\beta}_n(k)$. The parenthetical k will be dropped if its value is clear from the context.

For any fixed $k \le K$, consider the accumulated squared prediction error $\mathrm{APE}_n(k) = \sum_{t=k+1}^{n} e_t(k)^2$. Obviously, $\mathrm{APE}_n(k)$ is the same as the prediction error with refitting (2.2) when the model M_k is fixed through time t.

Expression $\mathrm{APE}_n(k)$ will lead us to a model selection criterion: choose that k which minimizes $\mathrm{APE}_n(k)$ over all $k \le K$.

For the remainder of this section σ^2 is supposed known and so, for simplicity, is taken to be 1. This is possible because, unlike many model selection criteria, the one based on APE does not require knowledge or an estimate of σ^2. The numbers $\{b_k\}$ which appear in the following theorem are normalized limiting (squared) bias terms defined by

$$b_k = \mathrm{tr}\{(E_{k,k^*} - D'_{k,k^*} C_k^{-1} D_{k,k^*})\zeta(k)\zeta(k)'\}$$

where for $k < k^*$ the principal submatrices C_k and C_{k^*} of C are written

$$C_{k^*} = \begin{bmatrix} C_k & D_{k,k^*} \\ D_{k,k^*} & E_{k,k^*} \end{bmatrix},$$

and $\beta(k^*) = (\beta(k)' \mid \zeta(k)')'$ is the corresponding partitioning of $\beta(k^*)$. It is shown in Section 5 (Lemma 5.3) that $b_1 \geq b_2 \geq \cdots \geq b_{k^*-1} > 0$.

THEOREM 3.1. *Under assumptions* (2.3) *and* (2.4), *as* $n \to \infty$, *let* $\hat{k}_n$ *denote the dimension selected by minimizing* $\mathrm{APE}_n(k)$. *Then we have the following bounds*:

(i) $\mathrm{pr}(\hat{k}_n < k^*) \leq O(\exp(-bn))$ *as* $n \to \infty$, *for* $b = \min(b_{k^*-1}/3, b_{k^*-1}^2/18)$.

(ii) $\mathrm{pr}(\hat{k}_n > k^*) \leq O(n^{-1/6})$ *as* $n \to \infty$.

Remark. The upper bound in (i) shows the interplay between the bias term b_k and the sample size n; the product of them determines the underfitting probability, not the sample size n alone.

COROLLARY 3.1. *The lower bounds* (2.2) *and* (2.6) *are attained for the* APE *selection rule.*

PROOF. Straightforward from Theorems 2.1, 2.2 and 3.1.

Remark. (a) Convergence in probability of the APE selection rule was established by Rissanen (1986*b*) under essentially the same conditions as we have used here. Other writers who have suggested the use of APE or a related criterion to select regression models include Hjorth (1982) and Dawid (1984, 1992). The latter describes a generalization of the use of APE as the prequential approach to statistical analysis. (b) There is no doubt that our assumptions could be weakened, but the derivations of the same results are expected to be much more involved. In the context of time series, Wax (1988) derived the weak consistency of an analogous estimator of the order of an autoregressive process without the Gaussian assumption, and Hemerly and Davis (1989) strengthened it to the a.s. consistency. Moreover, Wei (1992) obtained the a.s. consistency and asymptotic expansions of APE under stochastic regression models.

Now we turn to selection rules based on the residual sum of squares, which is $\mathrm{RSS}_n(k) = \sum_1^n r_{t,n}(k)^2$ where the ordinary residuals $r_{t,n}(k)$ are defined above. When $\sigma^2 = 1$ in the regression models M_k the *final prediction error* (FPE) criterion is $\mathrm{FPE}_{\alpha_n}(k) = \mathrm{RSS}_n(k) + \alpha_n k$ where (α_n) is a sequence of positive numbers. For AIC, $\alpha_n \equiv 2$. For BIC (Schwartz (1978)), $\alpha_n = \log n$. When σ^2 is not known, we may replace it by its usual estimate from the largest model M_K. Our results should still hold in that case.

Rissanen (1986*a*) introduced stochastic complexity (SC) of a set of data relative to a model as variant of his MDL and PMDL expressions, and in many cases it is asymptotically equivalent to the latter, whilst being easier to calculate. We refer to his paper for definitions of these quantities. For our regression models

with error variance equal to unity, SC takes a particularly simple form if the prior distribution for the parameter $\beta(k)$ is taken to be $N(0, \tau I_k)$ where $\tau > 0$ is a scale parameter, $k = 1, \ldots, K$. A simple calculation yields the expression

$$\text{SC}_n(k) = \frac{1}{2} n \log 2\pi + \frac{1}{2} \log \det(I_n + \tau X_k X_k') + \frac{1}{2} y'(I_n + \tau X_k X_k')^{-1} y. \tag{3.1}$$

From Lemma 5.5 in Section 5 we see that as $n \to \infty$,

$$\text{SC}_n(k) - \frac{1}{2} n \log 2\pi = k \log n + \text{RSS}_n(k) + O(1) \quad \text{a.s.}$$

and so any discussion of model selection based upon stochastic complexity is subsumed under that of BIC.

The FPE criterion has been discussed by Akaike (1970, 1974), Bhansali and Downham (1977), Atkinson (1980), and Shibata (1976, 1986*a*) amongst others. Geweke and Meese (1981) discuss the problem quite generally, but with random regressors, whilst Kohn (1983) considers selection in general parametric models. Shibata (1984) may be consulted for further details on some cases of FPE. The consistency of FPE's, with α_n's satisfying $\lim n^{-1}\alpha_n = 0$ and $\underline{\lim}(2 \log \log n)^{-1}\alpha_n > 1$, was established in a time-series context by Hannan and Quinn (1979). Moreover, the equivalence of BIC and APE has been shown by Hannan *et al.* (1989) for the finite-dimensional autoregressive models and by Wei (1992) for finite-dimensional stochastic regression models.

THEOREM 3.2. *Let $\hat{k}_n$ denote the dimension selected by* FPE_{α_n} *for some sequence α_n such that $n^{-1}\alpha_n \to 0$ as $n \to \infty$. Then*

(i) *$\hat{k}_n$ overfits with probability approaching unity as $n \to \infty$. More precisely, for any constant $0 < b < b_{k^*-1}/4$, $\text{pr}(\hat{k}_n < k^*) \leq O(\exp(-bn))$ as $n \to \infty$.*

(ii) *If $k^* < K$, and $\liminf(2 \log \log n)^{-1}\alpha_n > 2$, we have, for some $\gamma > 2$, $\text{pr}(\hat{k}_n > k^*) \leq O((\log n)^{-\gamma})$ as $n \to \infty$.*

We omit the proof of this theorem in this paper because Woodroofe (1982) and Haughton (1989) contain smilar bounds for BIC under more general models. Moreover, a lower bound, instead of an upper one, on the overfit probability (ii) is given in the Appendix II of Merhav *et al.* (1989) for BIC. Their result suggests that the overfit probability of BIC tends to zero slower than exponentially as n tends to infinity.

