UNE NOUVELLE APPROCHE DU CALCUL DE L'ESTIMATEUR DE MAXIMUM DE VRAISEMBLANCE EXACT D'UN MODELE AUTOREGRESSIF

Serge DEGERINE - Tuan Dinh PHAM

TIM 3 (Techniques de l'Informatique, des Mathématiques, de la Microélectronique et de la Microscopie quantitative)
Unité associée au CNRS n° 397 BP 68 38402 St Martin d'Hères cedex

RESUME

Nous présentons ici une méthode de calcul efficace de la fonction de log-vraisemblance d'un modèle autorégressif ainsi que de ses deux premières dérivées par rapport aux autocor - rélations partielles. Cela permet de contrôler la stabilité du filtre autorégressif dans la mise en oeuvre des techniques de maximisation de type gradient. Des essais, sur données simu - lées, ont montré la supériorité du maximum de vraisemblance exact par rapport aux méthodes de Burg ou du maximum de vraisemblance récursif lorsque le nombre d'observations est faible. La méthode de gradient utilisée ici est celle de Newton-Rhapson. Elle converge très vite à condition de l'initialiser dans un voisinage adéquat de la solution.

SUMMARY

We present here a method to compute efficiently the log likelihood function of an autoregressive model as well as its first and second derivatives with respect to the partial auto-correlations. This allows to control the stability of the auto-regressive filter in the implementation of maximisation tech - niques of gradient type. Some simulation results shows the better performance of the maximum likelihood estimator with respect to that of Burg or that of the recursive maximum like - lihood method, when the sample size is small. The gradient method used here is the Newton-Rhapson method. It conver - ges very fast provided that the initial value is already in some adequate neighbourhood of the solution.

INTRODUCTION

Le modèle autorégressif joue un rôle important dans l'ana - lyse statistique des signaux, en particulier comme technique spectrale de haute résolution.

Diverses méthodes d'ajustement de ce modèle existent dans la littérature: autocovariance empirique (Yule-Walker), maximum d'entropie (Burg, 1975), moindres carrés (cf. Marple, 1980), autocorrélation partielle empirique (cf. Dickinson, 1978), maximum de vraisemblance récursif (Kay, 1983). Ces méthodes, sauf Yule-Walker, sont équivalentes pour des séries longues mais peuvent se comporter différemment dans le cas de séries courtes. Alors le maximum de vraisemblance exact peut donner de meilleurs résultats. Son utilisation est devenue plus attractive grâce aux moyens de calcul de plus en plus performants. Ce calcul a été abordé à l'aide de techniques de gradient (Burg et all, 1982, Le Cadre et Lopez, 1984, Pham Dinh, 1987) et de type relaxation (Dégerine, 1987).

Une difficulté, dans l'utilisation des méthodes de gradient, est de contrôler la stabilité du polynôme autorégressif. Elle peut être surmontée grâce à la paramétrisation du modèle par les autocorrélations partielles. Nous proposons ici d'utiliser simultanément ces deux ensembles de paramètres. En effet la partie déterminant de la fonction de log-vraisemblance est une fonction simple des autocorrélations partielles et nous mon - trons que la partie quadratique (en les observations) est aussi quadratique en les coefficients autorégressifs.

Notre méthode n'est pas restreinte au cas d'une seule série mais s'applique surtout dans le cas d'un échantillon de séries courtes. Dans cette situation, l'ordre du modèle peut être égal à la longueur de la série, ce qui équivaut à l'estimation par maximum de vraisemblance d'une matrice de covariance de Toeplitz. La structure de Toeplitz est très courante: elle traduit une stationnarité temporelle ou spatiale (traitement d'antenne).

Nous précisons, dans le premier paragraphe, l'expression de la log-vraisemblance en fonction des autocorrélations partielles et des coefficients autorégressifs. Le deuxième para - graphe est consacré aux relations entre ces deux ensembles de paramètres et plus particulièrement entre les opérateurs de dérivation correspondants. Nous donnons, dans un dernier paragraphe, des résultats de simulation obtenus par la méthode de Newton-Rhapson dans le cas d'un échantillon de séries courtes.

