THÈSE DE DOCTORAT

DÉTECTION DES CHANGEMENTS DE POINTS MULTIPLES ET INFÉRENCE DU MODÈLE AUTORÉGRESSIF Á SEUIL.

soumis par

ELMI MOHAMED

Le 30 mars 2018

pour avoir le grade de

docteur de l'**Université Bourgogne Franche-Comté École Doctorale Carnot-Pasteur**

Discipline: **Mathématiques et applications**

Composition du jury :

Pierre BERTRAND	Université Clermont Auvergne (co-directeur)
Hermine BIERMÉ	Université de Poitiers (rapporteur)
Marianne CLAUSEL	Université de Lorraine (rapporteur)
Serguei DACHIAN	Université Lille 1 (examinateur)
Clément DOMBRY	Université Bourgogne Franche-Comté (examinateur)
Bruno SAUSSEREAU	Université Bourgogne Franche-Comté (co-directeur)

Laboratoire de Mathématiques de Besançon
16 route de Gray
25030 Besançon, France

École doctorale Carnot-Pasteur

Remerciements

Me voila enfin à la veille de ma soutenance, j'essayerai d'écrire quelques lignes pour remercier l'ensemble des personnes qui ont contribué à réaliser mes travaux de thèse. Je voudrais tout d'abord remercier les deux rapporteurs de cette thèse, les professeurs Hermine Biermé et Marianne Clausel qui ont eu la patience et la gentillesse de lire et de relire mon manuscrit, ainsi que les autres membres du jury, les professeurs Clément Dombry et Serguei Darchian qui ont eu l'amabilité de participer à ma soutenance.

Ce travail n'aurait jamais vu le jour sans mes deux directeurs de thèse, Messieurs Pierre Bertand et Bruno Saussereau. Franchement, à Monsieur Pierre Bertrand, je remercie énormément pour sa disponibilité, son écoute, ses conseils avisés et ses qualités humaines lors de mon séjour au laboratoire de Clermont Ferrand. A Monsieur Bruno Saussereau, je remercie pour l'excellence de son soutien scientifique et de la grande confiance qu'il a bien voulu m'accorder malgré la difficulté du sujet traité.

Un immense merci également à mes collègues du bureau 328 et à l'ensemble des membres du Laboratoire de Mathématiques de Besançon(LMB). Je remercie aussi le Doyen de la faculté de Sciences de l'Université de Djibouti Monsieur Ramadan Ali et le directeur des études du département de mathématiques et informatiques de l'université de Djibouti Monsieur Ibrahim Abdi. Je dis Merci à mon ami l'Honorable député Ali Soubaneh qui m'a vraiment soutenu sur mes périodes de doutes.

Je ne pourrais clôturer ces remerciements sans me retourner à ma famille, j'offre ce travail de thèse à mes Parents et à celle que je l'appelle toujours mon Bonheur, ma perle, ma femme BILAN, ainsi que tous mes enfants et frères : Hassan, Ahmed, Ibrahim, Hadi, Safa, Bahdon et à Warfa. Je suis content de vous dire tout simplement MERCI.

Résumé

Ma thèse est composée de deux parties : une première partie traite le problème de changement de régime et une deuxième partie concerne le processus autorégressif à seuil dont les innovations ne sont pas indépendantes. Toutefois, ces deux domaines de la statistique et des probabilités se rejoignent dans la littérature et donc dans mon projet de recherche. Dans la première partie, nous étudions le problème de changements de régime. Il existe plusieurs méthodes pour la détection de ruptures mais les principales méthodes sont : la méthode de moindres carrés pénalisés (PLS) et la méthode de derivée filtrée (FD) introduit par Basseville et Nikirov. D'autres méthodes existent telles que la méthode Bayésienne de changement de points.
Dans cette thèse, nous avons amélioré la méthode FD :

- Nous avons ajouté une deuxième étape nommée taux de fausses découvertes afin de mieux détecter les vrais instants de ruptures et d'avoir le moins de fausses alarmes possible. Ceci a fait l'objet d'une publication dans "International Journal of Statistics and Probability, Vol 1 N 1 p 12–23, 2014".

- L'algorithme de FD depend de deux paramètres : le seuil et la fenêtre de calcul. Nous avons donné des paramètres optimaux et ceci a fait l'objet d'une publication dans "International Journal of Statistics and Probability, Vol 3 N 3 p 29–43, 2014".

D'autre part, nous avons validé la nouvelle méthode de dérivée filtrée et taux de fausses découvertes (FDqV) sur des données réelles (des données du vent sur des éoliennes et des données du battement du coeur). Bien naturellement, nous avons donné une extension de la méthode FDqV sur le cas des variables aléatoires faiblement dépendantes.

Dans la deuxième partie, nous étudions le modèle autorégressif à seuil (en anglais Threshold Autoregessive Model (TAR)). Le TAR est étudié dans la littérature par plusieurs auteurs tels que Tong(1983), Petrucelli(1984, 1986), Chan(1993). Les applications du modèle TAR sont nombreuses par exemple en économie, en biologie, l'environnement, etc. Jusqu'à présent, le modèle TAR étudié concerne le cas où les innovations sont indépendantes. Dans ce projet, nous avons étudié le cas où les innovations sont non corrélées. Nous avons établi les comportements asymptotiques des estimateurs du modèle. Ces résultats concernent la convergence presque sûre, la convergence en loi et la convergence uniforme des paramètres.

Mots-clefs

séries temporelles, dérivée filtrée, taux de fausses découvertes, Programmation dynamique, modèle autorégressif et mobile moyenne, hors ligne, détection de points, modéle autorégressif à seuil, Processus de Poisson composé.

Abstract

This thesis has two parts : the first part deals the change points problem and the second concerns the weak threshold autoregressive model (TAR) ; the errors are not correlated.
In the first part, we treat the change point analysis. In the litterature, it exists two popular methods : The Penalized Least Square (PLS) and the Filtered Derivative introduced by Basseville end Nikirov.
Others methods such that the Bayesian change points exist in the litterature.
In this project, we improve the FD method :

- We added a second step nammed the False Discovery Rate (FDR), then we detect the true abrupt change point and to have less false alarms as possible as. These results were published in "International Journal of Statistics and Probability", vol 1 n 1 p12–23, 2014.
- The FD algorithm depends two parameters : the thresholdand the window for calculus. We gave the optimized parameters. This is published by "International Journal of Statistics and Probability", vol 3 n 3 p29–43.

In other hand, we gave the new method of filtered derivative and false discovery rate (FDqV) on real data (the wind turbines and heartbeats series). Also, we studied an extension of FDqV method on weakly dependent random variables.

In the second part, we spotlight the weak threshold autoregressive (TAR) model. The TAR model is studied by many authors such that Tong(1983), Petrucelli(1984, 1986). there exist many applications, for example in economics, biological and many others. The weak TAR model treated is the case where the innovations are not correlated.

Keywords

Time series, Filtered Derivative, False Discovery Rate, Dynamical programming, Autoregressive and moving average model, off-line, change points detection, threshold autoregressive model, compound Poisson Process.

Table des matières

Chapitre 1

Introduction

1.1 Motivations

1.1.1 Motivation pour la première partie

Dû au progrès technologique, la taille de jeux de données devient de plus en plus grande. Il s'avère que la détection de ruptures nécessite donc des méthodes rapides en calcul et peu coûteuses en mémoire. Il existe deux grandes méthodes souvent utilisées dans la littérature de détection de ruptures : la méthode de dérivée filtrée (FD) et la méthode de moindre carré pénalisée (PLS). Dans cette thèse, nous proposons d'améliorer la méthode FD. Dans un jeu de données de taille N, la méthode PLS a besoin d'une matrice $N \times N$ car elle est basée par la programmation dynamique et la méthode FD nécessite un vecteur de taille N car elle est basée par la méthode des moyennes mobiles. Lorsqu' on applique la méthode FD dans une série d'observations indépendantes dont les moyennes changent, on remarque qu'il existe parmi les instants estimés, des vrais instants mais aussi beaucoup de fausses alarmes. Pour séparer les vraies et les fausses ruptures, Bertrand, Fihima et Guillin ajoute une deuxième étape en faisant un test simple qui consiste à tester l'hypothèse nulle ; il n'y a pas de ruptures contre l'hypothèse alternative ; il y a une rupture entre deux moyennes consécutives et ont nommé la méthode de Dérivée Filtrée avec p-value (FDpV). Dans ce projet de recherche, nous remplaçons la deuxième étape par un test multiple : la méthode de Benjamini et Hochberg où le taux de fausses découvertes. Nous l'avons appelée la méthode de Dérivée Filtrée et le Taux de Fausses Découvertes (FDqV). La méthode FD dépend de deux paramètres à savoir le seuil et la fenêtre du calcul de moyennes, nous proposons dans cette thèse des paramètres optimaux afin de détecter l'ensemble des vraies ruptures et d'avoir le moins de fausses alarmes. Dans ce rapport, la méthode FDqV est appliquée aussi dans le cas où les variables aléatoires sont faiblement dépendantes. Nous l'appliquons sur des données réelles des battements du coeur, car on a remarqué que ces variables des battements du coeur ne sont pas indépendantes mais plutôt faiblement dépendantes.

1.1.2 Motivation pour la deuxième partie

Le modèle autorégressif à seuil est beaucoup étudié dans la littérature avec les hypothèses fortes sur les erreurs. D'autre part, Francq et Zakoian [42] ont considéré le problème de l'estimation du modèle autorégressif et moyenne mobile (ARMA) avec des hypothèses faibles sur les innovations. Bien naturellement, nous considérons le modèle autorégressif à seuil avec seulement les hypothèses d'ergodicité, de mélange et avec des erreurs non corrélées. Dans notre cas, nous établissons des théorèmes sur la convergence presque sûre,

la convergence uniforme et la convergence normale des paramètres du modèle. Nous avons aussi obtenu que le seuil du processus converge vers un Poisson composé dans le cas mélangeant. Ces recherches ouvrent des perspectives sur l'exploration du modèle TAR faible.

1.2 Revue de littérature des méthodes de détection de ruptures

1.2.1 La méthode de Dérivée Filtrée

modèle

Soit $X = (X_1, X_2 \ldots, X_n)$ une série indexée par le temps t=(1,2,...,n). On suppose qu'il existe une segmentation $\tau = (\tau_1, \tau_2, \ldots, \tau_K)$ tel que (X_t) est une suite des variables aléatoires indépendantes et identiquement distribuées (iid) pour $(\tau_k, \tau_{k+1}]$, où par convention $\tau_o = 1$ et $\tau_{K+1} = n$. Le modèle le plus simple est celui où X_t est une suite de variables aléatoires indépendantes gaussiennes avec $X_t \in \mathcal{N}(\mu(t), \sigma)$, avec $\mathcal{N}(\mu, \sigma)$ est la loi gaussienne de moyenne μ et de variance σ, t$\rightarrow \mu(t)$ est une fonction constante par morceaux, $\mu(t) = \mu_k$ pour tout k $\in (\tau_k, \tau_{k+1}]$.

la méthode de dérivée filtrée

La méthode de dérivée filtrée est introduite par Basseville et Nikirov [6] et elle utilise la méthode de moyennes mobiles. On choisit un seuil C et une fenêtre A pour le calcul de moyennes. On considère la fonction FD suivante :

$$FD(t, A) = \hat{\theta}(t + 1, t + A) - \hat{\theta}(t - A, t)$$

où $\hat{\theta}$ est la moyenne empirique du paramètre d'interêt θ sur l'intervalle $(t - A, t + A)$. Dans le cas où les observations sont complètement connues, on détecte le premier instant de rupture comme l'argument du maximum de la fonction FD dépassant le seuil choisi C. On pose $FD([\tau_1 - A : \tau_1 + A]) = 0$ et on recommence le même processus pour trouver le deuxième instant de rupture et ainsi de suite.
Finalement on obtient les instants de ruptures potentiels $(\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{Kmax})$ ou $Kmax$ est choisi par l'utilisateur.

La méthode de p-value

En 2011, Bertrand, Fihima et Guilin [11] ont remarqué qu'il existe beaucoup de fausses alarmes. Pour séparer les bons instants et les fausses alarmes, ils ont ajouté une deuxième étape qu'ils ont nommée "p-value". En effet, dans [11], ils ont fait un test statistique simple pour comparer si deux moyennes consécutives sont égales ou pas (voir en détails dans [11]. Á la fin de la deuxième étape, on obtient des bons instants mais également il existe encore des fausses alarmes. Nous avons eu l'idée d'utiliser un test multiple pour améliorer la méthode FD.

Le taux de fausses découvertes

En 2014, nous avons eu l'idée de remplacer la deuxième étape par un test multiple et utiliser la méthode de Benjamini et Hochberg[7] pour séparer les bons instants et les fausses alarmes. Nous avons eu des résultats meilleurs que ceux dans [11]. Cette nouvelle

méthode consiste à ajouter á la méthode FD, une deuxième étape qui fait un test multiple en choisissant un taux de fausses alarmes, voir plus en détails [35].

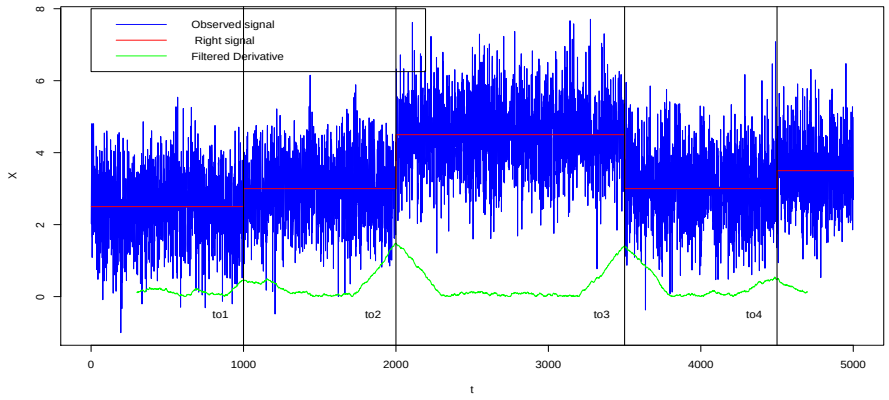


FIGURE 1.1 – *The right signal (red), the noisy signal (blue), and Filtered Derivative function (green).*

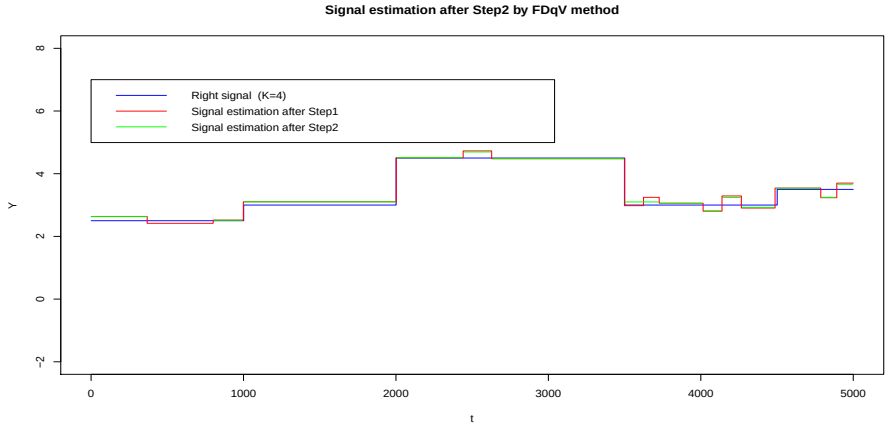


FIGURE 1.2 – *Signal reconstruction after Step2 by FDqV method*

1.2.2 La méthode des moindres carrés pénalisés

Cas où le nombre des instants de ruptures est connu.
Lorsque K, le nombre des instants de ruptures est connu, on utilise l'algorithme de la programmation dynamique. On pose $J(\tau_1^p, \tau_2^p, \ldots, \tau_r^p)$ la somme des carrés résiduels associés à la partition optimale contenant les r ruptures sur les p premières observations. Nous avons alors $J((\tau_1^p, \tau_2^p, \ldots, \tau_r^p) = \min_{r \leq j_1 \leq n-1}[J(\{\tau_1^h, \tau_2^h, \ldots, \tau_{r-1}^p\} + J(j_1 + 1, n)]$. Pour plus en détail, voir [4].

Cas où le nombre des instants de ruptures est inconnu.

Lorsque le nombre de ruptures K est inconnu, l'idée consiste alors à pénaliser l'ajout d'un point de ruptures en définissant une nouvelle fonction de contraste $U(K, \tau, X, \beta) = J(K, \tau, X, \beta) + \beta pen(K)$. Plusieurs auteurs ont proposé différents paramètres de pénalisations :

1. Le critère d'information de Schwartz en posant $\beta = \beta_n = \frac{\log n}{n}$ et $pen(K) = K$. L'inconvénient est que ceci surestime le nombre de ruptures. Voir plus en détails [56].

2. La pénalisation de Birgé et Massart, en posant $\beta = \beta_n = \frac{2\sigma^2}{n}$ et $pen(K) = K(1 + clog(\frac{n}{K}))$ avec $c = 2.5$. L'inconvénient est qu'il s'applique uniquement aux processus de variance constante et ne permet pas la détection de rupture sur la variance. Voir plus en détails [13].

3. La méthode adaptative de Lavielle [57] consiste à observer le tracé de la fonction J en fonction du pen(K). Pour un processus gaussien indépendant dépourvu de ruptures, on remarque que la fonction J(K) coïncide avec la fonction $f(K) = a \times K + b \times K \times \log(K) + e_K$, avec e_K une suite de variables aléatoires gaussiennes, centrées et indépendantes. L'algorithme de la méthode adaptative se décompose de cette manière :

   - On choisit $Kmax$, le nombre maximum de nombres de ruptures. $\forall\ 1 \leq K \leq Kmax$, on ajuste le modèle $f(K) = a \times K + b \times K \times \log(K) + e_K$ à la série J.
   - On évalue la probabilité que J suive ce modèle. C'est à dire qu'on estime la probabilité suivante

   $$P_K = \mathbb{P}(e_K \geq J - a \times K - b \times K \times \log K).$$

   - Le nombre estimé sera la plus grande valeur de K telle que la P-valeur P_K soit la plus petite qu'un seuil donné α.

   Notons dans l'algorithme, les coefficients changent, voir plus en détails [57].

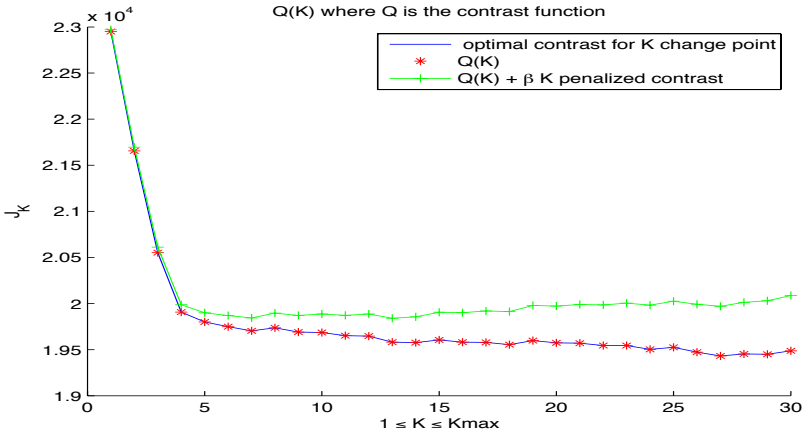


FIGURE 1.3 – *blue : Q(K) calculated with dynamical program method ; red : the penalized contrast function ; green : the optimal contrast function for K change points .*

1.2.3 La méthode Bayesienne

L'approche de la méthode bayesienne pour la détection des ruptures consiste à considérer un modèle probabiliste sur le vecteur des données $X = (X_1, \ldots, X_n)$. La distribution et la densité du vecteur X sont reliées par le vecteur des paramètres $\Theta = (\theta_1, \ldots, \theta_n)$, la densité est donc notée par $f(X/\Theta)$. Les paramètres inconnus sont représentés par des variables aléatoires, dont les densités de probabilités sont soit déterminés à priori par les informations dont on dispose soit exprimés par les lois non informatives. L'estimation est basé par la formule célèbre de Bayes :

$$f(X/\Theta) = \frac{L(X/\Theta)f(\Theta)}{\int L(X/\Theta)f(\Theta)}$$

où $f(\Theta/X)$ est l'expression de la densité de probabilité à posteriori de la variable Θ inconnue par rapport à X, $L(X/\Theta)$ est la fonction de vraisemblance des données par rapport aux paramètres Θ, et $f(\Theta)$ est la densité de probabilité jointe des θ_i, $1 \leq i \leq n$. Sous cette représentation, toute la distribution des paramètres de Θ par rapport aux données est disponible mais le but principal est de déterminer une valeur de Θ, une manière de trouver une telle valeur est de maximiser la densité de $f(\Theta/X)$. Dans ce cas on approxime la densité $f(\Theta/X)$ par $L(X/\Theta)f(\Theta)$.

Pour construire un modèle, il faut définir les distributions des données X en fonction du vecteur des paramètres X et déterminer la loi à priori de Θ. Ce choix n'est pas forcément simple en raison d'un manque des informations à priori et par exemple les distributions peuvent admettre des hyperparamètres à estimer ou à négliger tout simplement. Dans [20], on traite la question des modèles bayesiens hierarchiques, où plusieurs hyperparamètres sont introduits à l'aide des lois à priori. Pour éviter les calculs des densités marginales on utilise l'échantillonneur de Gibbs pour la simulation de termes conditionnellement aux autres variables. Des exemples sont fournis pour la détection des ruptures. Une fois l'expression de la densité à posteriori obtenue, on obtient l'estimateur de Θ en maximisant la fonction $L(X/\Theta)f(\Theta)$, ou en utilisant une approche numérique. On peut calculer l'estimateur de différentes manières par exemple en échantillonnant la variable Θ jusqu'à ce que l'algorithme converge vers le maximum de la distribution. L'inconvénient d'une telle méthode est le temps de calcul suffisamment important, pour plus en détails on peut consulter le livre de [19] et le livre [74] qui étudie l'inférence bayesienne paramétrique.

Un modèle bayesien diffère selon la nature des données de X(normales, exponentielles, continues ou discrètes), et selon le paramètre Θ (la longueur d'un segment, sa moyenne, sa variance ou un état associé aux X_i). Ainsi dans [55] et [31], le paramètre à estimer est le vecteur $\mathbf{R}$ des variables aléatoires indicatrices de la présence d'une rupture, dont les coefficients sont $R_i = 1$, si il y a une rupture ou $R_i = 0$, sinon. pour tout $1 < i < n$, et, par convention $R_1 = R_n = 1$. Détecter les événements dans le signal X revient donc à inférer $\mathbf{R}$. Le modèle présenté dans [55] est Bernoulli-Gaussien : les données X_i suivent une loi normale et sont supposés i.i.d dans un même segment, tandis que les R_i sont i.i.d de Bernouli de paramètre q :

$$f(\mathbf{R}/q) = \prod_{i=1}^{n} q^{R_i}(1-q)^{1-R_i}$$

Plutôt qu' à s'intéresser à estimer le vecteur $\mathbf{R}$ de position, on utilise une autre approche qui consiste à introduire une relation de récursion entre les segments du signal X, grâce à

laquelle on parvient à localiser les segments et à en déduire la position des ruptures. La méthode de [41] introduit ainsi la loi $B(t,s)$ des variables aléatoires $X_t,\ldots,X_s$ $(s \geq t)$, appartenant au même segment, et la probabilité $Q(t)$ que la variable aléatoire soit une rupture :

$$B(t,s) = P(X_t,\ldots,X_s,\ t \textit{ et } s \textit{ sont sur le même segment})$$

$$P(X_t,\ldots,X_n/X_{t-1} \textit{ est une rupture}).$$

Elle repose sur le fait que les paramètres θ_k des segments $1 \leq k \leq K+1$ sont indépendants les uns des autres. Le modèle fait également intervenir la distribution qui modélise la durée de l'intervalle entre deux ruptures, la loi binomiale négative est choisie a priori. Ainsi les positions des ruptures $\tau_1,\ldots,\tau_K$ sont estimées successivement et directement en partant de l'instant i = 1. La stratégie récursive est reprise dans [40] pour une application en ligne, et plus récemment dans [5], où l'algorithme BARD présenté permet de traiter des séries temporelles multivariées. Ces algorithmes sont adaptés selon les distributions des données, par exemple pour la loi normale et pour la loi de Student. D'autres méthodes ont été développées d'un point de vue non paramétrique, y compris dans un cadre bayesien, et permettent ainsi de s'affranchir de la dépendance au modèle. Cette alternative est intéressante en l'absence d'information à priori sur le système étudié, en particulier lorsque la normalité des données n'est pas garantie, ou bien quand le modèle doit être le plus généraliste possible, pour s'adapter à des lois de probabilités variées.

1.2.4 La méthode de type LASSO

Parmi les méthodes construites sur l'hypothèse que les observations suivent la loi normale, l'approche LASSO et ses variantes sont communément rencontrées. Le principe consiste à approcher le signal par une fonction. Dans le cas qui nous intéresse, la série temporelle X est vue comme une fonction constante par morceaux de K + 1 segments de coefficients $\mu_1,\ldots,\mu_K$ contaminée par un bruit ε de moyenne nulle :

$$X_i = \mu_k + \varepsilon_i \quad \tau_{k+1} \leq i \leq \tau_k, \quad 1 \leq k \leq K+1. \tag{1.2.1}$$

Les coefficients de la fonction à estimer sont notés $\beta = (\beta_1,...,\beta_n)$. Le problème de régression s'écrit généralement sous la forme d'une minimisation d'un critère de moindres carrés :

$$\min_{\beta\in\mathbb{R}^n} \sum_{i=1}^{n} (X_i - \beta_i)^2$$

cependant si la solution n'est pas constante par morceaux, il sera difficile de déterminer avec précision les sauts de moyenne significatifs. Afin de renforcer cette caractéristique, une pénalisation de la variation totale est ajoutée [75]. Le problème 1.2.1 devient un problème de régularisation :

$$\min_{\beta\in\mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu_i)^2 + \lambda \sum_{i=1}^{n} |\beta_{i+1} - \beta_i|, \tag{1.2.2}$$

La différence entre les coefficients successifs, notée $\Delta_i = \beta_{i+1} - \beta_i$, est pénalisée par la norme l_1, qui permet de sélectionner les différences les plus significatives en annulant certains termes Δ_i. Cette formulation est plus adaptée que la norme l_2 pour apporter

une contrainte de parcimonie sur les Δ_i, et est préférée à la norme l_0 pour faciliter la résolution. Le paramètre de régularisation λ contrôle la parcimonie des différences Δ_i, c'est-à-dire l'amplitude des sauts. Lorsque λ est nul, l'estimation $\hat{\beta}$ est la solution du problème des moindres carrés 1.2.1, et lorsqu'il est grand, le nombre de segments de $\hat{\beta}$ est faible. Ce problème d'optimisation convexe peut être résolu efficacement par la méthode LASSO (en anglais Least Absolute Shrinkage and Selection Operator), présentée dans [79]. L'expression 1.2.2 correspond au cas particulier du fused LASSO à une dimension [81]. On trouve parfois une pénalisation supplémentaire du nombre de valeurs prises par les coefficients de β, le problème est alors formulé de la façon suivante :

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} (X_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n} |\beta_{i+1} - \beta_i| + \lambda_2 \sum_{i=1}^{n} |\beta_i| \tag{1.2.3}$$

que l'on appelle le sparse fused LASSO.

Le problème 1.2.3 peut être exprimé par un problème dual équivalent. La résolution peut se faire par la méthode de Least Angle Regression (LAR)[33] ou par Alternating Direction Method of Multipliers (ADMM) [15]. Elle fait intervenir l'opérateur de seuillage doux, qui introduit un biais dans les estimateurs de plus grande valeur. Une manière de le corriger est par exemple de pondérer les termes de la norme l_1 par des poids adaptatifs [86]. On notera que ce biais sur l'amplitude des sauts de moyennes n'est pas gênant si l'objectif est simplement de localiser les changements. L'application de l'algorithme LASSO pour la détection de ruptures multiples est discuté dans [50], en particulier la question de l'estimation du nombre de ruptures, qui est contrôlée par le paramètre . Les auteurs remarquent en effet que la méthode a tendance à ajouter des sauts de moyenne à tort, bien que les vrais soient correctement estimés. L'algorithme Cachalot (CAtching CHAnge-points with LassO) est proposé pour effectuer une sélection du nombre K de ruptures a posteriori dans une procédure de programmation dynamique. La consistance de l'estimateur (1.2.2), dit également de moindre carrés et variation totale, est montrée dans [51], pour l'approximation du signal. En revanche on ne parvient à de tels résultats pour l'estimation des ruptures que sous certaines conditions. Ce genre de méthodes avec une pénalisation de la variation totale s'applique par exemple à la détection de ruptures et la segmentation [50], le débruitage de signal ou d'images [80], ou encore pour l'estimation de coefficients dans un processus auto-régressif [2]. Les résultats théoriques associés à l'algorithme LASSO ont été établis pour des erreurs ε_i centrées et distribuées normalement. En présence de bruit à queue lourde, qui introduit des valeurs aberrantes dans les observations, l'approche paramétrique LASSO a tendance à sur-segmenter le signal. En effet, le critère de moindres carrés dans le problème (1.2.2) est sensible aux fortes valeurs de X. Pour que le problème soit robuste à ce genre de phénomène, on peut remplacer ce critère, équivalent à l'application de la norme l_2, par la norme l_1 et ainsi contraindre la solution sur la parcimonie des résidus. Dans l'article [3] les auteurs présentent un ensemble de méthodes reposant sur des fonctions à support quadratique, dont font partie les normes l_1 et l_2, ainsi que leur mise en œuvre dans une série de problèmes d'optimisation. Le LASSO robuste y est introduit. Sa formulation avec la norme l_1 est la suivante :

$$\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^{n} |X_i - \beta_i| + \lambda \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i| \tag{1.2.4}$$

Les auteurs de [15] proposent de résoudre le problème dual par ADMM ou par la méthode du point intérieur, plus performante. Pour approcher le signal par une fonction,

une autre méthode est celle de la régression quantile, qui consiste à contraindre la solution afin que ses quantiles correspondent à ceux des données. Cette méthode est intéressante par exemple lorsqu'on ne peut pas supposer que les données suivent la distribution gaussienne. Dans notre problème de détection de rupture, on cherche à délimiter les portions du signal de moyenne ou de médiane constante. En choisissant pour quantile la médiane, on se ramène à la méthode de type LASSO robuste [34]. La préférence pour la norme l_1 est justifiée dans [72]. La méthode paramétrique LASSO peut être interprétée d'un point de vue bayésien [79]. En effet, en écrivant le problème 1.2.2 avec $AB = \beta$, $A \in \mathbb{R}^{n\times(n-1)}$ et $B \in \mathbb{R}^{n-1}$ les données sont générées selon la loi normale $\mathcal{N}(AB, I_n)$, où I_n est la matrice identité de dimension $n \times n$. On choisit de modéliser les coefficients b_i du vecteur B par la loi de Laplace de paramètre σ^2, afin que les différences $\beta_{i+1} - \beta_i$ soient fortement concentrées autour de 0. L'estimateur de 1.2.2 correspond alors à l'expression d'un mode de la densité de probabilité a posteriori. [67] développent ainsi un modèle bayésien à partir de l'algorithme LASSO, où λ est un hyperparamètre. Ce type d'approche constitue un ensemble de méthodes efficaces pour l'approximation d'un signal, pouvant notamment être employées pour la détection de ruptures. Cependant, comme toute méthode paramétrique, elles sont limitées par la dépendance au modèle des données, et le paramètre de régularisation λ doit être adapté à chaque application.

1.2.5 La méthode des noyaux

Le principe de la méthode du noyau est basé sur une transformation ϕ appliquée aux données de l'espace d'entrée E vers une espace F de dimension plus grande. Pour détecter les ruptures, on calcule une mesure de similarité entre les images d'observations. Comme la méthode de dérivée filtrée, cette méthode s'applique dans le cas où les données sont de grandes dimensions.
Dans le cas où la rupture est à la position τ, on fait un test d'homogénéité entre les deux segments dont les lois de probabilités sont F_1 et F_2. Quand la position est indéterminée, [50] propose un test sur une fenêtre glissante. Par conséquent pour déterminer la vraie rupture, on maximise la mesure d'homogénéité.
L'espace image F de la transformation ϕ où sont testées les hypothèses est appelée espace de Hilbert à noyau. On note $< .., .. >_F$ son produit scalaire.
On considère le noyau $h(X, y) =< \phi(X), \phi(y) >_F$: pour comparer X et y on traite les images des observations, où elles sont linéairement séparables. Cette opération est appelée l'astuce du noyau. La condition à respecter est que la matrice de Gram H, dont les coefficients sont les $h(X_i, y_j)$, est semi-définie positive. Cette condition est vérifiée par exemple dans les cas de noyaux linéaire et le noyau gaussien.
Deux paramètres sont caractérisés par les lois de probabilités dans F : la moyenne μ et la covariance $\sum$, définis par :

$$< \mu, f >= \mathbb{E}(f(X)), \quad \forall f \in F$$

$$< f, \sum g >_F = cov(f(X), g(X)), \quad \forall f, g \in F$$

pour une variable X de E.
Dans [46] se trouve les méthodes appelées divergence maximale moyenne. Cette méthode renvoie une mesure de similarité entre les deux moyennes de deux populations de E.
On note :

$$T_{n_1,n_2} = (n_1 + n_2)\|\hat{\mu}_1 - \hat{\mu}_2\|_F^2$$

où n_1, n_2 sont les effectifs de deux populations de E et $\hat{\mu}_1$ et $\hat{\mu}_2$ les mesures empiriques de μ_1 et μ_2.
Dans [65] se trouve une variante de la méthode de divergence maximale moyenne. Cette variante ajoute un terme de covariance et d'une normalisation. Nous avons :

$$T_{n,\tau,\delta} = \frac{n_1 n_2}{n} \|(\hat{\sum} + \gamma_n I)^{\frac{1}{2}} (\hat{\mu}_1 - \hat{\mu}_2)\|_F^2$$

où $n = n_1 + n_2$ et $\hat{\sum} = \frac{n_1}{n}\hat{\sum}_1 + \frac{n_2}{n}\hat{\sum}_2$
avec $\hat{\sum}_1$ et $\hat{\sum}_2$ sont les covariances empiriques de deux populations.

Pour détecter une rupture à une position inconnue l'algorithme de [50] parcourt le signal avec une fenêtre glissante de taille n, et la formule précédente est calculée à chaque mouvement. On désigne la position potentielle celle qui maximise la statistique. Dans [50], la distribution asymptotique sous l'hypothèse nulle et la consistence du test sous l'hypothèse alternative sont établis.
Dans [54], le test du noyau du rapport des densités est défini de cette manière : un estimateur de ce ratio $r(X, \theta)$, selon le paramètre θ est $\hat{\nu}_n = \frac{1}{n}\sum_1^n \theta_i \phi(X_i)$, avec $\theta_1, \theta_2, \ldots, \theta_n \geq 0$. La statistique de test est constante à partir de $\hat{\nu}_n$, $T_n = \frac{1}{n}\sum_1^n log(< \nu_n, \phi(X_i)) >_F$.

Une autre méthode à noyau est celle des machines à vecteurs de support. Elle consiste à calculer l'hyperplan qui définit la région de l'espace associée aux échantillons de X_1 et celle associée aux échantillons de X_2, afin d'obtenir une mesure de distance entre ces deux régions. Dans [30], la méthode de détection de changement par noyau est traité avec un noyau gaussien et pour une application à la détection séquentielle. On a deux vecteurs X_1 avec n_1 observations et X_2 avec n_2 observations, X_τ inclus. L'espace F est normalisé, de telle sorte que $\phi(E)$ soit un sous ensemble de l'hypersphère unitaire S centrée sur l'origine F. L'image de X_1 dans F est le vecteur d'apprentissage : il permet de construire l'hyperplan ω_1 par la résolution d'un problème d'optimisation. Cet hyperplan paramétrisé par (ω_1, p_1) sépare les échantillons du centre S avec la marge p_1 sans tenir compte des éventuelles valeurs aberrantes. La méthode est en effet insensible à ce genre de perturbations, en fixant un seuil ν. De la même façon on obtient l'hyperplan w_2, paramétrisé par (w_2, p_2) qui sépare les images des observations X_2 du centre de l'hypersphère. Une différence entre les images doit se traduire par une répartition des images d'observations dans des régions distinctes. La mesure de divergence suivante tient de l'écart entre les hyperplans ainsi que des dispersions des distributions,

$$T_{n_1,n_2} = \frac{\widehat{c_1 c_2}}{\widehat{c_1 p_1} + \widehat{c_2 p_2}}.$$

où c_i est le point d' intersection de S avec le vecteur prolongé w_i, et où p_i est le point d'intersection de w_i avec S dans le plan contenant les vecteurs w_1 et w_2, et inclus dans l'arc $\widehat{c_1 c_2}$. Les longueurs sont calculées à partir de produits scalaires et de normes de l'espace F sur les paramètres des hyperplans, ainsi que sur la matrice du noyau et des multiplicateurs de Lagrange issus de l'étape d'estimation des hyperplans. Cette statistique est calculée à chaque translation d'une fenêtre glissante, afin d'estimer l'instant de rupture. Une telle méthode n'a pas de résultat asymptotique. On fixe le seuil pour accepter ou rejeter l'hypothèse nulle. Les méthodes du noyau dépendent fortement du choix du noyau h.

1.2.6 La méthode mesure de distances

[64] présente une méthode basée sur la distance euclidienne. Cette méthode s'appelle la E-divise. Les observations sont i.i.d dans un même segment et avec seule condition sur les

distributions que le moment d'ordre a $\in [0,2]$ existe. On partage le signal en deux portions $S_k = (X_{\tau_{k-1}+1}, ...X_{\tau_k})$ et $S_{k+1} = (X_{\tau_k+1}, ...X_{\tau_{k+1}})$ de longueurs respectives m et n. Pour détecter l'existence d'une rupture au point τ_k, on procède de la façon suivante :

$$\xi(S_k, S_{k+1}, a) = \frac{2}{mn} \sum_{i=\tau_{k-1}}^{\tau_k} \sum_{j=\tau_{k-1}}^{\tau_k} |X_i - X_j|^a - \frac{1}{C_n^2} \sum_{\tau_{k-1}+1 \le i \le j \le \tau_k} |X_i - X_j|^a - \frac{1}{C_m^2} \sum_{\tau_k+1 \le i \le j \le \tau_{k+1}} |X_i - X_j|^a.$$

La statistique du test est $p(S_k, S_{k+1}) = \frac{mn}{m+n}\xi(S_k, S_{k+1}, a)$, et sa conséquence en distribution sous l' hypothèse nulle et sous l'hypothèse alternative est connue. On estime l'instant de rupture τ_k en maximisant la statistique locale.
Pour la détection de plusieurs ruptures on applique la méthode de segmentation binaire. On utilise un test comme critère d'arrêt. Ensuite on utilise les instants de ruptures à partir de ce test. On remarque qu'il y a des vraies ruptures mais aussi de fausses alarmes. La théorie de graphe est utilisée dans [25] pour établir la statistique du test. Le graphe est construit à partir de similarités entre les observations. la méthode traite la détection d'une seule rupture et pour des ruptures multiples, on utilise la bissection. Les noeuds du graphe G sont les observations et on se base sur l'idée que les variables générées selon la même distribution sont proches l'une de l'autre. Le graphe G permet de séparer les groupes d'observations de distributions différentes. La statistique $R_G(c)$ mesure alors le nombre d'arêtes connectant une observation i à une observation après i+1. Sous l'hypothèse nulle, $R_G(i)$ est petit. Le maximum de la statistique Z_G version standardisée de R_G donne la position du changement.

1.2.7 La méthode basée sur un test de rang

Le livre [59] est la référence des méthodes basées sur les tests. L'avantage des tests non paramétriques de rang est le fait qu'on a peu d'hypothèses sur les données. Le test de la somme des rangs de Wilcoxon est le plus connu de ce test. Il établit la comparaison des valeurs de deux populations données. Ce test d'homogénéité est sensible aux différences entre les rangs moyens des deux populations, ce qui revient à tester les médianes dans certaines cas.
Pour déterminer l'instant d'une seule rupture, on considère la statistique de test de Wilcoxon-Marn-Whitney

$$U_\tau = \sum_{i=1}^{\tau} \sum_{j=\tau+1}^{n} \mathbf{1}_{\{x_i \le x_j\}}.$$

L'hypothèse nulle H_0, étant que les observations du premier et du deuxième segment, délimitées par τ, suivent des distributions de même médiane. H_0 est rejetée pour des grandes valeurs de U_τ.
D'autre part, [70] introduit la statistique $T_\tau = \max_{1 \le n} U_\tau$. Ce test est applicable sur des données de distributions discrète et donne une version approchée pour les distributions continues. [62] propose aussi un autre test pour la détection d'un seul changement dans le cas des données multivariées.
Pour déterminer les instants de ruptures multiples, on considère la statistique de test suivante :

$$T(\tau_1, \ldots, \tau_K) = \frac{12}{n^2} \sum_{k=0}^{K} (\tau_{k+1} - \tau_k)(\widehat{R_h} - \frac{n}{2})^2 \tag{1.2.5}$$

où

$$\widehat{R_h} = (\tau_{k+1} - \tau_k)^{-1} \sum_{i=\tau_k+1}^{\tau_{k+1}} \sum_{j=1}^{n} \mathbf{1}_{\{x_j \le X_i\}},$$

avec la convention $\tau_0 = 1$ et $\tau_{K+1} = n$.
Le nombre de segment maximale K_{max} est déterminé à postériori, à l'aide d'une heuristique de pente sur la valeur de la statistique en fonction du nombre de ruptures.
L'autre avantage de ces méthodes traitées dans [45] et [62] est de pouvoir traiter des données censurées ou manquantes, en encadrant les observations par des valeurs limites lors du calcul des rangs.
La statistique 1.2.5 se calcule récursivement par conséquent [63] propose un algorithme de programmation dynamique pour un nombre d'événements K donné. Comme dans les autres méthodes de détection de ruptures, la complexité combinatoire reste un problème lorsque les données sont multivariées.

1.3 Revue de littérature du modèle autorégressif à seuil

Depuis les travaux de [85], les modèles autorégressifs linéaires et non-linéaires deviennent une branche importante de la statistique. Le modèle le plus populaire est le modèle autorégressif et moyenne mobile, en anglais Autoregressive Moving-Average Model(ARMA). C'est le cas le plus utilisé dans les modèles paramétriques des séries temporelles. Le modèle ARMA est souvent utilisé dans le système linéaire dynamique. Ceci est dût en raison de sa faisabilité pour l'approximation de plusieurs processus stationnaires. Depuis la naissance de séries temporelles jusqu'aux travaux de [14] qui ont marqué la maturité du modèle ARMA dans la théorie et dans la méthodologie, le modèle linéaire de séries temporelles a permis de développer et de construire les séries temporelles et par la suite, a eu beaucoup d'applications dans différentes domaines. Les quarante dernières années sont témoins de la popularité continue, voir [14], [17], [39] et beaucoup d'autres.
Cependant, aucun modèle statistique est une approximation exacte sur des données réelles. Les approximations linéaires sont la première étape pour approximer des données réelles. Malheureusement, le modèle ARMA n'approxime pas bien les phénomènes non-linéaires, par exemple les cycles asymétriques, la distorsion harmonique, la résonance du saut, la normalité et bien d'autres. Cette nécessité de prendre en compte la non-linéarité et plus particulièrement les changements de régimes tend à modifier profondément les approches des applications de séries temporelles. De nombreuses pistes ont été explorées pour modéliser la non-linéarité. La voie qui s'est cependant révélée la plus fructueuse est celle des modèles à changements de régimes qui ont l'avantage d'approximer les applications des exemples cités ci-dessus. Parmi les classes des modèles non-linéaires, il existe deux modèles populaires : le modèle GARCH, en anglais Conditionnal Héteroscedasticity introduit par [38] et le modèle autorégressif à seuil, en anglais Threshold Autoregressive (TAR) initié par [82]. Plus tard, le modèle TAR est devenu le modèle standard dans les séries temporelles non-linéaires. Actuellement, le modèle TAR est le modèle utilisé pour étudier les phénomènes non-linéaires dans différents domaines d'applications comme par exemple en économie, en science de l'environnement, en finance, en hydrologie, en physique et bien d'autres. Le mécanisme de transition du modèle TAR s'effectue à l'aide d'une variable aléatoire de transition observable, d'un seuil et d'une fonction de transition. La difficulté de ce type de modèle repose donc sur la définition de cette variable observable, il existe cependant des méthodes statistiques, telles que les tests de linéarité pour nous guider dans ce choix. Le modèle particulier qu'on considère dans cette thèse et qui est aussi considéré par [82], [32] et bien d'autres, est celui où on compare la variable de transition à un seuil : cette dernière est supérieure ou inférieure, alors la transition se réalise instantanément.