COROLLARY 3.2. (i) *The selection rules defined by* BIC *and* SC *all lead to predictors which achieve the lower bounds* (2.2) *and* (2.6);

(ii) *If $\lim(2 \log \log n)^{-1}\alpha_n < 1$, the selection rules defined by* FPE_{α_n} *do* not *achieve the lower bounds* (2.2) *and* (2.6) *unless $k^* = K$; in particular,* AIC *does* not *achieve the lower bounds unless $k^* = K$.*

4. Discussion

The results presented seem to suggest that if prediction is part of the objective of a regression analysis, then model selection carried out using APE, BIC, SC or an equivalent procedure has some desirable properties. Of course there is a qualification: in deriving these theorems we have assumed that the model generating our data is (i) fixed throughout the asymptotics; (ii) finite-dimensional; and (iii) belongs to the class of models being examined. Before commenting on these assumptions, let us see that our theorems are at least in general agreement with a number of analyses and simulations in the literature. The first paper to point out clearly that consistent model selection gives better predictions seems to be Shibata (1984), although he does not emphasize this conclusion. Atkinson's (1980) results also suggest the conclusion we have reached, but again this is not emphasized. The simulation results of Clayton *et al.* (1986) led them to conclude "that if the 'true' or 'approximately true' model is included among the alternatives considered, all reasonable model selection procedures will possess rather similar predictive capabilities". We feel that this conclusion is more a reflection of the limited scope of the simulations conducted rather than the true state of affairs. Indeed a close examination of the sample sizes and models these authors studied suggests that there was little opportunity for the procedures (not the models) to be distinguished, as far as the squared prediction error of the resulting choices is concerned. More recently, Rissanen (1989) reported clear differences between cross validation and SC, and to the extent that cross-validation and AIC perform similarly, Stone (1977), this is explained by Corollary 3.2.

Shibata (1981, 1983*a*, 1983*b*, 1984, 1986*a*, 1986*b*) presents a number of theorems demonstrating the optimality of AIC or other forms of FPE_{α_n} with bounded sequences (α_n), as well as arguments rebutting the criticisms that such procedures are unsatisfactory by virtue of their inconsistency under assumptions (i), (ii) and (iii). Shibata (1981), and Breiman and Freedman (1983) using random regressors, suppose the true model to be *infinite*-dimensional rather than *finite*-dimensional. Shibata (1981) also offers an optimality result for AIC valid under a "moving truth" assumption.

Clearly, the prediction optimality of BIC and its analogues like APE depend on the assumption that the true model is finite-dimensional, i.e., the bias term $b_k = 0$ for $k \geq k^*$. When the true model is assumed to be infinite-dimensional, i.e., $b_k > 0$ for all k, Breiman and Freedman (1983) showed that AIC's equivalent is optimal in terms of one-step further prediction. We now show by the following three simple examples that the decay rate of the bias term plays a determining role in the battle of AIC vs. BIC.

For simplicity, let us take the framework of Breiman and Freedman (1983) where an infinite-dimensional model with Gaussan $N(0,1)$ independent regressors is assumed with the error variance $\sigma^2 = 1$. Then the one-step ahead prediction error for the $(n+1)$-st observation based on model M_k is roughly $\mathrm{PE}(k) = b_k + kn^{-1}$. Moreover, AIC approximately minimizes $b_k + kn^{-1}$, while BIC minimizes $b_k + kn^{-1}\log n$. By the result of Breiman and Freedman (1983), asymptotically, $\mathrm{PE}(\hat{k}_{\mathrm{BIC}})/\mathrm{PE}(\hat{k}_{\mathrm{AIC}}) \geq 1$, where $\hat{k}_{\mathrm{AIC}}$ is the model selected by AIC, and similarly

for $\hat{k}_{\mathrm{BIC}}$.

Example 1. Assume $b_k = k^{-\alpha}$. Straightforward calculation shows that, as $n \to \infty$, $\mathrm{PE}(\hat{k}_{\mathrm{BIC}})/\mathrm{PE}(\hat{k}_{\mathrm{AIC}}) \to \infty$.

Example 2. Assume $b_k = e^{-k}$. Then as $n \to \infty$, $\mathrm{PE}(\hat{k}_{\mathrm{BIC}})/\mathrm{PE}(\hat{k}_{\mathrm{AIC}}) \to 2$.

Example 3. Assume $b_k = e^{-e^k}$. Then as $n \to \infty$, $\mathrm{PE}(\hat{k}_{\mathrm{BIC}})/\mathrm{PE}(\hat{k}_{\mathrm{AIC}}) \to 1$.

To summarize, as the decay rate of the bias term increases, the prediction performance of BIC catches up with that of AIC. And, as we have seen, BIC out-performs AIC when $b_k = 0$ for $k > k^*$, i.e. when the model is finite.

Finally, all three of APE, BIC and SC derive from general approaches to the model selection problem and have extensions to situations where one or more of (i), (ii) and (iii) are dropped, see Sawa (1978) for some remarks about this situation. When something is known about these extensions, it will be of interest to compare them with AIC or, more generally FPE_{α_n}.

5. Proofs

Most of the arguments given below are straightforward. We have tried to be explicit wherever possible, and have included some proofs which may be found elsewhere in order to keep this paper self-contained.

The proofs are presented in the following order: Theorem 3.1, Corollaries 3.1 and 3.2, Theorem 2.2, Theorem 2.3 and Theorem 2.1. We continue to use the notation introduced in Section 2 above. It is straightforward to show

LEMMA 5.1. *For $k < s < t \leq n$ and $c \in R(X_k(t))$, we have* $\mathrm{cov}(e_{s+1}(k), c'y(t)) = 0$.

It follows from the lemma that

COROLLARY 5.1. (a) *For all $k < s < t \leq n$, we have* $\mathrm{cov}(e_s(k), e_t(k)) = 0$. (b) *For all $k < t \leq n$, and $c \in R(X_k)$,* $\mathrm{cov}(e_t(k), c'y) = 0$.

