Available online at www.sciencedirect.com

SCIENCE DIRECT®

Statistics & Probability Letters 76 (2006) 1647–1654

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

A small-sample criterion based on Kullback's symmetric divergence for vector autoregressive modeling

Bezza Hafidi

Department of Mathematics, Faculty of Sciences Semlalia, Cadi-Ayyad University, P.B. 2390 Marrakech, Morocco

Received 10 April 2005; received in revised form 15 March 2006; accepted 3 April 2006
Available online 19 May 2006

Abstract

In this note, we propose a small-sample criterion KICc for selecting vector autoregressive models. KICc is an approximately unbiased estimator of the expected Kullback's symmetric divergence. A simulation study shows that KICc provides better model order choices than the KIC criterion in small samples.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Model selection; AIC; KIC; Kullback symmetric divergence

1. Introduction

The Kullback information criterion, KIC (Cavanaugh, 1999), is a variant of the Akaike information criterion, AIC (Akaike, 1973). It was designed as an asymptotically unbiased estimator of Kullback's symmetric divergence.

Although, both two criteria have proven to be widely applicable, it can have serious deficiencies especially for small-sample or when the number of fitted parameters is a moderate to large fraction of the sample size. In this setting, Hurvich and Tsai (1989) proposed a corrected Akaike information criterion, AICc, for linear and non-linear regression and autoregressive modeling. This corrected criterion has been extended in a large framework, including autoregressive moving average modeling (Hurvich et al., 1990), vector autoregressive modeling (Hurvich and Tsai, 1993), and multivariate regression modeling (Bedrick and Tsai, 1994). Recently, Cavanaugh (2004) proposed a corrected Kullback's information criterion (KICc) for linear model and Hafidi and Mkhadri (2006) proposed a corrected KICc for multiple and multivariate regression and univariate autoregressive (AR) modeling.

The vector AR model, is one of the most common and straightforward methods for modeling multivariate time series data. For vector AR models the potential for overfitting, and hence the potential usefulness of KICc, would seem to be even greater than in the univariate case. The reason is that the vector AR models contain many more unknown parameters than the corresponding univariate AR models. In this note, we derive and investigate a version of KICc for vector AR models selection.

E-mail address: b.hafidi@ucam.ac.ma.

0167-7152/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2006.04.007

In Section 2, we develop KICc for vector AR models. We present simulation results in Section 3. We end this note with a small conclusion in Section 4.

2. KICc for vector autoregressive models

Suppose that the candidate model for vector AR of order p, denoted AR(p), is given by

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + \varepsilon_t, \quad t = 1, \ldots, n, \tag{1}$$

where ε_t i.i.d. $\mathcal{N}(0, \Sigma)$, $Z_t = (z_{1t}, \ldots, z_{mt})$ is a $m \times 1$ observed vector at times $t = 1, \ldots, n$ and the Φ_j ($j = 1, \ldots, p$) are $m \times m$ matrices of unknown parameters.

Define the observation matrix $\mathbf{Y}$ as $\mathbf{Y} = (Z_1, \ldots, Z_n)^{\mathrm{t}}$, where "t" stands for the transpose. Thus, we can form a regression model from (1) by conditioning on the past and forming the design matrix, $\mathbf{X}$, where $\mathbf{X}$ is of full rank pm. The resulting regression model takes the form

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

where $\beta = (\phi_1, \ldots, \phi_p)^{\mathrm{t}}$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{t}}$ and the row of $\mathbf{X}$ have the following structure

$$X_t = (Z_{t-1}, \ldots, Z_{t-p})^{\mathrm{t}}.$$

The conditional least squares estimation of parameters are given as

$$\hat{\beta} = (\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{t}}\mathbf{Y} \quad \text{and} \quad \hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})^{\mathrm{t}}(\mathbf{Y} - \mathbf{X}\hat{\beta})/n.$$

