Technique de rééchantillonnage et estimation de l'ordre d'un modèle ARMA avec des données incomplètes

Abdelaziz El Matouat, Hassania Hamzaoui, Freedath Djibril Moussa

Technique de rééchantillonnage et estimation de l'ordre d'un modèle arma avec des données incomplètes

Abdelaziz El Matouat, Hassania Hamzaoui & Freedath Djibril Moussa

CERENE, Université du Havre, France
FLSH, Université de Fès, Maroc
FSDM, Université de Fès, Maroc

Résumé
Dans ce travail, nous nous interéssons à l'estimation de l'ordre d'un modèle ARMA relatif à des données incomplètes par les critères d'information. Nous avons utilisé la technique de rééchantillonnage pour améliorer la performance de ces critères. La qualité de notre démarche a été validée par une application à des échantillons simulés.

Abstract
In this paper, we are interested to estimate an ARMA model order from incomplete data by information criteria. We use a resampling scheme to improve the performances of those criteria. The quality of our approach is validated by application to simulated samples.

Mots clés: Critères d'information, rééchantillonnage, modèles ARMA, facteur de pénalisation, données incomplètes.

1 Introduction

Soit $\underline{X}$ une série d'observations de longueur n d'un modèle $ARMA(p,q)$ causal et inversible:

$$X_t - \phi_1 X_{t-1} \cdots - \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

où $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ sont les coefficients du modèle et $\{\varepsilon_t\}$ un bruit blanc de variance $\sigma^2(p,q)$. Les critères d'information permettent d'estimer l'ordre (p,q) du modèle en minimisant la quantité $IC(r,s,C_n) = \log(\hat{\sigma}^2(r,s)) + (r+s)C_n$, où $\hat{\sigma}^2(r,s)$ désigne une estimation de $\sigma^2(r,s)$ lorsque le modèle candidat est un $ARMA(r,s)$ et C_n un facteur de pénalisation. Les critères usuels sont le critrère AIC, $(C_n = \frac{2}{n})$ (1973), le critère BIC, $(C_n = \frac{\log(n)}{n})$ (1978) et le critère φ_β, $(C_n = \frac{n^\beta \log(\log(n))}{n})$, $0 < \beta < 1$ (1996). Le critère AIC conduit asymptotiquement à une surparamétrisation du modèle, et les critères BIC et φ_β assurent une estimation convergente de l'ordre. Mais l'application de ces critères peut ne pas aboutir à une sélection satisfaisante de l'ordre dans le cas des petits échantillons. Pour pallier cette difficulté, Chen et al. (1993) ont utilisé la technique de rééchantillonnage pour un modèle autorégressif avec le critère BIC; cette technique a également été étendue aux modèles $ARMA$ avec les critères BIC et φ_β par Ikama (1999).

Dans ce travail, nous nous intéressons à l'estimation de l'ordre d'un modèle $ARMA$ à partir de données incomplètes par la technique de rééchantillonnage. Pour un échantillon incomplet, la forme des critères d'information est la suivante:

$$IC_{cd}(r,s,C_n) = \log(\hat{\sigma}^2(r,s)) + \{(r+s) + trace\left[DF(I-DF)^{-1}\right]\}C_n$$

où I désigne la matrice identité, DF, la matrice jacobienne de l'application définie par l'algorithme EM, $\hat{\sigma}^2(r,s)$, une estimation de $\sigma^2(r,s)$ déterminée par l'algorithme EM, et C_n l'un des facteurs de pénalisation définis précédemment. Nous désignons par AIC_{cd}, BIC_{cd} et $\varphi_{\beta cd}$ les critères correspondants. Les facteurs de pénalisation de ces différents critères ne dépendent que de la taille de l'échantillon et de la fonction définie par l'algorithme EM. Nous reformulons les critères AIC_{cd}, BIC_{cd}, $\varphi_{\beta cd}$ en définissant des facteurs de pénalisation, qui intègrent la structure du modèle, par la technique de rééchantillonnage.

