Estimation d'un modèle autorégressif conditionnellement hétéroscédastique en présence de données manquantes

Pascal Bondon

Laboratoire des signaux et systèmes, CNRS UMR 8506
Supélec, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette Cedex, France
bondon@lss.supelec.fr

Résumé − Nous étudions le problème de l'estimation d'un modèle autorégressif conditionnellement hétéroscédastique (ARCH) en présence de données manquantes. Nous proposons un estimateur des moindres carrés en deux étapes qui est simple à calculer. On suppose que le processus décrivant les instants des données manquantes est stationnaire, faiblement mélangeant et indépendant du processus ARCH. Nous montrons que l'estimateur est fortement convergent et asymptotiquement gaussien. On présente une application à des données réelles boursières.

Abstract − A two-stage least squares estimator of the parameters of an autoregressive conditionally heteroscedastic (ARCH) model in the presence of missing data is proposed. The estimator is easy to obtain since it involves solving two sets of linear equations. The process which describes the dates of the missing data is assumed to be strictly stationary, weakly mixing and independent of the ARCH process. Strong consistency and asymptotic normality of the estimator are derived. An application to real data of a stock index is reported.

1 Introduction

Le modèle autorégressif conditionnellement hétéroscédastique (ARCH) introduit par Engle (1982) permet de modéliser l'hétéroscédasticité dans les actifs financiers. Ce modèle et ses nombreuses généralisations sont devenus aujourd'hui un outil indispensable pour les analystes des marchés financiers.

L'estimateur couramment utilisé pour un modèle ARCH est celui du quasi maximum de vraisemblance (EQMV) dont les propriétés ont été largement étudiées, voir par exemple Berkes et al. (2003), Francq et Zakoïan (2004), et Straumann (2005). Cet estimateur n'admet pas de forme explicite et son calcul numérique est difficile car la fonction de vraisemblance a tendance à être plate sauf pour de très grands échantillons. Bose et Mukherjee (2003) ont proposé un estimateur des moindres carrés en deux étapes qui s'exprime explicitement et présente donc l'avantage par rapport à l'EQMV de ne nécessiter aucune optimisation numérique. De plus, cet estimateur a les mêmes propriétés asymptotiques que l'EQMV.

La plupart des travaux sur les séries temporelles supposent que les données sont espacées régulièrement dans le temps. Cependant, dans les séries réelles, il n'est pas rare de rencontrer des données manquantes ou espacées irrégulièrement. Little et Rubin (2002) proposent des méthodes pour prendre en compte des données manquantes dans des problèmes d'estimation basés sur le maximum de vraisemblance. La monographie éditée par Parzen (1983) se concentre sur l'analyse de séries temporelles observées irrégulièrement.

Parzen (1963) propose de modéliser une telle série par une version modulée en amplitude de la série originale, i.e.,

$$X_t^\star = a_t X_t, \tag{1}$$

où X_t est définie pour tout t, a_t est donnée par

$$a_t = \begin{cases} 1 & \text{si } X_t \text{ est observée,} \\ 0 & \text{si } X_t \text{ est manquante,} \end{cases} \tag{2}$$

et $X_t^\star$ représente la valeur effectivement observée de X_t avec des zéros insérés dans la série lorsque X_t est manquante. En pratique, les données manquantes peuvent apparaître régulièrement ou aléatoirement. Jones (1962) et Parzen (1963) considèrent le cas d'un échantillonnage périodique où les données observées se composent de blocs successifs de A observations consécutives suivies de B données manquantes. Scheinok (1965) et Bloomfield (1970) quant à eux étudient le cas où X_t est observée ou non selon un schéma de Bernoulli. Ces quatre références se concentrent sur l'analyse spectrale non paramétrique de la série, alors que Dunsmuir et Robinson (1981b) privilégient une approche paramétrique. Dans le domaine temporel, Dunsmuir et Robinson (1981a) ainsi que Yajima et Nishino (1999) établissent des propriétés asymptotiques d'estimateurs non paramétriques des fonctions d'autocovariance et d'autocorrélation pour des séries satisfaisant (1) où (X_t) est un processus linéaire stationnaire avec