Jusqu'à présent, l'ensemble des régimes qui se trouvent dans la littérature ont permis

la modélisation des asymétries telles que les dynamiques distinctes dans les phases ascendantes et descendantes à l'aide de leurs différents régimes. Ils permettent également de s'interroger sur la stabilité temporelle des coefficients dans le temps. Cependant contrairement au modèle de rupture, le passage d'un régime à un autre n'est ni daté ni définitif étant déterminé de manière endogène en fonction d'un seuil. Un autre avantage du modèle TAR est qu'il tend à enrichir le débat relatif au traitement de la non-stationnarité. L'existence de plusieurs régimes dans un même modèle autorise un processus à être globalement stationnaire. En d'autres termes, tous les régimes ne sont pas obligatoirement caractérisés par la présence de racine unitaire dans leur polynôme autorégressif et réciproquement tous ne sont pas contraints à être stationnaires. En séries temporelles, cette question de la non-stationnarité versus la non-linéarité est également relativement importante, sachant que ces deux notions peuvent être confondues à l'issue d'un test de stationnarité classique.

1.3.1 Modèle TAR

Définition : Le processus $(X_t,\ t \in \mathbb{Z})$ sastisfait une représentation TAR à deux régimes d'ordre p_1 et p_2, si et seulement si :

$$X_t = \begin{cases} \alpha_{11} \times X_{t-1} + \ldots + \alpha_{1p} X_{t-p_1} + \varepsilon_t, \text{ pour } q_t \leq r_0 \\ \alpha_{21} \times X_{t-1} + \ldots + \alpha_{2p} X_{t-p_2} + \varepsilon_t, \text{ pour } q_t > r_0 \end{cases} \tag{1.3.1}$$

où de façon équivalente :

$$X_t = (\alpha_{11} \times X_{t-1} + \ldots + \alpha_{1p} X_{t-p_1})\mathbf{1}_{\{q_t \leq r_0\}} + (\alpha_{21} \times X_{t-1} + \ldots + \alpha_{2p} X_{t-p_2})\mathbf{1}_{\{q_t > r_0\}} + \varepsilon_t \tag{1.3.2}$$

avec (ε_t) un bruit blanc, r_0 la valeur du seuil, q_t la variable de transition et $\mathbf{1}(A)$ une variable aléatoire indicatrice qui prend la valeur 1 lorsque la contrainte A entre parenthèse est vérifiée et 0 sinon.

Le mécanisme de transition est gouverné par la comparaison d'une variable de transition observable q_t qui doit être préalablement définie et d'un seuil estimé r_0. Lorsque la valeur de la variable de transition est inférieure au seuil, la dynamique de la variable X_t est donc caractérisée par le processus autorégressif de paramètres $\alpha_{1,i}$ $(1 \leq i \leq p_1)$ et de manière équivalente par le processus autorégressif de paramètres $\alpha_{2,j}$ $1 \leq j \leq p_2)$ lorsque la valeur de q_t est supérieure au seuil. Le mécanisme de transition est brutal sachant que le passage d'un régime à l'autre se fait en une période. Il est également à nouveau possible de changer de régimes, dès lors que la valeur de la variable de transition devient supérieure ou inférieure à la valeur du seuil.

La difficulté majeure de cette modélisation porte donc sur le choix de transition. Le changement de régimes dépend de la variable observable. Le choix d'une mauvaise variable peut donc avoir de fortes implications. Habituellement, la variable de transition est soit une variable exogène, soit une variable endogène retardée, soit une fonction linéaire ou non des variables endogènes retardées. Ensuite pour sélectionner parmi un ensemble de variables potentielles la variable de transition la plus appropriée, il est possible de se référer à un critère statistique tel que la minimisation de la somme des carrés des résidus, ou bien encore au rejet du test de linéarité.

Lorsque la variable de transition sélectionnée est une variable endogène retardée X_{t-d} (où d est un entier positif), le modèle TAR devient un modèle SETAR, spécification qui a été développée par [47].

La seconde difficulté est de déterminer l'ordre p_1 et p_2 suivi par les dynamiques autorégressives de chaque régime. Par simplification dans un modèle TAR à multiples régimes, il est généralement supposé que l'ordre des processus autorégressifs de chaque régime est identique et déterminé à partir de l'estimation d'un modèle linéaire. Cette hypothèse ne repose sur aucune justification théorique, mais est souvent retenue d'un point de vue pratique. En revanche, dans un modèle contenant deux régimes, il est possible d'utiliser les critères d'information modifiées par [82] afin d'autoriser un ordre différent suivant les dynamiques autorégressives.

Estimation : Lorsque l'ordre des processus autorégressifs a été identifié, l'étape suivante consiste à estimer les coefficients des variables explicatives mais également le ou les paramètres à seuils. Les méthodes d'estimations usuelles du type MCO ne sont pas alors applicables dans cette situation, la définition des variables explicatives dépendant des seuils. La méthode du maximum de vraisemblance n'est pas non plus applicable étant donnée que la fonction de vraisemblance n'est pas dérivable en fonction de ces paramètres. La solution alors envisageable est d'utiliser les moindres carrés récursifs, voir [24] et [32] et la méthode de maximum conditionnelle, voir [73].

En effet, lorsque la valeur des seuils est fixée, il est possible d'estimer les coefficients des variables explicatives par les MCO. Il ne reste plus qu'à définir les valeurs possibles pour les variables de seuils et déterminer les seuils optimaux minimisant la somme des résidus du modèle. Les valeurs des seuils sont recherchées parmi les valeurs de la transition ; cependant un nombre de points minimum doit être conservé dans chaque régime. De même pour la méthode de maximum de vraisemblance conditionnelle, on fixe les seuils et on utilise la méthode de maximum de vraisemblance classique pour estimer les autres paramètres du modèle et pour trouver les estimateurs des seuils, on maximise à nouveau la méthode de vraisemblance. Notons que [73] considère la densité des erreurs.

Dans un autre point, le modèle ARMA faible est développé par [42], donnons une revue du modèle ARMA faible.

1.3.2 Modèle ARMA faible

Le modèle ARMA fort est constitué quand les erreurs sont i.i.d. et dans la littérature, on définit un ARMA semi-fort, lorsque les hypothèses sur les innovations sont une différence de martingale et un ARMA faible quand les erreurs sont uniquement décorrélés.

Répresentation du modèle ARMA faible

Soit $(X_t)_{t\in\mathbb{Z}}$ un processus stationnaire de seconde ordre tel que

$$X_t + \sum_{i=1}^{p} a_i X_{t-i} = \varepsilon_t + \sum_{i=1}^{q} b_i \varepsilon_{t-i} \tag{1.3.3}$$

où (ε_t) est une suite de variables aléatoires non corrélées sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ avec de moyenne nulle et de variance $\sigma^2 > 0$. Le polynome $\phi(z) = 1 + a_1 z + a_2 z^2 + \ldots + a_p z^p$ et $\psi(z) = 1 + b_1 z + \ldots + b_q z^p$ ont leurs racines en dehors du disque de l'unité et n'ont pas de racines communes.

Sans perte de généralité, assumons que a_p et b_q sont différents de zéro(par convention $a_0 = b_0 = 1$). Le processus peut-être interprété comme les innovations linéaires de (X_t) : $\varepsilon_t = X_t - \mathbb{E}(X_t/\mathcal{F}_{t-1})$ où $\mathcal{F}_{t-1}$ est l'espace de Hilbert engendré par $(X_s : s < t)$. De plus

assumons que (X_t) est une suite de processus strictement stationnaire.
Le paramètre $\theta_0 = (a_1, \ldots, a_p, b_1, \ldots, b_q)^{'}$ appartient à l'espace de paramètres Θ défini par :

$$\theta = \{\theta = (\theta_1, \ldots, \theta_p, \theta_{p+1}, \ldots, \theta_{p+q})^{'}; \phi(z) = 1+\theta_1 z+\ldots+\theta_p z^p \text{ et } \psi(z) = 1+\theta_{p+1} z+\ldots+\theta_{p+q} z^p\}$$

où les polynomes ϕ et ψ ont leurs racines en dehors de l'unité.
Pour tout $\theta \in \Theta$, soit $(\varepsilon_t(\theta))$ un processus stationnaire de second ordre(l'existence et l'unicité d'un tel processus est démontré dans le chapitre 3 de [16]) définit comme la solution de

$$\varepsilon_t(\theta) = X_t + \sum_{i=1}^{p} \theta_i X_{t-i} - \sum_{i=1}^{q} \theta_{p+1} \varepsilon_{t-i}(\theta). \tag{1.3.4}$$

Notons que $\varepsilon_t(\theta_0) = \varepsilon_t$ p.s. pour tout $t \in \mathbb{Z}$. L'assuption sur les moyennes mobiles du polynome ψ_θ implique qu'il existe une suite de constantes $(c_i(\theta))$ tel que $\sum_{i=1}^{\infty} |c_i(\theta)| < \infty$ et $\varepsilon_t(\theta) = X_t + \sum_{i=1}^{\infty} c_i(\theta) X_{t-i}$ pour tout $z \in \mathbb{Z}$. Notons enfin que pour tout $\theta \in \Theta$, $\varepsilon_t(\theta)$ est de carré intégrable et la fonction $\varepsilon_t(.)$ est continue.

Estimation des paramètres

On considère les observations $X_1, \ldots, X_n$ de longueur n et pour tout $0 \leq t \leq n$, les variables $\varepsilon_t(\theta)$ sont définis récursivement comme dans (1.3.4). Les valeurs initiales inconnues sont remplacées par zéro : $\varepsilon_0(\theta) = \ldots = \varepsilon_{1-q}(\theta) = X_0, \ldots = X_{1-p} = 0$.
Soit δ une constante strictement positive choisie tel que le vrai parmètre θ_0 appartient au compact Θ_δ où

$$\Theta_\delta := \{\theta \in \mathbb{R}^{p+q}, \textit{ les racines des polynomes } \phi_\theta(z) \textit{ et } \psi_\theta(z) \textit{ sont de module } \geq 1+\delta\}$$

Un estimateur de la vraie valeur θ_0 par la méthode de moindres carrés ordinaires(MCO) est toute solution p.s, mesurable $\hat{\theta}_n$ de

$$L_n(\hat{\theta}_n) = \min_{\theta \in \Theta} L_n(\theta) \tag{1.3.5}$$

où

$$L_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t^2(\theta)$$

Propriétés asymptotiques de l'estimateur du MCO

Les deux thèorèmes importants contenus dans [42] sont les suivants :

Theorem 1.3.1.
Soit $(X_t)_{t\in\mathbb{Z}}$ un processus strictement stationnaire, ergodique et de carré intégrable, et satisfaisant (1.3.3). Soit $\hat{\theta}_n$, une suite de MCO defini dans (1.3.5). Supposons que $\theta_0 \in \Theta_\delta$, alors

$$\hat{\theta}_n \to \theta_0 \textit{ p.s., quand } n \to \infty.$$

Soit $\mathcal{F}_{-\infty}^{t}$ et $\mathcal{F}_{t-k}^{\infty}$ les tribus engendrées par $\{X_u : u \leq t\}$ et $\{X_u : u \geq t+k\}$ respectivement. Les coefficients de mélanges $(\alpha_X(k))_{k\in\mathbb{Z}^*}$ sont définis par

$$\alpha_X(k) = \sup_{\{A\in\mathcal{F}_{-\infty}^{t},\ B\in\mathcal{F}_{t-k}^{\infty}\}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$$

Theorem 1.3.2.
Supposons que les hypothèses du théorème (1.3.1) sont vérifiées, de plus que $(X_t)_{t\in\mathbb{Z}}$ *satisfait* $\mathbb{E}|X_t|^{4+2\nu} < \infty$ *et* $\sum_{k\geq 0}\{\alpha_X(k)\}^{\frac{\nu}{2+\nu}} < \infty$ *pour un certain* ν *strictement positif, on a*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}(0, J^{-1}IJ^{-1})$$

où $J = J(\theta_0)$ *et* $I = I(\theta_0)$, *avec*

$$J(\theta) = \lim_{n\to\infty} \frac{\partial L_n^2(\theta)}{\partial\theta\partial\theta'} \ p.s.,$$

$$I(\theta) = \lim_{n\to\infty} V(\sqrt{n}\frac{\partial}{\partial\theta}L_n(\theta))$$

Jusqu'à présent, la littérature sur le modèle à changement de regimes traite le modèle TAR fort et [47] évoque le cas TAR semi-fort (l'erreur est une différence de martingale). Nous estimons donc que la connaissance du modèle TAR n'est pas complète et dans notre deuxième partie de thèse, nous traitons le cas TAR faible. Nous avons obtenu les mêmes résultats que ceux des théorèmes (1.3.1) et (1.3.2). Le point le plus remarquable étant l'apparition de la matrice de variance asymptotique sous la forme "sandwich". Nous avons aussi obtenu un théorème de convergence en loi de l'estimateur du seuil.

1.4 Objectifs et organisation des chapitres

1.4.1 Objectifs

Les objectifs de cette thèse se résument en plusieurs points :

1. Nous ajoutons une deuxième étape à la méthode de FD pour détecter l'ensemble des vraies ruptures et avoir le moins de fausses alarmes. Cette deuxième étape est la procédure de Benjamini et Hochberg qui consiste à faire un test multiple. Nous appliquons sur des suites de variables aléatoires indépendantes dont le paramètre d'intérêt, la moyenne change sur chaque segment mais aussi sur des variables aléatoires faiblement dépendantes. On applique aussi sur les paramètres pente et l'ordonnée à l'origine de la droite de régression linéaire. La méthode FDqV est validée sur des données simulés mais aussi sur des données réels des battements du coeur et de la vitesse du vent des éoliennes.

2. Nous comparons notre nouvelle méthode FDqV à la méthode FDpV établie par Bertrand, Fihima et Guillin[11]. Nous prouvons que notre méthode est plus performante en utilisant le critère de moyenne quadratique intégrée.

3. Dans les données des battements du coeur, nous avons avons remarqué que ces observations ne sont pas indépendantes mais plutôt faiblement dépendantes c'est pourquoi nous avons donné une extension de la méthode FDqV au modèle autorégressif car le meilleur modèle pour modéliser ces battements du coeur est le modèle autorégressif.

4. L'algorithme de la méthode FD fait intervenir deux paramètres : le seuil et la fenêtre du calcul des moyennes. Nous avons donné deux paramètres optimaux pour que les nombres de fausses alarmes soient proches de zéro et on détecte tous les vrais instants de ruptures, autrement dit pour rendre meilleur la méthode FD. Nous comparons la méthode FD avec les paramètres optimisées avec la méthode de PLS de Lavielle[56],

en utilisant comme le critère de la moyenne quadratique intégrée et sur les nombres de non détection des vraies ruptures et sur les nombres de fausses alarmes. En plus la méthode FD optimisée est clairement plus avantageuse que la méthode PLS traitée dans [56] sur la calcul computationnel et de la complexité.

5. La méthode FDpV proposée dans le papier de Betrand, Fihima et Guillin[11] pour détecter dans la droite de régression linéaire le paramètre pente est clairement faux. Ils considèrent que la fonction dérivée filtrée est une fonction "chapeau" comme dans le cas du paramètre moyenne. Nous avons corrigé et nous avons démontré que la fonction dérivée filtrée est une dérivée gaussienne lorsque les erreurs sont des variables aléatoires gaussiennes.

6. Dans la deuxième partie de ce projet, nous traitons le cas du modèle de changement de régimes à seuil avec les hypothèses faibles. Dans l'esprit de [42], nous exprimons les innovations sous la forme récursive des variables aléatoires. Nous avons obtenu des théorèmes similaires à ceux obtenus dans [24], [32] et [73] mais avec des hypothèses faibles sur les erreurs. Notre recherche ouvre une grande perspective pour la connaissance du modèle TAR faible et permettra de faire beaucoup d'applications comme par exemple en économie dynamique où les erreurs ne sont pas forcément indépendantes.

1.4.2 Organisation des chapitres

Le chapitre 2 donne une nouvelle méthode qu'on a nommée dérivée filtrée et taux de fausses découvertes, en anglais Filtered Derivative and False Discovery Rate(FDqV). Cette nouvelle méthode consiste à ajouter une deuxième étape à la dérivée filtrée de [8] où à remplacer la deuxième étape de la méthode de dérivée filtrée avec p-value(FDpV) de [11]. Nous comparons notre méthode avec la méthode de dérivée filtrée et la dérivée filtrée avec p-value en se basant sur le critère de l'erreur moyenne quadratique et celui du nombre de détection de vraies ruptures et du nombre de fausses alarmes. Nous donnons ci dessus les différentes figures obtenus dans le ce chapitre.

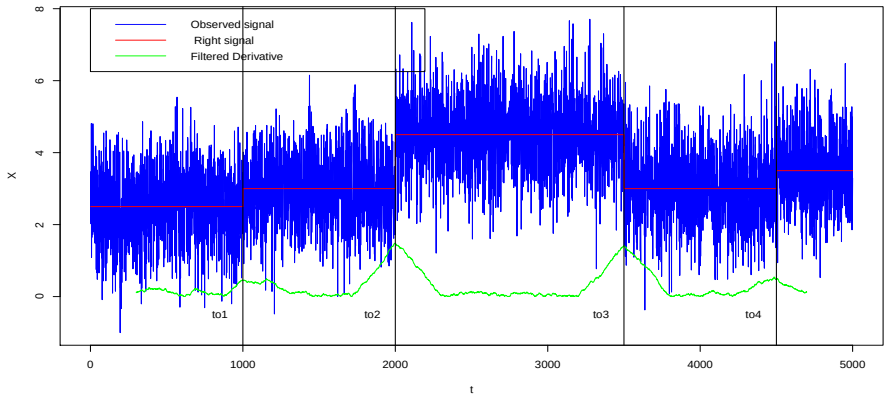


FIGURE 1.4 – *The right signal (red), the noisy signal (blue), and Filtered Derivative function (green).*

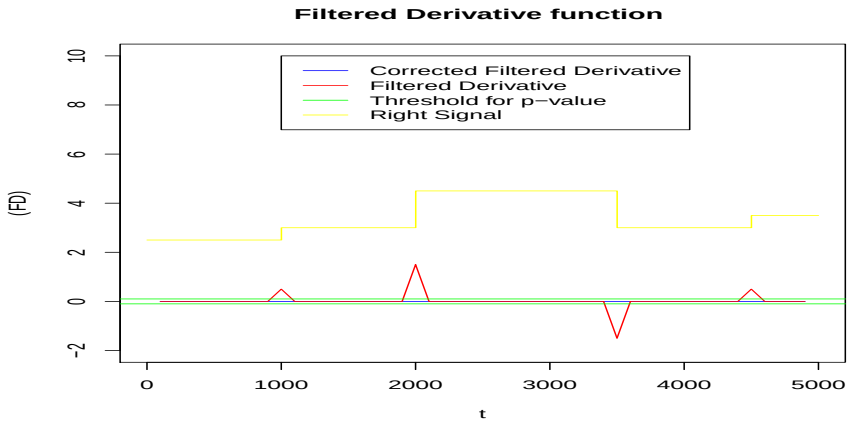


FIGURE 1.5 – *Filtered Derivative function without noise* ($\sigma = 0$).

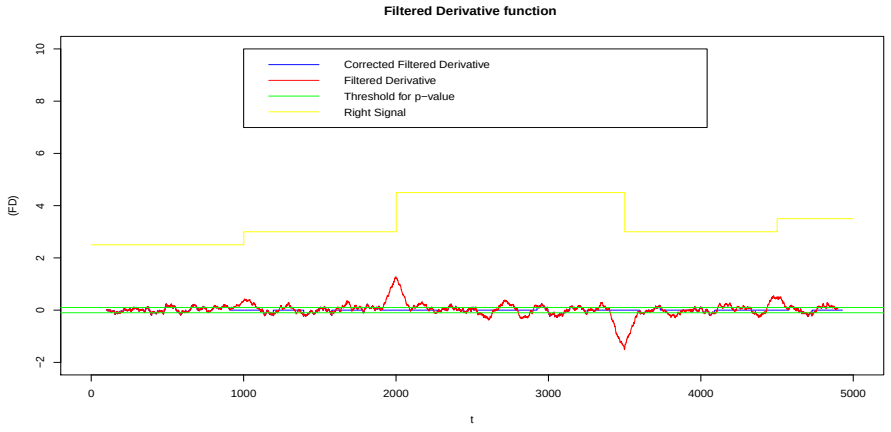


FIGURE 1.6 – *Filtered Derivative function with noise* ($\sigma = 1$).

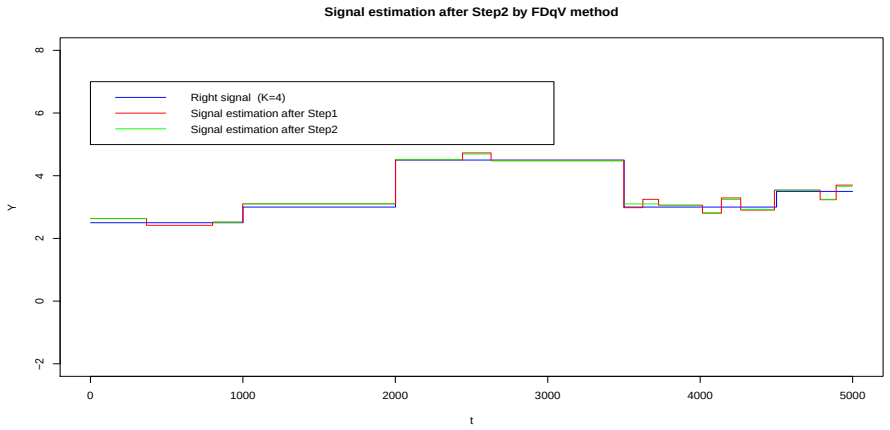


FIGURE 1.7 – *Signal reconstruction after Step2 by FDqV method*

Commentaires sur les figures

- Dans la figure (1.4), on a réprésenté le signal en bleu, le vrai signal en rouge et la dérivée filtrée en vert.
- Dans la figure (1.5), la droite horizontale en vert est le seuil C, la dérivée filtrée sans bruit est réprésenté en rouge et le vrai signal en jaune.
- Dans la figure (1.6), le seuil C est vert, la dérivée filtrée avec bruit en rouge et le vrai signal en jaune.
- Dans la figure (1.7), le vrai signal est en bleu, la réconstruction du signal à l'étape 1 est en rouge et la réconstruction du signal à l'étape 2 est en vert.

Le chapitre 3 traite l'optimisation des paramètres dont dépend la méthode de dérivée filtrée. En effet la fonction de dérivée filtrée dépend du seuil de détection et de la fenêtre du calcul des moyennes. Nous avons donné des paramètres optimaux du seuil et de la fenêtre. Par conséquent la dérivée filtrée avec les paramètres optimaux détecte mieux les vrais instants de ruptures et ont des fausses alarmes relativement proche de zéro. Nous comparons cette méthode avec la méthode de moindres carrés pénalisés adaptés de [57]. Une nouveauté dans la littérature et donc dans ce chapitre 3 est de contrôler le nombre de fausses alarmes et non la probabilité de fausses alarmes. Nous donnons une borne du nombre de fausses et une condition nécessaire à la détection des vrais ruptures. Un problème ouvert est l'optimisation de p-value dans l'étape 2 de FDpV [11] et le taux de fausses détection dans FDqV [35].

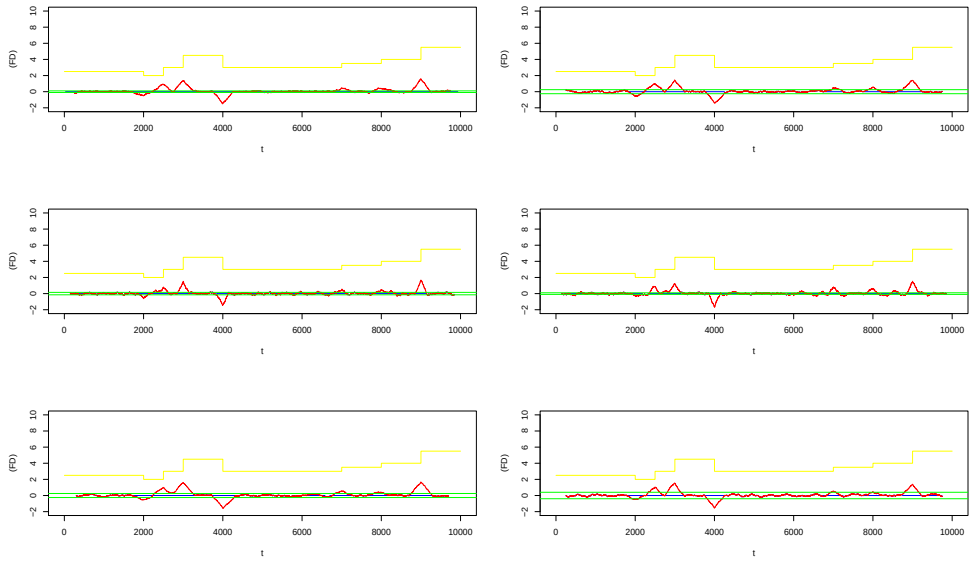


FIGURE 1.8 – *The Filtered Derivative with different parameters A=100 ; 150, 200 ; 250 and* $C_1 = 0.1; 0.15; 0.2, 0.25$.

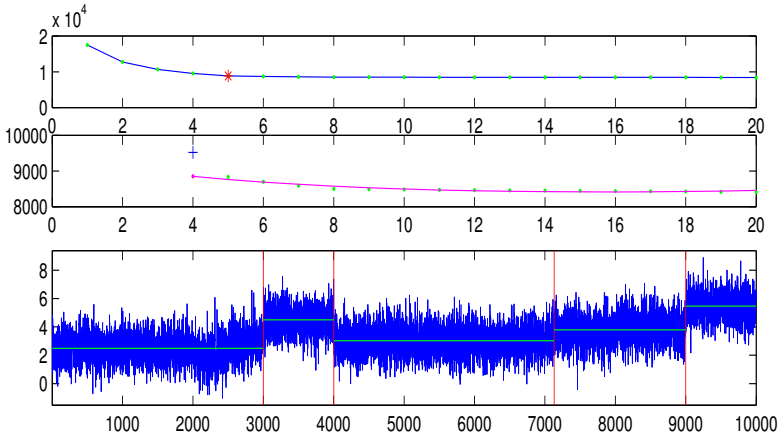


FIGURE 1.9 – *The adaptive method.*

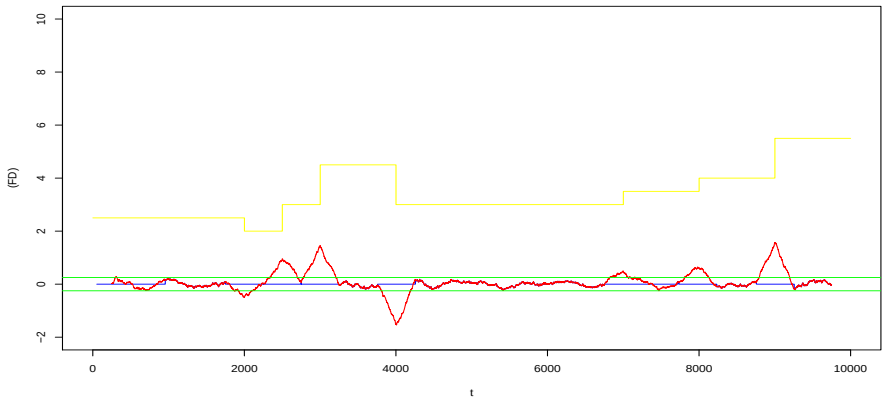
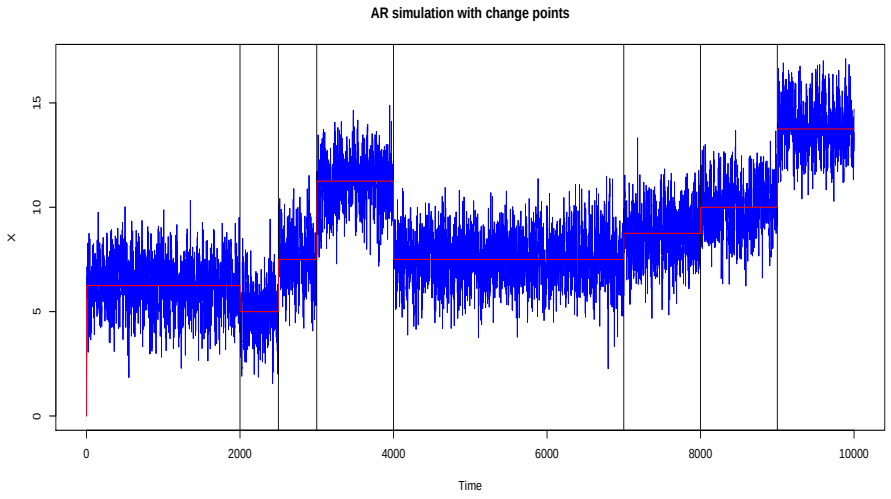


FIGURE 1.10 – *The filtered derivative with parameters optimized (A=250 et C=0.25).*

Le chapitre 4 évoque le cas où les variables aléatoires sont faiblement dépendantes. Nous avons traité dans les chapitres 2 et 3 le cas où les variables aléatoires sont indépendantes. Le modèle le plus simple est de considérer que les variables aléatoires suivent un modéle autorégressif, en anglais Autoregressive Model (AR). Nous détectons les instants de ruptures dans le modèle AR par la méthode FDqV. Nous comparons nos résultats à ceux obtenus dans [21]. Les deux méthodes détectent plus au moins correctement les vrais instants de ruptures mais dans le critère de complexité et du calcul computationnel, la méthode FDqV donne des résultats meilleurs [37] que ceux de [21].



Le chapitre 5 est écrit sous la forme d'une prépublication en collaboration avec Bruno Saussereau. Il contient un résultat important sur le TAR faible.

En effet rappelons que [42] ont développé dans la littérature le modèle ARMA faible, c'est à dire les erreurs sont supposés non corrélées et que le processus est supposé α-mélangeant, ils ont établi un théorème de convergence presque sûr des paramètres du modèle et un théorème de convergence en loi des paramètres, voir plus de détails dans [42]. En combinant les idées de [42] et [24], nous sommes les premiers à étudier le modèle TAR

faible : nous supposons que les erreurs sont non corrélées et le processus est α- mélangeant. Nous établissons la convergence presque sûre, la convergence uniforme et la convergence en loi des paramètres du modèle. D'autre part, le seuil du modèle converge en loi vers un processus de Poisson composé (CPP).

Voici les énoncés des principaux théorèmes contenus dans ce chapitre 5. On se restreindra au modèle suivant :

$$X_t = \begin{cases} \alpha_0 X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} \leq r_0 \\ \beta_0 X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} > r_0 \end{cases} \tag{1.4.1}$$

Le paramètre (α, β, r) est estimé par $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{r}_n)$, obtenu par la méthode des moindres carrés.

Comme annoncé, le bruit est dit faible, c'est à dire qu'il satisfait

(H1) La suite $(\varepsilon_t)_{t\in\mathbb{Z}}$ est stationnaire (au sens strict), admettant des moments d'ordre 4, centrée et non corrélée.

Le premier résultat à établir sera la consistance de notre estimateur. Cela se fera sous les hypothèses suivantes concernant le processus X :

(H2) Le processus $(X_t)_{t\in\mathbb{Z}}$ est ergodiques, stationnaire, admet des momoents d'ordre 4. De plus, pour tout t, la loi de X_t admet une densité π lipschitzienne et strictement positive sur tout intervalle borné.

Afin que le modèle soit effectivement un modèle TAR, nous supposerons que

(H3) $\alpha_0 \neq \beta_0$

Sous le contexte des hypopthèses ci-dessus, on aura le résultat suivant :

Théorème : Soit $(X_t)_{t\in\mathbb{Z}}$ le processus TAR satisfaisant (1.4.1). Supposons que les hypothèses **(H1)**, **(H2)** et **(H3)** sont vérifiées. Alors $\hat{\theta}_n \to \theta_0$ presque sûrement quand $n \to +\infty$.

Pour aller plus loin dans notre étude, il faudra ajouter d'autres hypothèses. Notre cadre de bruit non indépendant nous imposera de faire des hypothèses de mélange sur le processus X. Ceci est très naturel quand on se réfère aux différents travaux de Francq et Zakoian. Nous supposerons donc que

(H4) $(X_t)_{t\in\mathbb{Z}}$ satisfait la condition de mélange fort suivnate : il existe $\nu > 0$ tel que

$$\sum_{k=0}^{\infty} \{\alpha_X(k)\}^{\frac{\nu}{2+\nu}} < \infty \ . \tag{1.4.2}$$

Cette condition nous permettra d'utliser l'inégalité de Davydov. Nous aurons alors besoin d'augmenter notre hypothèse sur les moments de X :

(H5) $(X_t)_{t\in\mathbb{Z}}$ satisfies $\mathbb{E}|X_t|^{4+2\nu}$ with the real ν from Assumption **(H4)**.

Quand le bruit est iid, X satisfait une propriété de mélange géométrique qui est plus forte que nos conditions **(H4)** et **(H5)**. En effet le mélange géométrique signifie que $\alpha_X(k) = O(\rho^k)$ pour un $0 < \rho < 1$ tandis que notre hypothèse indique une décroissance en puissance de h.

Théorème : Sous les conditions **(H1)** à **(H5)**, nous avons

1. $n^\kappa(\hat{r}_n - r_0) = \mathrm{O}_{\mathbb{P}}(1)$ with $\kappa = (2+\nu)/(3+2\nu)$.
2. $\sup_{|r-r_0|\le \frac{B}{n}} \Big(|\hat{\alpha}_n(r) - \alpha_0| + |\hat{\beta}_n(r) - \beta_0|\Big) = \mathrm{o}_{\mathbb{P}}(1)$.

La normalité asymptotique sera aussi obtenue sous ce jeu d'hypothèse. Plus particulièrement on obtient le résultat suivant :

Théorème : On suppose que **(H1)** à **(H5)** sont vérifiées. L'estimateur $\hat{\lambda}_n(\hat{r}_n) = \begin{pmatrix} \hat{\alpha}_n(\hat{r}_n) \\ \hat{\beta}_n(\hat{r}_n) \end{pmatrix}$ satisfait $\sqrt{n}(\hat{\lambda}_n(\hat{r}_n) - \lambda_0) = \sqrt{n}(\hat{\lambda}_n(r_0) - \lambda_0) + \mathrm{o}_{\mathbb{P}}(1)$ et $\sqrt{n}(\hat{\lambda}_n(r_0) - \lambda_0)$ converge en loi vers une loi normale de moyenne nulle et de matrice de covariance sous la forme "sandwich" $J^{-1}IJ^{-1}$ avec

$$J = 2\begin{pmatrix} \mathbb{E}(X_1^2\mathbf{1}_{\{X_1\le r_0\}}) & 0 \\ 0 & \mathbb{E}(X_1^2\mathbf{1}_{\{X_1>r_0\}}) \end{pmatrix} \quad \text{et} \quad I = \lim_{n\to\infty}\left(\sqrt{n}\frac{\partial L_n(\lambda_0, r_0)}{\partial\lambda}\right).$$

Les résultats énoncés ci-dessus font appels à un mélange des techniques employées dans [42] et [24].

L'étude de la loi limite de l'estimateur du seuil $\hat{r}_n$ a nécessité de développer des techniques nouvelles. On réfère au chapitre 5 pour plus de détails et nous ne donnons dans cette introduction que les idées principales. Dans [24] il est montré que $n(\hat{r}_n - r_0) \to M_-$ où M_- est un minimum d'un processus de Poisson composé. Nous obtiendrons aussi ce type de résultat. L'idée, provenant de Chan, est de montrer que notre fonction des moindres carrés que l'on miminise, convergera dans l'espace de Skorohod, vers un Processus de Poisson. Ensuite il faut utiliser les résultats de Seijo et Sen (voir [76, 77]) qui montrent que la fonction Argmin est continue sur l'espace de Skorohod. Ainsi, $n(\hat{r}_n - r_0)$ sera lié à un minimum du processus de Poisson limite.

Mais dans les travaux de Chan, le bruit est fort, c'est à dire i.i.d. et la convergence vers le processus de Poisson s'obtient grâce à des techniques qui font intervenir le contexte i.i.d. et ne sont pas applicable à notre contexte. Donc, dans le cas d'un bruit faible, la situation est beaucoup plus difficile et très peu de résultats existent sur ce type de convergence (même dans un contexte de régression avec seuil et bruit faible). Nous avons pu surmonter cette difficulté technique en utilisant un travail assez récent de Chigansky et Klebaner [26]. Nous parvenons ainsi à établir un résultat de convergence analogue à celui du cas i.i.d. en imposant l'hypothèse supplémentaire suivante de mélange local :

(H6) Il existe un réel a avec $\nu/(2+\nu) < a < 1$ tel que pour tout r on a

$$\lim_{n\to\infty}\sum_{k=1}^{n}\sum_{j;|j-k|\le n^a, j\neq k}\mathbb{E}\Big(\mathbf{1}_{\{r<X_{k-1}\le r+1/n\}}\mathbf{1}_{\{r<X_{j-1}\le r+1/n\}}\Big) = 0\ . \tag{1.4.3}$$

Par stationarité, on notera que (1.4.3) peut s'écrire

$$\lim_{n\to\infty} n \sum_{h=1}^{n^a} \mathbb{E}\Big(\mathbf{1}_{\{r<X_1\leq r+1/n\}}\mathbf{1}_{\{r<X_{h+1}\leq r+1/n\}}\Big) = 0 \ .$$

Ce genre d'hypothèse apparaissait déjà dans l'article de Berman (voir [9]) qui traite la convergence de tableaux triangulaires vers des processus de Poisson composés. Sous indépendance et sous l'hypothèse **(H2)**, **(H6)** satisfaite.

A noter pour terminer que nous nous sommes restreint au cas d'un modèle TAR(1), c'est à dire qu'il y a qu'un décalage temporel d'ordre 1 contrairement au modèle (1.3.1) qui prend en compte des décalages jusqu'aux ordres p_1 et p_2. Ce cas général demeure compliqué et nous comptons l'étudier dans des travaux ultérieurs. Nous nous sommes focalisés sur le TAR(1) pour pouvoir bien mettre en avant les nouvelles techniques employées par rapport au cas fort.

Chapitre 2

Detection multiple change points by Filtered Derivative and False Discovery Rate

This chapter consists in article titled " **Detection multiple change points in Filtered Derivative and False Discovery Rate**, published in "International Journal of Statistics and Probability", Vol 1 N 1 p 12–23, 2104, a proceeding titled "**A real application of Filtered Derivative and False discovery Rate**", published in " 46ieme Journées De Statistiques", organized by la Société Française de la Statistique", 1 Jun 2014–06 jun 2014 at Rennes, France and a proceeding titled "**Multiple change point detection in linear regression by Filtered Derivative and False Discovery Rate**", p 0–5, published by International Statistics Institute, 26 July–31 July 2015, Rio, Brazil.

2.1 Multiple Change point detection by Filtered Derivative and False Discovery Rate for the paramater mean.

This article is published under reference "International Journal of Statistics and Probabilty", Vol 1 N 1 p 12–23, 2104"

Abstract

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a time series, that is a sequence of random variable indexed by the time $t = 1, 2, \ldots, n$. We assume the existence of a segmentation $\tau = (\tau_1, \tau_2, \ldots, \tau_n)$ such that X_i is a family of independent identically distributed (i.i.d) random variable for i $\in (\tau_k, \tau_k + 1]$, *and* $k = 0, \ldots, K$ where by convention $\tau_o = 0$ and $\tau_{K+1} = N$. In the literature, it exists two main kinds of change points detections : The change points on-line and the change points off-line. In this work, we consider only the change point analysis (off-line), when number of change points is unknown. The result obtained is based on Filtered Derivative method where we use a second step based on False Discovery Rate. We compare numerically this new method with the Filtered Derivative with p-Value.

2.1.1 Introduction

Change-point detection is an important problem in many applications, and it has been well-studied for a long time , see e.g. the textbooks [8, 18, 28], or [45, 53] for an updated

overview. Depending on the method of data acquisition, there exist two different kinds of change detection : *A posteriori* or off-line change-point detection arises when the series of observations is complete at the time we process the data, whereas in sequential analysis, the detection is performed on line. In this work, we only consider the *a posteriori* problem. In this century, the state to the art method was the Penalized Least Square Criterion (PLS) : When the number of change point is known, PLS minimizes a contrast function [4, 56]. When the number of change point is unknown, many authors use the penalized version of the contrast function [57, 58]. From a computational point of view, PLS methods use the dynamic programming algorithms and it needs to compute a matrix. Therefore, the time and memory complexity of PLS algorithm is of order $\mathcal{O}(n^2)$, where n denote the size of the dataset. Due to the *data deluge*, the size of datasets are larger and larger, then the computational complexity of statistical method has become a challenge. Cumulative sum can be iteratively computed and therefore leads to algorithms with both time and memory complexity of order $\mathcal{O}(n)$. Among these methods, the Filtered Derivative has been introduced by [6, 8]. The advantage of Filtered Derivative method is the time and memory complexity, both of order $\mathcal{O}(n)$. On the other hand, Filtered Derivative method leads to many false discoveries of change points. Recently, [11] have introduced a method called Filtered Derivative with p-value (FDpV) (see below for more details). Change detection by FDpV method has been successfully applied to real life large datasets ($n = 120,000$ or $n = 40,000$) of heartbeat series [35]. However, Step 2 of FDpV algorithm use single hypothesis tests, and therefore it does not allow to control the rate of false discoveries. In this work, we propose to replace the family of single hypothesis tests of Step 2 in FDpV method by the use of the False Discovery Rate. The False Discovery Rate (FDR) has been introduced for multiple tests [7]. Moreover, we investigate the effect of adding a Step 3, for taking advantage of the enlargement of windows when the number of potential change point decreases.

The rest of this paper is structured as follows : Section 1 describes the problem and the comparison criterions. Section 2 recall the methods (FDpV and PLS) used for off-line change detection. Section 3 described the new method proposed in this work (FDqV), then Section 4 contains the numerical comparison. Eventually, the choice of the extra-parameters for FDpV or FDqV method is discussed in Section 5.

2.1.2 Description of the Problem

In this section we describe the problem of change point analysis and we give some comparison's criterion. For sake of simplicity, we restrict ourselves to a toy model, since we still have checked on real life datasets the efficiency of FDpV method, see [11].

Change point analysis : a toy model

Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ be a series indexed by the time $\mathbf{t} = 1, 2, \ldots, N$. We assume that there exists a segmentation $\tau = (\tau_1, \ldots, \tau_K)$ such that X_t is a family of independent identically distributed (iid) random variables for $t \in (\tau_k, \tau_{k+1}]$, and $k = 0, \ldots, K$, where by convention $\tau_0 = 0$ and $\tau_{K+1} = N$. The most simple model is X_t a sequence of independent Gaussian variable with $X_t \in \mathcal{N}(\mu(t), \sigma)$, where $\mathcal{N}(\mu, \sigma)$ denote the Gaussian law with mean μ and standard deviation $\sigma = 1$, and $t \mapsto \mu(t)$ is a piecewise constant map, that is $\mu(t) = \mu_k$ for all time $t \in (\tau_k, \tau_{k+1}]$. We will use this model in all the sequel of this work.

Comparison criteria

Assume that we do not know in advance the number K of change points. We have to estimate the configuration of change $\tau = (\tau_1, \ldots, \tau_K)$ and the values of the mean $(\mu_0, \mu_1, \ldots, \mu_K)$. We denote the estimates by $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_K)$ and $(\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_K)$. Remark that the number of change points is unknown and estimated by $\widehat{K}$.

Criteria

1. The quality of estimation for one sample can be measured by two criteria :
   - $\hat{K} - K$
   - The integrated square error (ISE). Actually, we can reformulate the problem as an estimation of a noisy signal. The signal is
   $$s(t) = \sum_{k=0}^{K} \mu_k \times \mathbf{1}_{(\tau_k, \tau_{k+1}]}(t)$$
   where we have set by convention $\tau_0 = 0$ and $\tau_{K+1} = N$. The estimated signal is then
   $$\widehat{s}(t) = \sum_{k=0}^{\hat{K}} \widehat{\mu}_k \times \mathbf{1}_{(\widehat{\tau}_k, \widehat{\tau}_{k+1}]}(t)$$
   and the integral square error (ISE) by
   $$ISE = \sum_{i=1}^{N} \left\{ [\widehat{s}(t) - s(t)]^2 \right\}$$
2. However, a result on just one simulation is hazardous. So, we have to do M simulations, with e.g. $M = 1,000$ and calculate the mean integrated square error (MISE).
3. The second family of criterion is the time complexity and the memory complexity that is the mean CPU time for estimating $\widehat{s}$ and which quantity of memory is used.