Let us write $\lambda_t(k) = \boldsymbol{E}\{e_t(k)\}$ and $\mu_t(k) = \mathrm{Var}\{e_t(k)\} - 1$, $\epsilon_t = y_t - \boldsymbol{E}\{y_t\}$ and $H_n(k) = X_n(k)(X_n(k)'X_n(k))^{-1}X_n(k)'$, and define the following quantities:

$$V_n(k) = \sum_{t=k+1}^{n} \mu_t(k), \qquad B_n(k) = \sum_{t=k+1}^{n} \lambda_t(k)^2, \qquad N_n(k) = |H_n(k)\epsilon|^2,$$

$$N_n^{\dagger}(k) = \sum_{t=k+1}^{n} \mu_t(k) \left[\frac{(e_t(k) - \lambda_t(k))^2}{\mu_t(k) + 1} - 1 \right],$$

$$B_n^{\dagger}(k) = 2 \sum_{t=k+1}^{n} (e_t(k) - \lambda_t(k))\lambda_t(k).$$

It is clear from the proof of the result we state shortly that V is a *variance* term, B is a *bias* term, and N is a *noise* term, whilst $N^\dagger$ is a second noise term and $B^\dagger$ a part-noise part-bias term.

LEMMA 5.2. *With the above notation*

$$\sum_{t=k+1}^{n} e_t(k)^2 - \sum_{t=1}^{n} \epsilon_t^2 = V_n(k) + B_n(k) - N_n(k) + B_n^\dagger(k) + N_n^\dagger(k). \tag{5.1}$$

PROOF. It follows from Corollary 5.1 that $\{e_{k+1}(k), \ldots, e_n(k)\}$ are pairwise uncorrelated, and uncorrelated with $c'y$ for all $c \in R(X_k)$. Thus we can make an orthogonal transformation and obtain

$$|\epsilon|^2 = |H(k)\epsilon|^2 + \sum_{t=k+1}^{n} \frac{[e_t(k) - \boldsymbol{E}\{e_t(k)\}]^2}{\mathrm{Var}\{e_t(k)\}}. \tag{5.2}$$

The lemma then follows from this equation and the comparing two sides of (5.1). □

In the lemmas which follow, (2.1) and (2.2) will be assumed without comment. Moreover, to state our next result we need a little further notation. For $k < k^*$, write the principal $k \times k$ submatrix C_k of C given by (2.4) in the form

$$C_{k^*} = \begin{bmatrix} C_k & D_{k,k^*} \\ D'_{k,k^*} & E_{k,k^*} \end{bmatrix}$$

and we write $\beta(k^*) = (\beta(k)' \mid \zeta(k)')'$ and $X_{k^*}(n) = [X_k(n) \mid Z_k(n)]$.

LEMMA 5.3. $n^{-1}B_n(k) \to b_k$ *as* $n \to \infty$, *where*

$$b_k = \mathrm{tr}\{(E_{k,k^*} - D'_{k,k^*}C_k^{-1}D_{k,k^*})\zeta(k)\zeta(k)'\}$$

satisfies $b_1 \geq b_2 \geq \cdots \geq b_{k^*-1} > 0$.

PROOF. We begin by observing that for $k < k^*$, $\lambda_t(k) = A_k(t)'\zeta(k)$, where

$$A_k(t)' = z_t(k)' - x_t(k)'(X_k(t-1)'X_k(t-1))^{-1}X_k(t-1)'Z_k(t-1).$$

It follows that $\lambda_t(k)^2 = \mathrm{tr}\{A_k(t)A_k(t)'\zeta(k)\zeta(k)'\}$ and so

$$n^{-1}\sum_{t=k+1}^{n} \lambda_t(k)^2 = \mathrm{tr}\left\{n^{-1}\sum_{t=k+1}^{n} A_k(t)A_k(t)'\zeta(k)\zeta(k)'\right\}.$$

Using (2.4) and the notation introduced above, $t^{-1}X_k(t)'X_k(t) \to C_k$, $t^{-1}X_k(t)' \cdot Z_k(t) \to D_{k,k^*}$, and $t^{-1}Z_k(t)'Z_k(t) \to E_{k,k^*}$ as $t \to \infty$, and so it follows that

$$n^{-1}\sum_{t=k+1}^{n} A_k(t)A_k(t)' \to E_{k,k^*} - D_{k,k^*}C_k^{-1}D_{k,k^*}$$

as $n \to \infty$, giving the expression for b_k stated. The monotonicity of b_k can then be checked using the partial order of positive definite matrices. □

For the next lemma we need some notation paralleling that used in Lemma 5.2 above. Write $\bar{\lambda}_t(k) = \boldsymbol{E}\{r_t(k)\}$ and $\bar{B}_n(k) = \sum_1^n \bar{\lambda}_t(k)^2$. Furthermore, put $\bar{B}_n^\dagger(k) = 2\sum_1^n \bar{\lambda}_t(k)\epsilon_t$. By variants of the proofs of Lemmas 5.2 and 5.3 and by the law of iterative algorithm, we obtain

LEMMA 5.4.

$$\sum_1^n r_t(k)^2 - \sum_1^n \epsilon_t^2 = \bar{B}_n(k) - N_n(k) + \bar{B}_n^\dagger(k) \tag{5.3}$$

where for $k < k^*$, $n^{-1}\bar{B}_n(k) \to b_k$, *and* $\bar{B}_n^\dagger(k) = O((n \log\log n)^{1/2})$ *a.s. as* $n \to \infty$.

LEMMA 5.5. *In the notation introduced prior to equation* (3.1)

$$\log\det(I_n + \tau X_k(n)X_k(n)') + y(n)'(I_n + \tau X_k(n)X_k(n)')^{-1}y(n)$$
$$= k\log n + \sum_1^n r_t(k)^2 + O(1) \qquad a.s.\ n \to \infty.$$

PROOF. Straightforward from assumption (2.3) and Rao ((1973), p. 33). □

In the following lemmas we use the notation $\rho_k = \xi_{k+1} - X_k\gamma_k$, $\tilde{\rho}_k = \tilde{\xi}_{k+1} - \tilde{X}_k\gamma_k$ and $\eta_k = X_k(X_k'X_k)^{-1}\tilde{X}_k'\tilde{\rho}_k$, where $\gamma_k = (X_k'X_k)^{-1}X_k'\xi_{k+1}$. It is evident that γ_k is the regression coefficient of the $(k+1)$-st variable on the previous k, and so ρ_k and $\tilde{\rho}_k$ are essentially residuals when the current model is M_k, whereas η_k is part residual and part fitted value.