I - FONCTION DE LOG-VRAISEMBLANCE

Considérons le modèle autorégressif d'ordre p:

$$\sum_{j=0}^{p} a_j x_{t-j} = e_t, \quad a_0 = 1,$$

dans lequel e_t désigne un bruit blanc gaussien, centré, de variance σ^2 et pour lequel les coefficients autorégressifs a_j satisfont les conditions de stabilité habituelles. On suppose que l'on observe r séquences de longueur n, indépendantes, notées $x_{1i},\ldots,x_{ni}$, $i=1,\ldots,r$. La fonction de log-vraisem - blance s'écrit

$$L(R,\sigma^2;S) = -(1/2)\{n\ln(2\pi\sigma^2) + \ln(\det R) + \mathrm{Tr}(R^{-1}S)/\sigma^2\}$$

où $\sigma^2 R$ est la matrice de covariance commune à chaque séquence et S est la matrice de covariance empirique

$$S = (1/r)\sum_{i=1}^{r} (x_{1i},\ldots,x_{ni})(x_{1i},\ldots,x_{ni})'.$$

On utilise les notations $i\wedge j=\min(i,j)$ et $i\vee j=\max(i,j)$.

Lemme.–*La forme quadratique en les observations* $\mathrm{Tr}(R^{-1}S)$ *est aussi une forme quadratique en les paramètres* a_j

$$\mathrm{Tr}(R^{-1}S) = \sum_{i,j=0}^{p} a_i a_j Q_{ij}$$

où la matrice Q *est donnée par*

$$Q_{ij} = \sum_{k=1}^{n-|i-j|} S_{k,k+|i-j|} - \sum_{k=0}^{i\wedge j-1} S_{i-k,j-k} - \sum_{k=n+1}^{n+i\wedge j} S_{k-i,k-j}.$$

De plus Q *satisfait* $Q_{ij}=Q_{ji}=-Q_{n-j,n-i}$ *et*

$$Q_{ij} = \sum_{k=1}^{n-i-j} S_{k+i,k+j}, \quad i+j<n.$$

Preuve.–Posons $a_j=0$ pour $j>p$. On a $R^{-1}=A'A-B'B=AA'-BB'$ (cf. Godolphin et Unwin, 1983) où A et B sont les matrices triangulaires

$$A = \begin{bmatrix} a_0 & a_1 & \cdots & a_{n-1} \\ 0 & a_0 & \cdots & a_{n-2} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & a_0 \end{bmatrix}, \quad B = \begin{bmatrix} a_n & a_{n-1} & \cdots & a_1 \\ 0 & a_n & \cdots & a_2 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & a_n \end{bmatrix}.$$

Par suite $\mathrm{Tr}(R^{-1}S)=\mathrm{Tr}(A'SA)-\mathrm{Tr}(B'SB)$. Notant C la matrice nx2n, (A B'), on a

$$\mathrm{Tr}(R^{-1}S) = \mathrm{Tr}(C'SC) - \mathrm{Tr}(BSB') - \mathrm{Tr}(B'SB).$$

Or si J est la matrice de terme général $J_{ij}=1$ si $i+j=n+1$, 0 sinon, alors $B'=JBJ$ et comme J^2 est la matrice unité on obtient

$$\mathrm{Tr}(R^{-1}S) = \mathrm{Tr}(C'SC) - \mathrm{Tr}(BSB') - \mathrm{Tr}(BJSJB').$$

La trace de C'SC s'écrit

$$\sum_{k=1}^{2n} \sum_{u=1}^{n\wedge k} \sum_{s=1}^{n\wedge k} a_{k-u} a_{k-s} S_{us} = \sum_{i=0}^{n} \sum_{j=0}^{n} a_i a_j \Big(\sum_{k=i\vee j+1}^{n+i\wedge j} S_{k-i,k-j} \Big).$$

Notons que la dernière somme entre parenthèses ci-dessus est précisement

$$\sum_{k=1}^{n-|i-j|} S_{k,k+|i-j|}.$$

D'autre part la trace de BSB' est égale à

$$\sum_{k=1}^{n} \sum_{u=1}^{k} \sum_{s=1}^{k} a_{n-k+u} a_{n-k+s} S_{us} = \sum_{i=0}^{n} \sum_{j=0}^{n} a_i a_j \Big(\sum_{u=0}^{i\wedge j-1} s_{i-u,j-u} \Big).$$