des innovations conditionnellement homoscédastiques, i.e. (X_t) satisfait

$$X_t = \sum_{j=0}^{\infty} \beta_j \epsilon_{t-j}, \qquad \sum_{j=0}^{\infty} \beta_j^2 < \infty,$$

où $\mathrm{E}(\epsilon_t \mid \mathcal{F}_{t-1}^{\epsilon}) = 0$ et $\mathrm{E}(\epsilon_t^2 \mid \mathcal{F}_{t-1}^{\epsilon}) = \sigma^2$, $\mathcal{F}_t^{\epsilon}$ étant la σ-algèbre engendrée par $(\epsilon_j)_{j \le t}$. Ces résultats peuvent être utilisés pour mettre en oeuvre des estimateurs de type «Yule-Walker» pour un processus AR avec des données manquantes. Mais ils ne s'appliquent pas au cas d'une série ARCH, ni a son carré. D'autre part, sur le plan pratique, Jones (1962) montre que la vraisemblance gaussienne d'un processus ARMA avec données manquantes peut être calculée au moyen d'une représentation d'état adaptée. Le carré d'un processus ARCH admet une représentation AR dans laquelle les innovations sont conditionnellement hétéroscédastiques et la méthode de Jones (1962) ne s'applique pas.

Nous considérons ici le cas où (X_t) est une série ARCH et où (a_t) est un processus aléatoire. À notre connaissance, l'estimation des paramètres d'un modèle ARCH en présence de données manquantes n'a pas encore été étudiée. Cette estimation pose à la fois des problèmes pratiques et des problèmes théoriques. En effet, le modèle par espace d'état et filtre de Kalman proposé par Jones (1962) ne s'applique pas au calcul de la vraisemblance d'un processus ARCH avec des données manquantes, et les résultats d'estimation existants ne s'appliquent pas au cas de séries temporelles conditionnellement hétéroscédastiques.

Nous proposons ici un estimateur des moindres carrés en deux étapes qui généralise l'estimateur de Bose et Mukherjee (2003) au cas du processus (1). Nous établissons ses propriétés asymptotiques, puis nous considérons une application à des données réelles boursières.

2 Modèle et estimateur

Soit (X_t) la série ARCH(p) définie par l'équation

$$X_t = \sigma_t(\alpha)\epsilon_t, \tag{3}$$

où (ϵ_t) est une suite iid avec $\mathrm{E}\epsilon_0 = 0$, $\mathrm{E}\epsilon_0^2 = 1$, et $(\sigma_t(\alpha))$ est un processus positif satisfaisant

$$\sigma_t^2(\alpha) = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2. \tag{4}$$

Soit $\alpha = (\alpha_0,\alpha_1,\ldots,\alpha_p)'$ avec $\alpha_i \ge 0$ et $\alpha_p > 0$. D'après (Giraitis et al., 2000, Theorem 2.1), si $\sum_{i=1}^{p} \alpha_i < 1$, il existe une unique solution non anticipative strictement stationnaire et ergodique à (3)–(4) satisfaisant $\mathrm{E}X_0^2 < \infty$. Si $\alpha_0 = 0$, cette solution est $X_t = 0$. Dans la suite, on suppose $\alpha_0 > 0$ et $\sum_{i=1}^{p} \alpha_i < 1$.

Soit $Y_t = X_t^2$,

$$Z_t = (1,Y_{t-1},\ldots,Y_{t-p})',$$

et $\eta_t = \epsilon_t^2 - 1$. Alors (4) est équivalente à $\sigma_t^2(\alpha) = Z_t'\alpha$ et d'après (3),

$$Y_t = Z_t'\alpha + \sigma_t^2(\alpha)\eta_t, \tag{5}$$

où $\mathrm{E}(\sigma_t^2(\alpha)\eta_t \mid \mathcal{F}_{t-1}^X) = \sigma_t^2(\alpha)\,\mathrm{E}(\eta_t \mid \mathcal{F}_{t-1}^X) = 0$, $\mathcal{F}_t^X$ étant la σ-algèbre engendrée par $(X_j)_{j \le t}$. À partir de (5), Bose et Mukherjee (2003) proposent un estimateur des moindres carrés en deux étapes de α. Notre but est de généraliser cet estimateur dans le cas où la série (X_t) est observée irrégulièrement. Nous exprimons les données observées $(X_t^\star)_{1 \le t \le n}$ par (1) et nous faisons les deux hypothèses suivantes :

(H1) Le processus (a_t) est strictement stationnaire et faiblement mélangeant.

