Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 50 (2006) 1524–1550
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A corrected Akaike criterion based on Kullback's symmetric divergence: applications in time series, multiple and multivariate regression

B. Hafidi, A. Mkhadri*, [1]

Department of Mathematics, Faculty of Sciences-Semlalia, BP. 2390, Marrakech 40000, Morocco

Received 29 December 2003; received in revised form 14 January 2005; accepted 14 January 2005
Available online 2 March 2005

Abstract

The evaluation of an Akaike information criterion (AIC), KIC is considered. Kullback information criterion (KIC) is an approximately unbiased estimator for a risk function based on the Kullback's symmetric divergence. However, when the sample size is small, or when it is large and the dimension of the candidate model is relatively small, this criterion displays a large negative bias. To overcome this problem, corrected versions, KICc, of this criterion for univariate autoregressive models and for multiple and multivariate regression models are proposed. Thus, the methodology for AIC and AICc from McQuarrie and Tsai is extended to the KIC criterion. The performance of the new criterion relative to other criteria is examined in a large simulation study.
© 2005 Elsevier B.V. All rights reserved.

Keywords: AIC; KIC; AICc; Model selection; Regression; Time series

1. Introduction

The Akaike information criterion (AIC) (Akaike, 1973), was designed as an asymptotically unbiased estimator of a variant of the Kullback's (1968) directed divergence between the true model and a fitted approximating model. Another variant of this criterion,

* Corresponding author. Tel.: +212 044 434649; fax: +212 44 437409.
E-mail addresses: b.hafidi@ucam.ac.ma (B. Hafidi), mkhadri@ucam.ac.ma (A. Mkhadri).

[1] The second author is supported by TWAS Research Grant 01-159 RG/MATHS/AF/AC.

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.01.007

Kullback information criterion (KIC), was proposed by Cavanaugh (1999) as an asymptotically unbiased estimator of Kullback's symmetric divergence (the sum of the two directed divergences). However, when the sample size is small, or when the sample size is large and the dimension of the candidate model is relatively small, both estimators suffer from a large negative bias, which usually leads them to select over-parameterized models. In this setting, Hurvich and Tsai (1989) proposed a corrected Akaike information criterion (AICc), for linear and non-linear regression and for autoregressive modeling. The AICc has been extended in a number of directions, including autoregressive moving average modeling (Hurvich et al., 1990), vector autoregressive modeling (Hurvich and Tsai, 1993), and multivariate regression modeling (Bedrick and Tsai, 1994; Fujikoshi and Satoh, 1997). Further, the KIC criterion tends to underestimate the Kullback's symmetric divergence in small-sample applications, as indicated by Cavanaugh (1999).

In the present paper, we derive and investigate bias-corrected versions of KIC, denoted KICc, for general multiple regression and autoregressive time series models. Moreover, we extend KICc from multiple linear regression to multivariate regression models. In every case, the KICc was computed by assuming that the candidate family of models includes the true model. Furthermore, unlike the AICc criterion which is exactly unbiased for linear regression, our criterion is unbiased only asymptotically in every setting. However, the KICc may be written as the sum of KIC and a non-stochastic term, which means that the reduction in bias is achieved without any increase in the variance.

The signal-to-noise ratio is one of the important tools for evaluating model selection criteria. If the signal-to-noise ratio for one criterion is small (less than 0.5, say), the criterion tends to choose overfit models. On the other hand, if the signal-to-noise ratio is large (greater than 2, say), the criterion has a tendency to underfit. We derive the signal-to-noise ratio for KIC and KICc and we show that KIC has a small signal-to-noise ratio and therefore tends to overfit.

In Section 2 we develop the KICc in multiple regression, and we present simulation results for linear regression. We derive KICc and present simulation results for multivariate regression in Section 3. In Section 4, we present KICc for autoregressive models. In each section, we present the signal-to-noise ratios and the small sample probabilities of overfitting for these two criteria. We end the paper with a small conclusion in Section 5.

2. Derivation of the KICc criterion for multiple regression

Let $\mathbf{Y}$ be the vector of the observed data, θ_0 be the true parameter vector, which is unknown, and let θ be the parameter vector of the candidate model. $f(\mathbf{Y}|\theta_0)$ and $f(\mathbf{Y}|\theta)$ represent the generating and the candidate densities for data, respectively. We denote by $\hat{\theta}$ the maximum likelihood estimator of θ_0, and by d the dimension of θ. In this section, we derive a corrected version of KIC criterion for general multiple regression models. Moreover, the signal-to-noise ratio of overfitting and the overfitting probability for both the KIC and KICc criteria are outlined.

2.1. KICc for multiple regression

Suppose that the generating model for data is given by

$$\mathbf{Y} = \mu + \varepsilon_0, \tag{1}$$

where $\mathbf{Y}$ is an $n \times 1$ observation vector, μ is an $n \times 1$ vector, and ε is a Gaussian random vector with zero mean and variance–covariance matrix equal to $\sigma_0^2\mathbf{I}$. We assume that the candidate family of models is given by

$$\mathbf{Y} = g(\beta) + \varepsilon, \tag{2}$$

where β is a $p \times 1$ vector, ε is a Gaussian random vector with zero mean and variance–covariance matrix equal to $\sigma^2\mathbf{I}$, and g is a function assumed to be twice continuously differentiable in β. We assume that the candidate class of models includes the true model. This strong assumption is also used in the derivation of the KIC (Cavanaugh, 1999) and AIC criteria (Linhart and Zucchini, 1986). In this setting, the vector μ can be written as $\mu = g\left(\beta_0\right)$, where β_0 is an $p_0 \times 1$ unknown vector. If both the candidate and the generating models are linear, we have $g(\beta) = \mathbf{X}\beta$, $g\left(\beta_0\right) = \mathbf{X}_0\beta_0$, where $\mathbf{X}$ and $\mathbf{X}_0$ are $n \times p$ and $n \times p_0$ matrices of full rank, respectively. With this notation, the parameter vectors θ and θ_0 for the candidate and the generating models are $\left(\beta^{\mathrm{t}}, \sigma\right)^{\mathrm{t}}$ and $\left(\beta_0^{\mathrm{t}}, \sigma_0\right)^{\mathrm{t}}$, respectively, where "t" stands for the transpose and $d = p + 1$.

A measure of separation between the generating model $f\left(\mathbf{Y}|\theta_0\right)$ and the candidate model $f(\mathbf{Y}|\theta)$ is given by the Kullback's symmetric divergence (Kullback, 1968), defined by

$$J\left(\theta_0, \theta\right) = \left\{D\left(\theta_0, \theta\right) - D\left(\theta_0, \theta_0\right)\right\} + \left\{D\left(\theta, \theta_0\right) - D(\theta, \theta)\right\}, \tag{3}$$

where $D\left(\theta_0, \theta\right) = \mathbf{E}_{\theta_0}\{-2 \ln f(\mathbf{Y}|\theta)\}$ and $\mathbf{E}_{\theta_0}$ denotes the expectation with respect to $f\left(\mathbf{Y}|\theta_0\right)$.

The log-likelihood for the candidate model is given by

$$\ln f(\mathbf{Y}|\theta) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{Y} - g(\beta))^{\mathrm{t}}(\mathbf{Y} - g(\beta)).$$

Since the second term of (3) does not depend on θ and for the purpose of discriminating among various models, Cavanaugh (1999) proposed

$$K\left(\theta_0, \theta\right) = D\left(\theta_0, \theta\right) + \left\{D\left(\theta, \theta_0\right) - D(\theta, \theta)\right\}, \tag{4}$$

as a substitute for $J\left(\theta_0, \theta\right)$. However, he showed that, under the assumption that the candidate family includes the generating model, the criterion defined by

$$KIC = n\left(\ln \hat{\sigma}^2 + 1\right) + 3(p + 1) \tag{5}$$

is an asymptotically unbiased estimator of

$$\Omega\left(d, \theta_0\right) = \mathbf{E}_{\theta_0}\left\{K\left(\theta_0, \hat{\theta}\right)\right\}. \tag{6}$$

Now, we derive a corrected version of this criterion using the result of Hurvich and Tsai (1989) which established that the criterion defined by

$$AICc = n\left(\ln\hat{\sigma}^2 + 1\right) + \frac{2n(p+1)}{n-p-2} \tag{7}$$

is an asymptotically unbiased estimator of

$$\Delta\left(d, \theta_0\right) = \mathbf{E}_{\theta_0}\left\{D\left(\theta_0, \hat{\theta}\right)\right\}. \tag{8}$$

The following proposition is an adaptation of this result for the Kullback's symmetric divergence.

Proposition 2.1. *The criterion defined by*

$$KICc = n\left(\ln\hat{\sigma}^2 + 1\right) + \frac{(p+1)(3n-p-2)}{n-p-2}$$

is an asymptotically unbiased estimator of $\Omega(d, \theta_0)$.

