Econometric Theory, **18**, 2002, 913–925. Printed in the United States of America.
DOI: 10.1017.S0266466602184052

ON VARIABLE SELECTION IN LINEAR REGRESSION

PAUL KABAILA
La Trobe University

Shibata (1981, *Biometrika* 68, 45–54) considers data-generating mechanisms belonging to a certain class of linear regressions with errors that are independent and identically normally distributed. He compares the variable selection criteria AIC (Akaike information criterion) and BIC (Bayesian information criterion) using the following type of comparison. For each fixed possible data–generating mechanism, these criteria are compared as the data length increases. The results of this comparison have been interpreted as meaning that, in the context of the data-generating mechanisms considered by Shibata, AIC is better than BIC for large data lengths. Shibata's comparison is pointwise in the space of data–generating mechanisms (as the data length increases). Such comparisons are potentially misleading. We consider a simple class of data-generating mechanisms satisfying Shibata's assumptions and carry out a different type of comparison. For each fixed data length (possibly large) we compare the variable selection criteria for every possible data-generating mechanism in this class. According to this comparison, for this class of data-generating mechanisms no matter how large the data length AIC is not better than BIC.

1. INTRODUCTION

Shibata (1981) considers data-generating mechanisms belonging to a certain class of linear regressions with errors that are independent and identically normally distributed. He compares AIC (Akaike information criterion) and BIC (Bayesian information criterion) using the following type of comparison. For each fixed possible data-generating mechanism, these criteria are compared as the data length increases. The results of this comparison have been interpreted (see, e.g., Shibata, 1981, 1986) as meaning that, in the context of the data-generating mechanisms considered by Shibata (1981), AIC is better than BIC for large data lengths. Shibata's comparison is pointwise in the space of data-generating mechanisms (as the data length increases).

A pointwise limit of a sequence of functions may have features not evident in any element of the sequence, no matter how far we proceed along the sequence. In the context of statistical theory, Hajek (1971) states that "Especially misinformative can be those limit results that are not uniform. Then the limit

Address correspondence to: Paul Kabaila, Department of Statistical Science, La Trobe University, Victoria, 3086, Australia.

© 2002 Cambridge University Press 0266-4666/02 $9.50

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

may exhibit features that are not even approximately true for any finite n." Illustrations of this point, in the context of confidence regions and prediction regions, are provided by Kabaila (1995) (cf. Pötscher, 1991, 1995). The simplest illustration of this point, in the context of statistical theory, is provided by the famous example due to Hodges of a "superefficient" estimator (see, e.g., Lehmann, 1983). In other words, Shibata's comparison, which is pointwise in the space of data-generating mechanisms, is potentially misleading.

In Section 2 we describe a simple class of data-generating mechanisms satisfying Assumption 1 and Condition 2 of Shibata (1981). In Section 3 we describe the comparison of AIC and BIC for this simple class, using the approach of Shibata (1981). In Section 4, we carry out a different type of comparison of AIC and BIC in the context of this simple class of data-generating mechanisms. For each fixed data length (possibly large) we compare these criteria for every possible data-generating mechanism in this class. According to this comparison, for this class of data-generating mechanisms no matter how large the data length AIC is not better than BIC.

2. A SIMPLE CLASS OF DATA-GENERATING MECHANISMS SATISFYING ASSUMPTION 1 AND CONDITION 2

In this section we describe a simple class of data-generating mechanisms satisfying Assumption 1 and Condition 2 of Shibata (1981). We use an almost identical notation to that of Shibata (1981). Let ℓ_2 denote the Hilbert space of vectors $\alpha = [\alpha_1, \alpha_2, \ldots]'$, satisfying $\sum_{i=1}^{\infty} \alpha_i^2 < \infty$, with the usual inner product $\langle \cdot, \cdot \rangle$ and norm $\|\alpha\| = \langle \alpha, \alpha \rangle^{1/2}$. Suppose that f is a twice differentiable even function with period 2π. We can express f as follows:

$$f(t) = \sum_{k=0}^{\infty} \beta_{k+1} \frac{\cos(kt)}{k+1} \qquad \text{for all } t,$$

where $\beta = [\beta_1, \beta_2, \ldots]'$ is a vector in ℓ_2. This is the same f as that described in Example 3.2 of Shibata (1981), apart from a change of scale for t. We may write $f(t) = \langle x, \beta \rangle$ where

$$x = \left[1, \frac{\cos(t)}{2}, \frac{\cos(2t)}{3}, \frac{\cos(3t)}{4}, \ldots\right]'$$

is a vector in ℓ_2.

