A CRITERION FOR VECTOR AUTOREGRESSIVE MODEL SELECTION BASED ON KULLBACK'S SYMMETRIC DIVERGENCE

Abd-Krim Seghouane

National ICT Australia Limited,
Locked Bag 8001, Canberra ACT 2601, Australia.
E-mail: Abd-krim.seghouane@nicta.com.au
Phone: +61 2 6125 8621, Fax: +61 2 6125 8660

ABSTRACT

The Kullback Information Criterion, KIC and its univariate bias-corrected version, KIC_c are two recently developed criteria for model selection. In this paper, a small sample model selection criterion for vector autoregressive models is developed. The proposed criterion is named KIC_{vc}, where the notation "vc" stands for vector correction, and it can be considered as an extension of KIC for vector autoregressive models. KIC_{vc} is an unbiased estimator of a variant of the Kullback symmetric divergence, assuming that the true model is correctly specified or overfitted. Simulation results shows that the proposed criterion estimates the model order more accurately than any other asymptotically efficient method when applied to vector autoregressive model selection in small samples.

1. INTRODUCTION

Model selection is an important step in vector autoregressive modelling. Such selection is often facilitated by the use of a model selection criterion. In this paper, the problem of vector autoregressive (VAR) model selection when the sample size is small is considered. The VAR model is probably one of the most common and straightforward methods for modelling multivariate time series data. Another typical signal processing application of VAR models is clutter suppression in airborne radar signal processing [1].

A model selection criterion can be designed to estimate an expected overall discrepancy, a quantity which reflects the degree of similarity between a fitted approximating model and the generating or "true" model. Estimation of Kullback's information [3] is the key to deriving the Akaike Information criterion, AIC [2]. Whereas, from the estimation of Kullback's symmetric divergence [4] follows the Kullback Information Criterion, KIC [5].

National ICT Australia is funded by the Australian Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Centre of Excellence Program.

KIC serves as an *asymptotically* unbiased estimator of a variant of the Kullback's symmetric divergence between the generating model and the fitted approximating model under the assumption that the true model belongs to the class of candidate models. As the dimension of the candidate model, k, increases compared to the sample size, n, KIC becomes a strongly negatively biased estimate of the variant of the Kullback symmetric divergence and leads to the choice of over parameterized models. A bias corrected version of KIC, denoted KIC_c has been recently proposed for linear regression and univariate autoregressive models [6].

The correction of KIC proposed in [6] is not appropriate for VAR models. In this case, KIC_c will produce a biased estimate of the Kullback symmetric divergence and will lead to underfitting due to the choice of an under parameterized model. The reason is that the VAR models contain many more unknown parameters than the corresponding univariate AR models. In this paper, a new information criterion, denoted KIC_{vc}, is proposed as a bias correction of KIC for VAR models.

The remainder of this paper is organized as follows. In section II, we briefly review KIC and its corrected version KIC_c. Section III, is devoted to a development of the new proposed criterion, KIC_{vc} for VAR model selection. A numerical example of comparison is given in section IV and a conclusion is given in section V. A brief theoretical justification of the proposed criterion is also presented.

2. A BRIEF REVIEW OF THE DERIVATION OF KIC AND KIC_C

Suppose a collection of data $\mathbf{y}_n = (y_1, \ldots, y_n)^\top$ has been generated according to an unknown parametric model $p(\mathbf{y}|\theta_0)$. We try to find a parametric model which provides a suitable approximation for $p(\mathbf{y}|\theta_0)$.

Let $\mathcal{M}_k = \{p(\mathbf{y}|\theta_k)|\theta_k \in \Theta_k\}$ denote a k-dimensional parametric family and let $\hat{\theta}_k$ denote the vector of parameters estimate obtained by maximizing the likelihood func-

tion $p(\mathbf{y}_n|\theta_k)$ over Θ_k. For simplicity, we will assume $k = 1, 2, \ldots, k_{max}$, so the collection $\mathcal{M}_k$'s consists of nested families, i.e, $\Theta_1 \subset \Theta_2 \subset \ldots \subset \Theta_{k_{max}}$ of dimension 1 through (k_{max}) [2]. It is also assumed that the search is carried out in a parametric family of distribution including the true model, i.e, $\theta_0 \in \Theta_{k_{max}}$.

