Estimation de ruptures multiples dans les processus FARIMA

Martial COULON[1], Marie CHABERT[1], Ananthram SWAMI[2]

[1]Université de Toulouse, Institut de Recherche en Informatique de Toulouse (IRIT)
INP-ENSEEIHT, 2 rue Charles Camichel, BP 7122, 31071 Toulouse, France

[2]Army Research Lab Adelphi MD 20783 USA
{coulon,chabert}@enseeiht.fr,a.swami@ieee.org

Résumé – Cet article étudie l'estimation des instants de ruptures dans des processus FARIMA modélisant le trafic sur les réseaux de communication. Une procédure d'estimation de ruptures combinant l'estimateur temps-échelle du paramètre de longue dépendance, l'estimation du maximum de vraisemblance des paramètres ARMA et un critère de moindres carrés pénalisés est proposée. La résolution s'effectue par programmation dynamique. Les performances sont analysées de manière théorique, sur données réelles et simulées et confirment les propriétés de convergence et de robustesse de la méthode.

Abstract – This paper studies the estimation of abrupt change locations in FARIMA processes. Such processes allow to model the trafic in communication networks. The proposed strategy combines a time-scale estimation of the long range dependance parameter, the maximum likelihood estimation of the ARMA parameters and a penalized least square criterion. The resolution is performed through dynamic programming. The performance is theoretically studied. Experiments on synthetic and real data attest the convergence and the robustness of the method.

1 Introduction

1.1 Contexte

Les processus à longue dépendance (LRD pour *long range dependence*) ont fait l'objet de nombreuses études pendant les cinquante dernières années [1], [2]. En particulier, le trafic de données sur les réseaux ethernet et internet peut être décrit par des grandeurs variées dont certaines s'avèrent être LRD ou asymptotiquement auto-similaires au second ordre [3], [4] [5], [6]. Cet article s'intéresse à l'estimation des instants de variation brusque du trafic qui permet en particulier de réguler les techniques de contrôle et d'allocation de ressources [4]. La détection de changements peut être effectuée à partir de l'estimateur du paramètre de Hurst décrivant le degré de LRD dans les données [7]. Abry et Veitch ont développé un estimateur basé sur les ondelettes qui est asymptotiquement sans biais, convergent, gaussien et de variance quasi-minimale [8]. D'autre part, la détection de changements dans le spectre d'un processus paramétrique constant par morceaux, éventuellement FARIMA, a été étudié dans [9]. Toutefois, l'estimateur proposé, basé sur la pseudo-vraisemblance de Whittle, est très coûteux d'un point de vue calculatoire [1], et est donc difficilement applicable aux signaux de grande taille de type éthernet ou internet. Nous proposons une procédure d'estimation de ruptures combinant l'estimateur temps-échelle du paramètre de longue dépendance, l'estimation du maximum de vraisemblance des paramètres ARMA et un critère de moindres carrés pénalisé, la résolution s'effectuant par programmation dynamique. Une version étendue de cet article a été publiée dans [10].

1.2 Modèle de signal

Notons $r(k)$ la fonction d'autocorrélation d'un processus stationnaire tel que $r(k) \underset{k\longrightarrow+\infty}{\sim} ck^{-\beta}$ où c est une constante positive. Si $0 < \beta < 1$ (resp. $\beta > 1$), le processus est à longue (respectivement courte) dépendance. $H = 1 - \beta/2$ est le paramètre de Hurst [1]. Nous considérons dans la suite des processus FARIMA (*fractional auto-regressive integrated moving average*) qui se sont avérés appropriés pour des données telles que le trafic vidéo à débit variable [11]. Soit $y(n)$ un processus FARIMA(p, d, q) de paramètre LRD d (d est relié au paramètre de Hurst par $d = H - 1/2$.) et de paramètres ARMA $\mathbf{a} \triangleq [a(0), \ldots, a(p)]^T$ et $\mathbf{b} \triangleq [b(0), \ldots, b(q)]^T$ avec $a(0) = b(0) = 1$, $a(p) \neq 0$, et $b(q) \neq 0$. Le processus AR est supposé causal et stable c'est-à-dire toutes les racines de $\sum_{k=0}^{p_i} z^{-1} a_i(k)$ sont à l'intérieur du cercle unité. Le processus FARIMA $y(n)$ vérifie l'équation [1] (la notation $H(z^{-1})y(n)$ signifie $\sum_{k=0}^{\infty} h(k)y(n-k)$) :