Proof. From (4) and (6), we have

$$\Omega\left(d, \theta_0\right) = \mathbf{E}_{\theta_0}\left\{D\left(\theta_0, \hat{\theta}\right)\right\} + \mathbf{E}_{\theta_0}\left\{D\left(\hat{\theta}, \theta_0\right) - D\left(\hat{\theta}, \hat{\theta}\right)\right\}. \tag{9}$$

Now, we will compute the second term of this equation. First, we calculate $D\left(\theta_1, \theta_2\right)$ in the general case, where θ_1 and θ_2 are any two vectors, and then we apply the obtained result to compute $D\left(\hat{\theta}, \theta_0\right)$ and $D\left(\hat{\theta}, \hat{\theta}\right)$. We have

$$\begin{aligned}
D\left(\theta_1, \theta_2\right) &= \mathbf{E}_{\theta_1}\left\{-2\ln f\left(\mathbf{Y}|\theta_2\right)\right\} \\
&= \mathbf{E}_{\theta_1}\left\{n\ln\left(2\pi\sigma_2^2\right) + \frac{1}{\sigma_2^2}\left(\mathbf{Y} - g\left(\beta_2\right)\right)^{\mathrm{t}}\left(\mathbf{Y} - g\left(\beta_2\right)\right)\right\} \\
&= \mathbf{E}_{\theta_1}\left\{n\ln\left(2\pi\sigma_2^2\right) + \frac{1}{\sigma_2^2}\left(\mathbf{Y} - g\left(\beta_1\right)\right)^{\mathrm{t}}\left(\mathbf{Y} - g\left(\beta_1\right)\right)\right. \\
&\quad \left. + \frac{1}{\sigma_2^2}\left(g\left(\beta_1\right) - g\left(\beta_2\right)\right)^{\mathrm{t}}\left(g\left(\beta_1\right) - g\left(\beta_2\right)\right)\right\} \\
&= n\ln\left(2\pi\sigma_2^2\right) + \frac{1}{\sigma_2^2}\left\{n\sigma_1^2 + \left(g\left(\beta_1\right) - g\left(\beta_2\right)\right)^{\mathrm{t}}\left(g\left(\beta_1\right) - g\left(\beta_2\right)\right)\right\}.
\end{aligned} \tag{10}$$

From (10), we have

$$D\left(\hat{\theta}, \theta_0\right) = n\ln\left(2\pi\sigma_0^2\right) + \frac{1}{\sigma_0^2}\left\{n\hat{\sigma}^2 + \left(g\left(\hat{\beta}\right) - g\left(\beta_0\right)\right)^{\mathrm{t}}\left(g\left(\hat{\beta}\right) - g\left(\beta_0\right)\right)\right\}.$$

The linear expansion of order one of $g(\hat{\beta})$ at $\beta = \beta_0$ is given by

$$g\left(\hat{\beta}\right) \approx g\left(\beta_0\right) + \mathbf{V}\left(\hat{\beta} - \beta_0\right),$$

where $\mathbf{V} = \partial g/\partial\beta$ evaluated at $\beta = \beta_0$. This leads to

$$D\left(\hat{\theta}, \theta_0\right) \approx n \ln\left(2\pi\sigma_0^2\right) + \frac{1}{\sigma_0^2}\left\{n\hat{\sigma}^2 + \left(\hat{\beta} - \beta_0\right)\mathbf{V}^{\mathrm{t}}\mathbf{V}\left(\hat{\beta} - \beta_0\right)\right\}. \tag{11}$$

Similarly, from (10) we have

$$D\left(\hat{\theta}, \hat{\theta}\right) = n \ln 2\pi + n \ln \hat{\sigma}^2 + n. \tag{12}$$

From (11) and (12), we obtain

$$D\left(\hat{\theta}, \theta_0\right) - D\left(\hat{\theta}, \hat{\theta}\right) \approx n \ln \frac{\sigma_0^2}{\hat{\sigma}^2} + n\frac{\hat{\sigma}^2}{\sigma_0^2} + \frac{1}{\sigma_0^2}\left\{\left(\hat{\beta} - \beta_0\right)\mathbf{V}^{\mathrm{t}}\mathbf{V}\left(\hat{\beta} - \beta_0\right)\right\} - n.$$

Now, the distribution of $\left(n\hat{\sigma}^2/\sigma_0^2\right)$ is a chi-square with $(n - p)$ degrees of freedom, and the quadratic form $\left\{\left(\hat{\beta} - \beta_0\right)\left(1/\sigma_0^2\right)\left(\mathbf{V}^{\mathrm{t}}\mathbf{V}\right)\left(\hat{\beta} - \beta_0\right)\right\}$ is approximately distributed as chi-square with p degrees of freedom. Thus we have

$$\mathbf{E}_{\theta_0}\left\{D\left(\hat{\theta}, \theta_0\right) - D\left(\hat{\theta}, \hat{\theta}\right)\right\} \approx \mathbf{E}_{\theta_0}\left\{n \ln \frac{\sigma_0^2}{\hat{\sigma}^2}\right\}.$$

Furthermore, Cavanaugh (1997, Lemma 3) showed that

$$\mathbf{E}_{\theta_0}\left\{n \ln \frac{\sigma_0^2}{\hat{\sigma}^2}\right\} = (p + 1) + o(1). \tag{13}$$

According to (7), and substituting (13) in the second term of (9), we obtain the estimator of $\Omega(d, \theta_0)$ given in the proposition. □

Other equivalent forms of this criterion are

$$KICc = KIC + \frac{2(p+1)(p+2)}{n - p - 2},$$
$$KICc = AICc + (p + 1).$$

Since KICc is the sum of KIC and an additional non-stochastic penalty term, the reduction in bias is achieved without any increase in variance.

2.2. The signal-to-noise ratio and probability of overfitting

KIC can overfit in small samples for reasons similar to those of AIC. To examine KIC's properties in overfitting, we compute its signal-to-noise ratio and its small-sample

probability of overfitting. From the results of our experiments recorded in Table 2 in Section 2.3, we can see that KIC has a weak signal-to-noise ratio with respect to overfitting. Moreover, we assert that the signal-to-noise ratio has a close relationship to the probability of overfitting. We detail this relationship in Section 2.3.

We will compare the true model (1) with an overfit candidate model (2) of order p, $p > p_0$. Let $L = p - p_0$ represent the amount of overfitting. KIC will choose a model p_0 over model $p_0 + L$ if KIC_{p_0} is strictly less than KIC_{p_0+L}. Let $\Delta KIC = KIC_{p_0+L} - KIC_{p_0}$. Then, applying the relations (2A.13) and (2A.14) from McQuarrie and Tsai (1998), and scaling the first two terms of KIC by n, the signal-to-noise ratio of KIC is obtained as

$$\frac{\mathbf{E}[\Delta KIC]}{Sd[\Delta KIC]} = \frac{\sqrt{(n-p_0-L)(n-p_0+2)}}{\sqrt{2L}} \times \left\{ \ln\left(\frac{n-p_0-L}{n-p_0}\right) - \frac{L}{(n-p_0-L)(n-p_0)} + \frac{3L}{n} \right\}, \tag{14}$$

where $Sd[A]$ denotes the standard deviation of A. So, as $L \longrightarrow (n - p_0)$, the first two terms of (14), which result from the $\ln \hat{\sigma}$ term of KIC, converge to $-\infty$. While the last term, arising from the penalty term of KIC, converges to zero. Thus, the third term of (14) increases for small L and then decreases to 0 as L increases. Therefore, the signal-to-noise ratio of KIC increases for small L and decreases to $-\infty$ as L increases. Hence, KIC should exhibit small sample overfitting problems.

Similarly, the signal-to-noise ratio of KICc is

$$\frac{\mathbf{E}[\Delta KICc]}{Sd[\Delta KICc]} = \frac{\sqrt{(n-p_0-L)(n-p_0+2)}}{\sqrt{2L}} \left\{ \ln\left(\frac{n-p_0-L}{n-p_0}\right) - \frac{L}{(n-p_0-L)(n-p_0)} + \frac{2L(n-1)}{(n-p_0-L-2)(n-p_0-2)} + \frac{L}{n} \right\}.$$

The following lemma shows that the signal-to-noise ratio of KICc increases as L increases, so KICc should avoid overfitting. The proof is given in the Appendix A.

Lemma 2.2. *Let the true and candidate models be* (1) *and* (2) *respectively, where* $0 < p_0 < n - 6$. *For* $n \geqslant 6$, *the signal-to-noise ratio of KICc increases as the overfitting* L *increases, where* $0 < L < 2(n - p_0)/3$.

It is well known that the signal-to-noise ratio is related to the probability of over-fitting. Therefore, as in McQuarrie and Tsai (1998), we derive the probability of overfitting for KIC and KICc scaling the first two terms of them by n. (For other criteria see McQuarrie and Tsai, 1998.)