To determine which candidate model best approximates the generating unknown model $p(\mathbf{y}|\theta_0)$, we need a measure which provides a suitable reflection of the disparity between $p(\mathbf{y}|\theta_0)$ and an approximating model $p(\mathbf{y}|\hat{\theta}_k)$. The Kullback's symmetric divergence is one of such measure.

Kullback's symmetric divergence between two parametric densities $p(\mathbf{y}|\theta_k)$ and $p(\mathbf{y}|\theta_0)$ is defined as

$$\begin{aligned} 2J_n(\theta_0,\theta_k) &= 2I_n(\theta_0,\theta_k) + 2I_n(\theta_k,\theta_0) \\ &= E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_k)\} - E_{\theta_0}\{-2\ln p(\mathbf{y}|\theta_0)\} \\ &+ E_{\theta_k}\{-2\ln p(\mathbf{y}|\theta_0)\} - E_{\theta_k}\{-2\ln p(\mathbf{y}|\theta_k)\} \\ &= d_n(\theta_0,\theta_k) - d_n(\theta_0,\theta_0) + d_n(\theta_k,\theta_0) \\ &- d_n(\theta_k,\theta_k), \end{aligned}$$

where $I_n(\theta_i,\theta_j)$ is the directed Kullback divergence and $E_{\theta_j}\{.\}$ represents the expectation with respect to the density $p(\mathbf{y}|\theta_j)$. Since $d_n(\theta_0,\theta_0)$ does not depend on θ_k, any ranking of the candidate models according to $2J_n(\theta_0,\theta_k)$ would be identical to ranking them according to $K_n(\theta_0,\theta_k)$ defined by

$$K_n(\theta_0,\theta_k) = d_n(\theta_0,\theta_k) + d_n(\theta_k,\theta_0) - d_n(\theta_k,\theta_k). \quad (1)$$

Therefore, $K_n(\theta_0,\theta_k)$ would provide a suitable measure of a variant of Kullback's symmetric divergence between the generating model $p(\mathbf{y}|\theta_0)$ and the candidate model $p(\mathbf{y}|\hat{\theta}_k)$. Yet evaluating $K_n(\theta_0,\hat{\theta}_k)$ is not possible, since doing so requires the knowledge of θ_0.

Cavanaugh argues that $-2\ln p(\mathbf{y}_n|\hat{\theta}_k)$ is a biased estimator of $K_n(\theta_0,\hat{\theta}_k)$ and proposes an *asymptotic* bias correction [5] leading to

$$KIC = -2\ln p(\mathbf{y}_n|\hat{\theta}_k) + 3k. \quad (2)$$

If we write

$$\Omega_n(k,\theta_0) = E_{\theta_0}\left\{K_n(\theta_0,\hat{\theta}_k)\right\}, \quad (3)$$

one can establish that [5] [6]

$$\Omega_n(k,\theta_0) = E_{\theta_0}\{KIC\} + o(1).$$

Motivated by the fact that KIC is only asymptotically unbiased, a bias corrected version has been proposed in [6]

$$KIC_c \simeq -2\ln p(\mathbf{y}|\hat{\theta}_k) + \frac{(k+1)(3n-k-2)}{n-k-2} + \frac{k}{n-k} \quad (4)$$

that is exactly unbiased estimator of $K_n(\theta_0,\hat{\theta}_k)$ in the context of linear regression, i.e,

$$\Omega_n(k,\theta_0) = E_{\theta_0}\{KIC_c\}.$$

KIC is shown to outperform AIC in large sample linear regression and univariate autoregressive model selection and leads less frequently to overfitting than AIC and its corrected variant, AIC_c [2][7]. KIC_c is found to provide a better model order choice than KIC for small sample linear regression and univariate autoregressive model selection [6]. An extension of KIC for model selection in the presence of incomplete data has also been developed in [8].

