



Segmentation de processus avec un bruit autorégressif

Souhil Chakar

► To cite this version:

Souhil Chakar. Segmentation de processus avec un bruit autorégressif. Statistiques [math.ST]. Université Paris Sud - Paris XI, 2015. Français. NNT : 2015PA112196 . tel-01224745

HAL Id: tel-01224745

<https://theses.hal.science/tel-01224745>

Submitted on 5 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE MATHÉMATIQUES DE LA RÉGION PARIS-SUD
LABORATOIRE MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES, UMR 518
AGROPARISTECH/INRA

DISCIPLINE : MATHÉMATIQUES

THÈSE DE DOCTORAT

Soutenue le 22 septembre 2015 par

Souhil Chakar

Segmentation de processus avec un bruit autorégressif

Directeur de thèse : M. Stéphane Robin
Co-directeur de thèse : Mme. Émilie Lebarbier

Directeur de recherche (INRA)
Maître de conférence (AgroParisTech)

Composition du jury :

Président du jury :	M. Éric Moulines	Professeur (École Polytechnique)
Rapporteurs :	M. Jean-Marc Bardet	Professeur (Université Paris 1)
	M. Christophe Croux	Professeur (Katholieke Universiteit Leuven)
Examineurs :	M. Christophe Giraud	Professeur (Université Paris-Sud)
	M. Éric Moulines	Professeur (École Polytechnique)
Invités :	Mme. Céline Lévy-Leduc	Professeur (AgroParisTech)

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse. À Stéphane Robin, Émilie Lebarbier ainsi que Céline Lévy-Leduc. J'ai eu beaucoup de plaisir à travailler avec vous et je regrette déjà l'équipe que nous formions. Merci pour tout ce que j'ai appris à votre contact durant ces années, pour vos encouragements, votre disponibilité et votre patience. Sincèrement, je ne crois pas que j'aurais pu bénéficier d'un meilleur encadrement que celui-ci.

Je remercie les rapporteurs de cette thèse, Jean-Marc Bardet et Christophe Croux, d'avoir accepté de relire ce manuscrit ainsi que d'avoir apporté des observations importantes à ce travail. Merci aussi à Éric Moulines et Christophe Giraud d'avoir bien voulu faire partie du jury de cette thèse.

Je remercie tous ceux avec qui j'ai eu des échanges scientifiques, que ces échanges concernent cette thèse ou s'en éloignent. En particulier, je remercie Michel Koskas et sa disponibilité pour m'apprendre certains problèmes spécifiques de l'algorithmique ainsi qu'Alice Cleynen pour tous ces échanges sur la segmentation et la loi binomiale négative. Je remercie aussi tous ceux qui m'ont permis d'exposer mes travaux et ceux qui, lors de ces exposés, ont témoigné d'un intérêt.

Je tiens à remercier tous les membres et ex-membres du laboratoire MIA et du département MMIP, les collègues administratifs et informaticiens, ainsi que tous ceux avec qui j'ai pu être en contact à l'Agro. En particulier, je souhaite remercier mes ex-colocataires de bureau, Loïc et Marie. Ceux du bureau d'à côté, et encore à côté, Tristan, Marie, Pierre et Anna, pour leur bonne humeur communicative. Vincent pour son attention portée à la ponctuation et à la typographie. Liliane et Sarah pour leur aide dans la dernière année. Merci à Benjamin pour sa bonne humeur et à Francine pour sa bienveillance.

À Orsay, je remercie tous ceux qui m'ont facilité les choses d'une manière ou d'une autre. En particulier, Valérie Blandin-Lavigne.

Je remercie aussi tous ceux qui, à Lyon et à Grenoble, se sont enquis de l'avancement de ma thèse. En particulier, je remercie Cécile et Ester avec qui j'ai partagé mon bureau pendant un an. Merci aussi à Jean Bérard qui m'a orienté vers mes directeurs de thèse.

Je témoigne ma gratitude à l'ensemble de la « noblesse de canapé », sans qui cette thèse n'aurait pas été possible. On a frôlé la fronde, mais c'est fini. Merci à l'ensemble des mes amis, amis d'amis et tous les autres qui se sont intéressés à ce que je faisais.

Mes derniers remerciements vont à ma famille, en particulier à mes parents, à mes proches

et à ma compagne. Je ne peux pas les remercier complètement ici, au risque de revoir sérieusement la taille de ce manuscrit. Mention spéciale pour l'ensemble de leur œuvre.

Malgré les singularités, ces remerciements sont largement collectifs. Merci à tous, donc.

Sommaire

1	Introduction	6
1.1	Segmentation dans le cas d'observations indépendantes	7
1.2	Segmentation dans le cas d'observations dépendantes	14
1.3	Approche adoptée	16
2	Outils mathématiques et algorithmiques	18
2.1	Séries temporelles	19
2.2	Processus empirique et Delta-méthode fonctionnelle	23
2.3	Estimation robuste	26
2.4	Inégalités maximales	30
2.5	Sélection de modèle	32
2.6	Problèmes algorithmiques en segmentation	40
3	A robust approach for estimating change-points in the mean of an AR(1) process	50
3.1	Introduction	52
3.2	Robust estimation of the parameter ρ^*	54
3.3	Change-points and expectations estimation	55
3.4	Selecting the number of change-points	57
3.5	Numerical experiments	59
3.6	Conclusion	68
3.7	Proofs	68
4	A robust approach for estimating change-points in the mean of an AR(p) process	88
4.1	Introduction	90
4.2	Robust estimation of the autoregression coefficients	92
4.3	Change-points and expectations estimation	93
4.4	Selecting the number of change-points	94
4.5	Numerical experiments	96
4.6	Proofs	99
4.7	Tables and figures	109

5 Commentaires et discussion	124
5.1 Commentaires	125
5.2 Limites de l'approche adoptée	131
5.3 Perspectives	134
Bibliographie	140
Liste des illustrations	148
Liste des tableaux	150
Notations	151

Chapitre 1

Introduction

Contenu

1.1	Segmentation dans le cas d'observations indépendantes	7
1.1.1	Typologie des problèmes de segmentation	8
1.1.2	Modèles	8
1.1.3	Méthodes bayésiennes	10
1.1.4	Cadre fréquentiste	10
1.1.5	Sélection du nombre de ruptures	12
1.1.6	Problème algorithmique	13
1.2	Segmentation dans le cas d'observations dépendantes	14
1.2.1	Plusieurs types de ruptures	14
1.2.2	Ruptures dans la loi marginale	14
1.2.3	Autres types de ruptures	15
1.2.4	Sélection de modèle	16
1.3	Approche adoptée	16

L'objet de ce chapitre est, d'une part, de présenter la segmentation de processus et son rapport à la dépendance entre les observations et, d'autre part, de délimiter l'objet de cette thèse.

La segmentation de processus a de nombreux champs d'application, comme :

- la climatologie (par exemple Peterson *et al.* 1998 ; Aguilar *et al.* 2003 ; Beaulieu, Ouarda *et al.* 2007 ; Beaulieu, J. Chen *et al.* 2012),
- l'économétrie (Stock et Watson, 1996 ; Bai et Perron, 2003 ; Pesaran *et al.* 2006 ; Geweke et Jiang, 2011),
- la génomique (J. V. Braun et Müller, 1998 ; J. V. Braun, R. K. Braun *et al.* 2000 ; Picard, Robin *et al.* 2005 ; Hocking *et al.* 2013),
- la géodésie (Williams, 2003 ; Kenyeres et Bruyninx, 2004 ; Perfetti, 2006 ; Gazeaux *et al.* 2013),
- l'hydrométrie (Cobb, 1978 ; Jarušková, 1997 ; Perreault *et al.* 2000 ; Ehsanzadeh *et al.* 2011),
- l'imagerie médicale (Kaplan et Shishkin, 2000 ; Lavielle, 2005),
- la stylométrie (H. Chen et Zhang, 2015).

Formulé de façon générale, le problème est le suivant : étant donné des observations $y_1, \dots, y_n$, il s'agit d'identifier les segments sur lesquels la série $y_i, \dots, y_j$ ($i < j$) est homogène, en un sens à préciser. Les bornes de ces segments sont les *ruptures*. Elle marquent l'instant de la perte de cette propriété d'homogénéité. Les questions qui se posent sont alors :

- combien y a-t-il de ruptures ?
- quelle sont leurs positions ?
- quelles sont les propriétés de la série à l'intérieur de chaque segment ? ses propriétés globales en dehors de la segmentation ? On peut penser à des problèmes de sélection de modèle, d'estimation, de test, etc.

1.1 Segmentation dans le cas d'observations indépendantes

Dans le cas où les observations sont indépendantes, la segmentation se réduit à la détection de ruptures dans la loi marginale des observations. Formellement, on dispose de variables aléatoires indépendantes $y_1, \dots, y_n$, sur un même espace mesurable. Les ruptures sont alors les entiers $0 = t_0 < t_1 < \dots < t_m < t_{m+1} = n$ tels que pour tout entier $1 \leq k \leq m + 1$,

$$\begin{cases} y_{t_{k-1}+1}, \dots, y_{t_k} & \text{sont de même loi.} \\ y_i \text{ et } y_j & \text{sont de lois différentes si } k \leq m \text{ et } t_{k-1} < i \leq t_k < j \leq t_{k+1}. \end{cases}$$

1.1.1 Typologie des problèmes de segmentation

Historiquement, les problèmes de segmentation sont apparus comme le problème de la détection d'une rupture (Page, 1954, 1955). Il s'agit alors de tester

$$\begin{aligned}
 &H_0 : && y_1, \dots, y_n \text{ identiquement distribués} \\
 &\text{contre} \\
 &H_1 : \exists t \in \{1, \dots, n-1\}, \left\{ \begin{array}{ll} y_1, \dots, y_t & \text{identiquement distribués,} \\ y_{t+1}, \dots, y_n & \text{identiquement distribués,} \\ y_1 \text{ et } y_n & \text{sont de lois différentes.} \end{array} \right.
 \end{aligned}$$

Des conditions supplémentaires sur la loi des observations (appartenance à un certain modèle, etc.) peuvent être fixées. Si l'hypothèse nulle H_0 est rejetée, il est aussi question d'estimer l'instant de rupture t . La détection de rupture peut être *séquentielle* (*on-line*) ou *rétrospective* (*off-line*). Dans le premier cas, la taille n de la série croît et pour chaque n on teste la présence d'une rupture ou non (voir par exemple Basseville et Nikiforov, 1993). Typiquement, il s'agit de réaliser un compromis entre la détection à temps d'une rupture et la nécessité d'éviter les fausses alarmes. Dans le cas de la détection de rupture rétrospective, la totalité de la série $y_1, \dots, y_n$ est observée et, à posteriori, on teste la présence d'une rupture. La détection de rupture séquentielle et les problèmes spécifiques qu'elle pose ne seront pas abordés dans cette thèse.

Si le cas où au moins deux ruptures peuvent être présentes est possible, on parle alors de *détection de ruptures multiples* ou de *segmentation à ruptures multiples*. La méthodologie du cas où il y a au plus une rupture dans la série peut être adaptée. En particulier, comme Bai et Perron (1998), on peut tester l'absence de rupture contre la présence de m ruptures, ou bien la présence de m ruptures contre la présence de $m+1$ ruptures. Toutefois, en cas de ruptures multiples, l'estimation du nombre de ruptures par l'intermédiaire de ces tests mène à un problème complexe de tests multiples. Une méthode fréquemment adoptée consiste à

- dans le modèle à m ruptures, estimer les instants de ruptures ainsi que les autres paramètres du modèle,
- sélectionner un modèle à m de ruptures, à l'aide d'une procédure de sélection de modèle.