$$\Phi(z^{-1})(1 - z^{-1})^d y(n) = \Psi(z^{-1})\xi(n),$$

avec :

$$\Phi(z^{-1}) \triangleq \sum_{k=0}^{p} a(k)z^{-k}, \qquad \Psi(z^{-1}) \triangleq \sum_{k=0}^{q} b(k)z^{-k},$$

$$(1 - z^{-1})^d = \sum_{k=0}^{+\infty} (-1)^k \frac{\Gamma(d+1)}{k!\Gamma(d-k+1)} z^{-k}$$

et $\xi(n)$ est une séquence gaussienne indépendante et identiquement distribuée de moyenne et de variance finie ; $\Gamma(.)$ est la fonction Gamma standard. Le processus présente de la LRD si $0 < d < 1/2$ [1].

2 Méthode proposée

La LRD affecte seulement les corrélations à long terme. Nous ne chercherons donc pas à atteindre une résolution arbitrairement fine pour la segmentation. Les changements seront détectés à N_s échantillons près. Considérons la décomposition de l'ensemble $\{n = 1, ..., N\}$ en K segments élémentaires

$$\mathcal{I}_k \triangleq \{kN_s + 1 + \nu/2, \ldots, (k+1)N_s - \nu/2\}$$

de $N_s - \nu$ échantillons (i.e. $N \approx K \times (N_s + \nu)$), où ν garantit un intervalle de garde entre deux segments consécutifs afin d'assurer l'indépendance des estimations calculées sur chacun d'eux. On note $\overline{M}$ le nombre de changements. La méthode consiste tout d'abord à estimer les paramètres FARIMA sur chacun des K segments élémentaires, conduisant aux vecteurs estimés $\widehat{\boldsymbol{\theta}}_k \triangleq \left(\widehat{d}_k, \widehat{\mathbf{a}}_k^T, \widehat{\mathbf{b}}_k^T\right)^T$ pour $k = 1, \ldots, K$.

Soit $\widehat{\boldsymbol{\Theta}} = [\widehat{\boldsymbol{\theta}}_1 \widehat{\boldsymbol{\theta}}_2 ... \widehat{\boldsymbol{\theta}}_K]$ la matrice des vecteurs estimés. La deuxième étape consiste à détecter des changements de ces paramètres estimés c'est-à-dire sur chaque ligne de $\widehat{\boldsymbol{\Theta}}$ par minimisation d'un critère de moindres carrés. Une rupture détectée à la $l^{ème}$ colonne de $\widehat{\boldsymbol{\Theta}}$ suppose qu'une rupture s'est produite durant le segment $\mathcal{I}_l$ dans le processus observé $y(n)$. Le choix du nombre d'échantillons N_s des segments résulte d'un compromis entre la résolution souhaitée et la précision de l'estimation des paramètres FARIMA.

2.1 Estimation des paramètres FARIMA sur les segments élémentaires

Un algorithme en trois étapes est appliqué à chaque segment $\mathcal{I}_k$:

1. L'estimée du paramètre LRD, $\widehat{d}_k$, est calculée via la procédure d'Abry-Veitch [8] ;
2. – Si $\widehat{d}_k > 0$, le processus $(y(n))_{n \in \mathcal{I}_k}$ est considéré comme LRD. Il est alors filtré par le filtre FARIMA $(0, -\widehat{d}_k, 0)$ afin de supprimer le caractère LRD. Si l'estimée $\widehat{d}_k$ est suffisamment précise, le signal résultant $\widehat{z}_k(n)$ est proche d'un processus ARMA.
   – Si $\widehat{d}_k < 0$, le processus $(y(n))_{n \in \mathcal{I}_k}$ est SRD. Donc, le filtrage précédent ne doit pas être effectué car il peut conduire à de la LRD : dans ce cas $\widehat{z}_k(n) = y(n)$, $n \in \mathcal{I}_k$.
3. Les paramètres ARMA sont estimés à partir de $\widehat{z}_k(n)$, en utilisant le maximum de vraisemblance [12].