Let $SSE_p = n\hat{\sigma}_p$. We know that KIC overfits if $KIC_{p_0+L} < KIC_{p_0}$. Then the probability that KIC selects the overfitted model $p_0 + L$ is

$$\begin{aligned} P\left(KIC_{p_0+L} < KIC_{p_0}\right) &= P\left\{\ln \hat{\sigma}^2_{p_0+L} + \frac{3\,(p_0+L+1)}{n} < \ln\, \hat{\sigma}^2_{p_0} + \frac{3\,(p_0+1)}{n}\right\} \\ &= P\left\{\ln \frac{\hat{\sigma}^2_{p_0+L}}{\hat{\sigma}^2_{p_0}} < \frac{-3L}{n}\right\} \\ &= P\left\{\frac{\hat{\sigma}^2_{p_0} - \hat{\sigma}^2_{p_0+L}}{\hat{\sigma}^2_{p_0+L}} > \exp\left(\frac{3L}{n}\right) - 1\right\} \\ &= P\left\{\frac{SSE_{p_0} - SSE_{p_0+L}}{SSE_{p_0+L}} > \exp\left(\frac{3L}{n}\right) - 1\right\}. \end{aligned}$$

Since $SSE_{p_0} \sim \sigma_0^2\chi^2_{n-p_0}$ and it is independent of $SSE_{p_0} - SSE_{p_0+L}$ then, $SSE_{p_0} - SSE_{p_0+L} \sim \sigma_0^2\chi^2_L$ (cf. McQuarrie and Tsai, 1998, p. 66). Consequently, we have

$$\begin{aligned} P\left(KIC_{p_0+L} < KIC_{p_0}\right) &= P\left\{\frac{\chi^2_L}{\chi^2_{n-p_0-L}} > \exp\left(\frac{3L}{n}\right) - 1\right\} \\ &= P\left\{F_{L,n-p_0-L} > \frac{n-p_0-L}{L}\left(\exp\left(\frac{3L}{n}\right) - 1\right)\right\}. \end{aligned}$$

Similarly, we have for KICc

$$\begin{aligned} &P\left(KICc_{p_0+L} < KICc_{p_0}\right) \\ &\quad = P\left\{F_{L,n-p_0-L} > \frac{n-p_0-L}{L}\right. \\ &\qquad \left.\times\left(\exp\left(\frac{2L(n-1)}{(n-p_0-L-2)\,(n-p_0-2)} + \frac{L}{n}\right) - 1\right)\right\}. \end{aligned}$$

The quantiles of the two criteria are not the same, and are not easy to compare analytically. A small numerical comparison of them is performed in the next section.

2.3. Multiple regression example

To end this section, we evaluate the performance of the KICc criterion against KIC, AIC, AICc, SIC [SIC $= n\ \ln \hat{\sigma}^2 + (p+1)\ln(n)$, Schwarz, 1978] and HQ [HQ $= n\ \ln \hat{\sigma}^2 + 2(p+1)\ln(\ln(n))$, Hannan and Quinn, 1979] criteria in a small sample numerical experiment. For simplicity, we assume that the candidate and the true models are both linear. In this setting, we take $\mathbf{V} = \mathbf{X}$. We consider 1000 samples of sizes $n = 20$ and $n = 30$ generated from model (1), with $\mu = \mathbf{X}_0\beta_0$, $p_0 = 3$, $\beta_0 = (1, 2, 3)^{\mathrm{t}}$ and $\sigma_0^2 = 1$. We consider seven candidate models stored in an $n \times 7$ matrix $\mathbf{X}$, with a column of ones, followed by six columns of independent identically distributed normal random variables with zero mean and variance–covariance matrix equal to $\sigma_0^2\mathbf{I}$. The candidate models include the columns of $\mathbf{X}$ in a sequentially nested fashion; that is, columns 1 to p define the design matrix for the

candidate model with dimension p. The design matrix $\mathbf{X}_0$ for the true model was the first three columns of $\mathbf{X}$. Over the 1000 samples, the selections are summarized in Table 1.

We see from Table 1 that for $n = 20$ both the KICc and AICc criteria perform better than the other criteria, which exhibit a tendency to overfit the true model. Moreover, KICc outperforms AICc. When $n = 30$ the tendency to overfit for all other criteria decreases slightly, but KICc provides the best selection of p.

Fig. 1 provides some insight as to why KICc tend to outperform other criteria as a selection criterion. In this figure, simulated values of $\Omega(d, \theta_0)$ are plotted versus the order p for $p = 1, 2, \ldots, 7$. These values are obtained by averaging $K\left(\theta_0, \hat{\theta}\right)$ over the 1000 replications. The average values for other criteria are also plotted versus p. When $p > p_0$, the average curves of all criteria follows the simulated $\Omega(d, \theta_0)$ curve over the range of models. But, the average KICc curve tracks $\Omega(d, \theta_0)$ more effectively than other average criteria curves. On the other hand, we note that the corresponding shapes of AIC, SIC and HQ mirror the shapes of other criteria.

We present in Table 2 both the signal-to-noise ratios and the probabilities of overfitting calculated for all the criteria for sample sizes $n = 15, 25, 50$ and $p_0 = 6$. We denote by L the number of variables by which the model overfits the true model.

For $n = 15$, KICc and AICc have a greater signal-to-noise ratios than the others, and their probabilities of overfitting are almost zero. The other criteria have smaller signal-to-noise ratios and higher probabilities of overfitting. We conclude that the smallest signal-to-noise ratios correspond to the higher probabilities of overfitting, and that greater signal-to-noise ratios lead to small probabilities of overfitting. On the other hand, the probabilities of overfitting for KICc and AICc increase to a small degree as the sample size increases, whereas they decrease for all other criteria.

3. Derivation of KICc for multivariate regression

The multivariate regression model with q responses and m predictor variables, as opposed to the multiple regression with one response and q predictor variables, is one of the most useful tools for model selection. In this section, we extend the KICc for multiple linear regression to multivariate regression, and we present its signal-to-noise ratio in this framework.

3.1. KICc for multivariate regression

Suppose that the generating and the candidate models are both linear, where the candidate model is given by

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}, \quad \mathbf{U} \sim N_q(0, \Sigma), \tag{15}$$

with the rows of $\mathbf{Y}_{n\times q}$ correspond to q response variables on each of n individuals, $\beta_{m\times q}$ is a matrix of unknown parameters and $\mathbf{X}_{n\times m}$ is a known matrix of covariate values. The rows of the matrix $\mathbf{U}_{n\times q}$ are assumed to be independently distributed as a q-variate normal distribution with mean zero and covariance matrix Σ.

Table 1
Selected dimensions for linear regression

Selected dimensions												
$n = 20$							$n = 30$					
p	KICc	KIC	AIC	AICc	SIC	HQ	KICc	KIC	AIC	AICc	SIC	HQ
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	924	766	586	855	766	632	914	841	664	842	872	759
4	54	99	142	85	99	129	55	79	132	89	70	111
5	17	64	99	38	64	93	21	37	87	37	30	59
6	4	33	79	12	33	70	7	23	55	20	16	39
7	1	38	94	10	38	76	3	20	62	12	12	32

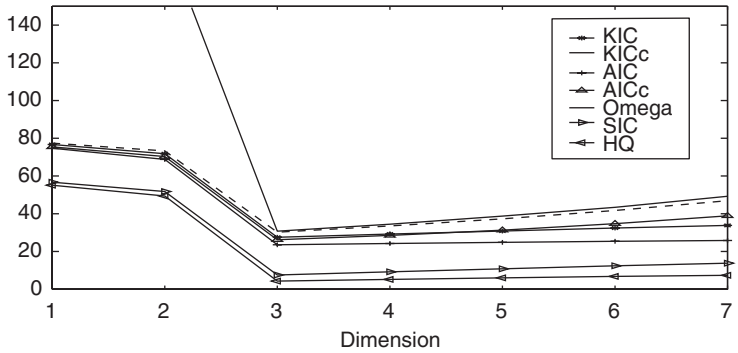

Fig. 1. Simulated $\Omega\left(d, \theta_0\right)$ and average of all criteria curves, with $p_0 = 3$ and $n = 20$ (multiple regression).

Ignoring the constant terms, the log-likelihood function

$$\ln f(\mathbf{Y}|\theta) = -0.5n \ln|\Sigma| - 0.5 \operatorname{tr}\left\{(\mathbf{Y} - \mathbf{X}\beta)\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\beta)^{\mathrm{t}}\right\},$$

where "tr" stands for trace, yields maximum likelihood estimates

$$\hat{\beta} = \left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{Y} \quad \text{and} \quad \hat{\Sigma} = \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)^{\mathrm{t}}\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)\Big/ n.$$

The true model has the following form:

$$\mathbf{Y} = \mathbf{X}_0\beta_0 + \mathbf{U}_0, \quad \mathbf{U}_0 \sim N_q\left(0, \Sigma_0\right), \tag{16}$$

where the matrices $\mathbf{X}_0$ and β_0 are $n \times m_0$ and $m_0 \times q$, respectively. The rows of $\mathbf{U}_0$ are independent with mean zero and covariance matrix Σ_0. The number of unknown parameters in the candidate model is $d = qm + 0.5q(q+1)$.

Then, the KIC criterion for model selection in multivariate regression is given by (cf. Cavanaugh, 1999)

$$KIC = n\left(\ln\left|\hat{\Sigma}\right| + q\right) + 3d.$$

Under the assumption that the family of candidate models includes the true model, the columns of $\mathbf{X}$ can be rearranged so that $\mathbf{X}_0\beta_0 = \mathbf{X}\beta^{\star}$, where $\beta^{\star} = \left(\beta_0^{\mathrm{t}}, \beta_1^{\mathrm{t}}\right)^{\mathrm{t}}$ and β_1 is an $(m - m_0) \times q$ matrix of zeros. The following proposition gives the KICc criterion for multivariate regression. Its proof is given in Appendix B.