Toutefois, certaines méthodes bayésiennes (Lavielle et Lebarbier, 2001) ou fréquentistes (comme le Lasso, pour *Least absolute shrinkage and selection operator*, voir Harchaoui et Lévy-Leduc, 2010) réalisent ces deux étapes simultanément.

Dans cette thèse, nous nous intéressons à la segmentation rétrospective à ruptures multiples.

1.1.2 Modèles

Dans la section 1.1.1, nous n'avons fixé aucune condition sur les lois marginales possibles des observations. Or il existe bien souvent une modélisation préalable de la série.

On peut se donner un modèle paramétrique $(P_\theta)_{\theta \in \Theta}$ et fixer comme condition

$$\forall k \in \{1, \dots, m+1\}, \begin{cases} \forall i \in \{t_{k-1}^* + 1, \dots, t_k^*\}, y_i \sim P_{\theta_k^*} \\ \text{et } k \leq m \Rightarrow \theta_k^* \neq \theta_{k+1}^* . \end{cases} \quad (1.1)$$

On peut ajouter à ce modèle un paramètre global, qui reste constant sur tous les segments. Par exemple, le modèle gaussien est ici :

$$\forall k \in \{1, \dots, m+1\}, \begin{cases} \forall i \in \{t_{k-1}^* + 1, \dots, t_k^*\}, y_i \sim \mathcal{N}(\mu_k^*, \sigma_k^{*2}) \\ \text{et } k \leq m \Rightarrow (\mu_k^*, \sigma_k^{*2}) \neq (\mu_{k+1}^*, \sigma_{k+1}^{*2}) . \end{cases} \quad (1.2)$$

Le modèle peut être restreint au cas où toutes les variances σ_k^{*2} sont égales, ce qui fait de cette variance commune un paramètre global du modèle. De même, on peut se restreindre au cas où toutes les espérances sont égales. Dans le premier cas, on segmente dans l'espérance, dans le second cas, dans la variance (voir figure 1.1).

Il peut être préférable dans certains cas d'étendre le modèle. On peut simplement demander à ce que certaines quantités associées aux lois marginales soient constantes sur les segments. Par exemple, par analogie avec le modèle gaussien, on peut considérer le modèle suivant :

$$\forall k \in \{1, \dots, m+1\}, \begin{cases} \forall i \in \{t_{k-1}^* + 1, \dots, t_k^*\}, y_i \text{ est de carré intégrable, } \mathbb{E}y_i = \mu_k^*, \text{Var}(y_i) = \sigma_k^{*2} \\ \text{et } k \leq m \Rightarrow (\mu_k^*, \sigma_k^{*2}) \neq (\mu_{k+1}^*, \sigma_{k+1}^{*2}) . \end{cases} \quad (1.3)$$

Tous les sous-modèles définis dans le cas gaussien peuvent être définis de façon analogue. Dans le cas où l'on segmente dans l'espérance, il s'agit d'un modèle de régression non-linéaire, où la

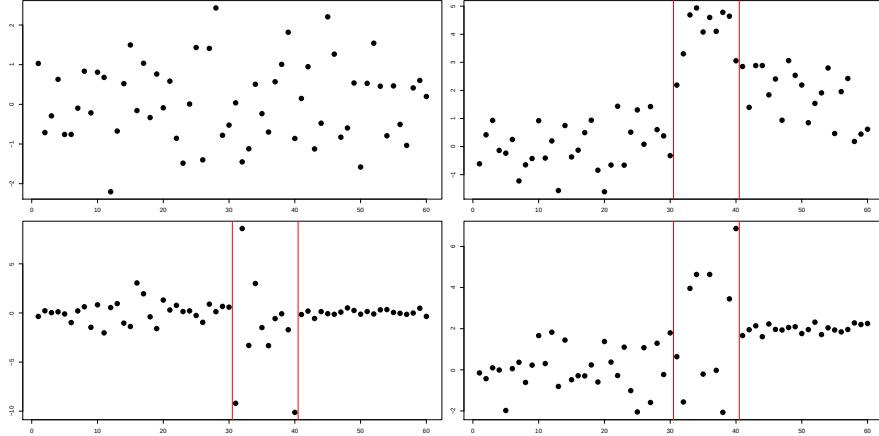


FIGURE 1.1 – Réalisation de 4 séries de 60 v.a.r. indépendantes de loi normale. Les droites verticales indiquent les ruptures éventuelles, localisées en 30 et 40. Haut, gauche : cas i.i.d. centré réduit. Haut, droite : variance constante égale à 1, ruptures dans l'espérance (0, 4 puis 2). Bas, gauche : espérance constante égale à 0, ruptures dans la variance (1, 16 puis 1/16). Bas, droite : ruptures simultanées dans l'espérance et la variance, avec les mêmes valeurs que précédemment.

fonction de régression est constante par morceaux. Si les instants de rupture étaient connus, il s'agirait d'un modèle linéaire.

1.1.3 Méthodes bayésiennes

Les méthodes bayésiennes en segmentation ont été introduites d'abord dans le cas où il y a au plus une rupture (Chernoff et Zacks, 1964), puis dans le cas des modèles à ruptures multiples (Barry et Hartigan, 1992). Dans le cadre bayésien, tous les paramètres du modèle, notamment le nombre et la position des ruptures, sont des variables aléatoires, qui suivent une loi à priori.

L'objectif de l'inférence est alors de fournir les lois à posteriori

$$P(t_1, \dots, t_m | y_1, \dots, y_n), \quad P(m | y_1, \dots, y_n) .$$

Le calcul de cette loi à posteriori n'est pas nécessairement trivial. Il y a trois façons de procéder :

- Les lois à priori sont choisies de façon à obtenir des expressions analytiques calculables pour les lois à posteriori (Rigaill *et al.* 2012). Cette approche est classique en statistique bayésienne et fait usage des lois conjuguées.
- On renonce à obtenir une expression analytique pour la loi à posteriori. Nous n'avons accès à cette loi de probabilité qu'à travers des échantillons simulés de cette loi. La simulation de tels échantillons peut se faire à l'aide de méthodes de Monte-Carlo par chaînes de Markov (Lavielle et Lebarbier, 2001), où la loi des échantillons simulés tend vers la loi à posteriori quand le temps d'exécution de l'algorithme tend vers l'infini. Il faut alors être attentif à ce que le temps d'exécution de l'algorithme soit suffisamment long (Fearnhead, 2006). Un autre algorithme, de type *forward-backward*, permet une simulation exacte de la loi à posteriori (Fearnhead, 2006).
- L'approche asymptotique, pour la loi du nombre de ruptures seulement, qui consiste à fournir une approximation des probabilités à posteriori quand le nombre d'observations n tend vers l'infini. Cette approche a été développée pour l'estimation du nombre de ruptures par Zhang et Siegmund (2007, voir section 2.5.4). Il s'agit de l'adaptation à la segmentation du critère de sélection de modèle proposé par Schwarz (1978).

Parmi ces trois approches seule la dernière sera abordée.

1.1.4 Cadre fréquentiste

Avant d'aborder les résultats asymptotiques établis en segmentation, il faut donner un sens à la théorie asymptotique dans ce cadre.

Pour pouvoir établir des résultats de consistance sur les estimateurs des ruptures, nous devons définir la suite des vrais instants de rupture quand la longueur de la série varie.

Un cadre fréquemment adopté et ayant permis d'établir de nombreux résultats asymptotiques (Bai, 1994 ; Bai et Perron, 1998 ; Lavielle et Moulines, 2000 ; Bardet, Kengne et Wintenberger, 2012) est le suivant. On considère $(y_{i,n})_{1 \leq i \leq n, n \geq n_0}$. Pour chaque $n \geq n_0$, $(y_{i,n})_{1 \leq i \leq n}$ est une série affectée par des ruptures. On suppose qu'il existe $0 = \tau_0^* < \tau_1^* < \dots < \tau_m^* < \tau_{m+1}^* = 1$, tels que, pour tout $n \geq n_0$, les ruptures de la série $(y_{i,n})_{1 \leq i \leq n}$ sont localisées en

$$t_{k,n}^* = \lfloor n\tau_k^* \rfloor \quad (1 \leq k \leq m) . \quad (1.4)$$

Ainsi, les instants de rupture sont des fractions de la longueur totale de la série. Les paramètres à estimer correspondant aux instants de rupture sont donc $\tau_1^*, \dots, \tau_m^*$, des réels compris entre 0 et 1. On considère que le nombre de ruptures ne varie pas avec la taille des séries.

C'est le cadre sous-jacent de la plupart des travaux concernant l'estimation des instants de rupture en statistique fréquentiste depuis le début des années 1990, même si $(y_{i,n})_{1 \leq i \leq n, n \geq n_0}$ n'est pas explicitement défini de façon systématique. En général, on définit une série $y_1, \dots, y_n$, puis on fait tendre n vers l'infini pour obtenir des résultats asymptotiques. En réalité, cela n'a de sens qu'à travers $(y_{i,n})_{1 \leq i \leq n, n \geq n_0}$.

Illustrons cette construction dans le cas gaussien. On considère le modèle défini à n fixé par l'équation (1.2), avec la contrainte supplémentaire d'avoir toutes les variances égales (à σ^{*2}). Soit $0 = \tau_0^* < \tau_1^* < \dots < \tau_m^* < \tau_{m+1}^* = 1$ et $(\epsilon_t)_{0 < t \leq 1, t \in \mathbb{Q}}$ une famille de v.a.r. i.i.d. gaussiennes centrées de variance σ^{*2} . Soit $\mu : (0, 1] \rightarrow \mathbb{R}$ définie par

$$\forall t \in (0, 1], \mu(t) = \sum_{k=0}^m \mu_k^* \mathbf{1}_{(\tau_k^*, \tau_{k+1}^*]}(t) .$$

Pour tout n , la loi de $(y_{i,n})_{1 \leq i \leq n}$ est donnée par :

$$(y_{i,n})_{1 \leq i \leq n} \stackrel{d}{=} \left(\mu \left(\frac{i}{n} \right) + \epsilon_{i/n} \right)_{1 \leq i \leq n} .$$

Estimation par maximum de vraisemblance

On se place dans le modèle défini par (1.1) et le cadre asymptotique défini par (1.4). Supposons que les lois de probabilité P_θ admettent une densité f_θ par rapport à une certaine mesure de référence. L'estimateur du maximum de vraisemblance des ruptures et des paramètres au sein de chaque segment est alors

$$(\widehat{t}_1, \dots, \widehat{t}_m, \widehat{\theta}_0, \dots, \widehat{\theta}_m) = \arg \max_{\substack{0=t_0 < t_1 < \dots < t_m < t_{m+1}=n \\ \theta_0, \dots, \theta_m \in \Theta}} \prod_{k=0}^m \prod_{i=t_k+1}^{t_{k+1}} f_{\theta_k}(y_k) . \quad (1.5)$$

La normalisation $\widehat{\tau}_k = \widehat{t}_k/n$ donne des estimateurs des ruptures τ_k^* . He et Severini (2010) ont démontré, sous certaines conditions de régularité, que les estimateurs $\widehat{\tau}_k$ sont consistants et convergent à l'ordre $1/n$. De plus, les estimateurs des paramètres $\widehat{\theta}_k$ sont eux aussi consistants et conjointement asymptotiquement normaux, y compris conjointement à un éventuel paramètre global.

