Selection of weak VARMA models by Akaïke's information criteria

Boubacar Mainassara, Yacouba

21 June 2010

Selection of weak VARMA models by Akaïke's information criteria

Y. Boubacar Mainassara [a]

[a] *Université Lille III, EQUIPPE-GREMARS, BP 60 149, 59653 Villeneuve d'Ascq cedex, France.*

Abstract

This article considers the problem of orders selections of vector autoregressive moving-average (VARMA) models and the sub-class of vector autoregressive (VAR) models under the assumption that the errors are uncorrelated but not necessarily independent. We relax the standard independence assumption to extend the range of application of the VARMA models, and allow to cover linear representations of general nonlinear processes. We propose a modified criterion to the corrected AIC (Akaïke information criterion) version (AICc) introduced by Tsai and Hurvich (1989). This modified criterion is an approximately unbiased estimator of the Kullback-Leibler discrepancy, originally used to derive AIC-based criteria. Moreover, this criterion requires the estimation of the matrice involved in the asymptotic variance of the quasi-maximum likelihood (QML) estimator of the models, which provide an additional information about models. Monte carlo experiments show that the proposed modified criterion estimates the models orders more accurately than the standard AIC and AICc in large samples and often in small samples.

Key words: AIC, discrepancy, Kullback-Leibler information, QMLE/LSE, order selection, structural representation, weak VARMA models.

1 Introduction

The class of vector autoregressive moving-average (VARMA) models and the sub-class of vector autoregressive (VAR) models are used in time series analysis

Email address: `mailto:yacouba.boubacarmainassara@univ-lille3.fr` (Y. Boubacar Mainassara).
URL: `http://perso.univ-lille3.fr/∼yboubacarmai/` (Y. Boubacar Mainassara).

and econometrics to describe not only the properties of the individual time series but also the possible cross-relationships between the time series (see Reinsel, 1997, Lütkepohl, 2005, 1993).

The parameters estimation is an important step of a VARMA(p, q) processes modeling. Usually, this estimation is carried out by quasi-maximum likelihood or by least squares procedures, given the orders p and q of the model. A companion to the problem of parameters estimation is the problem of model selection, which consists of choosing an appropriate model from a class of candidate models to characterize the data at the hand. The choice of p and q is particularly important because the number of parameters, $(p + q + 3)d^2$ where d is the number of the series, quickly increases with p and q, which entails statistical difficulties. If orders lower than the true orders of the VARMA(p, q) models are selected, the estimate of the parameters will not be consistent and if too high orders are selected, the accuracy of the estimation parameters is likely to be low.

This paper is devoted to the problem of the choice (by minimizing an information criterion) of the orders of VARMA models under the assumption that the errors are uncorrelated but not necessarily independent. Such models are called weak VARMA, by contrast to the strong VARMA models, that are the standard VARMA usually considered in the time series literature and in which the noise is assumed to be iid. Relaxing the independence assumption allows to extend the range of application of the VARMA models, and allows to cover linear representations of general nonlinear processes. The statistical inference of weak ARMA models is mainly limited to the univariate framework (see Francq and Zakoïan, 1998, 2000, 2005, 2007 and Francq, Roy and Zakoïan, 2005 for a review on weak ARMA models).

In the multivariate analysis, important advances have been obtained by Dufour and Pelletier (2005) who study the asymptotic properties of a generalization of the regression-based estimation method proposed by Hannan and Rissanen (1982) under weak assumptions on the innovation process, Francq and Raïssi (2007) who study portmanteau tests for weak VAR models, Boubacar Mainassara and Francq (2009) who study the consistency and the asymptotic normality of the QMLE for weak VARMA models and Boubacar Mainassara (2009a, 2009b) who studies portmanteau tests for weak VARMA models and studies the estimation of the asymptotic variance of QML/LS estimator of weak VARMA models. Dufour and Pelletier (2005) have proposed a modified information criterion

$$\log \det \tilde{\Sigma} + \dim(\gamma) \frac{(\log n)^{1+\delta}}{n}, \quad \delta > 0,$$

which gives a consistent estimates of the orders p and q of a weak VARMA models. Their criterion is a generalization of the information criterion proposed by Hannan and Rissanen (1982).

The choice amongst the models is often made by minimizing an information criterion. The most popular criterion for model selection is the Akaïke information criterion (AIC) proposed by Akaïke (1973). The AIC was designed to be an approximately unbiased estimator of the expected Kullback-Leibler information of a fitted model. Tsai and Hurvich (1989, 1993) derived a bias correction to the AIC for univariate and multivariate autoregressive time series under the assumption that the errors ϵ_t are independent identically distributed (*i.e.* strong models). The main goal of our paper is to complete the above-mentioned results concerning the statistical analysis of weak VARMA models, by proposing a modified version of the AIC criterion.

The paper is organized as follows. Section 2 presents the models that we consider here and summarizes the results on the QMLE/LSE asymptotic distribution obtained by Boubacar Mainassara and Francq (2009). For selftcontainedness purposes, Section 3 recalls useful the results concerning the general multivariate linear regression model, and Section 4 presents the definition and mains properties of the Kullback-Leibler divergence. In Section 5, we present the AIC_M criterion which we minimize to choose the orders for a weak VARMA(p, q) models and we establish his overfitting property. This section is also of interest in the univariate framework because, to our knowledge, this model selection criterion has not been studied for weak ARMA models. Numerical experiments are presented in Section 6. The proofs of the main results are collected in the appendix. We denote by $A \otimes B$ the Kronecker product of two matrices A and B (and by $A^{\otimes 2}$ when the matrix $A = B$), and by $\mathrm{vec}A$ the vector obtained by stacking the columns of A. The reader is refereed to Magnus and Neudecker (1988) for the properties of these operators. Let 0_r be the null vector of $\mathbb{R}^r$, and let I_r be the $r \times r$ identity matrix.

2 Model and assumptions

Consider a d-dimensional stationary process (X_t) satisfying a structural VARMA(p_0, q_0) representation of the form

$$A_{00}X_t - \sum_{i=1}^{p_0} A_{0i}X_{t-i} = B_{00}\epsilon_t - \sum_{i=1}^{q_0} B_{0i}\epsilon_{t-i}, \quad \forall t \in \mathbb{Z} = \{0, \pm 1, \dots\}, \qquad (1)$$

where ϵ_t is a white noise, namely a stationary sequence of centered and uncorrelated random variables with a non singular variance Σ_0. The structural forms are mainly used in econometrics to introduce instantaneous relationships between economic variables. Of course, constraints are necessary for the identifiability of these representations. Let $[A_{00} \dots A_{0p_0} B_{00} \dots B_{0q_0} \Sigma_0]$ be the

$d \times (p_0 + q_0 + 3)d$ matrix of all the coefficients, without any constraint. The parameter of interest is denoted θ_0, where θ_0 belongs to the parameter space $\Theta_{p_0,q_0} \subset \mathbb{R}^{k_0}$, and k_0 is the number of unknown parameters, which is typically much smaller that $(p_0 + q_0 + 3)d^2$. The matrices $A_{00}, \dots A_{0p_0}, B_{00}, \dots B_{0q_0}$ involved in (1) and Σ_0 are specified by θ_0. More precisely, we write $A_{0i} = A_i(\theta_0)$ and $B_{0j} = B_j(\theta_0)$ for $i = 0, \dots, p_0$ and $j = 0, \dots, q_0$, and $\Sigma_0 = \Sigma(\theta_0)$. We need the following assumptions used by Boubacar Mainassara and Francq (2009) to ensure the consistence and the asymptotic normality of the quasi-maximum likelihood estimator (QMLE).

A1: The functions $\theta \mapsto A_i(\theta)$ $i = 0, \dots, p$, $\theta \mapsto B_j(\theta)$ $j = 0, \dots, q$ and $\theta \mapsto \Sigma(\theta)$ admit continuous third order derivatives for all $\theta \in \Theta_{p,q}$.

For simplicity we now write A_i, B_j and Σ instead of $A_i(\theta)$, $B_j(\theta)$ and $\Sigma(\theta)$. Let $A_\theta(z) = A_0 - \sum_{i=1}^p A_i z^i$ and $B_\theta(z) = B_0 - \sum_{i=1}^q B_i z^i$.

A2: For all $\theta \in \Theta_{p,q}$, we have $\det A_\theta(z) \det B_\theta(z) \neq 0$ for all $|z| \leq 1$.

A3: We have $\theta_0 \in \Theta_{p_0,q_0}$, where Θ_{p_0,q_0} is compact.

A4: The process (ϵ_t) is stationary and ergodic.

A5: For all $\theta \in \Theta_{p,q}$ such that $\theta \neq \theta_0$, either the transfer functions

$$A_0^{-1} B_0 B_\theta^{-1}(z) A_\theta(z) \neq A_{00}^{-1} B_{00} B_{\theta_0}^{-1}(z) A_{\theta_0}(z)$$

for some $z \in \mathbb{C}$, or

$$A_0^{-1} B_0 \Sigma B_0' A_0^{-1'} \neq A_{00}^{-1} B_{00} \Sigma_0 B_{00}' A_{00}^{-1'}.$$

A6: We have $\theta_0 \in \overset{\circ}{\Theta}_{p_0,q_0}$, where $\overset{\circ}{\Theta}_{p_0,q_0}$ denotes the interior of Θ_{p_0,q_0}.

A7: We have $E\|\epsilon_t\|^{4+2\nu} < \infty$ and $\sum_{k=0}^\infty \{\alpha_\epsilon(k)\}^{\frac{\nu}{2+\nu}} < \infty$ for some $\nu > 0$.

The reader is referred to Boubacar Mainassara and Francq (2009) for a discussion of these assumptions. Note that (ϵ_t) can be replaced by (X_t) in **A4**, because $X_t = A_{\theta_0}^{-1}(L) B_{\theta_0}(L)\epsilon_t$ and $\epsilon_t = B_{\theta_0}^{-1}(L) A_{\theta_0}(L) X_t$, where L stands for the backward operator. Note that from **A1** the matrices A_0 and B_0 are invertible. Introducing the innovation process $e_t = A_{00}^{-1} B_{00}\epsilon_t$, the structural representation $A_{\theta_0}(L)X_t = B_{\theta_0}(L)\epsilon_t$ can be rewritten as the reduced VARMA representation

$$X_t - \sum_{i=1}^p A_{00}^{-1} A_{0i} X_{t-i} = e_t - \sum_{i=1}^q A_{00}^{-1} B_{0i} B_{00}^{-1} A_{00} e_{t-i}.$$

We thus recursively define $\tilde{e}_t(\theta)$ for $t = 1, \dots, n$ by

$$\tilde{e}_t(\theta) = X_t - \sum_{i=1}^{p} A_0^{-1} A_i X_{t-i} + \sum_{i=1}^{q} A_0^{-1} B_i B_0^{-1} A_0 \tilde{e}_{t-i}(\theta),$$

with initial values $\tilde{e}_0(\theta) = \dots = \tilde{e}_{1-q}(\theta) = X_0 = \dots = X_{1-p} = 0$. The gaussian quasi-likelihood is given by

$$\tilde{\mathrm{L}}_n(\theta) = \prod_{t=1}^{n} \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma_e}} \exp\left\{-\frac{1}{2}\tilde{e}_t'(\theta)\Sigma_e^{-1}\tilde{e}_t(\theta)\right\}, \; \Sigma_e = A_0^{-1} B_0 \Sigma B_0' A_0^{-1\prime}.$$

A quasi-maximum likelihood estimator (QMLE) of θ is a measurable solution $\hat{\theta}_n$ of

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \tilde{\mathrm{L}}_n(\theta).$$

We now use the matrix M_{θ_0} of the coefficients of the reduced form to that made by Boubacar Mainassara and Francq (2009), where

$$M_{\theta_0} = [A_{00}^{-1} A_{01} : \dots : A_{00}^{-1} A_{0p} : A_{00}^{-1} B_{01} B_{00}^{-1} A_{00} : \dots : A_{00}^{-1} B_{0q} B_{00}^{-1} A_{00} : \Sigma_{e0}].$$

Now we need an assumption which specifies how this matrix depends on the parameter θ_0. Let $\dot{M}_{\theta_0}$ be the matrix $\partial \mathrm{vec}(M_\theta)/\partial\theta'$ evaluated at θ_0.

A8: The matrix $\dot{M}_{\theta_0}$ is of full rank k_0.