2.1.3 Some methods for change point analysis

In this section, we recall some methods for change point analysis : The Penalized Least Square Error (PLS) and the Filtered Derivative with p-value (FDpV).
Step 1. The first step is the same as in FDpV : We compute the filtered derivative function $t \mapsto FD(t, A)$ and then select the potential change points as the local maxima of the function $t \mapsto |FD(t, A)|$ reaching a threshold C_1.

Penalized Least Square Method

Set $\mathcal{S}_\mathcal{K} = \{\tau \; such \; that \; length(\tau) = K, \; that \; is \; \tau = (\tau_1, \ldots, \tau_K)\}$ the set of all possible configuration of change of length K . Firstly, when the number of change points K is known, for each configuration of change $\tau \in \mathcal{S}_\mathcal{K}$, we can define

$$\widehat{\mu}_k = mean(X, (\tau_k, \tau_k + 1]) := \frac{1}{(\tau_{k+1} - \tau_k)} \sum_{i=\tau_k+1}^{\tau_{k+1}} X_i, \; for \; k = 0, \ldots, K \tag{2.1.1}$$

where $mean(X,\ Box)$ denotes the mean of the family X_t for the indices $t \in Box$. Next, we search the configuration of change $\widehat{\tau}_K \in \mathcal{S}_\mathcal{K}$ which minimizes the square error $\mathrm{Q}(\tau)$ defined by

$$\mathrm{Q}(\tau) = \sum_{k=0}^{K} \sum_{t=\tau_k+1}^{\tau_{k+1}} |X_t - \widehat{\mu}_k|^2 \tag{2.1.2}$$

and we denote it by $\widehat{\tau}^K$. Secondly, we consider that the number of change points K is unknown. We remark that the map $K \longmapsto \mathrm{Q}(\widehat{\tau}^K)$ is decreasing. So minimizing the function $\mathrm{Q}(\tau)$ with an unknown number of changes will lead to consider as optimal the trivial configuration of changes $\tau^\star = (1, 2, \ldots, N)$. To avoid this drawback, we add a penalty term proportional to the length of the change point configuration. Eventually we want to minimize

$$pen(K) = \mathrm{Q}(\widehat{\tau}^K) + \beta \times K \quad \text{for } K = 0 \ldots, N.$$

Different choices of the penalty coefficient β are possible. In [56], the following choice is proposed :

$$\beta_1 = \frac{2\sigma^2(logn)}{n}.$$

In [13, 58], the proposed choice is

$$\beta_2 = \frac{\sigma^2}{n} \times \left[2 + 5 \times \log(\frac{n}{K})\right]$$

where σ^2 is the variance assumed to be constant and known and n the size of the series. In Fig. 1 below, we have plotted the contrast function and the penalized contrast function [56]. We clearly see that the penalized contrast is almost horizontal, thus the minimum value is very fluctuating with respect to the choice of the parameter β.

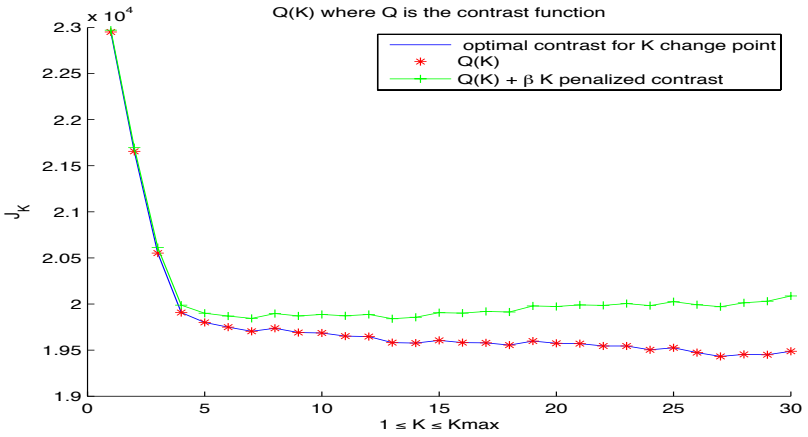


FIGURE 2.1 – *blue : Q(K) calculated with dynamical program method ; red : the penalized contrast function ; green : the optimal contrast function for K change points .*

Let us stress that both time and memory complexity of PLS method is $\mathcal{O}(n^2)$, see e.g. [56–58].

Description of the Filtered Derivative with p-Value Method (FDpV)

This method is a two steps procedure for change detection : Step 1 is based on Filtered Derivative and select a set of potential change points, whereas Step 2 calculate the p-value associated to each potential change point, for disentangling right change points and false alarms. More precisely, the method is defined as follows :

1. **Step 1 : Computation of the filtered derivative function**
   The filtered derivative function [6, 8, 10] is defined :

$$FD(t, A) = \widehat{\mu}(t+1, t+A) - \widehat{\mu}(t-A, t), \ \ for\ A < t < N - A \tag{2.1.3}$$

   where

$$\widehat{\mu}(t+1, t+A) := A^{-1} \sum_{j=t+1}^{t+A} X_j$$

   denote the empirical mean of the variables X_j on $(t+1, t+A)$.
   Next, remark that quantities $A \times FD(t, A)$ can be iteratively calculated by using

$$A \times FD(t+1, A) = A \times FD(t, A) + X(t+1+A) - 2X(t+1) + X(t-A). \tag{2.1.4}$$

   Thus, the computation of the whole function $t \longmapsto FD(t)$ for $t \in [A,\, n-A]$ requires $\mathcal{O}(n)$ operations and the storage of n real numbers.

2. **The determination of the potential change points.**
   Let us point that the absolute value of filtered derivative $|FD|$ presents hats at the vicinity of the change points see Fig. 2 below.

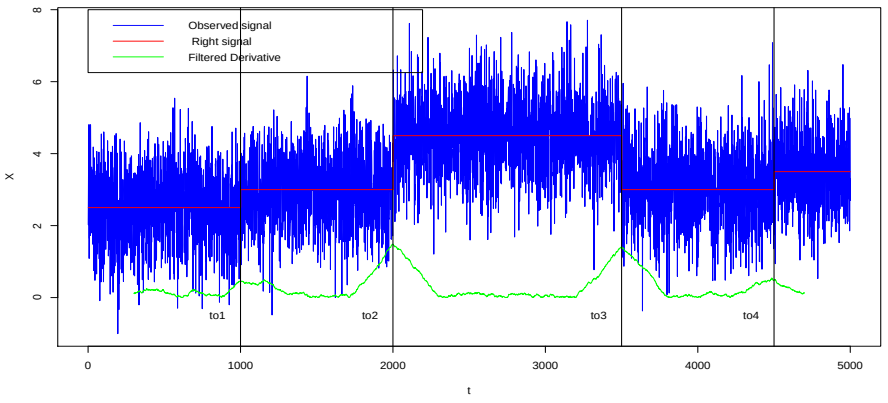


FIGURE 2.2 – *The right signal (red), the noisy signal (blue), and Filtered Derivative function (green).*

Potential change points τ_k^*, for $k = 1, \ldots, K^*$, are selected as local maxima of the absolute value of the filtered derivative $|FD(t, A)|$ where moreover $|FD(\tau_k^*, A)|$ exceed a given threshold C_1. In [10, 11], we have given the asymptotic distribution

of the maximum $|FD|$ under the null hypothesis. Therefore, we can fix the error type at level $p_1^\star$, and then we can deduce the threshold C_1 corresponding to $\Pr(\max|FD(\tau_k, A)| > C_1) = p_1^*$. We can remark the existence of many local maxima in the vicinity of each right change point (see Fig. 4 and [10, 11] for theoretical explanation). On the other hand, if there is no noise that is when $\sigma = 0$, we get hats of width $2A$ and hight $\mu_{k+1} - \mu_k$ at each change point τ_k, see Fig. 3 above.

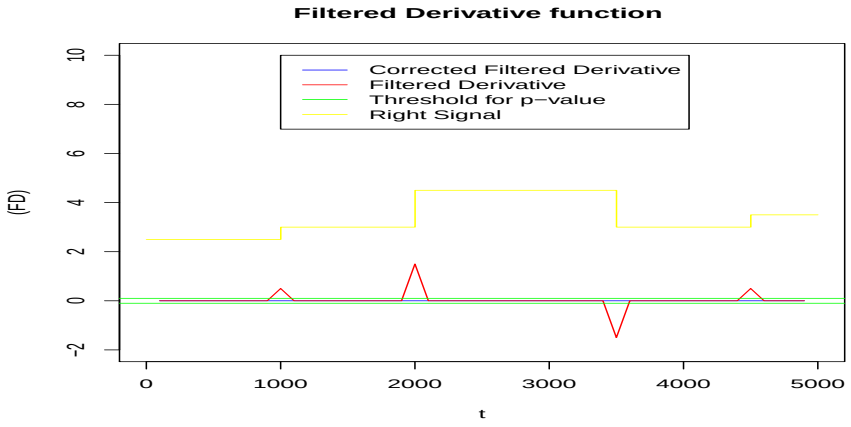


FIGURE 2.3 – *Filtered Derivative function without noise ($\sigma = 0$).*

For this reason, we select as first potential change point τ_k^* the global maximum of the function $|FD_k(t, A)|$, then we define the function FD_{k+1} by putting to 0 a vicinity of width $2A$ of the point τ_k^* and we iterate this algorithm while $|FD_k(\tau_k^*, A)| > C_1$, see [11]. When there is noise (e.g. $\sigma = 1$), we get the following landscape, see Fig. 4.

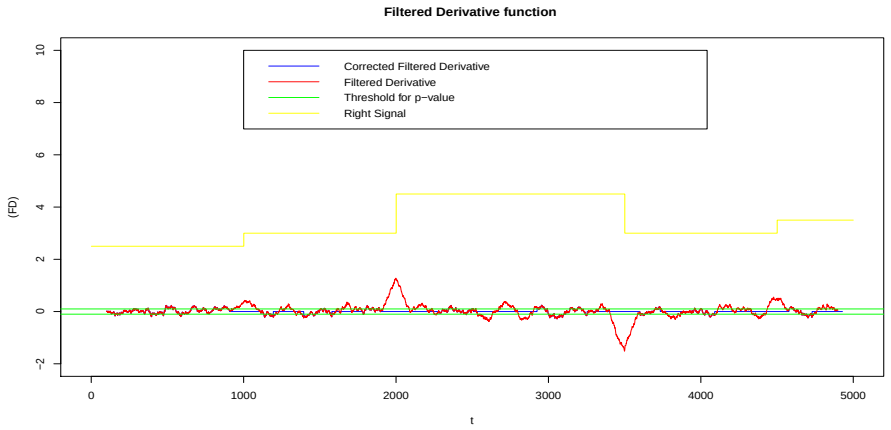


FIGURE 2.4 – *Filtered Derivative function with noise* ($\sigma = 1$).

3. **Step 2 : Elimination of false alarm by p-value.** A potential change point $\tau_k^\star$ can be an estimator of a right change point or a false alarm. In the first case, there exists an error of estimation on the location of the change. So we have to cancel a small vicinity of size ε_k around each point $\tau_k^\star$, [10, 11]. Then, for each segment, we calculate an estimation of the mean

$$\widehat{\mu}_k := mean(X, \tau_k + \varepsilon_k, \tau_{k+1} - \varepsilon_{k+1}). \tag{2.1.5}$$

Next, we have to eliminate false detection in order to keep (as possible) only the right change points. In [11], we use as Step 2 single hypothesis tests : For each potential change point $\tau_k^\star$, we test wether the parameter is the same for $t \in (\tau_{k-1}^\star + \varepsilon_{k-1}, \tau_k^\star - \varepsilon_k)$ and $t \in (\tau_k^\star + \varepsilon_k, \tau_{k+1}^\star - \varepsilon_{k+1})$ or not. More formally, for all $1 \leq k \leq K$, we apply the following hypothesis testing $(H_{0,k}) : \widehat{\mu}_k = \widehat{\mu}_{k+1}$ versus $(H_{1,k}) : \widehat{\mu}_k \neq \widehat{\mu}_{k+1}$ where $\widehat{\mu}_k$'s are defined by (2.1.5). By using this second single hypothesis test, we calculate the p-values $p_1^\star, \ldots, p_{K^\star}^\star$ associated to each potential change point $\tau_1^\star, \ldots, \tau_{K^\star}^\star$.

Calculation of p-value

We choose the statistic Student T. Indeed, under the null hypothesis, $t_k^\star$ has a Student distribution of degrees of freedom $d = N_k + N_{k-1} - 2$, where

$$t_k^\star \quad = \quad \frac{\widehat{\mu}_k - \widehat{\mu}_{k-1}}{\sqrt{\frac{S_{k-1}^2}{N_{k-1}} + \frac{S_k^2}{N_k}}}, \tag{2.1.6}$$

$\widehat{\mu}_k$'s are given by (2.1.5), $N_k = \Big\{(\tau_{k+1}^* - \varepsilon_{k+1}) - (\tau_k^* + \varepsilon_k)\Big\}$, and $S_k^2 = \left\{ \left(\frac{1}{N_k} \sum_{t=\tau_k+\varepsilon_k}^{\tau_{k+1}-\varepsilon_{k+1}} X_t^2 \right) - \overline{X_k}^2 \right\}$.

By construction, $d > 2A - (\varepsilon_{k-1} + 2\varepsilon_k + \varepsilon_{k+1})$, thus for $A > 30$ the distribution of $t_k^\star$ is approximatively Gaussian an we can set

$$p_k^\star \quad \approx \quad 2 \times \Big\{1 - \Phi(|t_k|)\Big\} \tag{2.1.7}$$

where Φ is the cumulative distribution function of the zero mean standard Gaussian law. Let us point a slight difficulty : Since τ_k^* maximizes the criterium $|FD_k(t, A)|$, τ_k^* is also a random variable. We avoid this drawback by canceling a small vicinity of size ε_k for each selected change point, see Formula (2.1.5) and [10] [Rem. 2.1, p. 178–179] for details.

In [11], we only keep the change points corresponding to a p-value lesser than a fixed threshold $p_2^\star$. Consequently, Step 2 is much more selective and it allows us to deduce an estimator of the piecewise constant map $t \longmapsto \mu(t)$, see Fig. 5 below.

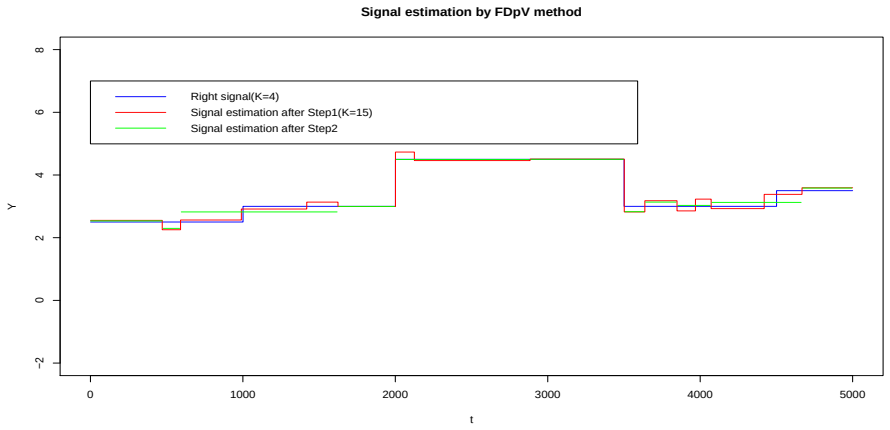


FIGURE 2.5 – *Signal reconstruction after Step2 by FDpV method*

2.1.4 A new Method for Change Point Analysis : Filtered Derivative with a q False Discovery Rate (FDqV)

We propose a new method derived from the FDpV one : We replace Step 2 of FDpV by False Discovery Rate method (FDR) and we call this method FDqV.

Step 1. The first step is the same as in FDpV : We compute the filtered derivative function $t \mapsto FD(t, A)$ and then select the potential change points as the local maxima of the function $t \mapsto |FD(t, A)|$ reaching a threshold C_1.

Step 2. The novelty of this work is the use of False Discovery Rate thresholding procedure [7]. The computation of p-value $p_k^\star$ is the same as for Step 2 of FDpV method. However, we then use a Bonferroni type multiple testing procedure :

- We tidy up p- value in the increasing order $p^*_{(1)} \leq \ldots \leq p^*_{(K^*)}$.
- We choose a threshold q corresponding to the rate of false alarms or FDR.
- We keep only the potential change points τ_i^* corresponding to a p-value $p^*_{(i)}$ such that $p^*_{(i)} \leq \dfrac{i}{K^*} q$.

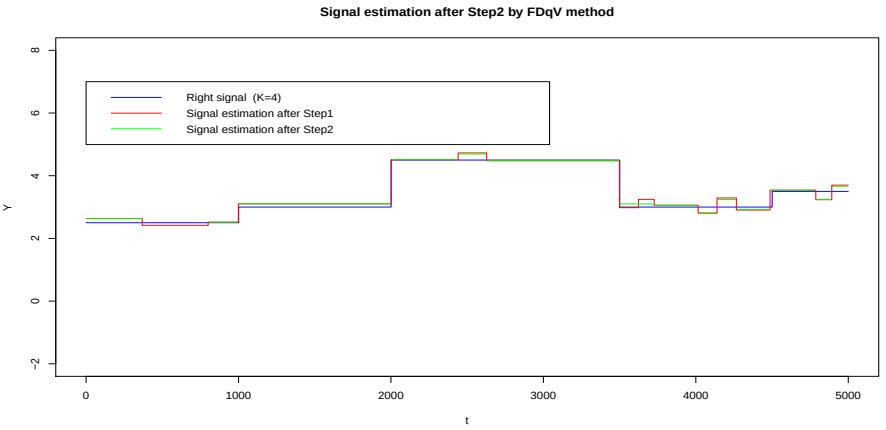


FIGURE 2.6 – *Signal reconstruction after Step2 by FDqV method*

Step 3. Let us point that Step 1 of FDpV or FDqV select potential change point as local maxima of the absolute value of the filtered derivative function, which is the difference of the mean estimated on sliding window of size A on a box at left and at right of the point t. Then Step 2 select some change points corresponding to a p-value smaller to a fixed threshold (FDpV) or a linear threshold (qFDR). In both case, the p-value is computed following the potential change point selected in Step 1. We recall here that these p-values are therefore calculated with windows larger than A. In other words, we have more information on the mean at Step 2 than at Step 1. This remark has suggested us to add a third step, which is the same as Step2 but with larger windows. More formally, Step 2 of FDqV can be seen as a map $FDR : (X, \tau^*, q) \longmapsto \widetilde{\tau}$, where $\widetilde{\tau} = (\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{K}})$ with $\widetilde{K} \leq K^*$. Thus, at Step 3, we plug $\widetilde{\tau}$ instead τ^* as input of Step 2, that is $FDR : (X, \widetilde{\tau}, q) \longmapsto \widehat{\tau}$, where $\widehat{\tau} = (\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{K}})$ with $\widehat{K} \leq \widetilde{K} \leq K^*$.

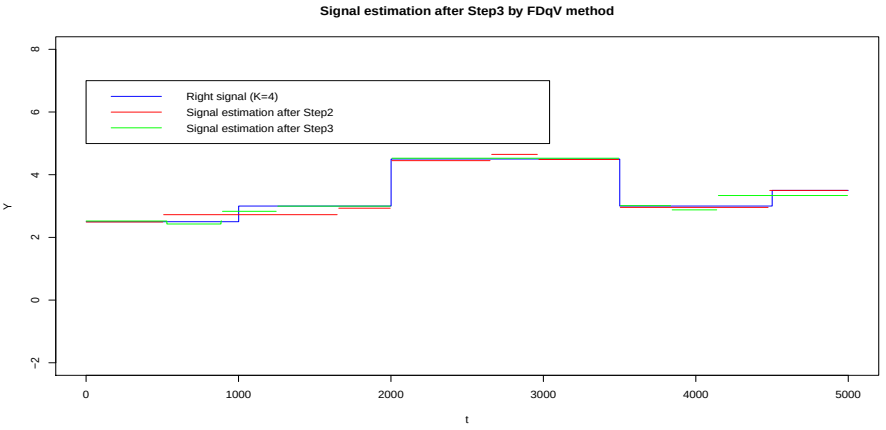


FIGURE 2.7 – *Signal reconstruction after Step3 by FDqV method*

To sum up, in [11] we have made simple statistics tests and we compare the means pairwisely. So we choose a threshold of critical probability to eliminate the false alarms. The novelty in this work is the use of a multiple test (FDR) with FDR fixed at level q. Both time and memory complexity of FDqV remain of order $\mathcal{O}(n)$.

2.1.5 Numerical Comparisons

In this section, we compare numerically the FDpV method and the new proposed method FDqV. We use Monte Carlo simulation and via MISE.

Simulations based on one realization

Firstly, we select the simulation for one realization, which corresponds to Fig. 2, and Fig. 4–8 above. For $n = 5,000$, we have simulated one replication of a sequence of Gaussian random variables $(X_1, \ldots, X_n)$ with variance $\sigma^2 = 1$ and mean $\mu_t = f(t)$ where f is a piecewise constant function with four change points at times $\tau = (1000, 2000, 3500, 4500)$ with means $\mu = (2.5, 3, 4.5, 3, 3.5)$. Both FDpV and FDqV method depend on extra-parameters, namely the window size A, the threshold C_1 corresponding to Step 1, the maximum number changes K_{max} for Step 1, the threshold p_2^* corresponding to Step 2 of FDpV, the uncertainties on the location of changes ε_k, and the threshold q of False Discovery Rate for Step 2 and Step 3 of FDqV. A brief discussion on the choice of the extra-parameters is postponed in Section 5. We have made the following choices : $A = 100$, $K_{max} = 15$, $C_1 = 0.1$, $p_2^* = 2 \times \Big\{1 - \Phi(1.5)\Big\} = 0.134$, $q = 0.1$.

Monte-Carlo Simulation

In this subsection, we made $M = 1,000$ simulations of independent iterations of sequences of Gaussian random variables $(X_0^j, \ldots, X_n^j)$ with variance $\sigma^2 = 1$ and mean $\mu_t = f(t)$, for $j = 1 \ldots, M$ and $t = 1, \ldots, N$. On each sample, we apply the FDpV method and the FDqV method with the extra-parameters given above. The mean value of $(\widehat{K} - K)$ is 3.38 with standard deviation (std) 1.64 for FDpV against a mean 2.84 with $std = 1.59$ for FDqV at Step 2, and mean $\hat{K} - K = 0.65$ with $std = 1.98$ for FDqV at Step3. Thus we can see that than the number of false discovery is smaller by FDqV. Note that at Step1, we have $mean(\hat{K} - K) = 12$.

Mean Integrate Square Error (MISE)

For $M = 1,000$ Monte Carlo simulations, we obtain the following values of MISE :

- for Filtered Derivative MISE=1419.12
- for FDpV MISE=189.59
- for FDqV (Step 2) MISE=148.75
- for FDqV (Step3) MISE=126.97

2.1.6 Numerical conclusion

We clearly see that the overestimation of the number of change points is smaller for the method FDqV than the for FPpV one. On the other hand, we can note that for the MISE criterion FDqV is better than FDpV, which is still better than Filtered Derivative. We

clearly see that the overestimation of the number of change points is smaller for the method FDqV than the for FPpV one. On the other hand, we can note that for the MISE criterion FDqV is better than FDpV, which is still better than Filtered Derivative.

2.1.7 How to choose the extra-parameters ?

In this section, we address the question of the choice of extra-parameters for FDpV method. Natural criteria are the error of type I and error of type II, so-called probability of false alarm (PFA), denoted α, and probability of non - detection (PND), denoted β.

Errors of type I and type II at Step 1

We stress that error of type II (PND) is more important than error of type I (PFA), at least for the ISE criterion : Indeed, just one change point missing increases strongly the error. On the other hand, as pointed out in [10], when there is more than one change, the notion of probability of non detection should be make more precise : For each right change point τ_k, we define the local PND as $\beta_{loc}(\tau_k) = \mathbf{P}(B_k)$ where $B_k = \Big\{\forall k \in [\tau_k - A, \tau_k + A], |D(A,k)| < C_1\Big\}$. Next, we can define the global PND as $PND_{global} = \mathbf{P}\Big(\bigcup_{k=1}^{K} B_k\Big)$. Next, we can obtain an upper bound for PND_{global}. On the one hand, let us denote by δ_k the size of the change on the mean at change point τ_k, more precisely $\delta_k = \mu_k - \mu_{k-1}$ for $k = 1, \ldots, K$. We have [10, Prop. 3. 2, p 222],

$$\mathbf{P}(B_k) \leq \Psi\left(\frac{\delta_k - C_1}{\sigma}\sqrt{\frac{A}{2}}\right) \times \Phi\left(\frac{C_1 - \delta_k/3}{\sigma}\sqrt{\frac{A}{2}}\right)^2 \tag{2.1.8}$$

$\Psi(x) = 1 - \Phi(x)$. Next, by remarking that the right side of (2.1.8)is a decreasing function of δ_k and setting $\delta = \inf_{k=1,\ldots,K} \delta_k$, we can deduce that

$$\mathbf{P}(B_k) \leq \beta^*(C_1, A) := \Psi\left(\frac{\delta - C_1}{\sigma}\sqrt{\frac{A}{2}}\right) \times \Phi\left(\frac{C_1 - \delta/3}{\sigma}\sqrt{\frac{A}{2}}\right)^2 \tag{2.1.9}$$

On the other hand, we obviously have

$$PND_{global} \leq \sum_{k=1}^{K} \mathbf{P}\Big(B_k\Big)$$

which combined with (3.4.5) gives us

$$PND_{global} \leq K \times \beta^*(C_1, A).$$

The right number K is unknown, but fixed. Thus, we will control the quantities $\beta^*(C_1, A)$, for instance we choose to set $\beta^*(C_1, A) = 10^{-4}$. This equation can be numerically solved, since the map $C_1 \longmapsto \beta^*(C_1, A)$ is decreasing, and we find an implicit function $A \mapsto C_1(A)$. After having controlled the error of type II (PND), we can control the error of type I (PFA). We know [10, Prop. 3. 1, p 221] that for all $\varepsilon > 0$ there exists a constant M_ε such that

$$\alpha \leq M_\varepsilon \times \alpha^*(C_1, A) := M_\varepsilon \times \left(\frac{n - A}{A}\right) \times \Psi\left(\frac{C_1}{\sigma}\sqrt{\frac{A}{2+\varepsilon}}\right). \tag{2.1.10}$$

For instance, we can set $\varepsilon = 0.1$, next we plug the implicit relationship between A and C_1 inside (2.1.10) and we obtain a function $A \longmapsto \alpha^*(C_1(A), A)$. The first idea is to make varying the parameter A in order to find the optimal value corresponding to a minimum of the map $A \longmapsto \alpha^*(C_1(A), A)$. Unfortunately enough, the map $A \longmapsto \alpha^*(C_1(A), A)$ is decreasing and reaches no minimum value.

Choice of the window A

From the preceding subsection, we can get the feeling that the larger the window size A is, the smaller type I and type II errors will be. This reasoning holds true as long as

$$2 \times A \quad < \quad L_0 := \inf\{|\tau_{k+1} - \tau_k|, k = 1, \ldots, K\}. \tag{2.1.11}$$

Thus, we have to choose a parameter $A < L_0/2$, even if we do not exactly know the quantity L_0. Fig 2.8–2.10 below illustrate the necessary condition (2.1.11). We consider the case without noise, that is $\sigma = 0$, with three change point at $\tau = (2000, 2200, 2600)$ thus $L_0 = 200$, and we make varying the window size A at $A = 100$, $A = 200$, and $A = 600$.

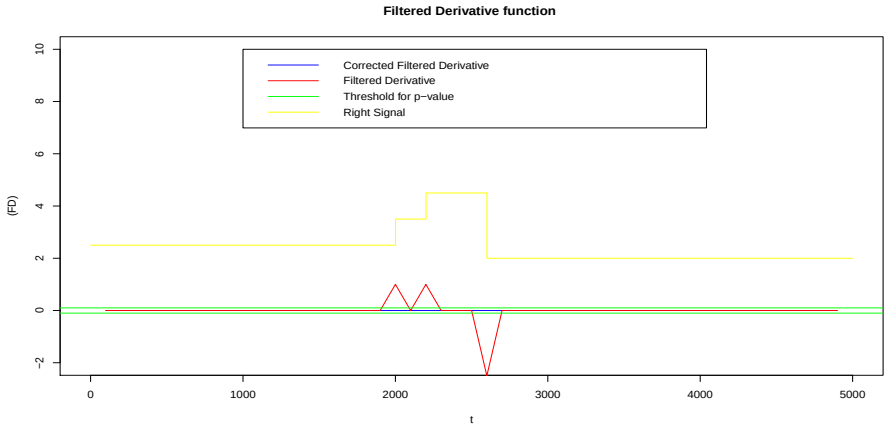


FIGURE 2.8 – *The Filtered Derivative without noise and A=100.*

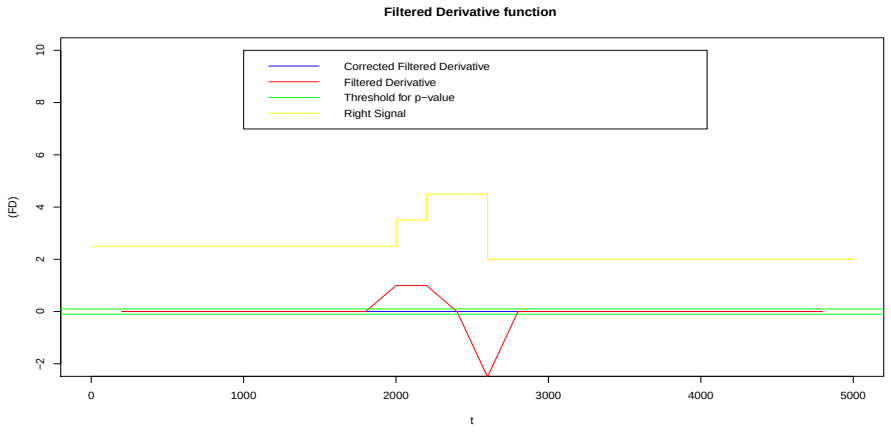


FIGURE 2.9 – *The Filtered Derivative without noise and A=200.*

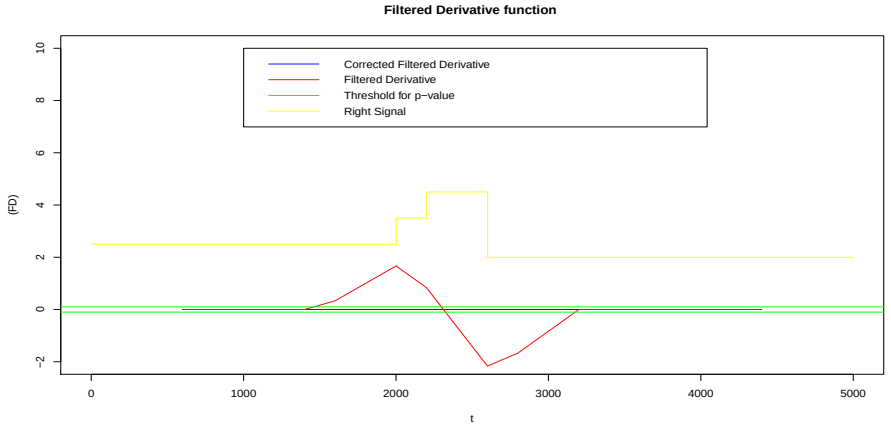


FIGURE 2.10 – *The Filtered Derivative without noise and A=600.*

In Fig. 2.8, we detect the three right change points. In Fig. 2.9 and Fig. 2.10, we only detect two change points. This plainly confirm the necessity of condition (2.1.11).

Error of type I and type II at Step 2

We can calculate t_k^* under both null and alternative assumption.

1. Under null assumption (H_0) : $\mu_k = \mu_{k+1}$, t_k^* approximatively follows a Gaussian law $\mathcal{N}(0,1)$.

2. Under alternative assumption (H_1) with $\mu_{k+1} - \mu_k = \delta$, then $t_k^* \sim \mathcal{N}(0,1) + \dfrac{\delta}{\sqrt{1/N_k + 1/N_{k+1}}}$.

On the other hand, the probability of one false alarm is

$$\alpha = PFA = \Pr(|\mathcal{N}(0,1)| > t_c) = 2 \times (1 - \Phi(t_c)).$$

For example, when $\delta = 0.5$, $t_c = 1.5$, $N_k = N_{k+1} = 100$ then $\beta_k = 0.0207$ and $\alpha = 0.1336$.

Actually, we want to select all the right change points with as few as possible false alarms. So, we want to control the probability of non detection PND. For a single change point, let us fix a critical level t_c, then

$$\begin{aligned} \beta_k = PND_k &= \Pr(|t_k^*| < t_c) \\ &= \Pr(-t_c < \mathcal{N}(0,1) + \frac{\delta}{\sqrt{1/N_k + 1/N_{k+1}}} < t_c) \\ &\simeq 1 - \Phi\left(\frac{\delta}{\sqrt{1/N_k + 1/N_{k+1}}} - t_c\right) \end{aligned}$$

2.1.8 Summary and conclusions

In this work, it clearly appears the power FDqV method than FDpV method. However, the FDqV method is established by the simulations, in the future we will valid by the real data. The questions not developed in the literature are :

- The FDpV and FDqV methods with the random variable weakly or strongly dependent.
- The Choice of parameters such the window and the threshold, which depend the both methods.

All these questions are very difficult so more detail is needed. We will try to do in forthcoming work

2.2 A real application of Filtered Derivative and False Discovery Rate

In this second part of this chapter, we give a real application of Filtered Derivative and False Discovery rate method. We precise that this proceeding is published by Statistical French society, 2014.

Résumé. Dans ce travail, nous donnons une application réelle de la méthode de dérivée filtrée avec le taux de fausses découvertes (FDqV). La FDqV utilise deux étapes, la première étape est la dérivée filtrée et la séconde étape utilise le taux de fausses découvertes pour éliminer les fausses alarmes et récupérer uniquement les vrais instants de ruptures. La domination de la FDqV par rapport á la dérivée filtrée avec p-value est clairement établie par le critère de l'erreur quadratique de la moyenne. Ici, nous utilisons des données fournis par EDF concernant des éolionnes implantés quelques part en France, nous détectons les instants de ruptures de la vitesse du vent sur une période donnée.

Mots-clés. Series Temporelles, Dérivée Filtrée, Taux de Fausses Découvertes

Abstract. In this work, we give a real application of the method of Filtered Derivative with False Discovery Rate (FDqV),[35] . This method use the Filtered Derivative (FD),[6, 35] as step 1 and a step 2 which use the False Discovery Rate [7] for elimination the false alarms at the end of step 2 and keep only as possible all right change points. The power of FDqV is provide in [35] for the criteria mean integrate square error (MISE) than Filtered Derivative with p-value (FDpV),[11]. Here we use a data given by electricity of France (EDF). It concerns wind turbines are implanted somewhere in France and we want to detect the breaks of the wind speed over a period.

Keywords. Time series, Filtered Derivative, False Discovery Rate

2.2.1 Introduction

In the literature, it exists two change points : The off-line detection or change points analysis and the on-line detection or sequential change points. Different methods for change point detection such that the penalized least square error [56–58], the filtered derivative [6], the filtered derivative with p-value [11] and the filtered derivative and false discovery rate [35] are used in the literature. In this work, we give an real application the filtered derivative and false discovery rate method. The rest of this paper is structured as followed : Section 2 describes the filtered derivative and false discovery rate, section 3 gives an real application of this method.

2.2.2 Recall method for change point analysis : Filtered Derivative and False Discovery Rate (FDqV)

Model

Let $\mathbf{X}= (X_1, X_2, \ldots, X_n)$ a sequence of independent random variable indexed by the time t= $1, 2, \ldots, n$. We suppose it exists a segmentation $\tau = (\tau_1, \tau_2, \ldots, \tau_K)$ with $\tau_k \in \{1, 2, \ldots, n\}$ and $\tau_1 < \tau_2 < \ldots \tau_K$. K denotes the number of changes. By convention, for the calculus of the mean, we set $\tau_o = 1$ and $\tau_{K+1} = n$. In other words, for $k = 0, \ldots K$, for i $=\tau_k + 1, \ldots \tau_{k+1}$, we have $X_k \sim \mathcal{N}(\mu_k, \sigma_k)$, where $\mathcal{N}(\mu, \sigma)$ is a gaussian law with mean μ and standard deviation σ.

The Filtered Derivative and False Discovery Rate (FDqV)

The FDqV method is introduced by [35] for a time series. In fact, The Filtered Derivative with False Discovery Rate has two steps : The Filtered Derivative and the False Discovery Rate

Step 1 : The Filtered Derivative

Step 1 is based on Filtered Derivative and select a set of potential change points, More precisely, we have

Computation of the filtered derivative function

Computation of the filtered derivative function, which is defined as the difference between the estimators of the mean computed in two sliding windows respectively at the right and at the left of the time t, both of size A, that is as the function :

$$FD(t, A) = \widehat{\mu}(t+1, t+A) - \widehat{\mu}(t-A, t), \ for\ A < t < N - A \tag{2.2.1}$$

where

$$\widehat{\mu}(t+1, t+A) := A^{-1} \sum_{j=t+1}^{t+A} X_j$$

denote the empirical mean of the variables X_j on the box $(t+1, t+A)$. This method consists in filtering data by computing the estimators of the parameter μ before applying a discrete derivation. So this construction explains the name of the algorithm, so called Filtered Derivative method [6]. Next, remark that quantities $A \times FD(t, A)$ can be iteratively calculated by using

$$A \times FD(t+1, A) = A \times FD(t, A) + X(t+1+A) - 2X(t+1) + X(t-A). \tag{2.2.2}$$

Thus, the computation of the whole function $t \longmapsto FD(t)$ for $t \in [A, n-A]$ requires $\mathcal{O}(n)$ operations and the storage of n real numbers. Let us point that the absolute value of filtered derivative $|FD|$ presents hats at the vicinity of the change points. Potential change points τ_k^*, for $k = 1, \ldots, K^*$, are selected as local maxima of the absolute value of the filtered derivative $|FD(t, A)|$ where moreover $|FD(\tau_k^*, A)|$ exceed a given threshold C_1. In [6,35], we have given the asymptotic distribution of the maximum $|FD|$ under the null hypothesis. Therefore, we can fix the error type at level $p_1^\star$, and then we can deduce the threshold C_1 corresponding to $\Pr(\max |FD(\tau_k, A)| > C_1) = p_1^*$. We can remark the existence of many local maxima in the vicinity of each right change point (see [6, 35] for theoretical explanation). On the other hand, if there is no noise that is when $\sigma = 0$, we get hats of width $2A$ and hight $\mu_{k+1} - \mu_k$ at each change point τ_k,. For this reason, we select as first potential change point τ_k^* the global maximum of the function $|FD_k(t, A)|$, then we define the function FD_{k+1} by putting to 0 a vicinity of width $2A$ of the point τ_k^* and we iterate this algorithm while $|FD_k(\tau_k^*, A)| > C_1$, see [6, 11, 35].

A potential change point $\tau_k^\star$ can be an estimator of a right change point or a false alarm. We want to eliminate false detection in order to keep (as possible) only the right change points. In [6], we use as Step 2 multiple hypothesis tests. More formally, consider K hypothesis tests for all $1 \leq k \leq K$, $(H_{0,k}) : \widehat{\mu}_k = \widehat{\mu}_{k+1}$ versus $(H_{1,k}) : \widehat{\mu}_k \neq \widehat{\mu}_{k+1}$ where $\widehat{\mu}_k$'s are defined as in the model. For each hypothesis test, we calculate the p-values $p_1^\star, \ldots, p_{K^\star}^\star$ associated to each potential change point $\tau_1^\star, \ldots, \tau_{K^\star}^\star$. After the calculation of p-value, we use a Bonferroni type multiple testing procedure :

1. We tidy up p- value in the increasing order $p^*_{(1)} \leq \ldots \leq p^*_{(K^*)}$.
2. We choose a threshold q corresponding to the rate of false alarms or FDR.
3. We keep only the potential change points τ^*_i corresponding to a p-value $p^*_{(i)}$ such that $p^*_{(i)} \leq \dfrac{i}{K^*}q$.

For more details see [35].

Simulation

For $n = 10,000$, we have simulated one replication of a sequence of Gaussian random variables $(X_1, \ldots, X_n)$ with variance $\sigma^2 = 1$ and mean $\mu_t = f(t)$ where f is a piecewise constant function with seven change points at times $\tau = (2000, 2500, 3000, 4000, 7000, 8000, 9000)$ with means $\mu = (2.5, 2, 3, 4.5, 3, 3.5, 4, 5.5)$. We have made the following choices : $A = 250$, $K_{max} = 20$, $C_1 = 0.25$, and $q = 10^{-6}$.

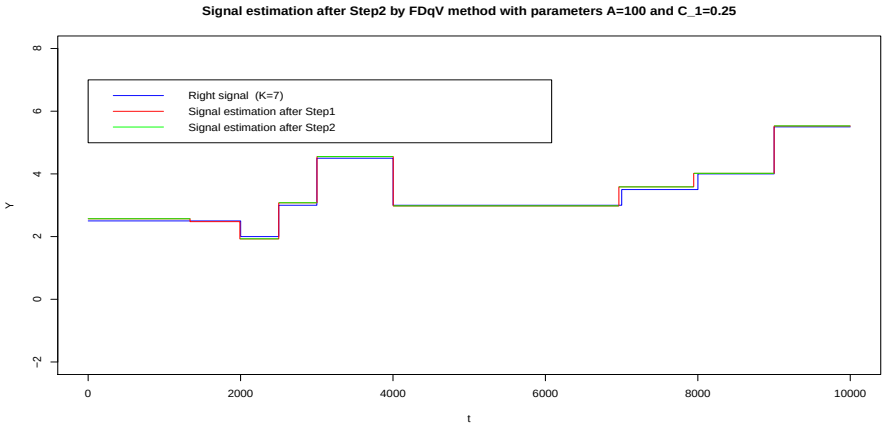


FIGURE 2.11 – *Signal reconstruction after Step2 by FDqV method*

2.2.3 A real application of Filtered Derivative and False Discovery Rate

In this paragraph, we want to apply the FDqV-method for a real application. The data concerns the wind speed of the wind turbines. We have 50598 observations and we want to detect abrupt changes of the wind speed over the time which corresponds when the wind speed change. We take the parameters followings A=144, Kmax=20, $C_1 = 0.1$ and $q = 10^{-6}$. The figure 2 corresponds the signal speed wind , the figure 3 give us when the instant of potential changes are produced and the last figure is the reconstruction of the signal by our method.

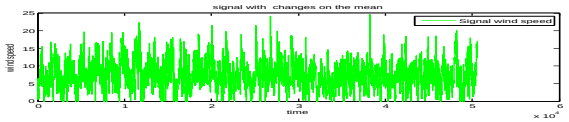


FIGURE 2.12 – *The signal of wind speed*

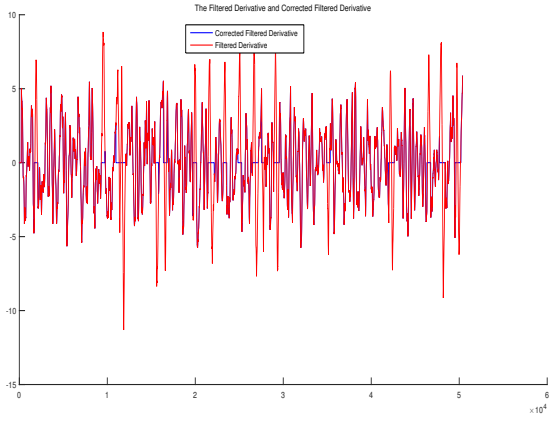


FIGURE 2.13 – *The Filtered Derivative and Corrected Filtered Derivative*

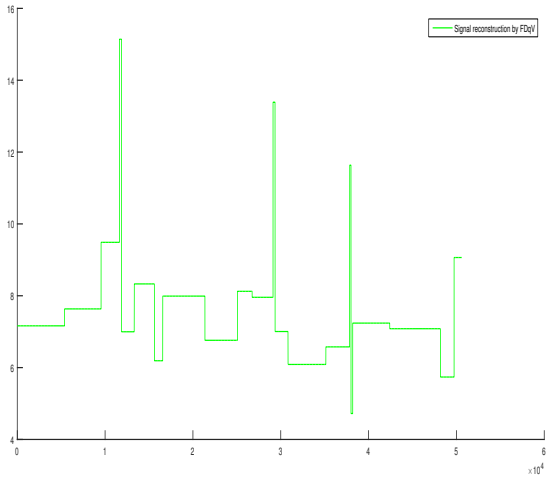


FIGURE 2.14 – *Signal estimation by FDqV method*

2.3 Multiple change points detection in linear regression by Filtered Derivative and False Discovery Rate method

In this third part of this chapter, we deal this new method with the simple linear regression. This work is published by International Statistical Institute(ISI,2015) at Rio, Brazil of 25 July until 31 July 2015. The reference of this proceeding is "Proceedings of the 60th ISI World Statistics , 26–31 July 2015, Rio de Janeiro, Brazil, p 0–5".

