Détection de ruptures faibles dans la moyenne des modèles CHARN

Marwa Ltaifa, Joseph Ngatchou Wandji

Détection de ruptures faibles dans la moyenne des modèles CHARN

Marwa Ltaifa [1] & Joseph Ngatchou-Wandji [2]

[1] *IECL, Université de Lorraine, France & LAMMDA, Université de Sousse, Tunisie.*
E-mail : marwa.ltaifa@univ-lorraine.fr
[2] *IECL,Université de Lorraine, France.*
E-mail : joseph.ngatchou-wandji@univ-lorraine.fr

Résumé. Dans [7] sont présentées des méthodes et stratégies pour détecter et estimer les localisations des ruptures dans la moyenne d'une grande classe de modèle autoregressifs conditionnellement non-linéaire. Nous faisons des simulations pour illustrer ces méthodes et les appliquer à la détection des ruptures dans des données du Covid-19.
Mots-clés. Séries chronologiques, Ruptures, Données du Covid-19 en France.

Abstract. In [7] is presented some methods and strategies for detectiong and estimating the locations of weak changes in the mean of classe of a Conditional Heteroscedastic AutoRegressif Non-linear model (CHARN) models. We present the results of some simulations for illustrating these methods which are apply to detecting changes in Covid-19 data.
Keywords. Time series, Breaks, Covid-19 data in France.

1 Introduction

Dans le présent travail on s'intéresse à l'étude des petites ruptures dans la moyenne des modèles CHARN. En santé, celles-ci peuvent être des signaux annonciateurs de maladies. En finance, elles peuvent annoncer une crise financières. En climatologie, elles peuvent signaler une tempête, une sécheresse, une inondation ou encore une canicule.
Des nouvelles méthodes sont proposées dans [7] pour détecter ce type de ruptures, ainsi que des stratégies pour estimer leurs localisations. Nous les appliquons à la détection des ruptures dans la moyenne des données sur le nombre de décès quotidiens du Covid-19 en France lors de la première vague. Au préalable, quelques résultats de simulation sur des données issues d'un modèle CHARN sont présentées.

2 Les méthodes

Les méthodes étudiées dans [7] reposent essentiellement sur la puissance théorique d'un test du rapport de vraisemblance pour discriminer entre modèles conditionnellement

hétéroscédastique non-linéaires CHARN). Plus précisément, soit une série d'observations $X_1, X_2, \ldots, X_n$ générée par le modèle CHARN(p) suivant

$$X_t = T(Z_{t-1}) + \gamma^\top \omega(t) + V(Z_{t-1})\varepsilon_t, \quad t \in \mathbb{Z}, \tag{1}$$

où $\gamma = (\gamma_1, \ldots, \gamma_k, \gamma_{k+1})^\top \in \mathbb{R}^{k+1}$ et pour $t_1, \ldots, t_k$, $1 < t_1 < \ldots < t_k < n$,
$\omega(t) = (\mathbb{1}_{[1,t_1[}(t), \mathbb{1}_{[t_1,t_2[}(t), \ldots, \mathbb{1}_{[t_k,n[}(t))^\top \in \{0,1\}^{k+1}$, $(X_t)_{t\in\mathbb{Z}}$ est un processus stationnaire par morceaux et ergodique, $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc centré réduit de densité f, pour tout $t \in \mathbb{Z}$, $Z_t = (X_t, \ldots, X_{t-p+1})^\top$, $p \in \mathbb{N}$, et T et V sont des fonctions réelles telles que $\inf_{x\in\mathbb{R}^p} V(x) > 0$. Parmi les travaux qui ont étudié cette classe de modèles nous pouvons citer par exemple [1], [2], [3] et [6].
Dans [7] un test du rapport de vraisemblance est construit pour tester

$$H_0 : \gamma = \gamma_0 \text{ contre } H_\beta^{(n)} : \gamma = \gamma_0 + \frac{\beta}{\sqrt{n}} = \gamma_n, \quad n > 1,$$