The true model of vector AR(p_0) has the following form:

$$\mathbf{Y} = \mathbf{X}_0\Phi_0 + \varepsilon_0,$$

where $\mathbf{Y} = (Z_1, \ldots, Z_n)$, $\Phi_0 = (\phi_1, \ldots, \phi_{p_0})^{\mathrm{t}}$, $\varepsilon_0 = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{t}}$, where the ε_t for $t = 1, \ldots, n$ are independent identically distributed normal random variables with mean vector zero and matrix covariance Σ_0, and the row of $\mathbf{X}_0$ have the following structure:

$$X_{0t} = (Z_{t-1}, \ldots, Z_{t-p_0})^{\mathrm{t}}$$

for $t = 1, \ldots, n$. For simplicity, we assume that all pre-sample values are zero, i.e. $Z_0 = Z_{-1} = \cdots = 0$. The same assumption is made by Akaike (1969) and a more detailed discussion is given in Priestley (1981, p. 375). Also, we assume that the candidate class of models includes the true model. This strong assumption is also used in the derivation of KIC criterion (Cavanaugh, 1999) and AIC criterion (Linhart and Zucchini, 1986). Under this assumption the columns of $\mathbf{X}$ can be rearranged so that $\mathbf{X}_0\beta_0 = \mathbf{X}\beta^\star$, where $\beta^\star = (\beta_0^{\mathrm{t}}, 0^{\mathrm{t}})^{\mathrm{t}}$ and 0 is the $(p - p_0)m \times m$ matrix of zeros.

Let θ_0 the true parameter vector which is unknown and θ the parameter vector of the candidate model, so that $f(\mathbf{Y}|\theta_0)$ and $f(\mathbf{Y}|\theta)$ represent the generating and candidate densities for data, respectively. With the previous notations, the parameter vectors θ and θ_0 for the candidate and the true models are $(\beta, \Sigma)^{\mathrm{t}}$ and $(\beta_0, \Sigma_0)^{\mathrm{t}}$, respectively.

A measure of separation between the generating and a candidate model is given by the Kullback's symmetric divergence (Kullback, 1968), defined by

$$J(\theta_0, \theta) = \{D(\theta_0, \theta) - D(\theta_0, \theta_0)\} + \{D(\theta, \theta_0) - D(\theta, \theta)\}, \tag{2}$$

where $D(\theta_0, \theta) = \mathbf{E}_{\theta_0}\{-2\ln f(\mathbf{Y}|\theta)\}$ and $\mathbf{E}_{\theta_0}$ denotes the expectation under $f(\mathbf{Y}|\theta_0)$. Since $D(\theta_0, \theta_0)$ does not depend on θ, for the purpose of discriminating among various models we propose

$$K(\theta_0, \theta) = D(\theta_0, \theta) + \{D(\theta, \theta_0) - D(\theta, \theta)\}, \tag{3}$$

as a substitute for $J(\theta_0, \theta)$. However, Cavanaugh (1999) showed that the criterion defined by

$$\mathrm{KIC} = n(\ln|\hat{\Sigma}| + m) + 3d \tag{4}$$

is an asymptotically unbiased estimator of

$$\Omega(d,\theta_0) = \mathbf{E}_{\theta_0}\{K(\theta_0,\hat{\theta})\} \tag{5}$$

$$= \mathbf{E}_{\theta_0}\{D(\theta_0,\hat{\theta})\} + \mathbf{E}_{\theta_0}\{D(\hat{\theta},\theta_0) - D(\hat{\theta},\hat{\theta})\}, \tag{6}$$

where $d = pm^2 + 0.5m(m+1)$ is a number of unknown parameter.

Now, Hurvich and Tsai (1993) established that the criterion defined by

$$\mathrm{AIC}c = n(\ln|\hat{\Sigma}| + m) + \frac{2dn}{n-pm-p-1} \tag{7}$$

is an approximately unbiased estimator of the first term of (6). Then, using the last result, we obtain an asymptotically unbiased estimator of the Kullback symmetric divergence defined in the following proposition.

Proposition. *The criterion defined by*

$$\mathrm{KIC}c = n(\ln|\hat{\Sigma}| + m) + \frac{d(3n-pm-p-1)}{n-pm-p-1},$$

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

Proof. The log-likelihood for the candidate model, ignoring the constant terms, is given by

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

where "tr" stands for trace. The expectation of $K(\theta_0,\hat{\theta})$ has the form