2.2 Estimation des instants de ruptures

Le problème revient ensuite à estimer des ruptures dans un signal multi-dimensionnel $(\widehat{\boldsymbol{\theta}}_k)_{k=1,\ldots,K}$. Le vecteur des estimées des instants de rupture $\widehat{\mathbf{l}} \triangleq \left(\widehat{l}_1, \ldots, \widehat{l}_{\overline{M}-1}\right)^T$ dans le signal $\left(\widehat{\boldsymbol{\theta}}_k\right)_{k=1,\ldots,K}$ est obtenu par minimisation d'un critère de moindres carrés pénalisé ou non selon que le nombre de ruptures est inconnu ou non. Dans les deux cas, l'optimisation du critère utilise un algorithme de programmation dynamique fournissant la solution exacte avec un gain de temps considérable par rapport à une recherche exhaustive [13].

3 Etude des performances

3.1 Analyse théorique

L'analyse des propriétés statistiques de l'estimateur des instants de rupture s'appuie sur une étude préalable des estimateurs des paramètres FARIMA. Il a été montré dans [10] que l'erreur d'estimation, calculée sur chaque segment $\mathcal{I}_k$, est asymptotiquement (c'est-à-dire lorsque le nombre N_s d'échantillons par segment tend vers l'infini) centrée, indépendante, gaussienne, de matrice de covariance constante par morceaux. Ces propriétés permettent ensuite, en s'appuyant sur les travaux présentés dans [14] et [15], d'étudier la convergence des estimations des instants de rupture. La notion de convergence pour l'estimateur d'instants qui prennent des valeurs entières est définie comme suit. Considérons que la longueur des signaux $y_i(n)$ croît linéairement avec N. On conserve le même nombre d'échantillons par segment $\mathcal{I}_k$, et donc le nombre K de ces segments $\mathcal{I}_k$ croît également linéairement avec N, puisque $N \approx KN_s$. On a ainsi $\bar{l}_i - \bar{l}_{i-1} = O(K)$, où $\bar{l}_i$ est la valeur réelle du $i^{\text{ème}}$ instant de rupture. Il existe alors $(\tau_1, \ldots, \tau_{\overline{M}})^T \in [0,1]^{\overline{M}}$ tel que $\lim_{K\to\infty} \bar{l}_i/K = \tau_i$, $\forall i \in \left\{1, \ldots, \overline{M}\right\}$. La convergence de $\widehat{l}_i$ signifie que [14]

$$\lim_{K\to\infty} \widehat{l}_i/K = \tau_i$$

Cette convergence a alors été montrée dans [10], lorsque le nombre de ruptures est connu ou inconnu.

3.2 Données synthétiques

Nous avons effectué un certain nombre de simulations destinées à illustrer l'étude théorique précédente. Considérons un signal synthétique $y(n)$ obtenu par concaténation de six FARIMA ($\overline{M} = 6$) d'ordre ARMA $p = 1$ et $q = 2$. Les $(y_i(n))_{i=1,\ldots,6}$ sont normalisés de manière à avoir même moyenne et même variance. Ainsi, la détection de ruptures ne pourrait pas être effectuée par un simple détecteur énergétique. Les séquences d'entrée $\xi_i(n)$ des processus FARIMA $y_i(n)$ sont gaussiennes. Le nombre d'échantillons est fixé à $N = 2^{15} = 32768$, et la taille d'un segment élémentaire est $N_s = 1024$. On compte donc $K = 32$ segments élémentaires $\mathcal{I}_k$. Les instants de rupture sont $t = [6450; 12721; 17145; 24932; 28432]$. La figure 1 présente la position de la rupture et la moyenne des estimées sur 200 itérations, avec $\overline{M}_{\max} = 6$ pour différents termes de pénalisation γ. Dans le cas $\gamma = 0$, le nombre de ruptures détectées est égal à $\overline{M}_{\max}$. De plus, les estimées sont relativement précises compte tenu du fait que $N_s = 1024$. En effet, soit l'algorithme trouve la position qui est la plus proche de la véritable position, soit il hésite entre deux positions situées de part et d'autre. Les performances sont tout à fait satisfaisantes