LEMMA 5.6.

$$\tilde{X}_{k+1}(X_{k+1}'X_{k+1})^{-1}X_{k+1}\epsilon = \tilde{X}_k(X_k'X_k)^{-1}X_k\epsilon + |\rho_k|^{-2}\langle\rho_k, \epsilon\rangle\tilde{\rho}_k.$$

PROOF. This is a straightforward consequence of the formula for the inverse of a partitioned matrix, see e.g. Rao ((1973), p. 33). □

If we write $N_{m,n}(k) = |\tilde{X}_k(X_k'X_k)^{-1}X_k'\epsilon|^2$ by analogy with the noise term introduced just before Lemma 5.2, then we have

COROLLARY 5.2.

$$N_{m,n}(k+1) = N_{m,n}(k) + 2|\rho_k|^{-2}\langle\eta_k, \epsilon\rangle\langle\rho_k, \epsilon\rangle + |\rho_k|^{-4}|\tilde{\rho}_k|^2\langle\rho_k, \epsilon\rangle^2.$$

Now let us write $\tilde{X}_{k^*} = [\tilde{X}_k \mid \tilde{Z}_k]$ and $\tilde{R}_k = \tilde{Z}_k - \tilde{X}_k(X_k'X_k)^{-1}X_k'Z_k$. Furthermore, for $k > k^*$, write

$$C_{k+1} = \begin{bmatrix} C_k & D_{k,k+1} \\ D_{k,k+1} & E_{k,k+1} \end{bmatrix}$$

and similarly for $\tilde{C}_{k+1}$. Finally, denote by $\Delta_{k,k+1}$ and Δ_k, the differences $\tilde{C}_k^{-1} \cdot \tilde{D}_{k,k+1} - C_k^{-1} D_{k,k+1}$ and $\tilde{C}_k^{-1}\tilde{D}_{k,k^*} - C_k^{-1} D_{k,k^*}$, respectively.

The following formulae bear a close resemblance to ones obtained in a similar context by Box and Draper (1959, 1963). There, however, the emphasis is on design: the choice of x vectors. It should be clear from the context whether or not $k \leq k^*$ is required to give a non-trivial result.

LEMMA 5.7. *As $m, n \to \infty$ we have*

(i) $m^{-1}\tilde{X}_k'\tilde{R}_k \to \tilde{C}_k\Delta_k$.

(ii) $m^{-1}\tilde{R}_k'\tilde{R}_k \to \tilde{E}_k - \tilde{D}_{k,k^*}'\tilde{C}_k^{-1}\tilde{D}_{k,k^*} + \Delta_k'\tilde{C}_k^{-1}\Delta_k$.

(iii) $m^{-1}|\tilde{\rho}_k|^2 \to \tilde{E}_{k,k+1} - \tilde{D}_{k,k+1}'\tilde{C}_k^{-1}\tilde{D}_{k,k+1} + \Delta_{k,k+1}'C_k^{-1}\Delta_{k,k+1}$.

(iv) $n^{-1}|\rho_k|^2 \to E_{k,k+1} - D_{k,k+1}'C_k^{-1}D_{k,k+1}$.

(v) $nm^{-2}|\eta_k|^2 \to \Delta_{k,k+1}'\tilde{C}_k C_k^{-1}\tilde{C}_k\Delta_{k,k+1}$.

PROOFS. These are all straightforward consequences of the relevant definitions. □

Next we extend some earlier notation, writing $B_{m,n}(k) = \mathrm{tr}\{\tilde{R}_k'\tilde{R}_k\zeta(k)\zeta(k)'\}$, and $S_{m,n}(k) = 2\langle\tilde{R}_k\zeta(k), \tilde{X}_k(X_k'X_k)^{-1}X_k'\epsilon\rangle$. Clearly the first term is the analogue of the bias term introduced prior to Lemma 5.2, and reduces to it if $m = n$ and $\tilde{X} = X$. For the definition of $\mathrm{PE}(k)$, see Section 2 above.

LEMMA 5.8. *In the notation just introduced, we have*

$$\mathrm{PE}(k) - m\sigma^2 = B_{m,n}(k) + N_{m,n}(k) - S_{m,n}(k).$$

PROOF. $\mathrm{PE}(k) - m\sigma^2 = |\tilde{X}_{k^*}\beta(k^*) - \tilde{X}_k\hat{\beta}(k)|^2$, where we may write

$$\begin{aligned}\tilde{X}_{k^*}\beta(k^*) - \tilde{X}_k\hat{\beta}(k) &= \tilde{X}_{k^*}\beta(k^*) - \tilde{X}_k(X_k'X_k)^{-1}X_k'(X_{k^*}\beta(k^*) + \epsilon) \\ &= (\tilde{Z}_k - \tilde{X}_k(X_k'X_k)^{-1}X_k'Z_k)\zeta(k) - \tilde{X}_k(X_k'X_k)^{-1}X_k'\epsilon.\end{aligned}$$

The result now follows upon taking the squared norm of this vector. □

LEMMA 5.9. *As $m, n \to \infty$ we have*

(i) $m^{-1}B_{m,n}(k) \to \mathrm{tr}\{(\tilde{E}_k - \tilde{D}_{k,k^*}'\tilde{C}_k^{-1}\tilde{D}_{k,k^*} + \Delta_k'\tilde{C}_k^{-1}\Delta_k)\zeta(k)\zeta(k)'\}$.

(ii) $m^{-1}n\boldsymbol{E}\{N_{m,n}(k)\} \to \mathrm{tr}(\tilde{C}_k C_k^{-1})$.

(iii) $m^{-1}nN_{m,n}(k) = O(\log\log n)$ *a.s.*

(iv) $m^{-1}nS_{m,n}(k) \to 0$ *a.s. if* $\Delta_k = 0$.

(v) $m^{-1}S_{m,n}(k) = O((n^{-1}\log\log n)^{1/2})$ *a.s. if* $\Delta_k \neq 0$.

PROOF. (i) is an immediate consequence of Lemma 5.7(iv); (ii) and (iii) are straightforward calculations; (iv) follows from the definitions, whilst (v) is a now-familiar form of the law of the iterated logarithm. □

PROOF OF THEOREM 3.1. (i) We begin by obtaining some probability inequalities concerning the terms in $\mathrm{APE}_n(k)$, cf. Lemma 5.2. Since $N_n(k) = |H_n(k)\epsilon|^2$ is a chi-squared r.v.,

$$\mathrm{pr}(N_n(k) > \beta_n) \le O(\exp(-\beta_n)) \quad \text{as} \quad n \to \infty.$$

Similarly, $B_n^\dagger(k)$ is a sum of independent zero mean normal r.v.'s whose variance is $O(n)$, and so $\mathrm{pr}(|B_n^\dagger(k)| > \gamma_n) \le O(\gamma_n^{-1} n^{1/2} \exp(-\gamma_n^2/2n))$.

