C. R. Acad. Sci. Paris, Ser. I 349 (2011) 695–698

Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistique/Économie mathématique

Estimation des ordres de modèles ARMA faibles multivariés

Estimating the orders of weak multivariate ARMA models

Yacouba Boubacar Mainassara

Université Lille III, EQUIPPE, BP 60 149, 59653 Villeneuve d'Ascq cedex, France

I N F O A R T I C L E

Historique de l'article :
Reçu le 7 août 2010
Accepté après révision le 19 avril 2011
Disponible sur Internet le 12 mai 2011

Présenté par Paul Deheuvels

R É S U M É

Dans cette Note, nous considérons le problème de sélection des ordres de modèles ARMA multivarié (VARMA) avec innovations linéaires non corrélées mais non nécessairement indépendantes. Ces modèles sont appelés VARMA faibles. Par opposition, nous appelons VARMA forts les modèles utilisés habituellement dans la littérature dans lesquels le terme d'erreur est supposé être un bruit iid. Cette sélection est fondée sur la minimisation d'un critère d'information, notamment celui introduit par Akaike (AIC pour Akaike's Information Criterion). Les fondements théoriques de ce critère AIC ne sont plus établis lorsque l'hypothèse de bruit iid est relâchée. Afin de remédier à ce problème, nous proposons un critère AIC modifié, et qui peut être très différent du critère AIC standard.

A B S T R A C T

In this Note, we consider the problem of order selection of vector autoregressive moving-average (VARMA) models under the assumption that the errors are uncorrelated, but not necessarily independent. These models are called weak VARMA by opposition to the standard VARMA models, also called strong VARMA models, in which the error terms are supposed to be iid. This selection is based on minimizing an information criterion, especially that introduced by Akaike. The theoretical foundations of the Akaike information criterion (AIC) are not more established when the iid assumption on the noise is relaxed. We propose a modified AIC criterion, and which may be very different from the standard AIC criterion.

1. Introduction

En statistique on est souvent confronté au problème de l'identification d'un modèle parmi m modèles, où les modèles candidats ont des paramètres θ_m de dimensions k_m. Le choix peut s'effectuer en estimant chaque modèle et en minimisant un critère de la forme : *mesure de l'erreur d'ajustement* + *terme de pénalisation*. On mesure souvent cette erreur d'ajustement par la somme des carrés des résidus, ou encore par -2 fois la log-quasi-vraisemblance. Le terme de pénalisation est une fonction croissante de la dimension k_m du paramètre $\hat{\theta}_m$. Ce terme est indispensable car la mesure de l'erreur d'ajustement est systématiquement minimale pour le modèle qui possède le plus grand nombre de paramètres, lorsque $\hat{\theta}_m$ minimise cette erreur d'ajustement et quand les modèles sont emboîtés.

Adresse e-mail : yacouba.boubacarmainassara@univ-lille3.fr.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2011.04.012

Dans l'étape d'identification de la traditionnelle méthodologie de Box et Jenkins, un des problèmes les plus délicats est celui de la sélection d'un petit nombre de valeurs plausibles pour les ordres p_0 et q_0 du modèle VARMA.

Parmi les méthodes d'identification, les plus populaires sont celles basées sur l'optimisation d'un critère d'information. Le critère le plus connu est sans doute le AIC introduit par Akaike [1]. Ce critère est fondé sur un estimateur de la divergence (ou information, ou entropie) de Kullback–Leibler. L'objectif de cette Note est d'étudier le problème de sélection des ordres p_0 et q_0 de modèles VARMA dont les termes d'erreur sont non corrélés mais peuvent contenir des dépendances non linéaires. Les fondements théoriques du critère AIC ne sont plus établis lorsque l'hypothèse de bruit iid est relâchée. Nous relâchons cette hypothèse standard d'indépendance pour étendre le champ d'application des modèles VARMA, ceci permettra aussi de couvrir une large classe de processus non linéaires. Ainsi nous introduisons les coefficients de mélange fort d'un processus vectoriel $Z = (Z_t)$ définis par $\alpha_Z(h) = \sup_{A\in\sigma(Z_u,\leqslant t),B\in\sigma(Z_u,\geqslant t+h)} |P(A\cap B) - P(A)P(B)|$.