Proposition 3.1. *The criterion defined by*

$$KICc = n\left(\ln\left|\hat{\Sigma}\right| + q\right) + \frac{d(3n - m - q - 1)}{n - m - q - 1}$$

is an asymptotically unbiased estimator of $\Omega\left(d, \theta_0\right)$.

Table 2
Signal-to-noise ratios and probabilities of overfitting for linear regression

Signal-to-noise ratios of overfitting							Probabilities of overfitting					
$n = 15$, True order $= 6$												
L	KICc	KIC	AIC	AICc	SIC	HQ	KICc	KIC	AIC	AICc	SIC	HQ
1	3.204	0.360	0.003	2.850	0.255	0.000	0.019	0.220	0.317	0.025	0.244	0.318
2	5.066	0.400	−0.067	4.600	0.264	−0.070	0.002	0.247	0.393	0.004	0.283	0.283
3	7.150	0.350	−0.173	6.627	0.198	−0.177	0.000	0.276	0.461	0.000	0.322	0.463
4	10	0.232	−0.310	9.455	0.074	−0.314	0.000	0.321	0.536	0.000	0.376	0.538
$n = 25$, True order $= 6$												
1	1.974	0.780	0.284	1.479	0.888	0.451	0.051	0.147	0.237	0.079	0.133	0.201
2	2.880	1.040	0.360	2.200	1.189	0.590	0.020	0.130	0.257	0.041	0.112	0.204
3	3.649	1.195	0.389	2.842	1.372	0.662	0.007	0.115	0.266	0.019	0.095	0.202
4	4.374	1.260	0.388	3.474	1.485	0.962	0.002	0.105	0.277	0.008	0.084	0.202
5	5.097	1.333	0.364	4.128	1.546	0.692	0.000	0.100	0.291	0.003	0.078	0.207
6	5.850	1.340	0.320	4.828	1.564	0.665	0.000	0.010	0.310	0.001	0.076	0.216
7	6.655	1.313	0.258	5.599	1.545	0.615	0.000	0.103	0.334	0.000	0.078	0.231
8	7.546	1.256	0.180	6.470	1.492	0.543	0.000	0.110	0.364	0.000	0.082	0.251
$n = 50$, True order $= 6$												
1	1.604	1.100	0.495	1.004	1.645	0.933	0.070	0.110	0.192	0.120	0.068	0.128
2	2.290	1.520	0.680	1.450	2.286	1.292	0.037	0.080	0.186	0.086	0.037	0.101
3	2.831	1.824	0.808	1.815	2.751	1.548	0.019	0.058	0.173	0.059	0.21	0.079
4	3.302	2.062	0.903	2.143	3.119	1.747	0.009	0.043	0.161	0.039	0.012	0.063
5	3.730	2.256	0.977	2.452	3.422	1.908	0.004	0.033	0.151	0.026	0.007	0.051
6	4.130	2.416	1.034	2.749	3.676	2.040	0.002	0.025	0.143	0.017	0.005	0.042
7	4.513	2.550	1.077	3.040	3.892	2.149	0.001	0.020	0.137	0.010	0.003	0.036
8	4.882	2.660	1.108	3.330	4.076	2.238	0.000	0.017	0.133	0.006	0.002	0.031

When there is a single response variable, this criterion reduces to the KICc criterion for multiple regression given in Section 2. Moreover, this criterion can be written as

$$KICc = KIC + \frac{2d(m+q+1)(n-m-q-1)}{n-q-2}$$

and

$$KICc = AICc + d,$$

where $AICc = n\left(\ln\left|\hat{\Sigma}\right| + q\right) + 2dn/(n-m-q-1)$ (cf. Bedrick and Tsai, 1994).

As in multiple regression, the reduction in bias for KICc is achieved without any increases in variance, since KICc is the sum of KIC and an additional constant term.

3.2. *The signal-to-noise ratio and probability of overfitting*

As in multiple regression and using both equations (4A.4) and (4A.5) from McQuarrie and Tsai (1998, pp. 182–183), the signal-to-noise ratio of KIC is

$$\begin{aligned}\frac{\mathbf{E}[\Delta KIC]}{Sd[\Delta KIC]} = {} & \frac{\sqrt{(n-m_0-L-(q-1)/2)\,(n-m_0-(q-1)/2+2)}}{\sqrt{2Lq}} \\ & \times\left\{q\ln\left(\frac{n-m_0-L-(q-1)/2}{n-m_0-(q-1)/2}\right)\right. \\ & \left.-\frac{Lq}{(n-m_0-L-(q-1)/2)\,(n-L-(q-1)/2)}+\frac{3Lq}{n}\right\}.\end{aligned}$$

As $L \longrightarrow (n-m_0-q)$, the first two terms, which derive from $\ln\hat{\Sigma}$, decrease to $-\infty$. Meanwhile, the last term, which arises from the penalty term, decreases to 0 as L increases. Therefore, the signal-to-noise ratio of KIC increases for small L, decreases to $-\infty$ as L increases resulting in small sample overfitting problems.

Analogously, the signal-to-noise ratio of KICc is

$$\begin{aligned}\frac{\mathbf{E}[\Delta KICc]}{Sd[\Delta KICc]} = {} & \frac{\sqrt{(n-m_0-L-(q-1)/2)\,(n-m_0-(q-1)/2+2)}}{\sqrt{2Lq}} \\ & \times\left\{q\ln\left(\frac{n-m_0-L-(q-1)/2}{n-m_0-(q-1)/2}\right)\right. \\ & -\frac{Lq}{(n-m_0-L-(q-1)/2)\,(n-L-(q-1)/2)} \\ & \left.+\frac{Lq(2n-q-1)}{(n-m_0-L-q-1)\,(n-m_0-q-1)}+\frac{Lq}{n}\right\}.\end{aligned}$$

The following lemma shows that the signal-to-noise ratio of KICc increases as L increases. So KICc should resist overfitting well. The proof is presented in Appendix B.

Lemma 3.2. *Let the true and candidate multivariate regression models be given by equations* (16) *and* (15), *where* $0 < m_0 \leqslant m < n - q - 6$. *For* $n \geqslant q + 5$ *and overfitting* L *where* $0 < L < (2/3)\,(n - m_0 - (q-1)/2)$, *the signal-to-noise ratio of KICc increases as* L *increases.*

We will now examine the small-sample probability of overfitting. Suppose that there is a true order m_0 and we fit a candidate model of order $m_0 + L$ where $L > 0$ and that those models are nested. The probability that KIC selected the overfitted model $m_0 + L$ is

$$P\left(KIC_{m_0+L} < KIC_{m_0}\right) = P\left\{n \sum_{i=1}^{q} \ln\left(1 + \frac{\chi_L^2}{\chi_{n-m_0-L+i}^2}\right) > 3Lq\right\}.$$

The detailed calculations are given in the Appendix B. Similarly the probability that KICc selected the overfitted model $m_0 + L$ is

$$P\left(KICc_{m_0+L} < KICc_{m_0}\right) = P\left\{n \sum_{i=1}^{q} \ln\left(1 + \frac{\chi_L^2}{\chi_{n-m_0-L+i}^2}\right) > \frac{Lq(2n-q-1)n}{(n-m_0-L-q-1)\,(n-m_0-q-1)} + Lq\right\}.$$

The probabilities of overfitting of the two criteria are not easy to compare analytically. We present a small numerical comparison of them in the next section.

3.3. Numerical experiments

To examine the performance of KICc criterion, we consider the same example as in Bedrick and Tsai (1994). We choose 1000 samples of sizes $n = 20$ and $n = 35$ generated from model (16), with $m_0 = 4$ and $\Sigma_0 = (1-\rho)\mathbf{I}_q + \rho\mathbf{J}_q$, where $\mathbf{J}_q$ is a $q \times q$ matrix of ones, $\mathbf{I}_q$ is $q \times q$ identity matrix, the selected values of ρ are 0.3, 0.6 and 0.8 and $q = 2, 4$. We consider eight candidate models stored in an $n \times 8$ matrix $\mathbf{X}$, with a column of ones, followed by seven columns of independent standard normal random variables. As in the multiple regression example, the candidate models include the columns of $\mathbf{X}$ in a sequentially nested fashion. For each of the q responses, the regression parameters are identical: (1, 2, 3, 4). Thus, we use the same matrix and values of ρ for each pair (n, q).

For each of the 1000 samples, all eight models in the candidate class are fit to the data. As in Bedrick and Tsai (1994), we have obtained the same simulation results for different values of the correlation ρ between responses. So we present only results for $\rho = 0.6$. The selections over the 1000 data sets are summarized in Table 3. In this context, the SIC and HQ criteria are defined by $SIC = n \ln\left|\hat{\Sigma}\right| + d\ln(n)$ and $HQ = n \ln\left|\hat{\Sigma}\right| + 2d\ln(\ln(n))$, respectively.

For $n = 20$, KICc performs better than the other criteria, except for the AICc which yields almost the same results as KICc. Moreover, the other criteria have a tendency to overfit the true model. On the other hand, when $n = 35$ all the criteria perform well, but KICc provides the best selection of m. The tendency to overfit the true model for AIC, KIC and HQ decreases as n increases here.