Suppose that we observe $Y = [Y(t_0), \ldots, Y(t_{n-1})]'$ where

$$Y(t_j) = f(t_j) + \varepsilon_j \qquad \text{for } j = 0, \ldots, n-1,$$

$t_j = 2\pi j/n$, and the ε_j's are independent and identically $N(0, \sigma^2)$ distributed. The data-generating mechanism is specified by (β, σ). It is convenient to express Y as follows:

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

$$Y = \sum_{k=1}^{\infty} \frac{\beta_k}{k} c_k + \varepsilon,$$

where $\varepsilon = [\varepsilon_0, \ldots, \varepsilon_{n-1}]'$ and the c_k's are n-vectors defined by

$$c_k = [\cos((k-1)t_0), \ldots, \cos((k-1)t_{n-1})]'.$$

Suppose that $n > 1$ and define N to be the greatest integer less than or equal to $n/2$. Some basic properties of the c_i's are described in Lemma A.1, which is stated and proved in Appendix A. Define $X(k)$ to be the $n \times k$ matrix with ith column $i^{-1}c_i$. The family of models fitted to the data is indexed by $j \in \{1, \ldots, K_n\}$ and the kth model in this class is

$$Y = X(k)\beta(k) + \varepsilon$$

with $\beta(k)$ an unknown k-vector. Note that this is a nested family of models. We make the following assumption about $\{K_n\}$.

Assumption A. $\{K_n\}$ is a sequence of positive integers satisfying $K_n < N$ for all n, $K_n = o(n)$, and $K_n \to \infty$ as $n \to \infty$.

Also define $M_n(j) = X(j)'X(j)$ where $j \leq N$. Observe that, by Lemma A.1, $M_n(j)$ is a $j \times j$ diagonal matrix with nonzero diagonal elements. Clearly Assumption 1 of Shibata (1981) is satisfied.

Now define $M_n = X_n'X_n$ where X_n is the $n \times \infty$ matrix with ith column $i^{-1}c_i$. Also define, for $\alpha \in \ell_2$, $\|\alpha\|_{M_n}^2 = \langle M_n\alpha, \alpha\rangle$. Let $\beta^{(n)}(j)$ be the value of $\beta^* = [\beta_1^*, \ldots \beta_j^*, 0, 0, \ldots]'$ minimizing $\|\beta - \beta^*\|_{M_n}^2$ where $j < N$. A convenient expression for $\|\beta - \beta^{(n)}(j)\|_{M_n}^2$ is derived in Appendix B. We now make the following assumption about β.

Assumption B. β has infinitely many nonzero components.

Let $\mathcal{B}$ denote the set of $\beta \in \ell_2$ that satisfy Assumption B. By Lemma A.1, the $N \times N$ matrix in the upper left corner of $2n^{-1}M_n$ is a diagonal matrix with diagonal elements $2, (\frac{1}{2})^2, \ldots, (1/N)^2$. By Theorem 3.1 of Shibata (1981), his Condition 2 is satisfied. A direct check that this condition is satisfied is provided in Appendix B. Shibata (1981) states that his Assumption 2 can be replaced by his Condition 2.

Define $\hat{\beta}(j) = [M_n(j)]^{-1}X(j)'Y$ the least squares estimator of $\beta(j)$. We now describe two criteria whose minimization with respect to j $(1 \leq j \leq K_n)$ has been suggested as providing the appropriate choice of j and consequently $\hat{\beta}(j)$.

$$AIC(j) = n \log \hat{\sigma}^2(j) + 2j, \qquad \text{(Akaike, 1974)}$$

$$BIC(j) = n \log \hat{\sigma}^2(j) + [\log(n)]\, j. \qquad \text{(Schwarz, 1978)}$$

Here $\hat{\sigma}^2(j) = n^{-1}(Y - X(j)\hat{\beta}(j))'(Y - X(j)\hat{\beta}(j))$. Let $\hat{j}_n$ and $\tilde{j}_n$ denote the smallest value of $j \in \{1, \ldots, K_n\}$ minimizing $AIC(j)$ and $BIC(j)$, respectively.

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

3. A COMPARISON POINTWISE IN THE SPACE OF DATA-GENERATING MECHANISMS

Define

$$L_n(j) = E\|[\hat{\beta}(j)',0,0,\ldots]' - \beta\|^2_{M_n},$$

which, for fixed j, is a measure of how well $[\hat{\beta}(j)',0,0,\ldots]'$ estimates β. Note that

$$L_n(j) = \|\beta - \beta^{(n)}(j)\|^2_{M_n} + j\sigma^2. \tag{1}$$

Let j_n^* denote the value of $j \in \{1,\ldots,K_n\}$ minimizing $L_n(j)$. Now consider a choice, say, $\bar{j}_n$, of j based on the data Y. The measure of efficiency for this choice chosen by Shibata (1981) is

$$\frac{d(\hat{\beta}(\bar{j}_n),\beta)}{L_n(j_n^*)}, \tag{2}$$

where

$$d(\hat{\beta}(j),\beta) = \|[\hat{\beta}(j)',0,0,\ldots]' - \beta\|^2_{M_n}.$$

His Theorem 2.2 implies the following result.