Comme la trace de BJSJB' est donnée par la même expression que la précédente, avec S_{ij} remplacé par $S_{n+1-i,n+1-j}$, on obtient le premier résultat du lemme. Le deuxième s'obtient, à partir de celui-ci, en regroupant les termes de S qui figurent dans les différentes sommes qui le composent.♦

Ainsi la partie quadratique, $\mathrm{Tr}(R^{-1}S)$, de la fonction log-vraisemblance est une expresssion simple des coefficients autorégressifs. Il n'en est pas de même pour la partie ln(detR) qui par contre s'exprime simplement à l'aide des autocorrélations partielles $\beta_1,\ldots,\beta_p$:

$$\ln(\det R) = -\sum_{k=1}^{p} k \ln(1-\beta_k^2).$$

Cependant les coefficients autorégressifs a_j sont soumis aux contraintes de stabilité alors que les autocorrélations partielles β_k sont libres dans]-1,1[. Il est donc préférable d'utiliser ces dernières comme paramètres du modèle.

Pour (β_k) fixé, la maximisation de $L(R,\sigma^2;S)$, par rapport à σ^2, est réalisée par

$$\sigma^2 = (1/n)\mathrm{Tr}(R^{-1}S).$$

Maximiser la log-vraisemblance équivaut donc à minimiser, par rapport à (β_k), la nouvelle fonction

$$L^*(\beta_1,\ldots,\beta_p;S) = (1/2)\{\ln \det(R) + n\ln[\mathrm{Tr}(R^{-1}S)/n]\}.$$

Cette minimisation ne peut être réalisée que par des méthodes itératives dont la plupart sont de type gradient.

II - CALCUL DES DERIVEES

Les coefficients autorégressifs sont obtenus, en fonction des autocorrélations partielles, par une récurrence faisant intervenir les paramètres des modèles autorégressifs d'ordre inférieur:

pour $k=1,\ldots,p$

$$a_0(k) = 1, \quad a_k(k) = -\beta_k,$$

$$a_j(k) = a_j(k-1) - \beta_k a_{k-j}(k-1), \quad j=1,\ldots,k-1,$$

et $a_j=a_j(p)$, $j=0,\ldots,p$.

Convenons de poser, pour $k=1,\ldots,p-1$, $a_j(k)=a_j(j)$ si $j>k$. La formule de dérivation

$$\frac{\partial}{\partial a_i(k-1)} = \sum_{j=1}^{p} [\partial a_j(k)/\partial a_i(k-1)] \frac{\partial}{\partial a_j(k)}$$

et la récurrence ci-dessus conduisent, pour $k=2,\ldots,p$, à

$$\frac{\partial}{\partial a_i(k-1)} = \frac{\partial}{\partial a_i(k)} + a_k(k) \frac{\partial}{\partial a_{k-i}(k)}, \quad i=1,\ldots,k-1,$$

$$\frac{\partial}{\partial a_k(k-1)} = \sum_{j=1}^{k} a_{k-j}(k-1) \frac{\partial}{\partial a_j(k)},$$

$$\frac{\partial}{\partial a_i(k-1)} = \frac{\partial}{\partial a_i(k)}, \quad i=k+1,\ldots,p.$$

Notant $\partial/\partial a(k)$ le vecteur $(\partial/\partial a_1(k),\ldots,\partial/\partial a_p(k))'$ ces relations s'écrivent

$$\partial/\partial a(k-1) = T(k-1)\partial/\partial a(k)$$

où $T(k-1)$ est la matrice de terme général

$$T_{ij}(k-1) = \begin{cases} \delta_{ij} + a_k(k) & \text{si } i+j=k \\ a_{k-j}(k-1) & \text{si } i=k \text{ et } j<k \\ \delta_{ij} & \text{sinon} \end{cases}$$

δ_{ij} désignant le symbole de Kronecker. Par suite pour toute fonction f de (a_i) on a

$$\frac{\partial^2 f}{\partial a_i(k-1)\partial a_j(k-1)} = \frac{\partial}{\partial a_j(k-1)} \sum_{u=1}^{p} T_{iu}(k-1)\frac{\partial f}{\partial a_u(k)}$$

$$= \sum_{u=1}^{p}\sum_{s=1}^{p} T_{iu}(k-1)T_{js}(k-1)\frac{\partial^2 f}{\partial a_u(k)\partial a_s(k)} + \sum_{u=1}^{p} \frac{\partial}{\partial a_j(k-1)} T_{iu}(k-1)\frac{\partial f}{\partial a_u(k)}$$