Abstract

Linear models are widely used in statistics to describe between two variables : X the explanatory variable and Y the response variable. In this work, we consider the simple linear regression model with change on the parameters slope and intercept. We use the Filtered Derivative and False Discovery Rate method (FDqV) for estimating the coefficients of linear regression. We indicate that it exists a previous work made for estimating the parameter slope by using the Filtered Derivative with p-value (FDpV). We specify in this work, that the estimating of the slope done by FDpV is wrong and we give the correct estimation.

2.3.1 Introduction

We study in this paper the problem of change point analysis in the case of simple linear regression. For an updated overview, the reader can see the book [6] or the article [4]. The goal is to detect or to estimate the instant of abrupt changes and the parameter (slope and intercept) corresponding of linear regression. Others methods for the problem of change points exist such that the Penalized Least Square Error (PLS), the Hierarchic Binary Splitting (HBS), but the both-methods are expensive in times of calculation. The PLS-method has the time and memory complexity of order $\mathcal{O}(N^2)$. Recently, an other method called the filtered derivative with p-value (FDpV) [11] is introduced for the detection of

change point analysis, this method for the detection of slope in linear regression is wrong, because the filtered derivative considered in [11] is a hat-function which presents an each instant of change point a maximum and this is not the case. We provide in this work the exactly order to detect of slope parameter in linear regression. In our method, the time and memory complexity are the order $\mathcal{O}(N)$. By consequent, the (FDqV)-method is more advantageous not only the criteria (time and memory complexity) but also the FDqV-method realize the best results for the detection of abrupt changes and so the estimation of corresponding parameters (for example the parameter mean) if we base ourselves on the criteria (Mean Integrate Square Error (MISE), Number of False Alarms (NFA) and Number of No-Detection(NND)). The dominance of FDqV than FDpV is well treated in [35] or [36]. The rest of this paper is structured as following : The section 1 describes the description of change points in simple linear regression, the section 2 recall back the FDqV-method for the mean. The section 3 apply the FDqV-method in the coefficients of linear regression.

2.3.2 Description of the problem

In this section, we describe the problem of change analysis and we study the change in the coefficients of linear regression. In fact, there are two types change points detection in linear regression : The model with a discontinuous change points and the model with a continuous change points.

Model discontinuous for change points in linear regression.

The model with a discontinuous change points is defined as :
Let (X_i, Y_i), $i = 1, 2 \ldots, n$, the observations where each Y_i describes the response of the explanatory variable X_i. A simple linear regression with no change is defined as $Y_t = aX_t + b + \varepsilon_t$, for $t = 1, 2, \ldots, n$, and a and b are the slope and the intercept of linear regression. Here, we suppose that the parameters a and b change. Then , we have : $\mathbf{X} = (X_1, X_2 \ldots, X_n)$ is a family of independent random variables indexed by the time

t=$1, 2, \ldots, n$. It exists a segmentation $\tau = (\tau_1, \tau_2, \ldots, \tau_K)$ with $\tau_K \in \{1, 2, \ldots, n\}$ and $0 < \tau_1 < \tau_2 < \ldots < \tau_K < n$. K denotes the number of changes. By convention, we set $\tau_o = 1$ and $\tau_{K+1} = n$, thus K can be equal to zero (this means no change) or any integer smaller than n. Thus

$$Y_t = \begin{cases} a_o \times X_t + b_o + \varepsilon_t, \ for\ t = 1, \ldots, \tau_1 \\ a_1 \times X_t + b_o + \varepsilon_t, \ for\ t = \tau_1 + 1, \ldots, \tau_2 \\ \vdots \\ a_K \times X_t + b_K + \varepsilon_t, \ for\ t = \tau_K + 1, \ldots, n \end{cases}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ denote the Gaussian law with mean 0 and standard deviation 1. For simplicity of the presentation, we have assumed that the variance σ^2 remains constant.

Model continuous for change points in linear regression.

The model continuous for change points in linear regression is defined as the discontinuous case but an each change point τ_k, for t=$1, \ldots, \tau_K$ there are a continuity constraint. For

more details, we have :

$$Y_t = \begin{cases} a_o \times X_t + b_o + \varepsilon_t, \ for\ t = 1, \ldots, \tau_1 \\ a_1 \times X_t + b_1 + \varepsilon_t, \ for\ t = \tau_1 + 1, \ldots, \tau_2 \\ \vdots \\ a_K \times X_t + b_K + \varepsilon_t, \ for\ t = \tau_K + 1, \ldots, n \end{cases}$$

With

$$\begin{cases} a_o \times X_t + b_o + \varepsilon_t = a_1 \times X_t + b_1 + \varepsilon_t, \ for\ (\tau_o, Y_{\tau_o}). \\ a_1 \times X_t + b_1 + \varepsilon_t = a_2 \times X_t + b_2 + \varepsilon_t, \ for\ (\tau_1, Y_{\tau_1}). \\ \vdots \\ a_K \times X_t + b_K + \varepsilon_t = a_n \times X_t + b_n + \varepsilon_t, \ for\ (\tau_K, Y_{\tau_K}). \end{cases}$$

For an illustration, we draw the following figures.

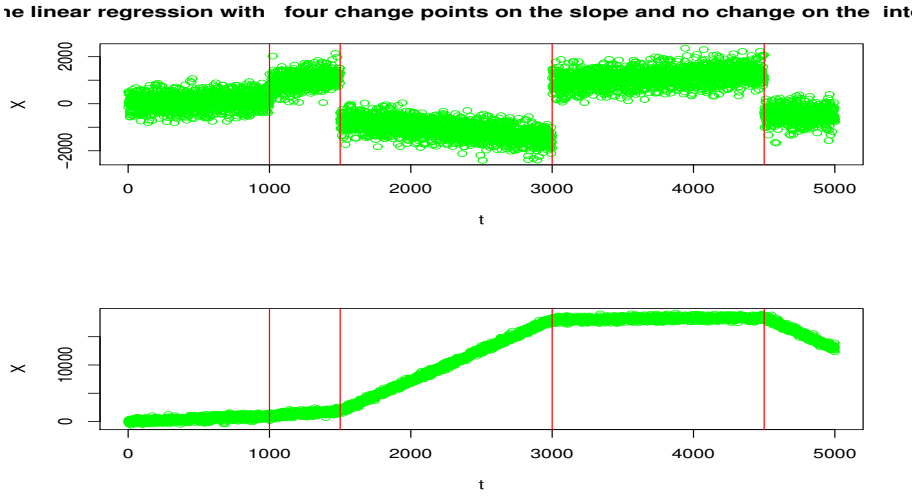


FIGURE 2.15 – First drawing : the linear regression with model discontinuous change points. Second drawing : the linear regression with model continuous change points

2.3.3 Recall the Filtered Derivative and False Discovery Rate Method (FDqV)

The Filtered Derivative and False Discovery Rate (FDqV) [35] is a method derived the Filtered Derivative. For reading of this section, see the article [36].

2.3.4 Detection the parameters of linear regression by FDqV-method

Our goal is to detect the abrupt changes and the estimation of both parameters slope and intercept each box.

Simulation : change on the slope

For $n = 5,000$, we have simulated a simple linear regression of the random variables $(X_t, Y_t),\ t = 1, 2, \ldots, n$ with four change points $\tau = (1000, 1500, 3000, 4500)$ with the cor-

responding slopes $a_t = (0.2, 0.8, -0.5, 0.3, -0.1)$ (here, the intercept remains constant).

The FDqV-method

We want to estimate the instant $(\tau_1, \tau_2 \ldots, \tau_K)$ and a_k an each box $[\tau_k + 1; \tau_{k+1}]$ for $1 < k < K$.

Step 1 : The Filtered Derivative (FD) for the slope

The FD for the slope is defined as :

$$FD(A,k) = \widehat{a}(k,A) - \widehat{a}(k-A,A) \tag{2.3.1}$$

where

$$\widehat{a}(k,A) = [A \times \sum_{t=k+1}^{k+A} X_t Y_t - \sum_{t=k+1}^{k+A} X_t \sum_{t=k+1}^{k+A} Y_t][A \times \sum_{k+1}^{k+A} X_t^2 - (\sum_{t=k+1}^{k+A} X_t)^2]^{-1} \tag{2.3.2}$$

is the estimator of the slope on the box [k+1,k+A] obtained by the least square method. We give below a lemma which allow us to well define the FD function.
lemma.

Let $Y_j = a_k \times X_j + b_k + e_j$ *for* $\tau_{k+1} - 1 \leq k \leq \tau_k$, $1 \leq j < n$ and e_j is the error of Gaussian law of mean zero and standard deviation σ. Then

$$\widehat{a}_k \sim \mathcal{N}(a_k, \sigma_{a_k}^2) \; with \; \frac{\sigma_{a_k}^2}{\sigma^2} = [\sum_{j=\tau_{k+1}-1}^{\tau_k} (X_j - \overline{X_k})^2]^{-1}$$

proof.
The proof is clear by using the hypothesis that e_j is a Gaussian law.

By definition, the Filtered Derivative function (FD) is mathematically defined in terms of difference between the estimators of the slope computed in two sliding windows respectively at the right and at the left of the time k, both of size A. We can re-define the FD using the lemma (2.3.4) as the difference between two Gaussian functions respectively at the right and the left of the time k, both of size A. Then we have exactly the definition of the Derivative Gaussian.

How to detect the potential change points.

In the figure (2.16), we clearly see that the Gaussian Derivative (GD) change sign on each box $[\tau_k - A, \tau_k + A]$ *for* $k = 1, \ldots, K$, so for the localization of abrupt changes, we use the following algorithm.

algorithm.

We choose a threshold C_1 and $Kmax$ corresponding the number of maximum of change points.

Step 1. We calculate the maximum of GD function and the argument of maximum and we set for $k = 1$,

$$\widehat{\tau}_1 = \frac{arg\max_{k\in[A;n-A]} GD(k) + arg\min_{k\in[A;n-A]} GD(k)}{2}$$

and on $[\widehat{\tau}_1 - A; \widehat{\tau}_1 + A]$ we put $GD(k) = 0$.

Step 2. While $(k < Kmax)$ and $(Cmax > C_1)$ do
k=k+1

$$\widehat{\tau}_k = \frac{arg\max_{k\in[A;n-A]} GD(k) + arg\min_{k\in[A;n-A]} GD(k)}{2}$$

and on $[\widehat{\tau}_k - A; \widehat{\tau}_k + A]$, we set $GD(k) = 0$

Step 3. We sort out in order increasing the instant of potential change points.

Thus, we Keep the instant of potential change points $\widehat{\tau}_1 < \widehat{\tau}_2 < \ldots < \widehat{\tau}_{\widehat{K}}$ with $K < \widehat{K}$. Below, we draw the Gaussian Derivative with noise and without noise.

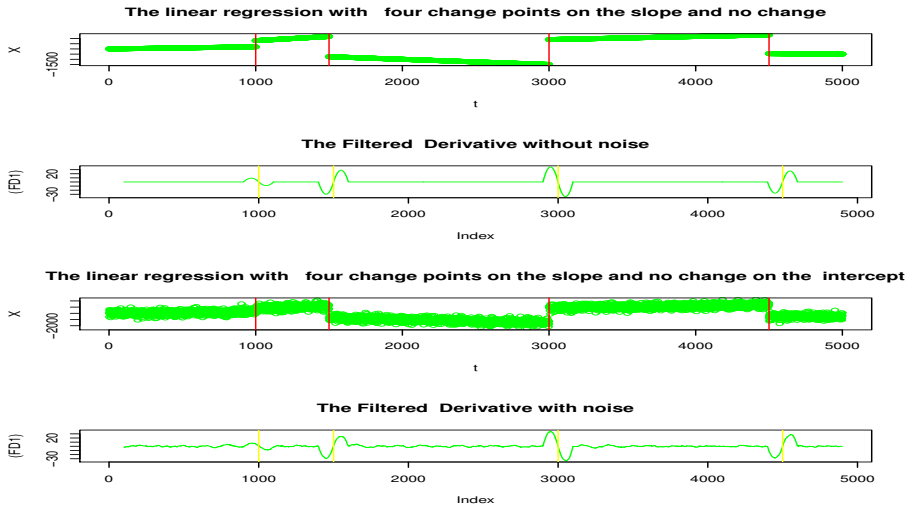


FIGURE 2.16 – First drawing : The observed signal, the third drawing : The Filtered derivative function for the slope without noise, the fourth drawing : The Filtered Derivative with noise.

Remark

We specify that the manner done in [11] for detection of change points in linear regression for the parameter slope is wrong. In fact, they use in step 1, a hat-function for the localization of change points. We realize in this work that the filtered derivative function is a GD and not a hat-function, see figure (2.16).

Step 2 : The False Discovery Rate

A the end of step 1, we have the instant of potential change points. Among these points, there are the right points and also the false detections. The false discovery rate allow us to separate these two kind points and keep the false detections at level close to zero. For this, we proceed in the same way as for the parameter mean describes in the section 2. Thus, we obtain the estimated instants $(\tau_1^* < \tau_2^* < \ldots < \tau_{K^*}^*)$ with $K \leq K^*$. Also, an each time detected, we can estimate the corresponding slopes a_k, for $k = 1; 2; \ldots; K^*$.

The FDqV-method for the intercept

In subsection, we want to apply the FDqV-method, the detection of abrupt changes and the estimation of the intercept on each box. The FD for the intercept is :

$$FD(A,k) = \widehat{b}(k,A) - \widehat{b}(k-A,A)$$

Where

$$\widehat{b}(k,A) = \frac{1}{A}\sum_{j=k+1}^{k+A} X_j - a \times \frac{1}{A}\sum_{k+1}^{k+A} X_j$$

Simulation : Change on the intercept

For $n = 5,000$, we have simulated a simple linear regression of the random variables (X_t, Y_t), $t = 1, 2, \ldots, n$ with four change points $\tau = (1000, 1500, 3000, 4500)$ on the intercept $b_t = (100, 500, 2000, -1000, 1500)$ (here, the slope remains constant). Below, we have the corresponding figure of this simulation and the filtered derivative.

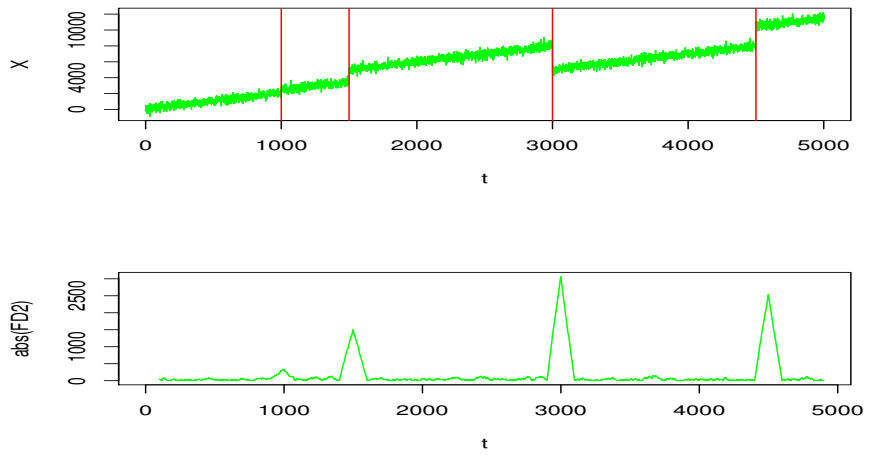


FIGURE 2.17 – First drawing : The observed signal, second drawing : The Filtered Derivative function for the intercept.

We see through this above figure that the filtered derivative is a hat-function, so in the first time we select the first instant $\widehat{\tau}_1$ of abrupt changes as the maximum value of the filtered derivative such that $|FD| > C_1$, where C_1 is the threshold chosen. We put around of the point $\widehat{\tau}_1$, $FD[\widehat{\tau}_1 - A, \widehat{\tau}_1 + A] = 0$, where A is the window size. We begin again the same procedure for a new function FD, Thus we keep the second estimated point τ_2. We apply the above procedure, while $(k < Kmax)$, where the $Kmax$ is the maximum number change points. Finally, we have the estimated instants $(\widehat{\tau}_1; \widehat{\tau}_2; \ldots \widehat{\tau}_{\widehat{K}})$. For more details see [35]. At the step 2, For eliminating the false alarms or having the number of false alarms at level close to zero, we did $\widehat{K}$ hypothesis where the null hypothesis $H_{(o,k)}$: $\widehat{b}_k = \widehat{b}_{k+1}$ versus the alternative hypothesis $H_{(1,k)}$: $\widehat{b}_k \neq \widehat{b}_{k+1}$ with $k = 1; 2; \ldots; \widehat{K}$. When, we did all hypothesis tests, we have the p-value $\widehat{p}_1; \widehat{p}_2; \ldots; \widehat{p}_{\widehat{K}}$. In the sequel, we apply the Benjamini and Hochberg's procedure [7], so we only keep the p-values $(p_1; p_2; \ldots; p_{K^*})$, with $K \leq K^*$. At the end of this procedure, we have the estimated instants and the corresponding intercept b_k for $k = (1; 2; \ldots; K^*)$. For more details see [35, 36].

2.3.5 Conclusions

In this work, we have given the estimating of the coefficients of linear regression. We corrected the mistake done in [11] concerning the estimating of the slope. It is logical to do the comparison of existing methods in the literature such the Penalized Square Error or the Hierarchic Binary Split. We already affirm that the FDqV-method is advantageous on times and memory complexity see [35, 36]. We have already done a real application of this method concerning the wind turbines for the parameter mean [36].

Chapitre 3

The parameters optimization of Filtered Derivative for change points analysis

In this chapter, we improve the new method "Filtered Derivative and False Discovery Rate Method". In fact, this method depends two parameters : the window A and the threshold C. We deal in this chapter the improvement of these parameters.
This article has published in "International Journal of Statistics and Probability, vol 3, n 3, P 29–43, 2014" doi :10.5539/ijsp.v3n3p29, URL : htpp//dx.doi.org/ijsp.v3n3p29.

Abstract

Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ be a time series. That is a sequence of random variable indexed by the time $t = (1, 2, \ldots, N)$, we suppose that the parameters of $\mathbf{X}$ are piecewise constant. In other words, it exists a subdivision $\tau = (\tau_1 < \tau_2 < \ldots < \tau_K)$ such that X_i is a family of independent and identically distributed (i.i.d) random variables for i $\in (\tau_k, \tau_{k+1}]$, and k= $0, 1, \ldots, K$ where by convention $\tau_o = 0$ and $\tau_{K+1} = N$. The preceding works such that [10] control the probability of false alarms for minimizing the probability of type I error of change point analysis. The novelty in this work is to control the number of false alarms. We give an bound of number of false alarms and the necessary condition for number of no detection. In other hand, we know the filtered derivative [6] depends the parameters such that the threshold and the window, we give in order to choose the optimal parameters. We compare the results of Filtered Derivative optimized parameters and the Penalized Square Error methods in particulary the adaptive method of [57] .

3.1 Introduction

The problem of change detection is much studied in the literature, it exists two types of change points detection : The on-line detection or sequential points analysis and the off-line detection or change points analysis. For an updated overview, we can see the textbooks [6, 28], or [45, 53]. Many applications use the change points analysis such as health, medicine and civil engineering and the sequential analysis such as fault detection, finance, surveillance and security system. Many methods exists but we often use : The penalized least square error (PLS) [56] and the filtered derivative (FD) [6]. The calculus of PLS need a matrix of size $\mathcal{O}(n^2)$, and that FD is of order $\mathcal{O}(n)$. To improve the FD-

method, two methods are developed : The filtered derivative with p-value (FDpV) [11] and the filtered derivative and false discovery rate (FDqV) [35]. For the PLS-method, many authors are proposed [56, 57] the choices of penalized parameter for performing the PLS-method. For FD-method, there were no papers which mentioned this. Recall, the algorithm FD depends the window and the threshold and by consequent his performance depends the optimization of these parameters. In this work, we give the reasonable choices these parameters. We organised our paper in this way : section 1 is the introduction, section 2 describes the art of change points detection and the criteria of measure . Section 3 recall the methods of change point analysis. In the section 4, we discuss how to control the number of false alarms and numbers of no-detections. The section 5 contains numerical comparison of FD and PLS adaptive methods. Finally the appendix contains all proofs, propositions and lemmas used in this work.

3.2 The art of change points detection

The following subsection describes the problem of abrupt changes and different criteria used in the literature.

3.2.1 Problem of change points detection : the model

- $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ is a family of independent random variables indexed by the time.
- There exists a subdivision $\tau = (\tau_1, \ldots, \tau_K)$ with $\tau_k \in \{1, \ldots, N\}$ and $0 < \tau_1 < \tau_2 \ldots < \tau_K < N$.
- A configuration a K change points $\tau = (\tau_1, \ldots, \tau_K)$ enlarged by convention by adding $\tau_0 = 0$ and $\tau_{K+1} = N$.
- Associated to the configuration of mean values $(\mu_0, \ldots, \mu_K)$ with $X_t \sim \mathcal{N}(\mu_k, 1)$, for $t \in (\tau_k,\ \tau_{k+1}]$ and for all $k = 0, \ldots, K$.
- For notational convenience, we also define the configuration of shifts, for $k = 1, \ldots, K$, $(\delta_1, \ldots, \delta_K)$ where $\delta_k = \mu_{k+1} - \mu_k$.
- Let us define the minimal distance between to consecutive change points by $L_0 = \inf\{|\tau_{k+1} - \tau_k|\ k = 0, \ldots, K\}$,
- and the minimal absolute value of the shifts by $\delta_0 = \inf\{|\delta_k|,\ k = 1, \ldots, K\}$.

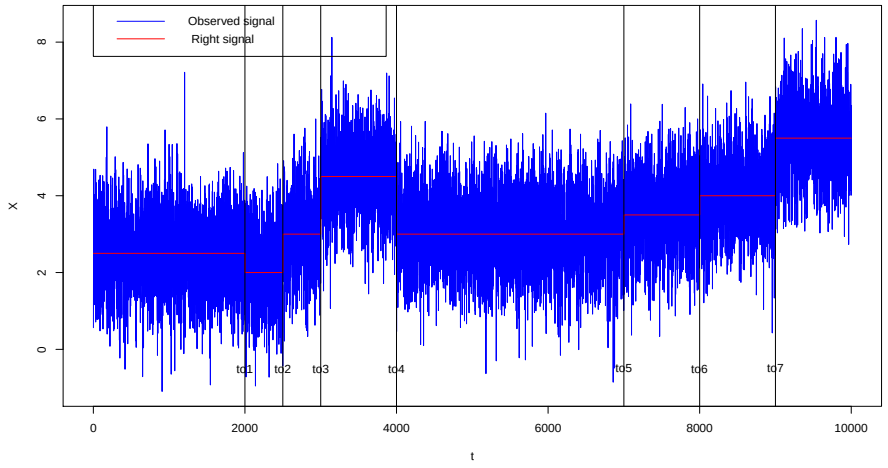
Let us also recall the definition of the cumulative distribution function for standard Gaussian law

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp^{\frac{-u^2}{2}} du \qquad \text{and} \qquad \Psi(x) = 1 - \Phi(x). \tag{3.2.1}$$

All this paper, we use a following simulation :

3.2.2 Simulation

For n=10000, we have done one realization of a sequence of Gaussian random variable $(X_1, \ldots, X_n)$ with variance $\sigma^2 = 1$ constant and mean μ have different values. we consider seven change points at time



$\tau = (2000, 2500, 3000, 4000, 7000, 8000, 9000)$ with means $\mu = (2.5, 2, 3, 4.5, 3, 3.5, 4, 5)$ and $\delta = (0.5, 1, 1.5, 1.5, 1.5, 0.5, 1)$. Below, we give a drawing for change points analysis.

We will use this model all this paper.

3.2.3 The criteria of measure

We suppose that the number K is unknown, and the goal of the off-line detection is to estimate the instants $\tau = (\tau_1, \ldots, \tau_K)$ and the values of the mean $(\mu_0, \mu_1, \ldots, \mu_K)$. we set $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\widehat{K}})$ and $(\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{\widehat{K}})$ the corresponding estimated.

Criterion

For measuring the quality of the estimation of parameters, we use the integrated square error (ISE). So we define the ISE as :

$$ISE = \sum_{i=1}^{N} \left\{ \left[\widehat{l}(t) - l(t) \right]^2 \right\}$$

with the signal

$$l(t) = \sum_{k=0}^{K} \mu_k \times \mathbf{1}_{(\tau_k, \tau_{k+1}]}(t)$$

and the estimated signal

$$\widehat{l}(t) = \sum_{k=0}^{\hat{K}} \widehat{\mu}_k \times \mathbf{1}_{(\widehat{\tau}_k, \widehat{\tau}_{k+1}]}(t)$$

As, a result for one replication is not significant, we make $M = 1,000$ replications and thus we use the mean integrated square error (MISE).

3.3 Methods of off-line detection

The most popular methods used are : The Penalized Least Square Error (PLS) [56], the Filtered Derivative with p-value (FDpV) [11] and the Filtered Derivative with False Dis-

covery Rate (FDqV) [35]. We make drawings with following methods using the simulation of the subsection (1.2).

3.3.1 PLS-method

For PLS-method, we have to search the instants of change points which minimize the contrast function defined as

$$\mathrm{Q}(\tau) = \sum_{k=0}^{K} \sum_{t=\tau_k+1}^{\tau_{k+1}} |X_t - \widehat{\mu}_k|^2 \tag{3.3.1}$$

Two cases are studied :

- The case where K is know [4] use the dynamical program method for estimating the instants of ruptures and the mean values corresponding.
- In the case where K is unknown, many authors as that[56, 57] or [13] proposed different values of penalized parameter for the performance of this method.
  In [56], the proposed choice of penalized parameter is :

  $$\beta_1 = \frac{2\sigma^2(logn)}{n}.$$

  The inconvenient is to over-estimate the number of change points.

  In [13], we have

  $$\beta_2 = \frac{\sigma^2}{n} \times \left[2 + 5 \times \log(\frac{n}{K})\right]$$

  We can only apply the times series with constant variance.
  [57] give an adaptive-method for estimating the number in following manner
  - For $1 < K < K_{max}$, we adjust the model $f(K) = a \times K + b \times log(K) + e_K$ with the contrast function and e_K a sequence independent of random variable following the gaussian law standard.
  - We evaluate the probability that $\mathrm{Q}(K)$ follows this model.
  - The estimated number of change points will be the highest value of K such that the corresponding p-value is the smaller value of a given threshold.

  For more details see [57]

For an illustration, we draw the following figure.

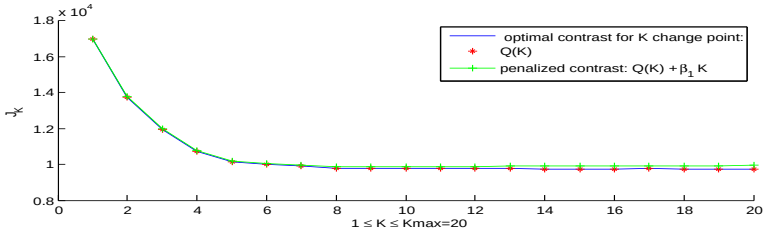


FIGURE 3.1 – *Blue with red crosses : the contrast function* $\mathrm{Q}(\widehat{\tau}^K)$ *; green : the penalized contrasted* $pen(K)$.

3.3.2 FDpV-method

The FDpV-method has two steps :

1. The step 1 is based for detection of potential changes points. For this we use the Filtered Derivative (Basseville M., & Nikirov, I. 1993) define as it follows :

$$FD(t, A) = \widehat{\mu}(t+1, t+A) - \widehat{\mu}(t-A, t), \ for \ A < t < N - A \tag{3.3.2}$$

where

$$\widehat{\mu}(t+1, t+A) := A^{-1} \sum_{j=t+1}^{t+A} X_j$$

is the classical empirical mean.

In this case Without noise, the function $j \to FD(j, A)$ presents a hat centered at $\tau = j$ that is the top of the hat corresponding at the right change point. The hight of the hat is exactly the size of the change on the mean, and the spread is 2A. When the signal is random, the true μ at the right (resp left) on $(\tau - A, \tau + A)$ is replaced by $\widehat{\mu}$ on $(\tau - A, \tau + A)$. The estimate mean $\widehat{\mu}$ is fluctuating around μ. In order to reduce the noise due to the sampling fluctuation, we filters the signal by replacing the true value mean at the right and the left at the point j by its estimated at the right and the left at point j on $(\tau - A, \tau + A)$, and we take the difference of these two quantities.

For detection change points, [6, 8] choose a threshold C_1 and keep only the instants $\widehat{\tau}_k$ for $k = 1, \ldots, \widehat{K}$, where $\widehat{K}$ is the length of potential change points and such that $\max |FD(\widehat{\tau}_k)|$ exceed the threshold C_1. At the end of the calculus, we have the instant of potential change points $\widehat{\tau}_k$ for $k = 1, \ldots, \widehat{K}$. We have $K < \widehat{K}$. We draw below a figure for illustration.

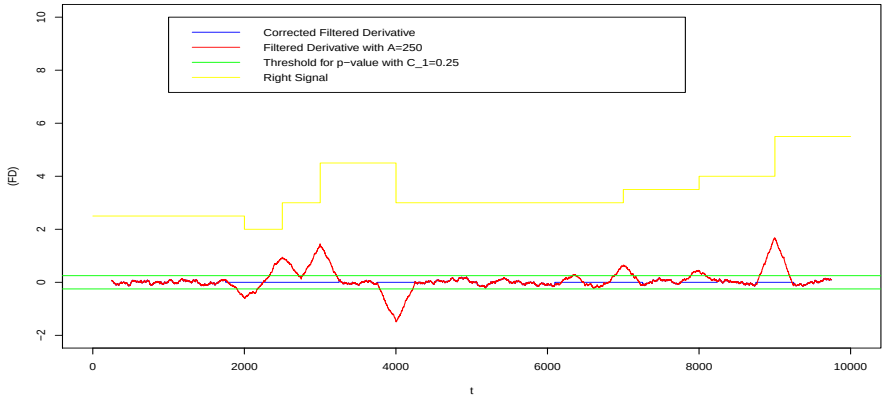


FIGURE 3.2 – *The right signal (red), the noisy signal (blue), and Filtered Derivative function (green).*

Remark 1

A natural question is coming on : Does it exist in order to choose the optimal parameters of filtered derivative A and C_1 ? The goal of this work is to give a response of this question.

2. Recently, [11] remark it exist false alarms at the end of the step 1, so for keeping only as possible the right change points, they have had an idea to add a second step. For this, they have compared pairwisely the means estimated $\widehat{\mu}_{k-1} := mean(X, \tau^*_{k-1}, \tau_k*)$ versus $\widehat{\mu}_k := mean(X, \tau^*_k, \tau^*_{k+1})$. In other words they have done a test hypothesis where :

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

In the sequel, they have calculated the p-value corresponding to each potential change points and they have chosen an critical p-value p^* for keeping only the p-value lesser than the critical p-value.

Remark 2

In [11], the critical p-value chosen is arbitrary ($p^* = 10^{-6}$), so we can say that the problem of optimal p-value is open and we will try to do in future work.

3.3.3 FDqV-method

In [35], we have proposed a method for change points detection, this method use also the filtered derivative as step 1, but we have added a step 2, which allow us to detect as possible all change points right. The difference between the FDpV and FDqV is : The first use a single hypothesis for keeping all change points right and the second use a multiple test. The power of FDqV is established in [35].
The algorithm of the step 2 of FDqV is :

- We put the p- values in this way $p^*_1 \leq \ldots \leq p^*_{K^*}$.
- We choose a critic value denote q^*.
- We eliminate all p-value such that $p^*_i > \frac{i}{K^*} q^*$.

At the end of the step 2, we obtain $(p^*_1, p^*_2, \ldots, p_{K^*})$ and the estimated instants $(\tau_1{}^*, \tau_2{}^*, \ldots, \tau_{K^*})$, with $K < K^*$.

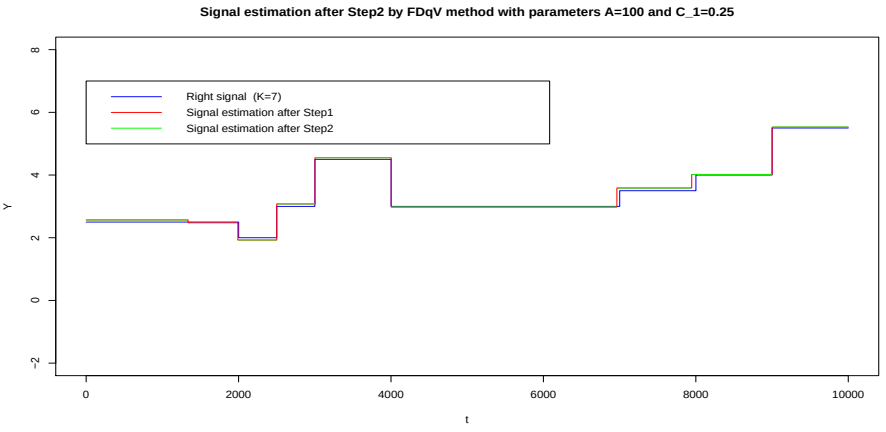


FIGURE 3.3 – *Signal reconstruction after Step2 by FDqV method*

Now, we start the main part of this article.

3.4 The choice of parameters for Filtered Derivative method

All change point method depends on extra-parameters, which have to be well chosen. The PLS method depends only on the penalization parameter β, different choices are possible see Section 3 above. The filtered derivative method depends on the parameters, namely the window size A and the threshold C_1. Both FDpV and FDqV method use filtered derivative as Step 1, so they depends on the same extra-parameters A and C_1. Moreover, FDpV and FDqV method add a step 2, which depends on another extra-parameter, that is the critical p-value p^* or the q-value q^*. In Subsection, 3.4 we discuss the different criterium. In Subsection, 3.4 we give the bound of the type II error.

Choice the extra-parameters of FD

As pointed out in Subsection, 3.2.3 the quality of a change point method can be evaluated by two criteria : i) the absolute value of the number of estimated change point minus the right number of change points $|\widehat{K} - K|$; ii) ISE or MISE. Both criterions lead to prefer detecting more potential change point than missing at least one. Indeed, the no detection of one change point could greatly impact the mean values $\widehat{\mu}_k$'s and after the ISE, but also the p-value $p_k^\star$'s. Stress that this phenomenon does not more exist when we restrict ourselves to FD method with the number of change as criterion.

So, the type II error or probability of no detection (PND) should be controlled at a level close to zero. However, the previous remark address to detect the right change point at the right times. As pointed out in (Bertrand, P. 2000), when there is more no detection, we have : For each right change point τ_k, we define the local PND as

$$\beta_{loc}(\tau_k) = \mathbf{P}\left(B_k\right) \quad \text{where} \quad B_k = \Big\{\forall k \in [\tau_k - A, \tau_k + A], |D(A,k)| < C_1\Big\}.$$

Then with these notations, we can write the global PND in this manner

$$PND_{global} \quad = \quad \mathbf{P}\Big(\bigcup_{k=1}^{K} B_k\Big). \tag{3.4.1}$$

On the other hand, we define the probability of false alarm or probability of type I error as following :

$$\alpha(A, C_1) = \mathbb{P}(\tau(C_1, A) \leq N - A)$$

Where $\tau(A, C_1)$ is the first hitting time of C_1 and

$$\tau(A, C_1) := \inf\{k \geq A \; such \; that \; FD(A,k) \geq C_1\} \tag{3.4.2}$$

However, the type I error is the probability of at least one false alarm and thus appears as a rough criterion see [10]

The type I and II errors at Step 1 (Filtered Derivative)

In the following proposition, we give an upper bound for PND_{global}

Proposition 1 Assume there are K change points and a configuration of change points $\tau = (\tau_1, \tau_2, \ldots, \tau_K)$, with means $(\mu_0, \ldots, \mu_K)$ and shifts $(\delta_1, \ldots, \delta_K)$ as described in Subsection 3.2.1. Then

$$PND_{global} \quad \leq \quad K \times \beta^*(C_1, A).$$

where PND_{global} is defined by (3.4.1) and

$$\beta^*(C_1, A) := \Psi\left(\frac{\delta - C_1}{\sigma}\sqrt{\frac{A}{2}}\right) \times \Phi\left(\frac{C_1 - \delta/3}{\sigma}\sqrt{\frac{A}{2}}\right)^2. \tag{3.4.3}$$

and Φ and Ψ are given by (3.2.1).

Proof. Following (Prop. 3. 2, p 222, [10]), we have, for each change point τ_k,

$$\mathbf{P}\left(B_k\right) \leq \Psi\left(\frac{\delta_k - C_1}{\sigma}\sqrt{\frac{A}{2}}\right) \times \Phi\left(\frac{C_1 - \delta_k/3}{\sigma}\sqrt{\frac{A}{2}}\right)^2. \tag{3.4.4}$$

Next, by remarking that the right side of (3.4.4) is a decreasing function of δ_k and setting $\delta = \inf_{k=1,\dots,K} \delta_k$, we can deduce that

$$\mathbf{P}\left(B_k\right) \leq \beta^*(C_1, A) := \Psi\left(\frac{\delta - C_1}{\sigma}\sqrt{\frac{A}{2}}\right) \times \Phi\left(\frac{C_1 - \delta/3}{\sigma}\sqrt{\frac{A}{2}}\right)^2. \tag{3.4.5}$$

By consequent, we obviously obtain

$$PND_{global} \leq \sum_{k=1}^{K} \mathbf{P}\Big(B_k\Big)$$

which combined with (3.4.5) gives us the bound (3.4.3). This finishes the proof of Proposition 1. K is unknown, but is not variable. Thus, we will monitor the quantities $\beta^*(C_1, A)$, for instance we choose to set $\beta^*(C_1, A) = 10^{-4}$ and we can write $\ln \beta^*(C_1, A) = f\left(\frac{C_1}{\delta}, \frac{\delta}{\sigma}\sqrt{\frac{A}{2}}\right)$. Thus after the variables change we have $f(x, y) = ln(10^{-4})$ with $x = \frac{C_1}{\delta}$ and $y = \frac{\delta}{\sigma}\sqrt{\frac{A}{2}}$ and $f(x, y) = \ln \Psi((1 - x) \times y) + 2 \ln \Phi((x - 1/3) \times y)$. This equation can be numerically solved, and we find couples solution of this equation. Since the map $C_1 \longmapsto \beta^*(C_1, A)$ is decreasing, and we find an implicit function $A \mapsto C_1(A)$. After having controlled the PND, we can control the PFA. We know (Prop. 3. 1, p 221, [10]) that for all $\varepsilon > 0$ there exists a constant M_ε such that

$$\alpha \leq M_\varepsilon \times \alpha^*(C_1, A) := M_\varepsilon \times \left(\frac{n - A}{A}\right) \times \Psi\left(\frac{C_1}{\sigma}\sqrt{\frac{A}{2 + \varepsilon}}\right). \tag{3.4.6}$$

For instance, we can set $\varepsilon = 0.1$, next we plug the implicit relationship between A and C_1 inside (3.4.6)and we obtain a function $A \longmapsto \alpha^*(C_1(A), A)$. The first idea is to make varying the parameter A in order to find the optimal value corresponding to a minimum of the map $A \longmapsto \alpha^*(C_1(A), A)$. Unfortunately, the map $A \longmapsto \alpha^*(C_1(A), A)$ is decreasing and reaches no minimum value.

3.4.1 Necessary condition of no-detection

In this subsection, we draw three figures for choosing a "good" window A. According to the figure below, we can choose A with the following condition :

$$2 \times A < L_0 := \inf\{|\tau_{k+1} - \tau_k|, k = 1, \dots, K\}. \tag{3.4.7}$$

With this drawings, we can say that A must verify $A < L_0/2$, because in the first, we detect all change points and others, we only detect two change points.

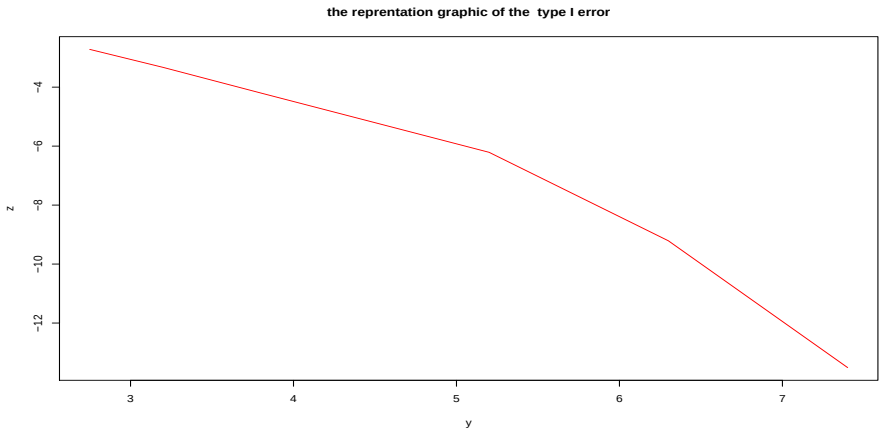


FIGURE 3.4 – *The graphic corresponding at the type I error , $y = \sqrt{\frac{A}{2}}$ and $z = \alpha^*(C_1(A), A)$.*

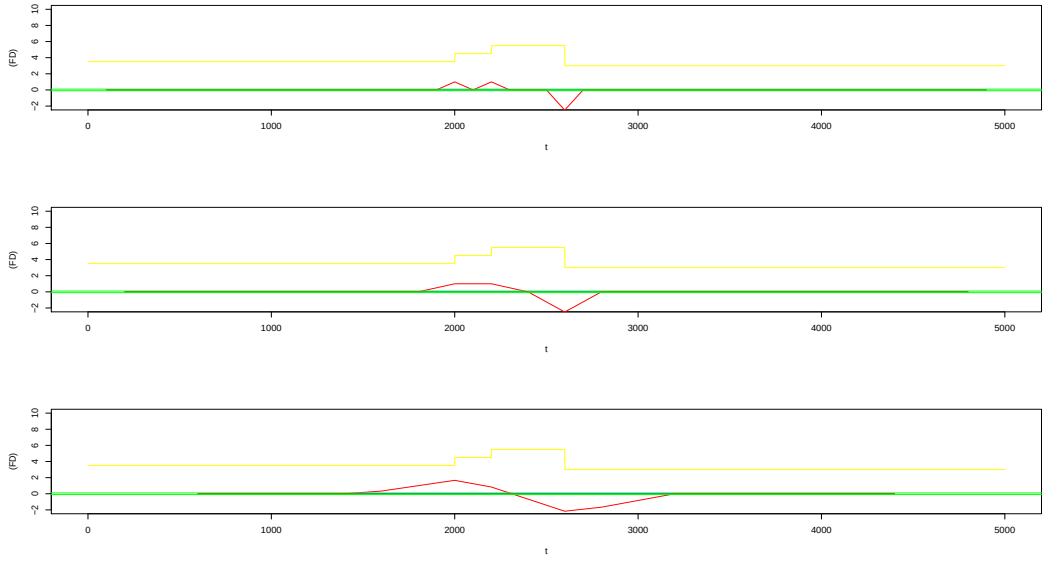


FIGURE 3.5 – *Red : the filtered derivative with A=100, A=200, A=600, Yellow : the right signal, Green : the threshold $C_1 = 0.1$.*

3.4.2 Control of number of false alarms

In this subsection, we want to control the number of false alarms (NFA) and not only the PFA (probability of false alarms). First, we can remark that the number of false alarm is always greater than the corresponding one when there is no change. Indeed, let us denote by $\widetilde{K}$ the number of change points select in step 1 (FD), then the number of false alarms is $(\widetilde{K} - K)$. By using [10], we have

$$FD(A,t) \quad = \quad \Gamma(A,t) + \sum_{k=1}^{K} \delta_k \times g\left(\frac{t-\tau_k}{A}\right) \qquad for\ all\ t$$

where

$$\Gamma(A,t) = A^{-1}[\widehat{S}_{t+A} + \widehat{S}_{t-A} - 2 \times \widehat{S}_t],\ and\ \widehat{S}_t = \sum_{k=1}^{t} X_k$$

and

$$g(x) \quad = \quad \begin{cases} 1 - |x| & : \quad when\ |x| \leq 1 \\ 0 & : \quad when\ |x| \geq 1 \end{cases}$$

Let us point that when there is no change, then $FD(A,t) = \Gamma(A,t)$ for all t, this implies that $(\widetilde{K} - K) \leq \widetilde{K}_0$, where $\widetilde{K}_0$ denotes the number of change points detected by FD when there is no change. For example, using the simulation the subsection(1.2), we can see that $\widetilde{K}_0 = 3$ (see drawings below and count the number τ^*). Thus, we can restrict ourselves

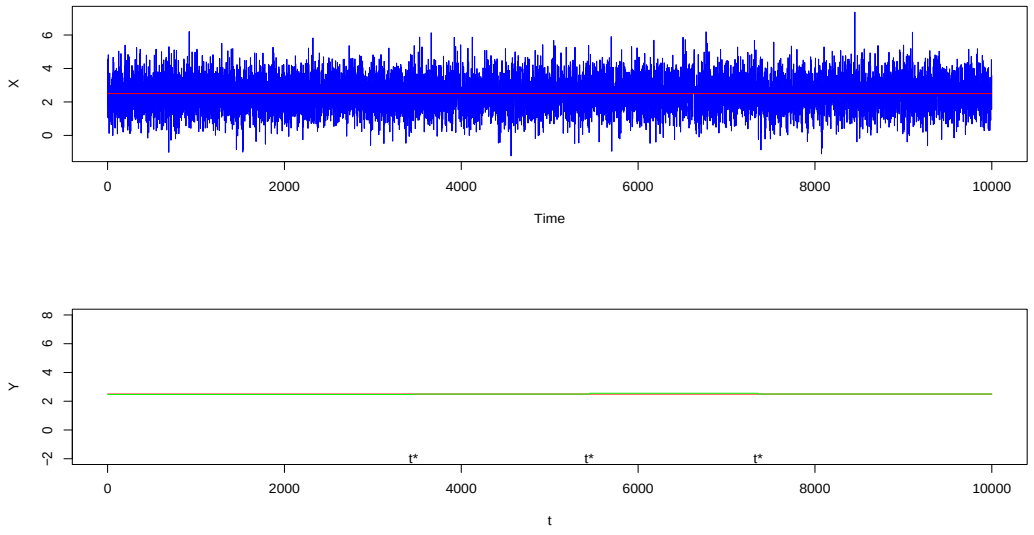


FIGURE 3.6 – *First drawing : The signal observed : blue, The right signal : red. Second drawing : The signal reconstruction : green.*

for estimation the number of false alarms to the case without change. In this case there are only false detection, and we denote by RL_k the real variable corresponding to the run

length before the kth false detection. For $k \geq 1$, we have $\widetilde{\tau}_k = \sum_{j=1}^{k} RL_j(\omega)$. Next, we denote by $M(\omega)$ the number of false alarms change points. With these notation, we can state the following theorem that allows us to give a bound of number of false alarms :

Theorem 1
Assume there is no change, then For all integer $l \leq N$

$$\mathbb{P}(M(\omega) \leq l-1) \quad \leq \quad \varphi(A, C_1, N, l) \tag{3.4.8}$$

where

$$\varphi(A, C_1, N, l) := \sum_{(L_1,\ldots,l_j),\sum l_j > N} \prod_{i=1}^{l} |l_j - 2A| \times \Psi(\frac{AC_1}{\sigma\sqrt{l_j}})$$

and $M(\omega)$ is defined as above.
Proof

See Appendix

3.5 Discussion

3.5.1 For FD method

The choice of parameter A

As stress above, the question of parameters which depends the FD method is important for its algorithm. In this work, we give the criteria for the choice of reasonable parameters A and C_1. In the preceding section, we have established that for detecting all right change points we must have $2 \times A < L_0$ with $L_0 := \inf\{|\tau_{k+1} - \tau_k|, k = 1, \ldots, K\}$, see also Fig. 5.