Estimation par moindres carrés

On se place dans le modèle (1.3), avec une variance constante, et le cadre asymptotique défini par (1.4). L'estimateur des moindres carrés des ruptures et des espérances est alors :

$$(\hat{t}_1, \dots, \hat{t}_m, \hat{\mu}_0, \dots, \hat{\mu}_m) = \arg \min_{\substack{0=t_0 < t_1 < \dots < t_m < t_{m+1}=n \\ \mu_0, \dots, \mu_m}} \sum_{k=0}^m \sum_{i=t_k+1}^{t_{k+1}} (y_i - \mu_k)^2. \quad (1.6)$$

À ma connaissance, les premiers résultats asymptotiques dans ce cadre sont ceux de Yao et Au (1989). Lavielle et Moulines (2000) fournissent des résultats similaires à ceux de He et Severini (2010) pour l'estimateur du maximum de vraisemblance dans ce cas.

Estimation par maximum de vraisemblance gaussienne

On se place dans le modèle (1.3) et le cadre asymptotique défini par (1.4). Les ruptures et les autres paramètres peuvent être estimés par maximisation de la vraisemblance gaussienne, qui est la fonction de vraisemblance associée au modèle (1.2). Même si les observations ne sont pas gaussiennes, Lavielle (1999) garantit sous certaines conditions la consistance des estimateurs et des vitesses de convergences de même ordre que celles de He et Severini (2010).

1.1.5 Sélection du nombre de ruptures

On a supposé jusqu'ici le nombre de ruptures m connu. Mais, dans la plupart des applications, il doit être estimé, ce qui nous amène à un problème de sélection de modèle. Des critères de différents types ont été proposés :

- des procédures séquentielles de tests (Bai et Perron, 1998),
- des approches par contraste pénalisé dans un cadre fréquentiste (Yao, 1988 ; C.-B. Lee, 1995 ; Lebarbier, 2005),
- des approches bayésiennes (Zhang et Siegmund, 2007 ; Rigai et al. 2012).

Une présentation plus détaillée de ces différentes approches est donnée dans la section 2.5. Si l'estimation des instants de rupture et des paramètres des lois marginales est un problème bien compris à m connu, l'estimation de ce dernier reste un problème difficile. La définition même du « bon » ou du « vrai » m à sélectionner n'est pas consensuelle. De plus, les résultats asymptotiques ne suffisent pas toujours à déterminer une procédure pratique, notamment quand des constantes de pénalisation restent à calibrer.

Dans ce travail, on porte une attention particulière au critère de Zhang et Siegmund (2007), qui est défini indépendamment de toute constante, et qui fournit de bons résultats pratiques (Picard, Lebarbier et al. 2011 ; Frick et al. 2014).

1.1.6 Problème algorithmique

L'estimation de ruptures multiples par maximum de vraisemblance ou moindres carrés fait apparaître un problème algorithmique. Cette question est centrale en segmentation. Par exemple, Picard (2007) et Jandhyala *et al.* (2013), dans leur exposé, introduisent l'estimation de ruptures multiples à travers le problème algorithmique que cela pose. En effet, l'exploration de l'ensemble des segmentations prendrait un temps $O(n^m)$, une complexité algorithmique souvent prohibitive. Pour la recherche de la segmentation optimale, trois approches peuvent alors être considérées :

- Établir un algorithme efficace donnant la solution optimale. Seul l'algorithme de programmation dynamique (Auger et Lawrence, 1989 ; Bai et Perron, 2003), ainsi que ses versions élaguées (Rigaill, 2010 ; Maidstone *et al.* 2014) permettent de recouvrer la segmentation optimale en un temps polynomial. L'algorithme de partition optimale de Jackson *et al.* (2005), ainsi que ses versions élaguées (Killick *et al.* 2012 ; Maidstone *et al.* 2014), fournissent eux aussi la segmentation optimale en réalisant simultanément la sélection du nombre de ruptures, avec certaines contraintes sur le critère de sélection de modèle. Tous ces algorithmes peuvent être utilisés pour peu que le critère à optimiser soit additif sur les segments (voir section 2.6).
- Définir une heuristique fournissant une solution sous-optimale acceptable. La segmentation binaire (voir par exemple Scott et Knott, 1974 ; Venkatraman, 1992 ; Bai, 1997) peut être vue comme une suite de tests sur la présence d'une rupture *supplémentaire* à celles qui ont déjà été estimées. Cette contrainte, le fait que les m ruptures estimées dans le modèle à m changements sont incluses dans les $m+1$ ruptures du modèle avec un changement supplémentaire, est très forte et entraîne une sous-exploration de l'ensemble des segmentations possibles, d'où la solution sous-optimale. Le problème est connu et a mené à des raffinements de la méthode (Olshen *et al.* 2004 ; Fryzlewicz, 2014) pour en réduire l'impact.

Davis, T. C. M. Lee *et al.* (2006) proposent un algorithme génétique fournissant une solution approchée à la segmentation par minimisation d'un critère MDL (*Minimum description length*). Toutefois, Eckley *et al.* (2011) font remarquer que ce critère peut tout aussi bien être minimisé par l'algorithme de programmation dynamique.

- Modifier le problème d'optimisation pour le rendre résoluble. En segmentation, le Lasso, tel que formulé par Harchaoui et Lévy-Leduc (2010), revient à remplacer la contrainte « m ruptures » par une contrainte convexe. Là aussi, la modification du problème d'optimisation fait que la segmentation proposée n'est pas optimale au sens du critère initial.

Les choix méthodologiques ont été orientés de façon à préserver l'applicabilité de l'algorithme de programmation dynamique, le plus souvent élaguée. Les algorithmes de partition

optimale n'ont pas été utilisés car le critère de sélection de modèle que nous avons utilisé ne le permettait pas.

1.2 Segmentation dans le cas d'observations dépendantes

Jusqu'à présent, nous avons considéré que les observations des séries temporelles considérées étaient indépendantes. Toutefois, dans un certain nombre d'applications, par exemple en géosciences ou en météorologie¹, cette condition est clairement violée.

1.2.1 Plusieurs types de ruptures

Différents types de ruptures peuvent affecter ces séries dépendantes :

- On pourra toujours considérer des ruptures dans la loi marginale des observations, comme dans la section 1.1, à la différence près que les observations ne sont plus considérées comme indépendantes. Voir Lavielle (1999), Lavielle et Moulines (2000) et, dans un cadre plus général que celui présenté ici, Bai et Perron (1998).
- On pourra considérer des ruptures dans la structure de dépendance de la série. Certains des modèles considérés sont détaillés dans la section 1.2.3.
- On pourra considérer des ruptures qui affectent à la fois la loi marginale des observations et la structure de dépendance de la série.

1.2.2 Ruptures dans la loi marginale

On se place dans le modèle (1.3), en supposant les variances constantes, mais sans supposer les observations indépendantes. Lavielle et Moulines (2000) démontrent que, sous certaines conditions, la dépendance entre les observations n'affecte pas la vitesse de convergence des estimateurs des moindres carrés des ruptures et des paramètres μ_k^* , tels que définis en (1.6).

Plus précisément, dans le modèle considéré par Lavielle et Moulines (2000),

$$\forall i \in \{1, \dots, n\}, y_i = \mu_k^* + \eta_i \text{ si } t_{k-1}^* < i \leq t_k^*, \quad (1.7)$$

où les ruptures vérifient la condition (1.4). Si la condition

$$\exists \phi < 2, \exists C > 0, \forall j > i > 0, \mathbb{E} \left[\left(\sum_{t=i}^j \eta_t \right)^2 \right] \leq C |j - i + 1|^\phi \quad (1.8)$$

est vérifiée, les estimateurs des moindres carrés des ruptures et des paramètres μ_k^* , *modulo* certaines restrictions de l'ensemble des paramètres considérés dans la minimisation du critère des moindres carrés, sont consistants, les estimateurs des ruptures τ_k^* convergeant à la vitesse $1/n$. La condition (1.8) est vérifiée pour une large classe de processus (η_i) . Si on se restreint

¹Références en préambule du chapitre 1.

aux processus centrés faiblement stationnaires (cf. définition 2.3), elle est vérifiée pour les processus à *courte mémoire*, c'est-à-dire tels que la fonction d'autocovariance associée au processus est absolument sommable, mais aussi pour certains processus à *longue mémoire*. Bai et Perron (1998) avaient déjà démontré le résultat sur les estimateurs des ruptures et des μ_k^* , notamment dans le cas où $(\eta_i)_{i \in \mathbb{Z}}$ est un processus centré faiblement stationnaire à courte mémoire. Lavielle (1999) a étendu ces résultats à un cadre plus général que les ruptures de moyennes ainsi qu'à d'autres critères que les moindres carrés.

Concernant la sélection du nombre de ruptures, Lavielle et Moulines (2000) donnent des conditions sur la pénalisation pour obtenir des critères pénalisés consistants. Là aussi, si $(\eta_i)_{i \in \mathbb{Z}}$ reste un processus centré faiblement stationnaire à courte mémoire, ces conditions ne sont pas affectées par la structure de dépendance.

Les résultats de Lavielle et Moulines (2000), ainsi que ceux de Lavielle (1999), pourraient donc plaider pour une non-prise en compte de la dépendance entre observations dans l'estimation des ruptures et la sélection de leur nombre dans un modèle de segmentation, pour peu que la perturbation $(\eta_i)_{i \in \mathbb{Z}}$ vérifie certaines conditions.

Toutefois, Lavielle et Moulines (2000) montrent que la variance asymptotique des estimateurs des ruptures peut être affectée par la structure de dépendance, y compris si $(\eta_i)_{i \in \mathbb{Z}}$ reste un processus centré faiblement stationnaire à courte mémoire. De même, Antoch *et al.* (1997) et Robbins *et al.* (2011) ont montré que la dépendance entre observations, y compris sous la forme d'une perturbation qui serait un processus centré faiblement stationnaire à courte mémoire, n'est pas sans effet sur la variance asymptotique de différentes statistiques de test d'une rupture dans l'espérance. La prise en compte de la relation de dépendance stochastique entre observations dans l'estimation des ruptures peut donc s'avérer importante.

1.2.3 Autres types de ruptures

On peut aussi considérer des ruptures qui affectent d'autres propriétés des séries que la loi marginale, c'est-à-dire qu'elles affectent la structure de dépendance des séries temporelles.

Supposons d'abord que $y_1, \dots, y_n$ sont n observations consécutives d'un *processus stationnaire par morceaux*, c'est-à-dire qu'il existe des ruptures $0 = t_0^* < t_1^* < \dots < t_m^* < t_{m+1}^* = n$ telles que, pour tout entier $k \in \{0, \dots, m\}$, $y_{t_k^*+1}, \dots, y_{t_{k+1}^*}$, sont des observations consécutives d'un processus faiblement stationnaire. Les ruptures vérifient (1.4) et les moments d'ordre un et deux du processus sur chaque segment ne varient pas avec n , de sorte qu'au k^e segment on peut associer l'espérance μ_k^* et la fonction d'autocovariance γ_k^* . Notons que si on impose l'égalité de toutes les fonctions d'autocovariance γ_k^* , on a un modèle de segmentation de l'espérance tel que décrit dans la section 1.2.2.

Ce modèle ne décrit pas les moments d'ordre deux de la série. En particulier, rien n'est dit de la covariance de deux observations séparées par une ou plusieurs ruptures. C'est par exemple le type de modèle considéré par Davis, T. C. M. Lee *et al.* (2006).

La classe de modèles considérée par Bardet, Kengne et Wintenberger (2012) est différente.