Dans la dernière expression ci-dessus, le premier terme est l'élément (i,j) de la matrice $T(k-1)f''(k)T(k-1)'$ où $f''(k)$ est la matrice de terme général $\partial^2 f/\partial a_i(k)\partial a_j(k)$. Quant au second terme il est égal à l'élément (i,j) de la matrice F' définie par

$$F'_{ij}(k) = \begin{cases} \partial f/\partial a_{k-i}(k) & \text{si } i<k \text{ et } j=k \\ \partial f/\partial a_{k-j}(k) & \text{si } i=k \text{ et } j<k \\ 0 & \text{sinon} \end{cases}$$

En conclusion, avec les notations introduites, nous pouvons énoncer:

Proposition.–*Soit* f *une fonction de* (a_i) *deux fois dérivable. Le vecteur des dérivées premières et la matrice des dérivées secondes de* f *par rapport à* (β_i) *sont donnés par* $-f'(1)$ *et* $f''(1)$ *issus de la récurrence*

$$f'(p) = (\partial f/\partial a_1,\ldots,\partial f/\partial a_p)',$$

$$f''(p) = (\partial^2 f/\partial a_i \partial a_j)_{i,j=1,\ldots,p},$$

et pour $k=p,\ldots,2$

$$f'(k-1) = T(k-1)f'(k),$$

$$f''(k-1) = T(k-1)f''(k)T(k-1)' + F'(k).$$

Appliquons ce résultat à la fonction

$$f = \frac{1}{2}\sum_{i,j=0}^{p} a_i a_j Q_{ij}.$$

Nous obtenons les dérivées de L^*

$$\frac{\partial L^*}{\partial \beta_i} = \frac{i\beta_i}{1-\beta_i^2} - \frac{1}{\sigma^2} g_i$$

$$\frac{\partial^2 L^*}{\partial \beta_i \partial \beta_j} = \frac{i(1+\beta_i^2)}{(1-\beta_i^2)^2}\delta_{ij} + \frac{1}{\sigma^2}G_{ij} - \frac{2}{n\sigma^4} g_i g_j ,$$

où σ^2 est égal à $2f/n$, g_i et G_{ij} désignant respectivement les composantes de $f'(1)$ et de $f''(1)$.

III - QUELQUES RESULTATS DE SIMULATION

Nous testons les performances de l'estimateur de maximum de vraisemblance exact (MVE) par rapport à ceux du maximum d'entropie (ME) et du maximum de vraisemblance récursif (MVR). Ces dernières méthodes, conçues dans le cas d'une seule série, sont transposables au cas où plusieurs séries sont observées.

Nous reprenons le modèle utilisé par Kay (1983), défini par les coefficients autorégressifs: $a_1=-2.761$, $a_2=3.812$, $a_3=-2.654$, $a_4=0.924$ et la variance du bruit $\sigma^2=0.010$. Les autocorrélations partielles correspondantes sont $\beta_1=0.716$, $\beta_2=-0.982$, $\beta_3=0.704$, $\beta_4=-0.924$.

Les méthodes sont appliquées sur 400 répétitions d'un échantillon de 10 séries de longueur 5. Nous indiquons, dans la table ci-après, la moyenne ainsi que l'écart quadratique moyen, observés sur ces 400 réalisations, des estimations des paramètres (β_i) et (a_i) obtenues par les trois méthodes. L'estimateur de maximum de vraisemblanece exact donne, pour chaque paramètre, de meilleurs résultats que les deux autres méthodes à la fois pour le biais et pour l'écart quadratique moyen. L'efficacité relative, mesurée par le rapport des écarts quadratiques moyens, est de l'ordre de 4 entre MVE et MVR sur l'ensemble des paramètres. Elle est de l'ordre de 7 entre MVE et ME pour les paramètres autorégressifs.