Under Assumptions **A1**–**A8**, Boubacar Mainassara and Francq (2009) showed the consistency ($\hat{\theta}_n \to \theta_0$ $a.s$ as $n \to \infty$) and the asymptotic normality of the QMLE:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Omega := J^{-1} I J^{-1}), \tag{2}$$

where $J = J(\theta_0)$ and $I = I(\theta_0)$, with

$$J(\theta) = \lim_{n\to\infty} \frac{2}{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log \mathrm{L}_n(\theta) \quad a.s$$

and

$$I(\theta) = \lim_{n\to\infty} \mathrm{Var} \frac{2}{\sqrt{n}} \frac{\partial}{\partial\theta} \log \mathrm{L}_n(\theta).$$

Note that, for VARMA models in reduced form, it is not very restrictive to assume that the coefficients $A_0, \dots, A_p, B_0, \dots, B_q$ are functionally independent of the coefficient Σ_e. Thus we can write $\theta = (\theta^{(1)\prime}, \theta^{(2)\prime})'$, where $\theta^{(1)} \in \mathbb{R}^{k_1}$ depends on $A_0, \dots, A_p$ and $B_0, \dots, B_q$, and where $\theta^{(2)} \in \mathbb{R}^{k_2}$ depends on Σ_e, with $k_1 + k_2 = k_0$. With some abuse of notation, we will then write $e_t(\theta) = e_t(\theta^{(1)})$.

A9: With the previous notation $\theta = (\theta^{(1)\prime}, \theta^{(2)\prime})'$, where $\theta^{(2)} = D \,\mathrm{vec}\, \Sigma_e$ for some matrix D of size $k_2 \times d^2$.

Let J_{11} and I_{11} be respectively the upper-left block of the matrices J and I, with appropriate size. Under Assumptions **A1**–**A9**, in a working paper, Boubacar Mainassara (2009) obtained explicit expressions of I_{11} and J_{11}, given by

$$\operatorname{vec} J_{11} = 2\sum_{i\geq 1} \mathcal{M}\left\{\lambda_i'\otimes\lambda_i'\right\}\operatorname{vec}\Sigma_{e0}^{-1} \quad \text{and}$$

$$\operatorname{vec} I_{11} = 4\sum_{i,j=1}^{+\infty} \mathbf{\Gamma}(i,j)\left(\left\{I_d\otimes\lambda_j'\right\}\otimes\left\{I_d\otimes\lambda_i'\right\}\right)\operatorname{vec}\left(\operatorname{vec}\Sigma_{e0}^{-1}\left\{\operatorname{vec}\Sigma_{e0}^{-1}\right\}'\right),$$

where

$$\mathcal{M} := E\left\{\left(I_{d^2(p+q)}\otimes e_t'\right)^{\otimes 2}\right\},$$

the matrices λ_i depend on θ_0 (see Boubacar Mainassara (2009) for the precise definition of these matrices) and

$$\mathbf{\Gamma}(i,j) = \sum_{h=-\infty}^{+\infty} E\left(\left\{e_{t-h}'\otimes\left(I_{d^2(p+q)}\otimes e_{t-j-h}'\right)\right\}\otimes\left\{e_t'\otimes\left(I_{d^2(p+q)}\otimes e_{t-i}'\right)\right\}\right).$$

We first define an estimator $\hat{J}_n$ of J by

$$\operatorname{vec}\hat{J}_n = \sum_{i\geq 1}\hat{\mathcal{M}}_n\left\{\hat{\lambda}_i'\otimes\hat{\lambda}_i'\right\}\operatorname{vec}\hat{\Sigma}_{e0}^{-1}, \text{ where } \hat{\mathcal{M}}_n := \frac{1}{n}\sum_{t=1}^{n}\left\{\left(I_{d^2(p+q)}\otimes\hat{e}_t'\right)^{\otimes 2}\right\}.$$

To estimate I consider a sequence of real numbers $(b_n)_{n\in\mathbb{N}^*}$ such that

$$b_n\to 0 \quad\text{and}\quad nb_n^{\frac{10+4\nu}{\nu}}\to\infty \quad\text{as}\quad n\to\infty,$$

and a weight function $f:\mathbb{R}\to\mathbb{R}$ which is bounded, with compact support $[-a,a]$ and continuous at the origin with $f(0)=1$. Let $\hat{I}_n$ an estimator of I defined by

$$\operatorname{vec}\hat{I}_n = 4\sum_{i,j=1}^{+\infty}\hat{\mathbf{\Gamma}}_n(i,j)\left(\left\{I_d\otimes\hat{\lambda}_i'\right\}\otimes\left\{I_d\otimes\hat{\lambda}_j'\right\}\right)\operatorname{vec}\left(\operatorname{vec}\hat{\Sigma}_{e0}^{-1}\left\{\operatorname{vec}\hat{\Sigma}_{e0}^{-1}\right\}'\right),$$

where

$$\hat{\mathbf{\Gamma}}_n(i,j) := \sum_{h=-T_n}^{+T_n} f(hb_n)\hat{\mathcal{M}}_{n\,ij,h} \quad\text{and}\quad T_n = \left[\frac{a}{b_n}\right],$$

where $[x]$ denotes the integer part of x, and where

$$\hat{\mathcal{M}}_{n\,ij,h} := \frac{1}{n}\sum_{t=1}^{n-|h|}\left(\left\{\hat{e}_{t-h}'\otimes\left(I_{d^2(p+q)}\otimes\hat{e}_{t-j-h}'\right)\right\}\otimes\left\{\hat{e}_t'\otimes\left(I_{d^2(p+q)}\otimes\hat{e}_{t-i}'\right)\right\}\right).$$

3 General multivariate linear regression model

Let $Z_t = (Z_{1t}, \dots, Z_{dt})'$ be a d-dimensional random vector of response variables, $X_t = (X_{1t}, \dots, X_{kt})'$ be a k-dimensional input variables and $B = (\beta_1, \dots, \beta_d)$ be a $k \times d$ matrix. We consider a multivariate linear model of the form $Z_{it} = X_t'\beta_i + \epsilon_{it}$, $i = 1, \dots, d$, or $Z_t' = X_t'B + \epsilon_t'$, $t = 1, \dots, n$, where the $\epsilon_t = (\epsilon_{1t}, \dots, \epsilon_{dt})'$ are uncorrelated and identically distributed random vectors with variance $\Sigma = E\epsilon_t\epsilon_t'$. The i-th column of B (*i.e.* β_i) is the vector of regression coefficients for the i-th response variable. Now, given the n observations $Z_1, \dots, Z_n$ and $X_1, \dots, X_n$, we define the $n \times d$ data matrix $\mathbf{Z} = (Z_1, \dots, Z_n)'$, the $n \times k$ matrix $\mathbf{X} = (X_1, \dots, X_n)'$ and the $n \times d$ matrix $\varepsilon = (\epsilon_1, \dots, \epsilon_n)'$. Then, we have the multivariate linear model $\mathbf{Z} = \mathbf{X}B + \varepsilon$. Now, it is well known that the QMLE of B is the same as the LSE and, hence, is given by

$$\hat{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}, \quad \text{that is,} \quad \hat{\beta}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_i, \quad i = 1, \dots, d,$$

where $\mathbf{Z}_i = (Z_{i\,1}, \dots, Z_{i\,n})'$ is the i-th column of $\mathbf{Z}$. We also have

$$\hat{\varepsilon} := \mathbf{Z} - \mathbf{X}\hat{B} = M_{\mathbf{X}}\mathbf{Z} = \varepsilon - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon = M_{\mathbf{X}}\varepsilon,$$

where $M_{\mathbf{X}} = I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a projection matrix. The usual unbiased estimator of the error covariance matrix Σ is

$$\begin{aligned} \Sigma^* &= \frac{1}{n-k}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n-k}(\mathbf{Z} - \mathbf{X}\hat{B})'(\mathbf{Z} - \mathbf{X}\hat{B}) \\ &= \frac{1}{n-k}\sum_{t=1}^{n}(Z_t - \hat{B}'X_t)(Z_t - \hat{B}'X_t)' \end{aligned}$$

or $\Sigma^* = (n-k)^{-1}\sum_{t=1}^{n}\hat{\epsilon}_t\hat{\epsilon}_t'$, where the $\hat{\epsilon}_t = Z_t - \hat{B}'X_t$ are the residual vectors. Note that the gaussian quasi-likelihood is given by

$$\mathrm{L}_n(B, \Sigma; \mathbf{Z}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma_e}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{n}(Z_t - B'X_t)'\,\Sigma^{-1}\,(Z_t - B'X_t)\right\},$$

whose maximization shows that the QML estimators of B is equal to $\hat{B}$ and that of Σ is $\hat{\Sigma} := n^{-1}\sum_{t=1}^{n}\hat{\epsilon}_t\hat{\epsilon}_t' = (n-k)n^{-1}\Sigma^*$. Because Σ^* is an unbiased estimator of the matrix Σ, by definition we have $E\{\Sigma^*\} = \Sigma$, we then deduce that

$$\frac{n}{n-k}E\left\{\hat{\Sigma}\right\} = \frac{1}{n-k}E\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n-k}E\varepsilon' M_{\mathbf{X}}\varepsilon = \Sigma. \tag{3}$$

An alternative is to use a result that $n\hat{\Sigma}$ have a asymptotic Wishart distribution[1] with matrix Σ and $n - d(p+q)$ degrees of freedom, so that $E\left\{\hat{\Sigma}^{-1}\right\} \approx n/[n - d(p+q) - d - 1]\Sigma^{-1}$. See Wei (1994, p. 406) and Anderson (2003, p. 296) for these results.

4 Kullback-Leibler discrepancy

Assume that, with respect to a σ-finite measure μ, the true density of the observations $X = (X_1, \ldots, X_n)$ is f_0, and that some candidate model m gives a density $f_m(\cdot, \theta_m)$ to the observations, where θ_m is a k_m-dimensional parameter. The discrepancy between the candidate and the true models can be measured by the Kullback-Leibler divergence (or information)

$$\Delta\left\{f_m(\cdot, \theta_m)|f_0\right\} = E_{f_0} \log \frac{f_0(X)}{f_m(X, \theta_m)} = E_{f_0} \log f_0(X) + \frac{1}{2} d\left\{f_m(\cdot, \theta_m)|f_0\right\},$$

where

$$d\left\{f_m(\cdot, \theta_m)|f_0\right\} = -2E_{f_0} \log f_m(X, \theta_m) = -2\int \left\{\log f_m(x, \theta_m)\right\} f_0(x)\mu(dx)$$

is sometimes called the Kullback-Leibler contrast (or the discrepancy between the approximating and the true models). Using the Jensen inequality, we have

$$\begin{aligned}\Delta\left\{f_m(\cdot, \theta_m)|f_0\right\} &= -\int \log \frac{f_m(x, \theta_m)}{f_0(x)} f_0(x)\mu(dx) \\ &\geq -\log \int \frac{f_m(x, \theta_m)}{f_0(x)} f_0(x)\mu(dx) = 0,\end{aligned}$$

with equality if and only if $f_m(\cdot, \theta_m) = f_0$. This is the main property of the Kullback-Leibler divergence. Minimizing $\Delta\left\{f_m(\cdot, \theta_m)|f_0\right\}$ with respect to $f_m(\cdot, \theta_m)$ is equivalent to minimizing the contrast $d\left\{f_m(\cdot, \theta_m)|f_0\right\}$. Let

$$\theta_{0,m} = \arg\inf_{\theta_m} d\left\{f_m(\cdot, \theta_m)|f_0\right\} = \arg\inf_{\theta_m} -2E \log f_m(X, \theta_m)$$

be an optimal parameter for the model m corresponding to the density $f_m(\cdot, \theta_m)$ (assuming that such a parameter exists). We estimate this optimal parameter by QMLE $\hat{\theta}_{n,m}$.

[1] The Wishart distribution arises in a natural way as a matrix generalization of the chi-square distribution.

5 Identification of VARMA models

Let

$$\begin{aligned}\tilde{\ell}_n(\theta) &= -\frac{2}{n}\log \tilde{\mathrm{L}}_n(\theta)\\ &= \frac{1}{n}\sum_{t=1}^{n}\left\{d\log(2\pi)+\log\det\Sigma_e+\tilde{e}_t'(\theta)\Sigma_e^{-1}\tilde{e}_t(\theta)\right\}.\end{aligned}$$

In Boubacar Mainassara and Francq (2009), it is shown that $\ell_n(\theta) = \tilde{\ell}_n(\theta) + o(1)$ a.s, where

$$\begin{aligned}\ell_n(\theta) &= -\frac{2}{n}\log \mathrm{L}_n(\theta)\\ &= \frac{1}{n}\sum_{t=1}^{n}\left\{d\log(2\pi)+\log\det\Sigma_e+e_t'(\theta)\Sigma_e^{-1}e_t(\theta)\right\},\end{aligned}$$

where

$$e_t(\theta) = A_0^{-1}B_0B_\theta^{-1}(L)A_\theta(L)X_t.$$

It is also shown uniformly in $\theta \in \Theta_{p,q}$ that

$$\frac{\partial \ell_n(\theta)}{\partial\theta} = \frac{\partial \tilde{\ell}_n(\theta)}{\partial\theta} + o(1) \quad a.s.$$

The same equality holds for the second-order derivatives of $\tilde{\ell}_n$. For all $\theta \in \Theta_{p,q}$, we have

$$-2\log \mathrm{L}_n(\theta) = nd\log(2\pi) + n\log\det\Sigma_e + \sum_{t=1}^{n} e_t'(\theta)\Sigma_e^{-1}e_t(\theta).$$

In view of Section 4, minimizing the Kullback-Leibler information of any approximating (or candidate) model, characterized by the parameter vector θ, is equivalent to minimizing the contrast $\Delta(\theta) = E\left\{-2\log \mathrm{L}_n(\theta)\right\}$. Omitting the constant $nd\log(2\pi)$, we find that

$$\Delta(\theta) = n\log\det\Sigma_e + n\mathrm{Tr}\left(\Sigma_e^{-1}S(\theta)\right),$$

where $S(\theta) = Ee_1(\theta)e_1'(\theta)$. In view of the following Lemma, the function $\theta \mapsto \Delta(\theta)$ is minimal for $\theta = \theta_0$.