2 Estimation des paramètres d'un modèle $ARMA$ pour des données incomplètes

Considérons une série d'observations complète $\underline{X} = (X_1, \ldots, X_n)$ d'un modèle $ARMA(p,q)$. Les estimations $\hat{\phi}_1, \ldots, \hat{\phi}_p$, $\hat{\theta}_1, \ldots, \hat{\theta}_q$ des coefficients et $\hat{\sigma}^2(p,q)$ de la variance sont obtenues par maximum de vraisemblance.
Lorsque les données sont incomplètes, nous notons $\underline{X} = (\underline{X}_{obs}, \underline{X}_{mis})$ où $\underline{X}_{obs}$ désigne la partie observée et $\underline{X}_{mis}$ la partie manquante. Les paramètres sont dans ce cas estimés par l'algorithme EM qui se décompose en deux étapes :

- l'étape E (Expectation) : calcul de l'espérance conditionnelle de la log-vraisemblance de l'échantillon $\underline{X}$ sachant la partie observée $\underline{X}_{obs}$.
- l'étape M (Maximization) : maximisation de l'espérance calculée à l'étape E.

Nous présentons maintenant la technique de rééchantillonnage relativement à un modèle $ARMA$ avec des données incomplètes.

3 Critères de sélection de l'ordre et technique de rééchantillonnage

Soit un échantillon incomplet $\underline{X} = (\underline{X}_{obs}, \underline{X}_{mis})$ d'un modèle $ARMA$ d'ordre (p,q) inconnu. Lorsque la taille n de l'échantillon est faible, la sélection de l'ordre par les critères d'information peut ne pas être satisfaisante. De plus, les pénalisations de la log-vraisemblance ne tiennent pas compte de l'information que pourrait apporter l'échantillon. Nous proposons d'utiliser la technique de rééchantillonnage pour définir une nouvelle pénalisation dans les critères afin d'améliorer leurs performances. Cette technique consiste à simuler des échantillons de taille N, $N >> n$, d'un modèle $ARMA$, pour estimer les paramètres et définir une nouvelle pénalisation dans la formulation des critères AIC_{cd}, BIC_{cd} et $\varphi_{\beta cd}$.

Proposition 3.1 *Soient $\underline{Y}$ un échantillon de taille N d'un modèle $ARMA(p,q)$, $K > p$ et $L > q$ des entiers fixés. On pose:*

$$a_N = \begin{cases} \max\limits_{\substack{p<r\leq K \\ q<s\leq L}} \frac{\log\hat{\sigma}^2(p,q)-\log\hat{\sigma}^2(r,s)}{(r+s)-(p+q)} & si\ p+q<K+L \\ 0 & sinon \end{cases}, \text{ et } b_N = \begin{cases} \infty & si\ p=q=0 \\ \min\limits_{\substack{0\leq r<p \\ 0\leq s<q}} \frac{\log\hat{\sigma}^2(r,s)-\log\hat{\sigma}^2(p,q)}{(p+q)-(r+s)} & si\ p+q>0, \end{cases}$$

Alors, lorsque $N \to \infty$

1. $a_N \to 0$ *et* $b_N \to b \geq 0$ *en probabilité.*
2. *Si* $a_N \leq b_N$, *alors pour tout* $C_{Nr} \in [a_N, b_N]$, $IC_{cd}(p,q,C_{Nr}) = \min\limits_{\substack{0\leq r\leq K \\ 0\leq s\leq L}} IC_{cd}(r,s,C_{Nr})$.

Cette proposition montre que si l'ordre (p,q) est connu, il réalise le minimum des critères $IC_{cd}(r,s,C_{Nr})$. Lorsque l'ordre est inconnu, on applique la technique de rééchantillonnage pour déterminer un facteur de pénalisation adéquat. Il s'agit de construire une suite d'ensembles non vides admissibles pour le facteur C_{Nr}. On commence par simuler des séries d'observations d'un $ARMA(k,l)$ pour $k = 0,\ldots,K_1$ et $l = 0,\ldots,L_1$, avec $K_1 < K$ et $L_1 < L$ deux entiers fixés. On détermine ensuite pour chaque couple (k,l) un intervalle admissible $I_N^{(k,l)} = [a_N^{(k,l)}, b_N^{(k,l)}]$ pour le facteur de pénalisation, par application de la proposition (3.1).