THEOREM 3.1. *Suppose Assumptions A and B are satisfied. Then*

$$\frac{d(\hat{\beta}(\hat{j}_n),\beta)}{L_n(j_n^*)} \xrightarrow{p} 1. \tag{3}$$

Shibata (1981, p. 54) states, "Furthermore it can be shown that the *BIC* method proposed by Schwarz (1978) and Akaike (1978) is also not optimal in our sense." In other words, (3) does not, in general, hold if $\hat{j}_n$ is replaced by $\tilde{j}_n$. Shibata (1986) has concluded that, in the context of a linear regression model that is supposed to approximate (in a certain sense) the data-generating mechanism, AIC is better than BIC for large sample sizes (cf. Barron, Birge, and Massart, 1999, p. 307).

The results of Shibata (1981) concern the behavior of a particular measure of efficiency as $n \to \infty$ *pointwise* in the space of data-generating mechanisms specified by (β,σ) where $\beta \in \mathcal{B}$. As with the example of Hodges's "superefficient" estimator this is not really appropriate. Let $\bar{j}_n$ denote a choice of j based on the data Y. The measure of efficiency (2) of this choice is based on $d(\hat{\beta}(\bar{j}_n),\beta)$. Assuming that one insists on such a basis, one should compare $\hat{j}_n$ with $\tilde{j}_n$ in the following way. One should fix n and examine $d(\hat{\beta}(\hat{j}_n),\beta)/d(\hat{\beta}(\tilde{j}_n),\beta)$ $(= [d(\hat{\beta}(\hat{j}_n),\beta)/L_n(j_n^*)]/[d(\hat{\beta}(\tilde{j}_n),\beta)/L_n(j_n^*)])$ as we vary over all data-generating mechanisms specified by (β,σ), where $\beta \in \mathcal{B}$, prior to letting $n \to \infty$.

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

4. A DIFFERENT COMPARISON OF THE VARIABLE SELECTION CRITERIA

The following result, which is the main result of the paper, shows that for each sufficiently large n there exists (β,σ), where $\beta \in \mathcal{B}$, such that (a) BIC is as good as or better than AIC for every data vector Y and (b) BIC significantly outperforms AIC for a set of Y's with probability exceeding 0.13. In other words, if we fix n and examine the performance of AIC relative to BIC as we vary over the possible data-generating mechanisms then AIC cannot be said to be better than BIC, irrespective of how large n is.

THEOREM 4.1. *Fix $\{K_n\}$ satisfying both Assumption A and $n^{1/2}\log(n)K_n = o(n)$. Also fix σ. For each sufficiently large n there exists $\beta \in \mathcal{B}$ such that*

$$\frac{d(\hat{\beta}(\hat{j}_n),\beta)}{d(\hat{\beta}(\tilde{j}_n),\beta)} \geq 1$$

and

$$P\left(\frac{d(\hat{\beta}(\hat{j}_n),\beta)}{d(\hat{\beta}(\tilde{j}_n),\beta)} > 2\right) > 0.13.$$

Proof. Fix $\{K_n\}$ satisfying both Assumption A and $n^{1/2}\log(n)K_n = o(n)$. The set of $\beta \in \mathcal{B}$ such that $\|\beta - \beta^{(n)}(K_n)\|^2_{M_n} \to 0$ as $n \to \infty$ is nonempty; roughly speaking, we can choose $\beta \in \mathcal{B}$ such that the nth component of β converges to zero as $n \to \infty$ arbitrarily fast. Fix both $\tilde{\beta}$ an element of this set and σ^2.

Restrict attention to n sufficiently large that $\log(n) > 2$. The key observation is that, for each fixed n, $\beta = (\bar{\beta}_1,\ldots,\bar{\beta}_{K_n},\tilde{\beta}_{K_n+1},\tilde{\beta}_{K_n+2},\ldots) \in \mathcal{B}$ for every $(\bar{\beta}_1,\ldots,\bar{\beta}_{K_n}) \in \mathbf{R}^{K_n}$. Note that β depends on n; however, for the sake of readability, this has not been included explicitly in the notation. As proved in Appendix C, for $j = 1,\ldots,K_n$,

$$\frac{d(\hat{\beta}(j),\beta)}{\sigma^2} = \sum_{i=1}^{j} W_i + \sum_{i=j+1}^{K_n}\left(\frac{\mu_i(\beta)}{\sigma}\right)^2 + \frac{\|\beta - \beta^{(n)}(K_n)\|^2_{M_n}}{\sigma^2}, \tag{4}$$

where $W_1, W_2,\ldots$ are independent and identically χ^2_1 distributed and

$$\mu_k(\beta) = \sqrt{\frac{n}{2}}\left[\frac{\beta_k}{k} + \sum_{i=1}^{\infty}\left(\frac{\beta_{in+k}}{in+k} + \frac{\beta_{in+2-k}}{in+2-k}\right)\right] \qquad \text{for } k = 2,\ldots,N. \tag{5}$$

Note that we can choose $(\bar{\beta}_1,\ldots,\bar{\beta}_{K_n})$ so that $\mu_2(\beta),\ldots,\mu_{K_n}(\beta)$ take any desired values.