Finally, $W_n(k) = V_n(k) + N_n^\dagger(k)$ is a sum of $n-k$ independent squared normals, the t-th of which is scaled by $\mu_t(k)$, and so

$$\begin{aligned}\mathrm{pr}(W_n(k) > \delta_n) &\le \exp(-\delta_n) \prod_{k+1}^{n} (1 - 2\mu_t(k))^{-1/2} \le \exp\left\{-\delta_n + \sum_{k+1}^{n} \mu_t(k)\right\} \\ &= \exp\{-\delta_n + k \log n + o(\log n)\} \\ &\le n^{k+1} \exp(-\delta_n), \quad \text{as} \quad n \to \infty.\end{aligned}$$

We now put these inequalities together, select (β_n), (γ_n) and (δ_n), and obtain (i). For simplicity, we drop subscripts n where no confusion will result. If $k < k^*$,

$$\begin{aligned}\mathrm{pr}(\hat{k} = k) &\le \mathrm{pr}\{\mathrm{APE}(k) < \mathrm{APE}(k^*)\} \\ &= \mathrm{pr}\{B(k) - N(k) + W(k) + B^\dagger(k) \\ &\qquad < B(k^*) - N(k^*) + W(k^*) + B^\dagger(k^*)\} \\ &\le \mathrm{pr}\{W(k^*) \ge B(k) + B^\dagger(k) - N(k)\} \\ &\qquad\qquad \text{since} \quad W(k) > 0 \quad \text{and} \quad N(k^*) > 0, \\ &\le \mathrm{pr}\{W(k^*) \ge nb_k + o(n) - \gamma_n - B_n\} \\ &\quad + P\{N(k) > B_n\} + P\{|B^\dagger(k)| > \gamma_n\} \\ &\le n^{k+1} \exp(-nb_k + o(n) + \gamma_n + \beta_n) \\ &\quad + O(\exp(-\beta_n)) + O(\gamma_n^{-1} n^{1/2} \exp(-\gamma_n^2/2n)).\end{aligned}$$

We now see that if $\beta_n = b_k n/3$ and $\gamma_n = b_k n/3$, the desired conclusion follows since b_k decreases as k increases to $k^* - 1$.

(ii) For the overfitting probability, we estimate $\mathrm{pr}(\hat{k} = k)$ for $k > k^*$, noting that in this case $\mathrm{APE}(k) = V(k) - N(k) + N^\dagger(k)$, i.e. the bias terms disappear. In this proof we bound $-N^\dagger(k)$ and $N^\dagger(k^*)$ from below by the same quantity, β_n say, and calculate the tail probability as in the first part of the proof. We find that

$$\begin{aligned}\mathrm{pr}(N^\dagger(k) < -\beta_n) &= \mathrm{pr}(-N^\dagger(k) > \beta_n) \\ &\le \exp(-\beta_n) \prod_{k+1}^{n} \{(1 + 2\mu_t(k))^{-1/2} \exp \mu_t(k)\} \\ &\le O(\exp(-\beta_n)).\end{aligned}$$

Similarly we have $\mathrm{pr}(N^\dagger(k^*) > \beta_n) \le O(\exp(-\beta_n))$, and since $N(k) - N(k^*)$ is a chi-squared r.v. on $k - k^*$ degrees of freedom,

$$\mathrm{pr}(N(k) - N(k^*) > \gamma_n) \le O(\gamma_n^{-1+(k-k^*)/2} \exp(-\gamma_n/2)).$$

Thus if $k > k^*$,

$$\begin{aligned}\operatorname{pr}(\hat{k} = k) &= \operatorname{pr}\{\mathrm{APE}(k) < \mathrm{APE}(k^*)\} \\ &= \operatorname{pr}\{V(k) - N(k) + N^\dagger(k) < V(k^*) - N(k^*) + N^\dagger(k^*)\} \\ &\le \operatorname{pr}\{V(k) - \beta_n - (N(k) - N(k^*)) < V(k^*) + \beta_n\} \\ &\quad + \operatorname{pr}\{N^\dagger(k) < -\beta_n\} + \operatorname{pr}\{N^\dagger(k) > \beta_n\} \\ &\le O(\gamma_n^{-1+(k-k^*)/2} \exp(-\gamma_n/2)) + 2O(\exp(-\beta_n)),\end{aligned}$$

where $\gamma_n = (k - k^*)\log n + o(\log n) - 2\beta_n$, since $V(k) = k \log n + o(\log n)$, and similarly for $V(k^*)$. If we take $\beta_n = \beta \log n$ for $\beta = 6^{-1}$, say, then we deduce that $\operatorname{pr}(\hat{k} > k) \le O(n^{-1/6})$. □

Corollary 3.2 can be shown by an argument similar to Theorems 2.1 and 2.3. Note that when the selection rule is not consistent, the inequality is sharp since the prediction error based on M_k for some $k > k^*$ is strictly larger than the one based on M_{k^*}, and underfitting does not cause any problem since all FPE's underfit with a probability vanishing exponentially fast (Theorem 3.1(i)).

Let $\{H_j : j = 1, \ldots, n\}$ be a set of pairwise orthogonal rank 1 projectors summing to the identity, such that for all $k = 1, \ldots, K$ we have $\sum_{p=1}^{k} H_p = H(k)$, where $R(H(k)) = R(X_k(n))$. Let $\epsilon = (\epsilon_i)$ be an n-tuple of iid $N(0, 1)$ random variables, F any function of $|H_i\epsilon|^2$ for a fixed $i \in \{1, \ldots, n\}$, and ξ, η fixed vectors.

LEMMA 5.10. $\boldsymbol{E}\{\langle x_i, H_i\epsilon\rangle F(|H_i\epsilon|^2)\} = 0$.

PROOF. The lemma is an immediate consequence of the symmetry of the normal distribution. □

COROLLARY 5.3. *Let f be a function of $|H_1\epsilon|^2, \ldots, |H_k\epsilon|^2$. Then if $1 \le i, j \le k$, we have*

$$\begin{aligned}&\boldsymbol{E}\{\langle \xi, H_i\epsilon\rangle f(|H_1\epsilon|^2, \ldots, |H_k\epsilon|^2)\} = 0, \\ &\boldsymbol{E}\{\langle \xi, H_i\epsilon\rangle\langle \eta, H_j\epsilon\rangle f(|H_1\epsilon|^2, \ldots, |H_k\epsilon|^2)\} = 0.\end{aligned}$$

PROOF. The identities follow from the lemma by a suitable conditioning. □

In the lemma which follows we use the expressions ρ_k and η_k defined prior to Lemma 5.6 above.