Proposition 3.2 *Soient* $(X_1,\cdots,X_n)$ *un échantillon relatif à un modèle ARMA causal d'ordre inconnu* (p,q), $K(>p)$ *et* $L(>q)$ *deux entiers. Soit* $\hat{\sigma}^2(K,L)$ *l'estimateur de la variance du bruit blanc du modèle* $ARMA(K,L)$. *Soient* $K_1 \leq p$ *et* $L_1 \leq q$ *deux entiers positifs. Pour* $k = 0,\ldots,K_1$ *et* $l = 0,\ldots,L_1$, *on considère* $\underline{Y}^{(k,l)} = (Y_1^{(k,l)},\ldots,Y_N^{(k,l)})$, $(N > n)$ *une série d'observations du modèle* $ARMA(k,l)$ *défini par:*

$$Y_t^{(k,l)} - \hat{\phi}_{k1}Y_{t-1}^{(k,l)} - \cdots - \hat{\phi}_{kk}Y_{t-k}^{(k,l)} = Z_t^* + \hat{\theta}_{l1}Z_{t-1}^* + \cdots + \hat{\theta}_{ll}Z_{t-l}^*$$

où $\hat{\phi}_{k1},\ldots,\hat{\phi}_{kk}$, $\hat{\theta}_{l1},\ldots,\hat{\theta}_{ll}$ *sont les paramètres estimés du modèle et où* $\{Z_t^*\}$ *est un bruit blanc de variance* $\hat{\sigma}^2(K,L)$. *Soient* $I_N^{(k,l)} = [a_N^{(k,l)}, b_N^{(k,l)}]$ *les intervalles obtenus en appliquant la proposition 3.1 (avec* $p=k, q=l$*) à la série* $\underline{Y}^{(k,l)}$. *Alors:*

1. $a_N = \max\limits_{\substack{0\leq k\leq K_1 \\ 0\leq l\leq L_1}} a_N^{(k,l)} \xrightarrow{P_n} 0$;
2. $b_N = \min\limits_{\substack{0\leq k\leq K_1 \\ 0\leq l\leq L_1}} b_N^{(k,l)} \xrightarrow{P_n} b \geq 0$;

   où $\xrightarrow{P_n}$ *désigne la convergence en probabilité conditionnellement à* $(X_1,\cdots,X_n)$.
3. *Si* $a_N \leq b_N$, *alors* $I_N = \bigcap\limits_{\substack{0\leq k\leq K_1 \\ 0\leq l\leq L_1}} I_N^{(k,l)}$ *converge en probabilité vers un ensemble non vide.*

Cette proposition nous permet de choisir une pénalisation C_{Nr} puisque l'intervalle I_N est non vide pour N suffisamment grand.

3.1 Forme des critères d'information

Avec la pénalisation C_{Nr}, nous aboutissons à la formulation suivante pour les critères de sélection de l'ordre d'un modèle $ARMA$ à partir d'un échantillon incomplet $\underline{X}$:

$$IC_{cdr}(r,s) = IC_{cd}(r,s,C_{Nr}) = \log\hat{\sigma}^2(r,s) + \{(r+s) + trace\left[DF(I-DF)^{-1}\right]\}C_{Nr}$$

où $C_{Nr} = \begin{cases} a_N + cb_N C_n & \text{si } a_N < b_N \\ b_N C_n & \text{sinon} \end{cases}$
avec c une constante telle que $C_{Nr} \in [a_N, b_N]$, si $a_N < b_N$ et C_n le facteur de pénalisation ordinaire.

3.2 Procédure de rééchantillonnage pour l'estimation de l'ordre d'un ARMA à partir de données incomplètes

Soit $\underline{X} = (\underline{X}_{obs}, \underline{X}_{mis})$ un échantillon incomplet d'un $ARMA(p,q)$ causal et inversible. Soient deux entiers $K > p$ et $L > q$.

Etape 1 Choisir des valeurs initiales $\phi_{K1}^{(0)}, \ldots, \phi_{KK}^{(0)}, \theta_{L1}^{(0)}, \ldots, \theta_{LL}^{(0)}$ pour les coefficients et $\sigma^{2(0)}(K,L)$ pour la variance, et construire un échantillon complété $\widetilde{X}$ par l'algorithme EM.

Etape 2 Avec $\widetilde{X}$, estimer les modèles $ARMA(k,l)$ pour $k = 1, ..., K$ et $l = 1, \ldots, L$.

Etape 3 Fixer deux entiers $K_1 \leq K$ et $L_1 \leq L$. Pour $k = 1, ..., K_1$ et $l = 1, \ldots, L_1$, générer un échantillon $\underline{Y}^{(k,l)} = (Y_1^{(k,l)}, \ldots, Y_N^{(k,l)})$ de l'$ARMA(k,l)$ de coefficients $\hat{\phi}_{k1}, \ldots, \hat{\phi}_{kk}$, $\hat{\theta}_{l1}, \ldots, \hat{\theta}_{ll}$ et de variance $\hat{\sigma}^2(K,L)$.