2. Modèle et hypothèses

Soit $X_t = (X_{1t},\dots,X_{dt})'$ un processus stationnaire au second ordre, vérifiant

$$A_{00}X_t - \sum_{i=1}^{p_0} A_{0i}X_{t-i} = B_{00}\epsilon_t - \sum_{j=1}^{q_0} B_{0i}\epsilon_{t-j}, \quad \forall t\in\mathbb{Z} \tag{1}$$

où le terme d'erreur $\epsilon_t = (\epsilon_{1t},\dots,\epsilon_{dt})'$ est un bruit blanc faible, c'est-à-dire une suite de variables aléatoires centrées ($E\epsilon_t = 0$), non corrélées, avec une matrice de covariance non singulière Σ_0. Pour l'estimation des paramètres du modèle (1), nous utiliserons la paramétrisation ainsi que la méthode du quasi-maximum de vraisemblance (QMV) définies dans Boubacar Mainassara et Francq [2], noté BMF dans la suite. Ces auteurs ont établi la convergence forte et la normalité asymptotique de l'estimateur $\hat\theta_n$ du QMV, sous les hypothèses essentielles d'ergodicité et de mélange suivantes : **A1 :** Le processus (ϵ_t) est stationnaire et ergodique ; **A2 :** Il existe un réel $\nu > 0$ tel que $E\|\epsilon_t\|^{4+2\nu} < \infty$ et les coefficients de mélange du processus (ϵ_t) vérifient $\sum_{k=0}^{\infty}\{\alpha_\epsilon(k)\}^{\frac{\nu}{2+\nu}} < \infty$. Soit les polynômes $A_\theta(z) = A_0 - \sum_{i=1}^{p_0} A_i z^i$ et $B_\theta(z) = B_0 - \sum_{i=1}^{q_0} B_i z^i$. Nous ferons également les hypothèses **A3 :** $\det A(z)\det B(z) = 0 \Rightarrow |z| > 1$; **A4 :** Pour tout $\theta\in\Theta_{p,q}$ tel que $\theta\neq\theta_0$, soient les fonctions de transfert $A_0^{-1}B_0B_\theta^{-1}(z)A_\theta(z) \neq A_{00}^{-1}B_{00}B_{\theta_0}^{-1}(z)A_{\theta_0}(z)$ pour un $z\in\mathbb{C}$ ou alors $A_0^{-1}B_0\Sigma B_0'A_0^{-1'} \neq A_{00}^{-1}B_{00}\Sigma_0 B_{00}'A_{00}^{-1'}$; **A5 :** Nous avons $\theta_0 \in \overset{\circ}{\Theta}_{p_0,q_0}$, où $\overset{\circ}{\Theta}_{p_0,q_0}$ est l'intérieur du sous espace compact Θ_{p_0,q_0} de l'espace des paramètres.

3. Identification de modèles VARMA faibles

Soit $\Sigma_\epsilon := A_0^{-1}B_0\Sigma B_0'A_0^{-1'}$. Pour tout $\theta\in\Theta_{p,q}$, nous avons

$$-2\log \mathrm{L}_n(\theta) = nd\log(2\pi) + n\log\det\Sigma_\epsilon + \sum_{t=1}^{n}\epsilon_t'(\theta)\Sigma_\epsilon^{-1}\epsilon_t(\theta),$$

où $\epsilon_t(\theta) = A_0^{-1}B_0B_\theta^{-1}(L)A_\theta(L)X_t$ et où $\mathrm{L}_n(\theta)$ est la quasi-vraisemblance. Minimiser l'information de Kullback–Leibler pour tout modèle candidat caractérisé par le paramètre θ, revient à minimiser l'écart (souvent appelé le contraste) entre le modèle ajusté et le vrai modèle $\Delta(\theta) := E\{-2\log \mathrm{L}_n(\theta)\}$. En omettant la constante $nd\log(2\pi)$, nous trouvons que $\Delta(\theta) = n\log\det\Sigma_\epsilon + n\,\mathrm{Tr}(\Sigma_\epsilon^{-1}S(\theta))$, où $S(\theta) = E\epsilon_1(\theta)\epsilon_1'(\theta)$. Nous énonçons le lemme suivant, qui montre que l'application $\theta\mapsto\Delta(\theta)$ est minimale en $\theta = \theta_0$:

Lemme 1. *Pour tout $\theta\in\bigcup_{p,q\in\mathbb{N}}\Theta_{p,q}$, nous avons $\Delta(\theta)\geqslant\Delta(\theta_0)$.*

Soit $X = (X_1,\dots,X_n)$ les observations satisfaisant la représentation (1). Posons $\hat\epsilon_t = \epsilon_t(\hat\theta_n)$ les résidus de l'estimateur du QMV. En vue du Lemme 1, il est naturel de minimiser le contraste moyen $E\Delta(\hat\theta_n)$. L'écart $\{\Delta(\hat\theta_n) - \Delta(\theta_0)\}$ s'interprète comme une perte de précision globale moyenne quand on utilise le modèle estimé à la place du vrai modèle.