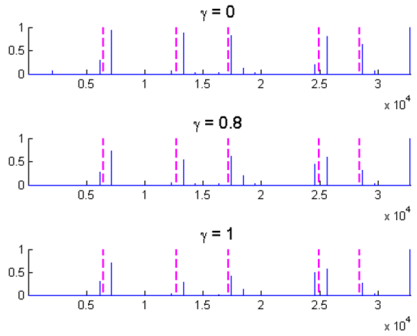

FIG. 1 – Positions de ruptures réelles (tirets) et estimées (traits pleins) obtenues sur 200 simulations.

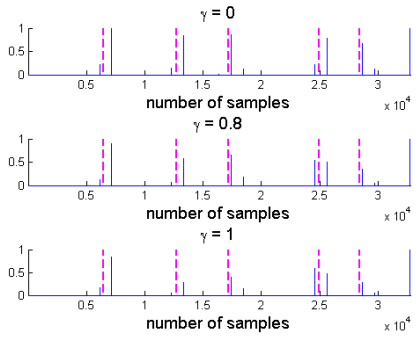

FIG. 2 – Positions de ruptures réelles (tirets) et estimées (traits pleins) obtenues sur 200 simulations, avec des entrées exponentielles.

puisque on ne dispose que de 32 échantillons du vecteur $\left(\widehat{\boldsymbol{\theta}}_k\right)$ et il y a 5 ruptures à estimer, et donc seulement 6 échantillons de $\left(\widehat{\theta}_k\right)$ en moyenne par segment. Pour un terme de pé nalisation croissant ($\gamma = 0.8$, 1, 1.5, respectivement), toujours avec $\overline{M}_{\max} = 6$, le nombre de ruptures détectées diminue. Ces estimations ont été obtenues avec un nombre maximum de ruptures égal au nombre effectif de ruptures. D'autres simulations ont montré qu'il n'y a pas d'inconvénient, excepté du point de vue de la complexité calculatoire, à surestimer $\overline{M}_{\max}$. La figure 2 donne les estimations de ruptures obtenues lorsque le processus d'entrée n'est plus gaussien, mais exponentiel. On peut constater que les résultats sont très similaires au cas gaussien. Ainsi, bien que l'analyse théorique de l'estimateur ne soit a priori valable que pour des entrées gaussiennes, cette figure montre une certaine robustesse de l'algorithme vis-à-vis de la loi du processus d'entrée. D'autre part, l'algorithme d'estimation proposé ainsi que les simulations précédentes supposent connus les ordres ARMA p et q. Cette hypothèse n'est pas toujours réaliste. Le comportement de l'algorithme a également été étudié lorsque les ordres des modèles sont inconnus, comme illustré sur la figure 3. Pour des ordres du modèle ARMA arbitrairement fixés à des valeurs non toutes nulles, l'estimation est presque équivalente au cas où les ordres sont connus. Toutefois,