Table 3
Selected dimensions for multivariate regression

		Selected dimensions											
$n = 20$								$n = 35$					
p	m	KICc	KIC	AIC	AICc	SIC	HQ	KICc	KIC	AIC	AICc	SIC	HQ
2	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0
	4	980	793	545	956	793	611	968	893	698	908	936	821
	5	19	97	153	37	97	146	26	63	119	66	43	92
	6	1	45	93	5	45	80	6	25	66	18	14	42
	7	0	38	89	2	38	82	0	7	55	5	4	22
	8	0	27	120	0	27	27	0	12	62	3	3	23
4	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0
	4	998	845	506	994	843	607	996	939	748	977	978	894
	5	2	89	127	6	89	121	3	46	113	19	17	66
	6	0	25	0	84	26	68	1	12	63	4	5	25
	7	0	20	101	0	20	79	0	1	36	0	0	9
	8	0	21	182	0	22	125	0	2	40	0	0	6

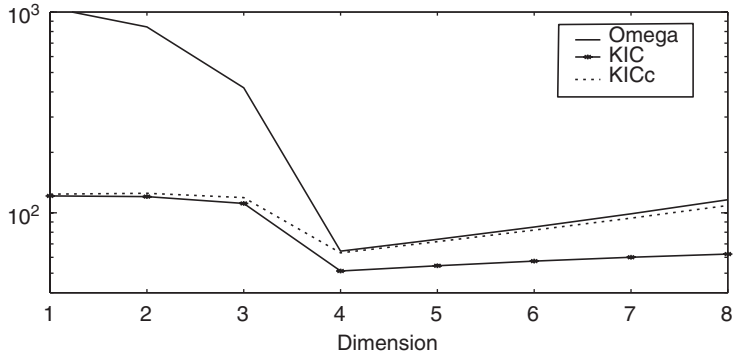

Fig. 2. Simulated $\Omega\left(d, \theta_0\right)$ and average KIC and KICc curves, with $m_0 = 4$, $q = 2$, $\rho = 0.6$ and $n = 20$ (multivariate regression).

In Fig. 2, we plot the average values of KIC, KICc and $\Omega\left(d, \theta_0\right)$ as a function of m. When $m < m_0$ the average symmetric Kullback–Leibler discrepancy ($\boldsymbol{\Omega}$) decreases rapidly; it then increases for $m > m_0$. Moreover, if $m > m_0$ the average KICc curve closely follows the simulated $\Omega\left(d, \theta_0\right)$ curve, while KIC drastically underestimates $\mathbf{E}_{\theta_0}\left(\Omega\left(d, \theta_0\right)\right)$.

Now, we examine both the signal-to-noise ratios and the probabilities of overfitting for bivariate regression model. We consider sample sizes $n = 20, 35, 50$, $q = 2$ and $m_0 = 5$. The results are given in Table 4.

When $n = 20$, KICc and AICc have a high signal-to-noise ratios, and their probabilities of overfitting are almost zero. The other criteria have smaller signal-to-noise ratios and higher probabilities of overfitting. However, the probabilities of overfitting for KICc and AICc increase as the sample size increases, whereas they decrease for the other criteria.

4. Derivation of KICc for univariate autoregressive models

The univariate autoregressive model is one of the most popular for time series data. It describes the present value as a linear combination of the past observations. In this section, we outline KICc for the univariate autoregressive model and we derive its small sample signal-to-noise ratio in this framework.

4.1. KICc for univariate autoregressive model

We first define the general autoregressive model of order p, denoted AR(p), as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t; \quad t = p+1, \ldots, n, \tag{17}$$

where ε_t i.i.d $N\left(0, \sigma^2\right)$, and $y_1, y_2, \ldots, y_n$ are an observed series of data. However, although we have n observations, because we use y_{t-p} to model y_t the effective series length is $T = n - p$.

Table 4
Signal-to-noise ratios and probabilities of overfitting for multivariate regression

Signal-to-noise ratios of overfitting							Probabilities of overfitting					
$n = 20$, True order $= 5$, $q = 2$												
L	KICc	KIC	AIC	AICc	SIC	HQ	KICc	KIC	AIC	AICc	SIC	HQ
1	3.294	0.950	0.300	2.644	0.947	0.426	0.013	0.141	0.271	0.025	0.141	0.239
2	4.906	1.232	0.350	4.024	1.228	0.522	0.001	0.109	0.285	0.003	0.110	0.239
3	6.375	1.368	0.336	5.343	1.363	0.536	0.000	0.093	0.302	0.000	0.093	0.244
4	7.881	1.410	0.277	6.748	1.405	0.497	0.000	0.086	0.328	0.000	0.087	0.260
$n = 35$, True order $= 5$, $q = 2$												
1	2.472	1.400	0.600	1.672	1.844	1.029	0.031	0.090	0.201	0.068	0.057	0.131
2	3.555	1.917	0.807	2.444	2.534	1.403	0.007	0.050	0.178	0.027	0.024	0.092
3	4.431	2.271	0.937	3.097	3.012	1.653	0.001	0.030	0.159	0.010	0.011	0.067
4	5.215	2.532	1.022	3.704	0.371	1.833	0.000	0.019	0.146	0.003	0.005	0.052
5	5.949	2.728	1.074	4.295	3.646	1.962	0.000	0.013	0.137	0.001	0.003	0.042
6	6.670	2.873	1.101	4.887	3.858	2.052	0.000	0.010	0.133	0.000	0.001	0.035
7	7.362	2.977	1.106	5.490	4.017	2.110	0.000	0.007	0.132	0.000	0.001	0.031
8	8.068	3.045	1.091	6.115	4.130	2.140	0.000	0.006	0.135	0.000	0.000	0.029
$n = 50$, True order $= 5$, $q = 2$												
1	2.282	1.580	0.720	1.422	2.364	1.346	0.037	0.075	0.179	0.088	0.034	0.095
2	3.257	2.191	0.989	2.055	3.287	1.864	0.010	0.037	0.146	0.043	0.009	0.054
3	4.028	2.630	1.176	2.573	3.956	2.235	0.002	0.019	0.121	0.020	0.003	0.032
4	4.698	2.975	1.317	3.039	4.487	2.524	0.000	0.010	0.102	0.009	0.001	0.020
5	5.307	3.256	1.426	3.477	4.925	2.758	0.000	0.006	0.088	0.004	0.000	0.013
6	5.879	3.489	1.510	3.900	5.293	2.951	0.000	0.003	0.079	0.001	0.000	0.009
7	6.423	3.683	1.575	4.314	5.606	3.110	0.000	0.002	0.071	0.000	0.000	0.006
8	6.950	3.845	1.622	4.727	5.873	3.241	0.000	0.001	0.066	0.000	0.000	0.004

Reducing the number of observations by p is just one of the possible methods for autoregressive modeling (see McQuarrie and Tsai, 1998, p. 90). Indeed, as explained in McQuarrie and Tsai (1998, p. 90), when $\{y_t\}$ is a zero mean time series, Priestley (1981) suggests an alternative method which consists of replacing past observations $y_0, y_{-1}, \ldots$ with 0. Box and Jenkins (1976) suggest backcasting to estimate the past observations. Each of these methods has its own merits; however, we prefer losing the first p observations primarily because these other methods require extra assumptions, such as zero mean, stationarity, or that the backcasted values are good estimates for $y_0, y_{-1}, \ldots$.

For now, we assume that the candidate model is an AR(p) model. Moreover, we suppose there is a finite true order p_0, and we further assume the true AR (p_0) model belongs to the set of candidate models of orders 1 to P, where $P \geqslant p_0$. Thus, we can form a regression model from (17) by conditioning on the past and forming the design matrix $\mathbf{X}$. The resulting regression model takes the form

$$\mathbf{Y} = \mathbf{X}\Phi + \varepsilon,$$

where $\mathbf{Y} = \left(y_{p+1}, \ldots, y_n\right)^{\mathrm{t}}$, $\Phi = \left(\phi_1, \ldots, \phi_p\right)^{\mathrm{t}}$, $\varepsilon = \left(\varepsilon_{p+1}, \ldots, \varepsilon_n\right)^{\mathrm{t}}$ and $\mathbf{X}$ is the $(n-p) \times p$ design matrix assumed to be of full rank p with elements $x_{ij} = y_{i-j}$ for $i = p+1, \ldots, n$, $j = 1, \ldots, p$. Since $\mathbf{X}$ is formed by conditioning on past values of $\mathbf{Y}$, least squares estimation of the parameters in this context is referred to as conditional least squares and is given by

$$\hat{\Phi} = \left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{Y} \quad \text{and} \quad \hat{\sigma}^2 = \left(\mathbf{Y} - \mathbf{X}\hat{\Phi}\right)^{\mathrm{t}}\left(\mathbf{Y} - \mathbf{X}\hat{\Phi}\right) \Big/ T.$$

The derivation of the KICc criterion for univariate autoregressive models is similar to that of multiple regression with the exception that the sample size for the autoregressive model is a function of the order of the candidate models. Then, following the same proof of KICc for multiple regression, we can show that the second term of (9) can be estimated by $(p+1)$. Furthermore, Hurvich and Tsai (1989) showed that $T\left(\ln \hat{\sigma}^2 + 1\right) + 2T(p+1)/(T-p-2)$ is an approximately unbiased estimator of the first term of (9). Therefore, we obtain an asymptotically unbiased estimator of the Kullback symmetric divergence defined in the following proposition:

Proposition 4.1. *The criterion defined by*

$$KICc = T\left(\ln \hat{\sigma}^2 + 1\right) + \frac{(p+1)(3T-p-2)}{T-p-2}$$

is an asymptotically unbiased estimator of $\Omega\,(d, \theta_0)$.