As proved in Appendix D, $\tilde{j}_n \leq \hat{j}_n$. We now choose $\mu_2(\beta),\ldots,\mu_{K_n}(\beta)$ so that AIC does not perform well by comparison with BIC. Choose $(\bar{\beta}_1,\ldots,\bar{\beta}_{K_n})$ such that $\mu_2(\beta) = \cdots = \mu_{K_n}(\beta) = 0$. It follows from (4) that

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

$$\frac{d(\hat{\beta}(\hat{j}_n),\beta)}{d(\hat{\beta}(\tilde{j}_n),\beta)} = \frac{\sum_{i=1}^{\hat{j}_n} W_i + \dfrac{\|\beta - \beta^{(n)}(K_n)\|^2_{M_n}}{\sigma^2}}{\sum_{i=1}^{\tilde{j}_n} W_i + \dfrac{\|\beta - \beta^{(n)}(K_n)\|^2_{M_n}}{\sigma^2}}.$$

Because $\tilde{j}_n \leq \hat{j}_n$, $d(\hat{\beta}(\hat{j}_n),\beta)/d(\hat{\beta}(\tilde{j}_n),\beta) \geq 1$. Now $\|\beta - \beta^{(n)}(K_n)\|^2_{M_n} \to 0$ as $n \to \infty$; remember, $\tilde{\beta}$ is fixed. Hence there exists $N_1 < \infty$ such that for all $n > N_1$, $\|\tilde{\beta} - \beta^{(n)}(K_n)\|^2_{M_n} < 0.01\sigma^2$. Thus

$$\frac{d(\hat{\beta}(\hat{j}_n),\beta)}{d(\hat{\beta}(\tilde{j}_n),\beta)} \geq \frac{\sum_{i=1}^{\hat{j}_n} W_i + 0.01}{\sum_{i=1}^{\tilde{j}_n} W_i + 0.01}$$

so that

$$P\left(\frac{d(\hat{\beta}(\hat{j}_n),\beta)}{d(\hat{\beta}(\tilde{j}_n),\beta)} > 2\right) \geq P\left(\frac{\sum_{i=1}^{\hat{j}_n} W_i + 0.01}{\sum_{i=1}^{\tilde{j}_n} W_i + 0.01} > 2\right)$$

for all $n > N_1$. As proved in Appendix E, $P(\tilde{j}_n = 1) \to 1$ as $n \to \infty$. We stress that this convergence is along a sequence of parameters that change with n. Thus

$$P\left(\frac{\sum_{i=1}^{\hat{j}_n} W_i + 0.01}{\sum_{i=1}^{\tilde{j}_n} W_i + 0.01} > 2\right) - P\left(\frac{\sum_{i=1}^{\hat{j}_n} W_i + 0.01}{W_1 + 0.01} > 2\right) \to 0 \qquad \text{as } n \to \infty.$$

As proved in Appendix E,

$$P\left(\frac{\sum_{i=1}^{\hat{j}_n} W_i + 0.01}{W_1 + 0.01} > 2\right) \geq P(W_2 > 2)P(W_1 \leq 1.99) + o(1). \tag{6}$$

Note that $P(W_2 > 2)P(W_1 \leq 1.99)$ is approximately equal to 0.1323923. ■

5. DISCUSSION

The purpose of this paper has been to make clear, in the most readable and concise form, the potentially misleading nature of the comparisons of variable

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

selection criteria put forward by Shibata (1981). For this reason we have dealt only with (a) two criteria AIC and BIC, (b) a very simple concrete class of data-generating mechanisms, and (c) a nested family of models.

Shibata (1980) has proved some properties of AIC and related criteria in a time series context. Inter alia he proves results similar to those of Shibata (1981). These and the results of Shibata (1983) concern the behavior of a particular measure of efficiency as $n \to \infty$ pointwise in the space of data-generating mechanisms. Clearly, the same care must exercised with the interpretation of these results as with the results of Shibata (1981). It should be noted that the methodology put forward by Shibata (1981) has been quite influential (see, e.g., Breiman and Freedman, 1983; Polyak and Tsybakov, 1990; and the references in these papers).