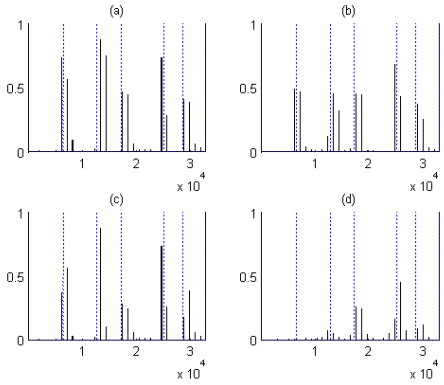

FIG. 3 – Estimation des ruptures avec des ordres variables (a) : ordres réels - (b) : $(p;q) = (2;2)$ - (c) $(p;q) = (3;1)$ (d) $(p;q) = (0;0)$.

lorsque $p = q = 0$, i.e. lorsque la partie ARMA n'est pas prise en compte, les performances chutent de manière conséquente. En effet, l'estimation de d est moins précise que celle des paramètres ARMA. L'estimation des paramètres ARMA est donc essentielle pour la segmentation. La segmentation peut donc être réalisée même lorsque les ordres du modèle ARMA sont inconnus. Notons toutefois que l'analyse théorique des performances n'est plus valable dans ce cas.

3.3 Données réelles

L'algorithme d'estimation de ruptures a été testé sur des signaux réels. Nous avons analysé les traces de trafic ethernet d'août 1989 des laboratoires Bellcore. Ces données contiennent des milliers d'observations qui représentent, pour chacune d'elles, le nombre d'octets envoyés sur une liaison éthernet sur une durée de 10 ms. Ces données sont décrites en détails dans [3]. Par contre nous ne disposons pas d'une expertise de ces données du point de vue de la segmentation. L'objectif est essentiellement d'étudier le comportement de l'algorithme en l'absence de connaissances a priori sur les paramètres et les ordres des modèles FARIMA. L'algorithme peut être utilisé en affectant des ordres ARMA arbitraires non nuls. L'algorithme a été testé avec différents ordres. La figure 4 montre les estimées des paramètres FARIMA, ainsi que la position des instants de rupture estimés, pour les 25 couples $(\widetilde{p}, \widetilde{q})$ possibles en faisant varier $\widetilde{p}$ et $\widetilde{q}$ entre 0 et 4. Dans cette figure, les marques à la i^{eme} ligne indiquent les positions de ruptures détectées pour les ordres $(\widetilde{p}_i, \widetilde{q}_i)$. Différents couples $(\widetilde{p}, \widetilde{q})$ permettent de détecter les mêmes ruptures. Les couples faisant intervenir $\widetilde{p} = 0$ ou $\widetilde{q} = 0$ négligent la partie AR ou la partie MA et laissent éventuellement passer davantage de ruptures. Les figures 5 et 6 montrent les histogrammes des positions de ruptures estimées obtenues avec les mêmes 25 couples $(\widetilde{p}, \widetilde{q})$ lorsqu'on prend en compte ou non l'estimateur du paramètre LRD pour la segmentation. Il apparaît clairement que considérer ce dernier permet de concentrer les estimations autour de quatre positions principales déjà identifiées sur la figure 4, alors qu'une segmentation

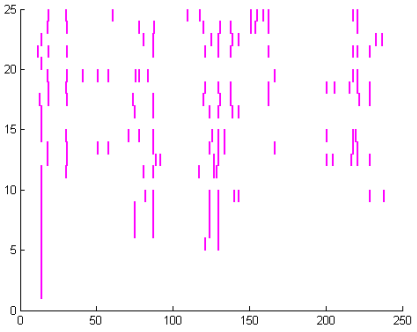

FIG. 4 – Ruptures estimées pour 25 couples $(\widetilde{p};\widetilde{q})$.

basée simplement sur la partie ARMA donne des estimations beaucoup plus éparses.

4 Conclusion

Nous avons étudié l'estimation de ruptures multiples dans les paramètres d'un processus FARIMA. Une estimation préliminaire des paramètres FARIMA est obtenue sur des segments élémentaires. La détection de ruptures se fait ensuite par minimisation d'un critère de moindres carrés. Le cas d'un nombre connu et celui d'un nombre inconnu de ruptures ont été étudiés. Dans le cas d'un nombre inconnu de ruptures, un terme de pénalisation dans le critère permet de régler la résolution de la segmentation. Dans les deux cas, la minimisation du critère se fait grâce à un algorithme de programmation dynamique. Une analyse statistique de l'estimateur des instants de rupture permet de montrer sa convergence. Les simulations montrent les bonnes performances de l'algorithme proposé, ainsi que sa robustesse à la loi des données et à la méconnaissance des ordres de la partie ARMA.