3. DERIVATION OF KIC_{VC}

Suppose that the generating model of the data $Z_1, ..., Z_n$ is an m-dimensional $VAR(k_0)$ process with zero mean

$$Z'_t = \sum_{j=1}^{k_0} Z'_{t-j}\phi'_{0j} + \varepsilon'_{0t} \quad t = 1, ..., n, \quad (5)$$

where $Z'_t = (z_{1t}, ..., z_{mt})$, ϕ'_{0j}, $j = 1, ..., k_0$ are $m \times m$ coefficient matrices and ε_{0t} are i.i.d normal random variables with mean zero and $m \times m$ variance-covariance matrix Σ_0. The k^{th} order approximating (or candidate) $VAR(k)$ model for the data $Z_1, ..., Z_n$ is

$$Z'_t = \sum_{j=1}^{k} Z'_{t-j}\phi'_j + \varepsilon'_t \quad t = 1, ..., n, \quad (6)$$

and the ε_t are i.i.d normal random variables with mean vector zero and $m \times m$ variance-covariance matrix Σ. The notations and assumptions adopted in this paper are similar to the ones of [9].

Lemma. Let

$$\begin{aligned} KIC_{vc} &= -2\ln p(\mathbf{y}|\hat{\theta}_k) + \frac{nm(2km+m+1)}{n-km-m-1} \\ &- \sum_{i=1}^{m} n\psi\left(\frac{n-m(k+1)+i}{2}\right) + nm\ln\left(\frac{n}{2}\right), \end{aligned} \quad (7)$$

where ψ is the Euler's psi function [10]. Then, under the model (6), KIC_{vc} is an exactly unbiased estimator of $K_n(\beta_0,\hat{\beta}_k)$

$$\Omega_n(k,\beta_0) = E_{\theta_0}\{KIC_{vc}\}.$$

Proof.

Due to the lack of space, only the most importants lines of proof are given here. For more detail contact the author. The definition of the notations used in the proof and that are not defined above can be found in [9].

The expression of the log-likelihood $-2\ln p(\mathbf{y}|\theta_k)$ gives

$$\begin{aligned} d_n(\theta_i,\theta_j) &= E_{\theta_i}\{-2\ln p(\mathbf{y}|\theta_j)\} \\ &= nm\ln 2\pi + n\ln(|\Sigma_j|) + n\text{tr}(\Sigma_j^{-1}\Sigma_i) \\ &+ \text{tr}\{\Sigma_j^{-1}(\beta_i-\beta_j)'E_{\theta_i}(X'X)(\beta_i-\beta_j)\}. \end{aligned}$$

where X and β are defined in [9]. Using this expression in (1) leads to

$$\begin{aligned} K_n(\theta_0,\hat{\theta}_k) &= n\ln(|\hat{\Sigma}_k|)+nm\ln(2\pi)-n\ln\left(\frac{|\hat{\Sigma}_k|}{|\Sigma_0|}\right) \\ &+ \text{tr}\{\Sigma_0^{-1}(\beta_0-\hat{\beta}_k)'E_{\theta_k}(X'X)(\beta_0-\hat{\beta}_k)\} \\ &+ \text{tr}\{\hat{\Sigma}_k^{-1}(\beta_0-\hat{\beta}_k)'E_{\theta_0}(X'X)(\beta_0-\hat{\beta}_k)\} \\ &+ n\text{tr}(\hat{\Sigma}_k^{-1}\Sigma_0)+n\text{tr}(\Sigma_0^{-1}\hat{\Sigma}_k)-nm \qquad (8) \end{aligned}$$

where $\hat{\beta}_k$ is the conditional least squares parameter estimate of β_0 and $\hat{\Sigma}_k$ is the maximum likelihood estimate of Σ_0. From results of [11] on $\hat{\beta}_k$ and $n\hat{\Sigma}_k$, pages 353-354, we have

$$E_{\theta_0}\{\text{tr}(\hat{\Sigma}_0^{-1}\Sigma_k)\}\approx\frac{(n-km)}{n}m,$$

and from [12], page 270 (Lemma 7.7.1), we have

$$E_{\theta_0}\{\text{tr}(\hat{\Sigma}_k^{-1}\Sigma_0)\}\approx bm,\quad b=n/(n-km-m-1).$$