Premièrement, car, même en absence de rupture, les séries peuvent être non-stationnaires et deuxièmement, car le comportement de la série aux ruptures est lui aussi différent. Bardet, Kengne et Wintenberger (2012) considèrent des séries $y_1, \dots, y_n$ qui sont des observations successives d'un processus $(y_i)_{i \in \mathbb{Z}}$, tel que, pour tout $t_{k-1}^* < i \leq t_k^*$,

$$y_i = f_{\theta_k^*}((y_j)_{j < i}) + m_{\theta_k^*}((y_j)_{j < i}) \epsilon_i \quad (1.9)$$

où $(\epsilon_i)_{i \in \mathbb{Z}}$ est une suite de v.a.r. i.i.d. centrées de variance 1 et $(f_\theta)_{\theta \in \Theta}$ et $(m_\theta)_{\theta \in \Theta}$ sont des familles de fonctions. Un tel processus est causal, au sens où chaque observation est fonction du passé et d'un aléa indépendant du passé. Davis, Hancock *et al.* (2010) ne proposent qu'un sous-modèle de (1.9), où l'équation (1.9) devient l'équation d'un processus autorégressif dont les coefficients et l'ordre varient entre les segments.

Bardet, Kengne et Wintenberger (2012) estiment les ruptures par quasi-maximum de vraisemblance : la fonction de vraisemblance de $(y_1, \dots, y_n)$ conditionnelle à $(y_j)_{j \leq 0}$ que l'on aurait si $(\epsilon_i)_{i \in \mathbb{Z}}$ était gaussien est maximisée en les paramètres θ_k et en les instants de rupture. Sous certaines hypothèses de régularité et d'identifiabilité du modèle, Bardet, Kengne et Wintenberger (2012) obtiennent les résultats classiques, évoqués dans le cas indépendant dans la section 1.1.4, de consistance des estimateurs des ruptures et des autres paramètres, ainsi que le taux de convergence en $1/n$ pour les estimateurs des ruptures τ_k^* , sous l'hypothèse (1.4).

1.2.4 Sélection de modèle

Le problème du choix du nombre de ruptures se pose également dans le cas dépendant. La sélection de modèle par contraste pénalisé a été étudiée pour des modèles de ruptures dans la loi marginale ou dans la structure de dépendance par Lavielle et Moulines (2000) et Bardet, Kengne et Wintenberger (2012) respectivement. Ces auteurs donnent des conditions suffisantes sur la forme de la pénalité pour garantir la consistance du nombre de ruptures sélectionné. Ces conditions sont cependant très larges et n'aboutissent pas à un critère explicite.

Un critère MDL, fondé sur la théorie de l'information, a également été proposé par Davis, T. C. M. Lee *et al.* (2006) dans le cas des modèles avec une autorégression spécifique à chaque segment. Davis et Yau (2013) généralisent ce critère à d'autres modèles et en démontrent la consistance sous certaines conditions.

Le critère de Zhang et Siegmund (2007) est explicitement restreint au cas où les observations sont indépendantes. Dans les chapitres 3 et 4, on propose une adaptation de ce critère pour certaines structures de dépendance sérielle qui conserve les garanties statistiques et l'efficacité algorithmique.

1.3 Approche adoptée

Dans cette thèse, on s'intéresse à la segmentation dans la moyenne de processus dépendants. On considère le modèle (1.7), en supposant de plus que $y_1, \dots, y_n$ sont n observations successives d'un processus $(y_i)_{i \in \mathbb{Z}}$. Le travail est guidé par deux idées :

- la prise en compte de la dépendance dans l'inférence,
- l'efficacité algorithmique.

Les algorithmes exacts efficaces (de type programmation dynamique) le sont à condition que le contraste à minimiser soit additif en les segments (voir la section 2.6). Or, la prise en compte de la dépendance viole en général cette contrainte. L'objectif de cette thèse est de faire entrer l'inférence pour un processus dépendant dans un cadre admissible pour la programmation dynamique, tout en conservant les garanties statistiques.

Il n'y a aucune rupture dans la structure de dépendance de $(\eta_i)_{i \in \mathbb{Z}}$. Donc, à priori, aucune raison de traiter ce cas comme dans la section 1.2.3. Néanmoins, les approches présentées dans la section 1.2.3 peuvent nous aider à prendre en compte la structure de dépendance de $(\eta_i)_{i \in \mathbb{Z}}$.

On considère que le processus $(\eta_i)_{i \in \mathbb{Z}}$ est centré et faiblement stationnaire et que sa fonction d'autocovariance appartient à une famille paramétrée $(\sigma^2 \gamma_\theta)_{\sigma^2 > 0, \theta \in \Theta}$. Si $(\eta_i)_{i \in \mathbb{Z}}$ est un processus causal, on peut exprimer la quasi-vraisemblance de $(y_1, \dots, y_n)$ conditionnellement à $(y_i)_{i \leq 0}$. Il s'agit de la vraisemblance conditionnelle de $(y_1, \dots, y_n)$ sachant $(y_i)_{i \leq 0}$ si le processus est supposé gaussien.

Dans cette quasi-vraisemblance, les termes correspondant aux ruptures t_k , aux espérances μ_k , à σ^2 et à θ sont liés. À priori, ce critère ne peut pas être maximisé en utilisant l'algorithme de programmation dynamique, qui s'impose dans le cas où les observations sont indépendantes (section 1.1.6).

Le premier problème qui se pose est la présence d'un paramètre θ global. Nous n'avons pas pour objectif ultime d'estimer ce paramètre. Il s'agit d'un paramètre de nuisance, dont l'estimation permettra possiblement d'améliorer l'estimation des autres paramètres.

Le but est alors de construire un tel estimateur de θ , tout en évitant d'avoir à segmenter la série, c'est-à-dire d'estimer les ruptures t_k^* sans connaître θ , ce que l'on cherche précisément à éviter. Nous nous sommes orientés vers des méthodes robustes.

Nous avons mis en œuvre cette méthodologie dans le cas où $(\eta_i)_{i \in \mathbb{Z}}$ est un processus $AR(1)$ dans le chapitre 3 et dans le cas où il s'agit d'un processus autorégressif d'ordre fini arbitraire dans le chapitre 4. Dans ces cas-là, du point de vue de l'inférence des ruptures, l'estimation préalable du paramètre θ , qui est ici le paramètre de l'autorégression, revient à décorréler la série, puis à adopter l'inférence classique du cas indépendant sur les séries ainsi décorrélées.

Dans le chapitre 5, on discute de la méthodologie adoptée, de méthodes alternatives et de résultats additionnels. Les limites de l'approche adoptée sont ensuite discutées avant de tracer quelques perspectives de travaux futurs concernant des résultats mathématiques employés à étendre, l'extension de la méthode employée à une classe de processus plus large ou des problèmes connexes en segmentation permettant de mettre en œuvre une approche similaire.

Chapitre 2

Outils mathématiques et algorithmiques

Contenu

2.1	Séries temporelles	19
2.1.1	Stationnarité, fonctions d'autocovariance et d'autocorrélation	19
2.1.2	Processus ARMA	20
2.1.3	Estimation dans les processus stationnaires d'ordre deux	21
2.2	Processus empirique et Delta-méthode fonctionnelle	23
2.2.1	Processus empirique	23
2.2.2	Delta-méthode fonctionnelle	24
2.2.3	Fonctions d'un processus gaussien	26
2.3	Estimation robuste	26
2.3.1	Point d'effondrement	27
2.3.2	Estimation robuste d'un paramètre d'échelle	27
2.3.3	Estimation robuste des fonctions d'autocovariance et d'autocorrélation d'un processus stationnaire du second ordre	28
2.4	Inégalités maximales	30
2.4.1	Inégalité de Kolmogorov	31
2.4.2	Inégalité de Hájek-Rényi	31
2.4.3	Généralisation de l'inégalité de Hájek-Rényi	31
2.4.4	Utilisation en segmentation	32
2.5	Sélection de modèle	32
2.5.1	Introduction	32
2.5.2	Sélection de modèle par critère pénalisé, cadre fréquentiste	33
2.5.3	Sélection de modèle bayésienne	35
2.5.4	Sélection du nombre de ruptures dans un modèle de segmentation	38
2.6	Problèmes algorithmiques en segmentation	40
2.6.1	Programmation dynamique	41
2.6.2	Programmation dynamique élaguée	45

2.1 Séries temporelles

Dans toute cette section, on s'intéressera à un processus aléatoire réel $X = (X_i)_{i \in \mathbb{Z}}$. On présente ici les propriétés des séries temporelles utilisées dans le reste du manuscrit. Notre référence principale à ce sujet est Brockwell et Davis (1991).

2.1.1 Stationnarité, fonctions d'autocovariance et d'autocorrélation

La loi de X est caractérisée par la famille des fonctions de répartition $(F_i)_{i \in \mathcal{I}}$, où $\mathcal{I} = \{i = (i_1, \dots, i_k) \in \mathbb{Z}; i_1 < \dots < i_k, k \in \mathbb{N}^*\}$ et $F_i(z) = P(X_{i_1} \leq z_1, \dots, X_{i_k} \leq z_k)$, pour tout $z = (z_1, \dots, z_k)$. Parmi ces lois, on souhaite caractériser celles qui sont invariantes par décalage d'indices.

Définition 2.1. X est dit strictement stationnaire si $\forall h \in \mathbb{Z}, \forall i \in \mathcal{I}, F_i = F_{i+h}$, où $i+h = (i_1+h, \dots, i_k+h)$ si $i = (i_1, \dots, i_k)$.

Il s'agit d'une propriété forte. Dans bien des cas, une notion plus faible de stationnarité suffit. Cette dernière étant liée aux moments d'ordre deux, il faut introduire la notion de processus du second ordre.

Définition 2.2. X est dit du second ordre si, pour tout $i \in \mathbb{Z}$, X_i est de carré intégrable.

Définition 2.3. Un processus du second ordre X est dit faiblement stationnaire, ou stationnaire à l'ordre deux, si :

1. $\forall i \in \mathbb{Z}, \mathbb{E}X_i = \mathbb{E}X_0$,
2. $\forall i \in \mathbb{Z}, \forall h \in \mathbb{Z}, \text{Cov}(X_i, X_{i+h}) = \text{Cov}(X_0, X_h)$.

On définit alors la fonction d'autocovariance $\gamma_X : \mathbb{Z} \rightarrow \mathbb{R}$ associée à X par

$$\forall h \in \mathbb{Z}, \gamma_X(h) = \text{Cov}(X_0, X_h),$$

ainsi que la fonction d'autocorrélation $\rho_X : \mathbb{Z} \rightarrow \mathbb{R}$ associée à X par

$$\rho_X = \frac{\gamma_X}{\gamma_X(0)}.$$

ρ_X n'est bien définie que si $\gamma_X(0) > 0$.

En l'absence de spécification, un processus dit *stationnaire* n'est que faiblement stationnaire. L'exemple le plus simple d'un tel processus est le bruit blanc.

Définition 2.4. On dit que X est un bruit blanc de variance $\sigma^2 > 0$ si X est un processus du second ordre faiblement stationnaire tel que $\mathbb{E}X_0 = 0$ et $\forall h \in \mathbb{Z}, \gamma_X(h) = \sigma^2 \delta_{0,h}$.

La stationnarité forte d'un processus du second ordre implique la stationnarité faible. La réciproque est fausse en général.

2.1.2 Processus ARMA

Les processus ARMA sont une classe importante de processus stationnaires. En effet, pour tout $\gamma : \mathbb{Z} \rightarrow \mathbb{R}$ fonction d'autocovariance associée à un processus stationnaire du second ordre et, pour tout $H \in \mathbb{N}$, il existe un processus ARMA X tel que $\forall h \in \mathbb{Z}, |h| \leq H \Rightarrow \gamma_X(h) = \gamma(h)$.