pour $\gamma_0 \in \mathbb{R}^{k+1}$ et $\beta \in \mathbb{R}^{k+1}$. Ces deux hypothèses se rapprochent lorsque la taille de l'échantillon grandit. On montre qu'elles sont contiguës au sens de Le Cam (voir [4]). Cette propriété permet l'étude de la puissance du test construit, et l'obtention d'une expression explicite de sa puissance. En effet, si $\psi_0 = (\rho_0^\top, \theta_0^\top)^\top \in \Theta \times \widetilde{\Theta} \subset \mathbb{R}^l \times \mathbb{R}^q$ est le vrai paramètre de nuisance du modèle (1), sous certaines hypothèses techniques, on montre que pour toute valeur de β, le test du rapport de vraisemblance construit est asymptotiquement optimal, de puissance asymptotique locale $\mathcal{P}_{k,t^k} = 1 - \Phi\left(u_\alpha - \varpi(\gamma_0, \beta)\right)$, où Φ est la fonction de répartition de la loi normale centrée réduite, u_α le quantile d'ordre $1-\alpha$, $\alpha \in (0,1)$ et ω est une fonction à valeurs réelles dont nous ne rappelons pas l'expression qui se trouve dans [7].
La première étape des méthodes décrites dans [7] consiste à déterminer, à partir du chronogramme, les m premières données $X_1, X_2, \ldots, X_m$ qui sont à peu près stationnaires, le nombre maximum K de ruptures potentielles et la distance minimale $h << n$ entre elles. Notons $\mathcal{P}_{0,t^0} = \alpha$, et considérons $\tau \in (0,1)$. Pour détecter la présence de rupture, prendre $k = 1$ et appliquer le test à tous les t_1 tels que $m \le t_1 \le n - h$.

1. Si $|\widehat{\mathcal{P}}_{1,t^1} - \mathcal{P}_{0,t^0}| \le \tau$ pour tous ces t_1, alors aucune rupture n'est détectée dans la série.
2. Si $|\widehat{\mathcal{P}}_{1,t^1} - \mathcal{P}_{0,t^0}| > \tau$ pour un t_1, alors, il existe au moins une rupture dans la série.

Pour estimer les localisations des ruptures, pour $k = 1, \ldots, K$, on suppose que $m < \tau_1^0 < \ldots < \tau_k^0 < n - h$, $\tau_j^0 - \tau_{j-1}^0 \ge h$, $j = 2, \ldots, k$, sont de potentielles localisations des ruptures obtenues du chronogramme. Soit C_j un ensemble arbitraire d'indices autour des τ_j^0, $j = 1, \ldots, k$. On considère $S_k = C_1 \times C_2 \times \ldots C_k$. Pour tout k-uplet $\tau^k = (\tau_1, \ldots, \tau_k) \in S_k$, on applique le problème de test ci-dessus avec $t_j = \tau_j$, $j = 0, \ldots, k+1$ et on calcule $\mathcal{P}_{k,t^k}$.
• À l'étape $k+1$:

1. Si $|\widehat{\mathcal{P}}_{k+1,t^{k+1}} - \mathcal{P}_{k,t^k}| \leq \tau$ et $|\widehat{\mathcal{P}}_{k,t^k} - \mathcal{P}_{k-1,t^{k-1}}| > \tau$, alors nous estimons le couple $(\widehat{k}, \widehat{t}^k)$ du nombre des ruptures et du vecteur des localisations par

$$(\widehat{k}, \widehat{t}^k) = \arg\max_{t^k \in S_k} \widehat{\mathcal{P}}_{k,t^k}.$$

2. Si $|\widehat{\mathcal{P}}_{k+1,t^{k+1}} - \mathcal{P}_{k,t^k}| > \tau$, nous répétons l'étape 1 en remplaçant k par $k+1$.