3.1. Estimation du contraste moyen

Nous allons dans cette section adapter aux modèles VARMA faibles, le critère d'information de Akaike corrigé (noté AICc) introduit et étendu au cadre de modèles VAR forts par Hurvich et Tsai [4,5], pour obtenir un estimateur approximativement sans biais de $E\Delta(\hat\theta_n)$. Notons que, pour les modèles VARMA sous la forme réduite, il n'est pas restrictif de supposer que les coefficients $A_0,\dots,A_{p_0},B_0,\dots,B_{q_0}$ sont fonctionnellement indépendants du coefficient Σ_ϵ. Par conséquent, nous écrivons $\theta = (\theta^{(1)'},\theta^{(2)'})'$, où $\theta^{(1)}\in\mathbb{R}^{k_1}$ dépend des $A_0,\dots,A_p$ et $B_0,\dots,B_q$, et où $\theta^{(2)}\in\mathbb{R}^{k_2}$ dépend uniquement de Σ_ϵ, avec $k_1+k_2 = k_0$. Posons $\hat\Sigma_\epsilon := n^{-1}\sum_{t=1}^{n}\hat\epsilon_t\hat\epsilon_t'$. Nous avons

$$E\Delta(\hat\theta_n) = En\log\det\hat\Sigma_\epsilon + nE\,\mathrm{Tr}\big(\hat\Sigma_\epsilon^{-1}S(\hat\theta_n)\big). \tag{2}$$

Le premier terme du second membre de l'équation (2) peut être estimé sans biais par $n\log\det\{n^{-1}\sum_{t=1}^{n}\hat{\epsilon}_t\hat{\epsilon}_t'\}$. Par conséquent, nous prenons uniquement en considération l'estimation du deuxième terme. En outre, en vue du Théorème 1 dans BMF, en faisant un développement de Taylor de $\epsilon_t(\theta)$ au voisinage de $\theta_0^{(1)}$ et en utilisant les propriétés d'orthogonalité entre $\epsilon_t(\theta_0)$ et toute combinaison linéaire des valeurs passées de $\epsilon_t(\theta_0)$ (en particulier $\partial\epsilon_t(\theta_0)/\partial\theta'$ et $\partial^2\epsilon_t(\theta_0)/\partial\theta\partial\theta'$), et le fait que les innovations sont centrées *i.e.* $E\epsilon_t(\theta_0)=0$, nous avons

$$S(\theta)=\Sigma_{\epsilon 0}+D\big(\theta^{(1)}\big)+O\big(\pi^4\big),\quad \text{où}\quad \Sigma_{\epsilon 0}:=A_{00}^{-1}B_{00}\Sigma_0B_{00}'A_{00}^{-1'}\quad\text{avec}$$
$$\pi=\big\|\theta^{(1)}-\theta_0^{(1)}\big\|,\quad \text{et } D\big(\theta^{(1)}\big)=E\left\{\frac{\partial\epsilon_t(\theta_0)}{\partial\theta^{(1)'}}\big(\theta^{(1)}-\theta_0^{(1)}\big)\big(\theta^{(1)}-\theta_0^{(1)}\big)'\frac{\partial\epsilon_t'(\theta_0)}{\partial\theta^{(1)}}\right\}.$$

Nous pouvons donc écrire l'écart moyen (2) comme

$$E\Delta(\hat{\theta}_n)=En\log\det\hat{\Sigma}_\epsilon+nE\,\mathrm{Tr}\big(\hat{\Sigma}_\epsilon^{-1}\Sigma_{\epsilon 0}\big)+nE\,\mathrm{Tr}\big(\hat{\Sigma}_\epsilon^{-1}D\big(\hat{\theta}_n^{(1)}\big)\big)+R_n,$$

où $R_n=nE\,\mathrm{Tr}(\hat{\Sigma}_\epsilon^{-1})O_P(n^{-2})$. En utilisant, comme dans une régression multivariée classique, la relation

$$\Sigma_{\epsilon 0}\approx n/\big\{n-d(p+q)\big\}E\{\hat{\Sigma}_\epsilon\}=dn/(dn-k_1)E\{\hat{\Sigma}_\epsilon\},$$

et la consistence de $\hat{\Sigma}_\epsilon$, nous déduisons

$$E\big\{\hat{\Sigma}_\epsilon^{-1}\big\}\approx\{E\hat{\Sigma}_\epsilon\}^{-1}\approx nd(nd-k_1)^{-1}\Sigma_{\epsilon 0}^{-1}. \tag{3}$$

