Change point detection by Filtered Derivative with p-Value : Choice of the extra-parameters

Pierre Raphaël Bertrand*†,
Doha Hadouni*†

1er février 2016

Résumé: La méthode de dérivée filtrée avec p-value (FDpV) est un algorithme de détection de rupture *a posteriori*. Elle consiste en deux étapes : dans la première, on utilise la fonction dérivée filtrée (FD) pour sélectionner un ensemble de points de rupture potentiels ; dans la seconde étape, on calcule la *p-value* pour chaque point de rupture potentiel afin de ne retenir que les vrais positifs et rejeter les faux positifs (fausses alarmes). Le calcul de la fonction dérivée filtrée (FD) est basée sur deux extra-paramètres : le seuil de détection et la taille de la fenêtre. Nous estimons les extra-paramètres optimaux de la fonction FD, afin de diminuer le nombre de fausses alarmes (FA) et des points de ruptures non-détectés (ND). Ensuite, nous calculons l'impact d'une non-détection et d'une fausse alarme sur l'erreur quadratique moyenne. Enfin, nous simulons quelques exemples par une méthode de Monte-Carlo.

Mots clès: Détection de point de rupture ; La méthode Dérivée Filtrée avec *p-Value* ; Extra-paramètres de la dérivée filtrée ; Erreur moyenne quadratique intégrée (MISE).

Abstract: The Filtered Derivative with p-Value method concern the off-line change point detection. It is a two-step procedure. In the first step, we use the Filtered Derivative (FD) to select a set of potential change points, using its extra-parameters - namely the threshold for detection, and the sliding window size. In the second one, we calculate the p-value for each change point in order to only retain the true positives (true change points) and discard the false positives (false alarms). We estimate the extra-parameters of the function FD, in order to have the fewest possible false positives and undetected change points (ND). After setting the extra-parameters, we need to know whether the absence of detection or the false alarm has more impact on the Mean Integrated Square Error (MISE), which prompts us to calculate the MISE in both cases. Finally, we simulate some examples with a Monte-Carlo method.

Keywords: Change points detection ; Filtered Derivative with p-Value ; Filtered Derivative extra-parameters ; Mean Integrated Square Error (MISE).

Introduction

La détection de rupture *a posteriori* dans une série chronologique est une méthode statistique pertinente pour les application en finance [5, 13], médecine [7], réchauffement climatique [12],

*Laboratoire de Mathématiques, UMR CNRS 6620, Université Clermont Auvergne, France.
†Recherche supported by grant ANR-12-BS01-0016-01 entitled *"Do Well B."*

physique-chimie [8], neuro-physiologie [10, 6]. La méthode FDpV a une complexité en $\mathcal{O}(n)$, en temps de calcul et en espace mémoire, avec n taille de la série [4, 11, 9], mais le problème des fausses découvertes n'est pas totalement résolu à la première étape (FD) même si grand nombre de ces points seront rejetés à la deuxième étape (pV). Toutefois, le calcul de leur p-value augmentera le temps d'exécution du programme. Ainsi, une méthode qui permet de choisir les extra-paramètres de la fonction FD semble primordiale. De plus, nous avons cherché à comprendre l'importance de l'impact des fausses alarmes et celui de la non-détection sur l'erreur quadratique moyenne intégrée (MISE).

1 Description du modèle et rappel de la méthode FDpV

1.1 Description du modèle

Dans cette présentation, nous utiliserons le modèle simplissime suivant : Soit $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ une série indexée par le temps $\mathbf{t} = 1, 2, \ldots, n$ telle que :

- $t \longmapsto \mu(t) = \mu_k$ une fonction par morceaux constante pour tout $t \in (\tau_k, \tau_{k+1}]$;
- $\tau = (\tau_1, \ldots, \tau_K)$ une configuration de K points de rupture, par convention $\tau_0 = 0$ et $\tau_{K+1} = n$;
- $\mu = (\mu_0, \ldots, \mu_K)$ une configuration de K moyennes ;
- $\delta = (\delta_1, \ldots, \delta_K)$ une configuration des décalages où $\delta_k = \mu_k - \mu_{k-1}$, pour $k = 1, \ldots, K$;
- $X_t \in \mathcal{N}(\mu_k, \sigma^2)$ pour $t \in (\tau_k, \tau_{k+1}]$ avec $k = 0, \ldots, K$.