Etape 4 Pour $k = 1, \ldots, K_1$, $l = 1, \ldots, L_1$ supposer que la série $\underline{Y}^{(k,l)}$ est issue d'un $ARMA(r,s)$ et calculer l'estimation $\hat{\sigma}^2_{k,l}(r,s)$ de la variance, pour $r = 1, \ldots, K$, $s = 1, \ldots, L$.

Etape 5 Pour $k = 1, ..., K_1$, pour $l = 1, \ldots, L_1$ calculer:

$$a_N^{(k,l)} = \begin{cases} \max\limits_{\substack{k<r\leq K \\ l<s\leq L}} \frac{\log(\hat{\sigma}^2_{k,l}(k,l)) - \log(\hat{\sigma}^2_{k,l}(r,s))}{(r+s)-(k+l)} & \text{si } k+l < K+L \\ 0 & \text{sinon} \end{cases} \quad \text{et}$$

$$b_N^{(k,l)} = \begin{cases} \infty & \text{si } k = l = 0 \\ \min\limits_{\substack{0\leq r\leq K_1 \\ 0\leq s\leq L_1}} \frac{\log(\hat{\sigma}^2_{k,l}(r,s)) - \log(\hat{\sigma}^2_{k,l}(k,l))}{(k+l)-(r+s)} & \text{si } k+l > 0. \end{cases}$$

Etape 6 Calculer $a_N = \max\limits_{\substack{0\leq k\leq K_1 \\ 0\leq l\leq L_1}} a_N^{(k,l)}$, $b_N = \min\limits_{\substack{0\leq k\leq K_1 \\ 0\leq l\leq L_1}} b_N^{(k,l)}$ et C_{Nr}.

Etape 7 Calculer les critères $IC_{cdr}(r,s)$, $r = 1, \ldots, K$, $s = 1, \ldots, L$ et déterminer les estimations $\hat{p}$ et $\hat{q}$.

4 Simulations

Pour étudier le comportement des critères AIC_{cdr}, BIC_{cdr} et $\varphi_{\beta cdr}$, nous simulons 100 échantillons de taille $n = 100$ du modèle $ARMA(2, 2)$ causal et inversible suivant:

$$X_t = 0.9X_{t-1} - 0.2X_{t-2} + \varepsilon_t - 0.7\varepsilon_{t-1} + 0.1\varepsilon_{t-2}$$

où$\{\varepsilon_t\}$ est un bruit blanc de variance 1. A partir de cette simulation, nous construisons des échantillons incomplets avec des proportions de valeurs manquantes $P_{mis} = 0,\ 0.2$ et 0.4. Nous estimons ensuite l'ordre du modèle par les nouveaux critères, les ordres varient de 1 à 8 pour p et q. Dans la procédure de rééchantillonnage, nous considérons des échantillons Y de taille $N = 1000$, $K = L = 8$, et $K_1 = L_1 = 5$. La pénalisation C_{Nr} est calculée avec $c = 1 - \frac{a_N}{b_N}$. Nous avons récapitulé les résultats obtenus dans des tableaux (voir page 6).
Nous avons également traité le cas où la taille des échantillons simulés est $n = 40$ et nous avons abouti aux mêmes résultats avec des fréquences d'estimation de l'ordre exact inférieures à celles du cas où $n = 100$.

Bibliographie

[1] Akaike, H.(1973) Information theory and an extension of the maximum likelihood principle, *Second international symposium of information entropy theory*, ed B. N. Petrov and F. Csaki, Akademia Kiado, Budapest, 267–281.
[2] Cavanaugh, J. E. et Shumway (1998) An Akaike Information Criterion for model selection in the presence of incomplete data, *Journal of statistical planning and inference*, 67, 45–65.
[3] Chen, C., Davis, A. R., Brockwell, P. J. et Bai, Z. D. (1993) Order determination for autoregressive processes using resampling methods, *Statistica sinica*, 3, 481–500.
[4] Dempster, A. P., Laird, N. M. et Rubin D. (1977) Maximum Likelihood from Icomplete Data via EM algorithm, *Journal of the Royal Statistical Society*, Series B, 39, n° 1, 1–38.
[5] El Matouat, A. et Hallin, M. (1996) Order selection, stochastic complexity and Kullback-Leibler information, *Times Series Analysis, Springer Verlag*, 11, 291–299.
[6] Hamzaoui, H., El Matouat, A. et Jarrar, A. (2002) Critère de sélection SIC pour des données incomplètes, *Actes des journées de la SFDS*, 244.
[7] Ikama, A. (1999) Technique de Rééchantillonnage et estimation de l'ordre pour des séries temporelles, *Thèse du Diplôme d'Etudes Supérieures*, Université de Fès.
[8] Meng, X. L. et Rubin, D. B. (1991) Using EM algorithm to obtain Asymptotic Variance-Covariance Matrices: the SEM algorithm, *Journal of the American Statistical Association*, 86, 899–909.
[9] Nishii, R. (1988) Maximum likelihood principle and model selection when the true model is unspecified, *Journal of the multivariate analysis*, 27, 392–403.
[10] Schwarz, G. (1978) Estimating the dimension of a model, *The annals of statistics*, 6, 461–464.