En utilisant des propriétés élémentaires sur la trace d'une matrice, nous avons

$$\mathrm{Tr}\big\{\Sigma_\epsilon^{-1}(\theta)D\big(\theta_n^{(1)}\big)\big\}=\mathrm{Tr}\left(E\left\{\frac{\partial\epsilon_t'(\theta_0)}{\partial\theta^{(1)}}\Sigma_\epsilon^{-1}(\theta)\frac{\partial\epsilon_t(\theta_0)}{\partial\theta^{(1)'}}\right\}\big(\theta^{(1)}-\theta_0^{(1)}\big)'\big(\theta^{(1)}-\theta_0^{(1)}\big)\right).$$

Maintenant, de cette dernière égalité appliquée en $\hat{\theta}_n$, en vue du Théorème 1 dans BMF et (3), nous avons

$$E\,\mathrm{Tr}\big(\hat{\Sigma}_\epsilon^{-1}S(\hat{\theta}_n)\big)=\frac{nd^2}{nd-k_1}+\frac{d}{2(nd-k_1)}\mathrm{Tr}\big(I_{11}J_{11}^{-1}\big)+O\left(\frac{1}{n^2}\right), \tag{4}$$

où $J_{11}=2E\{\partial\epsilon_t'(\theta_0)/\partial\theta^{(1)}\Sigma_{\epsilon 0}^{-1}\partial\epsilon_t(\theta_0)/\partial\theta^{(1)'}\}$ (voir le Théorème 2 dans BMF) et où I_{11} est le premier bloc supérieur gauche de la matrice I impliquée dans la variance asymptotique de l'estimateur du QMV. En utilisant (4) dans (2), sous les hypothèses précédentes, nous déduisons un estimateur approximativement sans biais de $E\Delta(\hat{\theta}_n)$ donné par

$$\mathrm{AIC}_M:=n\log\det\hat{\Sigma}_\epsilon+\frac{n^2d^2}{nd-k_1}+\frac{nd}{2(nd-k_1)}\mathrm{Tr}\big(\hat{I}_{11,n}\hat{J}_{11,n}^{-1}\big) \tag{5}$$

où les matrices $\hat{I}_{11,n}$ et $\hat{J}_{11,n}$ sont des estimateurs convergents des matrices I_{11} et J_{11}. Notons que l'estimation des matrices I_{11} et J_{11} est plus difficile et moins précise quand k_1 est très grand.

Nous obtenons des estimateurs $\hat{p}$ et $\hat{q}$ des ordres p_0 et q_0 en minimisant le critère modifié (5).

Dans le cas de modèles VARMA forts c'est-à-dire quand l'hypothèse d'ergodicité **A1** est remplacée par celle dont les termes d'erreur sont iid, nous avons $I_{11}=2J_{11}$ (voir remarque 3 dans BMF), alors $\mathrm{Tr}(I_{11}J_{11}^{-1})=2k_1$. Ce qui donne au critère AIC_M une forme plus conventionnelle

$$\mathrm{AIC}_M^*:=n\log\det\hat{\Sigma}_\epsilon+nd+\frac{nd}{nd-k_1}2k_1=\mathrm{AICc}.$$

3.2. Autre décomposition du contraste moyen

Dans la section précédente le contraste minimal a été approximé par $-2E\log\mathrm{L}_n(\hat{\theta}_n)$ (l'espérance est prise avec l'observation du vrai modèle X). Notons que l'étude de cette moyenne de l'écart est trop difficile en raison de la dépendance entre l'estimateur $\hat{\theta}_n$ et l'observation X. Une méthode alternative (légèrement différente de celle de la section précédente mais équivalente en interprétation) pour arriver à la quantité $E\Delta(\hat{\theta}_n)$ consiste à considérer $\hat{\theta}_n$ comme étant l'estimateur du QMV de θ fondé sur l'observation X. Soit $Y=(Y_1,\ldots,Y_n)$ l'observation indépendante de X et générée par le même modèle (1). Nous approximons la distribution de (Y_t) par $\mathrm{L}_n(Y,\hat{\theta}_n)$. Nous considérons donc l'écart moyen du modèle ajusté (modèle candidat Y) en $\hat{\theta}_n$. Ainsi, il est généralement plus facile de chercher un modèle qui minimise