(H2) Les processus (a_t) et (X_t) sont indépendants.

Soit $A_t = \prod_{i=0}^{p} a_{t-i}$ et pour $p+1 \le t \le n$,

$$Y_t^\star = A_t Y_t = A_t Z_t'\alpha + A_t\sigma_t^2(\alpha)\eta_t = Z_t^{\star\prime}\alpha + \sigma_t^{\star 2}(\alpha)\eta_t. \tag{6}$$

En ignorant le caractère aléatoire de $\sigma_t^{\star 2}(\alpha)$ et la dépendance en α, on obtient à partir de (6) l'estimateur préliminaire des moindres carrés de α,

$$\widehat{\alpha}_{\mathrm{pr}} = \left[\sum_{t=p+1}^{n} Z_t^\star Z_t^{\star\prime}\right]^{-1} \sum_{t=p+1}^{n} Z_t^\star Y_t^\star,$$

dont on montre la consistance forte et la convergence à la vitesse $\sqrt{n}$ vers une loi gaussienne sous les hypothèses (H1), (H2) et $\mathrm{E}X_0^8 < \infty$. Ensuite, en divisant (6) par $\sigma_t^{\star 2}(\alpha)$, on obtient

$$\frac{Y_t^\star}{\sigma_t^2(\alpha)} = \frac{Z_t^{\star\prime}}{\sigma_t^2(\alpha)}\alpha + A_t\eta_t.$$

Dans cette expression, les erreurs $A_t\eta_t$ sont conditionnellement homoscédastiques. En remplaçant α par $\widehat{\alpha}_{\mathrm{pr}}$ dans $\sigma_t^2(\alpha)$, on estime α par l'estimateur des moindres carrés ordinaires $\widehat{\alpha}$ donné par

$$\widehat{\alpha} = \left[\sum_{t=p+1}^{n} \frac{Z_t^\star Z_t^{\star\prime}}{\sigma_t^4(\widehat{\alpha}_{\mathrm{pr}})}\right]^{-1} \sum_{t=p+1}^{n} \frac{Z_t^\star Y_t^\star}{\sigma_t^4(\widehat{\alpha}_{\mathrm{pr}})},$$

et on montre le résultat suivant :

Théorème 1. Soit (X_t) le processus ARCH(p) défini par (3)–(4) où $\alpha_i > 0$ pour $i = 0,\ldots,p$, et (a_t) est un processus à valeurs dans $\{0,1\}$ satisfaisant (H1), (H2) et $\mathrm{E}A_0 \neq 0$. Alors, quand $n \to \infty$,

(i) $\widehat{\alpha} \xrightarrow{p.s.} \alpha$ si $\mathrm{E}X_0^6 < \infty$,

(ii) $\sqrt{n}(\widehat{\alpha} - \alpha) \xrightarrow{loi} \mathcal{N}(0,\Sigma)$ si $\mathrm{E}X_0^8 < \infty$,

où $\Sigma = (\mathrm{E}A_0)^{-1}\,\mathrm{Var}(\epsilon_0^2)\left\{\mathrm{E}\{(\alpha' Z_0)^{-2} Z_0 Z_0'\}\right\}^{-1}$.

Remarque 1. La variance asymptotique de $\widehat{\alpha}$ est celle de l'EQMV sans donnée manquante multipliée par le facteur $(\mathrm{E}A_0)^{-1}$. Quand (a_t) est une suite iid de Bernoulli avec $q = \mathsf{P}\{a_t = 1\}$, $(\mathrm{E}A_0)^{-1} = q^{-(p+1)}$. Ce facteur augmente quand q diminue et tend vers 1 quand q tend vers 1.