Other equivalent forms of this criterion are

$$KICc = KIC + \frac{2(p+1)(p+2)}{T-p-2},$$
$$KICc = AICc + (p+1),$$

where $KIC = T\left(\ln\hat{\sigma}^2 + 1\right) + 3(p+1)$ and $AICc = T\left(\ln\hat{\sigma}^2 + 1\right) + (2T(p+1))/(T-p-2)$. This KICc criterion has the same form as the KICc obtained in multiple regression with $n = T$.

4.2. The signal-to-noise ratio and probability of overfitting

Following Eqs. (3A.6) and (3A.7) from McQuarrie and Tsai (1998, pp. 131–132) and scaling the first two terms of KIC criterion by $T = n - p$, the signal-to-noise ratio of KIC is

$$\frac{\mathbf{E}[\Delta KIC]}{Sd[\Delta KIC]} = \frac{\sqrt{(n-2p_0-2L)(n-2p_0+2)}}{\sqrt{4L}}\left\{\ln\left(1-\frac{2L}{n-2p_0}\right) - \ln\left(1-\frac{L}{n-p_0}\right) - \frac{2L}{(n-2p_0-2L)(n-2p_0)} + \frac{3L(n+1)}{(n-p_0-L)(n-p_0)}\right\}.$$

Similarly, the signal-to-noise ratio of KICc criterion is

$$\frac{\mathbf{E}[\Delta KICc]}{Sd[\Delta KICc]} = \frac{\sqrt{(n-2p_0-2L)(n-2p_0+2)}}{\sqrt{4L}}\left\{\ln\left(1-\frac{2L}{n-2p_0}\right) - \ln\left(1-\frac{L}{n-p_0}\right) - \frac{2L}{(n-2p_0-2L)(n-2p_0)} + \frac{2Ln}{(n-2p_0-2L-2)(n-2p_0-2)} + \frac{L(p_0+1)}{(n-p_0-L)(n-p_0)}\right\}.$$

Thus, the probability that KIC prefers the overfitting model $(p_0 + L)$ is defined by (its proof is given in Appendix C)

$$P\left(KIC_{p_0+L} < KIC_{p_0}\right) = P\left\{F_{2L,n-2p_0-2L} > \frac{n-2p_0-2L}{2L}\left(\frac{n-p_0}{n-p_0-L} \times \exp\left(\frac{3L(n+1)}{(n-p_0-L)(n-p_0)}\right) - 1\right)\right\}.$$

Similarly, the probability that KICc prefers the overfitting model $(p_0 + L)$ is

$$P\left(KICc_{p_0+L} < KICc_{p_0}\right) = P\left\{F_{2L,n-2p_0-2L} > \frac{n-2p_0-2L}{2L}\left(\frac{n-p_0}{n-p_0-L} \times \exp\left(\frac{2Ln}{(n-2p_0-2L-2)(n-2p_0-2)} + \frac{L(p_0+1)}{(n-p_0-L)(n-p_0)}\right) - 1\right)\right\}.$$

As in the previous sections, the quantiles of the two criteria are not easy to compare analytically. We present a small numerical comparison of them in the next section.

4.3. Numerical illustrations

Now, we evaluate the performance of all criteria in a small sample, with $SIC = T \ln \hat{\sigma}^2 + (p+1)\ln(T)$ and $HQ = T \ln \hat{\sigma}^2 + 2(p+1)\ln(\ln(T))$. We consider 1000 samples generated from the second-order autoregressive model as in Hurvich and Tsai (1989)

$$y_t = 0.994 y_{t-1} - 0.8 y_{t-2} + \varepsilon_{0t}.$$

We choose $\sigma_0^2 = 1$ and two samples size $n = 23$ and $n = 35$ with the maximum order $p = 10$. Each vector $\mathbf{Y}$ of the 1000 samples was generated by starting at y_{-50} with $y_t = 0$ for all $t < -50$, but only observations $y_3, \ldots, y_n$ were retained for analysis. The row labeled 6–10, in Table 5, corresponds to the totals of models selected by each criteria for the dimensions 6–10.

With $n = 23$ we see from the first entry in Table 5 that KICc performs best, followed by SIC and AICc. These criteria do not exhibit the same tendency to overfit as AIC, KIC and HQ do. When $n = 35$, SIC outperforms KICc, but the latter is still better than all the other criteria. The tendency of all criteria to underestimate the correct dimension decreases and is approximately equal to zero.

Now, we present the calculation of the probability of overfitting and the signal-to-noise ratio for autoregressive models. We consider three sample sizes $n = 15$, $n = 25$ and $n = 50$. The true order is fixed at $p_0 = 5$, except for $n = 15$ where $p_0 = 2$. Table 6 shows that our results are similar to those obtained in the multiple and multivariate regression examples of the preceding sections.

5. Conclusion

In this paper, we have derived and investigated the KICc criterion, based on Kullback's symmetric divergence, for model selection in univariate autoregressive models and in multiple and multivariate regression models. Moreover, a large simulation study is undertaken to compare the performance of our criterion to other well know criteria. Our simulations show that KICc performs better in small samples and is asymptotically equivalent to KIC. Also, our simulation studies show that among the efficient criteria studied (AIC, KIC, AICc), KICc performed best in multiple and multivariate regression. For univariate autoregression, KICc was slightly outperformed by the consistent criterion SIC.

The signal-to-noise ratio study indicated that when the sample size is small, KICc has a greater signal-to-noise ratio, and its probability of overfitting is almost zero. In contrast, KIC has a tendency to overfit, because it has a low signal-to-noise ratio, which leads to a higher probability of overfitting.

Table 5
Selected dimensions for univariate autoregressive model

	Selected dimensions											
$n = 23$							$n = 35$					
p	KICc	KIC	AIC	AICc	SIC	HQ	KICc	KIC	AIC	AICc	SIC	HQ
1	33	21	13	19	38	23	2	2	0	1	8	2
2	930	878	741	900	903	877	934	887	755	864	953	914
3	33	64	105	63	43	62	53	79	116	101	34	63
4	4	27	61	17	14	14	9	19	58	25	5	15
5	0	4	24	1	1	3	2	8	27	7	0	5
6–10	0	6	56	0	1	9	0	5	44	2	0	1

Table 6
Signal-to-noise ratios and probabilities of overfitting for the univariate autoregressive model

Signal-to-noise ratios of overfitting							*Probabilities of overfitting*					
					$n = 15$, True order $= 2$							
L	KICc	KIC	AIC	AICc	SIC	HQ	KICc	KIC	AIC	AICc	SIC	HQ
1	1.590	0.744	0.283	1.503	0.380	0.104	0.075	0.174	0.277	0.082	0.252	0.332
2	2.856	0.913	0.295	2.742	0.377	0.015	0.107	0.158	0.303	0.019	0.279	0.399
3	5.110	0.856	0.172	4.982	0.206	−0.188	0.001	0.172	0.355	0.001	0.343	0.496
			$n = 25$, True order $= 5$									
1	1.709	0.667	0.223	1.606	0.509	0.201	0.066	0.188	0.294	0.074	0.221	0.301
2	2.761	0.852	0.246	2.622	0.608	0.196	0.170	0.019	0.319	0.022	0.221	0.336
3	4.009	0.899	0.195	3.847	0.580	0.111	0.166	0.003	0.350	0.004	0.236	0.380
4	5.842	0.818	0.068	5.669	0.439	−0.050	0.000	0.184	0.403	0.000	0.279	0.449
5	9.419	0.583	−0.151	9.249	0.170	−0.299	0.000	0.239	0.494	0.000	0.368	0.558
			$n = 50$, True order $= 5$									
1	0.873	0.895	0.406	0.816	1.195	0.679	0.153	0.150	0.245	0.163	0.111	0.187
2	1.284	1.246	0.558	1.203	1.654	0.935	0.106	0.110	0.233	0.116	0.070	0.156
3	1.639	1.500	0.662	1.540	1.978	1.111	0.069	0.082	0.217	0.079	0.045	0.131
4	1.978	1.698	0.737	1.865	2.224	1.239	0.043	0.063	0.204	0.051	0.031	0.113
5	2.318	1.855	0.789	2.193	2.414	1.333	0.026	0.050	0.195	0.031	0.023	0.101
6	2.672	1.978	0.823	2.537	2.557	1.398	0.014	0.041	0.189	0.018	0.017	0.092
7	3.050	2.071	0.839	2.905	2.659	1.435	0.007	0.035	0.187	0.010	0.014	0.087
8	3.462	2.135	0.837	3.310	2.721	1.447	0.003	0.031	0.188	0.005	0.012	0.086

Acknowledgements

The authors wish to thank an Associate Editor and an anonymous referee for extremely helpful comments which improve the presentation of this paper. We also thank Samuel Buttrey for reading the paper and for his valuable suggestions.