Références

[1] J. Beran, *Statistics for Long-Memory Processes*, Chapman&Hall, New York, 1994 (first edition in 1959).

[2] C.W.J. Granger, "Long memory relationships and the aggregation of dynamic models," *J. Econometrics*, vol. 14, pp. 227-238, Oct. 1980.

[3] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, "On the self-similar nature of ethernet traffic (Extended Version)," *IEEE/ACM Trans. on Networking*, vol. 2, no. 1, pp. 1-15, Feb. 1994.

[4] K. Park and W. Willinger, Eds, *Self-similar network traffic and performance evaluation*, Wiley Interscience Publication, 2000.

[5] W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson, "Self-similarity through high-variability : statistical analysis of ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71-96, Feb. 1997.

[6] O. Cappe, E. Moulines, J.-C. Pesquet, A.P. Petropulu, Xueshi Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal Proc. Mag.*, vol. 19, Issue 3, pp. 14-27, May 2002.

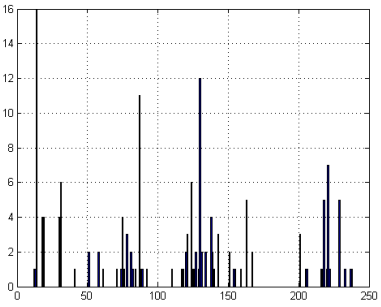

FIG. 5 – Histogramme des estimations de ruptures à l'aide d'une segmentation basée sur tous les paramètres FARIMA.

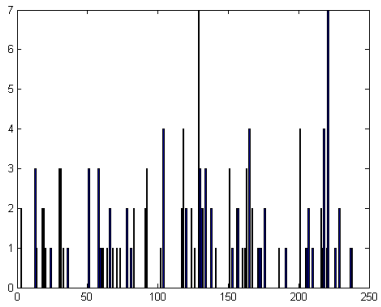

FIG. 6 – Histogramme des estimations de ruptures à l'aide d'une segmentation basée sur la partie ARMA seule.

[7] D. Veitch and P. Abry, "A statistical test for the time constancy of scaling exponents," *IEEE Trans. on Signal Processing*, vol. 49, pp. 2325-2334, Oct. 2001.

[8] P. Abry, P. Flandrin, M.S. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation and synthesis of Scaling Data", in [4].

[9] M. Lavielle and C. Ludeña, "The multiple change-points problem for the spectral distribution," *Bernoulli*, vol. 6, no. 5, pp. 845-869, 2000.

[10] M. Coulon, M. Chabert and A. Swami, "Detection of Multiple Changes in Fractional Integrated ARMA Processes," IEEE Trans. on Signal Processing, vol 57, pp.48-61, Jan. 2008.

[11] J. Beran, R. Sherman, M. Taqqu and W. Willinger, "LRD in VBR traffic", *IEEE Trans. Commun.*, 43, 1566–79, 1995.

[12] B. Porat, *Digital Processing of Random Signals : Theory and Methods*, Prentice-Hall, 1994.

[13] S.M. Kay, *Fundamentals of Statistical Signal Processing, Detection Theory*, Vol. II, Prentice-Hall PTR, Upper Saddle River, New Jersey 07458, 1998.

[14] M. Lavielle, "Detection of multiple changes in a sequence of dependent variables," *Stochastic Processes and Applications*, vol. 83, pp. 79-102, 1999.

[15] M. Lavielle and E. Moulines, "Least squares estimation of an unknown number of shifts in a time series", *Journal of Time Series Analysis*, vol. 21, no. 1, pp. 33-59, 2000.