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

Now, we will compute the second term of this equation. Indeed, we have

$$\begin{aligned} D(\hat{\theta},\theta_0) &= \mathbf{E}_{\theta}\{n\ln|\Sigma_0| + \mathrm{tr}\{(\mathbf{Y}-\mathbf{X}\beta_0)\Sigma_0^{-1}(\mathbf{Y}-\mathbf{X}\beta_0)^{\mathrm{t}}\}\}|_{\theta=\hat{\theta}} \\ &= n\ln|\Sigma_0| + \mathbf{E}_{\theta}\{\mathrm{tr}\{(\mathbf{Y}-\mathbf{X}\beta)\Sigma_0^{-1}(\mathbf{Y}-\mathbf{X}\beta)^{\mathrm{t}} + \Sigma_0^{-1}(\beta-\beta_0)^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}}\mathbf{X})(\beta-\beta_0)\}\}|_{\theta=\hat{\theta}} \\ &= n\ln|\Sigma_0| + n\,\mathrm{tr}(\Sigma_0^{-1}\hat{\Sigma}) + \mathrm{tr}\{\Sigma_0^{-1}(\hat{\beta}-\beta_0)^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}}\mathbf{X})(\hat{\beta}-\beta_0)\}. \end{aligned} \tag{8}$$

Similarly, we have

$$\begin{aligned} D(\hat{\theta},\hat{\theta}) &= \mathbf{E}_{\theta}\{n\ln|\Sigma| + \mathrm{tr}\{(\mathbf{Y}-\mathbf{X}\beta)\Sigma^{-1}(\mathbf{Y}-\mathbf{X}\beta)^{\mathrm{t}}\}\}|_{\theta=\hat{\theta}} \\ &= n\ln|\hat{\Sigma}| + nm \end{aligned} \tag{9}$$

The difference between (8) and (9) leads to

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

It is well known that $n\hat{\Sigma} \sim \mathrm{Wishart}_p(n-pm,\Sigma_0)$, then $\mathbf{E}_{\theta_0}(n\hat{\Sigma}) = (n-pm)\Sigma_0$ (cf. Muirhead, 1982, p. 137) and

$$\begin{aligned} \mathbf{E}_{\theta_0}(n\,\mathrm{tr}(\Sigma_0^{-1}\hat{\Sigma})) &= \mathrm{tr}\{\mathbf{E}_{\theta_0}(n\Sigma_0^{-1}\hat{\Sigma})\} \\ &= (n-pm)m. \end{aligned}$$

Furthermore, $\mathrm{vec}(\hat{\beta}) \sim \mathcal{N}_{pm^2}(\mathrm{vec}(\beta), \Sigma_0 \otimes (\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1})$, where $\mathrm{vec}(\beta)$ is the $pm^2 \times 1$ vector obtained by stacking the columns of β and $\otimes$ is the Kronecker product. It yields

$$\begin{aligned} \mathbf{E}_{\theta_0}\{\mathrm{tr}\{\Sigma_0^{-1}(\hat{\beta}-\beta_0)^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}}\mathbf{X})(\hat{\beta}-\beta_0)\}\} &= \mathrm{tr}\{\mathbf{E}_{\theta_0}\{\Sigma_0^{-1}(\hat{\beta}-\beta_0)^{\mathrm{t}}(\mathbf{X}^{\mathrm{t}}\mathbf{X})(\hat{\beta}-\beta_0)\}\} \\ &= \mathbf{E}_{\theta_0}\{\mathrm{vec}(\hat{\beta}-\beta_0)^{\mathrm{t}}\{\Sigma_0^{-1} \otimes (\mathbf{X}^{\mathrm{t}}\mathbf{X})\}\mathrm{vec}(\hat{\beta}-\beta_0)\} \\ &= pm^2 \end{aligned}$$

which leads to

$$\mathbf{E}_{\theta_0}\{D(\hat{\theta},\theta_0)-D(\hat{\theta},\hat{\theta})\} = -\mathbf{E}_{\theta_0}\left\{n\ln\frac{|\hat{\Sigma}|}{|\Sigma_0|}\right\}. \tag{10}$$

Moreover, we have

$$n\ln\frac{|n\hat{\Sigma}|}{|\Sigma_0|} = nm\ln n + n\ln\frac{|\hat{\Sigma}|}{|\Sigma_0|}. \tag{11}$$

From Muirhead (1982, p. 100), it is well known that $|n\hat{\Sigma}|/|\Sigma_0|$ has the same distribution as $\prod_{i=1}^{m}\chi^2_{(n-pm-i+1)}$, where the $\chi^2_{(n-pm-i+1)}$ denotes the independent χ^2 random variables with $(n-pm-i+1)$ degrees of freedom. Then

$$\mathbf{E}_{\theta_0}\left\{n\ln\frac{|n\hat{\Sigma}|}{|\Sigma_0|}\right\} = \mathbf{E}_{\theta_0}\left\{n\ln\prod_{i=1}^{m}\chi^2_{(n-pm-i+1)}\right\} = \mathbf{E}_{\theta_0}\left\{n\sum_{i=1}^{m}\ln\chi^2_{(n-pm-i+1)}\right\}. \tag{12}$$