LEMMA 5.11. *Let $\hat{k}_n$ denote the dimension selected by FPE_{α_n} and suppose that $l > k \ge k^*$. Then we have*

$$\lim_{m,n} m^{-1} n |\rho_k|^{-2} \boldsymbol{E}\{\langle \rho_k, \epsilon\rangle\langle \eta_k, \epsilon\rangle 1_{\{\hat{k}_n = l\}}\} = 0. \tag{5.4}$$

PROOF. We begin by replacing $\hat{k}_n$ by $\tilde{k}_n$, that k which minimizes $\mathrm{FPE}(k)$ over the range $\{k^*, k^* + 1, \ldots, K\}$. From Theorem 3.2 we know that $\operatorname{pr}(\hat{k}_n \ne \tilde{k}_n) \to 0$ as $n \to \infty$.

Now recall the definition of FPE(k) and note that if $k < l$, FPE(k) $\leq$ FPE(l) if and only if $\sum_{k+1}^{l} |H_p\epsilon|^2 \leq (l-k)\alpha$. Thus the event $\{\tilde{k} = l\}$ is the intersection of the two events: $\{\sum_{p=h+1}^{l} |H_p\epsilon|^2 \geq (l-h)\alpha; k^* \leq h < l\}$ and $\{\sum_{p=l+1}^{h} |H_p\epsilon|^2 \leq (h-l)\alpha, l < h \leq K\}$ whose indicators we denote by f_l and g_l respectively. Our aim is to show that

$$\boldsymbol{E}\{\langle \rho_k, \epsilon\rangle \langle \eta_k, \epsilon\rangle f_l g_l\} = 0 \tag{5.5}$$

and then deduce the conclusion of the lemma.

Since $\eta_k \in R(X_k)$, we may write $\langle \eta_k, \epsilon\rangle = \sum_{i=1}^{k} \langle \eta_k, H_i\epsilon\rangle$. Similarly, $\rho_k \in R(X_k)^{\perp}$ and so $\langle \rho_k, \epsilon\rangle = \sum_{j=k+1}^{n} \langle \rho_k, H_j\epsilon\rangle$. Thus our interim objective will be achieved if we can prove that for all i, j, $1 \leq i \leq k$, $k+1 \leq j \leq n$, we have

$$\boldsymbol{E}\{\langle \eta_k, H_i\epsilon\rangle \langle \rho_k, H_j\epsilon\rangle f_l g_l\} = 0. \tag{5.6}$$

Note that f_l is a function of $\{|H_p\epsilon|^2 : k^* < p \leq l\}$ whilst g_l is a function of $\{|H_p\epsilon|^2 : l < p \leq K\}$, and so if $i \leq k^*$ or $j > k$, (5.6) is trivially zero. If we take the case $k^* \leq i, j \leq l$, we can split off g_l by independence and use Corollary 5.3 to get the conclusion. Similarly if $k^* \leq i \leq l$ and $l < j \leq K$, we can again use independence this time splitting off $\langle \eta_k, H_i\epsilon\rangle f_l$, and again getting zero by the same corollary. Thus (5.6) and hence (5.5) are established.

The proof is completed by noting that $\lim_{m,n} m^{-1}n|\rho_k|^{-2}\boldsymbol{E}|\langle \eta_k, \epsilon\rangle \langle \rho_k, \epsilon\rangle|$ is finite, and so we can combine the result $\mathrm{pr}(\tilde{k}_n \neq \hat{k}_n) \to 0$ as $n \to \infty$ with (5.5) to obtain (5.4). □

PROOF OF THEOREM 2.2. We obtain (2.6) under each of the three conditions in turn; in all cases making use of Lemmas 5.8 and 5.9. Then by Lemma 5.8, the left-hand side of (2.6) will be $O(n)$ as $m, n \to \infty$, since the bias terms $nB_{m,n}(k)$ for $k < k^*$ are not all eliminated, and these are $O(n)$ as $m, n \to \infty$, and cannot be canceled by either of the noise terms. Thus (2.6) is trivially true. Now let us assume (B). By virtue of the result just established, we may also suppose that $\mathrm{pr}(\hat{k}_n < k^*) \to 0$ as $n \to \infty$. Otherwise we make no assumptions concerning the selection procedure $\hat{k}$. On the set $\{\hat{k} > k^*\}$, $B_{m,n}(\hat{k}) = S_{m,n}(\hat{k}) = 0$, and so $\mathrm{PE}(\hat{k}) - m\sigma^2 = N_{m,n}(\hat{k})$.

Our proof begins by observing that

$$\begin{aligned}
&\lim_{m,n} nm^{-1}\boldsymbol{E}||\rho_k|^{-2}\langle \eta_k, \epsilon\rangle \langle \rho_k, \epsilon\rangle| \\
&\quad \leq \lim_{m,n} nm^{-1}|\rho_k|^{-2}\{\boldsymbol{E}\langle \eta_k, \epsilon\rangle^2 \boldsymbol{E}\langle \rho_k, \epsilon\rangle^2\}^{1/2} \\
&\quad = \lim_{m,n} nm^{-1}|\rho_k|^{-2}\{|\eta_k|^2|\rho_k|^2\}^{1/2},
\end{aligned}$$

and this limit is zero by Lemma 5.7 and (B).

Repeated application of this result and Corollary 5.2 give a series of inequalities, which imply that for $k > k^*$:

$$\lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(k)1_{\{\hat{k}=k\}}\} \geq \lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(k^*)1_{\{\hat{k}=k\}}\},$$

whence $\lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(\hat{k})1_{\{\hat{k}\geq k^*\}}\} \geq \lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(k^*)1_{\{\hat{k}\geq k^*\}}\}$. Since $\mathrm{pr}(\hat{k}_n \geq k^*) \to 1$ as $n \to \infty$, and $N_{n,m}(k^*) \geq 0$, $\lim_{m,n} nm^{-1}\boldsymbol{E}\{N_{m,n}(k^*)\} = \mathrm{tr}\{\tilde{C}_{k^*}C_{k^*}^{-1}\}$ implies (2.6) in case (B).

Finally we consider case (C). The proof goes as for case (B), and in particular the selection rules $\hat{k}$ based on FPE_{α_n} for α_n such that $n^{-1}\alpha_n \to 0$ as $n \to \infty$, overfit with probability approaching unity by Theorem 3.2. The chain of inequalities leading to the final conclusion is also true, but this time the individual steps are justified by Theorem 3.1, and the proof is completed exactly as it was in case (B). Any other selection rule for which the same symmetry argument is valid also has the lower bound. □

PROOF OF THEOREM 2.3. (i) We begin by proving that the underfitting contribution to the left-hand side of (2.6) is asymptotically negligible. This follows from the readily checked fact that when $k < k^*$, $nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(k) - m\sigma^2)\} \leq O(n)$ as $m, n \to \infty$. Thus for all $k < k^*$,

$$nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}=k\}}\} \leq O(n)\sqrt{\mathrm{pr}(\hat{k}_n = k)} \to 0$$

as $m, n \to \infty$, and so $nm^{-1}\boldsymbol{E}\{\mathrm{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}<k^*\}}\} \to 0$ as $n, m \to \infty$.