Remarque 2. Un autre estimateur de α peut être obtenu en appliquant la méthode de Yule-Walker proposée par Dunsmuir et Robinson (1981a) au modèle autorégressif (5). En effet, il découle de (5) que

$$\alpha_0 = \left(1 - \sum_{i=1}^{p} \alpha_i\right) \mathrm{E}Y_0, \tag{7}$$

et $(\alpha_1, \ldots, \alpha_p)$ est l'unique solution des équations

$$\begin{pmatrix} \gamma_Y(0) & \cdots & \gamma_Y(p-1) \\ \vdots & \ddots & \vdots \\ \gamma_Y(p-1) & \cdots & \gamma_Y(0) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \begin{pmatrix} \gamma_Y(1) \\ \vdots \\ \gamma_Y(p) \end{pmatrix}, \tag{8}$$

où $\gamma_Y(k) = \mathrm{Cov}(Y_0, Y_k)$. Un estimateur $(\widehat{\alpha}_{\mathrm{yw},1}, \ldots, \widehat{\alpha}_{\mathrm{yw},p})$ de $(\alpha_1, \ldots, \alpha_p)$ est alors obtenu en remplaçant $\gamma_Y(k)$ par un estimateur approprié dans (8), par exemple

$$\widehat{\gamma}_Y(k) = \frac{\sum_{t=1}^{n-k} a_t a_{t+k} (Y_t - \widehat{\mu}_Y)(Y_{t+k} - \widehat{\mu}_Y)}{\sum_{t=1}^{n-k} a_t a_{t+k}},$$

où $\widehat{\mu}_Y = \sum_{t=1}^{n} a_t Y_t / \sum_{t=1}^{n} a_t$. Ensuite, $\widehat{\alpha}_{\mathrm{yw},0}$ est obtenu en remplaçant dans (7) α_i par $\widehat{\alpha}_{\mathrm{yw},i}$ et $\mathrm{E}Y_0$ par $\widehat{\mu}_Y$. Des résultats de simulations (non reproduits ici par manque de place) montrent que $\widehat{\alpha}$ se comporte mieux que $\widehat{\alpha}_{\mathrm{yw}} = (\widehat{\alpha}_{\mathrm{yw},0}, \widehat{\alpha}_{\mathrm{yw},1}, \ldots, \widehat{\alpha}_{\mathrm{yw},p})$ en termes de biais et variance. En particulier, $\widehat{\alpha}_{\mathrm{yw}}$ tend à surestimer α_0 et à sous-estimer les autres paramètres.

3 Application à une série réelle

À titre d'illustration, on présente les résultats numériques obtenus avec l'indice boursier journalier IPSA entre le 3 janvier 1994 et le 30 décembre 2004. Cet indice, noté (P_t) est composé des 40 principales capitalisations boursières du Chili. Parmi les 2869 jours, 127 données sont manquantes. Soit (R_t) le log-rendement défini par $R_t = \ln P_t - \ln P_{t-1}$. La figure 1 montre les 100 dernières valeurs de R_t, le diagramme Q-Q montre le caractère non-gaussien de la série, les fonctions d'autocorrélation et d'autocorrélation partielle empiriques suggèrent un modèle MA(1) ou un modèle AR(1).

Parmi les modèles ARMA(p,q) avec $1 \leq p + q \leq 2$, le modèle AR(1),

$$R_t = 0.220 R_{t-1} + X_t,$$

présente la valeur la plus petite du critère d'Akaike. La kurtosis de (X_t) est 8.958. La figure 2 montre la fonction d'autocorrélation empirique de (X_t) qui suggère bien un bruit blanc, ainsi que la fonction d'autocorrélation partielle empirique de (X_t^2) qui montre la présence d'hétéroscédasticité et indique qu'un modèle ARCH(p) avec $1 \leq p \leq 10$ peut être approprié pour (X_t).

On choisit d'ajuster un modèle ARCH(3) à la série incomplète des innovations (X_t). Le modèle ajusté est

$$\sigma_t^2 = 5.837\text{e-}5 + 0.203 X_{t-1}^2 + 0.176 X_{t-2}^2 + 0.181 X_{t-3}^2.$$

La figure 3 montre les intervalles de confiance à 90% des prédicteurs à un pas pour les 100 dernières valeurs de R_t obtenus au moyen du modèle AR(1)-ARCH(3) et du modèle AR(1). Sans surprise, les intervalles de prédiction non constants obtenus avec le modèle hétéroscédastique sont plus précis que les intervalles de prédiction constants.