Définition 2.5. B est l'automorphisme de décalage sur l'espace vectoriel des processus aléatoires (réels) du second ordre indicés par $\mathbb{Z}$, défini par $\forall i \in \mathbb{Z}, (BX)_i = X_{i-1}$.

On note $B^j = B \circ \dots \circ B$ (j fois) pour $j \geq 1$, $B^j = B^{-1} \circ \dots \circ B^{-1}$ ($-j$ fois) pour $j \leq -1$ et B^0 l'identité.

Pour tout polynôme (réel ou complexe) $P(z) = \sum_{j=0}^d a_j z^j$, on définit $P(B) = \sum_{j=0}^d a_j B^j$.

Proposition 2.1. Soit $(\psi_j)_{j \in \mathbb{Z}}$. Si $\sum_{j=-\infty}^{+\infty} |\psi_j| < +\infty$ et X stationnaire, alors pour tout $i \in \mathbb{Z}$, $\sum_{j=-\infty}^{+\infty} \psi_j X_{i-j}$ converge absolument, dans l'espace L^2 et presque-sûrement. Le processus $(\sum_{j=-\infty}^{+\infty} \psi_j X_{i-j})_{i \in \mathbb{Z}}$ est lui-même stationnaire.

La proposition 2.1 permet la définition suivante.

Définition 2.6. Soit $(\psi_j)_{j \in \mathbb{Z}}$. Si $\sum_{j=-\infty}^{+\infty} |\psi_j| < +\infty$, on peut définir $\Psi(B) = \sum_{j=-\infty}^{+\infty} \psi_j B^j$, endomorphisme de l'espace vectoriel des processus aléatoires stationnaires du second ordre indicés par $\mathbb{Z}$.

Définition 2.7. On dit que X est un processus ARMA(p, q) si c'est un processus stationnaire à l'ordre deux et s'il existe $\Phi(z) = 1 - \sum_{j=1}^p \phi_j z^j$ et $\Theta(z) = 1 + \sum_{j=1}^q \theta_j z^j$ des polynômes et $\epsilon = (\epsilon_i)_{i \in \mathbb{Z}}$ bruit blanc tel que

$$\Phi(B)X = \Theta(B)\epsilon, \quad (2.1)$$

ou, de façon équivalente,

$$\forall i \in \mathbb{Z}, X_i - \sum_{j=1}^p \phi_j X_{i-j} = \epsilon_i + \sum_{j=1}^q \theta_j \epsilon_{i-j}. \quad (2.2)$$

Remarque 2.1. Un tel processus n'existe pas nécessairement pour tout couple de polynômes (Φ, Θ) (cf. exemple 2.4, cas $|\rho| = 1$).

Exemple 2.1. Un processus ARMA(0,0) est un bruit blanc.

Exemple 2.2. Si $p = 0$, on parle de processus de *moyenne mobile* et on note $MA(q)$ au lieu de ARMA(0, q). Un tel processus existe pour tout polynôme Θ et pour tout bruit blanc ϵ .

Exemple 2.3. Si $q = 0$, on parle de processus *autorégressif* et on note $AR(p)$ au lieu de ARMA($p, 0$). L'étymologie est parlante : il s'agit d'une régression linéaire du processus sur lui-même, en l'occurrence sur son passé. Le présent s'explique linéairement par le passé et le bruit blanc ϵ joue le rôle de la perturbation.

Exemple 2.4 (Processus AR(1)). Ici, (2.2) devient

$$\forall i \in \mathbb{Z}, X_i - \rho X_{i-1} = \epsilon_i, \quad (2.3)$$

où ρ est un nombre réel.

- Si $|\rho| < 1$, l'unique solution stationnaire (à égalité presque-sûre près) de (2.3) est $X = (X_i)_{i \in \mathbb{Z}}$ défini par, pour tout $i \in \mathbb{Z}$, $X_i \stackrel{p.s.}{=} \sum_{j=0}^{\infty} \rho^j \epsilon_{i-j}$. Dans ce cas, $\forall h \in \mathbb{Z}, \rho_X(h) = \rho^{|h|}$.
- Si $|\rho| > 1$, l'égalité devient $X_i \stackrel{p.s.}{=} -\sum_{j=1}^{\infty} \rho^{-j} \epsilon_{i+j}$.
- Si $|\rho| = 1$, (2.3) n'a pas de solution stationnaire.

Une différence importante apparaît entre les cas $|\rho| < 1$ et $|\rho| > 1$. Dans le premier cas, X_i n'est pas corrélé avec $(\epsilon_j)_{j>i}$ (le « futur » du bruit blanc), alors que dans le second cas, X_i s'écrit linéairement à partir de $(\epsilon_j)_{j>i}$. Ce fait illustre la notion de causalité.

Définition 2.8. Soit X un processus ARMA(p, q) défini par (2.1). X est dit causal s'il existe une suite $(\psi_j)_{j \geq 0}$ telle que $\sum_{j=0}^{\infty} |\psi_j| < \infty$ et

$$\forall i \in \mathbb{Z}, X_i = \sum_{j=0}^{\infty} \psi_j \epsilon_{i-j} . \quad (2.4)$$

D'après la proposition 2.1, cette série converge absolument, dans l'espace L^2 et presque-sûrement.

Theorème 2.1 (Caractérisation de la causalité). Soit X un processus ARMA(p, q) défini par (2.1) et tel que les polynômes Φ et Θ n'ont pas de racine commune. Alors X est causal si et seulement si $\forall z \in \mathbb{C}, |z| \leq 1 \Rightarrow \Phi(z) \neq 0$. Les coefficients ψ_j de (2.4) s'obtiennent par le développement en série entière de $\frac{\Theta}{\Phi}$:

$$\forall z \in \mathbb{C}, |z| \leq 1 \Rightarrow \Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)} .$$

Theorème 2.2 (Unicité de la solution de l'équation ARMA). Soient ϵ , Φ et Θ définis comme dans la définition 2.7. Si $\forall z \in \mathbb{C}, |z| = 1 \Rightarrow \Phi(z) \neq 0$, alors (2.1) a une unique solution stationnaire X , définie par

$$X = \Psi(B)\epsilon ,$$

où Ψ est le développement en série de Laurent de $\frac{\Theta}{\Phi}$ au voisinage du cercle complexe unité.

2.1.3 Estimation dans les processus stationnaires d'ordre deux

Estimation de la fonction d'autocovariance et de la fonction d'autocorrélation

On dispose de n observations consécutives $X_1, \dots, X_n$, d'un processus faiblement stationnaire.

Définition 2.9. Soient $X_1, \dots, X_n$, n observations consécutives d'un processus faiblement stationnaire X . Soit un entier $0 \leq h < n$. L'estimateur de la méthode des moments de $\gamma_X(h)$ est

$$\widehat{\gamma}_X(h) = \frac{1}{n} \sum_{i=1}^{n-h} (X_i - \overline{X}_n) (X_{i+h} - \overline{X}_n) ,$$

où $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. L'estimateur de la méthode des moments de $\rho_X(h)$ est

$$\widehat{\rho}_X(h) = \frac{\widehat{\gamma}_X(h)}{\widehat{\gamma}_X(0)} .$$

Remarque 2.2. Il existe d'autres estimateurs de la fonction d'autocovariance et de la fonction d'autocorrélation. Voir, par exemple, la section 2.3.3.

Estimation dans les processus ARMA

Notations on notera ici, pour X un processus faiblement stationnaire et $k \geq 0, p \geq 1, q \geq 0$ entiers,

$$\Gamma_{q,p} = (\gamma_X(i-j-q))_{1 \leq i, j \leq p}, \quad (2.5)$$

$$R_{q,p} = (\rho_X(i-j-q))_{1 \leq i, j \leq p}, \quad (2.6)$$

$$\gamma_{k,p} = (\gamma_X(k), \dots, \gamma_X(p))^T \text{ si } p \geq k, \quad (2.7)$$

$$\rho_{k,p} = (\rho_X(k), \dots, \rho_X(p))^T \text{ si } p \geq k. \quad (2.8)$$

Si X est un processus $ARMA(p, q)$ vérifiant l'équation (2.2), on notera $\Phi = (\phi_1, \dots, \phi_p)^T$.

Équations de Yule-Walker et généralisation

Proposition 2.2. Si X est un processus $ARMA(p, q)$, avec $p \geq 1$ et $q \geq 0$, alors

$$\Gamma_{q,p} \Phi = \gamma_{1+q,p+q}, \quad (2.9)$$

$$R_{q,p} \Phi = \rho_{1+q,p+q}, \quad (2.10)$$

où $\Gamma_{q,p}$, $R_{q,p}$, $\gamma_{1+q,p+q}$ et $\rho_{1+q,p+q}$ sont définis en (2.5)–(2.8).

Remarque 2.3. Si on fixe $q = 0$, les équations (2.9) et (2.10) sont les équations de Yule-Walker. $\Gamma_{0,p}$ et $R_{0,p}$ sont des matrices symétriques définies positives. En particulier, elles sont inversibles. Si $q \geq 1$, $\Gamma_{q,p}$ et $R_{q,p}$ ne sont pas symétriques. Le résultat suivant est tiré d'Azencott et Dacunha-Castelle (1986, chapitre 11, paragraphe 2).

Proposition 2.3. Si X un processus $ARMA(p, q)$ avec $p \geq 1$ et $q \geq 0$ et $\Gamma_{q,p}$, $R_{q,p}$, $\gamma_{1+q,p+q}$ et $\rho_{1+q,p+q}$ sont définis en (2.5)–(2.8), alors $\Gamma_{q,p}$ et $R_{q,p}$ sont des matrices inversibles. En particulier,

$$\Phi = \Gamma_{q,p}^{-1} \gamma_{1+q,p+q} = R_{q,p}^{-1} \rho_{1+q,p+q}. \quad (2.11)$$

Utilisation des équations de Yule-Walker pour l'estimation On notera, pour X un processus faiblement stationnaire, $p \geq 1, q \geq 0$ entiers et pour $\bar{\gamma}_X(h)$ (resp. $\bar{\rho}_X(h)$) estimateur de $\gamma_X(h)$ (resp. $\rho_X(h)$) si $h \in \mathbb{Z}$,

$$\bar{\Gamma}_{q,p} = (\bar{\gamma}_X(i-j-q))_{1 \leq i, j \leq p},$$

$$\bar{R}_{q,p} = (\bar{\rho}_X(i-j-q))_{1 \leq i, j \leq p},$$

$$\bar{\gamma}_{q,p} = (\bar{\gamma}_X(q), \dots, \bar{\gamma}_X(p))^T \text{ si } p \geq q,$$

$$\bar{\rho}_{q,p} = (\bar{\rho}_X(q), \dots, \bar{\rho}_X(p))^T \text{ si } p \geq q.$$

Si X est un processus $ARMA(p, q)$, on peut alors estimer Φ à l'aide de l'équation (2.11), en remplaçant les $\gamma_X(h)$ par leurs estimateurs. On considère alors un des estimateurs suivants de Φ :

$$\bar{\Phi} = \bar{\Gamma}_{q,p}^{-1} \bar{\gamma}_{1+q,p+q} \text{ ou bien} \quad (2.12)$$

$$\bar{\Phi} = \bar{R}_{q,p}^{-1} \bar{\rho}_{1+q,p+q} . \quad (2.13)$$

Si les estimateurs des $\gamma_X(h)$ (resp. $\rho_X(h)$) sont les estimateurs de la méthode des moments (voir définition 2.9) et si $q = 0$, alors $\bar{\Gamma}_{q,p}$ et $\bar{R}_{q,p}$ sont des matrices symétriques définies positives. Le calcul de $\bar{\Phi}$ se fait alors à l'aide de l'algorithme de Durbin-Levinson (Brockwell et Davis, 1991, section 8.2). En dehors de ces conditions, on n'a pas de garantie sur l'inversibilité de $\bar{\Gamma}_{q,p}$ ou $\bar{R}_{q,p}$. On peut tenter de démontrer qu'asymptotiquement ces matrices sont presque-sûrement inversibles.