Also, from results of [11][12], we have

$$E_{\theta_0}\left\{\text{tr}\{\hat{\Sigma}_k^{-1}(\beta_0-\hat{\beta}_k)'E_{\theta_0}(X'X)(\beta_0-\hat{\beta}_k)\}\right\}\approx bkm^2,$$

and

$$E_{\theta_0}\left\{\text{tr}\{\Sigma_0^{-1}(\beta_0-\hat{\beta}_k)'E_{\theta_k}(X'X)(\beta_0-\hat{\beta}_k)\}\right\}\approx km^2.$$

By deducing from results of [13], page 100 (Theorem 3.2.15) and from [10], page 373, we have

$$E_{\theta_0}\{\ln|n\hat{\Sigma}_k|\}=\ln|\Sigma_0|+m\ln(2)+\sum_{i=1}^{m}\psi\left(\frac{n-m(k+1)+i}{2}\right).$$

Using the above results it is straightforward to establish the result of the lemma.
For easy computation it is possible to use the following approximation.
Corollary. Let

$$\begin{aligned} KIC_{vc} &= -2\ln p(\mathbf{y}|\hat{\theta}_k)+\frac{nm(2km+m+1)}{n-km-m-1} \\ &+\frac{nm}{n-mk-(m-1)/2}+\frac{2m^2k+m^2-m}{2}, \qquad (9) \end{aligned}$$

then, under the model (6),

$$\Omega_n(k,\beta_0)\simeq E_{\theta_0}\left\{KIC_{vc}\right\}.$$

Proof.
By putting $\alpha=(k+1)m-i$ and using the approximation result of $\psi\left(\frac{n-\alpha}{2}\right)$ developed in [6], we have

$$\begin{aligned} \psi\left(\frac{n-m(k+1)+i}{2}\right) &= \ln\left(\frac{n}{2}\right)-\frac{(k+1)m-i}{n} \\ &- \frac{1}{n-(k+1)m+i}. \end{aligned}$$

The sum of the second term of the right hand side is

$$\sum_{i=1}^{m}\frac{(k+1)m-i}{n}=\frac{2m^2k+m^2-m}{2n}.$$

From the following equality

$$\sum_{i=1}^{m}n-m(k+1)+i=m(n-mk-(m-1)/2),$$

and assuming that $n-m(p+1)$ is much larger that m, we have,

$$\sum_{i=1}^{m}\frac{1}{n-m(k+1)+i}\simeq\frac{m}{n-mk-(m-1)/2}.$$

Replacing these expressions in the equation (7) of the proposed criterion gives the easy computing approximation criterion of the corollary.
It is worth to mention that asymptotically KIC_{vc} converges to KIC and $KIC_{vc}(m=1)=KIC_c$. This motivates the use of KIC_{vc} for small sample applications.

4. SIMULATION RESULTS

Here, the results of a simulation study on the small sample performance of several criteria for the selection of bivariate AR model are presented. A thousand simulated realizations of size $n=50$ were generated from a VAR(2) ($p_0=2$) model. The model is bivariate with zero mean, as given by (6) with $m=2$. The VAR(2) model has

$$\Sigma_0=\begin{bmatrix}1 & -0.08\\ -0.08 & 1\end{bmatrix},\quad \phi_1=\begin{bmatrix}0.50 & -0.30\\ 0.20 & 0.65\end{bmatrix},$$

$$\phi_2=\begin{bmatrix}-0.50 & 0.30\\ 0 & -0.40\end{bmatrix}.$$

For each realization, the stepwise least squares algorithm [14] was used to fit candidate VAR model of orders 1 to 8 and various criteria are used to select from among the candidate models. The other criteria considered in this simulation are AIC, AIC_c, KIC, KIC_c and BIC [15]. Table 1 gives the frequency of model orders selected by the different criteria. It is clear that KIC_{vc} performs best, closely followed by BIC and KIC_c, while other criteria perform less effectively. This improved selection is due to its finite sample bias correction.
Figure 1 provides some insight as why KIC_{vc} tends to outperform KIC and KIC_c as a selection criterion. Simulated value of $E_{\theta_0}\{KIC\}, E_{\theta_0}\{KIC_c\}$ and of $E_{\theta_0}\{KIC_{vc}\}$ as given by equation (9) are obtained by averaging over the 1000 realizations. $\Omega_n(k,\beta_0)$ is obtained by averaging the exact expression of KIC_{vc} given by equation (7) using the

digamma function. These averages values are plotted versus k. Since KIC, KIC_c, KIC_{vc} and $\Omega_n(k, \beta_0)$ are obtained by adding a non stochastic penalty term to the log likelihood, the three criteria have the same variance. This is why in comparison only the mean values are emphasize. It can be noted that KIC and KIC_c are biased estimators of $\Omega_n(k, \beta_0)$ specially when k increases. This bias is the major factor for the bad performance of KIC and KIC_c compared with KIC_{vc}.