Références

I. Berkes, L. Horváth, and P. Kokoszka. GARCH processes: structure and estimation. *Bernoulli*, 9(2):201–227, 2003.

P. Bloomfield. Spectral analysis with randomly missing observations. *J. Roy. Statist. Soc. Ser. B*, 32:369–380, 1970.

A. Bose and K. Mukherjee. Estimating the ARCH parameters by solving linear equations. *J. Time Ser. Anal.*, 24(2):127–136, 2003.

W. Dunsmuir and P. M. Robinson. Asymptotic theory for time series containing missing and amplitude modulated observations. *Sankhyā Ser. A*, 43(3):260–281, 1981a.

W. Dunsmuir and P. M. Robinson. Parametric estimators for stationary time series with missing observations. *Adv. in Appl. Probab.*, 13(1):129–146, 1981b.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

C. Francq and J.-M. Zakoïan. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli*, 10(4):605–637, 2004.

L. Giraitis, P. Kokoszka, and R. Leipus. Stationary ARCH models: dependence structure and central limit theorem. *Econometric Theory*, 16(1):3–22, 2000.

R. H. Jones. Spectral analysis with regularly missed observations. *Ann. Math. Statist.*, 33:455–461, 1962.

R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2002.

E. Parzen. On spectral analysis with missing observations and amplitude modulation. *Sankhyā Ser. A*, 25:383–392, 1963.

E. Parzen, editor. *Proceedings of Time Series Analysis of Irregularly Observed Data*, volume 25 of *Lect. Notes in Statistics*. Springer Verlag, New York, 1983.

P. A. Scheinok. Spectral analysis with randomly missed observations: The binomial case. *Ann. Math. Statist.*, 36:971–977, 1965.

D. Straumann. *Estimation in conditionally heteroscedastic time series models*, volume 181 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 2005.

Y. Yajima and H. Nishino. Estimation of the autocorrelation function of a stationary time series with missing observations. *Sankhyā Ser. A*, 61(2):189–207, 1999.