Shevlyakov *et al.* (2014) proposent d'estimer la fonction d'autocovariance ou d'autocorrélation de façon robuste puis d'utiliser (2.12) ou (2.13) pour estimer Φ dans le cas $q = 0$. Dans le chapitre 4, nous proposons la même démarche avec $q = 1$.

2.2 Processus empirique et Delta-méthode fonctionnelle

On introduit ici des notions et énonce des résultats utilisés par la suite dans les preuves 3.7.1 et 4.6.1 de la normalité asymptotique d'estimateurs.

2.2.1 Processus empirique

La référence de cette section est en premier lieu Van der Vaart (2000, chapitre 19).

Définition 2.10. Soient $X_1, \dots, X_n$ des v.a.r. La fonction de répartition empirique est la fonction aléatoire $F_n : \mathbb{R} \rightarrow [0, 1]$ définie par

$$\forall t \in \mathbb{R}, F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \leq t} .$$

Si les X_i ont la même loi, de fonction de répartition F , le processus empirique est la fonction réelle aléatoire $\sqrt{n}(F_n - F)$.

D'après le théorème de Glivenko-Cantelli, si $(X_n)_{n \geq 1}$ est une suite de v.a.r. i.i.d. de même loi de fonction de répartition F , la fonction de répartition empirique converge presque-sûrement vers F pour la norme uniforme. C'est l'équivalent de la loi forte des grands nombres pour la fonction de répartition empirique. De même, le théorème 2.3 ci-dessous fournit un équivalent du théorème central-limite.

Définition 2.11. On note $D[-\infty, +\infty]$ l'espace vectoriel des fonctions bornées de $[-\infty, +\infty]$ dans $\mathbb{R}$ et càdlàg (continues à droite et ayant une limite à gauche en tout point, ayant une limite en $-\infty$ et en $+\infty$), muni de la norme uniforme.

En particulier, les fonctions de répartition sur $\mathbb{R}$ sont des éléments de $D[-\infty, +\infty]$.

Définition 2.12. Soit F une fonction de répartition sur $\mathbb{R}$. Un F -pont brownien est une variable aléatoire G_F sur $D[-\infty, +\infty]$ telle que :

- $\forall k \geq 1, \forall (t_1, \dots, t_k) \in \mathbb{R}^k, (G_F(t_1), \dots, G_F(t_k))$ est un vecteur gaussien centré,
- $\forall (s, t) \in \mathbb{R}^2, s \leq t \Rightarrow \text{Cov}(G_F(s), G_F(t)) = F(s)(1 - F(t))$.

Théorème 2.3 (Donsker). Soit $(X_n)_{n \geq 1}$ une suite de v.a.r. i.i.d. de fonction de répartition F . Alors $(\sqrt{n}(F_n - F))$ converge en loi dans $D[-\infty, +\infty]$ vers un F -pont brownien.

Toutefois, ce résultat est insuffisant dans le cas d'une suite de variables aléatoires dépendantes. Pour traiter ce cas, nous avons utilisé le théorème suivant prouvé par Csörgő et Mišćuk (1996).

Définition 2.13. Pour un entier naturel k , le k^e polynôme de Hermite est défini par

$$H_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2}.$$

Théorème 2.4. Soit $(Z_n)_{n \in \mathbb{Z}}$ une suite stationnaire de v.a.r. normales centrées réduites. Soit G une fonction borélienne et $(X_n)_{n \geq 1} = (G(Z_n))_{n \geq 1}$. Soit F la fonction de répartition de X_1 . Pour x réel et k entier naturel, on note $J_k(x) = \mathbb{E}(H_k(Z_1)(\mathbf{1}_{X_1 \leq x} - F(Z_1)))$. On définit

$$\begin{aligned} \forall x \in \mathbb{R}, m_x &= \min \{k; J_k(x) \neq 0\}, \\ m &= \min \{m_x; x \in \mathbb{R}, 0 < F(x) < 1\}. \end{aligned}$$

Supposons que F est continue et que

$$\sum_{h=1}^{\infty} |\gamma_Z(h)|^m < +\infty.$$

Soit W une v.a. sur $D[-\infty, +\infty]$, un processus gaussien centré — c'est-à-dire tel que pour tout $k \geq 1, \forall (t_l)_{1 \leq l \leq k} \in \mathbb{R}^k, (W(t_l))_{1 \leq l \leq k}$ est un vecteur gaussien centré — de fonction d'autocovariance :

$$(x, y) \mapsto \text{Cov}(W(x), W(y)) = \sum_{q=m}^{\infty} \frac{J_q(x)J_q(y)}{q!} \left\{ 1 + 2 \sum_{h=1}^{\infty} \gamma_Z(h)^q \right\}.$$

Alors $(\sqrt{n}(F_n - F))$ converge en loi dans $D[-\infty, +\infty]$ vers W .

2.2.2 Delta-méthode fonctionnelle

L'idée est de généraliser la Delta-méthode à des espaces de dimension infinie. La référence de cette section est Van der Vaart (2000, chapitres 3, 20, 21).

Delta-méthode en dimension finie

Théorème 2.5. Soit $\Theta \subset \mathbb{R}^k$, $\theta \in \Theta$ et une application $\phi : \Theta \rightarrow \mathbb{R}^m$ différentiable en θ . Soit T v.a. sur $\mathbb{R}^k$, (T_n) une suite de v.a. sur Θ et (r_n) une suite réelle de limite $+\infty$. Supposons que

$$r_n(T_n - \theta) \xrightarrow[n \rightarrow \infty]{d} T.$$

Alors

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow[n \rightarrow \infty]{d} \phi'_\theta(T),$$

où ϕ'_θ est la différentielle de ϕ en θ . De plus,

$$r_n(\phi(T_n) - \phi(\theta)) = \phi'_\theta(r_n(T_n - \theta)) + o_P(1).$$

Différentiabilité au sens de Hadamard

La généralisation du théorème 2.5 à des espaces de dimension infinie nécessite de se pencher plus avant sur la notion de différentiabilité et de différentielle dans un espace de dimension infinie. Il existe différentes notions de différentiabilité, qui sont équivalentes en dimension finie mais pas en dimension infinie. La notion dont nous avons besoin ici est la différentiabilité au sens de Hadamard.

Définition 2.14. Soit deux espaces vectoriels normés $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$. Soit $\theta \in \Theta \subset E$ et E_0 un sous-espace vectoriel de E . Soit $\phi : E \rightarrow F$. ϕ est dit différentiable au sens de Hadamard en θ tangentiellement à E_0 s'il existe $\phi'_\theta : E_0 \rightarrow F$ linéaire et continue telle que pour tout $h \in E_0$, pour tout $\varepsilon > 0$, pour toute application $g : (0, \varepsilon) \rightarrow E_0$ de limite h en 0 et telle que $\forall t \in (0, \varepsilon), \theta + tg(t) \in \Theta$,

$$\lim_{t \rightarrow 0} \left\| \frac{\phi(\theta + tg(t)) - \phi(\theta)}{t} - \phi'_\theta(h) \right\|_F = 0.$$

ϕ'_θ est appelée différentielle de ϕ en θ tangentiellement à E_0 .

Delta-méthode fonctionnelle

Théorème 2.6. Soit deux espaces vectoriels normés $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$. Soit $\theta \in \Theta \subset E$ et E_0 un sous-espace vectoriel de E . Soit $\phi : E \rightarrow F$ différentiable au sens de Hadamard en θ tangentiellement à E_0 . Soient T v.a. sur E_0 , (T_n) une suite de v.a. sur Θ et (r_n) une suite réelle de limite $+\infty$. Supposons que

$$r_n(T_n - \theta) \xrightarrow[n \rightarrow \infty]{d} T.$$

Alors

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow[n \rightarrow \infty]{d} \phi'_\theta(T),$$

où ϕ'_θ est la différentielle de ϕ en θ . De plus, si $E = E_0$, alors

$$r_n(\phi(T_n) - \phi(\theta)) = \phi'_\theta(r_n(T_n - \theta)) + o_P(1).$$

Typiquement, E peut être l'espace $D[-\infty, +\infty]$ (cf. définition 2.11) et $r_n(T_n - \theta)$ le processus empirique (cf. définition 2.10). D'un résultat de convergence du processus empirique (comme les théorèmes 2.3 et 2.4), on peut déduire la normalité asymptotique d'un estimateur *plug-in* ($\phi(F_n)$ est l'estimateur de $\phi(F)$). C'est notamment grâce à ce résultat qu'on obtient, sous certaines conditions, la normalité asymptotique des quantiles empiriques comme estimateurs des quantiles de la loi.

2.2.3 Fonctions d'un processus gaussien

Le résultat suivant (Arcones, 1994, théorème 4) est utilisé en association avec les théorèmes 2.4 et 2.6 pour démontrer la normalité asymptotique d'estimateurs.

Définition 2.15. Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$, de carré intégrable par rapport à une mesure gaussienne, et $X = (X^{(p)})_{1 \leq p \leq d}$ vecteur gaussien sur $\mathbb{R}^d$. Le rang d'Hermite de f par rapport à X est

$$\inf \{m \geq 1; \exists Q \text{ polynôme à } d \text{ variables, de degré } m \text{ tel que } \mathbb{E}[(f(X) - \mathbb{E}[X])Q(X)] \neq 0\}.$$

Théorème 2.7. Soit $(X_i)_{i \in \mathbb{Z}} = \left((X_i^{(p)})_{1 \leq p \leq d} \right)_{i \in \mathbb{N}^*}$ un processus gaussien centré stationnaire d -varié. Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ de carré intégrable par rapport à une mesure gaussienne et de rang d'Hermite m par rapport à X_0 . Supposons que

$$\forall (p, q) \in \{1, \dots, d\}^2, \sum_{k \in \mathbb{Z}} \left| \text{Cov}(X_0^{(p)}, X_k^{(q)}) \right|^m < \infty.$$

Alors

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)]) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \tilde{\sigma}^2),$$

où

$$\tilde{\sigma}^2 = \text{Var}(f(X_0)) + 2 \sum_{i=1}^{\infty} \text{Cov}(f(X_0), f(X_i)).$$

2.3 Estimation robuste

Un estimateur robuste est un estimateur dont les propriétés ne sont pas altérées par la présence de valeurs aberrantes dans les observations (c'est-à-dire de loi différente de celle des autres observations) ou par une relaxation des hypothèses du modèle.

Les *M-estimateurs* d'un paramètre θ du modèle, définis par

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \sum_{i=1}^n \rho(X_i, \theta),$$

si $X_1, \dots, X_n$ sont les observations et ρ une certaine fonction, jouent un rôle important dans l'estimation robuste. Les différentes propriétés de robustesse de l'estimateur peuvent être obtenues par un choix adéquat de la fonction ρ . Par exemple, si $\Theta = \mathbb{R}$, $X_1, \dots, X_n \in \mathbb{R}$ et $\rho(X_i, \theta) = |X_i - \theta|$, alors $\hat{\theta} = \text{med}_{i=1, \dots, n} X_i$. La médiane est un exemple simple d'estimateur robuste. Cela étant dit, il ne sera pas détaillé ici l'étude générale des M-estimateurs qui ne nous est pas utile en tant que telle, même si ce n'est pas sans raison que les estimateurs que nous considérons utilisent les quantiles (dont la médiane).