Lemma 1 *For all $\theta \in \bigcup_{p,q\in\mathbb{N}}\Theta_{p,q}$, we have*

$$\Delta(\theta) \geq \Delta(\theta_0).$$

Let $X = (X_1, \dots, X_n)$ be observation of a process satisfying the VARMA representation (1). Let $\hat{\theta}_n$ be the QMLE of the parameter θ of a candidate VARMA model. Let, $\hat{e}_t = \tilde{e}_t(\hat{\theta}_n)$ be the QMLE/LSE residuals when $p > 0$ or $q > 0$, and let $\hat{e}_t = e_t = X_t$ when $p = q = 0$. When $p + q \neq 0$, we have $\hat{e}_t = 0$ for $t \leq 0$ and $t > n$, and

$$\hat{e}_t = X_t - \sum_{i=1}^{p} A_0^{-1}(\hat{\theta}_n) A_i(\hat{\theta}_n) \hat{X}_{t-i} + \sum_{i=1}^{q} A_0^{-1}(\hat{\theta}_n) B_i(\hat{\theta}_n) B_0^{-1}(\hat{\theta}_n) A_0(\hat{\theta}_n) \hat{e}_{t-i},$$

for $t = 1, \dots, n$, with $\hat{X}_t = 0$ for $t \leq 0$ and $\hat{X}_t = X_t$ for $t \geq 1$.

In view of Lemma 1, it is natural to minimize an estimation of the theoretical criterion $E\Delta(\hat{\theta}_n)$. Note that $E\Delta(\hat{\theta}_n)$ can be interpreted as the average discrepancy when one uses the model of parameter $\hat{\theta}_n$. The Akaïke information criterion (AIC) is an approximately unbiased estimator of $E\Delta(\hat{\theta}_n)$. We will adapt to weak VARMA models the corrected AIC version (AICc) introduced by Tsai and Hurvich (1989) for the univariate strong AR models. Under Assumptions **A1**–**A9**, an approximately unbiased estimator of $E\Delta(\hat{\theta}_n)$ is given by

$$\mathrm{AIC}_M = n \log \det \hat{\Sigma}_e + \frac{n^2 d^2}{nd - k_1} + \frac{nd}{2(nd - k_1)} \left(\mathrm{vec}\, \hat{I}'_{11,n} \right)' \left(\mathrm{vec}\, \hat{J}^{-1}_{11,n} \right), \qquad (4)$$

with vec $\hat{J}_{11,n}$ and vec $\hat{I}_{11,n}$ are defined in Section 2. The AIC_M stands for AIC "modified".

Remark 1 In the standard strong VARMA case, *i.e.* when **A4** is replaced by the assumption that (ϵ_t) is iid, we have $I_{11} = 2J_{11}$, so that $\mathrm{Tr}\left(I_{11} J_{11}^{-1} \right) = 2k_1$. In this case, an approximately unbiased estimator of $E\Delta(\hat{\theta}_n)$ takes the following form

$$\begin{aligned}
\mathrm{AIC}_M &= n \log \det \hat{\Sigma}_e + \frac{n^2 d^2}{nd - k_1} + \frac{nd}{2(nd - k_1)} 2k_1 \\
&= n \log \det \hat{\Sigma}_e + \frac{nd}{nd - k_1} (nd + k_1) \\
&= n \log \det \hat{\Sigma}_e + nd + \frac{nd}{nd - k_1} 2k_1 = \mathrm{AICc}, \qquad (5)
\end{aligned}$$

which illustrates that the standard AIC and AIC_M differ only through the inclusion of the scale factor $nd/(nd - k_1)$ in the penalty term of AIC_M. This factor can play a substantial role in the performance of AIC_M if k_1 is non negligible fraction of the sample size n. In particular, this factor helps to reduce the bias of AIC, which may be substantial when n is not large. Consequently, use of this improved estimator of the discrepancy should lead to improved performance of AIC_M over AIC in terms of model selection.

Remark 2 Given a collection of competing families of approximating models, the one that minimizes $E\Delta(\hat{\theta}_n)$ might be preferred. For model selection, we then choose $\hat{p}$ and $\hat{q}$ as the set which minimizes the information criterion (4).

Remark 3 Consider the univariate case $d = 1$. We then have

$$\mathrm{AIC}_M = n\hat{\sigma}_e^2 + \frac{n}{n-(p+q)}\left(n + \frac{1}{\hat{\sigma}^4}\sum_{i,j,i'=1}^{+\infty}\hat{\gamma}(i,j)\left\{\hat{\lambda}_j\hat{\lambda}_{i'}^{-1'}\otimes\hat{\lambda}_i\hat{\lambda}_{i'}^{-1'}\right\}\right),$$

where $\hat{\sigma}_e^2$ is the variance estimate of the univariate process and where $\hat{\gamma}(i,j)$ are the estimators of $\gamma(i,j) = \sum_{h=-\infty}^{+\infty} E\left(e_t e_{t-i} e_{t-h} e_{t-j-h}\right)$ and $\hat{\lambda}_i'$ are the estimators of $\lambda_i' \in \mathbb{R}^{p+q}$ given in Section 2.

5.2 Other decomposition of the discrepancy

In Section 5.1, the minimal discrepancy (contrast) has been approximated by $-2E\log \mathrm{L}_n(\hat{\theta}_n)$ (the expectation is taken under the true model X). Note that studying this average discrepancy is too difficult because of the dependance between $\hat{\theta}_n$ and X. An alternative slightly different but equivalent interpretation for arriving at the expected discrepancy quantity $E\Delta(\hat{\theta}_n)$, as a criterion for judging the quality of an approximating model, is obtained by supposing $\hat{\theta}_n$ be the QMLE of θ based on the observation X and let $Y = (Y_1, \dots, Y_n)$ be independent observation of a process satisfying the VARMA representation (1) (*i.e.* X and Y independent observations satisfying the same process). Then, we may be interested in approximating the distribution of (Y_t) by using $\mathrm{L}_n(Y, \hat{\theta}_n)$. So we consider the discrepancy for the approximating model (model Y) that uses $\hat{\theta}_n$ and, thus, it is generally easier to search a model that minimizes

$$C(\hat{\theta}_n) = -2E_Y \log \mathrm{L}_n(\hat{\theta}_n), \tag{6}$$

where E_Y denotes the expectation under the candidate model Y. Since $\hat{\theta}_n$ and Y are independent, $C(\hat{\theta}_n)$ is the same quantity as the expected discrepancy $E\Delta(\hat{\theta}_n)$. A model minimizing (6) can be interpreted as a model that will do globally the best job on an independent copy of X, but this model may not be the best for the data at hand. The average discrepancy can be decomposed into

$$C(\hat{\theta}_n) = -2E_X \log \mathrm{L}_n(\hat{\theta}_n) + a_1 + a_2,$$

where

$$a_1 = -2E_X \log \mathrm{L}_n(\theta_0) + 2E_X \log \mathrm{L}_n(\hat{\theta}_n)$$

and

$$a_2 = -2E_Y \log \mathrm{L}_n(\hat{\theta}_n) + 2E_X \log \mathrm{L}_n(\theta_0).$$

The QMLE satisfies $\log \mathrm{L}_n(\hat{\theta}_n) \geq \log \mathrm{L}_n(\theta_0)$ almost surely, thus a_1 can be interpreted as the average over-adjustment (over-fitting) of this QMLE. Now, note that $E_X \log \mathrm{L}_n(\theta_0) = E_Y \log \mathrm{L}_n(\theta_0)$, thus a_2 can be interpreted as an average cost due to the use of the estimated parameter instead of the optimal parameter, when the model is applied to an independent replication of X. We now discuss the regularity conditions needed for a_1 and a_2 to be equivalent to $\mathrm{Tr}\left(I_{11}J_{11}^{-1}\right)$ in the following Proposition.

Proposition 1 *Under Assumptions* **A1**–**A9**, *a_1 and a_2 are both equivalent to $2^{-1}\,Tr\left(I_{11}J_{11}^{-1}\right)$, as $n \to \infty$.*

Remark 4

In view of Proposition 1, a_1 and a_2 are both equivalent to $2^{-1}\mathrm{Tr}\left(I_{11}J_{11}^{-1}\right)$ in the weak VARMA case. In this case, the AIC formula denoted by AIC_M^*

$$\mathrm{AIC}_M^* = -2\log \mathrm{L}_n(\hat{\theta}_n) + \mathrm{Tr}\left(\hat{I}_{11}\hat{J}_{11}^{-1}\right) \tag{7}$$

is an approximately unbiased estimate of the contrast $C(\hat{\theta}_n)$, where $\hat{I}_{11}$ and $\hat{J}_{11}$ are consistent estimators of the matrice I_{11} and J_{11} defined in Section 2. Model selection is then obtained by minimizing (7) over the candidate models.

In the standard strong VARMA case, *i.e.* when **A4** is replaced by the assumption that (ϵ_t) is iid, we have $I_{11} = 2J_{11}$, so that $\mathrm{Tr}\left(I_{11}J_{11}^{-1}\right) = 2k_1$. Therefore, a_1 and a_2 are both equivalent to $k_1 = \dim(\theta_0^{(1)})$. In this case, the AIC_M^* statistic formula takes the more AIC conventional form

$$\mathrm{AIC} = -2\log \mathrm{L}_n(\hat{\theta}_n) + 2k_1 \tag{8}$$

is an approximately unbiased estimate of the contrast $C(\hat{\theta}_n)$

Remark 5 Note that, under some regularity assumptions given in Section 2 and when **A4** is replaced by the assumption that (ϵ_t) is iid, it is shown in Findley (1993) that, a_1 and a_2 are both equivalent to k_1. In this case, the AIC formula

$$\mathrm{AIC} = -2\log \mathrm{L}_n(\hat{\theta}_n) + 2k_1 \tag{9}$$

is an approximately unbiased estimate of the contrast $C(\hat{\theta}_n)$. Model selection is then obtained by minimizing (9) over the candidate models.

For any models with k-dimensional parameter, the AIC$_M$ criterion given in (4) can be rewritten as

$$\mathrm{AIC}_M(k) = n \log \det \hat{\Sigma}_e(k) + \frac{n^2 d^2}{nd - k} + \frac{nd}{2(nd-k)} c_k,$$

where $c_k = \mathrm{Tr}\left(I_{11}(\hat{\theta}_{n,k}) J_{11}^{-1}(\hat{\theta}_{n,k})\right)$ and $\hat{\Sigma}_e(k) = \Sigma_e(\hat{\theta}_{n,k})$.

We define an overfitted model as a model that has more parameters than the true model. Overfitting is analysed here by comparing the model of true orders p_0 and q_0 and an overfitted model of orders $p' = p_0 + \ell_1$ and $q' = q_0 + \ell_2$, where the integers $\ell_1, \ell_2 > 0$. Recall that, for the true VARMA model in the reduced form, the number of unknown parameters in VAR and MA parts is $k_1 = d^2(p_0 + q_0)$. By analog, let $k_1' = d^2(p' + q')$ the number of parameters without any constraints of the overfitted model. Note that, $k_1' = k_1 + \ell$ where $\ell = d^2(\ell_1 + \ell_2)$ and let $c_\ell = c_{k_1'} - c_{k_1}$. The overfitting property of the AIC$_M$ criterion is described here through the probability of overfitting. The following Lemma gives the overfitting property of the VARMA models.

Lemma 2 *The AIC$_M$ criterion overfits if AIC$_M(k_1') <$ AIC$_M(k_1)$. The modified probability that the AIC$_M$ criterion selects the overfitted model is*

$$\mathbf{P}_M := P\left\{AIC_M(k_1 + \ell) < AIC_M(k_1)\right\} = P\left\{\chi^2_\ell > \frac{2\ell + c_\ell}{2}\right\}.$$

Remark 6 In the standard strong VARMA case, *i.e.* when **A4** is replaced by the assumption that (ϵ_t) is iid, we have $c_\ell = 2\ell$. In this case, the probability that the AIC$_M$ criterion selects the overfitted model takes the following form

$$\mathbf{P}_S := P\left\{\mathrm{AIC}_M(k_1 + \ell) < \mathrm{AIC}_M(k_1)\right\} = P\left\{\chi^2_\ell > \frac{2\ell + 2\ell}{2} = 2\ell\right\}.$$

Table 1
The calculated values for the standard version of asymptotic probabilities of overfitting by $\ell = d^2\ell_1$ parameters for strong bivariate VAR model.