Appendix A.

Proof of Lemma 2.2. Consider the signal-to-noise ratio of KICc as a function of L

$$STN_{KICc}(L) = \frac{\sqrt{(n-p_0-L)(n-p_0+2)}}{\sqrt{2L}}\left\{\ln\left(\frac{n-p_0-L}{n-p_0}\right) - \frac{L}{(n-p_0-L)(n-p_0)} + \frac{2L(n-1)}{(n-p_0-L-2)(n-p_0-2)} + \frac{L}{n}\right\},$$

where $0 < p_0 < n-6$, $n \geqslant 6$ and $0 < L < \left(\frac{2}{3}\right)(n-p_0)$. Using the signal-to-noise ratio of AICc defined in McQuarrie and Tsai (1998, p. 27), we can write

$$STN_{KICc}(L) = STN_{AICc}(L) + \frac{\sqrt{(n-p_0-L)(n-p_0+2)}}{\sqrt{2L}}\left(\frac{L}{n}\right).$$

Since $STN(L)$ is continuous in L, it is enough to show that its derivative with respect to L is positive: indeed, we have

$$\frac{\mathrm{d}}{\mathrm{d}L}STN_{KICc}(L) = \frac{\mathrm{d}}{\mathrm{d}L}STN_{AICc}(L) + \frac{\sqrt{n-p_0+2}\,(2n-2p_0-3L)}{n\sqrt{2L\,(n-p_0-L)}}.$$

Applying Theorem 2.1 from McQuarrie and Tsai (1998), $(\mathrm{d}/\mathrm{d}L)STN_{AICc}(L) > 0$, and since $0 < L < \left(\frac{2}{3}\right)(n-p_0)$, then the second term is also positive. Hence, the signal-to-noise increases as L increases.

Appendix B.

B.1. The probability of overfitting for multivariate regression

We will present detailed calculations for KIC criterion and the same method can be used for KICc. Let $SPE_m = n\hat{\Sigma}_m$. We know that KIC overfits if $KIC_{m_0+L} < KIC_{m_0}$.

Then the probability that KIC selected the overfitted model $m_0 + L$ is

$$\begin{aligned} P\left(KIC_{m_0+L} < KIC_{m_0}\right) &= P\left\{\ln\left|\hat{\Sigma}_{m_0+L}\right| + \frac{3\left((m_0+L)\,q + 0.5q(q+1)\right)}{n}\right. \\ &\qquad \left. < \ln\left|\hat{\Sigma}_{m_0}\right| + \frac{3\left(m_0 q + 0.5q(q+1)\right)}{n}\right\} \\ &= P\left\{\ln\frac{\left|SPE_{m_0+L}\right|}{\left|SPE_{m_0}\right|} < \frac{-3Lq}{n}\right\}. \end{aligned}$$

Or, we have (McQuarrie and Tsai, 1998, p. 182)

$$\ln\left(\frac{\left|SPE_{m_0+L}\right|}{\left|SPE_{m_0}\right|}\right) \sim \sum_{i=1}^{q} \log-\text{Beta}\left(\frac{n-m_0-L-q+i}{2}, \frac{L}{2}\right).$$

Then we have for KIC

$$P\left(KIC_{m_0+L} < KIC_{m_0}\right) = P\left\{n\sum_{i=1}^{q}\ln\left(1+\frac{\chi_L^2}{\chi_{n-m_0-L+i}^2}\right) > 3Lq\right\}.$$

B.2. Proof of Lemma 3.2

Using the same proof of Lemma 2.2, and applying Theorem 4.1 from McQuarrie and Tsai (1998), the result of Lemma 3.2 follows immediately.

B.3. Proof of Proposition 3.1

The log-likelihood of the candidate model is given by

$$\ln f(\mathbf{Y}\mid\theta) = -0.5nq\ln 2\pi - 0.5n\ \ln|\Sigma| - 0.5\,\mathrm{tr}\{(\mathbf{Y}-\mathbf{X}\beta)\Sigma^{-1}(\mathbf{Y}-\mathbf{X}\beta)^{\mathrm{t}}\}.$$

The expectation of $K\left(\theta_0, \hat{\theta}\right)$ has the form

$$\Omega\left(d, \theta_0\right) = \mathbf{E}_{\theta_0}\left\{D\left(\theta_0, \hat{\theta}\right)\right\} + \mathbf{E}_{\theta_0}\left\{D\left(\hat{\theta}, \theta_0\right) - D\left(\hat{\theta}, \hat{\theta}\right)\right\}.$$

Now, we will compute the second term of this equation. Indeed, ignoring the constant term in the log-likelihood, we obtain

$$\begin{aligned} D\left(\hat{\theta}, \theta_0\right) &= \mathbf{E}_\theta\left\{n\ \ln|\Sigma_0| + \mathrm{tr}\left\{\left(\mathbf{Y}-\mathbf{X}\beta_0\right)\Sigma_0^{-1}\left(\mathbf{Y}-\mathbf{X}\beta_0\right)^{\mathrm{t}}\right\}\right\}\Big|_{\theta=\hat{\theta}} \\ &= n\ln|\Sigma_0| + \mathbf{E}_\theta\left\{\mathrm{tr}\left\{(\mathbf{Y}-\mathbf{X}\beta)\Sigma_0^{-1}(\mathbf{Y}-\mathbf{X}\beta)^{\mathrm{t}}\right.\right. \\ &\qquad \left.\left. +\Sigma_0^{-1}\left(\beta-\beta_0\right)^{\mathrm{t}}\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)\left(\beta-\beta_0\right)\right\}\right\}\Big|_{\theta=\hat{\theta}} \\ &= n\ln|\Sigma_0| + n\,\mathrm{tr}\left(\Sigma_0^{-1}\hat{\Sigma}\right) + \mathrm{tr}\left\{\Sigma_0^{-1}\left(\hat{\beta}-\beta_0\right)^{\mathrm{t}}\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)\left(\hat{\beta}-\beta_0\right)\right\}. \end{aligned} \tag{18}$$

Similarly, we have

$$D\left(\hat{\theta},\hat{\theta}\right)=\mathbf{E}_{\theta}\left\{n\ln|\Sigma|+\operatorname{tr}\left\{(\mathbf{Y}-\mathbf{X}\beta)\Sigma^{-1}(\mathbf{Y}-\mathbf{X}\beta)^{\mathrm{t}}\right\}\right\}\Big|_{\theta=\hat{\theta}}$$
$$=n\ \ln\left|\hat{\Sigma}\right|+nq. \tag{19}$$

The difference between (18) and (19) leads to

$$D\left(\hat{\theta},\theta_0\right)-D\left(\hat{\theta},\hat{\theta}\right)=-n\ln\frac{\left|\hat{\Sigma}\right|}{|\Sigma_0|}+n\operatorname{tr}\left(\Sigma_0^{-1}\hat{\Sigma}\right)$$
$$+\operatorname{tr}\left\{\Sigma_0^{-1}\left(\hat{\beta}-\beta_0\right)^{\mathrm{t}}\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)\left(\hat{\beta}-\beta_0\right)\right\}-nq.$$

It is well known that $n\hat{\Sigma}\sim Wishart_q\ (n-m,\ \Sigma_0)$, then $\mathbf{E}_{\theta_0}\left(n\hat{\Sigma}\right)=(n-m)\Sigma_0$ (cf. Muirhead, 1982, p. 137) and

$$\mathbf{E}_{\theta_0}\left(n\operatorname{tr}\left(\Sigma_0^{-1}\hat{\Sigma}\right)\right)=\operatorname{tr}\left\{\mathbf{E}_{\theta_0}\left(n\Sigma_0^{-1}\hat{\Sigma}\right)\right\}$$
$$=(n-m)q.$$

Furthermore, $vec\left(\hat{\beta}\right)\sim N_{pm}\left(vec(\beta),\ \Sigma_0\otimes\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\right)$, where $vec(\beta)$ is the $qm\times 1$ vector obtained by stacking the columns of β and $\otimes$ is the Kronecker product. This yields

$$\mathbf{E}_{\theta_0}\left\{\operatorname{tr}\left\{\Sigma_0^{-1}\left(\hat{\beta}-\beta_0\right)^{\mathrm{t}}\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)\left(\hat{\beta}-\beta_0\right)\right\}\right\}$$
$$=\operatorname{tr}\left\{\mathbf{E}_{\theta_0}\left\{\Sigma_0^{-1}\left(\hat{\beta}-\beta_0\right)^{\mathrm{t}}\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)\left(\hat{\beta}-\beta_0\right)\right\}\right\}$$
$$=\mathbf{E}_{\theta_0}\left\{vec\left(\hat{\beta}-\beta_0\right)^{\mathrm{t}}\left\{\Sigma_0^{-1}\otimes\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)\right\}vec\left(\hat{\beta}-\beta_0\right)\right\}$$
$$=qm,$$

which leads to

$$\mathbf{E}_{\theta_0}\left\{D\left(\hat{\theta},\theta_0\right)-D\left(\hat{\theta},\hat{\theta}\right)\right\}=-\mathbf{E}_{\theta_0}\left\{n\ln\frac{\left|\hat{\Sigma}\right|}{|\Sigma_0|}\right\}. \tag{20}$$