The choice of C_1

In [10], we have $C_1 < \delta_o$ with $\delta_0 = \inf\{|\delta_k|,\ k = 1, \ldots, K\}$ where δ_k are the size of the average. An other hand in the theorem 1, we have obtained a bound of number of false detection and its average using the function φ. For N, L fixed and A supposed verified the condition above, we can choose C_1 optimal. We remark that if C_1 is increasing, the function Ψ is decreasing and consequently the average of number of false alarms is decreasing, so it should to choose C_1 the greatest possible and C_1 must verify the condition was given by [10].

3.5.2 simulation

We use the simulation of the subsection(1.2) and we make various drawings with different values C_1 and A.

Comment

We notice through these drawings above that the number of false alarms and the number of no detections vary according to parameters A and C_1. Thus, a choice of A and C minimizing these two points (NND and NFA) is imperative. It is what we are going to do after.

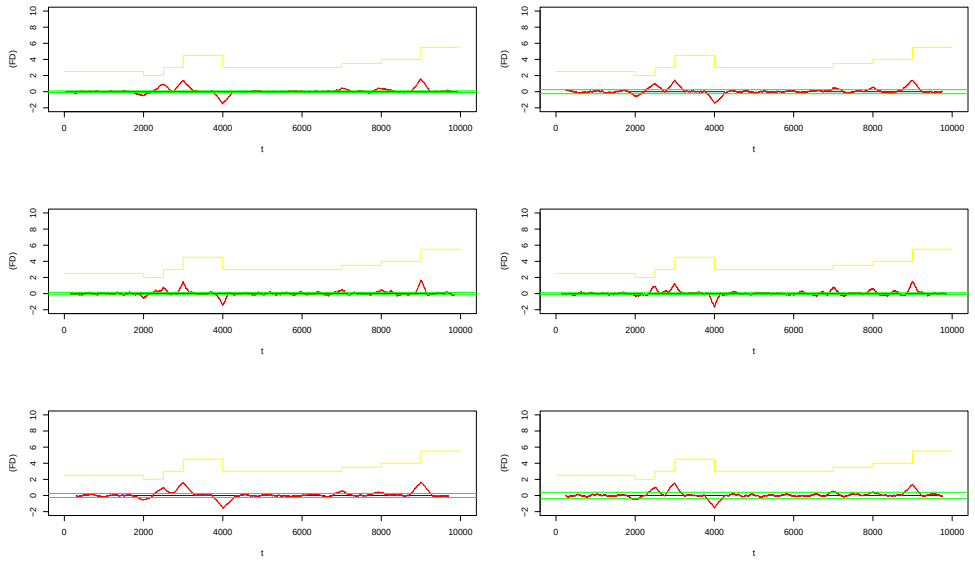


FIGURE 3.7 – *The Filtered Derivative with different parameters A=100; 150; 200; 250 and* $C_1 = 0.1; 0.15; 0.2; 0.25$.

Numerical estimation of NND and NFA

In this part, we want to have an estimation of NND and NFA . For this we make the calculation for Filtered Derivative method with different A=30 to 500 and $C_1 = 0.1$ to 1 and we choose Kmax=20 (we suppose that the maximum number change points is equal 20). For each couple (A,C_1), we make 1000 simulations for to have an exact number of no detection of change points and number of false alarms. Then, we can deduce the NND and NFA for each couple and we sum up the result in the followings arrays :
Table 1. Table of no detections of change points.

A/C_1	0.1	0.2	0.25	0.3	0.4	0.5	0.7
30	0.176	0.307	0.50	0.65	1.05	1.53	2.46
40	0.098	0.243	0.411	0.5	1.02	1.49	2.50
60	0.045	0.132	0.231	0.395	0.874	1.47	2.616
100	0.008	0.053	0.0117	0.219	0.710	1.457	2.766
180	0.005	0.008	0.025	0.071	0.470	1.489	2.92
220	0.002	0.003	0.011	0.055	0.469	1.489	2.934
250	0	0	0.01	0.028	0.369	0.161	2.954
350	0	0	0	1.029	1.143	1.447	2.99
450	0	0	0	0.003	0.716	1.508	2.994
500	0	0	0	0	0.11	1.48	3

Table 2. Table of number of false alarms.

A/C_1	0.1	0.2	0.25	0.3	0.4	0.5	0.7
30	13.51	13.281	13.471	13.622	14.026	14.506	14.796
40	13.166	13.217	13.385	13.469	14.002	14.461	5.810
60	13.019	13.105	13.205	13.369	13.848	10.077	1.714
100	12.834	13.028	13.143	13.052	5.168	2.163	1.283
180	12.976	12.217	6.762	2.948	1.433	1.518	1.0833
220	12.974	8.3033	3.614	1.686	1.269	1.407	1.035
250	12.974	5.752	2.429	1.281	1.227	1.393	1.041
350	10.822	2.331	1.189	0.010	0.291	1.335	1.005
450	4.333	1.296	1.0339	1.001	1.079	1.289	1
500	4.663	1.165	1.0219	0.994	1.030	1.167	0.986

Table 3. Table of integer square error.

A/C_1	0.1	0.2	0.25	0.3	0.4	0.5	0.7
30	7947.31	7947.31	7945.21	7924.43	8078.99	7947.31	7807.86
40	7779.01	7779.01	7871.04	7779.01	7779.01	7779.01	4897.58
60	2475.03	7592.16	7651.39	7737.42	7737.42	7737.42	2419.67
100	7641.06	7565.51	7641.06	7758.73	4531.81	1860.44	1775.77
180	7748.16	7748.18	5540.20	3187.57	1567.89	888.73	2406.88
220	7880.80	6395.04	3818.09	2235.87	1404.52	788.74	2099.12
250	7992.98	5261.26	3094.07	1875.92	1579.60	916.71	2134.77
350	7712.28	3208.19	1983.91	1874.15	1600.91	515.51	2140.14
450	5908.22	2102.07	1837.95	2144.79	1584.63	810.66	2201.55
500	5079.89	2332.44	2223.88	1961.04	1654.26	861.96	2080.03

3.6 Comparison the Filtered Derivative with parameters optimized and Penalized Least Square Error (the adapative method)

Monte Carlo simulation

For comparing the both-methods (The filtered derivative with parameters optimized and adaptive method of [56], we choose the simulation of the subsection(1.2). For FD-method, the optimal parameters chosen are $A_{opt} = 250$ and $C_{1,opt} = 0.25$.
The criteria of comparison are the number of false alarms, the number of no-detection, and the mean square error. Firstly, for one replication, we obtain :

- For adaptive method, NFA = 1, NND = 4, ISE = 28670 (see figure below).
- For filtered derivative with optimized parameters, NFA = 3, NND = 1, ISE = 3998.8 (see figure below).

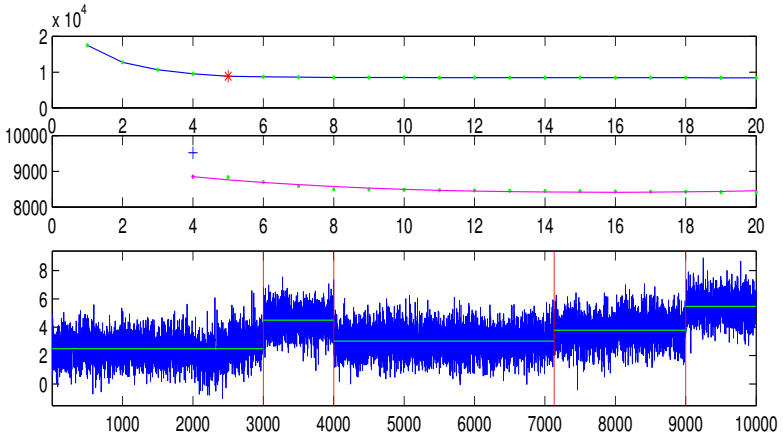


FIGURE 3.8 – *The adaptive method.*

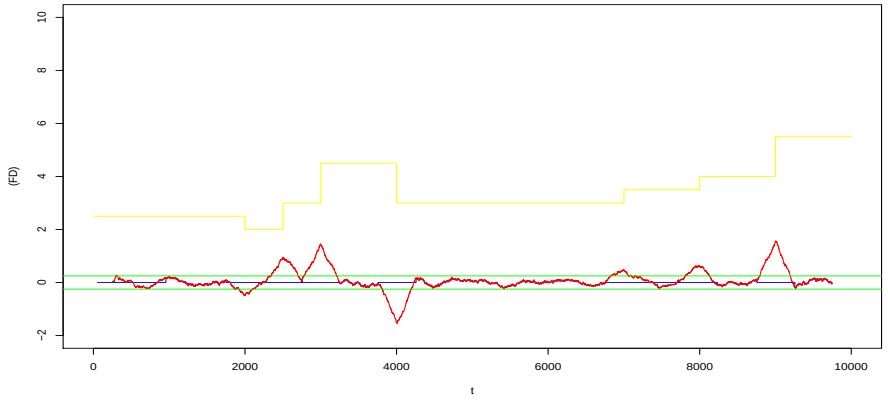


FIGURE 3.9 – *The filtered derivative with parameters optimized* A_{opt} *and* $C_{1,opt}$

The filtered derivative with parameters optimized
For $M = 1,000$ replications, we obtain :

- For FD-method, we obtain MISE= 3094.07, the number of false alarms NFA=2.429, the number of no detection NND=0.01.

- For adaptive-method, MISE= 29009, NFA=0.600, NND=2.250.

Numerical conclusion

It is clearly that the FD-method with parameters optimized is better than the PLS-method adaptive [57] for the criteria mean integrate square error. In other hand, the FD-method with parameters optimized has less no detection of points than the PLS-method adaptive but the firstly has many false alarms than the secondly. Stress that, for in forthcoming work, we will add in step 2 for FD-method with parameter optimized for having the number of false alarms at a level close to zero. In other words, we will optimize the FDqV [35] for the q-value corresponds the false discovery rate.

3.7 Conclusion

In this work, we gave the reasonable parameters of filtered derivative method. We obtained these parameters by doing the simulations but if we consider the theorem 1 and fix L, N and choose A in order (3.4.7) we can calculate C_1 theoretically. To do directly a theoretical calculus of A and C_1 is very difficult and not solution at this moment. In other hand, we can say that is better then to monitor the number of false alarms and number of no-detections that to control the probability of false alarms and the probability of no-detection as done in the preceding works. A natural sequel will have to make the same for FDqV-method for keeping as possible the right number of change points. It will be interesting to search in manner to adapt these results for the times series weakly and strongly dependent.

Appendix

In this subsection, we give some technical lemmas and proposition useful for the proof of the main theorem.

Lemma 1

Let $l \in \mathbb{N}$, then

$$\mathbb{P}(M(\omega) \leq l-1) = \mathbb{P}(\sum_{j=1}^{l} RL_j(\omega) > N)$$

Proof. We have

$$\{\sum_{j=1}^{l+1} RL_j(\omega) > N\} = \{M(\omega) \leq l\}$$

In fact, the event $\{\sum_{j=1}^{l+1} RL_j(\omega) > N\}$ means that the number of false alarms is not more than l.

and

$$\{\sum_{j=1}^{l} RL_j(\omega) \leq N\} = \{M(\omega) \geq l\}$$

$$\mathbb{P}(M(\omega) < l) = 1 - \mathbb{P}(M(\omega) \geq l) = \mathbb{P}(\sum_{j=1}^{l} RL_j(\omega) > N)$$

Finally, we have the result above

Lemma 2

We suppose that σ is constant and know, then $\Gamma(A,k)$ is a family Gaussian such that $\Gamma(A,k) \sim \mathcal{N}(0, \sigma\sqrt{\frac{2}{A}})$

Proof.

See [10].

Proposition 2

Let $l_j \in \mathbb{N}$,

we have

$$\mathbb{P}(RL_j \leq l_j) \leq |l_j - 2 \times A|\Psi(\frac{AC_1}{\sigma^2\sqrt{l_j}})$$

with

$$\Psi(x) = 1 - \Phi(x) \; and \; \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-u^2}{2}} du$$

Proof. From

$$(RL(A, C_1) > l_j) = \{max(\Gamma(A,k), \Gamma(A,k+1), \ldots, \Gamma(A,l_j)) \leq C_1\}$$

We can write

$$\mathbb{P}(RL_j \leq l_j) = 1 - \mathbb{P}(RL_j > l_j)$$

and

$$\mathbb{P}(RL_j > l_j) = \mathbb{P}(\max_{t\in[k,l_j]} \Gamma(A,t) \leq C_1)$$

By scaling, we obtain

$$\max_{t\in[k,l_j]} \Gamma(A,t) =^{\mathcal{L}} \sigma A^{-1}\sqrt{l_j}\rho(\frac{A}{l_j})$$

where

$$\rho(u) = \max[W(u + \frac{1}{A}) + W(u - \frac{1}{A}) - 2W(u)] \; for \; u = \frac{A}{n}, \ldots, 1 - \frac{A}{n}$$

ρ is the maximum of discrete Wiener Process, according to Lemma (3.3.2), we know

$$[W(u + \frac{1}{A}) + W(u - \frac{1}{A}) - 2W(u)] \sim \mathcal{N}(0, \sigma\sqrt{\frac{2}{A}})$$

Then

$$\mathbb{P}(\max_{t\in[k,l_j]} \Gamma(A,t) \leq C_1) = \mathbb{P}(\sigma A^{-1}\sqrt{l_j}\rho(\frac{A}{l_j}) \leq C_1)$$

According to the following remark from [28],
If $|I|$ is finite and $\forall\, i \in I$,$X_i \in \mathcal{N}(0, \sigma_i)$, then

$$\mathbb{P}(\sup_{i\in I} X_i \geq a) \leq |I|\Psi(\frac{a}{\sup_{i\in I}\sigma_i})$$

Finally, we get

$$\mathbb{P}(RL_j \leq l_j) \leq |l_j - 2 \times A|\Psi(\frac{AC_1}{\sigma^2\sqrt{l_j}})$$

Proof. [Proof of Theorem 1]
We prove the upper bound (3.4.8).

$$(\sum_{j=1}^{l} RL_j(\omega) > N) = \bigcup_{\{(l_1,\ldots,l_j),\sum l_j > N\}} \{\forall j = 1,\ldots,l, \ such\ that\ RL_j = l_j\}$$

$$\mathbb{P}(\sum_{j=1}^{l} RL_j(\omega) > N) = \mathbb{P}(\bigcup_{\{(l_1,\ldots,l_j),\sum l_j > N\}} \{\forall j = 1,\ldots,l, \ such\ that\ RL_j = l_j\})$$

By independence of RL_j, we can write

$$\mathbb{P}(\sum_{j=1}^{l} RL_j(\omega) > N) = \sum_{\{(l_1,\ldots,l_j),\sum l_j > N\}} \mathbb{P}(\forall j = 1,\ldots,l, \ such\ that\ RL_j = l_j)$$

$$\mathbb{P}(\sum_{j=1}^{l} RL_j(\omega) > N) = \sum_{\{(l_1,\ldots,l_j),\sum l_j > N\}} \mathbb{P}(\bigcap_{j=1}^{l} \{RL_j = l_j\})$$

Using again independence of RL_j

$$\mathbb{P}(\sum_{j=1}^{l} RL_j(\omega) > N) = \sum_{\{(l_1,\ldots,l_j),\sum l_j > N\}} \prod_{j=1}^{l} \mathbb{P}(\{RL_j = l_j\})$$

According to the proposition 2, we have

$$\mathbb{P}(\sum_{j=1}^{l} RL_j(\omega) > N) \leq \sum_{(l_1,\ldots,l_j),\sum l_j > N} \prod_{i=1}^{l} |l_j - 2A|\Psi(\frac{AC_1}{\sigma\sqrt{l_j}}) := \varphi(A, C_1, N, l)$$

Finally, using the lemma 1 we have the result (3.4.8)

Chapitre 4

Multiple change points detection in weakly dependent random variables using filtered derivative and false discovery rate method.

This article is published in "World Statistics Congress (WSC 2017)", International Institute Statistics, Marrakesh, 2017, Jul 16 – July 21.

Abstract

Let $\mathbf{X}=(X_1, X_2, \ldots, X_n)$ be a time series, that is a sequence of weakly dependent of random variable indexed by the time t=$1, 2, \ldots, n$. We assume that there exists a segmentation $(\tau_1, \tau_2, \ldots, \tau_K)$ such that X_i is a local stationary process for all time i $\in (\tau_k, \tau_{k+1}]$, for k=1,2,..., K, where K is unknown number of changes. The simplest model is to consider that X_i are autoregressive processes with change on the mean. We estimate the instant of breaks and means corresponding by using the filtered derivative and false discovery rate method(FDqV). As already established in the case of independent random variables, the FDqV has two steps : the first step compute the filtered derivative(FD) and then we select the potential change points as local maxima of the FD-function reaching a threshold and the second step (false discovery rate) eliminates false alarms and keeps as possible all right change points. We compare the FDqV method by a new method derived by the penalized least square method and give a real application with heartbeat series, because in the observed data, it exists a correlation between those.

4.1 Introduction

The problem of change points problem has been studied in the last forty years. It essentially exists two methods : the Penalized Least Square Method(PLSM) see for example [4] and the Filtered Derivative see for example [6,8,35]. Contrary others methods such that the law large of number, these methods do estimation for data of fixed size. Recently, many authors such that [56, 57] improved the PLSM and [35, 37] improved the FD method by adding a second step called the False Discovery Rate. The PLSM needs to compute a matrix n $\times$ n, where n is the size of datasets whereas the FD needs a matrix 1 $\times$ n. Due to the big

data, the size of datasets become larger and larger, then the computational complexity of statistical methods became a challenge and the FD method is more advantaged by PLSM for time and memory complexity. Many applications such that telecommunications, informatics and the data linked of health use the problem change points. In [36], we are done a real application by the Filtered Derivative and False Discovery(FDqV). Generally, the data of domain applications are a structural dependence between those. In this paper, we will do a comparison between the FDqV method and by a new method developed by [21] and we give an application of heartbeat series by the FDqV method. In the sequel, the paper is structured as follows : the section 2 presents the problem of change points, the section 3 describes the new method derived the PLSM, the section 4 gives the FDqV method for AR(1) processes, the section 5 compares the two methods, the section 6 gives an application of heartbeat series and the last section concludes the paper.

4.2 Description of problem.

Model

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a series indexed by the time $t = 1, 2, \ldots, n$. We assume there exists a segmentation $\tau = (\tau_1, \tau_2, \ldots, \tau_K)$ such that X_τ is a family of weakly dependent random variables for $(\tau_k, \tau_{k+1}]$, and $k = 0, 1, 2, \ldots, K$ where by convention $\tau_0 = 1$ and $\tau_{K+1} = n$. The simplest model is X_τ for a sequence of autoregressive processes with change on the mean.
More precisely, we consider the segmentation of autoregressive process with homogeneous auto-correlation coefficient ρ

$$X_t = \mu_k + \varepsilon_t,\ \tau_k + 1 \leq t \leq \tau_{k+1},\ 0 \leq k \leq K,\ 1 \leq t \leq n, \tag{4.2.1}$$

where (ε_t) is a zero-mean second-order stationary $AR(1)$ process defined as the solution of

$$\varepsilon_t = \rho * \varepsilon_{t-1} + \eta_t, \tag{4.2.2}$$

where $|\rho| < 1$ and η_t is a white noise with variance σ^2

Statistical challenge

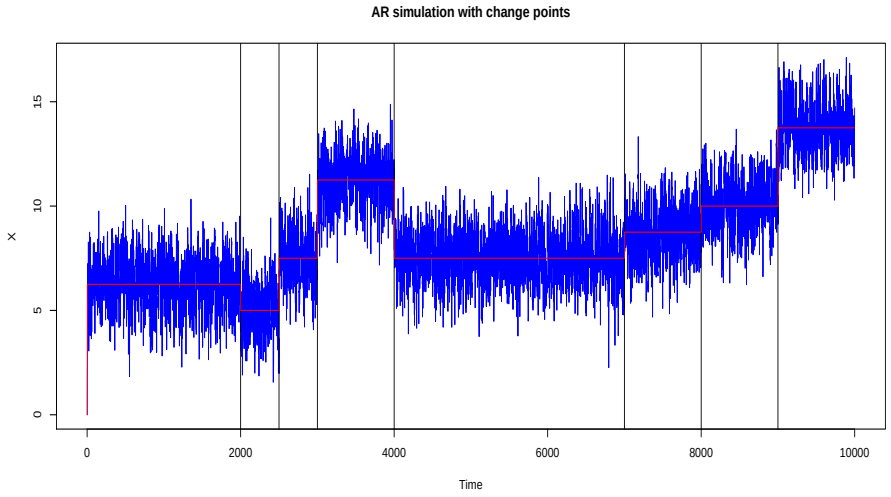
Assume that we do not know in advance the number K of change points. We have to estimate the configuration of change $\tau = (\tau_1, \ldots, \tau_K)$ and the values of the mean $(\mu_0, \mu_1, \ldots, \mu_K)$. We denote the estimates by $\widehat{\tau} = (\widehat{\tau}_1, \ldots, \widehat{\tau}_K)$ and $(\widehat{\mu}_0, \widehat{\mu}_1, \ldots, \widehat{\mu}_K)$. Remark that the number of change points is unknown and estimated by $\widehat{K}$.

Simulation

For n=10000, we have done one realization of a sequence of AR(1) process with change on means. We consider seven change points $\tau = (2000, 2500, 3000, 4000, 7000, 8000, 9000)$ with means $\mu = (2.5, 2, 3, 4.5, 3, 3.5, 4, 5)$. Below we give a drawing for an AR(1) process with change points.

4.3 A new method derived from Penalized Least Square

In this Section, we can see the paper of [21].



4.4 The Filtered Derivative and False Discovery Rate for AR(1) process

As in [35], the Filtered Derivative and False Discovery Rate has two steps : The first step allows to detect the potential change points and the second step distinguish the false and the right alarms.

The Filtered Derivative

The Filtered Derivative is a function defined as follows :

$$FD(t, A) = \widehat{\mu}(t+1, t+A) - \widehat{\mu}(t-A, t+1) \tag{4.4.1}$$

where $\widehat{\mu}(t+1, t+A) = \sum_{i=t+1}^{t+A} X_i$.
In the case without noise the FD function has a hat function around the instant of change points as showed in the following figure.

The instant of change points are local maximal of the absolute value of the filtered derivative function.
In this case of with noise, for estimating the instant of potential change points [6], and [35] choose a threshold C and only keep the instants such that $\max_{t\in[A,n-A]} |FD(t, A)|$ exceeds the threshold C. At the end of this step, we have among the potential change points, the false alarms. So for only detecting the right alarms, [35], adds a second step called the False Discovery Rate, for more details see [35]. In other hand, the FD function uses two parameters : the threshold C and the window A for calculating means. [37] gave the optimization of C and A.

4.5 Comparison of Two methods.

4.5.1 Comparison criteria

Assume that we do not know in advance the number K of change points. We have to estimate the configuration of change $\tau = (\tau_1, \ldots, \tau_K)$ and the values of the mean $(\mu_0, \mu_1, \ldots, \mu_K)$. We denote the estimates by $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}})$ and $(\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{\hat{K}})$. Remark that the number of change points is unknown and estimated by $\widehat{K}$.

Criterion

1. The quality of estimation for one sample can be measured by two criteria :
   - $\hat{K} - K$
   - The integrated square error (ISE). Actually, we can reformulate the problem as an estimation of a noisy signal. The signal is

     $$s(t) = \sum_{k=0}^{K} \mu_k \times \mathbf{1}_{(\tau_k, \tau_{k+1}]}(t)$$

     where we have set by convention $\tau_0 = 0$ and $\tau_{K+1} = N$. The estimated signal is then

     $$\widehat{s}(t) = \sum_{k=0}^{\hat{K}} \widehat{\mu}_k \times \mathbf{1}_{(\widehat{\tau}_k, \widehat{\tau}_{k+1}]}(t)$$

     and the integral square error (ISE) by

     $$ISE = \sum_{i=1}^{N} \left\{ [\widehat{s}(t) - s(t)]^2 \right\}$$

2. However, a result on just one simulation is hazardous. So, we have to do M simulations, with e.g. $M = 1,000$ and calculate the mean integrated square error (MISE).

3. The second family of criteria is the time complexity and the memory complexity that is the mean CPU time for estimating $\widehat{s}$ and which quantity of memory is used.

Simulation
We use the above simulation. The FDqV method depends the parameters C and A for the step 1 and the threshold q false discovery rate for step 2. We choose $C = 0.25$, $A = 250$ and $q = 0.01$ and Kmax=15.
Monte Carlo simulation
For one sample, we apply the FDqV method with parameters given above and we obtain the following change points : 1991 2497 3005 4000 7007 7991 9000.
For PLS method of [21], we obtain the following change points : 1990 2502 3002 4001 7001 8033 8999.
We made $M = 1,000$ simulations for autoregressive model above and set $\hat{K}$ is the length of estimated change points and K is the length of right abrupt changes :

1) For FDqV method, we obtain $mean(\hat{K} - K)$=0.0023.

2) For PLS method of [21], we obtain $mean(\hat{K} - K)$=0.0015.

Mean Integrate Square Error
For $M = 1,000$ simulations Monte Carlo, we obtain the following values for MISE

1) For FDqV method, MISE=1200.732

2) For PLS method of [21], MISE=2294.914

The mean time complexity
We use the computer with following characteristics : 2.40GHz processor and 2.93Go memory.

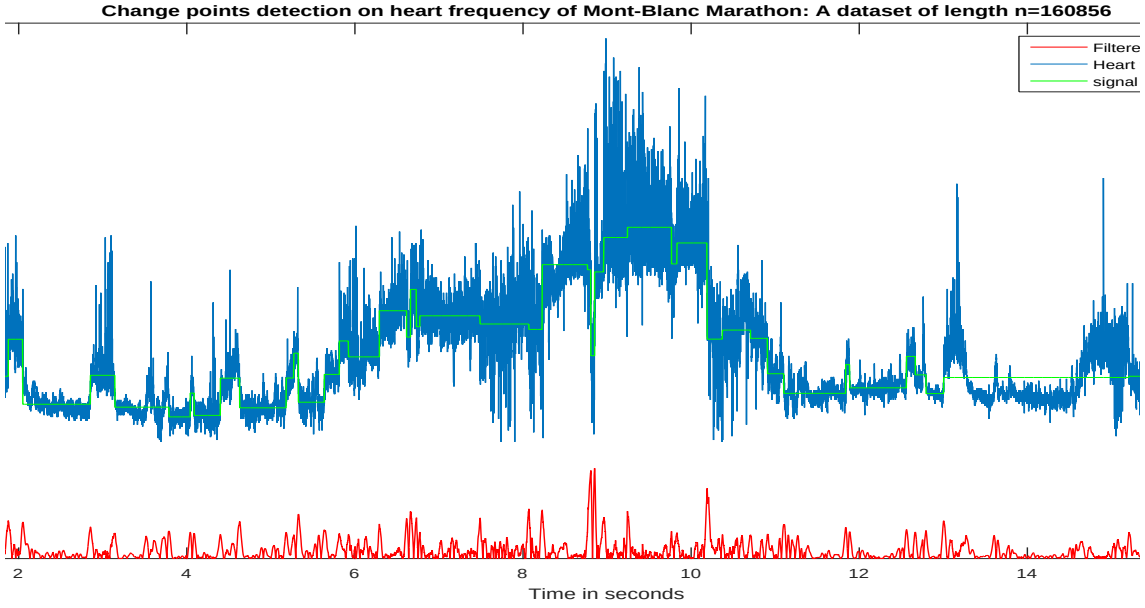
1) For FDqV method, CPU(Central Processing Unit)=32.07 seconds.

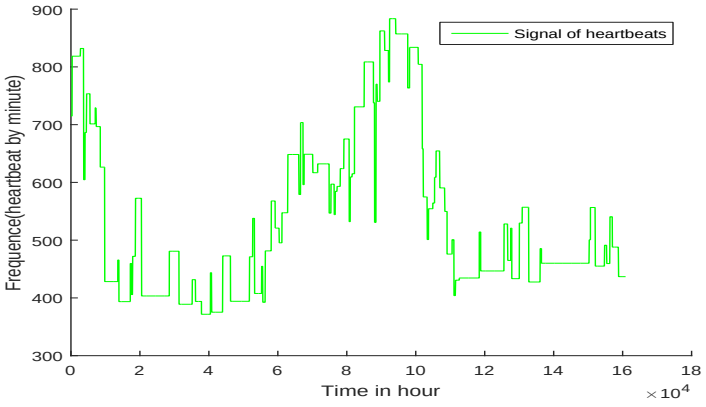
2) For PLS method of [21], CPU= 155.92 seconds.

Numerical conclusion
We clearly see that both methods give good results for the criteria $\hat{K} - K$ and the FDqV method is less expensive in time and memory complexity. For the criteria MISE, the FDqV method is best than the PLS method of [21]. For a big data, It will be reasonable using the FDqV method.

4.6 Application of real data heartbeats.

In this section, we apply the FDqV method on heartbeat series. Recall that the data of heartbeats aren't independent but weakly dependent, so for modelling that, we use the autoregressive model.
The data concerns the heart frequency of Mont-Blanc marathon. The size of data n=160856 and we take the threshold detection $C = 5$ beats by minute, the time resolution $A = 400$, $Kmax = 50$ and the false discovery rate $q = 0.01$. We detect the change of heart frequency of a marathon runner of Mont-Blanc in 2006 and we obtain the following figure.


Change points detection on heart frequency of Mont-Blanc Marathon: A dataset of length n=160856
Filtere
Heart
signal
2
4
6
8
10
12
14
Time in seconds


Signal of heartbeats
Frequence(heartbeat by minute)
Time in hour
900
800
700
600
500
400
300
0
2
4
6
8
10
12
14
16
18
×10^4

4.7 Conclusions.

In the precedent works such that [35, 37], we applied the FDqV method for a sequence of independent random variables and gave the optimized parameters. In this work, we apply the FDqV method for the variables weakly dependent and we clearly see that the Filtered Derivative is better for big data. Generally the PLS method needs a matrix by $n \times n$ and the time and memory complexity is order $\mathcal{O}(n^2)$ and the FDqV method use a vector $1 \times n$ and the time and memory complexity is order $\mathcal{O}(n)$, so for the development of big data of 21th century, it is advantageous using the FDqV method for detecting change points. For forthcoming work, we will do the FDqV method for the variables strong dependent.

Chapter 5

Inference of Threshold Autoregressive (TAR) models with dependent errors.

This chapter is a version of a forthcoming article in collaboration with Bruno Saussereau.

The Threshold Autoregressive (TAR) models are introduced by [82] and were studied by many authors such that [24, 68, 69] and references therein. This model captures the dynamic behavior of time series by switching the regimes. The TAR model plays an important role in nonlinear time series and have been widely used to nonlinear phenomena in various fields, for example economics, environment, hydrology, physics, population dynamics, biological sciences, and among others. The TAR process is able to capture asymmetric limit cycles, as the main motivation for these models was to describe limit cycles of cyclical time-series [84]. For an update overview on TAR models, we can see [83]. The popularity of TAR models are due to the fact there produce a simplified way of presenting a complex stochastic system in terns of decomposing it into a set smaller sub-system. Applications of TAR models include modeling exchanges rates and modeling arbitrage opportunities implied by the difference the spot and futures prices for a given markets. For example [78], presented strong evidence of presence of non-linearity in business cycles which confirmed that business cycles exhibit asymmetric behaviour. Others models which capture the dynamic complex functions are the Self-Exciting Threshold Autoregressive (SETAR) model, the Smooth Transition Autoregressive (STAR) model, the Logistic Smooth Transition Autoregressive (LSTAR) model. STAR models were fisrt proposed by [23] as a generalization of a non-linear two regimes univariate SETAR model. SETAR models are a special case of general univariate TAR models, where the state-dependent variable is the dependent variable itself. The LSTAR models have a logic distribution that approximate to the normal distribution and also have an advantage terms of being able to estimate theirs parameters using analytic derivatives. The LSTAR models also have distinct computational advantages over standard TAR models.
The main goal in TAR models or SETAR models is to study the asymptotic properties of the estimated parameters and the estimated threshold. In [24], the author showed that under some regularity conditions, the least squares estimators of a stationary ergodic TAR models is strongly consistent. Qian, in [73], establishes similar results as in [24] for the maximum likelihood estimators of the same model under some regularity conditions on the errors density, not necessarily Gaussian. Moreover, [60] provided a numerical

method to tabulate the limiting distribution of the estimated threshold in practice. In [47–49], the authors developed a statistical theory for threshold estimation in the context of regression. Under the assumption that the threshold effect is vanishingly small, he obtained the distribution and parameter free limit of the estimated threshold.

Up to date, all papers treat the case where the models rely on strong assumptions on the noise processes, such as independence or martingale difference. A natural sequel is therefore to investigate the case where the strong hypothesis on the non-linear innovation do not hold. In other words, we will work in the framework of a noise sequence that is no more an independent sequence of random variables but just a sequence of uncorrelated random variables. This implies that the TAR process is no longer Markovian and is no more geometrically mixing. Therefore we adopt the framework of [43] in which the authors studied the Autoregressive and Moving-Average (ARMA) models, under a mixing property (and a stronger moment condition) on the observed process. They called these models weak ARMA models in opposition to strong ARMA models when the noise is an independent and identically distributed (iid for short) sequence. In link with weak ARMA models, we name these models weak TAR models. After the pioneering work of [43], many articles have been devoted to the study of weak models when one works with non independent innovation process. Nevertheless, to our knowledge, no study has addressed the question of weak TAR models.

This chapter studies the least square estimation (LSE) of weak TAR models and the asymptotic properties of the estimators. Under reasonable mixing assumptions for the time series process, we will prove that the LSE is strongly consistent and we will study the asymptotic laws. Although the consistency result is not really affected by our context, the asymptotic distribution needs further attention. For the parameters arising in the autoregressive formulation, we will be able to adapt the techniques of [43] using the mixing assumptions and as usual in this case, an extra moment assumption. Indeed, we shall require that the process has moments of order strictly greater than 4 whereas in the classical case, the fourth order moments are sufficient to investigate the asymptotic normality.

The asymptotic behaviour of the law of the estimator of the threshold parameter is certainly the main novelty of our work. When the noise is strong (that means it is an iid sequence), the time series process is an ergodic Markov chain which is geometrically mixing. This is a stronger statement than the ones we do in our context because we will assume only $\alpha-$mixing property of the process. The results presented in [24, 60] heavily depend on the geometrically mixing property of the Markov chain. So we have to adapt their methodology in our case and this is feasible thanks to a weak convergence result of sums over triangular arrays to the compound Poisson limit under mixing assumptions. This technique is new in the time series context and we hope that it shall be used in other problems.

The remainder of this chapter is as follows. The next section introduces the model and its estimation procedure and gathers our main results. The Section 5.2 is devoted to proof of Theorem 5.1.4 in which the asymptotic distribution of the threshold estimator is stated. This is in this section that an original approach is conducted. In Section 5.4, some proofs are also gathered. They concern more classical techniques. An Appendix (see Section 5.5) contains the proof of the consistency for which the approach is classical. We decided, nevertheless, to give the details for the sake of completeness. Our numerical illustrations are gathered in Section 5.3.

5.1 Model, assumptions and main results

A time series $\{X_t\}_{t\in\mathbb{Z}}$ is said to be a weak TAR model if it satisfies

$$X_t = \begin{cases} \alpha_0\, X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} \le r_0 \\ \beta_0\, X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} > r_0 \end{cases} \tag{5.1.1}$$

where the noise $\varepsilon = (\varepsilon_t)_{t\in\mathbb{Z}}$ is a weak noise, that is it satisfies the following assumption.

(H1) The sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ is strictly stationary, square integrable, and satisfies

- for any t, $\mathbb{E}(\varepsilon_t) = 0$
- for any s, t, $\mathbb{E}(\varepsilon_t\varepsilon_s) = \delta_{s,t}\,\sigma^2$ where $\delta_{t,s} = 1$ if $s = t$ and 0 otherwise
- for any t, $\mathbb{E}\varepsilon_t^4 < \infty$

In the model formulation (5.1.1), r_0 is called the true threshold parameter and it is unknown. Without loss of generality, we assume that there exist two finite constants $\underline{r}$, $\overline{r}$ such that $r_0 \in [\underline{r}, \overline{r}] := I$. When $r_0 = -\infty$ or $r_0 = +\infty$, the model reduces to a weak autoregressive (AR) model which is not of our interest.
The true parameter is denoted by $\theta_0 = (\alpha_0, \beta_0, r_0)' \in \mathbb{R}^2 \times [\underline{r}, \overline{r}] = \mathbb{R}^2 \times I$ and a generic parameter is $\theta = (\alpha, \beta, r) \in \mathbb{R}^2 \times I$. We assume that the parameter space Θ is a compact subspace of $\mathbb{R}^2 \times I$. We assume that σ is known (and equals 1). If σ is not known, it can be estimated by classical method as soon as we know how to estimate the parameter θ.
We will also use the restricted parameter $\lambda = (\alpha, \beta)' \in \Lambda$. This goes without saying that $\theta' = (\alpha, \beta, r) = (\lambda', r)$ and $\Theta = \Lambda \times I$ and Λ is also assumed to be compact.
For any $\theta = (\alpha, \beta, r) \in \Theta$, we denote

$$\begin{aligned} \epsilon_t(\theta) &= X_t - (\alpha + (\beta - \alpha)\mathbf{1}_{\{X_{t-1}>r\}})X_{t-1} \\ &= X_t - A_{t-1}(\theta)X_{t-1}\ , \end{aligned} \tag{5.1.2}$$

where obviously

$$A_{t-1}(\theta) = \alpha + (\beta - \alpha)\mathbf{1}_{\{X_{t-1}>r\}} = \beta + (\alpha - \beta)\mathbf{1}_{\{X_{t-1}\le r\}}\ . \tag{5.1.3}$$

We suppose that a sample $\{X_1, \ldots, X_n\}$ is a sample from the model (5.1.1) with the true parameter θ_0. Given the initial value $\mathcal{X}_0 = \{X_t; t \le 0\}$ (that we may assume equal to 0 or equivalently $\epsilon_t(\theta) = 0$ for $t \le 0$), we consider the following sum of squares errors:

$$L_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \epsilon_t^2(\theta)\ .$$

The minimizer $\hat{\theta}_n$ of $L_n(\theta)$ is called the least squares estimator of θ_0, that is,

$$\hat{\theta}_n = \inf_{\theta\in\Theta} L_n(\theta)$$

Since the function $L_n(\theta)$ is discontinuous in r, a manner to obtain $\hat{\theta}_n$ is as follows:

- for fixed r, one minimizes $L_n(\theta) = L_n(\lambda, r)$ and get its minimizer $\hat{\lambda}_n(r) = (\hat{\alpha}_n(r), \hat{\beta}_n(r))'$ and minimum $L_n^*(r) = L_n(\hat{\alpha}_n(r), \hat{\beta}_n(r), r)$,

- since $L_n^*(r)$ only takes a finite number of possible values, the one with the smallest r can be chosen as $\hat{\theta}_n$.

In all this work, we assume that $|\alpha| + |\beta| < 1$. This condition in sufficient to ensure the invertibility of model (see Theorem A1 in [61]). Even if we do not use the invertibility in our proof since the main assumptions will be put on the process X, this condition is necessary to prove that the initial values $\mathcal{X}_0$ will not affect the asymptotic properties of the estimator $\hat{\theta}_n$ (see [60] for further details).
Before stating the convergence result of $\hat{\theta}_n$ towards θ, one shall need further assumptions on the process $(X_t)_{t\in\mathbb{Z}}$.

(H2) The process $(X_t)_{t\in\mathbb{Z}}$ is ergodic, strictly stationary and has fourth order moments. Moreover, for any t, the probability distribution function of X_t is absolutely continuous. Its density is denoted by π and is bounded away from 0 and ∞ over each bounded set. The function π is also assumed to be Lipschitzian.

Supposing that $|\alpha| + |\beta| < 1$ implies the ergodicity in **(H2)** as it was noticed in [24]. See also the work of Chan and Tong [22] for some general sufficient conditions for stanionarity and ergodicity.
We shall also need the following hypotheses.

(H3) The threshold r_0 in $\mathbb{R}$ is the discontinuity point of autoregressive function, that is

$$(\beta_0 - \alpha_0) \neq 0 \ .$$

The above hypotheses is natural because if $\alpha_0 = \beta_0$, then our model becomes a simple auto-regressive AR(1) model.
Under the above hypothesis, we can state the following consistency result.

Theorem 5.1.1. *Let $(X_t)_{t\in\mathbb{Z}}$ be the TAR process satisfying (5.1.1). We assume that **(H1)**, **(H2)** and **(H3)** hold. Then, $\hat{\theta}_n \to \theta_0$, a.s. as $n \to +\infty$*

The proof of this result is classical. The fact that the noise is a weak noise does not affect the arguments of [24] and [73]. Nevertheless, for the sake of completeness, a proof is proposed in the Appendix (see Section 5.5).
Next, we study the limiting distribution of $\hat{\theta}_n$.
We shall need some notations concerning the mixing property that will be assumed. First we recall that for two random variables X and Y, the mixing coefficient $\alpha(X, Y)$ is defined by

$$\alpha(X, Y) = \sup_{A\in\sigma(X), B\in\sigma(Y)} |P(A \cap B) - P(A)P(B)|.$$

where $\sigma(X)$ is the sigma-filed generated by X. We will make use of the Davydov inequality (see [29] or [44]) that states that for p, q and r three positive numbers such that $1/p + 1/q + 1/r = 1$, there exists a constant K such that we have

$$\mathrm{Cov}(X, Y) \leq K\|X\|_{\mathbb{L}^p}\|Y\|_{\mathbb{L}^q}|\alpha(X, Y)|^{\frac{1}{r}} \ . \tag{5.1.4}$$

Now, let $\mathcal{F}_{-\infty}^t$ and $\mathcal{F}_{t+k}^{\infty}$ be the σ-fields generated by $\{X_u : u \leq t\}$ and $\{X_u : u \geq t + k\}$ respectively. The strong mixing property coefficients $(\alpha_X(k))_{k\in\mathbb{N}^*}$ of the stationary process (X_t) are defined by

$$\alpha_X(k) = \sup_{A\in\mathcal{F}_{-\infty}^t, B\in\mathcal{F}_{t+k}^{\infty}} |P(A \cap B) - P(A)P(B)|. \tag{5.1.5}$$

We formulate the following hypothesis.