ℓ_1	1	2	3	4	5
$\mathbf{P}_S$	0.0915782	0.04238011	0.02034103	0.00999978	0.004995412
ℓ_1	6	7	8	9	10
$\mathbf{P}_S$	0.002524130	0.001286361	0.0006599276	0.0003403570	0.0001763029

Table 2
The calculated values for the standard version of asymptotic probabilities of overfitting by $\ell = d^2(\ell_1 + \ell_2)$ parameters for strong bivariate VARMA model.

(ℓ_1, ℓ_2)	$(1,0)$	$(0,1)$	$(1,1)$	$(1,2)$	$(2,1)$
$\mathbf{P}_S$	0.0915782	0.0915782	0.04238011	0.02034103	0.02034103
(ℓ_1, ℓ_2)	$(2,2)$	$(2,3)$	$(3,2)$	$(3,3)$	$(4,4)$
$\mathbf{P}_S$	0.00999978	0.004995412	0.004995412	0.0003403570	3.617021e-06

From Tables 1 and 2, it is clear that the AIC_M criterion is not consistent in the strong case, since his probability of overfitting is not zero.

6 Numerical illustrations

In this section, by means of Monte Carlo experiments, we present the results of simulations study on small and large sample performance of several **AIC** criteria introduced in this paper. The numerical illustrations of this section are made with the software R (see http://cran.r-project.org/). We generate VAR, VMA and VARMA models, with several choices of their innovation process (ϵ_t). Firstly, we consider the strong case in which (ϵ_t) is defined by

$$\begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} \sim \text{IID}\,\mathcal{N}(0, I_2). \tag{10}$$

The same experiment is repeated for three weak choices for (ϵ_t). In the first one, we assume that (ϵ_t) is an ARCH(1) model:

$$\begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} = \begin{pmatrix} h_{11,t} & 0 \\ 0 & h_{22,t} \end{pmatrix} \begin{pmatrix} \eta_{1,t} \\ \eta_{2,t} \end{pmatrix}, \quad \text{with} \quad \begin{pmatrix} \eta_{1,t} \\ \eta_{2,t} \end{pmatrix} \sim \text{IID}\,\mathcal{N}(0, I_2), \tag{11}$$

and where

$$\begin{pmatrix} h^2_{11,t} \\ h^2_{22,t} \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix} + \begin{pmatrix} 0.45 & 0 \\ 0.4 & 0.25 \end{pmatrix} \begin{pmatrix} \epsilon^2_{1,t-1} \\ \epsilon^2_{2,t-1} \end{pmatrix}.$$

In two other sets of experiments, we assume that (ϵ_t) is defined by

$$\begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} = \begin{pmatrix} \eta_{1,t}\eta_{2,t-1}\eta_{1,t-2} \\ \eta_{2,t}\eta_{1,t-1}\eta_{2,t-2} \end{pmatrix}, \quad \text{with} \quad \begin{pmatrix} \eta_{1,t} \\ \eta_{2,t} \end{pmatrix} \sim \text{IID}\,\mathcal{N}(0, I_2), \tag{12}$$

and then by

$$\begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} = \begin{pmatrix} \eta_{1,t}(|\eta_{1,t-1}|+1)^{-1} \\ \eta_{2,t}(|\eta_{2,t-1}|+1)^{-1} \end{pmatrix}, \quad \text{with} \quad \begin{pmatrix} \eta_{1,t} \\ \eta_{2,t} \end{pmatrix} \sim \mathrm{IID}\,\mathcal{N}(0, I_2), \tag{13}$$

These noises are direct extensions of those defined by Romano and Thombs (1996) in the univariate case.

We used the spectral estimator $\hat{I}^{\mathrm{SP}} := \hat{\mathbf{\Phi}}_r^{-1}(1)\hat{\Sigma}_{\hat{u}_r}\hat{\mathbf{\Phi}}_r'^{-1}(1)$ of the matrix I defined in Theorem 3 of Boubacar Mainassara and Francq (2009). In this theorem, the AR order $r = r(n)$ is automatically selected by BIC criterion in the weak models (in this case, Theorem 3 requires that $r \to \infty$), using the function `VARselect()` of the **vars** R package. In the strong case we can be shown that, the AR spectral estimator is consistent with any fixed value of r (or $r = o(n^{1/3})$ as in Theorem 3 and we took $r = 1$. The matrix J can easily be estimated by its empirical counterpart. The reader is referred to Section 4 in Boubacar Mainassara and Francq (2009) for a discussion of these estimators involved in our modified criterion.

The corresponding relative rejection frequencies to the orders chosen are displayed in bold type in Tables 3, 4, 5, 7. . . ,14.

6.1 Strong and weak VAR case

We simulated N independent trajectories of different sizes of a bivariate VAR(1) model with the strong Gaussian and weak noise above-mentioned. We took $N = 1,000$ when the sample size $n \leq 2000$ and $N = 1,00$ in the opposite case. For each of these N replications, we will fit 6 bivariate candidates models (*i.e.* VAR(k) models with $k = 1, \dots, 6$). The quasi-maximum likelihood (QML) method was used to fit VAR models of order $1, \dots, 6$. The standard and modified versions of **AIC** criteria were used to select among the candidate models. To generate the strong and weak VAR(1) models, we consider the bivariate model of the form:

$$\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{pmatrix} \begin{pmatrix} X_{1t-1} \\ X_{2t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}, \qquad \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{14}$$

Table 3 displays the relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria of a strong (Model I) candidates models, over the $N = 1,000$ independent replications. In view of the observed relative frequency, the order $p = 1$ (*i.e.* VAR(1) model) is selected by all versions of the **AIC** criteria and they have the similar performance.

Table 4 displays the relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria of a strong (Model I) and weak (Model II, with error term (12)) candidates models, over the N independent replications. Table 4 shows that the standard **AIC** criteria clearly did not perform well here when $n \geq 500$, and they have tendency to overestimate the order p. When $n = 500$ the order $p = 1$ is selected by all versions of the **AIC** criteria, but the modified criterion has better performed. As expected, when $n \geq 2000$ the standard **AIC** criteria select a weak VAR(2) model. By contrast, a VAR(1) model is selected by a modified criterion for all values of n and its performance is increasing with n.

Table 5 displays the relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria of a weak VAR(k) candidates models for $k = 1, \dots, 6$, firstly with error term (11) (Model III) and secondly with error term (13) (Model IV). In view of the observed relative frequency, a VAR(1) model is selected by all versions of the **AIC** criteria and they have the same performance in Model IV. By contrast, Table 5 shows that a modified criterion has clearly hight performance in Model III.

Table 6 displays the modified version of asymptotic probabilities of overfitting by $\ell := d^2\ell_1$ parameters for bivariate VAR models of various versions of **AIC** criteria. Table 6 shows clearly that the AIC_M criterion is not consistent in the weak and strong cases, since his probability of overfitting is not zero. As expected, the asymptotic probabilities of overfitting of the standard versions of the **AIC** criteria are very strong than the modified criterion in the weak case. By contrast, they are similar in the strong case for all versions of the **AIC** criteria. The asymptotic probabilities of overfitting of the modified version is decreasing with the sample size n.

6.2 Strong and weak vector moving average (VMA) case

We simulated N independent trajectories of different sizes of bivariate VMA(1) model with the strong Gaussian and the weak noise above-mentioned. We took $N = 1,000$ when the sample size $n \leq 2000$ and $N = 1,00$ in the opposite case. For each of these N replications of VMA(1) model, we will fit 6 candidates models (*i.e.* VMA(k) models with $k = 1, \dots, 6$). The quasi-maximum likelihood (QML) method was used to fit candidates bivariate VMA models of order $1, \dots, 6$; standard and modified versions of **AIC** criteria were used to select among the candidates models.

To generate the strong and weak VMA(1) models, we consider the bivariate

Table 3
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model I		
n	p	AIC	AICc	AIC_M
	1	**84.9**	**91.1**	**90.6**
	2	9.3	7.4	7.8
	3	3.0	1.3	1.2
50	4	0.8	0.1	0.2
	5	1.1	0.1	0.2
	6	0.9	0.0	0.0
	1	**86.9**	**90.4**	**90.9**
	2	8.9	7.5	7.1
	3	2.7	1.6	1.4
100	4	0.7	0.3	0.4
	5	0.4	0.0	0.0
	6	0.4	0.2	0.2
	1	**88.6**	**89.4**	**89.6**
	2	6.7	7.0	6.8
	3	2.8	2.4	2.4
200	4	1.0	0.7	0.7
	5	0.5	0.4	0.4
	6	0.4	0.1	0.1

I: Strong VAR(1) model (14)-(10)

model of the form

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} - \begin{pmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{pmatrix} \begin{pmatrix} \epsilon_{1,t-1} \\ \epsilon_{2,t-1} \end{pmatrix}, \qquad \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{15}$$

Table 7 displays the relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria of a strong (Model I) VMA(k) candidates models, for $k = 1, \dots, 6$, over the N independent replications. Table 7 shows that the standard **AIC** criteria have overfit the order q in the small sample size (*i.e.* $n = 50$) and selected a VMA(6) model. By

Table 4
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model I			Criteria Model II		
n	p	AIC	AICc	AIC$_M$	AIC	AICc	AIC$_M$
	1	**89.3**	**90.0**	**89.6**	**46.7**	**47.3**	**64.1**
	2	6.9	6.5	6.9	38.5	38.7	24.6
	3	2.1	2.0	2.0	9.8	9.6	6.9
500	4	1.1	1.0	0.9	2.5	2.2	2.7
	5	0.5	0.4	0.5	1.1	1.1	0.9
	6	0.1	0.1	0.1	1.4	1.1	0.8
	1	**87.7**	**87.7**	**87.9**	40.6	40.7	**69.3**
	2	8.1	8.1	8.1	**42.7**	**42.8**	22.3
	3	2.7	2.7	2.7	11.9	11.8	5.7
$2,000$	4	1.0	1.0	0.9	3.5	3.5	2.1
	5	0.4	0.4	0.3	0.6	0.7	0.3
	6	0.1	0.1	0.1	0.7	0.5	0.3
	1	**88.0**	**88.0**	**88.0**	34.0	34.0	**61.0**
	2	8.0	8.0	7.0	**44.0**	**44.0**	25.0
	3	3.0	3.0	3.0	16.0	16.0	10.0
$5,000$	4	0.0	0.0	0.0	3.0	3.0	2.0
	5	0.0	0.0	1.0	3.0	3.0	2.0
	6	1.0	1.0	1.0	0.0	0.0	0.0
	1	**87.0**	**87.0**	**87.0**	34.0	34.0	**72.0**
	2	7.0	7.0	7.0	**43.0**	**43.0**	20.0
	3	4.0	4.0	4.0	17.0	17.0	6.0
$10,000$	4	0.0	0.0	0.0	5.0	5.0	1.0
	5	2.0	2.0	2.0	1.0	1.0	1.0
	6	0.0	0.0	0.0	0.0	0.0	0.0

I: Strong VAR(1) model (14)-(10), II: Weak VAR(1) model (14)-(12)

contrast, the modified criterion selected a VMA(1) model. In view of the observed relative frequency, when $n > 50$, the order $q = 1$ (*i.e.* VMA(1) model) is selected by all versions of the **AIC** criteria, but the modified criterion has

Table 5
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model III			Criteria Model IV		
n	p	AIC	AICc	AIC_M	AIC	AICc	AIC_M
	1	**67.0**	**67.9**	**75.1**	**91.9**	**92.5**	**91.1**
	2	22.2	21.8	15.5	5.2	5.0	6.1
	3	6.8	6.6	5.8	1.9	1.7	2.0
500	4	1.6	1.5	2.0	0.6	0.5	0.5
	5	1.7	1.7	1.2	0.4	0.3	0.3
	6	0.7	0.5	0.4	0.0	0.0	0.0
	1	**62.9**	**63.3**	**78.0**	**92.3**	**92.5**	**90.6**
	2	22.3	22.2	15.0	4.8	4.7	6.2
	3	9.4	9.3	4.6	2.2	2.2	2.3
$2,000$	4	3.4	3.5	1.7	0.4	0.4	0.6
	5	1.3	1.2	0.5	0.3	0.3	0.3
	6	0.7	0.5	0.2	0.0	0.0	0.0
	1	**67.0**	**67.0**	**79.0**	**92.0**	**92.0**	**91.0**
	2	16.0	16.0	10.0	5.0	5.0	6.0
	3	9.0	9.0	6.0	1.0	1.0	1.0
$5,000$	4	2.0	2.0	2.0	1.0	1.0	1.0
	5	3.0	3.0	1.0	0.0	0.0	0.0
	6	3.0	3.0	2.0	1.0	1.0	1.0
	1	**67.0**	**67.0**	**82.0**	**92.0**	**92.0**	**88.0**
	2	17.0	17.0	10.0	5.0	5.0	7.0
	3	11.0	11.0	7.0	2.0	2.0	4.0
$10,000$	4	3.0	3.0	1.0	1.0	1.0	1.0
	5	1.0	1.0	0.0	0.0	0.0	0.0
	6	1.0	1.0	0.0	0.0	0.0	0.0

III: Weak VAR(1) model (14)-(11), IV: Weak VAR(1) model (14)-(13)

clearly hight performance.