2.3.1 Point d'effondrement

La référence de cette partie est Maronna *et al.* (2006, paragraphe 3.2.5.).

On observe $\mathbf{v} = (v_1, \dots, v_n) \in \mathbb{R}^n$, réalisation d'une variable aléatoire dont la loi dépend d'un paramètre $\theta \in \Theta$. On suppose que $\Theta \subset \mathbb{R}$. Soit $\widehat{\theta}_n$ un estimateur de θ . On souhaite mesurer la proportion d'observations X_i que l'on peut modifier arbitrairement tout en conservant une borne sur les déviations induites sur l'estimation. La proportion maximale d'observations que l'on peut ainsi modifier est le *point d'effondrement* (*breakdown point* en anglais) de $\widehat{\theta}_n$. Pour $\mathbf{v} = (v_1, \dots, v_n), \mathbf{u} = (u_1, \dots, u_n) \in \mathbb{R}^n$ et $I \subset \{1, \dots, n\}$, on notera

$$\mathbf{v} \setminus_I \mathbf{u} = (u_i \mathbf{1}_{i \in I} + v_i \mathbf{1}_{i \notin I})_{1 \leq i \leq n} . \quad (2.14)$$

Définition 2.16 (Point d'effondrement). *Soient $\widehat{\theta}_n : \mathbb{R}^n \rightarrow \Theta$ et $\mathbf{v} \in \mathbb{R}^n$. Le point d'effondrement de $\widehat{\theta}_n$ en $\mathbf{v}$ est*

$$\varepsilon_n^*(\widehat{\theta}_n, \mathbf{v}) = \frac{1}{n} \max \left\{ m \geq 0; \exists K \subset \mathring{\Theta}, K \text{ compact}, \forall \mathbf{u} \in \mathbb{R}^n, \forall I \subset \{1, \dots, n\}, |I| = m \Rightarrow \widehat{\theta}_n(\mathbf{v} \setminus_I \mathbf{u}) \in K \right\} ,$$

où $\mathbf{v} \setminus_I \mathbf{u}$ est défini en (2.14) et $\mathring{\Theta}$ est l'intérieur de Θ . Si $\varepsilon_n^*(\widehat{\theta}_n, \cdot)$ est constante sur $\mathbb{R}^n$, l'unique valeur prise par cette fonction est appelée point d'effondrement de $\widehat{\theta}_n$ et est notée $\varepsilon_n^*(\widehat{\theta}_n)$. Dans ce cas, si la suite $(\varepsilon_n^*(\widehat{\theta}_n))_n$ a une limite, la limite est appelée point d'effondrement de $\widehat{\theta}$ et est notée $\varepsilon^*(\widehat{\theta})$.

Le point d'effondrement d'un estimateur consistant est inférieur ou égal à 1/2.

2.3.2 Estimation robuste d'un paramètre d'échelle

Soit $\mathbf{X} = (X_i)_{i \in \mathbb{N}}$ où les X_i sont des v.a.r. i.i.d. telles que

$$\forall i \geq 1, X_i = \mu + \sigma u_i ,$$

où $\mu \in \mathbb{R}$ et $\sigma > 0$ sont inconnus et les v.a.r. u_i sont i.i.d. de loi connue de fonction de répartition F . On appelle ce modèle le modèle de *localisation-dispersion* (Maronna *et al.* 2006, équation (2.66)). Le but est d'estimer σ .

Plusieurs estimateurs de ce type ont été proposés :

- La *distance interquartile* $IQR_n(\mathbf{X}) = c_{IQR,F} \left(X_{(\lceil \frac{3}{4}n \rceil)} - X_{(\lceil \frac{1}{4}n \rceil)} \right)$. La constante positive $c_{IQR,F}$ est choisie en fonction de F pour obtenir la consistance de l'estimateur. Le point d'effondrement de l'estimateur IQR est 1/4.
- La *déviations absolue médiane à la médiane* $MAD_n(\mathbf{X}) = c_{MAD,F} \text{med}_{i=1, \dots, n} \left| X_i - \text{med}_{j=1, \dots, n} X_j \right|$. La constante $c_{MAD,F} > 0$ est choisie en fonction de F pour obtenir la consistance de l'estimateur. Le point d'effondrement de l'estimateur MAD est 1/2.

Rousseeuw et Croux (1993) introduisent deux nouveaux estimateurs,

$$S_n(\mathbf{X}) = c_{S,F} \operatorname{med}_{i=1,\dots,n} \operatorname{med}_{j=1,\dots,n} |X_i - X_j|, \quad (2.15)$$

$$Q_n(\mathbf{y}) = c_{Q,F} \{ |X_i - X_j|; 1 \leq i < j \leq n \}_{(\frac{1}{2} \lfloor \frac{n}{2} \rfloor (\lfloor \frac{n}{2} \rfloor + 1))}, \quad (2.16)$$

où $E_{(k)}$ est la k^e statistique d'ordre de $x_1, \dots, x_r$ et $E = \{x_1, \dots, x_r\}$. Les constantes strictement positives $c_{S,F}$ et $c_{Q,F}$ sont choisies en fonction de F pour obtenir la consistance de l'estimateur. À titre d'exemple, si $F = \Phi$ est la fonction de répartition de la loi normale centrée réduite,

$$c_{S,\Phi} = \operatorname{med} \left| \sqrt{\Phi^{-1}\left(\frac{3}{4}\right) + u_1} \right|^{-1} \approx 1.1926, \quad (2.17)$$

$$c_{Q,\Phi} = \frac{1}{\sqrt{2}\Phi^{-1}\left(\frac{5}{8}\right)} \approx 2.2191. \quad (2.18)$$

S_n et Q_n ont le même point d'effondrement $\frac{1}{n} \lfloor \frac{n}{2} \rfloor$, soit asymptotiquement $1/2$. La complexité en temps du calcul de ces deux estimateurs est $O(n \log n)$ (Croux et Rousseeuw, 1992). Dans la section 2.3.3, ces estimateurs servent de base pour construire des estimateurs des fonctions d'autocovariance et d'autocorrélation d'un processus stationnaire du second ordre. Rousseeuw et Croux (1993) font remarquer que $\frac{1}{2} \lfloor \frac{n}{2} \rfloor (\lfloor \frac{n}{2} \rfloor + 1) \approx \lfloor \frac{1}{4} C_n^2 \rfloor$. Lévy-Leduc *et al.* (2011) utilisent alors cette approximation pour définir Q_n :

$$Q_n(\mathbf{y}) = c_{Q,F} \{ |X_i - X_j|; 1 \leq i < j \leq n \}_{(\lfloor \frac{1}{4} C_n^2 \rfloor)}. \quad (2.19)$$

Nous reprenons cette définition.

2.3.3 Estimation robuste des fonctions d'autocovariance et d'autocorrélation d'un processus stationnaire du second ordre

Dans cette section sont présentés les estimateurs de Ma et Genton (2000, un pour la fonction d'autocovariance, un pour la fonction d'autocorrélation). On se donne $\mathbf{X} = (X_i)_{i \in \mathbb{Z}}$, un processus stationnaire du second ordre (non-nécessairement centré), et on cherche à estimer $\gamma_X(h)$ et $\rho_X(h)$ à partir de n observations successives de $\mathbf{X}$.

Estimation robuste de la fonction d'autocovariance

Cet estimateur de la fonction d'autocovariance utilise l'identité suivante, valable pour deux v.a.r. de carré intégrable A et B :

$$\operatorname{Cov}(A, B) = \frac{1}{4} \{ \operatorname{Var}(A + B) - \operatorname{Var}(A - B) \}.$$

Cette égalité est notamment vraie pour $A = X_i$ et $B = X_{i+h}$. Ainsi, pour tout $i \in \mathbb{Z}$,

$$\gamma_X(h) = \frac{1}{4} \{ \operatorname{Var}(X_i + X_{i+h}) - \operatorname{Var}(X_i - X_{i+h}) \}.$$

Cela conduit à estimer préalablement les variances $\text{Var}(X_i + X_{i+h})$ et $\text{Var}(X_i - X_{i+h})$. L'estimateur Q_n (cf. équation (2.19)) permet de faire cela de façon robuste.

On se donne $\mathbf{u}^h = (X_1, \dots, X_{n-h})$ et $\mathbf{v}^h = (X_{h+1}, \dots, X_n)$. L'estimation robuste de $\text{Var}(X_i + X_{i+h})$ et de $\text{Var}(X_i - X_{i+h})$ se fait respectivement par $Q_{n-h}^2(\mathbf{u}^h + \mathbf{v}^h)$ et $Q_{n-h}^2(\mathbf{u}^h - \mathbf{v}^h)$. Ma et Genton (2000) proposent donc comme estimateur robuste de $\gamma_X(h)$:

$$\widehat{\gamma}_n(h, \mathbf{X}) = \frac{1}{4} \{Q_{n-h}^2(\mathbf{u}^h + \mathbf{v}^h) - Q_{n-h}^2(\mathbf{u}^h - \mathbf{v}^h)\} . \quad (2.20)$$

Ma et Genton (2000) annoncent un point d'effondrement de $1/4$ quand $\mathbf{X}$ est un processus ARMA, mais dans un sens différent que celui de la définition 2.16. En effet, ils considèrent que les perturbations affectent le processus d'innovation intervenant dans l'équation ARMA vérifiée par $\mathbf{X}$ et non les observations de $\mathbf{X}$ directement.

Estimation robuste de la fonction d'autocorrélation

Pour estimer $\rho_X(h)$, on pourrait utiliser

$$\frac{\widehat{\gamma}_n(h, \mathbf{X})}{Q_n(\mathbf{X})} .$$

Toutefois, l'identité suivante est valable pour deux v.a.r. de carré intégrable A et B :

$$\frac{1}{2} \{\text{Var}(A) + \text{Var}(B)\} = \frac{1}{4} \{\text{Var}(A+B) + \text{Var}(A-B)\} ,$$

et donc

$$\text{Var}(A) = \frac{1}{4} \{\text{Var}(A+B) + \text{Var}(A-B)\}$$

si A et B ont la même variance. On peut donc utiliser

$$\frac{1}{4} \{Q_{n-h}^2(\mathbf{u}^h + \mathbf{v}^h) + Q_{n-h}^2(\mathbf{u}^h - \mathbf{v}^h)\}$$

pour estimer de façon robuste $\gamma_X(0)$. Ainsi,

$$\widehat{\rho}_n(h, \mathbf{X}) = \frac{Q_{n-h}^2(\mathbf{u}^h + \mathbf{v}^h) - Q_{n-h}^2(\mathbf{u}^h - \mathbf{v}^h)}{Q_{n-h}^2(\mathbf{u}^h + \mathbf{v}^h) + Q_{n-h}^2(\mathbf{u}^h - \mathbf{v}^h)} \quad (2.21)$$

est un estimateur robuste de $\rho_X(h)$. Il a le même point d'effondrement que $\widehat{\gamma}_n(h)$ et présente l'avantage d'être compris entre -1 et 1 , tout comme $\rho_X(h)$. Notons aussi que, dans le cas gaussien, $\widehat{\rho}_n(h, \mathbf{X})$ ne dépend pas de la constante $c_{Q,\Phi}$.

Propriétés asymptotiques

Lévy-Leduc *et al.* (2011) établissent le résultat suivant.