Turning now to the overfitting contribution, we begin by proving that in the chain of inequalities used to prove the lower bound in cases (B) and (C), the terms dropped—the second and third terms of the right-hand side of Corollary 5.2—all have absolute expectations which are $O(mn^{-1})$. The argument at the beginning of the proof of case (B) of Theorem 2.2 shows this for the second term, for even without the hypothesis (B) we get a constant at that stage by Lemma 5.7(v). Similarly for the third terms,

$$\lim_{m,n} nm^{-1}\boldsymbol{E}\{|\rho_k|^{-4}|\tilde{\rho}_k|^2\langle\rho_k, \epsilon\rangle^2\} = O(1)$$

by Lemma 5.7. Thus we may use the consistency hypothesis and get

$$\begin{aligned}
&\lim_{m,n} nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}>k^*\}}\} \\
&\quad = \sum_{k=k^*+1}^{K} \lim_{m,n} nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}=k\}}\} \\
&\quad = \sum_{k=k^*+1}^{K} \lim nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(k^*) - m\sigma^2)1_{\{\hat{k}=k\}}\} \\
&\quad = \lim_{m,n} nm^{-1}\boldsymbol{E}(\mathrm{PE}(k^*) - m\sigma^2) = \mathrm{tr}\{\tilde{C}_{k^*}C_{k^*}^{-1}\},
\end{aligned}$$

the second last step following from our assumption that $\mathrm{pr}(\hat{k}_n = k) \to 0$ as $n \to \infty$ for all $k > k^*$. This completes the proof of (i).

(ii) Now we suppose that $\hat{k}$ is obtained by minimizing FPE_{α_n} for a sequence $\alpha_n < 2\log\log n$. We know from Theorem 3.2 that $\mathrm{pr}(\hat{k} < k^*) = o(n^{-1})$ and so

need only consider overfitting. By Shibata (1984), $\liminf \mathrm{pr}(\hat{k}_n = k^* + 1) > 0$. We next simplify $\lim_{m,n} nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(\hat{k}) - m\sigma^2)\}$ in the now familiar way, noting that (as in the proof of Theorem 2.2) it coincides with

$$\lim_{m,n} nm^{-1}\boldsymbol{E}\{(\mathrm{PE}(\hat{k}) - m\sigma^2)1_{\{\hat{k}\geq k^*\}}\}$$
$$\geq \mathrm{tr}\{\tilde{C}_{k^*}C_{k^*}^{-1}\} + \lim_{m,n} nm^{-1}\boldsymbol{E}\{|\rho_{k^*}|^{-4}|\tilde{\rho}_{k^*}|^2\langle\rho_{k^*},\epsilon\rangle^2 1_{\{\hat{k}=k^*+1\}}\}.$$

Now the second term above is zero only if $\rho_{k^*} = 0$, which implies $k^* = K$, since we have assumed all design matrices to be of full rank. Thus the inequality (2.6) is strict for selection rules based on FPE_{α_n} with $\liminf(2\log\log n)^{-1}\alpha_n < 1$. □

PROOF OF THEOREM 2.1. Since ϵ_t is independent of $\hat{k}_{t-1}$ and $\hat{\beta}_{t-1}$ for all $t > 1$,

$$\boldsymbol{E}\left\{\sum_1^n (y_t - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2\right\} = n\sigma^2 + \sum_1^n \boldsymbol{E}(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2.$$

Write

$$U_n = \sum_1^n \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2 1_{\{\hat{k}_{t-1}<k^*\}}\},$$
$$V_n = \sum_1^n \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2 1_{\{\hat{k}_{t-1}=k^*\}}\},$$
$$W_n = \sum_1^n \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2 1_{\{\hat{k}_{t-1}>k^*\}}\}.$$

We deal with each of these three components in turn. Let us temporarily denote $x_t'(X_k(t-1)'X_k(t-1))^{-1}X_k(t-1)'\epsilon(t-1)$ by $d'\epsilon$. Then

$$U_n = \sum_{k=1}^{k^*-1}\sum_{t=1}^n \boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(\hat{k}_{t-1}))^2 1_{\{\hat{k}_{t-1}=k\}}\}$$
$$= \sum_{k=1}^{k^*-1}\sum_{t=1}^n \boldsymbol{E}\{(\lambda_t(k) - d'\epsilon)^2 1_{\{\hat{k}_{t-1}=k\}}\}$$
$$\leq 2\sum_{k=1}^{k^*-1}\sum_{t=1}^n [\lambda_t(k)^2\mathrm{pr}(\hat{k}_{t-1} = k) + 2\boldsymbol{E}\{(d'\epsilon)^2 1_{\{\hat{k}_{t-1}=k\}}\}].$$

Now for $k < k^*$, $\sum_1^n \lambda_t(k)^2 = b_k n + o(1)$ as $n \to \infty$, whilst $\mathrm{pr}(\hat{k}_{t-1} = k) \leq O(t^{-2}(\log t)^{-c})$ as $n \to \infty$, $c > 1$. Summing by parts we thus conclude that

$$\sum_{k=1}^{k^*-1}\sum_{t=1}^n \lambda_t(k)^2\mathrm{pr}(\hat{k}_{t-1} = k) = O(1) \quad \text{as} \quad n \to \infty.$$

Furthermore, $\boldsymbol{E}\{(d'\epsilon)^4\} = 3\boldsymbol{E}\{(d'\epsilon)^2\}$, and since $\boldsymbol{E}(d'\epsilon)^2 = |d|^2\sigma^2 = \mu_t(k)\sigma^2$,

$$\sum_{k=1}^{k^*-1}\sum_{t=1}^{n}\boldsymbol{E}\{(d'\epsilon)^2 1_{\{\hat{k}_{t-1}=k\}}\} \le \sum_{k=1}^{k^*-1}\sum_{t=1}^{n}\sqrt{3}\sigma^2\mu_t(k)\{\mathrm{pr}(\hat{k}_{t-1}=k)\}^{1/2}$$
$$= O(1) \quad \text{as} \quad n \to \infty,$$

as argued above, but this time using $\sum_1^n \mu_t(k) = k\log n(1+o(1))$ as $n \to \infty$. Thus $U_n = O(1)$ as $n \to \infty$.