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

où E_Y est l'espérance de l'observation Y du modèle candidat. Puisque $\hat{\theta}_n$ et Y sont indépendants, $C(\hat{\theta}_n)$ est la même quantité que l'écart moyen $E\Delta(\hat{\theta}_n)$. Le modèle minimisant (6) peut être interprété comme le modèle qui aura une meilleure approximation globale sur une copie indépendante d'observations de X, mais ce modèle peut ne pas être le meilleur pour les

données à portée de main. Cette moyenne du contraste peut être décomposée comme suit $C(\hat{\theta}_n) = -2E_X \log \mathrm{L}_n(\hat{\theta}_n) + a_1 + a_2$, où $a_1 = -2E_X \log \mathrm{L}_n(\theta_0) + 2E_X \log \mathrm{L}_n(\hat{\theta}_n)$ et $a_2 = -2E_Y \log \mathrm{L}_n(\hat{\theta}_n) + 2E_X \log \mathrm{L}_n(\theta_0)$. L'estimateur du QMV satisfait $\log \mathrm{L}_n(\hat{\theta}_n) \geqslant \log \mathrm{L}_n(\theta_0)$ presque sûrement, c'est-à-dire un ajustement anormal des données au modèle, quand ce sont ces mêmes données qui ont ajusté le modèle. Alors, le terme a_1 s'interprète comme *le sur-ajustement moyen* de cet estimateur du QMV. Notons que $E_X \log \mathrm{L}_n(\theta_0) = E_Y \log \mathrm{L}_n(\theta_0)$, donc le terme a_2 s'interprète comme *le coût moyen d'utilisation* (sur replications indépendantes de l'observation X) du modèle estimé à la place du vrai modèle. La proposition suivante montre que les termes a_1 et a_2 sont équivalents :

Proposition 1. *Sous les hypothèses ci-dessus, quand $n \to \infty$, nous avons a_1 (sur-ajustement moyen) et a_2 (perte de précision sur replications indépendantes) sont tous deux égaux au nombre $2^{-1} \operatorname{Tr}(I_{11} J_{11}^{-1})$.*

En vue de cette proposition 1, dans le cas de modèles VARMA faibles, nous pouvons approximer le moyen $C(\hat{\theta}_n)$ par le critère modifié

$$\mathrm{AIC}^* := -2 \log \mathrm{L}_n(\hat{\theta}_n) + \operatorname{Tr}\bigl(\hat{I}_{11} \hat{J}_{11}^{-1}\bigr). \tag{7}$$

Nous obtenons des estimateurs $\hat{p}$ et $\hat{q}$ des ordres p_0 et q_0 en minimisant le critère modifié (7).

Dans le cas de modèles VARMA forts, en vue de la proposition 1, nous obtenons que a_1 et a_2 sont tous deux égaux à $k_1 = \dim(\theta_0^{(1)})$. Ainsi, nous obtenons les mêmes résultats montrés dans Findley [3]. Ceci permet d'approximer le contraste moyen (6) par le critère $\mathrm{AIC} = -2 \log \mathrm{L}_n(\hat{\theta}_n) + 2k_1$.

Références

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csáki (Eds.), 2nd International Symposium on Information Theory, Akadémia Kiado, Budapest, 1973, pp. 267–281.
[2] Y. Boubacar Mainassara, C. Francq, Estimating structural VARMA models with uncorrelated but non-independent error terms, Journal of Multivariate Analysis 102 (2011) 496–505.
[3] D.F. Findley, The overfitting principles supporting AIC, Statistical Research Division Report RR 93/04, Bureau of the Census, 1993.
[4] C.M. Hurvich, C.-L. Tsai, Regression and time series model selection in small samples, Biometrika 76 (1989) 297–307.
[5] C.M. Hurvich, C.-L. Tsai, A corrected Akaïke information criterion for vector autoregressive model selection, Journal of Time Series Analysis 14 (1993) 271–279.