Table 8 displays the relative frequency (in %) of the order selected by various

Table 6
Modified version of asymptotic probabilities of overfitting by $\ell = d^2\ell_1$ parameters for bivariate VAR models of various versions of **AIC** criteria.

Length	Order	$\mathbf{P}_W$ Model I			$\mathbf{P}_W$ Model II		
n	ℓ_1	$\mathbf{P}_W^{AIC}$	$\mathbf{P}_W^{AICc}$	$\mathbf{P}_W^{AIC_M}$	$\mathbf{P}_W^{AIC}$	$\mathbf{P}_W^{AICc}$	$\mathbf{P}_W^{AIC_M}$
	1	0.076	0.072	0.075	0.492	0.486	0.325
	2	0.040	0.037	0.039	0.370	0.359	0.221
500	3	0.018	0.015	0.014	0.243	0.231	0.144
	4	0.015	0.010	0.013	0.172	0.157	0.091
	5	0.003	0.002	0.002	0.109	0.099	0.059
	1	0.101	0.101	0.100	0.557	0.557	0.283
	2	0.046	0.044	0.045	0.406	0.403	0.204
2000	3	0.027	0.025	0.026	0.274	0.271	0.120
	4	0.014	0.013	0.012	0.172	0.168	0.077
	5	0.008	0.008	0.009	0.126	0.121	0.056

I: Strong VAR(1) model (14)-(10)

II: Weak VAR(1) model (14)-(12)

standard and modified versions of the **AIC** criteria of a strong (Model I) and weak (Model II, with error term (12)) VMA(k) candidates models, for $k = 1, \dots, 6$, over the N independent replications. Table 8 shows that the standard **AIC** criteria have overfit the order q in the small sample size ($n = 20$ and $n = 50$). In view of the observed relative frequency, the order $q = 1$ (*i.e.* VMA(1) model) is selected by all versions of the **AIC** criteria in Models I and II. As expected in Model II, the observed relative frequency of the standard **AIC** criteria is very smaller than a modified one. Table 8 shows also that the standard **AIC** criteria clearly did not perform well here, and they have tendency to overestimate the order $q = 3$. By contrast, in Model I all versions of the **AIC** criteria have the same performance.

Table 9 displays the relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria of a weak VMA(k) candidates models for $k = 1, \dots, 6$, firstly with error term (11) (Model III) and secondly with error term (13) (Model IV). In view of the observed relative frequency, a VMA(1) model is selected by all versions of the **AIC** criteria and they have the same performance in Model IV. By contrast, Table 9 shows that a modified criterion has clearly hight performance in Model III.

Table 7
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model I		
n	q	AIC	AICc	AIC$_M$
	1	1.8	5.8	**53.4**
	2	0.0	0.0	1.3
	3	1.8	4.5	9.7
50	4	7.4	12.0	9.4
	5	25.0	28.8	13.5
	6	**64.0**	**48.9**	12.7
	1	**65.3**	**72.5**	**85.3**
	2	0.2	0.1	0.5
	3	6.5	5.1	5.2
100	4	3.2	2.4	1.4
	5	5.9	4.8	2.9
	6	18.9	15.1	4.7
	1	**92.4**	**93.8**	**94.2**
	2	0.0	0.0	0.0
	3	3.1	2.7	3.2
200	4	1.8	1.9	1.2
	5	1.4	1.0	0.8
	6	1.3	0.6	0.6

I: Strong VMA(1) model (15)-(10)

6.3 Strong and weak VARMA case

We simulated N independent trajectories of different sizes of a bivariate VARMA$(1,1)$ model in echelon form or, more precisely, an ARMA$_E(0,1)$, with the strong Gaussian and weak noise above-mentioned. We took $N = 1,000$ when the sample size $n \leq 2000$ and $N = 1,00$ in the opposite case. For each of these N replications of both models, we have 9 candidates models (*i.e.* VARMA$(1,1)$, VARMA$(2,2)$,VARMA$(2,1)$, VARMA$(1,2)$, VARMA$(1,3)$, VARMA$(3,1)$, VARMA$(3,2)$, VARMA$(2,3)$ and VARMA$(3,3)$

Table 8
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model I			Criteria Model II		
n	q	AIC	AICc	AIC_M	AIC	AICc	AIC_M
	1	**95.1**	**95.8**	**95.6**	**57.3**	**58.5**	**73.4**
	2	0.0	0.0	0.0	0.0	0.0	0.0
	3	2.9	2.6	3.0	34.8	34.4	19.7
500	4	1.4	1.1	1.0	4.5	4.1	4.3
	5	0.4	0.4	0.3	2.0	1.8	1.6
	6	0.2	0.1	0.1	1.4	1.2	1.0
	1	**95.0**	**95.0**	**95.2**	**54.7**	**55.1**	**77.5**
	2	0.0	0.0	0.0	0.0	0.0	0.0
	3	3.1	3.1	3.1	37.9	37.8	17.2
2000	4	1.6	1.6	1.4	4.8	4.7	2.8
	5	0.2	0.2	0.2	1.5	1.3	1.4
	6	0.1	0.1	0.1	1.1	1.1	1.1

I: Strong VMA(1) model (15)-(10)

II: Weak VMA(1) model (15)-(12)

models). These candidates models are constrained in echelon form (*i.e.* an $\text{ARMA}_E(0,k)$ for $k = 1,2,3$). The quasi-maximum likelihood method was used to fit candidates bivariate VARMA models and standard and modified versions of **AIC** criteria were used to select among the candidates models. To generate the strong and weak VARMA$(1,1)$ model, we consider the bivariate model of the form

$$\begin{aligned}\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} &= \begin{pmatrix} 0 & 0 \\ 0 & 0.225 \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} \\ &\quad - \begin{pmatrix} 0 & 0 \\ -0.313 & 0.750 \end{pmatrix} \begin{pmatrix} \epsilon_{1,t-1} \\ \epsilon_{2,t-1} \end{pmatrix}. \end{aligned} \tag{16}$$

Table 10 displays the relative frequency (in %) of the orders selected by various standard and modified versions of the **AIC** criteria of a strong (Model I) candidates VARMA models, over the N independent replications. Table 10

Table 9
Relative frequency (in %) of the order selected by various standard and modified versions of the criteria **AIC**.

Length	Order	Criteria Model III			Criteria Model IV		
n	q	AIC	AICc	AIC$_M$	AIC	AICc	AIC$_M$
	1	**75.9**	**77.2**	**82.8**	**96.0**	**96.5**	**95.6**
	2	0.0	0.0	0.0	0.0	0.0	0.0
	3	17.2	16.5	12.2	2.2	2.1	3.0
500	4	3.7	3.5	2.9	1.3	1.1	1.2
	5	2.0	1.9	1.7	0.5	0.3	0.2
	6	1.2	0.9	0.4	0.0	0.0	0.0
	1	**72.0**	**72.3**	**84.8**	**96.1**	**96.3**	**95.6**
	2	0.0	0.0	0.0	0.0	0.0	0.0
	3	19.4	19.3	10.9	2.8	2.7	3.2
2000	4	5.4	5.4	2.6	0.5	0.5	0.6
	5	2.4	2.2	1.1	0.4	0.4	0.4
	6	0.8	0.8	0.6	0.2	0.1	0.2

III: Weak VMA(1) model (15)-(11)

IV: Weak VMA(1) model (15)-(13)

shows that a standard AICc and a modified AIC$_M$ have performed in the small samples sizes ($n = 20$ and $n = 50$) and selected the true orders of the strong model. By contrast, when $n = 20$ a standard AIC overfit the order q and selected an VARMA$(1, 3)$, but did not perform well. In view of the observed relative frequency in Tables 11 and 12, the true orders $(1, 1)$ (*i.e.* VARMA$(1, 1)$ model) are selected by all versions of the **AIC** criteria. They have similar performance, with a slight advantage to the standard versions.

Tables 13 and 14 display the relative frequency (in %) of the orders selected by various standard and modified versions of the **AIC** criteria of weak candidates VARMA models, firstly with error term (11) (Model III) and secondly, with error term (13) (Model IV). In view of the observed relative frequency, the true orders $(1, 1)$ are selected by all versions of the **AIC** criteria. They have similar performance, with a slight advantage to the standard versions.

Table 15 displays the modified version of asymptotic probabilities of overfitting by $\ell = d^2(\ell_1 + \ell_2)$ parameters for bivariate VARMA models of various versions of **AIC** criteria. Table 15 shows clearly that the AIC$_M$ criterion is not consistent in the weak and strong VARMA cases, since his probability of overfitting is not zero. The modified asymptotic probabilities of overfitting of the standard and modified versions of the **AIC** criteria are similar in the two cases. Note that the asymptotic probabilities of overfitting of the AIC_M criterion decreases when n is large.

7 Conclusion

The results in Section 6 suggest that the relative frequency of the orders selected by the standard (AIC and AICc) and the modified AIC_M versions are comparable, with a slight advantage the modified version, in the strong VAR and VMA models cases. In the weak VAR and VMA models cases, the modified version performs better than the standard versions, which often overestimate the order. By contrast, in the strong and weak VARMA models cases, the standard and modified **AIC** criteria have often the same performance, with a slight advantage to the standard versions. This may be due to the fact that in these cases, the VARMA models may be more parsimonious in terms of the number of parameters involved (to ensure the identifiability problem) than an appropriate finite order VAR models.

Our modified **AIC** criterion proposed, has two major strength: first, AIC_M is an approximately unbiased estimator of the Kullback-Leibler discrepancy $(E\Delta(\hat{\theta}_n))$, originally used to derive AIC-based criteria. Secondly, the AIC_M requires the estimation of the matrice I and J involved in the asymptotic variance of the QML estimator (*i.e.* $\Omega := J^{-1}IJ^{-1}$) of the models, which provide an additional information about models. It can be noted that the performance of the AIC_M criterion increases with n. This fact can be justified by the increasing of estimation accuracy of the matrice I and J involved in Ω.

8 Appendix

Proof of Lemma 1: We have

$$\begin{aligned}\Delta(\theta) = & \, n \log \det \Sigma_e + n\mathrm{Tr}\left(\Sigma_e^{-1}\left\{Ee_1(\theta_0)e_1'(\theta_0) + 2Ee_1(\theta_0)\left\{e_1(\theta) - e_1(\theta_0)\right\}'\right.\right. \\ & \left.\left. + \, E\left(e_1(\theta) - e_1(\theta_0)\right)\left(e_1(\theta) - e_1(\theta_0)\right)'\right\}\right).\end{aligned}$$

Now, using the fact that the linear innovation $e_t(\theta_0)$ is orthogonal to the linear past (*i.e.* to the Hilbert space H_{t-1} generated by the linear combinations of the X_u for $u < t$), it follows that $Ee_1(\theta_0)\left\{e_1(\theta) - e_1(\theta_0)\right\}' = 0$, since $\{e_t(\theta) - e_t(\theta_0)\}$ belongs to the linear past H_{t-1}. We thus have

$$\begin{aligned}\Delta(\theta) = & \, n \log\det\Sigma_e + n\mathrm{Tr}\left(\Sigma_e^{-1}\Sigma_{e0}\right) \\ & +n\mathrm{Tr}\left\{\Sigma_e^{-1}E\left(e_1(\theta)-e_1(\theta_0)\right)\left(e_1(\theta)-e_1(\theta_0)\right)'\right\}.\end{aligned}$$

Moreover

$$\begin{aligned}\Delta(\theta_0) = & \, n\log\det\Sigma_{e0} + n\mathrm{Tr}\left(\Sigma_{e0}^{-1}S(\theta_0)\right) = n\log\det\Sigma_{e0} + n\mathrm{Tr}\left(\Sigma_{e0}^{-1}\Sigma_{e0}\right) \\ = & \, n\log\det\Sigma_{e0} + nd.\end{aligned}$$