(H4) $(X_t)_{t\in\mathbb{Z}}$ satisfies the strong mixing condition: there exists $\nu > 0$ such that

$$\sum_{k=0}^{\infty} \{\alpha_X(k)\}^{\frac{\nu}{2+\nu}} < \infty \ . \tag{5.1.6}$$

The above strong mixing condition will be used hereafter by means of Davydov's inequality which has just been recalled in (5.1.4). Therefore, the following moment condition will also be needed.

(H5) $(X_t)_{t\in\mathbb{Z}}$ satisfies $\mathbb{E}|X_t|^{4+2\nu} < \infty$ with the real ν from Assumption **(H4)**.

When the noise is iid, the time series process satisfies a geometric mixing property which is stronger than our assumptions **(H4)** and **(H5)** and plays an important role in the work of [24, 60] and all related works using their techniques.
The following result is an auxiliary result that gives, among other, a rather sharp speed of almost-sure convergence of our estimator $\hat{r}_n$ toward r_0. Hence it completes the consistency result proved in Theorem 5.1.1. These results will be useful for the study of the asymptotic laws of $\hat{\theta}_n - \theta$.

Theorem 5.1.2. *Under conditions **(H1)** to **(H5)**, it holds*

1. $n^{\kappa}(\hat{r}_n - r_0) = \mathrm{O}_{\mathbb{P}}(1)$ *with* $\kappa = (2+\nu)/(3+2\nu)$.
2. $\sup_{|r-r_0|\leq \frac{B}{n}} \left(|\hat{\alpha}_n(r) - \alpha_0| + |\hat{\beta}_n(r) - \beta_0|\right) = \mathrm{o}_{\mathbb{P}}(1)$.

The proof of this result is done in Section 5.4. It mainly follows the arguments of [24] but we will focus to show how the mixing property intervenes at different stages of reasoning.
In order to state the asymptotic normality result we need the following notation. We recall that we have denoted $\lambda = (\alpha, \beta)$ the restricted parameter being such that $\theta = (\lambda, r) \in \Lambda \times I = \Theta$.

Theorem 5.1.3. *If Assumptions **(H1)** to **(H5)** hold, then* $\hat{\lambda}_n(\hat{r}_n) = \begin{pmatrix} \hat{\alpha}_n(\hat{r}_n) \\ \hat{\beta}_n(\hat{r}_n) \end{pmatrix}$ *satisfies*

$$\sqrt{n}(\hat{\lambda}_n(\hat{r}_n) - \lambda_0) = \sqrt{n}(\hat{\lambda}_n(r_0) - \lambda_0) + \mathrm{o}_{\mathbb{P}}(1)$$

and $\sqrt{n}(\hat{\lambda}_n(r_0) - \lambda_0)$ *has a normal limiting distribution with mean 0 and covariance matrix* $J^{-1}IJ^{-1}$ *with*

$$J = 2\begin{pmatrix} \mathbb{E}(X_1^2 \mathbf{1}_{\{X_1 \leq r_0\}}) & 0 \\ 0 & \mathbb{E}(X_1^2 \mathbf{1}_{\{X_1 > r_0\}}) \end{pmatrix} \quad \text{and} \quad I = \lim_{n\to\infty} \left(\sqrt{n}\frac{\partial L_n(\lambda_0, r_0)}{\partial \lambda}\right).$$

Asymptotic behaviour of $\hat{r}_n$

Now, we study the limiting distribution of $n(\hat{r}_n - r_0)$. The arguments follow the ideas of [24]. Nevertheless, many precisions about this method are given in [60] and we further need to give more specific details due to our context.
In order to explain our the strategy, we consider the following profile sum of squares errors function defined for $s \in \mathbb{R}$ by

$$\tilde{\phi}_n(s) = nL_n\left(\hat{\lambda}_n(r_0 + s/n), r_0 + s/n\right) - nL_n\left(\hat{\lambda}_n(r_0), r_0\right). \tag{5.1.7}$$

Suppose that the sequence of processes $((\tilde{\phi}_n(s)_{s\in\mathbb{R}})_{n\geq 1}$ converges in the Skorohod space $\mathbb{D}(\mathbb{R})$ of càdlàg functions on $\mathbb{R}$ (details will be given hereafter) to a process $(\phi(s))_{s\in\mathbb{R}}$. Then one uses a continuity result on the Skorohod space that have been established in [76, 77] that asserts that the argmin will also converge to the argmin of the process $(\phi(s))_{s\in\mathbb{R}}$ (if it exists).
To prove this convergence, we show that the sequence $(\tilde{\phi}_n(\cdot))_{n\geq 1}$ can be approximated in $\mathbb{D}(\mathbb{R})$ by the sequence of processes $(\phi_n(\cdot))_{n\geq 1}$ defined by

$$\phi_n(s) = nL_n\left(\alpha_0, \beta_0, r_0 + s/n\right) - nL_n(\alpha_0, \beta_0, r_0). \tag{5.1.8}$$

Using (5.1.2) and easy calculus, one may write ϕ_n as

$$\phi_n(s) = \sum_{t=1}^{n} \zeta_{1,t}(s)\mathbf{1}_{\{s<0\}} + \sum_{t=1}^{n} \zeta_{2,t}(s)\mathbf{1}_{\{s\geq 0\}} \tag{5.1.9}$$

with

$$\zeta_{1,t}(s) = \Big(2X_tX_{t-1}(\alpha_0 - \beta_0) + X_{t-1}^2(\beta_0^2 - \alpha_0^2)\Big)\mathbf{1}_{\{r_0+s/n<X_{t-1}\leq r_0\}} \tag{5.1.10}$$

$$\zeta_{2,t}(s) = \Big(2X_tX_{t-1}(\beta_0 - \alpha_0) + X_{t-1}^2(\alpha_0^2 - \beta_0^2)\Big)\mathbf{1}_{\{r_0<X_{t-1}\leq r_0+s/n\}}\ . \tag{5.1.11}$$

We also denote

$$\Gamma_{1,t}(s) = 2X_tX_{t-1}(\alpha_0 - \beta_0) + X_{t-1}^2(\beta_0^2 - \alpha_0^2) \tag{5.1.12}$$

$$\Gamma_{2,t}(s) = 2X_tX_{t-1}(\beta_0 - \alpha_0) + X_{t-1}^2(\alpha_0^2 - \beta_0^2)\ . \tag{5.1.13}$$

We will prove that the ϕ_n converges to a two sided compound poisson process ϕ in the Skorohod space. In order to define the limiting process, one introduces $F_1(.|r_0)$ the conditional distribution of $\Gamma_{1,t}$ given $X_{t-1} = r_0^-$ and $F_2(.|r_0)$ the conditional distribution of $\Gamma_{2,t}$ given $X_{t-1} = r_0^+$. This measure exists and is the limiting conditional distribution of $\Gamma_{2,t}$ given $\{r_0 < X_{t-1} \leq r_0 + \delta\}$ as $\delta \downarrow 0$. Analogously, $F_1(.|r_0)$ exists as the limiting conditional distribution of $\Gamma_{1,t}$ given $\{r_0 - \delta < X_{t-1} \leq r_0\}$ as $\delta \downarrow 0$. The existence of this limit follows from a result of Neveu (see [66] page 124). By stationarity, $F_2(.|r_0)$ is also the conditional distribution of $\Gamma_{2,2} = 2X_2X_1(\beta_0 - \alpha_0) + X_1^2(\alpha_0^2 - \beta_0^2)$ given $X_1 = r_0^+$.
We define a two-sided compound Poisson process (CPP) $(\phi(s))_{s\in\mathbb{R}}$ as follows:

$$\phi(s) = \phi_1(-s)\mathbf{1}_{\{s<0\}} + \phi_2(s)\mathbf{1}_{\{s\geq 0\}},$$

where $\{\phi_1(s), s \geq 0\}$ and $\{\phi_2(s), s \geq 0\}$ are two independent Poisson processes with $\phi_1(0) = \phi_2(0) = 0$ a.s., with the same jump rate $\pi(r_0) > 0$, where $\pi(x)$ is the density of X_1.
As soon as we have proved that ϕ_n converges to the two sided compound poisson process ϕ in the Skorohod space, we use Theorem 3.1 of [76]. Then it exists a unique random interval $[M_-, M_+]$ on which the process ϕ attains its global minimum a.s. and then $n(\hat{r}_n - r_0)$ converges to M_-.
Now we can state our convergence result but we need an additional mixing assumption of the process X.

(H6) There exists a real a with $\nu/(2+\nu) < a < 1$, a constant C and a real $0 < \beta < 1$ such that for any $u, r \in I$ we have

$$\lim_{n\to\infty} \sum_{k=1}^{n} \sum_{j;|j-k|\leq n^a, j\neq k} \mathbb{E}\Big(\mathbf{1}_{\{r<X_{k-1}\leq r+1/n\}}\mathbf{1}_{\{r<X_{j-1}\leq r+1/n\}}\Big) = 0 \quad \text{and} \tag{5.1.14}$$

$$\sum_{k=1}^{n} \sum_{j;|j-k|\leq n^a, j\neq k} \mathbb{E}\left(\mathbf{1}_{\{r_0+r/n<X_{j-1}\leq r_0+u/n\}}\mathbf{1}_{\{r_0+r/n<X_{k-1}\leq r_0+u/n\}}\right) \leq C(u-r)^{\beta} . \tag{5.1.15}$$

Let us make few comments about this assumption. By stationarity, (5.1.14) is equivalent to

$$\lim_{n\to\infty} n \sum_{h=1}^{n^a} \mathbb{E}\Big(\mathbf{1}_{\{r<X_1\leq r+1/n\}}\mathbf{1}_{\{r<X_{h+1}\leq r+1/n\}}\Big) = 0 . \tag{5.1.16}$$

This is a local mixing assumption and is clearly satisfied in the independent case and if **(H2)** is satisfied. Indeed one may write that

$$n \sum_{h=1}^{n^a} \mathbb{E}\Big(\mathbf{1}_{\{r<X_1\leq r+1/n\}}\mathbf{1}_{\{r<X_{h+1}\leq r+1/n\}}\Big) \leq C \times n \times n^a \times \frac{1}{n^2}$$

and this tends to 0 as $n \to \infty$ since $a < 1$.
In the context of [24, 60], it is deduced from a conditional argument and the Markovian context which implies that the process is geometrically mixing.
In the same way, (5.1.16) is equivalent to

$$n \sum_{h=1}^{n^a} \mathbb{E}(\mathbf{1}_{\{r_0+r/n<X_1\leq r_0+u/n\}}\mathbf{1}_{\{r_0+r/n<X_{h+1}\leq r_0+u/n\}}) \leq C(u-r)^{\beta} . \tag{5.1.17}$$

Once again, under independence, the above condition is clearly satisfied. One can also check that if we assume that for any $h > 1$, the random vector (X_1, X_h) has a continuous density, then (5.1.17) holds true with $\beta = 2$.
Hence Assumption **(H6)** is a technical assumption but we strength the fact that this condition is written in the same spirit of Assumption (II) in [9] so it is quite natural in our non Markovian context.
Now we can state our other main result as follows:

Theorem 5.1.4.
*We suppose that Assumptions **(H1)** to **(H6)** hold and that the density π is Lipschitz. Then $n(\hat{r}_n - r_0) \to M_-$ and $n(\hat{r}_n - r_0)$ is asymptotically independent of $\sqrt{n}(\hat{\alpha}_n(r_0) - \alpha_0, \hat{\beta}_n(r_0) - \beta_0)'$ which is always asymptotically normally distributed (regardless of whether r_0 is known or not).*

Before turning to the proof of this result, we briefly indicate how we can simulate the distribution of M_- (see [60] and references therein for further details). We know that two factors determine the distribution of M_-, that is the jump distributions $F_1(./r_0)$ and $F_2(./r_0)$. We can simulate M_- via simulating the compound Poisson process on the interval $[-T, T]$ for any given $T > 0$ large enough since the expectations of the jumps are positives. Modifying algorithm 6.2 pp. 183 in [27] for one-sided compound Poisson process, we have an algorithm for a two-sided compound Poisson process as follows:

- Step 1. Generate two i.i.d Poisson random variables N_1 and N_2 with the parameter $\pi(r_0)T$ as the total number of jumps on the intervals $[-T, 0]$ and $[T, 0]$, respectively.
- Step 2. Given N_1 and N_2, generate $\{U_1, \ldots, U_{N_1}\}$ and $\{V_1, \ldots, V_{N_2}\}$ as two independent jump time sequences, where $U_i \sim U[-T, 0]$, *i.i.d.* and $V_i \sim U[0, T]$, *i.i.d.*. Here $U[a, b]$, denotes the uniform distribution on the interval $[a, b]$.

- Step 3. Given N_1 and N_2, generate $\{Y_1, \ldots, Y_{N_1}\}$ and $\{Z_1, \ldots, Z_{N_2}\}$ as two independent jump-size sequences from $F_1(./r_0)$ and $F_2(./r_0)$, respectively.

For $s \in [-T, T]$, with $T > 0$ large enough, the trajectory of (5.2.1) is given by

$$\phi(s) = I(s < 0) \sum_{i=1}^{N_1} I(U_i > s) Y_i + I(s \geq 0)) \sum_{j=1}^{N_2} I(U_i > s) Z_j.$$

Then, we take the smallest minimizer of $\phi(s)$ on $[-T, T]$ as M_-. By repeating the above algorithm B times and using the nonparametric kernel method, we can get the density of M_- numerically.
Now we present in the following section, the arguments that lead us to the proof of Theorem 5.1.4.

5.2 Proof of Theorem 5.1.4

As mentioned before, the proof follows the argument of [60]. We will point out the main difference due to our context. Some intermediary results will be proved in Subsection 5.4.3.
First of all we remind that on the Skorohod space $\mathbb{D}(\mathbb{R})$, one uses the metric $d(.,.)$ defined as $d(x,y) = \sum_{k=1}^{\infty} 2^{-k} \min(1, d_k(x,y))$ for $x, y \in \mathbb{D}(\mathbb{R})$, where $d_k(.,.)$ is the Skorohod metric on $\mathbb{D}([-k,k])$ (see Section 16 of [12] for further details).
The proof of the following Lemma is given in Subsection 5.4.3.

Lemma 5.2.1. *Under the assumptions of Theorem 5.1.4, $d(\tilde{\phi}_n, \phi_n)$ converges to 0 in probability.*

Thanks to the above Lemma and Theorem 3.1 in [12], $\tilde{\phi}_n$ will converges to ϕ in $\mathbb{D}(\mathbb{R})$ as soon as ϕ_n converges weakly to ϕ. So we study the asymptotic behaviour of the sequence of processes $(\phi_n)_{n\geq 1}$ in the Skorohod space. For this purpose we consider the truncated process $(\phi_n^M(s))_{s\in\mathbb{R}}$ defined by

$$\phi_n^M(s) = \sum_{t=1}^{n} \zeta_{1,t}^M(s) \mathbf{1}_{\{s<0\}} + \sum_{t=1}^{n} \zeta_{2,t}^M(s) \mathbf{1}_{\{s\geq 0\}} \tag{5.2.1}$$

with

$$\zeta_{1,t}^M(s) = \chi_M\Big(2X_tX_{t-1}(\alpha_0 - \beta_0) + X_{t-1}^2(\beta_0^2 - \alpha_0^2)\Big)\mathbf{1}_{\{r_0 + s/n < X_{t-1} \leq r_0\}} \tag{5.2.2}$$

$$\zeta_{2,t}^M(s) = \chi_M\Big(2X_tX_{t-1}(\beta_0 - \alpha_0) + X_{t-1}^2(\alpha_0^2 - \beta_0^2)\Big)\mathbf{1}_{\{r_0 < X_{t-1} \leq r_0 + s/n\}} \tag{5.2.3}$$

and $\chi_M(x) = x\mathbf{1}_{\{|x|\leq M\}}$.
We will only deal with the case of positive times.
We denote

$$\Gamma_t^M = \chi_M\big(2X_tX_{t-1}(\beta_0 - \alpha_0) + X_{t-1}^2(\alpha_0^2 - \beta_0^2)\big)\ .$$

Remark that Γ_t^M is bounded by M and that $\Gamma_t^M = \chi_M(\Gamma_{2,t})$.
We will prove that for each $M > 0$, $(\phi_n^M(\cdot))_{n\geq 1}$ converges weakly to a two sided compound Poisson process. The convergence in the Skorohod space will be a consequence of two properties: the convergence of finite dimensional distributions and the tightness of $(\phi_n^M(\cdot))_{n\geq 1}$. First we prove that finite dimensional distributions converge.

At this point we do not follow exactly the method from [60]. Indeed, the arguments used in [24,60] and all the works related to these papers are mainly based on the Markov property of the process X and on the exponential convergence rate to its invariant distribution. Unfortunately, we can no more use this kind of arguments and we have to replace them in our context of process which is $\alpha-$mixing.
We prove the result for positive times $s \geq 0$ (the negative times can be treated exactly in the same way).
We remark that

$$\phi_n^M(s) = \sum_{t=1}^{n} \chi_M \Big(2X_t X_{t-1}(\beta_0 - \alpha_0) + X_{t-1}^2(\alpha_0^2 - \beta_0^2) \Big) \mathbf{1}_{\{r_0 < X_{t-1} \leq r_0 + s/n\}}$$

and so it is a triangular array and since the process X is mixing, there is a dependence structure that makes the convergence to a compoud Poisson process less usual. So far as we know, there are no result about compound Poisson approximation with such a dependence structure but hopefully we will be able to adapt the techniques of [26]. In this paper the author studied the convergence of sums over triangular arrays with a certain dependence structure. They applied their result in a Markovian framework but we will be able to apply and adapt their methodology when we only suppose the mixing properties **(H5)** and **(H6)** on the process X. This result is new to our knowledge and we think that this technique can be applied in other interesting problems.

Lemma 5.2.2. *Under the assumptions of Theorem 5.1.4, the finite dimensional distribution of $(\phi_n^M(s))_{s\geq 0}$ converges to those of the compound Poisson process $(\phi^M(s))_{s\geq 0}$ defined as*

$$\phi^M(s) = \sum_{i=1}^{N^M(s)} Y_i^M$$

where $(N^M(s))_{s\geq 0}$ is a Poisson process with jump rate $\pi(r_0)$ and $(Y_i^M)_{i\geq 1}$ is an i.i.d. sequence with distribution $\mathbb{Q}^M$ where $\mathbb{Q}^M$ is the distribution induced by the law of Γ_2^M given $X_1 = r_0^+$.

We specify the following fact. The measure $\mathbb{Q}^M$ is induced by the distribution of Γ_2^M given $X_1 = r_0^+$. By stationarity, it is also the measure induced by Γ_t^M given $X_{t-1} = r_0^+$. This measure exists and is the limiting conditional distribution of Γ_2^M given $\{r_0 < X_1 \leq r_0 + \delta\}$ as $\delta \downarrow 0$. The existence of this limit follows from a result of Neveu (see [66] page 124).We follow and adapt the proof of Theorem 1.1 from [26].

Proof. The proof is quite long so it is divided in several steps. A generic constant is denoted by C and may change from line to line all along this proof. We have to study the limit of $\sum_{i=1}^K \lambda_i \phi_n^M(s_i)$ for any $K \geq 1$, for any times $0 \leq s_1 < \cdots < s_K$ and any reals $\lambda_1, \ldots, \lambda_K$. We will only deal with the linear combination of the increments of ϕ_n^M

$$\begin{aligned} S_n &= c_1(\phi_n^M(s_2) - \phi_n^M(s_1)) + c_2(\phi_n^M(s_4) - \phi_n^M(s_3)) \\ &= \sum_{t=1}^{n} \Big[c_1 \Gamma_t^M \mathbf{1}_{\{r_0 + s_1/n < X_{t-1} \leq r_0 + s_2/n\}} + c_2 \Gamma_t^M \mathbf{1}_{\{r_0 + s_3/n < X_{t-1} \leq r_0 + s_4/n\}} \Big] \\ &:= \sum_{t=1}^{n} Y_{n,t} \ , \end{aligned}$$

for any $0 \leq s_1 \leq s_2 < s_3 \leq s_4 \leq T$ and any real numbers c_1 and c_2. The general case can be easily deduced from this.

Step 1: preliminaries
The characteristic function of the linear combination $c_1(\phi^M(s_2)-\phi^M(s_1))+c_2(\phi^M(s_4)-\phi^M(s_3))$ of the independent increments of the compound Poisson process ϕ^M is given for any $\xi\in\mathbb{R}$ by

$$\Psi(\xi)=\exp\left\{\pi(r_0)(s_2-s_1)(\varphi(c_1\xi)-1)\right\}\times\exp\left\{\pi(r_0)(s_4-s_3)(\varphi(c_2\xi)-1)\right\}$$

where φ is the characteristic function of $\mathbb{Q}^M$. It solves the initial value problem

$$\Psi'(\xi)=\left[c_1\pi(r_0)(s_2-s_1)\varphi'(c_1\xi)+c_2\pi(r_0)(s_4-s_3)\varphi'(c_2\xi)\right]\Psi(\xi)$$

with $\Psi(0)=0$. Moreover, since $\mathbb{E}|S_n|<\infty$,

$$\Psi_n(\xi)=\mathbb{E}(e^{i\xi S_n})=\mathbb{E}\left(\exp\left\{i\xi\sum_{t=1}^{n}Y_{n,t}\right\}\right)$$

is continuously differentiable and $\Delta_n=\Psi-\Psi_n$ satisfies for $\xi\in\mathbb{R}$:

$$\Delta_n'(\xi)=\left[c_1\pi(r_0)(s_2-s_1)\varphi'(c_1\xi)+c_2\pi(r_0)(s_4-s_3)\varphi'(c_2\xi)\right]\Delta_n(\xi)+r_n(\xi)$$

with initial condition $\Delta_n(0)=0$ and where

$$r_n(\xi)=\left[c_1\pi(r_0)(s_2-s_1)\varphi'(c_1\xi)+c_2\pi(r_0)(s_4-s_3)\varphi'(c_2\xi)\right]\Psi_n(\xi)-\Psi_n'(\xi)\ .$$

We denote

$$\bar{\varphi}(\xi)=\pi(r_0)(s_2-s_1)\varphi(c_1\xi)+\pi(r_0)(s_4-s_3)\varphi(c_2\xi)$$

and we obtain the expression of $\Delta_n(\xi)$:

$$\Delta_n(\xi)=\int_0^{\xi}\exp\left\{\bar{\varphi}(\xi)-\bar{\varphi}(z)\right\}r_n(z)dz\ ,$$

and since $\bar{\varphi}$ is bounded, it follows that there exists a constant C such that

$$|\Delta_n(\xi)|\leq e^C\int_0^{\xi}|r_n(z)|dz\ .$$

So we will obtain the convergence in law of $(S_n)_{n\geq1}$ as soon as we prove that $\sup_{z\geq0}|r_n(z)|\leq c_n$ with $(c_n)_{n\geq1}$ a sequence of positive numbers that tends to 0.
Step 2: decomposition of r_n
First, with $\Phi_{n,k}(\xi)=\mathbb{E}(e^{i\xi Y_{n,k}})$ we write

$$r_n=\left(\bar{\varphi}'\Psi_n-\Psi_n\sum_{k=1}^{n}\Phi_{n,k}'\right)+\left(\Psi_n\sum_{k=1}^{n}\Phi_{n,k}'-\Psi_n'\right):=A_n^1+A_n^2 \tag{5.2.4}$$

Arguing as in [26] on may write that

$$\begin{aligned}\Psi_n'(\xi)&=\frac{d}{d\xi}(\mathbb{E}(e^{i\xi S_n}))\\&=\mathbb{E}\left(\frac{d}{d\xi}e^{i\xi\sum_{k=1}^{n}Y_{n,k}}\right)\\&=\sum_{k=1}^{n}\mathbb{E}\left(iY_{n,k}e^{i\xi\sum_{j=1}^{n}Y_{n,k}}\right)\end{aligned}$$

$$= \sum_{k=1}^{n} \mathbb{E}\left(iY_{n,k} e^{i\xi Y_{n,k}} e^{i\xi \sum_{j=1;j\neq k}^{n} Y_{n,k}} \right)$$
$$= J_n^1(\xi) + J_n^2(\xi)$$

where

$$J_n^1(\xi) = \sum_{k=1}^{n} \mathbb{E}\left(iY_{n,k} e^{i\xi Y_{n,k}} \exp\left(i\xi \sum_{j;|j-k|>n^a} Y_{n,j} \right) \right)$$
$$J_n^2(\xi) = \sum_{k=1}^{n} \mathbb{E}\left(iY_{n,k} e^{i\xi Y_{n,k}} \left[\exp\left\{ i\xi \sum_{j;j\neq k} Y_{n,j} \right\} - \exp\left\{ i\xi \sum_{j;|j-k|>n^a} Y_{n,j} \right\} \right] \right)$$

with a real $0 < a < 1$ to be fixed later. Reporting the above notations in (5.2.4) we obtain

$$A_n^2 = \Psi_n \sum_{k=1}^{n} \Phi'_{n,k} - J_n^1 - J_n^2 \ . \tag{5.2.5}$$

For $1 \leq k \leq n$ we have

$$\begin{aligned}
&\left| \mathbb{E}\left(iY_{n,k} e^{i\xi Y_{n,k}} \exp\left(i\xi \textstyle\sum_{j;|j-k|>n^a} Y_{n,j} \right) \right) - \Phi'_{n,k}(\xi)\Psi_n(\xi) \right| \\
&\quad = \left| \mathbb{E}\left(iY_{n,k} e^{i\xi Y_{n,k}} \exp\left(i\xi \textstyle\sum_{j;|j-k|>n^a} Y_{n,j} \right) \right) - \mathbb{E}(iY_{n,k} e^{i\xi Y_{n,k}}) \mathbb{E} e^{i\xi \sum_{j=1}^{n} Y_{n,j}} \right| \\
&\quad \leq \left| \mathbb{E}\left(Y_{n,k} e^{i\xi Y_{n,k}} \exp\left(i\xi \textstyle\sum_{j;|j-k|>n^a} Y_{n,j} \right) \right) - \mathbb{E}(Y_{n,k} e^{i\xi Y_{n,k}}) \mathbb{E} \exp\left(i\xi \textstyle\sum_{j;|j-k|>n^a} Y_{n,j} \right) \right| \\
&\qquad + \left| \mathbb{E}(Y_{n,k} e^{i\xi Y_{n,k}}) \right| \left| \mathbb{E} \exp\left(i\xi \textstyle\sum_{j;|j-k|>n^a} Y_{n,j} \right) - \mathbb{E} \exp\left(i\xi \textstyle\sum_{j=1}^{n} Y_{n,j} \right) \right| \\
&\quad \leq R_{n,k}^1(\xi) + R_{n,k}^2(\xi)
\end{aligned}$$

with obvious notations . We deduce that

$$\begin{aligned}
|A_n^2(\xi)| &\leq \left| \Psi_n \sum_{k=1}^{n} \Phi'_{n,k} - J_n^1 \right| + |J_n^2(\xi)| \\
&\leq \sum_{k=1}^{n} R_{n,k}^1(\xi) + \sum_{k=1}^{n} R_{n,k}^2(\xi) + |J_n^2(\xi)| \ .
\end{aligned}$$

We finally obtain the following decomposition for r_n:

$$|r_n(\xi)| \leq \left| \bar{\varphi}' \Psi_n - \Psi_n \sum_{k=1}^{n} \Phi'_{n,k} \right| + \sum_{k=1}^{n} R_{n,k}^1(\xi) + \sum_{k=1}^{n} R_{n,k}^2(\xi) + |J_n^2(\xi)| := \sum_{l=1}^{4} r_n^l(\xi) \ . \tag{5.2.6}$$

Now we prove in the following steps that each of the four terms in the decomposition of r_n tends to to 0.

Step 3: convergence of r_n^4 (and r_n^3)

We use the fact that for any $u, v \in \mathbb{R}$, $|e^{iu} - e^{i(u+v)}| \leq 2\mathbf{1}_{v\neq 0}$. Then we have

$$\begin{aligned}
r_n^4(\xi) &\leq \sum_{k=1}^{n} \mathbb{E}|Y_{n,k}| \left| \exp\left(i\xi \sum_{j;j\neq k} Y_{n,j} \right) - \exp\left(i\xi \sum_{j;|j-k|>n^a} Y_{n,j} \right) \right| \\
&\leq 2 \sum_{k=1}^{n} \mathbb{E}|Y_{n,k}| \mathbf{1}_{\sum_{j;|j-k|\leq n^a, j\neq k} Y_{n,j} \neq 0}
\end{aligned}$$

$$\leq 2\sum_{k=1}^{n}\mathbb{E}|Y_{n,k}|\sum_{j;|j-k|\leq n^a, j\neq k}\mathbf{1}_{Y_{n,j}\neq 0}\ .$$

Since Γ_k^M is bounded, $|Y_{n,k}| \leq C(\mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} + \mathbf{1}_{\{r_0+s_3/n<X_{k-1}\leq r_0+s_4/n\}})$. Moreover the event $\{Y_{n,j}\neq 0\}$ is equal to $\{r_0+s_1/n<X_{j-1}\leq r_0+s_2/n\}\cup\{r_0+s_3/n<X_{j-1}\leq r_0+s_4/n\}$. Therefore

$$\begin{aligned}r_n^4(\xi) &\leq C\sum_{k=1}^{n}\sum_{j;|j-k|\leq n^a, j\neq k}\mathbb{E}\Big([\mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}}+\mathbf{1}_{\{r_0+s_3/n<X_{k-1}\leq r_0+s_4/n\}}]\\ &\qquad\times\mathbf{1}_{\{r_0+s_1/n<X_{j-1}\leq r_0+s_2/n\}\cup\{r_0+s_3/n<X_{j-1}\leq r_0+s_4/n\}}\Big)\\ &\leq 4C\sum_{k=1}^{n}\sum_{j;|j-k|\leq n^a, j\neq k}\mathbb{E}\Big(\mathbf{1}_{\{r_0<X_{k-1}\leq r_0+T/n\}}\mathbf{1}_{\{r_0<X_{j-1}\leq r_0+T/n\}}\Big)\end{aligned}$$

and this tends to 0 as $n\to\infty$ by the mixing type assumption **(H6)**. The term r_n^3 can be treated in the same manner.

Step 3: convergence of r_n^2

Remind that $r_n^2(\xi)=\sum_{k=1}^n R_{n,k}^1(\xi)$ and observe that

$$R_{n,k}^1(\xi)=\left|\mathrm{Cov}\Big(Y_{n,k}e^{i\xi Y_{n,k}};\exp\left(i\xi\textstyle\sum_{j;|j-k|>n^a}Y_{n,j}\right)\Big)\right|.$$

Thanks to the Davydov inequality one may write for $1/p+1/q+1/r=1$ that

$$R_{n,k}^1\leq\|Y_{n,k}e^{i\xi Y_{n,k}}\|_p\left\|\exp\left(i\xi\textstyle\sum_{j;|j-k|>n^a}Y_{n,j}\right)\right\|_q[\alpha_X(n^a)]^{1/r}$$

and since $\exp\left(i\xi\sum_{j;|j-k|>n^a}Y_{n,j}\right)$ is bounded, one uses the above inequality with $q=+\infty$. Moreover, by Assumption **(H4)**, $\sum_{k=0}^{\infty}\{\alpha_X(k)\}^{\frac{\nu}{2+\nu}}<\infty$ and since the sequence of mixing coefficient is decreasing, one may find C such that $k\{\alpha_X(k)\}^{\frac{\nu}{2+\nu}}\leq C$ (see [44] Exercice 3.9). Then $\alpha_X(k)\leq Ck^{-(2+\nu)/\nu}$. This yields with $1/p+1/r=1$:

$$R_{n,k}^1\leq C\|Y_{n,k}e^{i\xi Y_{n,k}}\|_p\left(\frac{1}{n}\right)^{a\frac{2+\nu}{\nu}\frac{1}{r}}$$

By Assumption **(H2)** (especially the fact that X_t has a bounded density), $\|Y_{n,k}e^{i\xi Y_{n,k}}\|_p\leq Cn^{-1/p}$ and if we choose a such that $1>a>\nu/(2+\nu)$ we obtain that

$$R_{n,k}^1\leq C\left(\frac{1}{n}\right)^{a\frac{2+\nu}{\nu}\frac{1}{r}+\frac{1}{p}}$$

and

$$a\frac{2+\nu}{\nu}\frac{1}{r}+\frac{1}{p}=1+\frac{1}{r}\left(a\frac{2+\nu}{\nu}-1\right)>1$$

and thus $r_n^2(\xi)\to 0$ as $n\to\infty$.

Step 3: convergence of r_n^1

Since $|\Psi_n|\leq 1$ one just has to prove that $\lim_{n\to\infty}|\bar{\varphi}'-\sum_{k=1}^n\Phi'_{n,k}|=0$. Remind that $\bar{\varphi}(\xi)=\pi(r_0)(s_2-s_1)\varphi(c_1\xi)+\pi(r_0)(s_4-s_3)\varphi(c_2\xi)$ where $\varphi(\xi)=\int_{\mathbb{R}}e^{i\xi u}d\mathbb{Q}^M(v)$ where $\mathbb{Q}^M$ is the measure induced by the distribution of Γ_2^M given $X_1=r_0^+$. By stationarity, it is also the measure induced by Γ_t^M given $X_{t-1}=r_0^+$. This measure exists and is the limiting

conditional distribution of Γ_2^M given $\{r_0 < X_1 \leq r_0 + \delta\}$ as $\delta \downarrow 0$. The existence of this limit follows from of result of Neveu (see [66] page 124).
We recall that

$$Y_{n,k} = c_1 \Gamma_t^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} + c_2 \Gamma_t^M \mathbf{1}_{\{r_0+s_3/n<X_{k-1}\leq r_0+s_4/n\}}$$

and since $\mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} \times \mathbf{1}_{\{r_0+s_3/n<X_{k-1}\leq r_0+s_4/n\}} = 0$, one remarks that

$$\begin{aligned}
\Phi'_{n,k}(\xi) &= \mathbb{E}(iY_{n,k} e^{i\xi Y_{n,k}}) \\
&= \mathbb{E}\left(ic_1 \Gamma_k^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} e^{ic_1\xi\Gamma_k^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}}} \right) \\
&\qquad + \mathbb{E}\left(ic_2 \Gamma_k^M \mathbf{1}_{\{r_0+s_3/n<X_{k-1}\leq r_0+s_4/n\}} e^{ic_2\xi\Gamma_k^M \mathbf{1}_{\{r_0+s_3/n<X_{k-1}\leq r_0+s_4/n\}}} \right) \\
&= \mathbb{E}\left(ic_1 \Gamma_k^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} e^{ic_1\xi\Gamma_k^M} \right) + \mathbb{E}\left(ic_2 \Gamma_k^M \mathbf{1}_{\{r_0+s_3/n<X_{k-1}\leq r_0+s_4/n\}} e^{ic_2\xi\Gamma_k^M} \right) .
\end{aligned}$$

So we only need to prove that

$$\lim_{n\to\infty} \left| \sum_{k=1}^n \mathbb{E}\left(ic_1 \Gamma_k^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} e^{ic_1\xi\Gamma_k^M} \right) - c_1 \pi(r_0)(s_2 - s_1)\varphi'(c_1\xi) \right| = 0 . \tag{5.2.7}$$

We denote $\mathbb{Q}_n^M$ the probability measure induced by the conditional distribution of Γ_2^M given $\mathbf{1}_{\{r_0+s_1/n<X_1\leq r_0+s_2/n\}} = 1$. By construction,

$$\int_{\mathbb{R}} ic_1 v e^{ic_1\xi v} d\mathbb{Q}_n^M(v) = \mathbb{E}\left(ic_1 \Gamma_2^M e^{ic_1\xi\Gamma_2^M} \,\middle|\, \mathbf{1}_{\{r_0+s_1/n<X_1\leq r_0+s_2/n\}} = 1 \right) \xrightarrow[n\to\infty]{} c_1\varphi'(c_1\xi).$$

So (5.2.7) will be a consequence of the following estimation

$$\begin{aligned}
\Pi_n := \Big| &\mathbb{E}\left(ic_1 \Gamma_k^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} e^{ic_1\xi\Gamma_k^M} \right) \\
&- c_1\pi(r_0)(s_2 - s_1)\mathbb{E}\left(i\Gamma_2^M e^{ic_1\xi\Gamma_2^M} \,\middle|\, \mathbf{1}_{\{r_0+s_1/n<X_1\leq r_0+s_2/n\}} = 1 \right) \Big| \quad \leq \frac{C}{n^2}.
\end{aligned} \tag{5.2.8}$$

We proceed as follows. Using a conditioning argument and the stationarity

$$\begin{aligned}
&\mathbb{E}\left(ic_1 \Gamma_k^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} e^{ic_1\xi\Gamma_k^M} \right) \\
&\qquad = \mathbb{E}\left(\mathbb{E}\left(ic_1 \Gamma_k^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} e^{ic_1\xi\Gamma_k^M} \,\middle|\, \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} \right) \right) \\
&\qquad = \mathbb{E}\left(c_1 \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} \mathbb{E}\left(i\Gamma_k^M e^{ic_1\xi\Gamma_k^M} \,\middle|\, \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} \right) \right) \\
&\qquad = \mathbb{E}\left(c_1 \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} \mathbb{E}\left(i\Gamma_2^M e^{ic_1\xi\Gamma_2^M} \,\middle|\, \mathbf{1}_{\{r_0+s_1/n<X_1\leq r_0+s_2/n\}} \right) \right) .
\end{aligned}$$

If we denote $N(Z; d\gamma)$ the conditional kernel of Γ_2^M given the random variable $Z = \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}}$ one obtains

$$\begin{aligned}
&\mathbb{E}\left(ic_1 \Gamma_k^M \mathbf{1}_{\{r_0+s_1/n<X_{k-1}\leq r_0+s_2/n\}} e^{ic_1\xi\Gamma_k^M} \right) \\
&\quad = \int_{\mathbb{R}\times\mathbb{R}} c_1 z \mathbb{E}(i\gamma e^{i\xi\gamma} | Z = z) N(z; d\gamma) \\
&\quad = c_1 \mathbb{P}(Z = 1) \int_{\mathbb{R}} \mathbb{E}(i\gamma e^{i\xi\gamma} | Z = 1) N(z; d\gamma) + 0
\end{aligned}$$

$$= c_1\mathbb{P}(r_0 + s_1/n < X_{k-1} \leq r_0 + s_2/n) \times \mathbb{E}(i\gamma e^{i\xi\gamma}|Z = 1)$$
$$= c_1\mathbb{P}(r_0 + s_1/n < X_{k-1} \leq r_0 + s_2/n) \times \mathbb{E}\left(i\Gamma_2^M e^{ic_1\xi\Gamma_2^M}\Big|\mathbf{1}_{\{r_0+s_1/n<X_1\leq r_0+s_2/n\}} = 1\right).$$

Now we return to the proof of (5.2.8). Since the density π is Lipschitz, we have

$$\begin{aligned}\Pi_n &\leq |c_1|\left|\mathbb{P}(r_0 + s_1/n < X_{k-1} \leq r_0 + s_2/n) - \pi(r_0)\frac{(s_2 - s_1)}{n}\right| \\ &\qquad \times \left|\mathbb{E}\left(i\Gamma_2^M e^{ic_1\xi\Gamma_2^M}\Big|\mathbf{1}_{\{r_0+s_1/n<X_1\leq r_0+s_2/n\}} = 1\right)\right| \\ &\leq c_1\int_{r_0+s_1/n}^{r_0+s_2/n}|\pi(x) - \pi(r_0)|dx \leq \frac{C}{n^2}\end{aligned}$$

and (5.2.8) is proved and the convergence of r_n^1 to 0 follows. □

Now we prove the tightness. One has to estimate some moments of the increments of the process and one has to adapt some technical arguments form [12].

Lemma 5.2.3. *Under the assumptions of Theorem 5.1.4, the sequence $(\phi_n^M(s))_{s\in\mathbb{R}}$ is tight in the Skorohod space.*

Proof. First we recall that for $r < s < u \leq T$ (we restrict ourselves to positive times for simplicity)

$$\phi_n^M(s) - \phi_n^M(r) = \sum_{t=1}^{n}\Gamma_t^M\,\mathbf{1}_{\{r_0+r/n<X_{t-1}\leq r_0+s/n\}}$$

where Γ_t^M is bounded by M. We may write that

$$\begin{aligned}&\mathbb{E}\left(|\phi_n^M(s) - \phi_n^M(r)||\phi_n^M(u) - \phi_n^M(s)|\right) \\ &\qquad \leq 2\sum_{t=1}^{n}\sum_{t'=1}^{t-1}\mathbb{E}\left(|\Gamma_t^M||\Gamma_{t'}^M|\mathbf{1}_{\{r_0+r/n<X_{t-1}\leq r_0+s/n\}}\mathbf{1}_{\{r_0+s/n<X_{t'-1}\leq r_0+u/n\}}\right) \\ &\qquad \leq 2M^2\sum_{t=1}^{n}\sum_{t'=1}^{t-1}\mathbb{E}\left(\mathbf{1}_{\{r_0+r/n<X_{t-1}\leq r_0+u/n\}}\mathbf{1}_{\{r_0+r/n<X_{t'-1}\leq r_0+u/n\}}\right) \\ &\qquad \leq 2M^2\sum_{t=1}^{n}\sum_{t'=1}^{t-1}|\mathrm{Cov}(\mathbf{1}_{\{r_0+r/n<X_{t-1}\leq r_0+u/n\}};\mathbf{1}_{\{r_0+r/n<X_{t'-1}\leq r_0+u/n\}})| \\ &\qquad\qquad + \mathbb{E}(\mathbf{1}_{\{r_0+r/n<X_{t-1}\leq r_0+u/n\}})\mathbb{E}(\mathbf{1}_{\{r_0+r/n<X_{t'-1}\leq r_0+u/n\}}) \\ &\qquad \leq C(u-r)^2 + 2M^2\sum_{t=1}^{n}\sum_{t'=1}^{t-1}|\mathrm{Cov}(\mathbf{1}_{\{r_0+r/n<X_{t-1}\leq r_0+u/n\}};\mathbf{1}_{\{r_0+r/n<X_{t'-1}\leq r_0+u/n\}})| \\ &\qquad \leq C(u-r)^2 + 2M^2\sum_{h=1}^{n}n|\mathrm{Cov}(\mathbf{1}_{\{r_0+r/n<X_0\leq r_0+u/n\}};\mathbf{1}_{\{r_0+r/n<X_h\leq r_0+u/n\}})|\,.\end{aligned} \tag{5.2.9}$$

We strength the fact that in [24] or [60], the authors can prove that

$$\mathbb{P}\left(\{r_0 + r/n < X_{t-1} \leq r_0 + t/n\} \cap \{r_0 + r/n < X_{t'-1} \leq r_0 + t/n\}\right) \leq C(t-r)^2/n^2$$

because they used a conditional argument due to the Markovian context. This is no more possible in our case. Thus we use the Davydov inequality and an argument that we used in the previous proof. More precisely we recall that by Assumption **(H4)**,

$\sum_{k=0}^{\infty}\{\alpha_X(k)\}^{\frac{\nu}{2+\nu}} < \infty$ and since the sequence of mixing coefficient is decreasing, one may find C such that $k\{\alpha_X(k)\}^{\frac{\nu}{2+\nu}} \leq C$ (see [44] Exercice 3.9). Then $\alpha_X(k) \leq Ck^{-(2+\nu)/\nu}$. With the real a from Assumption **(H6)**, we introduce the real p defined by

$$p = \frac{2a(2+\nu) - 2\nu}{a(2+\nu) - 2\nu} \ .$$

We notice that $p > 2$ and we may use the Davydov inequality in order to obtain that

$$\begin{aligned} n \sum_{h=n^a}^{n} |\mathrm{Cov}(\mathbf{1}_{\{r_0+r/n<X_0\leq r_0+u/n\}};&\mathbf{1}_{\{r_0+r/n<X_h\leq r_0+u/n\}})| \\ &\leq Cn \sum_{h=n^a}^{n} \|\mathbf{1}_{\{r_0+r/n<X_0\leq r_0+u/n\}}\|_{\mathbb{L}^p}^2 (\alpha_X(n^a))^{(p-2)/p} \\ &\leq Cn^2 \left(\frac{u-r}{n}\right)^{\frac{2}{p}} n^{-a\frac{2+\nu}{\nu}\frac{p-2}{p}} \\ &\leq C(u-r)^{\frac{2}{p}} \end{aligned} \tag{5.2.10}$$

where we have used the Assumption **(H2)** on the boundedness of the density of the stationary process X.
Moreover, Assumption **(H6)** implies that

$$n \sum_{h=1}^{n^a} |\mathrm{Cov}(\mathbf{1}_{\{r_0+r/n<X_0\leq r_0+u/n\}};\mathbf{1}_{\{r_0+r/n<X_h\leq r_0+u/n\}})| \leq C(u-r)^{\beta} \ . \tag{5.2.11}$$

Reporting (5.2.10) and (5.2.11) into (5.2.9), we finally obtain (with $\tilde{\beta} = \beta \wedge (2/p)$)

$$\mathbb{E}\left(|\phi_n^M(s) - \phi_n^M(r)||\phi_n^M(u) - \phi_n^M(s)|\right) \leq C(u-r)^{\tilde{\beta}} \ . \tag{5.2.12}$$

The estimation (5.2.12) looks like (13.14) page 143 in [12] that implies the tightness condition (13.13) in Theorem 13.15 from [12]. Unfortunately it is not exactly the same, especially from the fact that $\tilde{\beta} < 1$. So we have to adapt the arguments from [12]. This can be done in the following way. The arguments in the proof of Theorem 13.15 can be repeated but one uses, instead of Theorem 10.3, the following trick. We follow the proof of Theorem 10.3, case 1, but with the set $\{i/k\}_{0\leq i\leq k}$ instead of the set of dyadic rationals. With the notations of [12] that we do not repeat here, one obtains that for a $0 < \theta < 1$ (see the top of page 110)

$$\begin{aligned} \mathbb{P}(L(\phi_n^M - \phi_n) \geq \lambda) &\leq \sum_{k=1}^{\infty}\sum_{i=1}^{k-1} \mathbb{P}\left[m\left(\frac{i-1}{k}, \frac{i}{k}, \frac{i+1}{k}\right) \geq C\lambda\theta^k\right] \\ &\leq C\sum_{k=1}^{\infty} k\frac{1}{\lambda\theta^k}\frac{1}{k^{1/a}} \end{aligned}$$

and since the above series is convergent (remind that $0 < \theta < 1$) the arguments from [12] are still valid. □

Now we can end the proof of Theorem 5.1.4.