Table.–Moyenne et écart quadratique moyen des estimateurs (les valeurs de eqm sont multipliées par 10^5)

méthodes		ME		MVR		MVE	
paramètres		moy.	eqm	moy.	eqm	moy.	eqm
β_1	0.716	0.716	53	0.714	116	0.717	22
β_2	−0.982	−0.980	10	−0.977	14	−0.981	5
β_3	0.704	0.651	1960	0.691	1124	0.698	314
β_4	−0.924	−0.882	1020	−0.892	640	−0.926	137
a_1	−2.761	−2.626	7290	−2.701	4080	−2.752	1020
a_2	3.812	3.575	17556	3.690	9672	3.799	2434
a_3	−2.654	−2.460	11156	−2.551	6101	−2.648	1588
a_4	0.924	0.882	1020	0.892	640	0.926	203

Le maximum de vraisemblance exact a été réalisé ici par la méthode de Newton-Rhapson. D'autre part, afin d'éliminer les problèmes de bord dans les contraintes $-1<\beta_i<1$, nous avons utilisé la transformation de Fisher

$$z_i = (1/2)\ln \frac{1+\beta_i}{1-\beta_i} \quad , \quad \frac{\partial}{\partial z_i} = [1-\beta_i^2]\frac{\partial}{\partial \beta_i} .$$

Cependant il est important d'initialiser la méthode par une valeur suffisamment proche de la solution sinon elle ne converge pas. Ceci se produit lorsque la matrice des dérivées secondes n'est pas positive (fonction non convexe) ou encore lorsque le pas, au cours des itérations, est trop grand. Dans les essais de l'exemple que nous avons présenté, la matrice des dérivées secondes n'est jamais positive au point donné par la matrice de covariance empirique (Yule-Walker). De même, souvent, la méthode n'aboutit pas à partir de celui donné par ME ou MVR. Nous avons donc pris, comme valeur initiale, celle obtenue après quelques itérations de la méthode de relaxation proposée par Dégerine (1987) pour le calcul de MVE dans cette situation. Pour un même critère d'arrêt, alors que la méthode de relaxation exige en moyenne une trentaine d'itérations, une dizaine suffisent pour initialiser la méthode de Newton-Rhapson qui converge alors vers la même valeur avec, en moyenne, trois itérations. Néanmoins il faut souligner que la méthode ne modifie pas de façon importante les résultats donnés par la valeur d'initialisation.

CONCLUSION

Nous disposons d'une méthode efficace pour calculer la log-vraisemblance d'un modèle autorégressif ainsi que ses deux premières dérivées dans le cas d'une seule série ou d'un échantillon.

Nous avons montré, sur un exemple, que l'estimateur de maximum de vraisemblance peut améliorer de façon importante les propriétés statistiques données par d'autres estimateurs comme celui du maximum d'entropie ou encore celui du maximum de vraisemblance récursif.

La mise en oeuvre des techniques de gradient, pour calculer cet estimateur, doit cependant tenir compte de la non convexité de la fonction de log-vraisemblance.

REFERENCES

BURG, J. P. (1975) Maximum entropy spectral analysis. Ph. D. dissertation, Dep. Geophys. Stanford Univ., Stanford, CA.

BURG, J.P., LUENBERGER, D. G. and WENGER, D. L.(1982) Estimation of structured covariance matrices. Proc. IEEE, 70, 963–974.

DEGERINE, S. (1987) Maximum likelihood estimation of autocovariance matrices from replicated short time series. J. Time Series Anal. 18, (à paraître).

DICKINSON, B. W. (1978) Autoregressive estimation using residual energy ratios. IEEE-IT, 24, 503–506.

GODOLPHIN, E. J. and UNWIN, J. M. (1983) Evaluation of the covariance matrix for the maximum likelihood estimator of a gaussian autoregressive-moving average process. Biometrika, 70, Vol.1, 279–84.

KAY, S. M. (1983) Recursive maximum likelihood estimation of autoregressive processes. IEEE-ASSP, 31, 56–65.

LE CADRE, J. P. et LOPEZ, P. (1984) Estimation d'une matrice interspectrale de structure imposée. Applications. Traitemant du Signal, 1, 3–17.

MARPLE, L. (1980) A new autoregressive spectrum analysis algorithm. IEEE-ASSP, 28, 441–454.

PHAM, D. T. (1987) Exact maximum likelihood and Lagrange multiplier statistic for ARMA models. J. Time Series Anal. 18, Vol.1, 61–78.