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)

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

Substituting this relation into (12) yields

$$\mathbf{E}_{\theta_0}\left\{n\sum_{i=1}^{m}\ln\chi^2_{(n-pm-i+1)}\right\} = n\sum_{i=1}^{m}\left\{\ln(n-pm-i+1)-\frac{1}{n-pm-i+1} + \mathrm{O}\left(\frac{1}{(n-pm-i+1)^2}\right)\right\}. \tag{13}$$

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

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

Substituting this equation into (13) yields

$$\mathbf{E}_{\theta_0}\left\{n\sum_{i=1}^{m}\ln\chi^2_{(n-pm-i+1)}\right\} = nm\ln n + \sum_{i=1}^{m}(1-pm-i) - \sum_{i=1}^{m}\frac{n}{n-pm-i+1} \tag{14}$$

$$+\sum_{i=1}^{m}\left\{\mathrm{O}\left(\frac{n}{(n-pm-i+1)^2}\right)+\mathrm{O}\left(\frac{(1-pm-i)^2}{n}\right)\right\}. \tag{15}$$

As $n\to\infty$ and when pm is held constant, each of the two terms in (15) for each i is asymptotically o(1). However, each of these terms is also o(1) when pm 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 (14) converge to m, and since $\sum_{i=1}^{m}(1-pm-i) = m-(pm^2+0.5m(m+1))$ we establish that

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

Substituting this equation into (12), and using (11) and (10) we get

$$\mathbf{E}_{\theta_0}\{D(\hat{\theta},\theta_0)-D(\hat{\theta},\hat{\theta})\} = (pm^2+0.5m(m+1)) + \mathrm{o}(1). \tag{16}$$

According to (7), and substituting (16) in the second term of (6), we obtain the estimator of $\Omega_{\mathbf{Y}}(d, \theta_0)$ given in the proposition. $\square$

3. Monte Carlo results

In this section, we evaluate the performance of our criterion against KIC, AIC, AICc, SIC (Schwarz, 1978) and HQ (Hannan and Quinn, 1979) criteria in a small-sample for the selection of bivariate vector AR models. We present some examples of simulation in which the data are generated by different manners, from each of two models: a vector AR(1) and a vector AR(2). One thousand sample of size $n = 20, 30$ and 40 are generated from model(1), with zero mean, $m = 2$ and the maximum order $p = 6$ and $p = 8$.

In the first example, the data are generated using a vector AR(1) with

$$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Phi_1 = \begin{pmatrix} -1.0 & 0.96 \\ -1.5 & 1.4 \end{pmatrix}.$$

The AR(2) model, used in the second example, is based on

$$\Sigma_0 = \begin{pmatrix} 1 & -0.08 \\ -0.08 & 1 \end{pmatrix}, \quad \Phi_1 = \begin{pmatrix} -1.0 & 0.96 \\ -1.5 & 1.4 \end{pmatrix}, \quad \Phi_2 = \begin{pmatrix} -0.5 & 0.3 \\ 0.0 & -0.4 \end{pmatrix}.$$

In the third example, the components of AR (2) are fixed as

$$\Sigma_0 = \begin{pmatrix} 2 & 1.4 \\ 1.4 & 1 \end{pmatrix}, \quad \Phi_1 = \begin{pmatrix} -0.2 & 0.1 \\ 0.5 & 0.2 \end{pmatrix}, \quad \Phi_2 = \begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix}.$$

Each matrix $\mathbf{Y}$ of the 1000 samples was generated by starting at y_{-50} with $y_t = 0$ for all $t < -50$. But only observations $y_1, \ldots, y_n$ were kept.

Tables 1–3 give the frequency of the orders selected by each of the criteria AIC, AICc, KIC, KICc, SIC and HQ, for the AR(1) and AR(2) models, with

$$\mathrm{SIC} = n \ln|\hat{\Sigma}| + d \ln(n) \quad \text{and} \quad \mathrm{HQ} = n \ln|\hat{\Sigma}| + 2d \ln(\ln(n)).$$

We see from Table 1 that for $n = 20$, the KICc criterion performs well 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. When the sample size increase, all the criteria perform well, but KICc and SIC provide the best selection of p. The tendency to overfit the true model for KIC, AIC, SIC and HQ decreases as n increases here.