An interesting problem for future research is to find a broad class of data-generating mechanisms $\mathcal{C}$ ($\mathcal{C} \subset \mathcal{B}$) such that in large samples (a) AIC is at least as good as BIC for every element of $\mathcal{C}$ and (b) AIC is better than BIC for at least one element of $\mathcal{C}$. The proof of Theorem 4.1 becomes inapplicable if $\|\beta - \beta^{(n)}(K_n)\|_{M_n}^2 \to \infty$ as $n \to \infty$. This then raises the following question, which is difficult to answer, for future research. Could it be that $\mathcal{C}$ can be taken to consist of those elements of $\mathcal{B}$ satisfying a condition of the type $\|\beta - \beta^{(n)}(K_n)\|_{M_n}^2 \to \infty$ as $n \to \infty$?

REFERENCES

Akaike, H. (1974) A new look at statistical model identification. *IEEE Transactions on Automatic Control* 19, 716–723.

Akaike, H. (1978) A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics A* 30, 9–14.

Barron, A., L. Birge, & P. Massart (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields* 113, 301–413.

Bickel, P.J. & K.A. Doksum (1977) *Mathematical Statistics*. Oakland, California: Holden-Day.

Breiman, L. & D. Freedman (1983) How many variables should be entered in a regression equation? *Journal of the American Statistical Association* 78, 131–136.

Hajek, J. (1971) Limiting properties of likelihoods and inference. In V.P. Godambe & D.A. Sprott (eds.), *Foundations of Statistical Inference: Proceedings of the Symposium on the Foundations of Statistical Inference, Held at University of Waterloo, Ontario, Canada, from March 31 to April 9, 1970*, pp. 142–161. Toronto, Canada: Holt, Rinehart, and Winston.

Hamming, R.W. (1973) *Numerical Methods for Scientists and Engineers*, 2nd ed.. New York: McGraw-Hill.

Kabaila, P. (1995) The effect of model selection on confidence regions and prediction regions. *Econometric Theory* 11, 537–549.

Lehmann, E.L. (1983) *Theory of Point Estimation*. New York: Wiley.

Polyak, B.T. & A.B. Tsybakov (1990) Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory of Probability and Its Applications* 35, 293–306.

Pötscher, B.M. (1991) Effects of model selection on inference. *Econometric Theory* 7, 163–181.

Pötscher, B.M. (1995) Comment on "The effect of model selection on confidence regions and prediction regions" by P. Kabaila. *Econometric Theory* 11, 550–559.

Schwarz, G. (1978) Estimating the dimension of a linear model. *Annals of Statistics* 6, 461–464.

Shibata, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8, 147–164.

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

Shibata, R. (1981) An optimal selection of regression variables. *Biometrika* 68, 45–54.
Shibata, R. (1982) Correction to "An optimal selection of regression variables." *Biometrika* 69, 492.
Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics* 35, 415–423.
Shibata, R. (1986) Regression variables, selection of. In S. Kotz & N.L. Johnson (eds.), *Encyclopedia of Statistical Sciences*, vol. 7. New York: Wiley.

APPENDIX A

In this Appendix we prove the following result.

LEMMA A.1. *Suppose* $n > 1$ *and define* N *to be the greatest integer less than or equal to* $n/2$. *Then*

(a) $c_1, \ldots, c_{N+1}$ *are orthogonal* n*-vectors i.e.,* $c_i' c_j = 0$ *for* $i \neq j$ *and* $i, j \in \{1, \ldots, N+1\}$.
(b) $c_{k+n} = c_k$ *for all* $k \geq 1$.
(c) $c_\ell = c_{n+2-\ell}$ *for* $\ell = 2, \ldots, N$ *and, for* n *odd,* $c_{N+1} = c_{N+2}$.
(d)

$$c_k' c_k = \begin{cases} n & \text{for } k = 1, \\ \dfrac{n}{2} & \text{for } k = 2, \ldots, N \end{cases}$$

and

$$c_{N+1}' c_{N+1} = \begin{cases} \dfrac{n}{2} & \text{for } n \text{ odd}, \\ n & \text{for } n \text{ even}. \end{cases}$$

Proof.

(a) Note that

$$\sum_{k=0}^{n-1} \exp\left[i \left(\frac{2\pi \ell}{n} \right) k \right] = \begin{cases} n & \text{for } \ell \text{ an integer multiple of } n, \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$\sum_{k=0}^{n-1} \cos\left(\frac{2\pi \ell}{n} k \right) = \begin{cases} n & \text{for } \ell \text{ an integer multiple of } n, \\ 0 & \text{otherwise.} \end{cases}$$

We now prove that $c_{r+1}' c_{s+1} = 0$ for $r \neq s$ and $r, s \in \{0, 1, \ldots, N\}$. This follows from the observation that

$$c_{r+1}' c_{s+1} = \sum_{k=0}^{n-1} \cos\left(\frac{2\pi r}{n} k \right) \cos\left(\frac{2\pi s}{n} k \right) = \frac{1}{2} [A(r,s) + B(r,s)],$$

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

where

$$A(r,s) = \sum_{k=0}^{n-1} \cos\left(\frac{2\pi(r-s)}{n}k\right) \qquad \text{and} \qquad B(r,s) = \sum_{k=0}^{n-1} \cos\left(\frac{2\pi(r+s)}{n}k\right).$$

We remark that for n even, the result follows from the orthogonality relations described, e.g., by Hamming (1973, p. 523).