3 Ruptures et estimation des localisations

Dans ce paragraphe, nous utilisons le logiciel R pour étudier les performances des résultats théoriques obtenus dans [7]. Nous appliquons ces résultats au modèle (1) pour $p = 1$, $Z_t = X_{t-1}$, $T(Z_{t-1}) = \rho_1 + \rho_2 X_{t-1} e^{\rho_3 X_{t-1}^2}$, $\gamma = \gamma_0 + \beta/\sqrt{n}$ et $V(Z_{t-1}) = (\theta_1 + \theta_2 X_{t-1}^2 e^{-\theta_3 X_{t-1}^2})^{1/2}$, où les ρ_j, θ_j et γ_0 sont des paramètres prenant certaines valeurs à préciser, n est la taille de l'échantillon, $(\varepsilon_t)_t$ est un bruit blanc standard de densité f, et β est un réel arbitraire. Le niveau nominal considéré est $\alpha = 0.05$ et le nombre de réplications est $N = 5000$.

3.1 Simulations

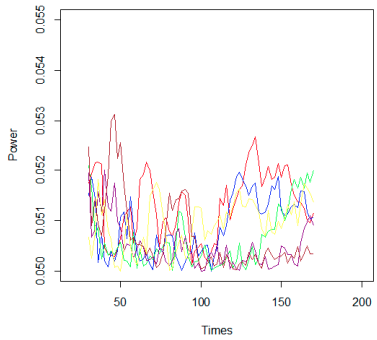
Plusieurs situations sont considérées pour évaluer la performance de notre méthode. Nous commençons par le cas d'une série d'observations stationnaires. C'est-à-dire une série sans ruptures. Nous considérons ensuite le cas d'une série comportant une rupture.

3.1.1 Aucun point de rupture dans les données

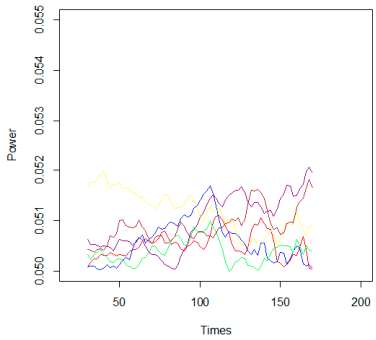
Nous calculons d'abord la puissance locale asymptotique dans le cas où le modèle ne présente aucune rupture, c'est-à-dire dans le cas où $k = 0$. Nous considérons le modèle ci-dessus lorsque $n = 200$, $\gamma_0 = 0$ et f la densité gaussienne standard. Puis, nous représentons graphiquement la puissance dans le cas où $\rho_1 = \rho_2 = 0$, $\theta_1 = 1$ et $\theta_2 = 0$ et dans le cas où $\rho_1 = 0.5$, $\rho_2 = 0$, $\theta_1 = 1$, $\theta_2 = 0$. Les deux courbes sont construites sur la Figures 1.
On observe que la puissance du test ne dépasse pas 0.053 pour $\alpha = 0.05$. Donc on accepte l'hypothèse nulle d'absence de rupture, ce qui est bien le cas ici.

3.1.2 Un seul point de rupture

Nous prenons le cas où $\rho_1 = \rho_2 = 0$, $\theta_1 = 1$ et $\theta_2 = 0$ et le cas où $\rho_1 = 0.5$, $\rho_2 = 0$, $\theta_1 = 1$, $\theta_2 = 0$. Pour chaque cas considéré, nous traçons les courbes de la puissance du test en faisant varier les instants de rupture (voir les Figures 2 (a) et 2 (b)). Par exemple pour le premier échantillon considéré, nous prenons $t_1 = 30$ c'est-à-dire $n_1(n) = 30$ et $n_2(n) = 170$ et nous traçons la puissance du test pour $\beta = (0; -0.5)$ et pour $\beta = (0; 0.8)$. Ensuite, nous prenons $t_1 = 60$, c'est-à-dire $n_1(n) = 60$ et $n_2(n) = 140$ et nous traçons la courbe de la puissance pour $\beta = (0; 0.4)$ et pour $\beta = (0; -0.4)$ ainsi de suite. On constate