Proof. By Lemma 5.2.2 and Lemma 5.2.3, it follows that for any $M > 0$, ϕ_n^M converges to ϕ^M in $\mathbb{D}(\mathbb{R})$. Using Therorem 3.2 in [12], if

$$\lim_{M\to\infty} \phi^M = \phi \quad \text{in } \mathbb{D}(\mathbb{R}) \tag{5.2.13}$$

and

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}(d(\phi_n^M,\phi_n) > \eta) = 0 \quad \text{for any } \eta > 0, \tag{5.2.14}$$

then ϕ_n converges to ϕ in the Skorohod space.
The convergence (5.2.13) follows from the fact that the measures $\mathbb{Q}^M$ converges to $\mathbb{Q}$ as $M \to \infty$ and this is easy to prove that the process ϕ^M (which is a compound Poisson process) converges to the compound Poisson process ϕ using the Theorem 16 page 134 in [71]. This theorem can be applied easily because the compound Poisson processes are Levy processes and then the Aldous condition (see [1]) for convergence in the Skorohod space is satisfied.
Moreover, one remarks that for any $\eta > 0$, for M sufficiently big, we have

$$\begin{aligned}
\mathbb{P}\left(\sup_{s\le T} |\phi_n(s) - \phi_n^M(s)| > \eta\right) &\le \mathbb{P}\left(\sum_{t=1}^{n} |\Gamma_t| \mathbf{1}_{|\Gamma_t|\ge M} \mathbf{1}_{\{r_0 < X_{t-1} \le r_0 + T/n\}} > \eta\right) \\
&\le \mathbb{P}\left(\bigcup_{t=1}^{n} \{|\Gamma_t| \ge M\} \cap \{r_0 < X_{t-1} \le r_0 + T/n\}\right) \\
&\le n\mathbb{P}\big(\{|\Gamma_1| \ge M\} \cap \{r_0 < X_0 \le r_0 + T/n\}\big).
\end{aligned}$$

Since $\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}(\{|\Gamma_1| \ge M\} \cap \{r_0 < X_0 \le r_0 + T/n\}) = 0$, (5.2.14) holds true.
The remaining arguments are the same from those given by [24, 60]. □

5.3 Simulation studies

In this section, we simulate a TAR model with different noise processes;

Type I: the white noise Gaussian process.

Type II: the process $\varepsilon_t = \eta_{t-1}\eta_t$, where η_t is i.i.d standard Gaussian process.

Type III: the process $\varepsilon_t = \eta_{t-1}\eta_t^2$, where η_t is i.i.d standard Gaussian process.

Type IV: the process $\varepsilon_t = (|\eta_{t-1}| + 1)^{-1}\eta_t$, where η_t is i.i.d standard Gaussian process.

We stregnth the fact that only the third noise is not a martingale difference sequence.
Simulation of TAR model with white noise Gaussian process (noise of Type I)
We simulate the following two regimes TAR model:

$$X_t = \begin{cases} -0.5 \times X_{t-1} + \varepsilon_t, \ for\ X_{t-1} \le 0.4 \\ 0.9 \times X_{t-1} + \varepsilon_t, \ for\ X_{t-1} > 0.4 \end{cases} \tag{5.3.1}$$

To have the performance of the Least Squares Estimation of $\theta_0 = (-0.5, 0.9, 0.4)$ in finite samples. We made M=1000 replications of (5.3.1) in each sample. We use sample sizes n = 600, n=1200, n=1500 and n=2000 and we suppose that $\varepsilon_t \sim$ i.i.d $\mathcal{N}(0, 0.8)$. For estimating the parameters, we use the function tar lies in the package TSA of software R.
We summarize bias results in the following table.

Table 1. Table of bias.

n/*parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.09472	0.05747	0.00924
1200	0.01394	0.00886	0.00596
1500	0.02416	0.01094	0.01494
2000	0.00204	0.00098	0.00078

Let $\hat{\alpha}(\hat{r})$, $\hat{\beta}(\hat{r})$ and $\hat{r}$ be estimators of α_0, β_0 and r_0 respectively. We calculate the standard deviations these parameters as following:

$$std(\hat{\alpha}(\hat{r})) = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}((\hat{\alpha}_j(\hat{r}) - \overline{\hat{\alpha}}))^2}.$$

$$std(\hat{\beta}(\hat{r})) = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}((\hat{\beta}_j(\hat{r}) - \overline{\hat{\beta}}))^2}.$$

$$std(\hat{r}) = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(\hat{r}_j - \overline{\hat{r}}))^2}.$$

We summarize standard deviations results in the following table.

Table 2. Table of standard deviations.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.24479	0.04578	0.06168
1200	0.07202	0.03252	0.03116
1500	0.08035	0.01623	0.02226
2000	0.07351	0.01518	0.01806

For calculating the asymptotic standard deviations of $\hat{\alpha}$ and $\hat{\beta}$, we use the theorem 3.2 of [60]. We summarize the results in the following table.

Table 3. Table of asymptotic standard deviations.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$
600	0.10493	0.02344
1200	0.07712	0.01464
1500	0.07036	0.01306
2000	0.06064	0.01261

Simulation of TAR model with noise of Type II
We simulate the following two regimes TAR model:

$$X_t = \begin{cases} -0.5 \times X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} \leq 0.4 \\ 0.9 \times X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} > 0.4 \end{cases} \tag{5.3.2}$$

$\varepsilon_t = \eta_{t-1}\eta_t$, where η_t is standard Gaussian process.

Here, the noise is martingale difference and as before we establish two tables, the first concerns the bias and the second concerns the empirical standard deviation. We add another table which summarizes the asymptotic standard deviation. We made $M = 1000$ replications of (5.3.2).

Table 4. Table of bias.

n/*parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.00730	0.01556	0.00574
1200	0.00573	0.00991	0.03819
1500	0.00481	0.00735	0.02950
2000	0.00264	0.00667	0.01767

Table 5. Table of standard deviations.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.06104	0.02797	0.02831
1200	0.04787	0.02613	0.01114
1500	0.03450	0.02446	0.00402
2000	0.02081	0.01575	0.00356

For the asymptotic standard deviation, by using the theorem (5.1.3), we establish the following table.

Table 6. Table of asymptotic standard deviations.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$
600	0.04380	0.05016
1200	0.05153	0.03557
1500	0.02857	0.03321
2000	0.02744	0.01781

We calculate the root mean of squared errors(MSE) as follows:

$$MSE(\hat{\alpha}(\hat{r})) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{\alpha}_i(\hat{r}) - \alpha)^2}$$

$$MSE(\hat{\beta}(\hat{r})) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_i(\hat{r}) - \beta)^2}$$

$$MSE(\hat{r}) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{r}_i - r)^2}$$

Table 7. Table of the root mean squared errors.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.00033	0.00047	0.00018
1200	0.00072	0.00055	0.00058
1500	0.00057	0.00022	0.00002
2000	0.00016	0.00055	0.00003

Simulation TAR model with noise of Type III
We simulate the following two regimes TAR model:

$$X_t = \begin{cases} -0.5 \times X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} \leq 0.4 \\ 0.9 \times X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} > 0.4 \end{cases} \tag{5.3.3}$$

$\varepsilon_t = \eta_{t-1}\eta_t^2$, where η_t is standard Gaussian process.

Table 8. Table of bias.

n/*parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.00459	0.00678	0.00871
1200	0.00391	0.00578	0.00789
1500	0.00271	0.00403	0.00509
2000	0.00197	0.00072	-0.00036

Table 9. Table of standard deviations.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.09557	0.03460	0.01560
1200	0.035286	0.03639	0.00609
1500	0.04157	0.02385	0.00496
2000	0.04709	0.02468	0.00289

Table 10. Table of asymptotic standard deviations.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$
600	0.00033	0.00086
1200	0.00064	0.00108
1500	0.00037	0.00080
2000	0.00045	0.00089

Table 11. Table of the root mean squared errors.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.09795	0.03439	0.01796
1200	0.04112	0.03801	0.00578
1500	0.04095	0.02496	0.00556
2000	0.03847	0.02394	0.00323

Simulation of TAR model with noise of Type IV
We simulate the following two regimes TAR model:

$$X_t = \begin{cases} -0.5 \times X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} \leq 0.4 \\ 0.9 \times X_{t-1} + \varepsilon_t, \text{ for } X_{t-1} > 0.4 \end{cases} \tag{5.3.4}$$

$\varepsilon_t = (|\eta_{t-1}| + 1)^{-1}\eta_t$, where η_t is standard Gaussian process.

Table 11. Table of bias.

n/*parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	-0.0252	-0.00164	-0.00674
1200	-0.0003	-0.00723	-0.00393
1500	-0.00907	0.00485	-0.00202
2000	0.00277	-0.00287	0.0028

Table 12. Table of standard deviations.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.06452	0.02437	0.00999
1200	0.03080	0.02490	0.00890
1500	0.02030	0.02353	0.00430
2000	0.01153	0.01979	0.00547

Table 13. Table of asymptotic standard deviations.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$
600	0.03911	0.05145
1200	0.01726	0.02380
1500	0.01554	0.01994
2000	0.01448	0.01709

Table 14. Table of the root mean squared errors.

n/ *parameters*	α_0= -0.5	$\beta_0 = 0.9$	$r_0 = 0.4$
600	0.06618	0.02312	0.01053
1200	0.02922	0.02301	0.00943
1500	0.02733	0.02136	0.00440
2000	0.02661	0.02038	0.00317

Conclusion of simulations.
All tables summarize the bias, the empirical standard deviation and the asymptotic standard deviation. It is clearly that for the larger sample size , we obtain the closer of the empirical standard deviation and the asymptotic standard deviation on the whole. The asymptotic standard deviation of estimated parameters are computed by using the theorem (5.1.3). In other case, our theorem (5.1.4) is similar of theorem of [60]. Consequently, we left the asymptotic standard deviation of $\hat{r}_n$ (the results simulations are similar).

5.4 Proofs

5.4.1 Proof of Theorem 5.1.2

Before proving this result, we need some preliminary materials which are stated in the following lemmas and proposition.
For any $r, r_0 \in [\underline{r}, \overline{r}] := I$, we denote when $r_0 < r$

$$Q(r_0, r) = \mathbb{E}(\mathbf{1}_{r_0 < X_{t-1} \leq r}) = \int_{r_0}^{r} \pi(x)dx.$$

Lemma 5.4.1. *Suppose that the assumptions of Theorem 5.1.2 hold, then there exists some positive constants $m < M$ and C independent of n such that*

(i) $m(r - r_0) < Q(r_0, r) < M(r - r_0)$

(ii) $\mathrm{Var}(\mathbf{1}_{\{r_0 < X_{t-1} \leq r\}}) \leq Q(r_0, r)$

(iii) $\mathrm{Var}(\sum_{t=1}^{n} \mathbf{1}_{\{r_0 < X_{t-1} \leq r\}}) \leq n \sum_{k \geq 0} (\alpha(k))^{\frac{\nu}{2+\nu}} Q(r_0, r)^{\frac{2}{2+\nu}}$.

Proof. Since **(H2)** is true, one may write that there exists $m < M$ such that for any x, $m \leq \pi(x) \leq M$ and consequently (i) is true. Moreover

$$\mathrm{Var}(\mathbf{1}_{\{r_0 < X_{t-1} \leq r\}}) = \mathbb{E}(\mathbf{1}_{\{r_0 < X_{t-1} \leq r\}}) - (\mathbb{E}(\mathbf{1}_{\{r_0 < X_{t-1} \leq r\}}))^2 = Q(r_0, r)(1 - Q(r_0, r)) \leq Q(r_0, r)$$

and (ii) is also easily proved.
The third point is a litlle bit more difficult and this is the first time when the fact that the noise is no more independant interferes.

By symmetry and stationarity of the process X one has

$$\begin{aligned}\mathrm{Var}\left(\sum_{t=1}^{n}\mathbf{1}_{\{r_0<X_{t-1}\le r\}}\right) &= \sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{Cov}(\mathbf{1}_{\{r_0<X_{t-1}\le r\}},\mathbf{1}_{\{r_0<X_{s-1}\le r\}})\\ &= \sum_{k=1-n}^{n-1}(n-|k|)\mathrm{Cov}(\mathbf{1}_{\{r_0<X_t\le r\}},\mathbf{1}_{\{r_0<X_{t-k}\le r\}})\\ &= n\times\sum_{k\in\mathbb{Z}}\tfrac{n-k}{n}\mathbf{1}_{|k|\le n-1}c_k \qquad (5.4.1)\end{aligned}$$

where $c_k=\mathrm{Cov}(\mathbf{1}_{\{r_0<X_t\le r\}},\mathbf{1}_{\{r_0<X_{t-k}\le r\}})$.
The mixing coefficients will be used via the Davydov inequality (5.1.4). There exists a constant C such that

$$\begin{aligned}|c(k)| &= |\mathrm{Cov}(\mathbf{1}_{\{r_0<X_t\le r\}},\mathbf{1}_{\{r_0<X_{t-k}\le r\}})|\\ &\le C(\mathbb{E}\mathbf{1}_{\{r_0<X_t\le r\}})^{\frac{1}{2+\nu}}(\mathbb{E}\mathbf{1}_{\{r_0<X_{t-k}\le r\}})^{\frac{1}{2+\nu}}(\alpha(k))^{\frac{\nu}{2+\nu}}\\ &\le C(\mathbb{E}\mathbf{1}_{\{r_0<X_1\le r\}})^{\frac{2}{2+\nu}}(\alpha(k))^{\frac{\nu}{2+\nu}}\\ &\le CQ(r_0,r)^{\frac{2}{2+\nu}}(\alpha(k))^{\frac{\nu}{2+\nu}}. \qquad (5.4.2)\end{aligned}$$

By Assumption **(H4)**, $\sum_{k\ge0}(\alpha(k))^{\frac{\nu}{2+\nu}}<\infty$ and we deduce that there exists a constant C independent of n such that

$$\left|\sum_{k\in\mathbb{Z}}\tfrac{n-k}{n}\mathbf{1}_{|k|\le n-1}c_k\right|\le CQ(r_0,r)^{\frac{2}{2+\nu}}$$

and using this inequality into (5.4.1) yields (iii). □

Lemma 5.4.2. *Suppose that the assumptions of Theorem 5.1.2 hold. There exists a positive constant C independent of n such that for all $0<\delta<1$ and for all u, u_1, $u_2\in[0,\delta]$, with $u_1<u_2$, it holds*

$$\mathrm{Var}\left(\frac{1}{n}\sum_{t=1}^{n}(\varepsilon_tX_{t-1}-\mathbb{E}\varepsilon_tX_{t-1})\mathbf{1}_{\{r_0<X_{t-1}\le r_0+u\}}\right)\le\frac{C}{n}Q(r_0,r_0+u)^{\frac{1}{2+\nu}} \qquad (5.4.3)$$

$$\mathrm{Var}\left(\frac{1}{n}\sum_{t=1}^{n}|\varepsilon_tX_{t-1}-\mathbb{E}\varepsilon_tX_{t-1}|\mathbf{1}_{\{r_0+u_1<X_{t-1}\le r_0+u_2\}}\right)\le\frac{C}{n}(Q(0,r_0+u_2)-Q(0,r_0+u_1))^{\frac{1}{2+\nu}}. \qquad (5.4.4)$$

Proof. We only prove (5.4.4). The estimation (5.4.3) can be proved following the same arguments. We have

$$\begin{aligned}&\mathrm{Var}\left(\sum_{t=1}^{n}|\varepsilon_tX_{t-1}-\mathbb{E}\varepsilon_tX_{t-1}|\mathbf{1}_{\{r_0+u_1<X_{t-1}\le r_0+u_2\}}\right)\\ &\le 2\sum_{j=1}^{n}\sum_{k=1}^{j}\mathrm{Cov}(|\varepsilon_jX_{j-1}-\mathbb{E}\varepsilon_jX_{j-1}|\mathbf{1}_{\{r_0+u_1<X_{j-1}\le r_0+u_2\}},|\varepsilon_kX_{k-1}-\mathbb{E}\varepsilon_kX_{k-1}|\mathbf{1}_{\{r_0+u_1<X_{k-1}\le r_0+u_2\}})\\ &\le 2\sum_{j=1}^{n}\sum_{k=1}^{j}c(j,k) \qquad (5.4.5)\end{aligned}$$

with obvious notations. Using (5.1.2) with for $\theta = \theta_0$:

$$\varepsilon_t = X_t - \left(\beta_0 + (\alpha_0 - \beta_0)\mathbf{1}_{\{X_{t-1} \le r_0\}}\right) X_{t-1}$$

we obtain that $c(j,k) = \mathrm{cov}(Y_j, Y_k)$ with

$$Y_j = \Big| X_{j-1}X_j - X_{j-1}\left(\beta_0 + (\alpha_0 - \beta_0)\mathbf{1}_{\{X_{j-1} \le r_0\}}\right) \\ - \mathbb{E}\left[X_{j-1}X_j - X_{j-1}\left(\beta_0 + (\alpha_0 - \beta_0)\mathbf{1}_{\{X_{j-1} \le r_0\}}\right)\right]\Big| \mathbf{1}_{\{r_0+u_1 < X_{j-1} \le r_0+u_2\}} .$$

Thanks to the Davydov inequality, one obtains for $k \le j$

$$|c(j,k)| \le (\alpha(j-k-1))^{\frac{\nu}{2+\nu}} \|Y_j\|_{\mathbb{L}^{2+\nu}} \|Y_k\|_{\mathbb{L}^{2+\nu}} . \tag{5.4.6}$$

Since Θ is compact, there exists a constant C such that

$$Y_j \le C\left(1 + |X_j| + \mathbb{E}|X_j|\right) \mathbf{1}_{\{r_0+u_1 < X_{j-1} \le r_0+u_2\}} . \tag{5.4.7}$$

Using (5.4.7), the stationarity of the process X that admits moments of order $4+2\nu$ and Holder's inequality, one deduces that

$$\begin{aligned} \|Y_j\|_{\mathbb{L}^{2+\nu}} &= \|Y_1\|_{\mathbb{L}^{2+\nu}} \\ &\le C(1 + \|X_1\|_{\mathbb{L}^{4+2\nu}}) \|\mathbf{1}_{\{r_0+u_1 < X_0 \le r_0+u_2\}}\|_{\mathbb{L}^{4+2\nu}} \\ &\le C\big(Q(0, r_0+u_2) - Q(0, r_0+u_1)\big)^{1/(4+2\nu)} . \end{aligned}$$

Substituting the above inequality in (5.4.6) and (5.4.6) into (5.4.5) yield

$$\begin{aligned} &\mathrm{Var}\left(\sum_{t=1}^{n} |\varepsilon_t X_{t-1} - \mathbb{E}\varepsilon_t X_{t-1}| \mathbf{1}_{\{r_0+u_1 < X_{t-1} \le r_0+u_2\}}\right) \\ &\qquad \le C\big(Q(0, r_0+u_2) - Q(0, r_0+u_1)\big)^{\frac{1}{2+\nu}} \sum_{j=1}^{n}\sum_{k=1}^{j} (\alpha(j-k))^{\frac{\nu}{2+\nu}} \\ &\qquad \le C\big(Q(0, r_0+u_2) - Q(0, r_0+u_1)\big)^{\frac{1}{2+\nu}} \sum_{j=1}^{n}\sum_{k=1}^{\infty} (\alpha(k))^{\frac{\nu}{2+\nu}} \\ &\qquad \le Cn\textstyle\sum_{k=0}^{\infty} (\alpha(k))^{\frac{\nu}{2+\nu}} \big(Q(0, r_0+u_2) - Q(0, r_0+u_1)\big)^{\frac{1}{2+\nu}} \end{aligned}$$

and we obtain (5.4.4). □

Proposition 5.4.3.
We suppose that the assumptions of Theorem 5.1.2. Then for each $\epsilon > 0$, $\eta > 0$, there exist a constant $B < \infty$ such that for all $0 < \delta < 1$ and for all n large enough,

$$(i)\ \mathbb{P}\left(\sup_{\{\frac{B}{n^\kappa} < |r-r_0| \le \delta\}} \left|\frac{\sum_{t=1}^{n} \mathbf{1}_{\{r_0 < X_{t-1} \le r\}}}{nQ(r)} - 1\right| < \eta\right) > 1 - \epsilon$$

$$(ii)\ \mathbb{P}\left(\sup_{\{\frac{B}{n^\kappa} < |r-r_0| \le \delta\}} \left|\frac{\sum_{t=1}^{n} \varepsilon_t X_{t-1} \mathbf{1}_{\{r_0 < X_{t-1} \le r\}}}{nQ(r)}\right| < \eta\right) > 1 - \epsilon$$

where $\kappa = \frac{2+\nu}{3+2\nu}$.

Proof. For any $B > 0$ and $0 < \delta < 1$ we choose a partition of the interval $(\frac{B}{n^\kappa}, \delta]$ as follows. We fix a $b > 1$ and we let $I_i = (\frac{b^i B}{n^\kappa}, \frac{b^{i+1} B}{n^\kappa}]$ for all $i \ge 0$.

Proof of (i)**:** For simplicity we denote $Q(u) = Q(r_0, u)$ for any $u > r_0$. For any $\eta_1 > 0$ we have

$$\begin{aligned}\mathbb{P}\left(\sup_{i\geq 0}\left|\frac{\sum_{t=1}^n \mathbf{1}_{\{r_0<X_{t-1}\leq r_0+\frac{b^iB}{n^\kappa}\}}}{nQ(\frac{b^iB}{n^\kappa})}-1\right|>\eta_1\right) &= \mathbb{P}\left(\bigcup_{i\geq 0}\left|\frac{\sum_{t=1}^n \mathbf{1}_{\{r_0<X_{t-1}\leq r_0+\frac{b^iB}{n^\kappa}\}}}{nQ(\frac{b^iB}{n^\kappa})}-1\right|>\eta_1\right)\\ &\leq \sum_{i\geq 0}\mathbb{P}\left(\left|\frac{\sum_{t=1}^n \mathbf{1}_{\{r_0<X_{t-1}\leq r_0+\frac{b^iB}{n^\kappa}\}}}{nQ(\frac{b^iB}{n^\kappa})}-1\right|>\eta_1\right)\\ &\leq \frac{1}{\eta_1^2}\sum_{i\geq 0}\frac{Var\left(\frac{1}{n}\sum_{t=1}^n \mathbf{1}_{\{r_0<X_{t-1}\leq r_0+\frac{b^iB}{n^\kappa}\}}\right)}{Q^2(\frac{b^iB}{n^\kappa})}\end{aligned}$$

and by (i) and (iii) from Lemma 5.4.1, we deduce that

$$\begin{aligned}\mathbb{P}\left(\sup_{i\geq 0}\left|\frac{\sum_{t=1}^n \mathbf{1}_{\{r_0<X_{t-1}\leq r_0+\frac{b^iB}{n^\kappa}\}}}{nQ(\frac{b^iB}{n^\kappa})}-1\right|>\eta_1\right) &\leq \frac{C}{\eta_1^2}\sum_{i\geq 0}\frac{Q^{\frac{2}{2+\nu}}(\frac{b^iB}{n^\kappa})}{nQ^2(\frac{b^iB}{n^\kappa})}\\ &\leq \frac{C}{\eta_1^2}\sum_{i\geq 0}\frac{1}{n(\frac{mb^iB}{n^\kappa})^{\frac{2+2\nu}{2+\nu}}}\\ &\leq \frac{C}{mB\eta_1^2}\sum_{i\geq 0}\frac{1}{b^i}\end{aligned}$$

because $n^{1-\kappa(2+2\nu)/(2+\nu)} = n^{1/(3+2\nu)} \geq 1$. Consequently we obtain that

$$\mathbb{P}\left(\sup_{i\geq 0}\left|\frac{\sum_{t=1}^n \mathbf{1}_{\{r_0<X_{t-1}\leq r_0+\frac{b^iB}{n^\kappa}\}}}{nQ(\frac{b^iB}{n^\kappa})}-1\right|>\eta_1\right) \leq \frac{C}{mB\eta_1^2(1-b^{-1})}\ . \tag{5.4.8}$$

Let $0 < x \leq y \leq bx \leq \delta$. We denote $Q_n(x) = \frac{1}{n}\sum_{t=1}^n \mathbf{1}_{\{r_0<X_{t-1}\leq r_0+x\}}$. We have $Q_n(x) \to Q(x)$ almost surely as $n \to \infty$ by the ergodicity property of the process X or equivalently $\frac{Q_n(x)}{Q(x)} - 1 \to 0$ almost surely. Then $\forall \eta_1 > 0$, $\exists \delta$, such that $|x| \leq \delta$ and $|\frac{Q_n(x)}{Q(x)} - 1| < \eta_1$. By the increasing property of Q_n and Q, we have

$$(1-\eta_1)\frac{Q(x)}{Q(bx)} - 1 \leq \frac{Q_n(x)}{Q(bx)} - 1 \leq \frac{Q_n(y)}{Q(y)} - 1 \leq \frac{Q_n(bx)}{Q(x)} - 1 \leq \frac{Q(bx)}{Q(x)}(1+\eta_1) - 1 \tag{5.4.9}$$

Again by the increasing property of Q:

$$(1-\eta_1)\frac{Q(\frac{b^iB}{n^\kappa})}{Q(\frac{b^{i+1}B}{n^\kappa})} - 1 \leq \frac{Q_n(u)}{Q(u)} - 1 \leq (1+\eta_1)\frac{Q(\frac{b^{i+1}B}{n^\kappa})}{Q(\frac{b^iB}{n^\kappa})} - 1$$

and thus $\forall \eta > 0$, one can choose $\eta_1 > 0$ and $b > 1$ sufficiently small such that

$$\sup_i\left\{\left|(1-\eta_1)\frac{Q(\frac{b^iB}{n^\kappa})}{Q(\frac{b^{i+1}B}{n^\kappa})} - 1\right| \vee \left|(1+\eta_1)\frac{Q(\frac{b^{i+1}B}{n^\kappa})}{Q(\frac{b^iB}{n^\kappa})} - 1\right|\right\} < \eta. \tag{5.4.10}$$

Now let

$$A_n = \left\{\sup_i\left|\frac{Q_n(\frac{b^iB}{n^\kappa})}{Q(\frac{(b^iB)}{n^\kappa})} - 1\right| < \eta_1 \ \vee \sup_i\left|\frac{Q_n(\frac{b^{i+1}B}{n^\kappa})}{Q(\frac{(b^{i+1}B)}{n^\kappa})\}} - 1\right| < \eta_1\right\}.$$

Then on A_n, (5.4.9) and (5.4.10) imply that

$$\begin{aligned}\sup_{\frac{B}{n^\kappa}<u\le\delta}\left|\frac{Q_n(u)}{Q(u)}-1\right| &= \sup_i\sup_{u\in I_i}\left|\frac{Q_n(u)}{Q(u)}-1\right| \\ &\le \sup_i\left|(1-\eta_1)\frac{Q(\frac{b^iB}{n^\kappa})}{Q(\frac{b^{i+1}B}{n^\kappa})}-1\right| \vee \left|(1+\eta_1)\frac{Q(\frac{b^{i+1}B}{n^\kappa})}{Q(\frac{b^iB}{n^\kappa})}-1\right| \\ &< \eta\end{aligned}$$

and by (5.4.8), we choose B sufficiently large and such that for any $n \ge n_0 = [\frac{B}{\delta}]+1$ it holds

$$\begin{aligned}\mathbb{P}\left(\sup_{\frac{B}{n^\kappa}<u\le\delta}\left|\frac{Q_n(u)}{Q(u)}-1\right|>\eta\right) &\le \mathbb{P}\left(\sup_i\sup_{u\in I_i}\left|\frac{Q_n(u)}{Q(u)}-1\right|>\eta\right) \\ &\le \mathbb{P}\left(\sup_i\left|(1-\eta_1)\frac{Q(\frac{b^iB}{n^\kappa})}{Q(\frac{b^{i+1}B}{n^\kappa})}-1\right| \vee \left|(1+\eta_1)\frac{Q(\frac{b^{i+1}B}{n^\kappa})}{Q(\frac{b^iB}{n^\kappa})}-1\right|>\eta\right) \\ &\le \mathbb{P}(A_n^c) \\ &< \epsilon \qquad (5.4.11)\end{aligned}$$

and we have proved (i) of Proposition 5.4.3.

Proof of (ii) We need the following notations:

$$\begin{aligned}R_n(u) &= \frac{1}{n}\sum_{t=1}^n \varepsilon_t X_{t-1}\mathbf{1}_{\{r_0<X_{t-1}\le r_0+u\}} \\ r_n(u) &= \mathbb{E}R_n(u) \\ R_n^*(u_1,u_2) &= \frac{1}{n}\sum_{t=1}^n |\varepsilon_t X_{t-1}-\mathbb{E}\varepsilon_t X_{t-1}|\mathbf{1}_{\{r_0+u_1<X_{t-1}\le r_0+u_2\}} \\ R^*(u_1,u_2) &= \mathbb{E}R_n^*(u_1,u_2)\end{aligned}$$

For $\frac{b^iB}{n^\kappa} < u \le \frac{b^{i+1}B}{n^\kappa}$, $i \ge 0$ we have

$$\begin{aligned}|R_n(u)-r_n(u)| &\le \left|R_n(u)-r_n(u)-(R_n(b^iB/n^\kappa)-r_n(b^iB/n^\kappa))\right| + \left|R_n(b^iB/n^\kappa)-r_n(b^iB/n^\kappa)\right| \\ &\le R_n^*(b^iB/n^\kappa,u) + \left|R_n(b^iB/n^\kappa)-r_n(b^iB/n^\kappa)\right|.\end{aligned}$$

The increasing property of R_n^* and Q imply

$$\sup_{\frac{b^iB}{n^\kappa}<u\le\frac{b^{i+1}B}{n^\kappa}}\left|\frac{R_n(u)-r_n(u)}{G(u)}\right| \le \frac{R_n^*(\frac{b^iB}{n^\kappa},\frac{b^{i+1}B}{n^\kappa})}{Q(\frac{b^iB}{n^\kappa})} + \left|\frac{R_n(\frac{b^iB}{n^\kappa})-r_n(\frac{b^iB}{n^\kappa})}{Q(\frac{b^iB}{n^\kappa})}\right|. \qquad (5.4.12)$$

Similarly to (5.4.8), we use (5.4.3) and we obtain that for any $\eta_1 > 0$

$$\begin{aligned}\mathbb{P}\left(\sup_i\left|\frac{R_n(\frac{b^iB}{n^\kappa})-r_n(\frac{b^iB}{n^\kappa})}{Q(\frac{b^iB}{n^\kappa})}\right|>\eta_1\right) &\le \frac{C}{\eta_1^2}\sum_{i\ge0}\frac{Q^{1/(2+\nu)}(\frac{b^iB}{n^\kappa})}{nQ^2(\frac{b^iB}{n^\kappa})} \\ &\le \frac{C}{\eta_1^2}\sum_{i\ge0}\frac{1}{n(\frac{mb^iB}{n^\kappa})^{\frac{3+2\nu}{2+\nu}}}\end{aligned}$$

$$\leq \frac{C}{mB\eta_1^2}\sum_{i\geq 0}\frac{1}{b^i} \tag{5.4.13}$$

since we chose $\kappa = \frac{2+\nu}{3+2\nu}$. With (5.4.4), the same arguments yield to

$$\mathbb{P}\left(\sup_i \left|\frac{R_n^*(\frac{b^{i+1}B}{n^\kappa}, \frac{b^iB}{n^\kappa}) - R^*(\frac{b^{i+1}B}{n^\kappa}, \frac{b^iB}{n^\kappa})}{Q(\frac{b^iB}{n^\kappa})}\right| > \eta_1\right) \leq \frac{Cb}{m_1B_1\eta_1^2}\ . \tag{5.4.14}$$

The estimations (5.4.14) and (5.4.13) imply (ii) thanks to same arguments done in the proof of (i). □

Now we can start the proof of Theorem 5.1.2.

Proof. Since $\hat{\theta}_n$ is consistent by Theorem 5.1.1, we restrict the parameter space to an open neighborhood V_δ of θ_0.
The proof is divided in two steps.
Step 1: We prove that for any ϵ, there exists $B > 0$ such that with probability greater than $1-\epsilon$

$$L_n(\alpha_0, \beta_0, r) - L_n(\alpha_0, \beta_0, r_0) > 0 \text{ for } |r - r_0| > \frac{B}{n^\kappa} \text{ and } \theta \in V_\delta. \tag{5.4.15}$$

Remark that if (5.4.15) holds true then $n^\kappa(\hat{r}_n - r_0) = \mathrm{O}_{\mathbb{P}}(1)$. Consequently the first point (i) of Theorem 5.1.2 is proved.
Now we turn to the proof of (5.4.15). We denote $A_t^r = \left(\beta_0 + (\alpha_0 - \beta_0)\mathbf{1}_{\{X_{t-1}\leq r\}}\right)$ and thus $\varepsilon_t(\alpha_0, \beta_0, r) = X_t - A_t^r X_{t-1}$. One may write

$$\begin{aligned}
&n(L_n(\alpha_0, \beta_0, r) - L_n(\alpha_0, \beta_0, r_0)) \\
&\qquad = \sum_{t=1}^n \varepsilon_t^2(\alpha_0, \beta_0, r) - \sum_{t=1}^n \varepsilon_t^2(\alpha_0, \beta_0, r_0) \\
&\qquad = \sum_{t=1}^n (X_t - A_t^r X_{t-1})^2 - \sum_{t=1}^n (X_t - A_t^{r_0} X_{t-1})^2 \\
&\qquad = \sum_{t=1}^n (X_t - A_t^r X_{t-1} - X_t + A_t^{r_0} X_{t-1})(X_t - A_t^r X_{t-1} + X_t - A_t^{r_0} X_{t-1}) \\
&\qquad = \sum_{t=1}^n (-A_t^r X_{t-1} + A_t^{r_0} X_{t-1})(2X_t - A_t^r X_{t-1} - A_t^{r_0} X_{t-1}) \\
&\qquad = \sum_{t=1}^n (-A_t^r X_{t-1} + A_t^{r_0} X_{t-1})(2\varepsilon_t - A_t^r X_{t-1} + A_t^{r_0} X_{t-1}) \\
&\qquad = \sum_{t=1}^n (A_t^{r_0} - A_t^r)^2 X_{t-1}^2 + 2\varepsilon_t (A_t^{r_0} - A_t^r) X_{t-1} \\
&\qquad = C + D
\end{aligned}$$

with

$$\begin{aligned}
C &= \sum_{t=1}^n (\alpha_0 - \beta_0)^2 X_{t-1}^2 + 2\varepsilon_t(\alpha_0 - \beta_0) X_{t-1} \mathbf{1}_{\{X_{t-1}\leq r_0,\ X_{t-1} > r\}} \\
D &= \sum_{t=1}^n (\beta_0 - \alpha_0)^2 X_{t-1}^2 + 2\varepsilon_t(\beta_0 - \alpha_0) X_{t-1} \mathbf{1}_{\{X_{t-1} > r_0,\ X_{t-1}\leq r\}}.
\end{aligned}$$

We suppose that $r > r_0$, (the other cas can be treated analogously). We write that

$$n(L_n(\alpha_0,\beta_0,r) - L_n(\alpha_0,\beta_0,r_0)) = \sum_{t=1}^{n}(\beta_0-\alpha_0)^2 X_{t-1}^2 + 2\varepsilon_t(\beta_0-\alpha_0)X_{t-1}\mathbf{1}_{\{r_0<X_{t-1}\le r\}}. \tag{5.4.16}$$

Since V_δ is a neighborhood of θ_0, there exists $\rho > 0$ such that

$$\sum_{t=1}^{n}(\beta_0-\alpha_0)^2 X_{t-1}^2\mathbf{1}_{\{r_0<X_{t-1}\le r\}} \ge \rho^2\sum_{t=1}^{n}\mathbf{1}_{\{r_0<X_{t-1}\le r\}}.$$

By Proposition 5.4.3, we have with probability greater than $1-\epsilon$

$$\sum_{t=1}^{n}(\beta_0-\alpha_0)^2 X_{t-1}^2\mathbf{1}_{\{r_0<X_{t-1}\le r\}} \ge n\rho^2(1-\eta)Q(r). \tag{5.4.17}$$

We also have

$$\left|\sum_{t=1}^{n}2\varepsilon_t(\beta_0-\alpha_0)X_{t-1}\mathbf{1}_{\{r_0<X_{t-1}\le r\}}\right| \le 2\upsilon\left|\sum_{t=1}^{n}\varepsilon_t X_{t-1}\mathbf{1}_{\{r_0<X_{t-1}\le r\}}\right|$$

for some constant υ independent of n. By Proposition 5.4.3, it follows

$$\left|\sum_{t=1}^{n}2\varepsilon_t(\beta_0-\alpha_0)X_{t-1}\mathbf{1}_{\{r_0<X_{t-1}\le r\}}\right| \le 2\upsilon n\eta Q(r)\ . \tag{5.4.18}$$

Injecting (5.4.18) and (5.4.17) into (5.4.16) yield

$$\frac{L_n(\alpha_0,\beta_0,r) - L_n(\alpha_0,\beta_0,r_0)}{Q(r)} \ge -2\upsilon\eta + \rho^2(1-\eta)$$

with probability greater than $1-\epsilon$. We can choose $\eta > 0$ such that $-2\upsilon\eta + \rho^2(1-\eta) > 0$ and we obtain (5.4.15) and then the proof of the first point of the theorem is complete.
Step 2: We prove (ii).
For the proof of the second point of this theorem, we need the following lemma whose proof is similar to the proof of Lemma 5.5.2 from the Appendix.

Lemma 5.4.4. *Let $\lambda = (\alpha,\beta)$ be such that $\theta = (\lambda, r) \in \Lambda \times I = \Theta$. For any $\lambda \in \Lambda$, $U_\lambda(\eta)$ denotes the open ball in Λ centered in λ of radius η. Under the assumptions of Theorem 5.1.2, we have*

$$\mathbb{E}\left(\sup_{r\in\mathbb{1}}\sup_{\tilde\theta^*\in U_{\tilde\theta}(\eta)}|\varepsilon_t^2(\tilde\theta^*,r) - \varepsilon_t^2(\tilde\theta,r)|\right) \to 0 \text{ as } \eta\to 0.$$

Thanks to the above Lemma, for any given open neighborhood V of $(\alpha_0,\beta_0) \in \Lambda$ and $(\alpha^*,\beta^*) \in V^c = \Lambda\setminus\{V\}$, we have

$$\begin{aligned}\mathbb{E}\varepsilon_t^2(\alpha^*,\beta^*,r) &= \mathbb{E}(\varepsilon_t(\alpha^*,\beta^*,r) - \varepsilon_t(\alpha_0,\beta_0,r_0) + \varepsilon_t(\alpha_0,\beta_0,r_0))^2\\ &= \mathbb{E}(\varepsilon_t(\alpha^*,\beta^*,r) - \varepsilon_t(\alpha_0,\beta_0,r_0))^2 + \mathbb{E}\varepsilon_t^2(\alpha_0,\beta_0,r_0)\\ &> \sigma^2.\end{aligned}$$

Using Lemma 5.4.4, arguing as in the proof of Theorem 5.1.1 yield that there exists $\delta_0 > 0$ such that a.s. for sufficiently large n

$$\inf_{r\in I}\inf_{(\alpha^*,\beta^*)\in V^c} L_n(\alpha^*,\beta^*,r) \ge \delta_0 + \sigma^2. \tag{5.4.19}$$

By (5.4.16), on may prove that as $n \to \infty$:

$$\mathbb{E}\left(\sup_{|r-r_0|\leq \frac{B}{n^\kappa}} |L_n(\alpha_0,\beta_0,r) - L_n(\alpha_0,\beta_0,r_0)|\right) = \mathrm{O}(1/n). \tag{5.4.20}$$

By the ergodic theorem $L_n(\alpha_0,\beta_0,r_0) \to \sigma^2$ almost-surely. Thus (5.4.20) implies that for n large enough, there exists δ_0 such that

$$\sup_{|r-r_0|\leq \frac{B}{n^\kappa}} |L_n(\alpha_0,\beta_0,r)| < \delta_0 + \sigma^2 \tag{5.4.21}$$

with probability greater than $1-\epsilon$. By (5.4.19) and (5.4.21), we have for n sufficiently large

$$\inf_{r\in I}\inf_{(\alpha^*,\beta^*)\in V^c} L_n(\alpha^*,\beta^*,r) \geq \delta_0 + \sigma^2 > \sup_{|r-r_0|\leq \frac{B}{n^\kappa}} |L_n(\alpha_0,\beta_0,r)|$$

with probability greater than $1-\epsilon$. Define the set

$$D = \left\{\inf_{r\in I}\inf_{(\alpha^*,\beta^*)\in V^c} L_n(\alpha^*,\beta^*,r) > \sup_{|r-r_0|\leq \frac{B}{n^\kappa}} |L_n(\alpha_0,\beta_0,r)|\right\}.$$

Then on D (that satisfies $\mathbb{P}(D) \geq 1-\epsilon$), for n sufficiently large enough,

$$(\hat{\alpha}_n(r),\hat{\beta}_n(r)) \in V \text{ for } r \in [r_0 - \tfrac{B}{n^\kappa}, r_0 + \tfrac{B}{n^\kappa}].$$

By the arbitrariness of V and then fact that $\kappa < 1$ it follows that

$$\begin{aligned}\sup_{|r-r_0|\leq \frac{B}{n}} &\Big(|\hat{\alpha}_n(r)-\alpha_0| + |\hat{\beta}_n(r)-\beta_0|\Big) \\ &\leq \sup_{|r-r_0|\leq \frac{B}{n^\kappa}} \Big(|\hat{\alpha}_n(r)-\alpha_0| + |\hat{\beta}_n(r)-\beta_0|\Big) = \mathrm{o}_{\mathbb{P}}(1)\end{aligned}$$

and the proof of (ii) is complete. □

5.4.2 Proof of Theorem 5.1.3

To prove this theorem, we need further properties on the time series process.