(c) Observe that

$$\cos((n+2-\ell-1)t_j) = \cos\left(2\pi j - \frac{(\ell-1)2\pi j}{n}\right) = \cos((\ell-1)t_j).$$

(d) Obviously $c_1'c_1 = n$. Now

$$c_k'c_k = \frac{1}{2}[A(k-1,k-1) + B(k-1,k-1)] = \frac{1}{2}[n + B(k-1,k-1)].$$

The result now follows from the observation that

$$B(k-1,k-1) = \begin{cases} n & \text{for } 2k-2 \text{ an integer multiple of } n, \\ 0 & \text{otherwise.} \end{cases}$$

■

APPENDIX B

In this Appendix we derive a convenient expression (used in Section 4) for $\|\beta - \beta^{(n)}(j)\|^2_{M_n}$. Observe that for $\alpha \in \ell_2$, $\|\alpha\|^2_{M_n} = \langle M_n\alpha, \alpha\rangle = (X_n\alpha)'(X_n\alpha)$. Using the properties of the c_k's stated in Section 2, we find that for n even

$$X_n\alpha = \left(\sum_{i=0}^{\infty}\frac{\alpha_{in+1}}{in+1}\right)c_1 + \sum_{k=2}^{N}\left[\frac{\alpha_k}{k} + \sum_{i=1}^{\infty}\left(\frac{\alpha_{in+k}}{in+k} + \frac{\alpha_{in+2-k}}{in+2-k}\right)\right]c_k + \left(\sum_{i=0}^{\infty}\frac{\alpha_{in+N+1}}{in+N+1}\right)c_{N+1}$$

so that

$$(X_n\alpha)'(X_n\alpha) = n\left(\sum_{i=0}^{\infty}\frac{\alpha_{in+1}}{in+1}\right)^2 + \frac{n}{2}\sum_{k=2}^{N}\left[\frac{\alpha_k}{k} + \sum_{i=1}^{\infty}\left(\frac{\alpha_{in+k}}{in+k} + \frac{\alpha_{in+2-k}}{in+2-k}\right)\right]^2 + n\left(\sum_{i=0}^{\infty}\frac{\alpha_{in+N+1}}{in+N+1}\right)^2.$$

Similarly, for n odd,

$$(X_n\alpha)'(X_n\alpha) = n\left(\sum_{i=0}^{\infty}\frac{\alpha_{in+1}}{in+1}\right)^2 + \frac{n}{2}\sum_{k=2}^{N+1}\left[\frac{\alpha_k}{k} + \sum_{i=1}^{\infty}\left(\frac{\alpha_{in+k}}{in+k} + \frac{\alpha_{in+2-k}}{in+2-k}\right)\right]^2.$$

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

Now $\beta^{(n)}(j)$ is the value of $\beta^* = [\beta_1^*,\ldots\beta_j^*,0,0,\ldots]'$ minimizing $\|\beta - \beta^*\|_{M_n}^2$ where $j < N$. Restrict attention to $j < N$. Define

$$a(\beta,\beta^*) = n\left(\beta_1 - \beta_1^* + \sum_{i=1}^{\infty}\frac{\beta_{in+1}}{in+1}\right)^2 + \frac{n}{2}\sum_{k=2}^{j}\left[\frac{\beta_k - \beta_k^*}{k} + \sum_{i=1}^{\infty}\left(\frac{\beta_{in+k}}{in+k} + \frac{\beta_{in+2-k}}{in+2-k}\right)\right]^2.$$

For n even,

$$\|\beta - \beta^*\|_{M_n}^2 = a(\beta,\beta^*) + \frac{n}{2}\sum_{k=j+1}^{N}\left[\frac{\beta_k}{k} + \sum_{i=1}^{\infty}\left(\frac{\beta_{in+k}}{in+k} + \frac{\beta_{in+2-k}}{in+2-k}\right)\right]^2 + n\left(\sum_{i=0}^{\infty}\frac{\beta_{in+N+1}}{in+N+1}\right)^2,$$

and, for n odd,

$$\|\beta - \beta^*\|_{M_n}^2 = a(\beta,\beta^*) + \frac{n}{2}\sum_{k=j+1}^{N+1}\left[\frac{\beta_k}{k} + \sum_{i=1}^{\infty}\left(\frac{\beta_{in+k}}{in+k} + \frac{\beta_{in+2-k}}{in+2-k}\right)\right]^2.$$