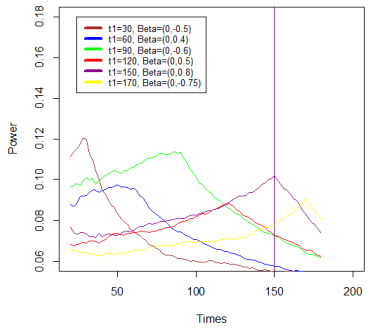


(a) $n = 200, \gamma_0 = 0, \rho_1 = \rho_2 = 0, \theta_1 = 1, \theta_2 = 0$

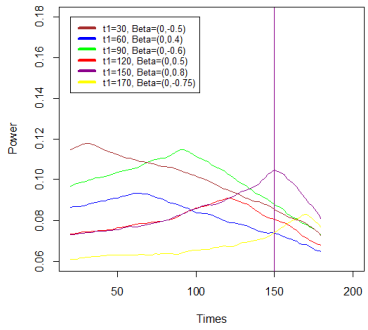


(b) $n = 200, \gamma_0 = 0, \rho_1 = 0.5, \rho_2 = 0, \theta_1 = 1, \theta_2 = 0$

FIGURE 1 – Pas de rupture



(a) $n = 200$, $\rho_1 = \rho_2 = 0$, $\theta_1 = 1$ et $\theta_2 = 0$



(b) $n = 200$, $\rho_1 = 0.5$, $\rho_2 = 0$, $\theta_1 = 1$, $\theta_2 = 0$

FIGURE 2 – Une rupture

que la puissance du test est maximale à l'instant de rupture. Donc notre test détecte bien les instants de rupture qui sont les temps donnant la plus grande puissance.

3.2 Application aux données réelles

Nous cherchons dans cette partie les points de rupture dans la série sur le décès quotidiens de COVID-19 en France dans la période du 27/02/2020 au 10/07/2020. La Figure 3 (a) correspond au chronogramme des données brutes. Nous voulons savoir si les potentiels points de rupture dans cette série, tracés en vert dans la figure, représentent réellement des points de rupture. Ces points sont $t_1 = 31, t_2 = 38, t_3 = 55, t_4 = 86$ et $t_5 = 116$ correspondants respectivement aux dates suivantes : 28/03/2020, 04/04/2020, 21/04/2020,

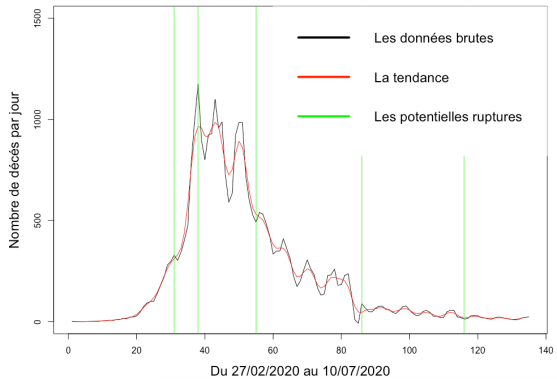
22/05/2020 et 21/06/2020. Nous commençons tout d'abord par modéliser la série que nous étudions.
D'après le graphique, celle-ci présente une tendance et ne semble pas présenter de saisonnalité. Les résultats de [7] ne peuvent pas s'appliquer directement à ces séries. Nous considérons alors la série corrigée de la tendance par la méthode des moyennes mobiles d'ordre 5. Cette série est représentée dans la Figure 3. (b). Sur chaque intervalle $[t_{i-1}, t_i)$ la série résiduelle semble stationnaire. Nous ajustons à cette série un modèle de la forme