	AIC_{cd}					BIC_{cd}					$\varphi_{\beta cd}$				
$P_{mis}=0$	1	2	3	4	5-8	1	2	3	4	5-8	1	2	3	4	5-8
1	8	4	2	0	1	2	5	4	0	0	3	5	1	1	0
2	3	51	2	1	1	1	69	1	2	1	1	72	2	0	0
3	1	4	5	1	0	4	3	0	0	0	2	4	0	1	0
4	2	2	0	0	1	3	2	1	0	0	1	1	1	1	0
5-8	4	3	3	1	0	1	0	0	1	0	2	1	1	0	0
	AIC_{cdr}					BIC_{cdr}					$\varphi_{\beta cdr}$				
$P_{mis}=0$	1	2	3	4	5-8	1	2	3	4	5-8	1	2	3	4	5-8
1	2	1	1	0	1	1	3	1	0	0	0	2	1	1	0
2	1	59	2	1	0	3	77	2	1	0	5	78	2	1	0
3	3	2	4	0	2	4	1	0	0	0	4	1	2	0	0
4	1	5	1	4	0	1	2	1	0	0	1	1	0	0	0
5-8	4	1	3	1	1	2	0	0	1	0	1	0	0	0	0
	AIC_{cd}					BIC_{cd}					$\varphi_{\beta cd}$				
$P_{mis}=0.2$	1	2	3	4	5-8	1	2	3	4	5-8	1	2	3	4	5-8
1	3	4	5	5	3	1	5	2	1	1	4	6	3	1	0
2	3	39	3	0	3	4	54	3	0	0	5	57	2	0	0
3	8	9	4	1	1	3	6	2	1	1	2	7	4	0	0
4	2	2	2	0	1	5	5	1	0	0	3	1	2	0	0
5-8	1	0	0	0	2	2	0	3	0	0	2	0	1	0	0
	AIC_{cdr}					BIC_{cdr}					$\varphi_{\beta cdr}$				
$P_{mis}=0.2$	1	2	3	4	5-8	1	2	3	4	5-8	1	2	3	4	5-8
1	4	5	2	1	2	5	4	4	0	0	7	5	1	1	0
2	3	46	4	1	0	11	63	2	0	0	9	64	2	0	0
3	7	7	6	0	1	2	2	3	0	0	4	3	1	0	0
4	2	2	3	0	0	1	0	1	0	0	2	0	0	0	0
5-8	2	1	0	0	1	0	1	0	1	0	0	1	0	0	0
	AIC_{cd}					BIC_{cd}					$\varphi_{\beta cd}$				
$P_{mis}=0.4$	1	2	3	4	5-8	1	2	3	4	5-8	1	2	3	4	5-8
1	2	5	8	2	1	6	3	5	1	0	5	12	3	1	0
2	4	26	9	2	0	11	35	4	2	1	10	40	4	0	0
3	10	6	5	0	2	8	7	3	0	0	5	6	3	1	0
4	5	3	2	1	1	4	4	2	2	0	3	2	1	1	0
5-8	3	1	1	0	1	2	1	0	0	0	1	0	2	0	0
	AIC_{cdr}					BIC_{cdr}					$\varphi_{\beta cdr}$				
$P_{mis}=0.4$	1	2	3	4	5-8	1	2	3	4	5-8	1	2	3	4	5-8
1	4	5	5	3	2	5	10	5	2	0	6	8	1	2	0
2	5	33	6	1	1	8	44	4	1	0	9	47	4	1	1
3	9	4	3	0	2	6	5	2	0	0	3	3	5	2	0
4	5	2	3	1	0	3	1	1	0	0	1	1	3	0	0
5-8	2	1	2	0	1	2	0	1	0	0	1	1	0	1	0

Fréquences des ordres estimés pour $n = 100$