Clearly, $\|\beta - \beta^*\|_{M_n}^2$ is minimized with respect to β^* by setting $a(\beta,\beta^*) = 0$. Hence, for n even,

$$\|\beta - \beta^{(n)}(j)\|_{M_n}^2 = \frac{n}{2}\sum_{k=j+1}^{N}\left[\frac{\beta_k}{k} + \sum_{i=1}^{\infty}\left(\frac{\beta_{in+k}}{in+k} + \frac{\beta_{in+2-k}}{in+2-k}\right)\right]^2 + n\left(\sum_{i=0}^{\infty}\frac{\beta_{in+N+1}}{in+N+1}\right)^2,$$

and, for n odd,

$$\|\beta - \beta^{(n)}(j)\|_{M_n}^2 = \frac{n}{2}\sum_{k=j+1}^{N+1}\left[\frac{\beta_k}{k} + \sum_{i=1}^{\infty}\left(\frac{\beta_{in+k}}{in+k} + \frac{\beta_{in+2-k}}{in+2-k}\right)\right]^2.$$

That Condition 2 of Shibata (1981) is satisfied may be checked directly as follows. Suppose Assumption B is satisfied. Fix $j > 0$ and let ℓ denote the smallest $i \in \{j+1, j+2,\ldots\}$ such that $\beta_i \neq 0$. Consider $N > j$. Then

$$\|\beta - \beta^{(n)}(j)\|_{M_n}^2 \geq \frac{n}{2}\left[\frac{\beta_\ell}{\ell} + \sum_{i=1}^{\infty}\left(\frac{\beta_{in+\ell}}{in+\ell} + \frac{\beta_{in+2-\ell}}{in+2-\ell}\right)\right]^2.$$

The right-hand side diverges to infinity as $n \to \infty$.

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

APPENDIX C

Define $Z = D_n' Y$ where D_n is an $n \times n$ orthonormal matrix whose jth column is proportional to c_j $(1 \leq j \leq N+1)$. Define $\mu(\beta) = D_n' X_n \beta$ and observe that $Z = \mu(\beta) + \eta$ where η is a random n-vector with independent and identically $N(0, \sigma^2)$ distributed components. Using the results of Appendix B, it may be shown that $\mu_k(\beta)$, the kth component of $\mu(\beta)$, is given by (5). Using the standard methodology described by, e.g., Bickel and Doksum (1977, p. 255) and the results of Appendix B, it may be shown that

$$d(\hat{\beta}(j), \beta) = \sum_{i=1}^{j} \eta_i^2 + \sum_{i=j+1}^{K_n} \mu_i^2(\beta) + \|\beta - \beta^{(n)}(K_n)\|_{M_n}^2$$

for $j = 1, \ldots, K_n$. The result (4) follows immediately from this.

APPENDIX D

In this Appendix we prove the following result.

LEMMA D.1. *Let $\bar{j}_n$ denote the smallest value of $j \in \{1, \ldots, K_n\}$ minimizing $n \log \hat{\sigma}^2(j) + cj$ where $c > 2$. Then $\bar{j}_n \leq \hat{j}_n$.*

Proof. Restrict attention to $j \in \{1, \ldots, K_n\}$. By the definition of $\hat{j}_n$,

$$AIC(\hat{j}_n) \leq AIC(j) \qquad \text{for all } j > \hat{j}_n.$$

In other words,

$$n \log \hat{\sigma}^2(\hat{j}_n) - n \log \hat{\sigma}^2(j) + 2(\hat{j}_n - j) \leq 0 \qquad \text{for all } j > \hat{j}_n.$$

Thus

$$n \log \hat{\sigma}^2(\hat{j}_n) - n \log \hat{\sigma}^2(j) + c(\hat{j}_n - j) \leq 0 \qquad \text{for all } j > \hat{j}_n.$$

In other words,

$$[n \log \hat{\sigma}^2(\hat{j}_n) + c\hat{j}_n] - [n \log \hat{\sigma}^2(j) + cj] \leq 0 \qquad \text{for all } j > \hat{j}_n.$$

Hence $\bar{j}_n \leq \hat{j}_n$. ■

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

APPENDIX E

In this Appendix we derive some properties of $\tilde{j}_n$ and $\hat{j}_n$ for large n. We assume that $\{K_n\}$ satisfies both Assumption A and $n^{1/2}\log(n)K_n = o(n)$. We also assume that $\|\beta - \beta^{(n)}(K_n)\|^2_{M_n} \to 0$ as $n \to \infty$. The random variable $\tilde{j}_n$ is the smallest value of $j \in \{1,\ldots,K_n\}$ minimizing

$$\frac{n}{\sigma^2}\exp\left(\frac{BIC(j)}{n}\right) = \exp\left(\frac{\log(n)j}{n}\right)\sum_{i=j+1}^{n}\left(\frac{\mu_i(\beta)+\eta_i}{\sigma}\right)^2, \tag{E.1}$$

where $\mu_i(\beta)$ and η_i are as defined in Appendix C. Now suppose that $\mu_2(\beta) = \cdots = \mu_{K_n}(\beta) = 0$. Thus (*E*.1) is equal to

$$\exp\left(\frac{\log(n)j}{n}\right)\left[\sum_{i=j+1}^{K_n}\left(\frac{\eta_i}{\sigma}\right)^2 + \sum_{i=K_n+1}^{n}\left(\frac{\mu_i(\beta)+\eta_i}{\sigma}\right)^2\right].$$