$$X_t = \mu + (\beta_i/\sqrt{n}) + \sigma_i \varepsilon_t,$$

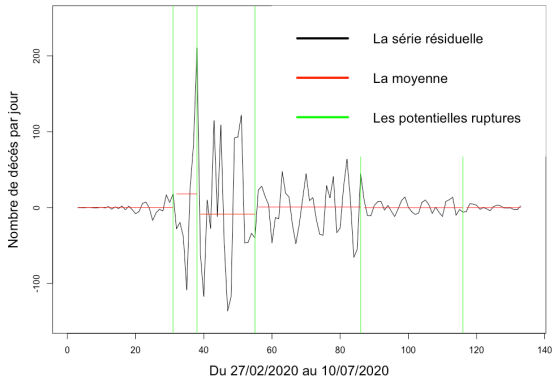
où pour tout $i \in \{1, \ldots, 6\}$, $\mu + (\beta_i/\sqrt{n})$ et σ_i sont respectivement la moyenne et la variance de X_t sur l'intervalle $[t_{i-1}, t_i)$, $\beta_1 = 0$ et pour tout $i = 2, \ldots, 6$, $\beta_i \in \mathbb{R}$, ε_t est un bruit blanc gaussien centré réduit, $t_0 = 1$, $t_6 = n$ et 1 et n sont respectivement 27/02/2020 et 10/07/2020. Cela correspond à notre problème de test pour $T(x) = \mu$, $V(x) = \sigma_i$ sur chaque intervalle $[t_{i-1}, t_i)$, $\gamma_0 = (\mu, \mu, \ldots, \mu)^\top$, $\gamma_n = (\mu, \mu + (\beta_2/\sqrt{n}), \mu + (\beta_3/\sqrt{n}), \ldots, \mu + (\beta_6/\sqrt{n}))^\top$ et $\beta = (0, \beta_2, \beta_3, \ldots, \beta_6)^\top$. Nous calculons la puissance du test autour des t_i et nous prenons les dates donnant la plus grande puissance. Nous obtenons ainsi les dates : $\widehat{t}_1 = 35, \widehat{t}_2 = 40, \widehat{t}_3 = 56, \widehat{t}_4 = 88$ et $\widehat{t}_5 = 1117$, à compter à partir du 27/02/2020. Ces dates correspondent respectivement aux dates 02/03/2020, 06/04/2020, 22/04/2020, 21/05/2020 et 25/06/2020, différentes mais assez proches des dates potentielles de rupture. Elles sont représentées dans la Figure 4.
Une interprétation possible des dates estimées est la suivante : Au tour du 02/03/2020, le nombre de décès augmente drastiquement, atteint son pic et redescend autour du 06/04/2020, puis oscille significativement jusqu'aux environs du 22/04/2020, et un peu moins significativement entre les première et deuxième phases du déconfinement, qui ont lieu le 11/05/2020 et le 02/06/2020 respectivement, jusqu'au 21/05/2020, date à partir laquelle il se stabilise avant de se réduire considérablement à partir du 25/06/2020, peu après la troisième phase du déconfinement, qui a lieu le 02/06/2020.

3.3 Comparaison avec d'autres méthodes

D'autres méthodes issues pour le CUSUM test sont étudiées dans [5] dans le but est de détecter les ruptures lorsqu'elle se produit dans les premières ou les dernières observations. Nous voulons dans ce paragraphe comparer les résultats obtenues par nos méthodes à ceux obtenues par [5] implémentées sous R. Pour cela, nous simulons une série d'observations générées par le modèle (1) pour $n = 200, \rho_1 = 0.5, \rho_2 = 0, \theta_1 = 1$ et $\theta_0 = 0$. Nous prenons le cas où on a qu'un seul point de rupture à l'instant $t = 30, 60, 90, 120, 150, 170$ et le paramètre $\beta = (0, \beta_2)$, où β_2 est une valeur arbitraire dans $\mathbb{R}$. Nous notons notre méthode NEW et les deux méthodes de [5] par SCUSUM et RCUSUM. Les résultats sont affichés dans la Table 1.
Nous remarquons que lorsque β_2 prend de petites valeurs, les SCUSUM et RCUSUM donnent de mauvais résultats. Ces deux méthodes n'estiment pas bien les localisations



(a) La série des données brutes



(b) La série résiduelle

FIGURE 3 – Nombre de décès quotidiens de COVID-19 en France du 27/02/2020 au 10/07/2020.