We have

$$n\ln\frac{\left|n\hat{\Sigma}\right|}{|\Sigma_0|}=nq\ln n+n\ln\frac{\left|\hat{\Sigma}\right|}{|\Sigma_0|}. \tag{21}$$

Next, from Muirhead (1982, p. 100), we know that $\left|n\hat{\Sigma}\right| \Big/ |\Sigma_0|$ has the same distribution as $\prod_{i=1}^{q} \chi^2_{(n-m-i+1)}$, where the $\chi^2_{(n-m-i+1)}$ denotes the independent χ^2 random variables with $(n-m-i+1)$ degrees of freedom . Then

$$\mathbf{E}_{\theta_0}\left\{n \ln \frac{\left|n\hat{\Sigma}\right|}{|\Sigma_0|}\right\} = \mathbf{E}_{\theta_0}\left\{n \ln \prod_{i=1}^{q} \chi^2_{(n-m-i+1)}\right\}$$
$$= \mathbf{E}_{\theta_0}\left\{n \sum_{i=1}^{q} \ln \chi^2_{(n-m-i+1)}\right\}. \tag{22}$$

Taking a second order expansion of $\ln \chi^2_{\mathrm{df}}$ about its degrees of freedom df and evaluating the expectation of the result, we obtain (Bickel and Doksum, 1977, p. 310)

$$\mathbf{E}\left\{\ln \chi^2_{\mathrm{df}}\right\} = \ln\mathrm{df} - 1/\mathrm{df} + O\left(1/\mathrm{df}^2\right).$$

Substituting this relation into (22) yields

$$\mathbf{E}_{\theta_0}\left\{n \sum_{i=1}^{q} \ln \chi^2_{(n-m-i+1)}\right\} = n \sum_{i=1}^{q} \left\{\ln(n-m-i+1) - \frac{1}{n-m-i+1} + O\left(\frac{1}{(n-m-i+1)^2}\right)\right\}. \tag{23}$$

Now, taking a first order expansion of $\ln(n-m-i+1)$ about n for $i = 1, \ldots, q$, it leaves

$$\ln(n-m-i+1) = \ln n + \frac{1-m-i}{n} + O\left(\frac{1-m-i}{n}\right)^2.$$

Substituting this equation into (23) yields

$$\mathbf{E}_{\theta_0}\left\{n \sum_{i=1}^{q} \ln \chi^2_{(n-m-i+1)}\right\}$$
$$= nq \ln n + \sum_{i=1}^{q} (1-m-i) - \sum_{i=1}^{q} \frac{n}{n-m-i+1} \tag{24}$$
$$+ \sum_{i=1}^{q} \left\{O\left(\frac{n}{(n-m-i+1)^2}\right) + O\left(\frac{(1-m-i)^2}{n}\right)\right\}. \tag{25}$$

As $n \to \infty$ and holding m constant, each of the two terms in (25), is asymptotically $o(1)$ for each i. However, each of these terms is also $o(1)$ when m is allowed to grow at a rate less than $\sqrt{n}$ when $n \to \infty$. Thus, when $n \to \infty$ the last term on the right of (24) converges to q, and since $\sum_{i=1}^{q} (1-m-i) = q - (qm + 0.5q(q+1))$ we establish that

$$\mathbf{E}_{\theta_0}\left\{n \sum_{i=1}^{q} \ln \chi^2_{(n-m-i+1)}\right\} = nq \ln n - (qm + 0.5q(q+1)) + o(1).$$

Substituting this equation into (22), and using (21) and (20) we get

$$\mathbf{E}_{\theta_0}\left\{D\left(\hat{\theta},\theta_0\right)-D\left(\hat{\theta},\hat{\theta}\right)\right\}=(qm+0.5q(q+1))+o(1). \tag{26}$$

To evaluate the first term of (9), we use the result of Bedrick and Tsai (1994) which establishes that an exact estimator of (8) is given by

$$AICc=n\left(\ln\left|\hat{\Sigma}\right|\right)+q)+\frac{2nd}{(n-(m+q+1))},$$

where $d=qm+0.5q(q+1)$. Now, from this result and (26), we obtain an asymptotically unbiased estimator of (9) given in the Proposition 3.1. □

Appendix C.

In this appendix, we give the detailed calculations for the probability of overfitting for KIC in the autoregressive context. Scaling the first two terms of the KIC criterion, defined in Section 4, by $T=n-p$, we then obtain

$$KIC=\left(\ln\hat{\sigma}^2+1\right)+\frac{3(p+1)}{n-p}.$$

Let $SSE_p=T\hat{\sigma}^2$. Using equations (3A.6) from McQuarrie and Tsai (1998, pp. 131–132), the probability that KIC selected the overfitted model p_0+L is

$$\begin{aligned}
P\left(KIC_{p_0+L}<KIC_{p_0}\right)&=P\left\{\ln\hat{\sigma}^2_{p_0+L}+\frac{3\left(p_0+L+1\right)}{n-p_0-L}<\ln\hat{\sigma}^2_{p_0}+\frac{3\left(p_0+1\right)}{n-p_0}\right\}\\
&=P\left\{\ln\frac{SSE_{p_0+L}}{SSE_{p_0}}<\frac{-3L(n+1)}{\left(n-p_0-L\right)\left(n-p_0\right)}\right.\\
&\quad\left.+\ln\left(1-\frac{L}{n-p_0}\right)\right\}\\
&=P\left\{\frac{SSE_{p_0}}{SSE_{p_0+L}}-1>\frac{n-p_0}{n-p_0-L}\right.\\
&\quad\left.\times\exp\left(\frac{3L(n+1)}{\left(n-p_0-L\right)\left(n-p_0\right)}\right)-1\right\}\\
&=P\left\{\frac{SSE_{p_0}-SSE_{p_0+L}}{SSE_{p_0+L}}>\frac{n-p_0}{n-p_0-L}\right.\\
&\quad\left.\times\exp\left(\frac{3L(n+1)}{\left(n-p_0-L\right)\left(n-p_0\right)}\right)-1\right\}.
\end{aligned}$$

Since $SSE_p \sim \sigma_0^2\chi^2_{T-p}$ and is independent of $SSE_p - SSE_{p+L}$ then $SSE_p - SSE_{p+L} \sim \sigma_0^2\chi^2_{2L}$. So

$$\begin{aligned}
&P\left(KIC_{p_0+L} < KIC_{p_0}\right)\\
&\quad = P\left\{\frac{\chi^2_{2L}}{\chi^2_{n-2p_0-2L}} > \frac{n-p_0}{n-p_0-L}\exp\left(\frac{3L(n+1)}{(n-p_0-L)\,(n-p_0)}\right) - 1\right\}\\
&\quad = P\left\{F_{2L,n-2p_0-2L} > \frac{n-2p_0-2L}{2L}\right.\\
&\qquad \left.\times\left(\frac{n-p_0}{n-p_0-L}\exp\left(\frac{3L(n+1)}{(n-p_0-L)\,(n-p_0)}\right) - 1\right)\right\}.
\end{aligned}$$

Now, following a similar method, we can deduce the probability of overfitting for KICc.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium Information Theory. Akademia Kiado, Bubapest, pp. 267–281.

Bedrick, E.J., Tsai, C.L., 1994. Model selection for multivariate regression in small samples. Biometrics 50, 226–231.

Bickel, P.J., Doksum, K.A., 1977. Mathematical Statistics: Basic Ideas and Selected Topics. Prentice-Hall, New Jersey.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis, Forecasting and Control. Holden-Day, San Franscisco. (revised version).

Cavanaugh, J.E., 1997. Unifying the derivations of the Akaike and corrected Akaike information criteria. Statist. Probab. Lett. 33, 201–208.

Cavanaugh, J.E., 1999. A large-sample model selection criterion based on Kullback's symmetric divergence. Statist. Probab. Lett. 44, 333–344.

Fujikoshi, Y., Satoh, K., 1997. Modified AIC and Cp in multivariate linear regression. Biometrika 84, 707–716.

Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregressive. J. Roy. Statist. Soc. B 41, 190–195.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Hurvich, C.M., Tsai, C.L., 1993. A corrected Akaike information criterion for vector autoregressive model selection. J. Time Ser. 14, 271–279.

Hurvich, C.M., Shumway, R.H., Tsai, C.L., 1990. Improved estimators of Kullback–Leibler information for autoregressive model selection in small samples. Biometrika 77, 709–719.

Kullback, S., 1968. Information Theory and Statistics. Dover, New York.

Linhart, H., Zucchini, W., 1986. Model Selection. Wiley, New York.

McQuarrie, A.D.R., Tsai, C.L., 1998. Regression and Time Series Model Selection. World Scientific Publishing, Singapore.

Muirhead, R.J., 1982. Aspects of Multivariate Statistical Theory. Wiley, New York.

Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, New York.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.