Because $\sum_{i=K_n+1}^{n}((\mu_i(\beta)+\eta_i)/\sigma)^2$ does not depend on j, $\tilde{j}_n$ is the smallest value of $j \in \{1,\ldots,K_n\}$ minimizing

$$\sum_{i=j+1}^{K_n}\left(\frac{\eta_i}{\sigma}\right)^2 + \left[\exp\left(\frac{\log(n)j}{n}\right)-1\right]\sum_{i=j+1}^{K_n}\left(\frac{\eta_i}{\sigma}\right)^2 + \left[\exp\left(\frac{\log(n)j}{n}\right)-1\right]\sum_{i=K_n+1}^{n}\left(\frac{\mu_i(\beta)+\eta_i}{\sigma}\right)^2.$$

The second term is a nonnegative random variable with expected value less than or equal to

$$\left[\exp\left(\frac{\log(n)K_n}{n}\right)-1\right]K_n$$

which converges to zero uniformly in $j \in \{1,\ldots,K_n\}$ as $n \to \infty$. Hence this second term is the $o_p(1)$ term uniformly in $j \in \{1,\ldots,K_n\}$. It is straightforward to show that the third term is equal to $\log(n)j + o_p(1)$, where the $o_p(1)$ term is uniform in $j \in \{1,\ldots,K_n\}$, using the fact that

$$\sum_{i=K_n+1}^{n}\left(\frac{\mu_i(\beta)}{\sigma}\right)^2 = \frac{\|\beta-\beta^{(n)}(K_n)\|^2_{M_n}}{\sigma^2}.$$

Thus $\tilde{j}_n$ is the smallest value of $j \in \{1,\ldots,K_n\}$ minimizing

$$\sum_{i=j+1}^{K_n}\left(\frac{\eta_i}{\sigma}\right)^2 + \log(n)j + o_p(1),$$

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press

where the $o_p(1)$ term is uniform in $j \in \{1,\ldots,K_n\}$. Now $\sum_{i=1}^{K_n}(\eta_i/\sigma)^2$ does not depend on j and so $\tilde{j}_n$ is the smallest value of $j \in \{1,\ldots,K_n\}$ minimizing

$$\sum_{i=1}^{j}\left[\log(n) - \left(\frac{\eta_i}{\sigma}\right)^2\right] + o_p(1),$$

where the $o_p(1)$ term is uniform in $j \in \{1,\ldots,K_n\}$. A standard argument shows that $P(\tilde{j}_n = 1) \to 1$ as $n \to \infty$. Similarly, it may be proved that $\hat{j}_n$ is the smallest value of $j \in \{1,\ldots,K_n\}$ minimizing

$$\sum_{i=1}^{j}\left[2 - \left(\frac{\eta_i}{\sigma}\right)^2\right] + o_p(1),$$

where the $o_p(1)$ term is uniform in $j \in \{1,\ldots,K_n\}$. Now define $W_i = (\eta_i/\sigma)^2$. Observe that

$$\frac{\sum_{i=1}^{\hat{j}_n} W_i + 0.01}{W_1 + 0.01} \geq \frac{\sum_{i=1}^{\min(\hat{j}_n,2)} W_i + 0.01}{W_1 + 0.01}.$$

Suppose $\delta > 0$. Now

$$\left\{\frac{\sum_{i=1}^{\min(\hat{j}_n,2)} W_i + 0.01}{W_1 + 0.01} > 1 + \delta\right\} = \{\hat{j}_n \geq 2\} \cap \{W_2 > \delta(W_1 + 0.01)\}$$

$$\supset \{AIC(2) < AIC(1)\} \cap \{W_2 > \delta(W_1 + 0.01)\}.$$

Thus

$$P\left(\frac{\sum_{i=1}^{\hat{j}_n} W_i + 0.01}{\sum_{i=1}^{\tilde{j}_n} W_i + 0.01} > 1 + \delta\right) \geq P(W_2 > 2, W_2 > \delta(W_1 + 0.01)) + o(1). \quad \textbf{(E.2)}$$

Also,

$$P(W_2 > 2, W_2 > \delta(W_1 + 0.01)) \geq P\left(W_2 > 2, W_1 \leq \frac{2}{\delta} - 0.01\right). \quad \textbf{(E.3)}$$

The result (6) now follows from (*E*.2) and (*E*.3) when $\delta = 1$.

https://doi.org/10.1017/S0266466602184052 Published online by Cambridge University Press