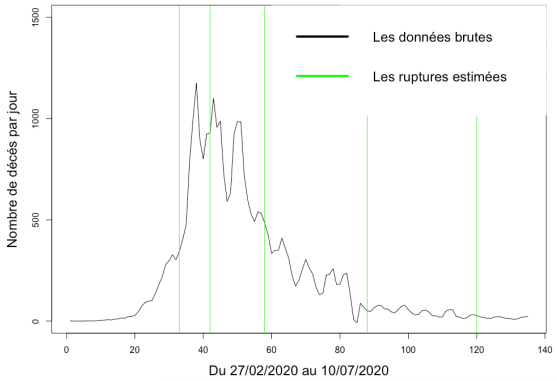


FIGURE 4 – Les dates de rupture trouvées

des points de rupture. Si β_2 prend de grandes valeurs, les deux méthodes donnent des estimations plus proches de l'instant exact de rupture, mais RCUSUM est plus performante SCUSUM. Dans tous les cas, il est clair que la méthode NEW est la plus efficace. Elle estime bien la localisation des points de rupture et donne des résultats plus proches que les deux autres méthodes, quelle que soit la valeur de β_2.
De plus, lorsque nous prenons $\beta = (0, 0)$ c'est-à-dire lorsque la série ne présente aucune rupture (lorsqu'elle est stationnaire), les SCUSUM et RCUSUM estiment des localisations des points de rupture qui n'existent pas. La méthode NEW est donc bien plus efficace et plus performante que les deux autres sur l'exemple considéré.

Méthodes	t exact $(0, \beta_2)$	30 $(0, -0.5)$	60 $(0, 0.4)$	90 $(0, -0.6)$	120 $(0, 0.5)$	150 $(0, 0.8)$	170 $(0, -0.75)$
NEW		30	60	91	120	149	170
SCUSUM		99	99	99	100	101	100
RCUSUM		91	60	69	143	79	126

Méthodes	t exact $(0, \beta_2)$	30 $(0, -5)$	60 $(0, 4)$	90 $(0, -6)$	120 $(0, 5)$	150 $(0, 8)$	170 $(0, -2)$
NEW		30	60	90	120	150	171
SCUSUM		76	81	93	113	136	105
RCUSUM		29	63	82	80	159	44

Méthodes	t exact $(0, \beta_2)$	30 $(0, -25)$	60 $(0, 24)$	90 $(0, -26)$	120 $(0, 25)$	150 $(0, 28)$	170 $(0, -22)$
NEW		30	60	90	120	150	170
SCUSUM		36	62	91	120	148	163
RCUSUM		38	61	90	121	150	166

TABLE 1 – $n = 200, \rho_1 = 0.5, \rho_2 = 0, \theta_1 = 1$ et $\theta_0 = 0$.

Références

[1] Amano, T. (2012). *Asymptotic Optimality of Estimating Function Estimator for CHARN Model*. Advances in Decision Sciences.

[2] Bardet, J. M., & Kengne, W. (2014). *Monitoring procedure for parameter change in causal time series*. Journal of Multivariate Analysis, 125, 204-221.

[3] Bardet, J. M., & Wintenberger, O. (2009) *Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes*. The Annals of Statistics, 37(5B), 2730-2759.

[4] Dreosebeke, J-J. & Fine, Inférence non paramétrique : Les statistiques de rangs. (1996). Ed. de l'Université de Bruxelles ; Ed. Ellipses.

[5] Horváth, L., Miller, C., & Rice, G. (2020). *A new class of change point test statistics of Rényi type*. Journal of Business & Economic Statistics, 38(3), 570-579.

[6] Kengne, W. C. (2012). *Testing for parameter constancy in general causal time-series models*. Journal of Time Series Analysis, 33(3), 503-518.

[7] Ngatchou-Wandji, J., & Ltaifa, M. (2021). *On detecting weak changes in the mean of CHARN models*. arXiv preprint arXiv :2101.08597.