Properties of $\partial L_n(\theta)/\partial\theta$

In order to investigate the asymptotic normality of our estimator, we need the following lemmas. First, we notice that in our simple case of a TAR process, we have by (5.1.2)

$$\frac{\partial\varepsilon_t(\theta)}{\partial\alpha} = -X_{t-1}\mathbf{1}_{\{X_{t-1}\leq r\}} \quad \text{and} \quad \frac{\partial\varepsilon_t(\theta)}{\partial\beta} = -X_{t-1}\mathbf{1}_{\{X_{t-1}> r\}}\,. \tag{5.4.22}$$

Lemma 5.4.5. *For any $(\lambda,r) \in \Lambda \times I$, the random variable $\frac{\partial L_n(\lambda,r)}{\partial\lambda}$ exists and belongs to L^2.*

Proof. Since $\mathbb{E}|\varepsilon_t(\lambda,r)|^4 < \infty$, we have

$$\left\|\frac{\partial L_n(\alpha,\beta,r)}{\partial\alpha}\right\|_{\mathbb{L}^2} = \left\|\frac{1}{n}\sum_{t=1}^{n}\frac{2\partial\varepsilon_t(\alpha,\beta,r)}{\partial\alpha}\varepsilon_t(\alpha,\beta,r)\right\|_{\mathbb{L}^2}$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{t=1}^{n}\left\|\frac{2\partial\varepsilon_t(\alpha,\beta,r)}{\partial\alpha}\varepsilon_t(\theta)\right\|_{\mathbb{L}^2}\\
&\leq \frac{2}{n}\sum_{t=1}^{n}\|X_{t-1}\varepsilon_t(\theta)\|_{\mathbb{L}^2}\\
&\leq \frac{2}{n}\sum_{t=1}^{n}\|X_{t-1}\|_{\mathbb{L}^4}\,\|\varepsilon_t(\theta)\|_{\mathbb{L}^4}\\
&\leq C\ .
\end{aligned}
$$

The proof is similar for the derivative with respect to β. □

Lemma 5.4.6. *Under the assumptions of theorem 5.1.3, the matrix*

$$I = \lim_{n\to\infty}\left(\sqrt{n}\frac{\partial L_n(\lambda_0, r_0)}{\partial\lambda}\right)$$

exists.

Proof. We proceed as in the proof of Lemma 5.4.1. We prove that the limit of

$$I^n_{\alpha,\beta} = \mathrm{Cov}\left(\sqrt{n}\frac{\partial L_n(\theta_0)}{\partial\alpha},\sqrt{n}\frac{\partial L_n(\theta_0)}{\partial\beta}\right)$$

exists. The two other cases ($I^n_{\alpha,\alpha}$ and $I^n_{\beta,\beta}$ with obvious notations) can be treated in the same way. We denote

$$Y_t^\alpha = \varepsilon_t(\theta_0)\frac{\partial\varepsilon_t(\theta_0)}{\partial\alpha} \quad \text{and} \quad Y_t^\beta = \varepsilon_t(\theta_0)\frac{\partial\varepsilon_t(\theta_0)}{\partial\beta}.$$

By stationarity one may write that

$$
\begin{aligned}
I^n_{\alpha,\beta} &= \frac{4}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathrm{Cov}(Y_t^\alpha, Y_s^\beta)\\
&= \frac{4}{n}\sum_{k=1-n}^{k=n-1}(n-|k|)Cov(Y_t^\alpha, Y_{t-k}^\beta)\\
&= 4\sum_{k\in\mathbb{Z}}\frac{n-k}{n}\mathbf{1}_{\{|k|\leq n-1\}}c(k) \qquad (5.4.23)
\end{aligned}
$$

where $c_k = Cov(Y_t^\alpha, Y_{t-k}^\beta)$. Then the dominated convergence Lebesgue theorem yields

$$\lim_{n\to\infty} I^n_{\alpha,\beta} = \sum_{k\in\mathbb{Z}} c(k)$$

provided that $\sum_{k\in\mathbb{Z}}|c(k)| < \infty$.
In order to prove that $\sum_{k\in\mathbb{Z}}|c(k)| < \infty$, we first suppose that $k \geq 0$. Using (5.1.2) and (5.4.22) one obtains that

$$
\begin{aligned}
c(k) &= \mathrm{Cov}(Y_t^\alpha, Y_{t-k}^\beta)\\
&= \mathrm{Cov}\Big(X_tX_{t-1}\mathbf{1}_{\{X_{t-1}\leq r_0\}} - (\beta_0 + (\alpha_0-\beta_0)\mathbf{1}_{X_{t-1}\leq r_0})X_{t-1}^2\mathbf{1}_{\{X_{t-1}\leq r_0\}},\\
&\qquad X_{t-k}X_{t-k-1}\mathbf{1}_{\{X_{t-k-1}\leq r_0\}} - (\beta_0 + (\alpha_0-\beta_0)\mathbf{1}_{X_{t-k-1}\leq r_0})X_{t-k-1}^2\mathbf{1}_{\{X_{t-k-1}\leq r_0\}}\Big).
\end{aligned}
$$

Since Θ is compact, there exists a constant C such that

$$\begin{aligned}\mathbb{E}|Y_t^\alpha|^{2+\nu} &= \mathbb{E}\left|X_tX_{t-1}\mathbf{1}_{\{X_{t-1}\leq r_0\}} - (\beta_0 + (\alpha_0-\beta_0)\mathbf{1}_{X_{t-1}\leq r_0})X_{t-1}^2\mathbf{1}_{\{X_{t-1}\leq r_0\}}\right|^{2+\nu}\\ &\leq C\ \mathbb{E}|X_1|^{4+2\nu}\\ &\leq C\end{aligned}$$

and the Davydov inequality implies that

$$\begin{aligned}|c(k)| &\leq C\|Y_t^\alpha\|_{\mathbb{L}^{2+\nu}}\|Y_{t-k}^\beta\|_{\mathbb{L}^{2+\nu}}(\alpha(k-1))^{\frac{\nu}{2+\nu}}\\ &\leq C(\alpha(k-1))^{\frac{\nu}{2+\nu}}.\end{aligned}$$

It follows that $\sum_{k\in\mathbb{Z}}|c(k)|<\infty$. □

Lemma 5.4.7.
Under the assumptions of Theorem 5.1.3, the random vector $\sqrt{n}\frac{\partial L_n(\theta_0)}{\partial\lambda}$ *has a limiting distribution normal with mean* 0 *and covariance matrix* I.

Proof. We follow the arguments of Lemma 4 in [43].
We have

$$\mathbb{E}\left(\frac{\partial\varepsilon_t^2(\theta_0)}{\partial\lambda}\right) = \mathbb{E}\left(\varepsilon_t\frac{\partial\varepsilon_t^2(\theta_0)}{\partial\lambda}\right) = 0$$

because, by (5.4.22), $\partial\varepsilon_t^2(\theta_0)/\partial\lambda$ belongs to the Hilbert space $\mathcal{H}_X(t-1)$ generated by $\{X_r\ ; r\leq t-1\}$. Hence $\sqrt{n}\frac{\partial}{\partial\lambda}L_n(\theta_0)$ is centred. We have

$$\sqrt{n}\frac{\partial}{\partial\lambda}L_n(\theta_0) = \frac{2}{\sqrt{n}}\sum_{t=1}^n Y_t$$

where

$$Y_t = (Y_t^\alpha, Y_t^\beta)' = \left(\varepsilon_t(\theta_0)\frac{\partial\varepsilon_t(\theta_0)}{\partial\alpha}\ ,\ \varepsilon_t(\theta_0)\frac{\partial\varepsilon_t(\theta_0)}{\partial\beta}\right)'$$

and

$$\begin{aligned}Y_t^\alpha &= (X_t-\alpha_0X_{t-1})X_{t-1}\mathbf{1}_{\{X_{t-1}\leq r_0\}}\\ Y_t^\beta &= (X_t-\beta_0X_{t-1})X_{t-1}\mathbf{1}_{\{X_{t-1}>r_0\}}\ .\end{aligned}$$

The process $(Y_t)_{t\in\mathbb{Z}}$ is stationary and since it is a function of a finite number of values of the process $(X_t)_{t\in\mathbb{Z}}$, it also satisfies a mixing property of the form (5.1.6). The central limit theorem for strongly mixing processes (see [52]) implies the expected result. □

Lemma 5.4.8.
Almost surely the matrix

$$J = \lim_{n\to\infty}\begin{pmatrix}\frac{\partial^2L_n(\theta_0)}{\partial\alpha^2} & \frac{\partial^2L_n(\theta_0)}{\partial\alpha\partial\beta}\\ \frac{\partial^2L_n(\theta_0)}{\partial\alpha\partial\beta} & \frac{\partial^2L_n(\theta_0)}{\partial\beta^2}\end{pmatrix} = 2\begin{pmatrix}\mathbb{E}(X_1^2\mathbf{1}_{\{X_1\leq r_0\}}) & 0\\ 0 & \mathbb{E}(X_1^2\mathbf{1}_{\{X_1>r_0\}})\end{pmatrix}$$

exists and is strictly positive definite.

Proof. By (5.4.22) it is clear that $\frac{\partial^2 L_n(\theta_0)}{\partial\alpha\partial\beta} = 0$. Using again (5.4.22), one has

$$\begin{aligned}\frac{\partial^2 L_n(\theta_0)}{\partial\alpha^2} &= \frac{\partial}{\partial\alpha}\left(\frac{2}{n}\sum_{t=1}^{n}\varepsilon_t(\theta)\frac{\partial\varepsilon_t(\theta)}{\partial\alpha}\right)(\theta_0)\\ &= \frac{2}{n}\sum_{t=1}^{n}\left(\frac{\partial\varepsilon_t(\theta_0)}{\partial\alpha}\right)^2\\ &= \frac{2}{n}\sum_{t=1}^{n}\left(X_{t-1}\mathbf{1}_{\{X_{t-1}\le r_0\}}\right)^2\\ &\to 2\mathbb{E}(X_1^2\mathbf{1}_{\{X_1\le r_0\}}) \quad \text{as } n\to\infty\end{aligned}$$

by the ergodic theorem. Similar arguments hold true for the limit of $\frac{\partial^2 L_n(\theta_0)}{\partial\beta^2}$. The matrix J is clearly strictly positive definite. □

Now we are able to start the proof of Theorem 5.1.3.

Proof of Theorem 5.1.3

Proof. By Proposition 5.4.9 which is stated below, we deduce that

$$\sqrt{n}(\hat\lambda_n(\hat r_n) - \lambda_0) = \sqrt{n}(\hat\lambda_n(r_0) - \lambda_0) + o_{\mathbb{P}}(1).$$

Now we prove the asymptotic normality. On a neighborhood of θ_0, using a standard technique of Taylor expansion, we have

$$0 = \sqrt{n}\begin{pmatrix}\frac{\partial L_n(\hat\alpha_n,\hat\beta_n,r_0)}{\partial\alpha}\\ \frac{\partial L_n(\hat\alpha_n,\hat\beta_n,r_0)}{\partial\beta}\end{pmatrix} = \sqrt{n}\begin{pmatrix}\frac{\partial L_n(\theta_0)}{\partial\alpha}\\ \frac{\partial L_n(\theta_0)}{\partial\beta}\end{pmatrix} + \sqrt{n}\begin{pmatrix}\frac{\partial^2 L_n(\alpha_n^*)}{\partial\alpha^2} & 0\\ 0 & \frac{\partial^2 L_n(\beta_n^*)}{\partial\beta^2}\end{pmatrix}\begin{pmatrix}\hat\alpha_n-\alpha_0\\ \hat\beta_n-\beta_0\end{pmatrix}$$

where α_n^* is between $\hat\alpha_n$ and α_0, and β_n^* is between $\hat\beta_n$ and β_0. We remark that in our simple case we have

$$\begin{pmatrix}\frac{\partial^2 L_n(\alpha_n^*)}{\partial\alpha^2} & 0\\ 0 & \frac{\partial^2 L_n(\beta_n^*)}{\partial\beta^2}\end{pmatrix} = \begin{pmatrix}\frac{\partial^2 L_n(\theta_0)}{\partial\alpha^2} & 0\\ 0 & \frac{\partial^2 L_n(\theta_0)}{\partial\beta^2}\end{pmatrix}$$

and then by Lemma 5.4.8 converges almost surely to the matrix J which is invertible. By Lemma 5.4.7 we deduce that $\begin{pmatrix}\hat\alpha_n-\alpha_0\\ \hat\beta_n-\beta_0\end{pmatrix}$ has a normal limiting distribution with mean 0 and covariance matrix $J^{-1}IJ^{-1}$. □

5.4.3 Proofs of auxilliary results from Section 5.2

Convergence rate of $\hat\lambda_n(r)$

We shall need in the following a result concerning the convergence rate of $\hat\lambda_n(r) - \hat\lambda(r_0)$ where we recall that $\hat\lambda_n(r) = (\hat\alpha_n(r),\hat\beta_n(r))'$ and that we have denoted $\lambda = (\alpha,\beta)' \in \Lambda$ and $\Theta = \Lambda \times I$.

Proposition 5.4.9. *Under the assumptions of Theorem 5.1.4, for any fixed $B \in (0,\infty)$, we have*

$$\sqrt{n}\sup_{|r-r_0|\le B/n}\left(|\hat\alpha_n(r)-\hat\alpha_n(r_0)| + |\hat\beta_n(r)-\hat\beta_n(r_0)|\right) = o_{\mathbb{P}}(1)\ . \tag{5.4.24}$$

Proof. The proof is exactly the same than the proof of Theorem 2.2(ii) from [60] as soon as the following Lemma is true. □

Lemma 5.4.10. *Under the assumptions of Theorem 5.1.4, there exists $\varepsilon > 0$ such that any $\eta > 0$ and any $0 < B < \infty$ we have*

$$\mathbb{E} \sup_{\lambda\in\Lambda} \sup_{|r-r_0|\leq\eta} \left| \frac{\partial\epsilon_t^2(\lambda,r)}{\partial\lambda} - \frac{\partial\epsilon_t^2(\lambda,r_0)}{\partial\lambda} \right| \leq C\ \eta^{\frac{1}{2}+\varepsilon} \tag{5.4.25}$$

$$\mathbb{E} \sup_{\lambda\in\Lambda} \sup_{|r-r_0|\leq\eta} \left| \frac{\partial^2\epsilon_t^2(\lambda,r)}{\partial\lambda\partial\lambda'} - \frac{\partial^2\epsilon_t^2(\lambda,r_0)}{\partial\lambda\partial\lambda'} \right| \leq C\ \eta^{\frac{1}{2}+\varepsilon} \tag{5.4.26}$$

$$\mathbb{E} \sup_{|\lambda-\lambda_0|\leq\eta} \sup_{r\in I} \left| \frac{\partial\epsilon_t^2(\lambda,r)}{\partial\lambda} - \frac{\partial\epsilon_t^2(\lambda_0,r)}{\partial\lambda} \right| \leq C\ \eta \tag{5.4.27}$$

$$\mathbb{E} \sup_{|r-r_0|\leq B/n} \left| \frac{\partial L_n(\lambda_0,r)}{\partial\lambda} - \frac{\partial L_n(\lambda_0,r_0)}{\partial\lambda} \right| \leq C\ (1/n)^{\frac{1}{2}+\varepsilon} \tag{5.4.28}$$

$$\mathbb{E} \sup_{|\lambda-\lambda_0|\leq B/\sqrt{n}} \sup_{|r-r_0|\leq B/n} \left| \frac{\partial^2 L_n(\lambda,r)}{\partial\lambda\partial\lambda'} - \frac{\partial^2 L_n(\lambda_0,r_0)}{\partial\lambda\partial\lambda'} \right| \leq C\ (1/n)^{\frac{1}{2}+\varepsilon}\ . \tag{5.4.29}$$

Proof.

- Using (5.1.3) and (5.1.2) one may write that

$$\frac{\partial\epsilon_t^2(\alpha,\beta,r)}{\partial\alpha} - \frac{\partial\epsilon_t^2(\alpha,\beta,r_0)}{\partial\alpha} = (\mathbf{1}_{\{X_{t-1}\leq r\}} - \mathbf{1}_{\{X_{t-1}\leq r_0\}}) \times \left[\beta X_{t-1}^2 + X_t X_{t-1}((\alpha-\beta)-1)\right]$$

  and thus (5.4.25), for the derivative with respect to α, follows from Assumption **(H5)** and Hölder's inequality (choose $\varepsilon = 1/(2+\nu)$). The derivative with respect to β can be treated in a similar way.

- We recall that

$$\frac{\partial^2\epsilon_t^2(\alpha,\beta,r)}{\partial\alpha^2} = 2X_{t-1}\mathbf{1}_{\{X_{t-1}\leq r\}}\ ,\ \frac{\partial^2\epsilon_t^2(\alpha,\beta,r)}{\partial\beta^2} = 2X_{t-1}\mathbf{1}_{\{X_{t-1}\ r\}} \text{ and } \frac{\partial^2\epsilon_t^2(\alpha,\beta,r)}{\partial\alpha\partial\beta} = 0\ ,$$

  and arguing as before leads to (5.4.26).

- We have

$$\left|\frac{\partial\epsilon_t^2(\alpha,\beta,r)}{\partial\alpha} - \frac{\partial\epsilon_t^2(\alpha_0,\beta_0,r)}{\partial\alpha}\right| = 2X_{t-1}^2\mathbf{1}_{\{X_{t-1}\leq r\}}\,|(\beta-\beta_0) + ((\alpha-\alpha_0)-(\beta-\beta_0))\mathbf{1}_{\{X_{t-1}\leq r\}}|$$

$$\left|\frac{\partial\epsilon_t^2(\alpha,\beta,r)}{\partial\beta} - \frac{\partial\epsilon_t^2(\alpha_0,\beta_0,r)}{\partial\beta}\right| = 2X_{t-1}^2\mathbf{1}_{\{X_{t-1}>r\}}|\beta-\beta_0|$$

  and consequently (5.4.27) holds true.

- The proof of (5.4.28) and (5.4.29) follows from the previous estimations.

□

Proof of Lemma 5.2.1

Proof. We have to prove that for any $B > 0$ it holds

$$\sup_{|s|\leq B} |\tilde{\phi}_n(s) - \phi_n(s)| = \mathrm{o}_{\mathbb{P}}(1) \ . \tag{5.4.30}$$

Using the expressions (5.1.7)and (5.1.8), a Taylor expansion yields

$$\begin{aligned}\frac{1}{n}(\tilde{\phi}_n(s) - \phi_n(s)) = &\frac{\partial L_n(\lambda^*, r_0 + s/n)}{\partial \lambda}(\hat{\lambda}_n(r_0 + s/n) - \tilde{\theta}_0) \\ &- \frac{\partial L_n(\lambda^\sharp, r_0)}{\partial \lambda}(\hat{\lambda}_n(r_0) - \tilde{\theta}_0)\end{aligned}$$

Where λ^* lies between $\hat{\lambda}_n(r_0 + s/n)$ and λ_0, and $\lambda^\sharp$ lies between $\hat{\lambda}_n(r_0)$ and λ_0. Another Taylor's expansion implies that for any $\lambda \in \Lambda$:

$$\begin{aligned}\frac{\partial L_n(\lambda, r)}{\partial \lambda} &= \frac{\partial L_n(\hat{\lambda}_n(r), r)}{\partial \lambda} + \frac{\partial^2 L_n(\bar{\lambda}, r)}{\partial \lambda \partial \lambda'}(\hat{\lambda}_n(r) - \lambda) \\ &= 0 + \frac{\partial^2 L_n(\bar{\lambda}, r)}{\partial \lambda \partial \lambda'}(\hat{\lambda}_n(r) - \lambda)\end{aligned}$$

where $\bar{\lambda}$ is between λ and $\hat{\lambda}_n(r)$. Thus we have

$$\left|\frac{1}{n}(\tilde{\phi}_n(s) - \phi_n(s))\right| \leq \sup_{\lambda\in\Lambda}\sup_{r\in I}\left|\frac{\partial^2 L_n(\lambda, r)}{\partial \lambda \partial \lambda'}\right| \times \left[|\hat{\lambda}_n(r_0 + s/n) - \lambda_0|^2 + |\hat{\lambda}_n(r_0) - \lambda_0|^2\right]$$

and (5.4.24) implies (5.4.30) because $\sup_{\lambda\in\Lambda}\sup_{r\in I}\left|\frac{\partial^2 L_n(\lambda,r)}{\partial\lambda\partial\lambda'}\right| = \mathrm{O}_{\mathbb{P}}(1)$. □

5.5 Appendix: proof of consistency

We shall need the following two lemmas concerning the identifiability of the parametric model and the continuity of the noise process.

Lemma 5.5.1.
If the conditions in Theorem 5.1.1 hold, then $\mathbb{E}\varepsilon_t^2(\theta) \geq \mathbb{E}\varepsilon_t^2$ for all $\theta \in \Theta$ and the equality holds if and only if $\theta = \theta_0$

Proof. By (5.1.2), $\varepsilon_t(\theta) - \varepsilon_t$ belongs to the Hilbert space $H_X(t-1)$ generated by $\{X_s\}_{s\leq t-1}$. Therefore the linear innovation ε_t is not correlated with $\varepsilon_t(\theta) - \varepsilon_t$. Thus we may write

$$\begin{aligned}\mathbb{E}\varepsilon_t^2(\theta) &= \mathbb{E}(\varepsilon_t(\theta) - \varepsilon_t)^2 + 2\mathbb{E}\{\varepsilon_t(\varepsilon_t(\theta) - \varepsilon_t)\} + \mathbb{E}\varepsilon_t^2 \\ &= \mathbb{E}(\varepsilon_t(\theta) - \varepsilon_t)^2 + \mathbb{E}\varepsilon_t^2 \\ &\geq \mathbb{E}\varepsilon_t^2\end{aligned}$$

If the equality holds for some $\theta^* = (\alpha^*, \beta^*, r^*)$ then,

$$\varepsilon_t(\theta^*) - \varepsilon_t = 0 \ a.s.,$$

But we have

$$\varepsilon_t(\theta^*) - \varepsilon_t = a_1 + a_2 + a_3 + a_4 \tag{5.5.1}$$

with

$$\begin{aligned} a_1 &= \mathbf{1}_{\{X_{t-1}\leq r^*,\ X_{t-1}\leq r_0\}}(\alpha^* - \alpha_0)X_{t-1} \\ a_2 &= \mathbf{1}_{\{X_{t-1}\leq r^*,\ X_{t-1}> r_0\}}(\alpha^* - \beta_0)X_{t-1} \\ a_3 &= \mathbf{1}_{\{X_{t-1}> r^*,\ X_{t-1}\leq r_0\}}(\beta^* - \alpha_0)X_{t-1} \\ a_4 &= \mathbf{1}_{\{X_{t-1}> r^*,\ X_{t-1}> r_0\}}(\beta^* - \beta_0)X_{t-1}\ . \end{aligned}$$

First we suppose that $r_0 < r^*$ and consequently $a_3 = 0$. By assumption **(H2)**, the probability distribution is bounded away from 0 over each bounded subset. Thus we have

$$(\alpha^*-\alpha_0)\mathbf{1}_{\{X_{t-1}\leq r^*,\ X_{t-1}\leq r_0\}}+(\alpha^*-\beta_0)\mathbf{1}_{\{X_{t-1}\leq r^*,\ X_{t-1}> r_0\}}+(\beta^*-\beta_0)\mathbf{1}_{\{X_{t-1}> r^*,\ X_{t-1}> r_0\}} = 0.$$

Using the orthogonality among the indicator functions we obtain that

$$\begin{aligned} 0 &= |\alpha^* - \alpha_0|\mathbb{P}(X_{t-1}\leq r^*,\ X_{t-1}\leq r_0) + |\alpha^* - \beta_0|\mathbb{P}(X_{t-1}\leq r^*,\ X_{t-1}> r_0) \\ &\qquad + |\beta^* - \beta_0|\mathbb{P}(X_{t-1}> r^*,\ X_{t-1}> r_0) \\ &= |\alpha^* - \alpha_0|\mathbb{P}(X_{t-1}\leq r_0) + |\alpha^* - \beta_0|\mathbb{P}(r_0< X_{t-1}\leq r^*) + |\beta^* - \beta_0|\mathbb{P}(X_{t-1}> r^*)\ . \end{aligned}$$

By Assumption **(H2)** we have $\mathbb{P}(X_{t-1}\leq r_0) > 0$, $\mathbb{P}(r_0 < X_{t-1}\leq r^*) > 0$ and $\mathbb{P}(X_{t-1} > r^*) > 0$ since X_{t-1} has a positive and continuous density. Consequently we have $\alpha^* = \alpha_0 = \beta_0$. This is in a contradiction with **(H3)**. Similar results hold when $r_0 > r^*$. So we obtain that $\theta^* = \theta_0$. □

Lemma 5.5.2. *If the assumptions of Theorem 5.1.1 hold, then for any $\theta \in \Theta$*

$$\mathbb{E}\sup_{\theta^*\in V_\delta}|\varepsilon_t^2(\theta^*) - \varepsilon_t^2(\theta)| \to 0$$

when $(V_\delta)_{\delta>0}$ are open neighborhoods of θ shrinking to θ.

Proof. We have

$$\epsilon_t^2(\theta) = R_{1t}(\theta) + R_{2t}(\theta) + R_{3t}(\theta) + R_{4t}(\theta)$$

with

$$\begin{aligned} R_{1t}(\theta) &= (X_t - \alpha X_{t-1})^2\mathbf{1}_{\{X_{t-1}\leq r,\ X_{t-1}\leq r_0\}} \\ R_{2t}(\theta) &= (X_t - \alpha X_{t-1})^2\mathbf{1}_{\{X_{t-1}\leq r,\ X_{t-1}> r_0\}} \\ R_{3t}(\theta) &= (X_t - \beta X_{t-1})^2\mathbf{1}_{\{X_{t-1}> r,\ X_{t-1}\leq r_0\}} \\ R_{4t}(\theta) &= (X_t - \beta X_{t-1})^2\mathbf{1}_{\{X_{t-1}> r,\ X_{t-1}> r_0\}}\ . \end{aligned}$$

Thus we have

$$\epsilon_t^2(\theta^*) - \epsilon_t^2(\theta) = \sum_{j=1}^4 R_{jt}(\theta^*) - R_{jt}(\theta)\ .$$

Using the equation (5.1.1) satisfied by X_t we remark that

$$R_{1t}(\theta) = \varepsilon_t^2\mathbf{1}_{\{X_{t-1}\leq r,\ X_{t-1}\leq r_0\}}+2\varepsilon_t(\alpha_0-\alpha)X_{t-1}\mathbf{1}_{\{X_{t-1}\leq r,\ X_{t-1}\leq r_0\}}+(\alpha_0-\alpha)^2X_{t-1}^2 1_{\{X_{t-1}\leq r,\ X_{t-1}\leq r_0\}}.$$

So we have

$$\begin{aligned} R_{1t}(\theta^*) - R_{1t}(\theta) = \varepsilon_t^2(&\mathbf{1}_{\{X_{t-1}\leq r^*,\ X_{t-1}\leq r_0\}} - \mathbf{1}_{\{X_{t-1}\leq r,\ X_{t-1}\leq r_0\}}) \\ &+ 2\epsilon_t X_{t-1}(\alpha_0 - \alpha^*)\mathbf{1}_{\{X_{t-1}\leq r^*,\ X_{t-1}\leq r_0\}} - (\alpha_0 - \alpha)\mathbf{1}_{\{X_{t-1}\leq r,\ X_{t-1}\leq r_0\}}) \end{aligned}$$

$$
\begin{aligned}
&+ X_{t-1}^2((\alpha_0-\alpha^*)^2\mathbf{1}_{\{X_{t-1}\le r^*,\ X_{t-1}\le r_0\}} - (\alpha_0-\alpha)^2\mathbf{1}_{\{X_{t-1}\le r,\ X_{t-1}\le r_0\}}) \\
&= \Delta^1(\alpha,\alpha^*,r,r^*) + \Delta^2(\alpha,\alpha^*,r,r^*) + \Delta^3(\alpha,\alpha^*,r,r^*)
\end{aligned}
$$

with obvious notations. We may write that

$$
\begin{aligned}
|\Delta^3(\alpha,\alpha^*,r,r^*)| \le\ & X_{t-1}^2(\alpha_0-\alpha^*)^2|\mathbf{1}_{\{X_{t-1}\le r^*,\ X_{t-1}\le r_0\}} - \mathbf{1}_{\{X_{t-1}\le r,\ X_{t-1}\le r_0\}}| \\
& + X_{t-1}^2((\alpha_0-\alpha)^2 - (\alpha_0-\alpha^*)^2)\mathbf{1}_{\{X_{t-1}\le r,\ X_{t-1}\le r_0\}}.
\end{aligned}
$$

Then we have

$$
\lim_{\delta\to 0}\mathbb{E}\left(\sup_{\theta^*\in V_\delta}|\Delta^3(\alpha,\alpha^*,r,r^*)|\right) = 0
$$

and we may have similar convergences for $\Delta^1(\alpha,\alpha^*,r,r^*)$ and $\Delta^2(\alpha,\alpha^*,r,r^*)$. The other terms $R_{jt}(\theta^*) - R_{jt}(\theta)$ for $j = 2,3,4$ can be treated in the same way and the lemma is proved. □

Now, we start the proof of theorem 5.1.1.

Proof. Let V be an open neighborhood of the true value θ_0 and $V^c = \Theta \setminus V$. First we prove that

$$
\inf_{\theta\in V^c} L_n(\theta) > \inf_{\theta\in V} L_n(\theta). \tag{5.5.2}
$$

Indeed, by Lemma 5.5.1, Lemma 5.5.2 and the compactness of V^c, we have

$$
\inf_{\gamma\in V^c}\mathbb{E}\varepsilon_t^2(\gamma) = \mathbb{E}\varepsilon_t^2(\gamma_0) > \sigma^2
$$

Let $U_\theta(\rho_0)$ be a subset of V^c. By Lemma 5.5.2, we obtain

$$
\mathbb{E}\inf_{\gamma^*\in U_\theta(\rho_0)}\varepsilon_t^2(\gamma^*) \ge \mathbb{E}\varepsilon_t^2(\gamma_0) - \phi_0 = 2\phi_0 + \sigma^2 \tag{5.5.3}
$$

with

$$
\phi_0 = \frac{1}{3}(\mathbb{E}\varepsilon_t^2(\gamma_0) - \sigma^2) > 0.
$$

Since V^c is compact, there a finite partition $U_{\theta_1}(\rho_0), U_{\theta_2}(\rho_0), \ldots, U_{\theta_m}(\rho_0)$, such that $V^c = \cup_{i=1}^m U_{\theta_i}(\rho_0)$. By the ergodic theorem and (5.5.3), we have almost surely that for n sufficiently large and $1 \le j \le m$:

$$
\begin{aligned}
\inf_{\rho^*\in V^c} L_n(\rho^*) &= \inf_{\rho^*\in\cup_{i=1}^m U_{\theta_i}(\rho_0)} L_n(\rho^*) \\
&= \min_{1\le i\le m}\inf_{\rho^*\in U_{\theta_i}(\rho_0)}\frac{1}{n}\sum_{t=1}^n\varepsilon_t^2(\rho^*) \\
&\ge \min_{1\le i\le m}\frac{1}{n}\sum_{t=1}^n\inf_{\rho^*\in U_{\theta_i}(\rho_0)}\varepsilon_t^2(\rho^*) \\
&\ge \sigma^2 + \phi_0\ .
\end{aligned}
$$

Thus for sufficiently large n

$$
\inf_{\theta\in V} L_n(\theta) \le L_n(\theta_0) \le \sigma^2 + \frac{\phi_0}{2}.
$$

We then deduce that $\hat{\theta}_n \in V$ a.s. Since V is a arbitrary subset, we have the inequality (5.5.2) and the proof is complete. □

Bibliographie

[1] David Aldous, *Stopping times and tightness*, Ann. Probability **6** (1978), no. 2, 335–340.

[2] G. B. Angelosante D. et Giannakis, *Group lassoing change-points in piecewise-constant ar processes.* ., EURASIP Journal on Advances in Signal Processing. **1** (2012).

[3] Burke J. V. Aravkin A. Y. and G. Pillonetto, *Sparse/-robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory.*, The Journal of Machine Learning Research. **14** (1997), no. 1.

[4] J. Bai and P. Perron, *Estimating and testing linear models with multiple structural changes.*, Econometrica. **66** (1998), 47–78.

[5] L. Bardwell and P. Fearnhead, *Bayesian detection of abnormal segments in multiple time series.*, arXiv preprint arXiv:1412.5565. (2014).

[6] M . Basseville and I. Nikirov, *The detection of abrupt changes-theory and applications. information and system sciences series.*, Prentice-Hall, Englewood Cliffs, Nj. (1993), 47–78.

[7] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.*, J. Roy. Statist. Soc. Ser. B, Journal of the Royal Statistical Society. Series B. **57** (1995), no. 1, 289–300.

[8] A. Benveniste and M. Basseville, *Detection of abrupt changes in signals and dynamical systems: some statistical aspects.*, In Analysis and optimization of systems, Lecture Notes in Control and Inform. Sci. Springer, Berlin . **62** (1984), no. 1, 145–155.

[9] Simeon M. Berman, *A compound Poisson limit for stationary sums, and sojourns of Gaussian processes*, Ann. Probab. **8** (1980), no. 3, 511–538.

[10] P. Bertrand, *A local method for estimating change points: the "hat-function".*, Statistics . **34** (2000), no. 1, 215–235.

[11] P. Bertrand, M. Fhima, and A. Guillin, *Off-line detection of multiple change points by the filtered derivative with p-value method.*, Sequential Analysis . **30** (2011), no. 2, 172–207.

[12] Patrick Billingsley, *Convergence of probability measures*, Second, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Inc., New York, 1999. A Wiley-Interscience Publication.

[13] L. Birgé and P. Massart, *Minimal penalties for gaussian model selection.*, Probab. Theory Related Fields. **138** (2007), no. 1-2, 33–73.

[14] G.E.P. Box and G.M. Jenkins, *Forecasting and control*, Time Series Analysis, Holden-Day, San Francisco. (1970).

[15] Parikh N. Chu E. Peleato B. et Eckstein J. Boyd S., *Distributed optimization and statistical learning via the alternating direction method of multipliers.*, Foundations and Trends® in Machine Learning. **3** (2011), no. 1, 1–22.

[16] Peter J. Brockwell and Richard A. Davis, *Time series: theory and methods*, Second, Springer Series in Statistics, Springer-Verlag, New York, 1991. MR1093459

[17] P.J. Brockwell and R. A. Davis, *Time series: Theory and methods (2nd). springer-verlag, new york.*

[18] B. Brodsky and B. Darkhovsky, *Nonparametric methods in change-points problems.*, Kluwer Academic Publishers, the Netherlands. (1993).

[19] Gelman A. Jones G. Brooks S. and X. Meng, *Handbook of markov chain monte carlo.*, Chapman Hall/CRC Handbooks of Modern Statistical Methods. Taylor Francis. (2011).

[20] Gelfand A. E. Carlin B. P. and A. F. Smith, *Hierarchical bayesian analysis of changepoint problems.*, Applied statistics. (1992).

[21] Lebarbier E Lévy-Leduc C Robin S. Chakar S, *A robust approach for estimating change-points in the mean of an ar(1) process.*, Bernouilli. **3** (2017.), no. 2, 1408–1447.

[22] K. S. Chan and H. Tong, *On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations*, Adv. in Appl. Probab. **17** (1985), no. 3, 666–678.

[23] ———, *On estimating thresholds autoregressive models.*, Journal of Times Series Analysis . **7** (1986), 179–190.

[24] S Chan K, *Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model .*, Annals of Statistics. **21** (1993), 520–533.

[25] Zhang N. Chen H. and al., *Graph-based change-point detection.*, The Annals of Statistics. **43** (2015), no. 1, 139–176.

[26] Pavel Chigansky and Fima C. Klebaner, *Compound Poisson approximation for triangular arrays with application to threshold estimation*, Electron. Commun. Probab. **17** (2012), no. 29, 10.

[27] R Cont and P. Tankov, *Financial modelling with jump processes.* (2004.)

[28] M. Csorgo and L. Horv´ ath, *Limit theorem in change-point analysis.*, J. Wiley, New York. (1997).

[29] Ju. A. Davydov, *The convergence of distributions which are generated by stationary random processes*, Teor. Verojatnost. i Primenen. **13** (1968), 730–737.

[30] Davy M. et Doncarli C. Desobry F., *An online kernel change detection algorithm.*, Signal Processing, IEEE Transactions on. **58** (2012).

[31] Tourneret J.-Y. Dobigeon N. and J. D. Scargle, *Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model.*, Signal Processing, IEEE Transactions on. **55** (2007), no. 2, 414–423.

[32] L Dong and S. Ling, *On the least estimations of threshold autoregressive and moving-average models.*, Statistics and its Interface. **4** (2011), 183–196.

[33] Hastie T.-Johnstone I. Tibshirani R. Efron B. and al., *Least angle regression.*, The annals of Statistics. **32** (2004), no. 2, 407–499.

[34] P. H. Eilers and R. X. De Menezes, *Quantile smoothing of array cgh data.*, Bioinformatics. **21** (2005), no. 1, 1146–1153.

[35] M. Elmi, *Multiple change point detection by filtered derivative and false discovery rate method.*, International Journal of Statistics and Probability **3** (2013Decembre), no. 1, 12–23.

[36] ———, *A real application of filtered derivative and false discovery rate.*, Paper-sjds14.sfds.asso.fr submission 23,Rennes,France.Société Francaise de la Statistique. **3** (201401Juin).

[37] ———, *Multiple change point detection in weakly random variable using filtered derivative and false discovery rate method.*, World Statistics Congress(WSC), Marrakech. **3** (2017July 16).

[38] R. F. Engle, Econometrica, 9871008., Title = Autoregressive conditional heteroscedasticity with estimates of the variance of U.K., Volume = 50, Url = , Year = 1982.

[39] J. Fan and Q. Yao, *Nonlinear time series: Nonparametric and parametric methods.*, Springer-Verlag, New York. (2003).

[40] P. Fearnhead, *Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model.* **55** (2007), no. 2, 414–423.

[41] P. Fearnhead and Z. Liu, *On-line inference for multiple changepoint problems.*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69** (2007), no. 4, 589–605.

[42] C Francq and J M. Zakoian, *Estimating linear representations of nonlinear processes.*, Series des documents de travail du Crest, Centre de Recherche en Économie et Statistique. INSEE. **21** (1995November), no. 9562.

[43] Christian Francq and Jean-Michel Zakoïan, *Estimating linear representations of nonlinear processes*, J. Statist. Plann. Inference **68** (1998), no. 1, 145–165.

[44] Christian Francq and Jean-Michel Zakoïan, *Garch models: Structure, statistical inference and financial applications*, Wiley, 2010.

[45] E. Gombay and S. Liu, *A nonparametric test for change in randomly censored data..*, The Canadian Journal of Statistics/La Revue Canadienne de Statistique. (2000), 113–321.

[46] Borgwardt K. M. Rasch M. Schölkopf B. Gretton A. and A. J. Smola, *A kernel method for the two-sample-problem.*, In Advances in neural information processing systems. (2006), 513–520.

[47] B.E. Hansen, *Inference in tar models.*, Studies in Nonlinear Dynamics and Econometrics. **2** (1997), 1–14.

[48] ———, *Sample splitting and threshold estimation.*, Econometrics. **68** (2000), 575–603.

[49] ———, *Threshold autoregressions in economics.*, Statistics and its Interface. **4** (2011), 123–127.

[50] C. Harchaoui Z. et Lévy-Leduc, *Catching change-points with lasso.*, In Advances in Neural Information Processing Systems, . (2008), 617–624.

[51] Z Harchaoui and C. Lévy-Leduc, *Segmentation temporelles de signaux à l'aide du lasso.*, Colloque Gretsi, 11-14 septembre 2007, Troyes. (2007), 401–404.

[52] Norbert Herrndorf, *A functional central limit theorem for weakly dependent sequences of random variables*, Ann. Probab. **12** (1984), no. 1, 141–153.

[53] M. Huskovà and S. Meintanis, *Change point analysis based on the empirical characteristic functions of ranks.*, J. Wiley, New York. **25** (2006a), 421–436.

[54] Suzuki T. Kanamori T. and M. Sugiyama, *f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models.*, Information Theory, IEEE Transactions on. **58** (2012).

[55] M. Lavielle and E. Lebarbier, *An application of mcmc methods for the multiple change-points problem.*, Signal Processing. **81** (2001), no. 1, 39–53.

[56] M. Lavielle and E. Moulines, *Least-squares estimation of an unknown number of shifts in a time series.*, J. Time Ser. Anal. **21** (2000), no. 1, 33–59.

[57] M. Lavielle and G. Teyssière, *Detection of multiple change points in multivariate time series.*, Lithuanian Math. J. **46** (2006), 287–306.

[58] E. Lebarbier, *Detecting multiple change-points in the mean of gaussian process by model selection.*, Signal Processing. **85** (2005), 717–736.

[59] E. L. Lehmann and H. J. D'Abrera **1**.

[60] Dong Li, Shiqing Ling, and Wai Keung Li, *Asymptotic theory on the least squares estimation of threshold moving-average models*, Econometric Theory **29** (2013), no. 3, 482–516.

[61] Shiqing Ling and H. Tong, *Testing for a linear ma model against threshold ma models.*, The annals of statistics. **33** (2005.), no. 6, 2529–2552.

[62] Lévy-Leduc C. et Cappé O. Lung-Yut-Fong A., *Robust changepoint detection based on multivariate rank statistics. in acoustics, speech and signal processing (icassp).*, IEEE International Conference on. (2011), 3608–3611.

[63] ———, *Robust retrospective multiple change-point estimation for multivariate data..*, In Statistical Signal Processing Workshop (SSP). (2011), 405–408.

[64] D. S. Matteson and N. A. James, *A nonparametric approach for multiple change point analysis of multivariate data.*, Journal of the American Statistical Association,. **109** (2014), no. 505, 334–345.

[65] Harchaoui Z. Moulines E. and F. Bach, *Testing for homogeneity with kernel fisher discriminant analysis.*, Advances in Neural Information Processing Systems. (2008).

[66] Jacques Neveu, *Mathematical foundations of the calculus of probability*, Translated by Amiel Feinstein, Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1965.

[67] T. Park and G. Casellas, *The bayesian lasso.*, Journal of the American Statistical Association. **103** (2008), no. 482, 681–686.

[68] D Joseph Petrucelli, *On the consistency of least squares estimators for a threshold ar(1) model .*, Journals of Time Series Analysis. **21** (1986), 267–279.

[69] D Joseph Petrucelli and S. W Woolford, *A threshold ar(1) model.*, Journals of Applied probability. **21** (1984), 267–279.

[70] A. Pettitt, *A non-parametric approach to the change-point problem.*, Applied statistics. **1** (1979), 126–135.

[71] David Pollard, *Convergence of stochastic processes*, Springer Series in Statistics, Springer-Verlag, New York, 1984.

[72] Koenker R. Portnoy S. and al., *The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators.*, Statistical Science. **12** (1997), no. 4.

[73] L Qian, *On maximum likelihood for a threshold autoregressive.*, Journal of Statistical Planning and Inference. **7** (1998), 21–46.

[74] C. P. Robert, *Le choix bayésien.*, Springer. (2006).

[75] L Rudin, S Osher, and Joel S., *Nonlinear total variation based noise removal algorithms.*, Physics D, Nonlinear Phenomena. **60** (1992), no. 1, 295–268.

[76] E. Seijo and B. Sen, *A continuous mapping theorem for the smallest argmax functional. .*, Electronic Journal of Statistics. **5** (2011), 421–439.

[77] Emilio Seijo and Bodhisattva Sen, *Change-point in stochastic design regression and the bootstrap*, Ann. Statist. **39** (2011), no. 3, 1580–1607.

[78] T. Terasvirta and H. M. Anderson, *Characterizing nonlinearities in business cycles using smooth transition autoregressive models.*, Applied Econometrics . **7** (1992December), S119–S136.

[79] R. J. Tibshirani and T. B. Arnold, *Regression shrinkage and selection via the lasso.*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.

[80] Ryan J. Tibshirani and Jonathan Taylor, *The solution path of the generalized lasso.*, Stanford University. (2011).

[81] Saunders M. Rosset-S. Zhu J. Tibshirani R. and K. Knight, *Sparsity and smoothness via the fused lasso.*, Journal of the Royal Statistical Society. Series B (Methodological) **67** (2005), no. 1, 91–108.

[82] H Tong, *Threshold models in non-linear time series analysis .*, Lecture Notes in Statistics, Springer-Verlag: New York. **62** (1983).

[83] ———, *Non-linear time series : A dynamical system approach .*, Oxford University. Press, New York. (1990).

[84] S Tsay R, *Testing and modeling threshold autoregressive processes .*, J. Amer. Statist. Assoc. **84** (1989), 231–240.

[85] U. Yule, *On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers.*, Philos. Trans. R. Soc. Lond. Ser. A. **226** (1927), 267–298.

[86] Liu Y. Qin-P. et Wang Z. Zou C., *The adaptive lasso and its oracle properties.*., Journal of the American statistical association. **101** (2006), no. 476, 374–382.