1.2 La méthode Dérivée Filtrée avec p-Value (FDpV)

La méthode FDpV est basée sur deux étapes. La première consiste à utiliser la fonction Dérivée Filtrée (FD) pour sélectionner les points de ruptures potentiels. La seconde étape calcule la *p-value* pour chaque point de rupture potentiel afin de ne garder que les points de ruptures réels. La méthode est définie précisément dans ce qui suit.

Etape 1 : Dérivée Filtrée

La première étape dépend de deux paramètres : la taille de la fenêtre A et le seuil de séléction C_1.

1. **Calcul de la fonction Dérivée Filtrée :**

   Définition 1.1 *La fonction Dérivée Filtrée est définie par la formule suivante :*

   $$FD(t, A) = \widehat{\mu}(X, [t+1, t+A]) - \widehat{\mu}(X, [t-A+1, t]), \quad \text{pour } A < t < n - A, \tag{1.1}$$

   avec $\widehat{\mu}(X, [u, v]) := \frac{1}{(v-u+1)} \times \sum_{t=u}^{v} X_t$ *la moyenne empirique des variables* X_t *avec* $t \in [u, v]$

   Cette méthode consiste à filtrer les données en calculant les estimateurs du paramètres μ avant d'appliquer une dérivation discrète. Ce qui explique le nom *"méthode de la Dérivée Filtrée "* [2, 1]. La quantité $A \times FD(t, A)$ peut être calculée de manière itérative en utilisant

   $$A \times FD(t+1, A) = A \times FD(t, A) + X(t+1+A) - 2X(t+1) + X(t-A+1)$$

2. **Détermination des points de rupture potentiels**

Définition 1.2 *Les points de rupture potentiels notés τ_k^*, pour $k = 1, \ldots, K^*$ sont les maximums locaux de la valeur absolue de la fonction dérivée filtrée $|FD(t, A)|$.*

Pour la réalisation de la détection, on suit l'algorithme suivant :

(a) Sélectionner le point de rupture potentiels τ_k^* qui est le maximum global de la fonction $|FD_k(t, A)|$,

(b) Définir les valeurs de la fonction FD_{k+1} à zero afin que la largeur de l'amplitude du point de rupture τ_k^* soit égale à $2A$.

(c) Itérer cet algorithme tant que $|FD_k(\tau_k^*, A)| > C_1$.

Etape 2 : Calcul de la p-Value

Un point de rupture potentiel peut être une fausse alarme ou une estimation d'un point de rupture réel. Dans le deuxième cas, il existe alors une erreur d'estimation ε sur la location du point de rupture qu'il faut rectifier pour chaque point de rupture τ_k^*, [3, 4]. Donc, pour chaque intervalle, on calcule un estimateur de la moyenne

$$\widehat{\mu}_k := \widehat{\mu}(X, [\tau_k + \varepsilon_k, \tau_{k+1} - \varepsilon_{k+1}]).$$

Ensuite, on élimine les fausses alarmes pour ne garder que les vrais points de rupture. Dans [3], on applique le test d'hypothèses suivant :

$$(H_{0,k}) : \widehat{\mu}_k = \widehat{\mu}_{k+1} \quad \text{versus} \quad (H_{1,k}) : \widehat{\mu}_k \neq \widehat{\mu}_{k+1} \quad \text{pour tout } 1 \leq k \leq K$$

Enfin, on calcule les *p-values* $p_1^*, ..., p_K^*$ associées chaque point de rupture potentiel $\tau_1^*, ..., \tau_K^*$ tel que

$$p_k^\star \;=\; 2 \times \Big\{1 - St_d(|t_k|)\Big\} \;\simeq\; 2 \times \Big\{1 - \Phi(|t_k|)\Big\}, \tag{1.2}$$

avec : - St_d la fonction de répartition de la loi de Student de degré $d = N_k + N_{k-1} - 2$ tel que $N_k = \Big\{(\tau_0^* - \varepsilon_0) - (\tau_k^* + \varepsilon_0)\Big\}$,
- Φ la fonction de répartition de la loi Gaussienne
- t_k^* suit la loi de Student de degré d avec