Thus, we obtain

$$\begin{aligned}\Delta(\theta)-\Delta(\theta_0) = & -n\log\det\left(\Sigma_e^{-1}\Sigma_{e0}\right) - nd + n\mathrm{Tr}\left(\Sigma_e^{-1}\Sigma_{e0}\right) \\ & +n\mathrm{Tr}\left\{\Sigma_e^{-1}E\left(e_1(\theta)-e_1(\theta_0)\right)\left(e_1(\theta)-e_1(\theta_0)\right)'\right\} \\ \geq & -n\log\det\left(\Sigma_e^{-1}\Sigma_{e0}\right) - nd + n\mathrm{Tr}\left(\Sigma_e^{-1}\Sigma_{e0}\right),\end{aligned}$$

with equality if and only if $e_1(\theta) = e_1(\theta_0)$ a.s. Using the elementary inequality $\mathrm{Tr}(A^{-1}B) - \log\det(A^{-1}B) \geq \mathrm{Tr}(A^{-1}A) - \log\det(A^{-1}A) = d$ for all symmetric positive semi-definite matrices of order $d\times d$, it is easy see that $\Delta(\theta)-\Delta(\theta_0)\geq 0$. The proof is complete. $\square$

Justification of (4). Using a Taylor expansion of the functions $\partial\log\mathrm{L}_n(\hat\theta_n)/\partial\theta^{(1)}$ around $\theta_0^{(1)}$, it follows that

$$\left(\hat\theta_n^{(1)}-\theta_0^{(1)}\right) \overset{o_p\left(\frac{1}{\sqrt{n}}\right)}{=} -\frac{2}{n}J_{11}^{-1}\frac{\partial\log\mathrm{L}_n(\theta_0)}{\partial\theta^{(1)}} = -\frac{2}{n}\sum_{t=1}^{n}J_{11}^{-1}\frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\Sigma_{e0}^{-1}e_t(\theta_0), \quad (17)$$

where $a\overset{c}{=}b$ signifies $a=b+c$ and where $J_{11}=J_{11}(\theta_0)$ with

$$J_{11}(\theta) = \lim_{n\to\infty}\frac{2}{n}\frac{\partial^2\log\mathrm{L}_n(\theta)}{\partial\theta^{(1)}\partial\theta^{(1)'}} \quad a.s.$$

We have

$$E\Delta(\hat\theta_n) = En\log\det\hat\Sigma_e + nE\mathrm{Tr}\left(\hat\Sigma_e^{-1}S(\hat\theta_n)\right), \quad (18)$$

where $\hat\Sigma_e = \Sigma_e(\hat\theta_n)$ is the estimated error variance matrix under $\hat\theta_n$, with $\Sigma_e(\theta) = n^{-1}\sum_{t=1}^n e_t(\theta)e_t'(\theta)$. Then the first term on the right-hand side of (18) can be estimated without bias by $n\log\det\left\{n^{-1}\sum_{t=1}^n e_t(\hat\theta_n)e_t'(\hat\theta_n)\right\}$. Hence, only an estimate for the second term needs to be considered. Moreover, in view of (2), a Taylor expansion of $e_t(\theta)$ around $\theta_0^{(1)}$ yields

$$e_t(\theta) = e_t(\theta_0) + \frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}(\theta^{(1)}-\theta_0^{(1)}) + R_t, \quad (19)$$

where

$$R_t = \frac{1}{2}(\theta^{(1)} - \theta_0^{(1)})' \frac{\partial^2 e_t(\theta^*)}{\partial\theta^{(1)}\partial\theta^{(1)'}}(\theta^{(1)} - \theta_0^{(1)}) = O_P\left(\pi^2\right),$$

with $\pi = \left\|\theta^{(1)} - \theta_0^{(1)}\right\|$ and θ^* is between $\theta_0^{(1)}$ and $\theta^{(1)}$. We then obtain

$$\begin{aligned}
S(\theta) = S(\theta_0) &+ E\left\{\frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}(\theta^{(1)} - \theta_0^{(1)})e_t'(\theta_0)\right\} + ER_t e_t'(\theta_0) \\
&+ E\left\{e_t(\theta_0)(\theta^{(1)} - \theta_0^{(1)})' \frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\right\} + D\left(\theta^{(1)}\right) \\
&+ ER_t\left\{(\theta^{(1)} - \theta_0^{(1)})' \frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\right\} + Ee_t(\theta_0)R_t \\
&+ E\left\{\frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}(\theta^{(1)} - \theta_0^{(1)})\right\} R_t + ER_t^2,
\end{aligned}$$

where

$$D\left(\theta^{(1)}\right) = E\left\{\frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}(\theta^{(1)} - \theta_0^{(1)})(\theta^{(1)} - \theta_0^{(1)})' \frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\right\}.$$

Using the orthogonality between $e_t(\theta_0)$ and any linear combination of the past values of $e_t(\theta_0)$ (in particular $\partial e_t(\theta_0)/\partial\theta'$ and $\partial^2 e_t(\theta_0)/\partial\theta\partial\theta'$), and the fact that $Ee_t(\theta_0) = 0$, we have

$$S(\theta) = S(\theta_0) + D(\theta^{(1)}) + O\left(\pi^4\right) = \Sigma_{e0} + D(\theta^{(1)}) + O\left(\pi^4\right),$$

where $\Sigma_{e0} = \Sigma_e(\theta_0)$. Thus, we can write the expected discrepancy quantity in (18) as

$$\begin{aligned}
E\Delta(\hat\theta_n) = En \log\det\hat\Sigma_e &+ nE\mathrm{Tr}\left(\hat\Sigma_e^{-1}\Sigma_{e0}\right) + nE\mathrm{Tr}\left(\hat\Sigma_e^{-1}D(\hat\theta_n^{(1)})\right) \\
&+ nE\left\{\mathrm{Tr}\left(\hat\Sigma_e^{-1}\right)O_P\left(\frac{1}{n^2}\right)\right\}.
\end{aligned} \tag{20}$$

As in the classical multivariate regression model, an analog of (3) is

$$\Sigma_{e0} \approx \frac{n}{n - d(p+q)}E\left\{\hat\Sigma_e\right\} = \frac{dn}{dn - k_1}E\left\{\hat\Sigma_e\right\}.$$

Thus, using the last approximation and from the consistency of $\hat\Sigma_e$, we obtain

$$E\left\{\hat\Sigma_e^{-1}\right\} \approx \left\{E\hat\Sigma_e\right\}^{-1} \approx nd(nd - k_1)^{-1}\Sigma_{e0}^{-1}. \tag{21}$$

Using the elementary property on the trace, we have

$$\begin{aligned}
\mathrm{Tr}\left\{\Sigma_e^{-1}(\theta)D\left(\theta_n^{(1)}\right)\right\} &= \mathrm{Tr}\left(\Sigma_e^{-1}(\theta)E\left\{\frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}(\theta^{(1)}-\theta_0^{(1)})(\theta^{(1)}-\theta_0^{(1)})'\frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\right\}\right) \\
&= E\left(\mathrm{Tr}\left\{\frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\Sigma_e^{-1}(\theta)\frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}(\theta^{(1)}-\theta_0^{(1)})'(\theta^{(1)}-\theta_0^{(1)})\right\}\right) \\
&= \mathrm{Tr}\left(E\left\{\frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\Sigma_e^{-1}(\theta)\frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}\right\}(\theta^{(1)}-\theta_0^{(1)})'(\theta^{(1)}-\theta_0^{(1)})\right).
\end{aligned}$$

Now, using (2), (21) and the last equality, we have

$$\begin{aligned}
E\mathrm{Tr}\left\{\hat{\Sigma}_e^{-1}D\left(\hat{\theta}_n^{(1)}\right)\right\} &= \frac{1}{n}\mathrm{Tr}\left(E\left\{\frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\hat{\Sigma}_e^{-1}\frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}\right\}\right. \\
&\quad \left. En(\hat{\theta}_n^{(1)}-\theta_0^{(1)})'(\hat{\theta}_n^{(1)}-\theta_0^{(1)})\right) \\
&= \frac{d}{nd-k_1}\mathrm{Tr}\left(E\left\{\frac{\partial e_t'(\theta_0)}{\partial\theta^{(1)}}\Sigma_{e0}^{-1}\frac{\partial e_t(\theta_0)}{\partial\theta^{(1)'}}\right\}J_{11}^{-1}I_{11}J_{11}^{-1}\right) \\
&= \frac{d}{2(nd-k_1)}\mathrm{Tr}\left(I_{11}J_{11}^{-1}\right),
\end{aligned}$$

where $J_{11} = 2E\left\{\partial e_t'(\theta_0)/\partial\theta^{(1)}\Sigma_{e0}^{-1}\partial e_t(\theta_0)/\partial\theta^{(1)'}\right\}$ (see Theorem 3 in Boubacar Mainassara and Francq, 2009). Thus, using (21), we have

$$\begin{aligned}
E\mathrm{Tr}\left(\hat{\Sigma}_e^{-1}S(\hat{\theta}_n)\right) &= E\mathrm{Tr}\left(\hat{\Sigma}_e^{-1}\Sigma_{e0}\right) + E\mathrm{Tr}\left\{\hat{\Sigma}_e^{-1}D\left(\hat{\theta}_n^{(1)}\right)\right\} \\
&\quad + E\left\{\mathrm{Tr}\left(\hat{\Sigma}_e^{-1}\right)O_P\left(\frac{1}{n^2}\right)\right\} \\
&= \frac{nd}{nd-k_1}\mathrm{Tr}\left(\Sigma_{e0}^{-1}\Sigma_{e0}\right) + \frac{d}{2(nd-k_1)}\mathrm{Tr}\left(I_{11}J_{11}^{-1}\right) + O\left(\frac{1}{n^2}\right) \\
&= \frac{nd^2}{nd-k_1} + \frac{d}{2(nd-k_1)}\mathrm{Tr}\left(I_{11}J_{11}^{-1}\right) + O\left(\frac{1}{n^2}\right).
\end{aligned}$$

Therefore, using the last expression in (20), we deduce an approximately unbiased estimator of $E\Delta(\hat{\theta}_n)$ given by

$$\mathrm{AIC}_M = n\log\det\hat{\Sigma}_e + \frac{n^2d^2}{nd-k_1} + \frac{nd}{2(nd-k_1)}\mathrm{Tr}\left(\hat{I}_{11,n}\hat{J}_{11,n}^{-1}\right),$$

where $\hat{J}_{11,n}$ and $\hat{I}_{11,n}$ are defined in Section 2. Using $\mathrm{Tr}\,(AB) = \mathrm{vec}\,(A')'\,\mathrm{vec}\,(B)$, we then obtain

$$\mathrm{AIC}_M = n \log \det \hat{\Sigma}_e + \frac{n^2 d^2}{nd - k_1} + \frac{nd}{2(nd - k_1)} \left(\mathrm{vec}\, \hat{I}'_{11,n}\right)' \left(\mathrm{vec}\, \hat{J}^{-1}_{11,n}\right).$$

The justification is complete. $\square$

Proof of Remark 3: For $d = 1$, we have

$$\mathrm{AIC}_M = n\hat{\sigma}_e^2 + \frac{n}{n-(p+q)} \left\{ n + \frac{1}{2} \left(\mathrm{vec}\, \hat{I}'_{11}\right)' \left(\mathrm{vec}\, \hat{J}^{-1}_{11}\right) \right\}.$$

In view of Section 2, we obtained

$$\mathrm{vec}\, \hat{J}_{11} = 2 \sum_{i' \geq 1} \left\{ \hat{\lambda}_{i'} \otimes \hat{\lambda}_{i'} \right\}' \quad \text{and} \quad \mathrm{vec}\, \hat{I}_{11} = \frac{4}{\hat{\sigma}^4} \sum_{i,j=1}^{+\infty} \hat{\gamma}(i,j) \left\{ \hat{\lambda}_j \otimes \hat{\lambda}_i \right\}',$$

where $\hat{\gamma}(i,j)$ are the estimators of $\gamma(i,j) = \sum_{h=-\infty}^{+\infty} E\left(e_t e_{t-i} e_{t-h} e_{t-j-h}\right)$ and $\hat{\lambda}'_i$ are the estimators of $\lambda'_i \in \mathbb{R}^{p+q}$ given in Section 2. Using the last expressions of $\mathrm{vec}\, \hat{J}_{11}$ and $\mathrm{vec}\, \hat{I}_{11}$, we then have

$$\mathrm{AIC}_M = n\hat{\sigma}_e^2 + \frac{n}{n-(p+q)} \left(n + \frac{1}{\hat{\sigma}^4} \sum_{i,j,i'=1}^{+\infty} \hat{\gamma}(i,j) \left\{ \hat{\lambda}_j \otimes \hat{\lambda}_i \right\} \left\{ \hat{\lambda}_{i'} \otimes \hat{\lambda}_{i'} \right\}^{-1'} \right).$$