Turning now to the overfitting term V_n, we find only the quadratic form $(d'\epsilon)^2$, as the bias term vanishes. Thus we can argue as above, giving

$$W_n = \sum_{k=k^*+1}^{K}\sum_{t=1}^{n}\boldsymbol{E}\{(d'\epsilon)^2 1_{\{\hat{k}_{t-1}=k\}}\}$$
$$\le \sqrt{3}\sigma^2 \sum_{k=k^*+1}^{K}\sum_{t=1}^{n}\mu_t(k)\{\mathrm{pr}(\hat{k}_{t-1}=k)\}^{1/2} = O(1),$$

since $\mathrm{pr}(\hat{k}_{t-1}=k) \le O(\log t^{-\alpha})$ as $t \to \infty$, where $\alpha > 2$.

Finally, we examine the term corresponding to getting the model correct. Since $\mathrm{pr}(\hat{k}_{t-1} \neq k^*) \le A(t^{-2}(\log t)^{-c}) + B(\log t)^{-\alpha}$ for large t,

$$V_n = \sum_{t=1}^{n}\boldsymbol{E}\{(x_t'\beta^* - x_t'\hat{\beta}_{t-1}(k^*))^2 1_{\{\hat{k}_{t-1}=k^*\}}\}$$
$$= \sum_{t=1}^{n}\boldsymbol{E}\{(d'\epsilon)^2\} - \sum_{t=1}^{n}\boldsymbol{E}\{(d'\epsilon)^2 1_{\{\hat{k}_{t-1}\neq k^*\}}\}$$
$$= k^*\log n(1+o(1)) + O(1) \quad \text{as} \quad n \to \infty. \qquad \square$$

Acknowledgements

We would like to thank Jorma Rissanen for his inspiration and for many useful discussions. Special thanks are due to David Freedman for his criticisms of the manuscript.

References

Akaike, H. (1970). Statistical predictor identification, *Ann. Inst. Statist. Math.*, **22**, 202–217.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Control*, **19**, 716–723.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model, *Biometrika*, **67**, 413–418.

Bhansali, R. H. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion, *Biometrika*, **64**, 547–551.

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a regression surface design, *J. Amer. Statist. Assoc.*, **54**, 622–654.

Box, G. E. P. and Draper, N. R. (1963). The choices of a second order rotatable design, *Biometrika*, **50**, 335–352.

Breiman, L. A. and Freedman, D. F. (1983). How many variables should be entered in a regression equation?, *J. Amer. Statist. Assoc.*, **78**, 131–136.

Clayton, M. K., Geisser, S. and Jennings, D. (1986). A comparison of several model selection procedures, *Bayesian Inference and Decision* (eds. P. Goel and A. Zellner), 425–439, Elsevier, New York.

Dawid, A. P. (1984). Present position and potential developments: some personal views, Statistical theory—The prequential approach (with discussion), *J. Roy. Statist. Soc. Ser. A*, **147**, 278–292.

Dawid, A. P. (1992). Prequential data analysis, *Current Issues in Statistical Inference: Essays in Honor of D. Basu, Institute of Mathematical Statistics, Monograph*, **17** (eds. M. Ghosh and P. K. Pathak).

Geweke, J. and Meese, R. (1981). Estimating regression models of finite but unknown order, *Internat. Econom. Rev.*, **22**, 55–70.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression, *J. Roy. Statist. Soc. Ser. B*, **41**, 190–195.

Hannan, E. J., McDougall, A. J. and Poskitt, D. S. (1989). Recursive estimation of autoregressions, *J. Roy. Statist. Soc. Ser. B*, **51**, 217–233.

Haughton, D. (1989). Size of the error in the choice of a model to fit data from an exponential family, *Sankhyā Ser. A*, **51**, 45–58.

Hemerly, E. M. and Davis, M. H. A. (1989). Strong consistency of the predictive least squares criterion for order determination of autoregressive processes, *Ann. Statist.*, **17**, 941–946.

Hjorth, U. (1982). Model selection and forward validation, *Scand. J. Statist.*, **9**, 95–105.

Kohn, R. (1983). Consistent estimation of minimal dimension, *Econometrica*, **51**, 367–376.

Lai, T., Robbins, H. and Wei, C. Z. (1979). Strong consistency of least squares estimates in multiple regression II, *J. Multivariate Anal.*, **9**, 343–361.

Merhav, N., Gutman, M. and Ziv, J. (1989). On the estimation of the order of a Markov chain and universal data compression, *IEEE Trans. Inform. Theory*, **39**, 1014–1019.

Nishi, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression, *Ann. Statist.*, **12**, 758–765.

Rao, C. R. (1973). *Linear Statistical Inference*, 2nd ed., Wiley, New York.

Rissanen, J. (1984). Universal coding, information prediction, and estimation, *IEEE Trans. Inform. Theory*, **30**, 629–636.

Rissanen, J. (1986*a*). Stochastic complexity and modeling, *Ann. Statist.*, **14**, 1080–1100.

Rissanen, J. (1986*b*). A predictive least squares principle, *IMA J. Math. Control Inform.*, **3**, 211–222.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*, World Books, Singapore.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models, *Econometrica*, **46**, 1273–1291.

Schwartz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461–464.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, **63**, 117–126.

Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45–54.

Shibata, R. (1983*a*). Asymptotic mean efficiency of a selection of regression variables, *Ann. Inst. Statist. Math.*, **35**, 415–423.

Shibata, R. (1983*b*). A theoretical view of the use of AIC, *Times Series Analysis: Theory and Practice 4* (ed. O. D. Anderson), 237–244, Elsevier, Amsterdam.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables, *Biometrika*, **71**, 43–49.

Shibata, R. (1986*a*). Selection of the number of regression variables; a minimax choice of generalized FPE, *Ann. Inst. Statist. Math.*, **38**, 459–474.

Shibata, R. (1986*b*). Consistency of model selection and parameter estimation, *Essays in Time Series and Allied Processes: Papers in Honour of E. J. Hannan, J. Appl. Probab.*, **23A**, 127–141.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. Roy. Statist. Soc. Ser. B*, **39**, 44–47.

Wax, M. (1988). Order selection for AR models by predictive least squares, *IEEE Trans. Acoust. Speech Signal Process.*, **36**, 581–588.

Wei, C. Z. (1992). On the predictive least squares principle, *Ann. Statist.*, **20**, 1–42.

Woodroofe, M. (1982). On model selection and the arc sine laws, *Ann. Statist.*, **10**, 1182–1194.