In the second example (Table 2), we consider $p = 8$. For $n = 20$, KICc and AICc perform better than all other criteria, which exhibit a strong tendency to overfit the true model. On the other hand, when n increases, the tendency to overfit for all other criteria decreases, but KICc provides the best selection of p.

Other examples of simulation where $p = 10, 12$, not reported here, give the same results as in the second example. In the third example (Table 3), SIC is the best, followed by KICc and AICc which perform similarly at selecting the correct model order, although KICc exhibits substantially less underfitting than AICc.

Table 1
Selected dimensions for vector autoregressive model AR(1), example 1

Selected dimensions																		
	$n = 20$						$n = 30$						$n = 40$					
p	KICc	KIC	AIC	AICc	SIC	HQ	KICc	KIC	AIC	AICc	SIC	HQ	KICc	KIC	AIC	AICc	SIC	HQ
1	998	937	621	991	937	715	992	964	794	968	981	905	990	976	829	953	991	945
2	2	24	85	9	24	70	7	23	87	25	10	56	9	21	89	41	8	43
3	0	10	32	0	10	27	1	8	34	6	5	17	1	3	32	6	1	9
4	0	3	30	0	3	22	0	3	28	1	2	11	0	0	23	0	0	2
5	0	8	57	0	8	46	0	0	30	0	0	6	0	0	12	0	0	1
6	0	18	175	0	18	120	0	2	27	0	2	5	0	0	15	0	0	0

Table 2
Selected dimensions for vector autoregressive model AR(2), example 2

Selected dimensions																		
	$n=20$						$n=30$						$n=40$					
p	KICc	KIC	AIC	AICc	SIC	HQ	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	0	0	0	0	0	0
2	998	595	159	995	593	223	997	930	637	979	966	823	993	965	763	971	989	926
3	2	30	24	5	30	26	3	41	96	19	24	73	7	24	79	23	10	47
4	0	13	16	0	13	15	0	13	45	1	5	32	0	9	48	6	1	15
5	0	8	19	0	8	18	0	3	41	1	2	15	0	2	30	0	0	5
6	0	24	25	0	24	27	0	1	31	0	0	13	0	0	30	0	0	2
7	0	37	87	0	36	83	0	0	52	0	0	12	0	0	30	0	0	2
8	0	293	670	0	296	608	0	12	98	0	3	32	0	0	28	0	0	3

Table 3
Selected dimensions for vector autoregressive model AR(2), example 3

Selected dimensions						
	AR(2), $n=40$					
p	KICc	KIC	AIC	AICc	SIC	HQ
1	32	18	2	13	28	7
2	964	938	762	964	985	906
3	4	32	112	22	13	54
4	0	7	54	0	0	15
5	0	4	32	1	1	12
6	0	1	38	0	0	6

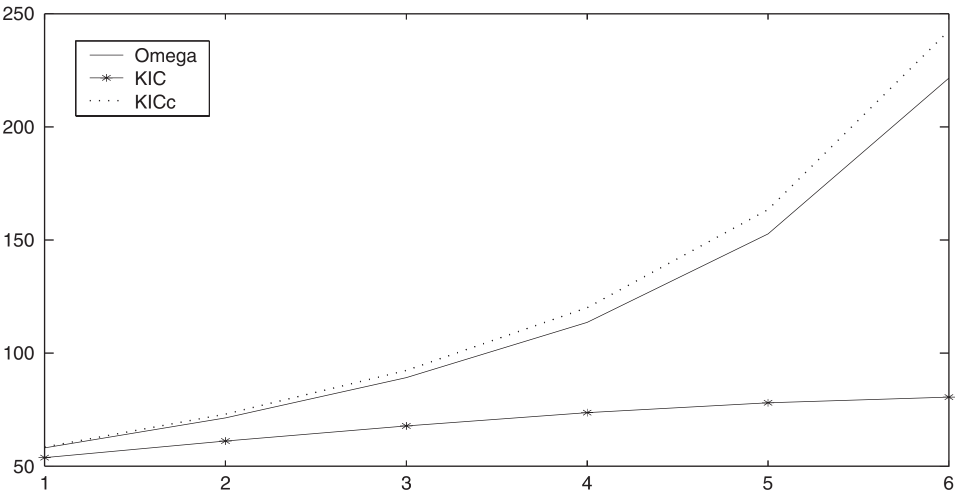

Fig. 1. Simulated $\Omega(d,\theta_0)$ and average KIC and KICc curves of a bivariate AR(1).