Using $(A \otimes B)(C \otimes D) = AC \otimes BD$, we have

$$\mathrm{AIC}_M = n\hat{\sigma}_e^2 + \frac{n}{n-(p+q)} \left(n + \frac{1}{\hat{\sigma}^4} \sum_{i,j,i'=1}^{+\infty} \hat{\gamma}(i,j) \left\{ \hat{\lambda}_j \hat{\lambda}_{i'}^{-1'} \otimes \hat{\lambda}_i \hat{\lambda}_{i'}^{-1'} \right\} \right).$$

The proof is complete. $\square$

Proof of Proposition 1: Using a Taylor expansion of the quasi log-likelihood, we obtain

$$-2 \log \mathrm{L}_n(\theta_0) = -2 \log \mathrm{L}_n(\hat{\theta}_n) + \frac{n}{2} (\hat{\theta}_n^{(1)} - \theta_0^{(1)})' J_{11} (\hat{\theta}_n^{(1)} - \theta_0^{(1)}) + o_P(1).$$

Taking the expectation (under the true model) of both sides, and in view of (2) we shown that

$$\begin{aligned} E_X n(\hat{\theta}_n^{(1)} - \theta_0^{(1)})' J_{11} (\hat{\theta}_n^{(1)} - \theta_0^{(1)}) &= \mathrm{Tr} \left\{ J_{11} E_X n(\hat{\theta}_n^{(1)} - \theta_0^{(1)})' (\hat{\theta}_n^{(1)} - \theta_0^{(1)}) \right\} \\ &\to \mathrm{Tr} \left(I_{11} J_{11}^{-1} \right), \end{aligned}$$

we then obtain $a_1 = 2^{-1}\mathrm{Tr}\left(I_{11}J_{11}^{-1}\right) + o(1)$. Now a Taylor expansion of the discrepancy yields

$$\begin{aligned}
\Delta(\hat{\theta}_n) &= \Delta(\theta_0) + (\hat{\theta}_n^{(1)} - \theta_0^{(1)})' \left.\frac{\partial \Delta(\theta)}{\partial \theta^{(1)}}\right|_{\theta=\theta_0} \\
&\quad + \frac{1}{2}(\hat{\theta}_n^{(1)} - \theta_0^{(1)})' \left.\frac{\partial^2 \Delta(\theta)}{\partial \theta^{(1)} \partial \theta^{(1)'}}\right|_{\theta=\theta_0} (\hat{\theta}_n^{(1)} - \theta_0^{(1)}) + o_P(1) \\
&= \Delta(\theta_0) + \frac{n}{2}(\hat{\theta}_n^{(1)} - \theta_0^{(1)})' J_{11} (\hat{\theta}_n^{(1)} - \theta_0^{(1)}) + o_P(1),
\end{aligned}$$

assuming that the discrepancy is smooth enough, and that we can take its derivatives under the expectation sign. We then deduce that

$$E_Y - 2\log \mathrm{L}_n(\hat{\theta}_n) = E_X \Delta(\hat{\theta}_n) = E_X \Delta(\theta_0) + \frac{1}{2}\mathrm{Tr}\left(I_{11}J_{11}^{-1}\right) + o(1),$$

which shows that a_2 is equivalent to a_1. The proof is complete. $\square$

Proof of Lemma 2: We denote by $|A|$, the determinant of the matrix A. The probability that the AIC_M criterion selects the overfitted model is

$$\begin{aligned}
P\left\{\mathrm{AIC}_M(k_1') < \mathrm{AIC}_M(k_1)\right\} &= P\left\{ n \log\left|\hat{\Sigma}_e(k_1')\right| + \frac{n^2 d^2}{nd - k_1'} + \frac{ndc_{k_1'}}{2(nd - k_1')} \right. \\
&\qquad \left. < n\log\left|\hat{\Sigma}_e(k_1)\right| + \frac{n^2 d^2}{nd - k_1} + \frac{ndc_{k_1}}{2(nd - k_1)} \right\} \\
&= P\left\{\mathrm{AIC}_M(k_1 + \ell) < \mathrm{AIC}_M(k_1)\right\} \\
&= P\left\{ n\log\left\{ \frac{\left|n\hat{\Sigma}_e(k_1+\ell)\right|}{\left|n\hat{\Sigma}_e(k_1)\right|} \right\} < \frac{n^2 d^2}{nd - k_1} \right. \\
&\qquad + \frac{ndc_{k_1}}{2(nd-k_1)} - \frac{n^2 d^2}{nd - (k_1+\ell)} \\
&\qquad \left. - \frac{nd(c_{k_1} + c_\ell)}{2\left[nd - (k_1+\ell)\right]} \right\} \\
&= P\left\{ n\log\left\{ \frac{\left|n\hat{\Sigma}_e(k_1+\ell)\right|}{\left|n\hat{\Sigma}_e(k_1)\right|} \right\} \right. \\
&\qquad < \frac{-n^2 \ell d^2}{(nd - k_1)\left[nd - (k_1+\ell)\right]} \\
&\qquad \left. + \frac{nd(k_1 c_\ell - \ell c_{k_1}) - n^2 d^2 c_\ell}{2(nd-k_1)\left[nd - (k_1+\ell)\right]} \right\}.
\end{aligned}$$

Let $q_1 = k_1/d$ and $q_2 = (k_1+\ell)/d$, we denote

$$\frac{\left|n\hat{\Sigma}_e(k_1+\ell)\right|}{\left|n\hat{\Sigma}_e(k_1)\right|} = \frac{\left|n\hat{\Sigma}_e(k_1+\ell)\right|}{\left|n\hat{\Sigma}_e(k_1+\ell) + n\left\{\hat{\Sigma}_e(k_1) - \hat{\Sigma}_e(k_1+\ell)\right\}\right|} \sim U_{d,\ell,n-q_2},$$

where $U_{d,\ell,n-q_2}$ is the U-statistic (see Anderson, 2003, chap. 8), a generalized version of the F-statistic used for the univariate case. From Theorem 3.2.15 in Muirhead (1982, p. 100), the distribution of the determinants $|n\hat{\Sigma}_e(k_1)|$ and $|n\hat{\Sigma}_e(k_1+\ell)|$ are respectively the product of independent χ^2 random variables,

$$\frac{\left|n\hat{\Sigma}_e(k_1)\right|}{|\Sigma_{e0}|} \sim \prod_{i=1}^{d} \chi^2_{n-q_1-i+1} \quad \text{and} \quad \frac{\left|n\hat{\Sigma}_e(k_1+\ell)\right|}{|\Sigma_{e0}|} \sim \prod_{i=1}^{d} \chi^2_{n-q_2-i+1}.$$

Note that in view of Theorem 7.3.2 (see Anderson, 2003, p. 260) $n\left\{\hat{\Sigma}_e(k_1) - \hat{\Sigma}_e(k_1+\ell)\right\} \sim \mathrm{W}_d\left(\ell/d, \Sigma_{e0}\right)$, where the subscript on W denoting the size of the matrix Σ_{e0}. Using the previous results and Lemma 8.4.2 (see Anderson, 2003, p. 305), it follows that the distribution of the ratio $|n\hat{\Sigma}_e(k_1+\ell)|/|n\hat{\Sigma}_e(k_1)|$ is the multivariate $Beta_d$ distribution [2] i.e. the product of independents *Beta* distributions (see Anderson, 2003, Section 5.2):

$$\frac{\left|n\hat{\Sigma}_e(k_1+\ell)\right|}{\left|n\hat{\Sigma}_e(k_1)\right|} \sim \prod_{i=1}^{d} Beta\left(\frac{n-q_2-i+1}{2}, \frac{\ell}{2d}\right).$$

Expressed in terms of independent χ^2, we obtain

$$\left\{\frac{\left|n\hat{\Sigma}_e(k_1+\ell)\right|}{\left|n\hat{\Sigma}_e(k_1)\right|}\right\}^{-1} = \frac{\left|n\hat{\Sigma}_e(k_1)\right|}{\left|n\hat{\Sigma}_e(k_1+\ell)\right|} \sim \prod_{i=1}^{d}\left(1 + \frac{\chi^2_{\ell/d}}{\chi^2_{n-q_2-i+1}}\right).$$

Thus the probability of overfitting for AIC_M criterion can be rewrite as

$$\begin{aligned} P\left\{\mathrm{AIC}_M(k_1+\ell) < \mathrm{AIC}_M(k_1)\right\} = P\Bigg\{ & -n\sum_{i=1}^{d}\log\left(1 + \frac{\chi^2_{\ell/d}}{\chi^2_{n-q_2-i+1}}\right) \\ & < \frac{-n^2\ell d^2}{(nd-k_1)\left[nd-(k_1+\ell)\right]} \\ & + \frac{nd(k_1 c_\ell - \ell c_{k_1}) - n^2 d^2 c_\ell}{2(nd-k_1)\left[nd-(k_1+\ell)\right]}\Bigg\}. \end{aligned}$$

[2] The multivariate beta distribution generalizes the usual beta distribution in much the same way that the Wishart distribution generalizes the χ^2 distribution.

Recall that, $\log(1+x) \simeq x$ for small value of $|x|$. Using the fact that $\chi^2_{n-q_2-i+1}/n \to 1$ a.s. as $n \to \infty$ for k_1, ℓ fixed and $1 \leq i \leq d$; it follows that

$$
\begin{aligned}
n \sum_{i=1}^{d} \log\left(1 + \frac{\chi^2_{\ell/d}}{\chi^2_{n-q_2-i+1}}\right) &= n \sum_{i=1}^{d} \log\left(1 + \frac{(1/n)\chi^2_{\ell/d}}{(1/n)\chi^2_{n-q_2-i+1}}\right) \\
&\to n \sum_{i=1}^{d} \frac{(1/n)\chi^2_{\ell/d}}{(1/n)\chi^2_{n-q_2-i+1}} \to \sum_{i=1}^{d} \chi^2_{\ell/d} = \chi^2_{\ell}. \qquad (22)
\end{aligned}
$$

Note that, as $n \to \infty$, for k_1, ℓ and d fixed, we have

$$
\begin{aligned}
&\frac{-n^2 \ell d^2}{(nd-k_1)\left[nd-(k_1+\ell)\right]} + \frac{nd(k_1 c_\ell - \ell c_{k_1}) - n^2 d^2 c_\ell}{2(nd-k_1)\left[nd-(k_1+\ell)\right]} \\
&= \frac{-2n^2 \ell d^2 + nd(k_1 c_\ell - \ell c_{k_1}) - n^2 d^2 c_\ell}{2(nd-k_1)\left[nd-(k_1+\ell)\right]} \to -\frac{2\ell + c_\ell}{2}. \qquad (23)
\end{aligned}
$$

In view of (22) and (23), we deduce the following asymptotic probability of overfitting

$$
P\left\{\mathrm{AIC}_M(k_1+\ell) < \mathrm{AIC}_M(k_1)\right\} = P\left\{\chi^2_\ell > \frac{2\ell + c_\ell}{2}\right\}.
$$

The proof is complete. $\square$

References

Akaïke, H. (1973) Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory,* Eds. B. N. Petrov and F. Csáki, pp. 267–281. Budapest: Akadémia Kiado.

Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis 3nd Edn.* New Jersey, Wiley.

Boubacar Mainassara, Y. (2009a) Multivariate portmanteau test for structural VARMA models with uncorrelated but non-independent error terms. *Working Papers,* http://perso.univ-lille3.fr/∼yboubacarmai/yac/VARMAPORTM.pdf

Boubacar Mainassara, Y. (2009b) Estimating the asymptotic variance matrix of structural VARMA models with uncorrelated but non-independent error terms. *Working Papers,* http://perso.univ-lille3.fr/∼yboubacarmai/yac/VARMAEstCOV2.pdf

Boubacar Mainassara, Y. and Francq, C. (2009) Estimating structural VARMA models with uncorrelated but non-independent error terms. *Working Papers,* http://perso.univ-lille3.fr/∼cfrancq/pub.html.

Dufour, J-M., and Pelletier, D. (2005) Practical methods for modelling weak VARMA processes: identification, estimation and specification with a macroeconomic application. *Technical report, Département de sciences économiques and CIREQ, Université de Montréal, Montréal, Canada.*

Findley, D.F. (1993) The overfitting principles supporting AIC, Statistical Research Division Report RR 93/04, Bureau of the Census.

Francq, C. and Raïssi, H. (2007) Multivariate Portmanteau Test for Autoregressive Models with Uncorrelated but Nonindependent Errors, *Journal of Time Series Analysis* 28, 454–470.

Francq, C., Roy, R. and Zakoïan, J-M. (2005) Diagnostic checking in ARMA Models with Uncorrelated Errors, *Journal of the American Statistical Association* 100, 532–544.

Francq, and Zakoïan, J-M. (1998) Estimating linear representations of nonlinear processes, *Journal of Statistical Planning and Inference* 68, 145–165.