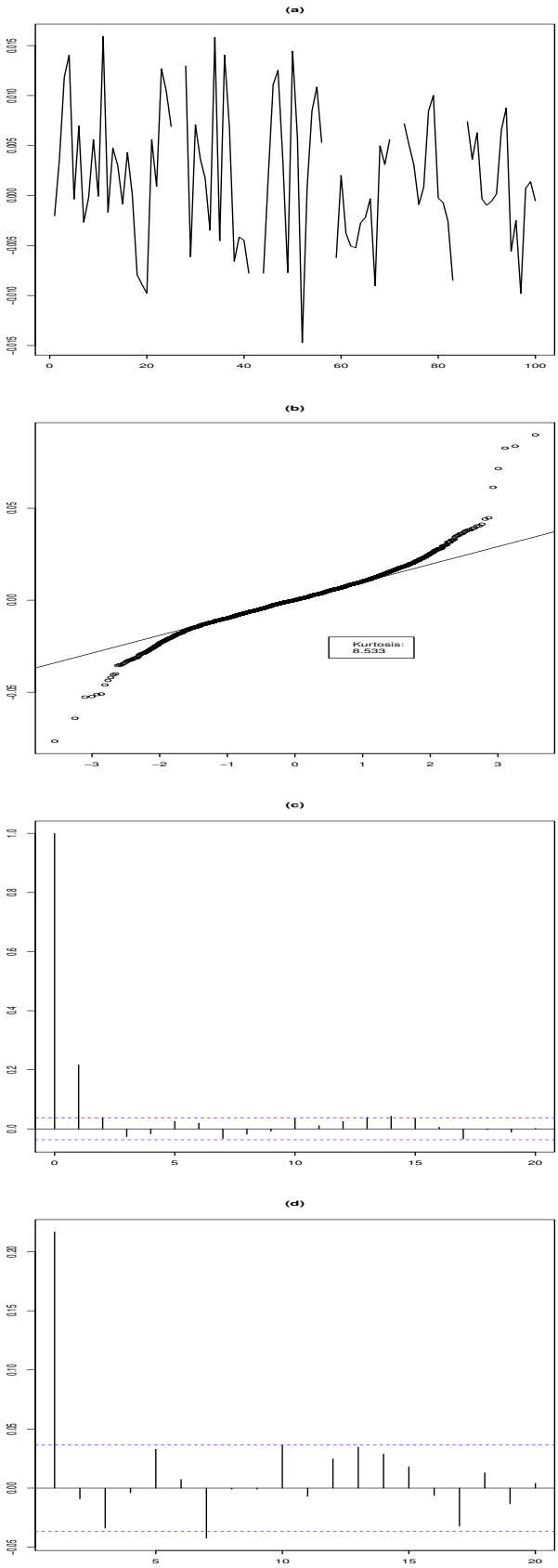

FIG. 1 – *Série (R_t); (a) 100 dernières valeurs, (b) Diagramme Q-Q, (c) Fonction d'autocorrélation empirique, (d) Fonction d'autocorrélation partielle empirique.*

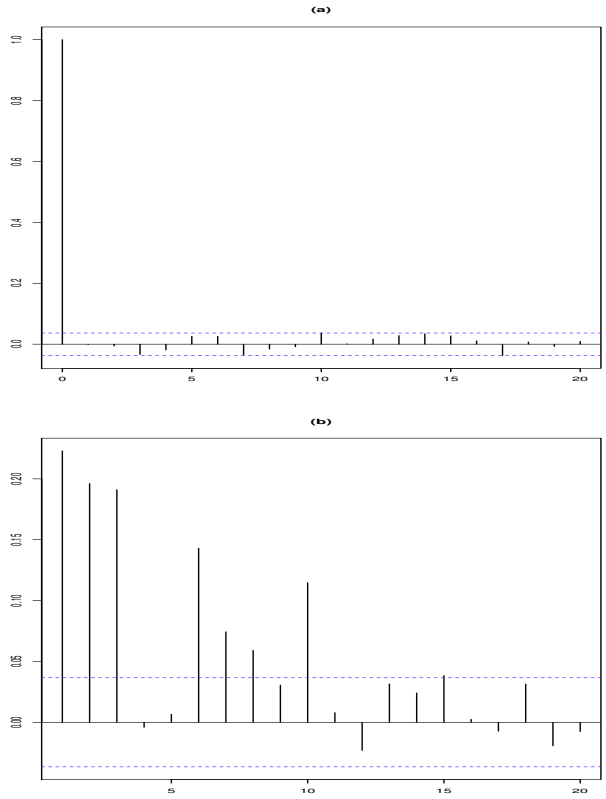

FIG. 2 – *(a) Fonction d'autocorrélation empirique de (X_t), (b) Fonction d'autocorrélation partielle empirique de (X_t^2).*

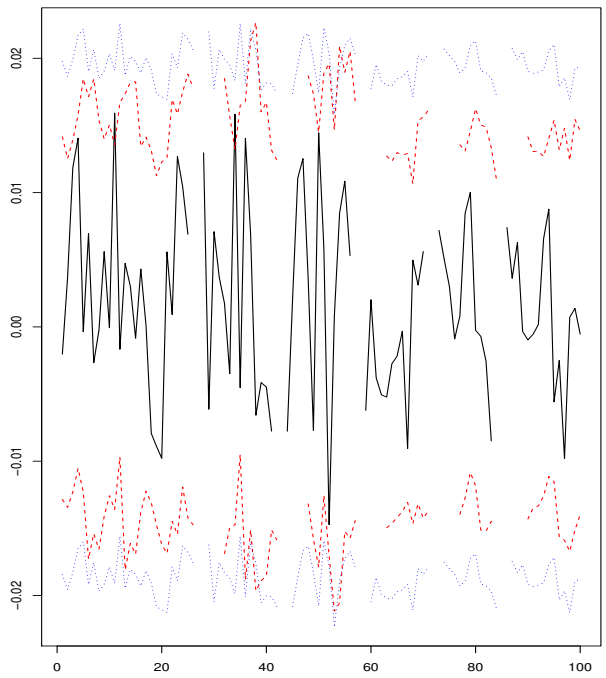

FIG. 3 – *Intervalles de prédiction à un pas pour les 100 dernières valeurs de R_t; modèle AR(1) : trait pointillé bleu; modèle AR(1)-ARCH(3) : trait hachuré rouge.*