In both AR(1) and AR(2) and for small sample sizes, KICc performs well than KIC criterion which has a tendency to overfit greater than KICc.

Figs. 1 and 2 give some insights on KICc's success relative to KIC. We plotted the average values of KIC, KICc and $\Omega(d,\theta_0)$ as function of p for AR(1) with $n=20$ and AR(2) (Example 3).

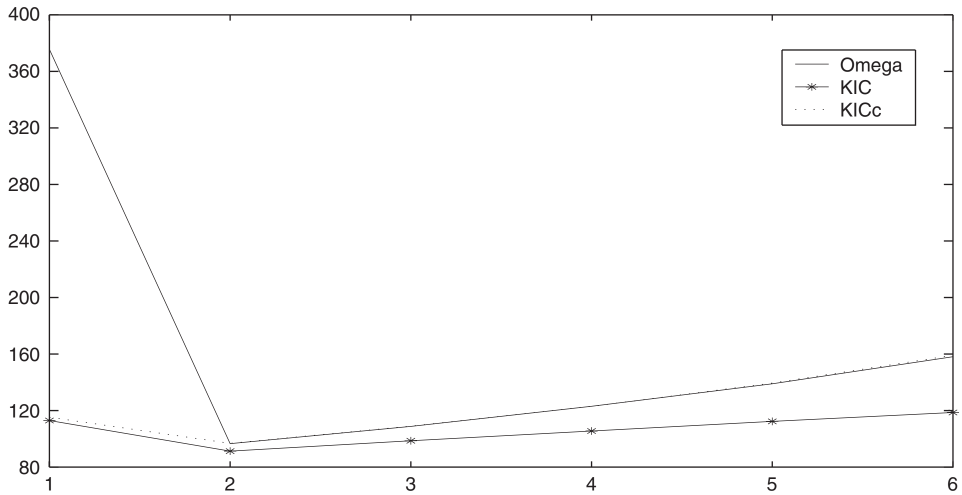

Fig. 2. Simulated $\Omega(d, \theta_0)$ and average KIC and KICc curves of a bivariate AR(2).

For the AR(1) (Fig. 1), we see that KICc exhibits less bias overall. While KIC exhibits a negative bias for all model orders, and this bias becomes more severe as the model order increases.

For the AR(2) (Fig. 2), we state that the curve of the average values of KICc, as well as Ω, have a clearly defined minimum at the true order $p = 2$, and closely follows the simulated $\Omega(d, \theta_0)$ curve. The average value of KIC is also minimized at $p = 2$, but the minimum is not sharply defined.

4. Conclusion

In this note, we have derived and investigated the KICc criterion, based on Kullback's symmetric divergence, for model selection in vector autoregressive models. Moreover, a small simulation study is undertaken to compare the performance of our criterion to other well known criteria. Our simulations indicate 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 is the best performed. Whereas, KICc was slightly outperformed by the consistent criterion SIC. Unlike multivariate regression, parameters counts increase more rapidly in vector autoregressive models. Indeed, in vector autoregressive models of order p (AR(p)), the parameters increase by m^2 when the order increases by 1, potentially a very large number. In multivariate regression with rank($\mathbf{X}$) $= p$, we had $n - p$ degrees of freedom. Whereas in a vector AR model of the same order, we are reduced to $n - p - pm$ degrees of freedom. This means that the probabilities of overfitting are much smaller for vector AR.

Acknowledgements

The author wish to thank the referee for comments which improved the presentation of this paper. I also thank A. Mkhadri for reading the paper and for his valuable suggestions.

References

Akaike, H., 1969. Fitting autoregressive models for prediction. Ann. Inst. Statist. Math. 21, 243–247.

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, NJ.

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

Cavanaugh, J.E., 2004. Criteria for linear model selection based on Kullback's symmetric divergence. Austral. New Zealand J. Statist. 46, 257–274.

Hafidi, B., Mkhadri, A., 2006. A corrected Akaike criterion based on Kullback's symmetric divergence: Applications in time series, multiple and multivariate regression. Comput. Statist. Data Anal. 50, 1524–1550.

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 Kulback-Leibler information for autoregressive model selection in small samples. Biometrika 77, 709–719.

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