Francq, and Zakoïan, J-M. (2000) Covariance matrix estimation of mixing weak ARMA models, *Journal of Statistical Planning and Inference* 83, 369–394.

Francq, and Zakoïan, J-M. (2005) Recent results for linear time series models with non independent innovations. In *Statistical Modeling and Analysis for Complex Data Problems,* Chap. 12 (eds P. Duchesne and B. Rémillard). New York: Springer Verlag, 241–265.

Francq, and Zakoïan, J-M. (2007) HAC estimation and strong linearity testing in weak ARMA models, *Journal of Multivariate Analysis* 98, 114–144.

Hannan, E. J. and Rissanen (1982) Recursive estimation of mixed of Autoregressive Moving Average order, *Biometrika* 69, 81–94.

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

Hurvich, C. M. and Tsai, C-L. (1993) A corrected Akaïke information criterion for vector autoregressive model selection. *Journal of Time Series Analysis* 14, 271–279.

Lütkepohl, H. (1993) *Introduction to multiple time series analysis.* Springer Verlag, Berlin.

Lütkepohl, H. (2005) *New introduction to multiple time series analysis.* Springer Verlag, Berlin.

Magnus, J.R. and H. Neudecker (1988) *Matrix Differential Calculus with Application in Statistics and Econometrics*. New-York, Wiley.

Muirhead, R.J. (1982) *Aspects of Multivariate Statistical Theory.* Wiley, New-York, Wiley.

Reinsel, G. C. (1997) *Elements of multivariate time series Analysis.* Second edition. Springer Verlag, New York.

Romano, J. L. and Thombs, L. A. (1996) Inference for autocorrelations under weak assumptions, *Journal of the American Statistical Association* 91, 590–600.

Wei, W.H. (1994) *Time Series Analysis: Univariate and Multivariate Methods 2nd Edn.* New York, Addison-Wesley.

Table 10
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model I		
n	(p,q)	AIC	AICc	AIC_M
	$(1,1)$	27.5	**64.1**	**62.1**
	$(2,2)$	0.2	0.0	0.5
	$(2,1)$	1.6	2.0	4.1
	$(1,2)$	19.3	22.2	14.9
20	$(3,3)$	2.8	0.0	0.8
	$(3,2)$	3.5	0.0	1.3
	$(3,1)$	0.6	0.0	1.8
	$(2,3)$	11.0	0.1	2.1
	$(1,3)$	**33.5**	11.6	12.4
	$(1,1)$	**39.9**	**61.9**	**58.3**
	$(2,2)$	0.1	0.0	0.3
	$(2,1)$	1.5	1.5	2.4
	$(1,2)$	11.8	12.0	9.6
50	$(3,3)$	9.7	1.3	6.3
	$(3,2)$	3.0	2.0	3.2
	$(3,1)$	0.2	0.3	0.4
	$(2,3)$	24.5	13.6	15.5
	$(1,3)$	9.3	7.4	4.0
	$(1,1)$	**52.2**	**62.6**	**56.9**
	$(2,2)$	0.1	0.1	0.1
	$(2,1)$	0.9	1.0	2.2
	$(1,2)$	6.3	6.7	6.9
100	$(3,3)$	12.5	6.4	10.7
	$(3,2)$	1.7	1.3	2.2
	$(3,1)$	0.2	0.1	0.6
	$(2,3)$	22.5	18.7	18.0
	$(1,3)$	3.6	3.1	2.4

I: Strong VARMA(1,1) model (16)-(10)

Table 11
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model I			Criteria Model II		
n	(p,q)	AIC	AICc	AIC_M	AIC	AICc	AIC_M
	$(1,1)$	**82.0**	**83.4**	**74.4**	**62.6**	**63.7**	**59.2**
	$(2,2)$	0.1	0.1	1.1	1.2	0.9	2.9
	$(2,1)$	0.8	0.8	2.1	1.6	1.5	3.9
	$(1,2)$	1.9	1.9	4.6	20.0	19.6	17.2
500	$(3,3)$	11.4	10.3	12.9	9.2	9.2	10.8
	$(3,2)$	0.3	0.2	0.5	0.6	0.5	0.7
	$(3,1)$	0.0	0.0	0.3	0.0	0.0	0.7
	$(2,3)$	2.1	1.9	2.4	1.7	1.7	2.1
	$(1,3)$	1.4	1.4	1.7	3.1	2.9	2.5
	$(1,1)$	**90.6**	**91.1**	**80.2**	**79.1**	**79.3**	**73.5**
	$(2,2)$	0.8	0.7	1.7	0.3	0.3	1.5
	$(2,1)$	0.2	0.2	1.6	1.9	1.9	4.8
	$(1,2)$	1.7	1.5	4.2	11.9	11.8	10.4
2000	$(3,3)$	5.5	5.3	10.5	4.6	4.5	4.9
	$(3,2)$	0.2	0.2	0.1	0.2	0.2	0.4
	$(3,1)$	0.0	0.0	0.1	0.0	0.0	0.3
	$(2,3)$	0.4	0.4	0.1	0.6	0.6	1.4
	$(1,3)$	0.6	0.6	1.5	1.4	1.4	2.8

I: Strong VARMA$(1,1)$ model (16)-(10)

II: Weak VARMA$(1,1)$ model (16)-(12)

Table 12
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model I			Criteria Model II		
n	(p,q)	AIC	AICc	AIC_M	AIC	AICc	AIC_M
	$(1,1)$	**82.0**	**83.4**	**74.4**	**81.0**	**81.0**	**74.0**
	$(2,2)$	0.1	0.1	1.1	0.0	0.0	0.0
	$(2,1)$	0.8	0.8	2.1	3.0	3.0	6.0
	$(1,2)$	1.9	1.9	4.6	11.0	11.0	7.0
$5,000$	$(3,3)$	11.4	10.3	12.9	4.0	4.0	6.0
	$(3,2)$	0.3	0.2	0.5	0.0	0.0	2.0
	$(3,1)$	0.0	0.0	0.3	0.0	0.0	2.0
	$(2,3)$	2.1	1.9	2.4	0.0	0.0	1.0
	$(1,3)$	1.4	1.4	1.7	1.0	1.0	2.0
	$(1,1)$	**90.6**	**91.1**	**80.2**	**75.0**	**75.0**	**70.0**
	$(2,2)$	0.8	0.7	1.7	0.0	0.0	3.0
	$(2,1)$	0.2	0.2	1.6	0.0	0.0	5.0
	$(1,2)$	1.7	1.5	4.2	20.0	20.0	11.0
$10,000$	$(3,3)$	5.5	5.3	10.5	1.0	1.0	4.0
	$(3,2)$	0.2	0.2	0.1	0.0	0.0	1.0
	$(3,1)$	0.0	0.0	0.1	0.0	0.0	3.0
	$(2,3)$	0.4	0.4	0.1	1.0	1.0	1.0
	$(1,3)$	0.6	0.6	1.5	3.0	3.0	2.0

I: Strong VARMA$(1,1)$ model (16)-(10)

II: Weak VARMA$(1,1)$ model (16)-(12)

Table 13
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model III			Criteria Model IV		
n	(p, q)	AIC	AICc	AIC_M	AIC	AICc	AIC_M
	$(1, 1)$	**74.3**	**75.6**	**67.9**	**82.1**	**83.0**	**75.0**
	$(2, 2)$	0.3	0.3	1.0	0.3	0.3	1.0
	$(2, 1)$	0.8	0.8	2.9	0.3	0.2	1.4
	$(1, 2)$	8.4	8.2	11.4	1.2	1.1	4.3
500	$(3, 3)$	8.5	8.0	7.0	12.5	11.8	13.2
	$(3, 2)$	0.4	0.4	0.7	0.1	0.1	0.6
	$(3, 1)$	0.0	0.0	0.4	0.0	0.0	0.1
	$(2, 3)$	6.1	5.9	6.0	2.3	2.3	1.8
	$(1, 3)$	1.2	0.8	2.7	1.2	1.2	2.6
	$(1, 1)$	**84.0**	**84.2**	**73.4**	**90.9**	**90.9**	**87.3**
	$(2, 2)$	0.3	0.3	1.8	0.0	0.0	0.9
	$(2, 1)$	0.8	0.8	3.4	0.3	0.3	0.9
	$(1, 2)$	7.8	7.8	8.4	1.8	1.8	3.2
2000	$(3, 3)$	3.2	3.2	7.2	6.7	6.7	5.8
	$(3, 2)$	0.1	0.1	0.4	0.1	0.1	0.1
	$(3, 1)$	0.0	0.0	0.7	0.0	0.0	0.0
	$(2, 3)$	0.5	0.5	0.5	0.0	0.0	0.5
	$(1, 3)$	3.3	3.1	4.2	0.2	0.2	1.3

III: Weak VARMA$(1, 1)$ model GARCH (16)-(11)

IV: Weak VARMA$(1, 1)$ model (16)-(13)

Table 14
Relative frequency (in %) of the order selected by various standard and modified versions of the **AIC** criteria.

Length	Order	Criteria Model III			Criteria Model IV		
n	(p,q)	AIC	AICc	AIC_M	AIC	AICc	AIC_M
	$(1,1)$	**82.0**	**82.0**	**79.0**	**95.0**	**95.0**	**85.0**
	$(2,2)$	0.0	0.0	0.0	0.0	0.0	2.0
	$(2,1)$	0.0	0.0	1.0	0.0	0.0	1.0
	$(1,2)$	8.0	8.0	9.0	1.0	1.0	2.0
$5,000$	$(3,3)$	3.0	3.0	4.0	2.0	2.0	6.0
	$(3,2)$	0.0	0.0	1.0	0.0	0.0	0.0
	$(3,1)$	0.0	0.0	0.0	0.0	0.0	1.0
	$(2,3)$	0.0	0.0	1.0	0.0	0.0	0.0
	$(1,3)$	7.0	7.0	5.0	2.0	2.0	3.0
	$(1,1)$	**89.0**	**89.0**	**84.0**	**96.0**	**96.0**	**87.0**
	$(2,2)$	1.0	1.0	3.0	0.0	0.0	0.0
	$(2,1)$	0.0	0.0	1.0	0.0	0.0	1.0
	$(1,2)$	8.0	8.0	8.0	3.0	3.0	7.0
$10,000$	$(3,3)$	0.0	0.0	0.0	1.0	1.0	5.0
	$(3,2)$	0.0	0.0	0.0	0.0	0.0	0.0
	$(3,1)$	0.0	0.0	0.0	0.0	0.0	0.0
	$(2,3)$	0.0	0.0	1.0	0.0	0.0	0.0
	$(1,3)$	2.0	2.0	3.0	0.0	0.0	0.0

III: Weak VARMA$(1,1)$ model GARCH (16)-(11)

IV: Weak VARMA$(1,1)$ model (16)-(13)

Table 15
Modified version of asymptotic probabilities of overfitting by $\ell = d^2(\ell_1 + \ell_2)$ parameters for bivariate VARMA models of various versions of **AIC** criteria.

Length	Order	$\mathbf{P}_W$ Model I			$\mathbf{P}_W$ Model II		
n	(ℓ_1, ℓ_2)	$\mathbf{P}_W^{AIC}$	$\mathbf{P}_W^{AICc}$	$\mathbf{P}_W^{AIC_M}$	$\mathbf{P}_W^{AIC}$	$\mathbf{P}_W^{AICc}$	$\mathbf{P}_W^{AIC_M}$
	$(1,0)$	0.009	0.009	0.028	0.057	0.056	0.104
	$(0,1)$	0.025	0.024	0.060	0.237	0.231	0.228
	$(1,1)$	0.003	0.002	0.021	0.086	0.079	0.136
	$(0,2)$	0.016	0.016	0.036	0.130	0.117	0.139
500	$(2,0)$	0.002	0.002	0.011	0.016	0.016	0.057
	$(1,2)$	0.027	0.024	0.036	0.073	0.065	0.103
	$(2,1)$	0.006	0.005	0.011	0.043	0.039	0.076
	$(2,2)$	0.126	0.114	0.140	0.127	0.121	0.168
	$(1,0)$	0.007	0.007	0.026	0.083	0.081	0.124
	$(0,1)$	0.020	0.017	0.058	0.258	0.251	0.221
	$(1,1)$	0.009	0.009	0.022	0.103	0.102	0.133
	$(0,2)$	0.010	0.010	0.036	0.147	0.146	0.141
2000	$(2,0)$	0.001	0.001	0.006	0.026	0.025	0.069
	$(1,2)$	0.005	0.005	0.011	0.071	0.071	0.096
	$(2,1)$	0.005	0.005	0.011	0.053	0.052	0.082
	$(2,2)$	0.057	0.055	0.108	0.073	0.071	0.122

I: Strong VARMA$(1,1)$ model (16)-(10)

II: Weak VARMA$(1,1)$ model (16)-(12)