$$t_k^* = \frac{\widehat{\mu}_k - \widehat{\mu}_{k-1}}{\sqrt{\frac{S_{k-1}^2}{N_{k-1}} + \frac{S_k^2}{N_k}}},$$

où $S_k = \sqrt{\Big(\frac{1}{N_k}\sum_{t=\tau_k+\varepsilon_0}^{\tau_{k+1}-\varepsilon_0} X_t^2\Big) - \widehat{\mu}_k^2}$ est l'écart type.
Le point de rupture est validé si $p_k^* < p^*$ (p^* un seuil de détection).

2 Choix des extra-paramètres de l'étape 1

Toutes les méthodes d'analyse de points de ruptures dépendent d'extra-paramètres. La méthode FDpV dépend de deux extra-paramètres : la taille de la fenêtre A et le seuil C_1.

2.1 Monte-Carlo simulation

Ces simulations Monte-Carlo sont faites pour $M = 1.000$. Soit $(X_1^j, X_2^j, \ldots, X_n^j)$ une séquence de variables aléatoires gaussiennes simulées où $n = 5.000$ et $j = 1, \ldots, M$, avec une variance $\sigma^2 = 1$ et une moyenne $\mu_t = f(t)$ où f est une fonction par morceaux constante avec un certain nombre de points de rupture en des temps différents τ et avec différentes moyennes μ. Dans chaque exemples[1], nous appliquons la méthode FDpV pour différentes valeurs des extra-paramètres A and C_1. On fait varier le paramètre A entre 20 et 220 en rajoutant 10 et le paramètre C_1 varie entre 0.1 et 1 en rajoutant 0.05.

Example 1

Les figures 1 and 2 représentent respectivement l'évolution du nombre moyen des points de rupture non-détectés et l'évolution du nombre moyen des fausses alarmes en fonction des extra-paramètres A and C_1 où $n = 5000$; $\tau = (1000, 1250, 1500, 2000, 3500, 4000, 4500)$ et $\delta\mu = (-0.5, -1, -1.5, 0.5, 1, 1.5)$ avec $\delta_0 = 0.5$.

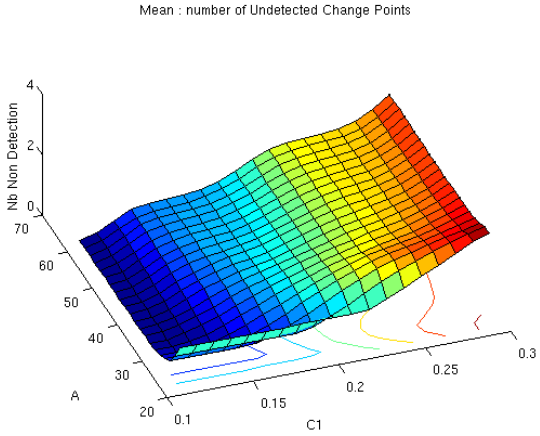


FIGURE 1 – The mean of the number of undetected change points

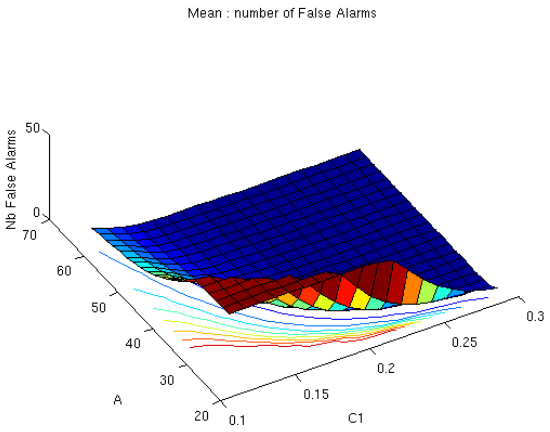
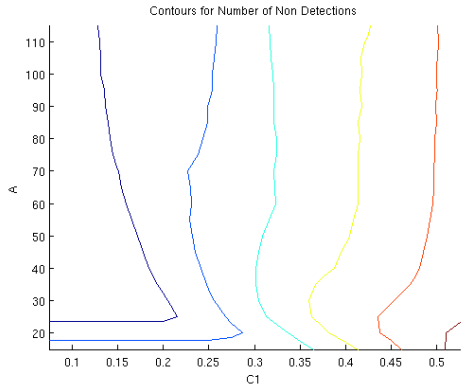