Table 1. Frequency of the model order estimated by each criterion for 1000 realizations of sample size N=50.

N	order	AIC	AICc	KIC	KIC_c	KIC_{vc}	BIC
50	$< p_0$	1	2	12	44	31	27
50	$= p_0$	748	941	946	955	**962**	959
50	$> p_0$	251	57	42	1	7	14

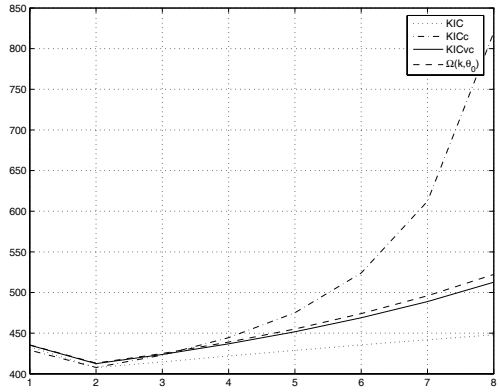

Fig. 1. Averages of KIC, KIC_c, KIC_{vc} and $\Omega_n(k, \beta_0)$ for $N = 50$ and the true model order $p_0 = 2$.

5. CONCLUSION

KIC_{vc} outperforms classical criteria in small sample VAR model selection. As a result, KIC_{vc} serves as an effective tool for selecting a vector autoregressive model of appropriate order when the sample size is small. The bias of KIC_{vc}, in comparison to KIC is reduced leading to improved order selection as shown by a simulation example.

6. REFERENCES

[1] J. Li, G. Liu and P. Stoica, "Moving target feature extraction for airborne high-range resolution phased array radar," *IEEE Transactions on Signal Processing*, Vol. 49, pp. 277-289, 2001.

[2] H. Akaike, "A new look at the model identification," *IEEE Transactions on Automatic and Control*, Vol. 19, pp. 716-723, 1974.

[3] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematics Statistics,* Vol. 22, pp. 76-86, 1951.

[4] H. Jeffreys, "An invariant form of the prior probability in estimation problems," *The Royal Statistical Society*, Vol. 186, pp. 453-469, 1946.

[5] J. E. Cavanaugh, "A large-sample model selection criterion based on Kullback's symmetric divergence," *Statistics and Probability Letters*, Vol. 42, pp. 333-343, 1999.

[6] A. K. Seghouane and M. Bekara, "A small sample model selection criterion based on Kullback's symmetric divergence," *IEEE Transactions on Signal Processing*, Vol. 52, pp. 3314-3323, 2004.

[7] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, Vol. 76, pp. 297-307, 1989.

[8] A. K. Seghouane, M. Bekara and G. Fleury, "A criterion for model selection in the presence of incomplete data based on Kullback's symmetric divergence," *Signal Processing*, In press.

[9] C. M. Hurvich and C. L. Tsai, "A corrected Akaike information criterion for vector autoregressive model selection," *Journal of Time Series Analysis*, Vol. 14, pp. 271-279, 1993.

[10] S. Kotz, N. L. Johnson and C. B. Read, eds. *Encyclopedia of Statistical Sciences*, Vol. 2, New York, Wiley Series, 1982.

[11] W. S. Wei, *Time Series Analysis*, New York, Addison-Wesley, 1989.

[12] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Second Edition, New York, Wiley Series, 1984.

[13] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, New York, Wiley Series, 1982.

[14] A. Neumaier and T Schneider, "Estimation of parameters and eigenmodes of multivariate autoregressive models," *ACM Trans. Math. Softw.*, Vol. 27, pp. 27-57, 2001.

[15] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statictics*, Vol. 6, pp. 461-464, 1978.