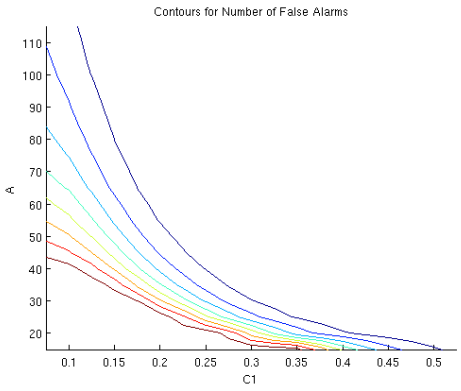
FIGURE 2 – The mean of the number of false alarm

Ces figures 3(a) et 3(b), ci dessous, représentent les contours du nombre moyen des points de rupture non-détectés et ceux du nombre moyen des fausses alarmes. Ces contours aident à définir un ensemble admissible d'extra-paramètres. La zone hachurée dans la figure 3(c) correspond à l'intersection de zero ou un point de rupture non détecté et zero ou une fausse alarme. Dans ce cas, le choix d'une taille de fenêtre $A \geq 70$ et d'un seuil $C_1 \in [0.15, 0.22]$ assure un nombre moyen des points de rupture non-détectés et un nombre moyen de fausses alarmes inférieur à 1.

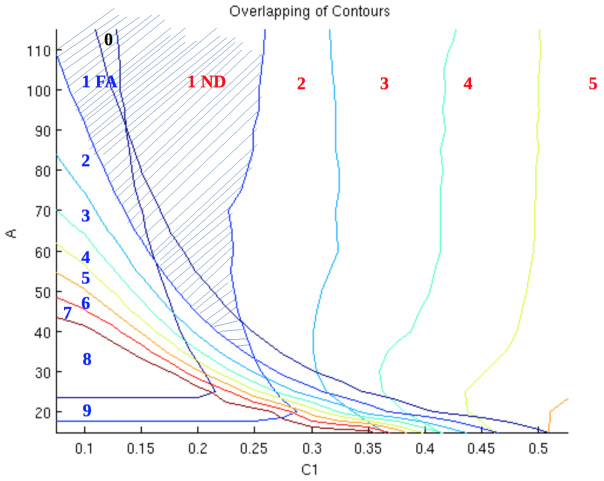
1. L'algorithme de la méthode FDpV et les simulations Monte-Carlo ont été implémentées par Guillaume Paugam (ingénieur en informatique), en tant que membre du projet ANR "Do Well B.", grant ANR-12-BS01-0016-01.



(a) The contour of the mean number of undetected change points.



(b) The contour of the mean number of false alarm.



(c) The overlap contour of 3(a) and 3(b).

FIGURE 3 – The contours of the figures 1 and 2 and their overlap

2.2 L'impact du MISE

Définition 2.1 (Mean Integrated Square Error) *Soit τ_1 et τ_2 deux points de rupture. L'erreur moyenne quadratique intégrée entre τ_1 et τ_2 est définie par*

$$MISE(\tau_1, \tau_2) := \mathbb{E}\left(\sum_{t=\tau_1}^{\tau_2} |\widehat{s}(t) - s(t)|^2\right).$$

où $s(t) = \sum_{k=0}^{K} \mu_k \times \mathbf{1}_{(\tau_k, \tau_{k+1}]}(t)$ est le signal traité et $\widehat{s}(t) = \sum_{k=0}^{\widehat{K}} \widehat{\mu}_k \times \mathbf{1}_{(\widehat{\tau}_k, \widehat{\tau}_{k+1}]}(t)$ le signal estimé avec $\widehat{\mu}_k = \mu_k \pm \epsilon$.

Proposition 2.2 *On définit $\Delta MISE_{ND} = MISE_{withND} - MISE_{withoutND}$ et $\Delta MISE_{FA} = MISE_{withFA} - MISE_{withouFA}$*

- *Si $\left(\dfrac{\mu_2 - \mu_1}{\sigma}\right)^2 > 2$, alors $\Delta MISE_{ND} > \Delta MISE_{FA}$, l'impact des non détections est plus important que celui des fausses alarmes.*
- *Si $\left(\dfrac{\mu_2 - \mu_1}{\sigma}\right)^2 < \dfrac{8}{\tau_3 - \tau_1}$, alors $\Delta MISE_{ND} < \Delta MISE_{FA}$, l'impact des non détections est moins important que celui des fausses alarmes.*

Conclusion

Notre analyse suggère une méthode générale pour optimiser les paramètres de la fonction dérivée filtrée dans la première étape de la méthode FDpV. En donnant les valeurs précises à choisir pour la taille de la fenêtre et le seuil de détection, $A > 70$ et $C_1 \in [0,15, 0,22]$, on réduit le nombre de fausses découvertes. Ainsi, dans la deuxième étape, on calcule moins de p-valeurs, ce qui nous permet de gagner en temps de calcul et en mémoire. Enfin, nous avons étudié l'impact des fausses alarmes et les points de rupture non détectés sur l'erreur quadratique moyenne intégrée (MISE) afin de déterminer celle qui a un impact plus important.

Références

[1] M. Basseville and I. V. Nikiforov, *Detection of abrupt changes : theory and application*, Prentice Hall Information and System Sciences Series, Prentice Hall Inc., Englewood Cliffs, NJ, 1993. MR MR1210954 (95g :62153)

[2] A. Benveniste and M. Basseville, *Detection of abrupt changes in signals and dynamical systems : some statistical aspects*, Analysis and optimization of systems, Part 1 (Nice, (1984), Lecture Notes in Control and Inform. Sci., vol. 62, Springer, Berlin, 1984, pp. 145–155. MR MR876686

[3] P. R. Bertrand, *A local method for estimating change points : the "hat-function"*, Statistics **34** (2000), no. 3, 215–235. MR MR1802728 (2001j :62032)

[4] P. R. Bertrand, M. Fhima, and A Guillin, *Off-line detection of multiple change points by the filtered derivative with p-value method*, Sequential Analysis **30 (2)** (2011), 172–207.

[5] S. Bianchi, A. Pantanella, and A. Painese, *Efficient markets and behacioral finance : a comprehensive multifractal model*, Advances in Complex Systems **18** (2015).

[6] S. Grun, M. Diesmann, and A. Aertsen, *Unitary events in multiple single neuron spiking activity : I. detection and significance.*, Neural Comput. **14** (Jan, 2002).

[7] N. Khalfa, P.R. Bertrand, G. Boudet, A. Chamoux, and V. Billat, *Heart rate regulation processed through wavelet analysis and change detection.*, Some case studies, Acta Biotheoretica. **60** (2012), 109–129.

[8] S.C. Lim and L.P. Teo, *Modeling single-file diffusion by step fractional brownian motion and generalized fractional langevin equation.*, Journal of Statistical Mechanics : Theory and Experiment **2009** (2009), no. 8.

[9] M. Messer, M. Kirchener, J. Shiemann, J. Roeper, R. Neininger, and G. Schneider, *A multiple filter test for the detection of rate changes in renewal processes with varying variance*, The Annals of Applied Statistics **8** (2015), no. 4, 2027–2067.

[10] G. Schneider, *Messages of oscillatory correlograms : A spike train model*, Neural Comput **20** (2008), 1211–1238.

[11] Y. S. Soh and V. Chandrasekaran, *High-dimensional change-point estimation : Combining filtering with convex optimization*, Article in Applied and Computational Harmonic Analysis (January 2015).

[12] W. Wang, I. Bobojonov, W.K. Hardle, and M. Odening, *Testing for increasing weather risk*, Stochasitic Environmental Research and Risk Assessment **27** (2013), 1565–1574.

[13] W. Xiao, W. Zhang, and X. Zhang, *Parameter identification for drift fractional brownian motions with application to the chinese stock markets*, Communications in Statistics - Simulation and Computation **44** (2015), 2117–2136.