Theorème 2.8. *Soit $\mathbf{X} = (X_i)_{i \in \mathbb{Z}}$ un processus gaussien univarié centré stationnaire. Soit $h \in \mathbb{N}$. Soient $Q_n(\mathbf{X})$ et $\widehat{\gamma}_n(h, \mathbf{X})$ définis resp. en (2.19) et (2.20). Soit*

$$\text{IF}(x, Q, \Phi) = c_{Q,\Phi} \left(\frac{1/4 - \Phi(x + 1/c_{Q,\Phi}) + \Phi(x - 1/c_{Q,\Phi})}{\int_{\mathbb{R}} \phi(y) \phi(y + 1/c_{Q,\Phi}) dy} \right) ,$$

où Φ est la fonction de répartition de la loi normale centrée réduite, $\phi = \Phi'$ et $c_{Q,\Phi}$ est définie en (2.18). Soit

$$\psi : (x, y) \mapsto \left\{ (\gamma_X(0) + \gamma_X(h)) \operatorname{IF} \left(\frac{x+y}{\sqrt{2(\gamma_X(0) + \gamma_X(h))}}, Q, \Phi \right) \right. \\ \left. - (\gamma_X(0) - \gamma_X(h)) \operatorname{IF} \left(\frac{x-y}{\sqrt{2(\gamma_X(0) - \gamma_X(h))}}, Q, \Phi \right) \right\} .$$

Supposons que la fonction d'autocovariance de $\mathbf{X}$ vérifie

$$\sum_{h=1}^{\infty} |\gamma_X(h)| < \infty .$$

Alors

$$\sqrt{n} \left(Q_n(\mathbf{X}) - \sqrt{\gamma_X(0)} \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \tilde{\sigma}_Q^2) , \quad (2.22)$$

$$\sqrt{n} (\hat{\gamma}_n(h, \mathbf{X}) - \gamma(h)) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \tilde{\sigma}_{\hat{\gamma}}^2(h)) , \quad (2.23)$$

où

$$\tilde{\sigma}_Q^2 = \gamma_X(0) \mathbb{E} \left[\operatorname{IF}^2 \left(\frac{X_1}{\sqrt{\gamma_X(0)}}, Q, \Phi \right) \right] + 2\gamma_X(0) \sum_{k=1}^{\infty} \mathbb{E} \left[\operatorname{IF} \left(\frac{X_1}{\sqrt{\gamma_X(0)}}, Q, \Phi \right) \operatorname{IF} \left(\frac{X_{k+1}}{\sqrt{\gamma_X(0)}}, Q, \Phi \right) \right] , \quad (2.24)$$

$$\tilde{\sigma}_{\hat{\gamma}}^2(h) = \mathbb{E} [\psi^2(X_1, X_{1+h})] + 2 \sum_{k=1}^{\infty} \mathbb{E} [\psi(X_1, X_{1+h}) \psi(X_{k+1}, X_{k+1+h})] . \quad (2.25)$$

De plus,

$$\sqrt{n-h} (\hat{\gamma}_n(h, \mathbf{X}) - \gamma(h)) = \frac{1}{\sqrt{n-h}} \sum_{i=1}^{n-h} \psi(X_i, X_{i+h}) + o_P(1) .$$

On en déduit le corollaire suivant par le lemme de Slutsky :

Corollaire 2.1. *Sous les mêmes hypothèses que le théorème 2.8 et si $\hat{\rho}_n(h, \mathbf{X})$ est l'estimateur défini en (2.21), alors*

$$\sqrt{n} (\hat{\rho}_n(h, \mathbf{X}) - \rho_X(h)) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left(0, \frac{\tilde{\sigma}_{\hat{\gamma}}^2(h)}{\gamma_X^2(0)} \right) , \quad (2.26)$$

où $\tilde{\sigma}_{\hat{\gamma}}^2(h)$ est défini en (2.25).

2.4 Inégalités maximales

Dans cette section, on énonce l'inégalité de Kolmogorov, sa généralisation qu'est l'inégalité de Hájek-Rényi, une généralisation de cette dernière, ainsi que leur utilisation en segmentation. Dans toute cette section, $(\eta_i)_{i \geq 1}$ est une suite de variables aléatoires réelles, centrées, de carré intégrable. Pour tout $n \geq 1$, on définit $S_n = \sum_{i=1}^n \eta_i$.

2.4.1 Inégalité de Kolmogorov

Théorème 2.9 (Inégalité de Kolmogorov). *Si les variables aléatoires η_i sont mutuellement indépendantes, alors*

$$\forall n \geq 1, \forall x > 0, P\left(\max_{i=1,\dots,n} |S_i| \geq x\right) \leq \frac{1}{x^2} \sum_{i=1}^n \text{Var } \eta_i.$$

La démonstration de cette inégalité figure dans bien des ouvrages de probabilités. L'article original (Kolmogorov, 1928) étant en allemand, les non-germanophones peuvent consulter par exemple Durrett (2010, Théorème 2.5.2.). L'inégalité de Bienaymé-Tchebychev nous aurait seulement assuré $P(|S_n| \geq x) \leq \frac{1}{x^2} \sum_{i=1}^n \text{Var } \eta_i$. L'inégalité maximale du théorème 2.9 est un résultat bien plus puissant. L'inégalité de Kolmogorov est un ingrédient essentiel de la démonstration de la loi forte des grands nombres.

2.4.2 Inégalité de Hájek-Rényi

Hájek et Rényi (1955) généralisent l'inégalité de Kolmogorov :

Théorème 2.10 (Inégalité de Hájek-Rényi). *Si les variables aléatoires η_i sont mutuellement indépendantes, alors, pour toute suite décroissante de réels positifs $(c_n)_{n \geq 1}$,*

$$\forall m > n \geq 1, \forall x > 0, P\left(\max_{i=n,\dots,m} (c_i |S_i|) \geq x\right) \leq \frac{1}{x^2} \left(c_n^2 \sum_{i=1}^n \text{Var } \eta_i + \sum_{i=n+1}^m c_i^2 \text{Var } \eta_i \right).$$

Le théorème 2.9 est un cas particulier du théorème 2.10, avec $n = 1$ et $\forall n \geq 1, c_n = 1$. Si $\forall n \geq 1, c_n = \frac{1}{n}$, on obtient un contrôle sur les déviations maximales des moyennes :

$$\forall m > n \geq 1, \forall x > 0, P\left(\max_{i=n,\dots,m} \left(\frac{1}{i} |S_i|\right) \geq x\right) \leq \frac{1}{x^2} \left(\frac{1}{n^2} \sum_{i=1}^n \text{Var } \eta_i + \sum_{i=n+1}^m \frac{1}{i^2} \text{Var } \eta_i \right).$$

Corollaire 2.2. *Sous les mêmes hypothèses que le théorème 2.10,*

$$\forall n \geq 1, \forall x > 0, P\left(\max_{i \geq n} (c_i |S_i|) \geq x\right) \leq \frac{1}{x^2} \left(c_n^2 \sum_{i=1}^n \text{Var } \eta_i + \sum_{i=n+1}^{\infty} c_i^2 \text{Var } \eta_i \right).$$

En particulier, si $\forall n \geq 1, c_n = \frac{1}{n}$, on obtient là aussi un contrôle sur les déviations maximales des moyennes :

$$\forall n \geq 1, \forall x > 0, P\left(\max_{i \geq n} \left(\frac{1}{i} |S_i|\right) \geq x\right) \leq \frac{1}{x^2} \left(\frac{1}{n^2} \sum_{i=1}^n \text{Var } \eta_i + \sum_{i=n+1}^{\infty} \frac{1}{i^2} \text{Var } \eta_i \right).$$

2.4.3 Généralisation de l'inégalité de Hájek-Rényi

Cette inégalité de Hájek-Rényi a fait l'objet de nombreuses généralisations, notamment au cas où les variables aléatoires η_i sont dépendantes. Nous n'en énoncerons qu'une.

Proposition 2.4 (Inégalité de Hájek-Rényi pour les processus linéaires). *Soit $\epsilon = (\epsilon_i)_{i \in \mathbb{Z}}$ une suite de v.a.r. i.i.d. centrées de variance σ^2 .*

Supposons qu'il existe une suite réelle $(a_j)_{j \geq 0}$ telle que :

- $\forall i \geq 1, \eta_i = \sum_{j=0}^{\infty} a_j \epsilon_{i-j}$ presque-sûrement,
- $\sum_{j=0}^{\infty} j |a_j| < +\infty$.

Alors, pour toute suite décroissante de réels positifs $(c_n)_{n \geq 1}$,

$$\forall m > n \geq 1, \forall x > 0, P\left(\max_{i=n, \dots, m} (c_i |S_i|) \geq x\right) \leq \frac{A\sigma^2}{x^2} \left(nc_n^2 + \sum_{i=n+1}^m c_i^2\right),$$

où A est une constante réelle qui ne dépend que de la suite $(a_j)_{j \geq 0}$.

Résultat démontré par Bai (1994). De plus, Lavielle et Moulines (2000) démontrent un théorème qui permet notamment la généralisation de l'inégalité de Hájek-Rényi aux processus à longue mémoire (voir par exemple Beran *et al.* 2013, pour une introduction à ces processus).

2.4.4 Utilisation en segmentation

L'inégalité de Kolmogorov, et surtout les inégalités de Hájek-Rényi, interviennent régulièrement en segmentation ; qu'il s'agisse d'une rupture simple (Bai, 1994) ou de ruptures multiples (Lavielle, 1999 ; Lavielle et Moulines, 2000), de détection *on-line* (Bardet et Kengne, 2014) ou *off-line* (Bai, 1994).

Ici, $y_1, \dots, y_n$ sont des observations telles que $y_i = \mu(i) + \eta_i$ pour tout $1 \leq i \leq n$ où $\mu : \{1, \dots, n\} \rightarrow \mathbb{R}$ a une ou des ruptures $\lfloor n\tau \rfloor$, et η est une suite de v.a.r. centrées (avec éventuellement des hypothèses supplémentaires).

Dans ce contexte, l'inégalité sert principalement à démontrer le taux de convergence de l'estimateur des moindres carrés $\hat{\tau}_n$ de la rupture $\tau \in (0, 1)$, à savoir $\hat{\tau}_n - \tau = O_P(n^{-1})$ sous certaines conditions sur η . Cette démarche peut se généraliser à d'autres estimateurs (Lavielle, 1999).

Dans les chapitres 3 et 4, l'utilisation d'une inégalité de Hájek-Rényi généralisée est cachée derrière les références aux résultats de Lavielle et Moulines (2000). Il est fait référence à ce résultat en particulier pour ne pas multiplier les références dans l'article. Toutefois, l'aspect longue-mémoire n'est pas utilisé et on aurait pu se contenter de faire référence à un résultat plus faible. L'inégalité de Kolmogorov est, elle, utilisée dans la preuve de la proposition 3.4, qui concerne la sélection du nombre de ruptures par critère pénalisé.

2.5 Sélection de modèle

Un catalogue non-exhaustif de principes de sélection de modèle est établi dans les sections 2.5.2 et 2.5.3. Leur application à la segmentation est discutée dans la section 2.5.4.

2.5.1 Introduction

La question générale de la sélection de modèle est : parmi un ensemble donné de modèles, lequel choisir au vu des observations ? Notamment, on s'attache à estimer la dimension du modèle.

On dispose d'observations $\mathbf{y} = (y_1, \dots, y_n)$, un n -uplet de variables aléatoires sur un ensemble E , et de $M + 1$ modèles $\mathcal{M}_0, \dots, \mathcal{M}_M$, sous-modèles de $\mathcal{M}$. La loi de $\mathbf{y}$ dépend d'un paramètre $s^* \in S$. La loi de $\mathbf{y}$ appartient au modèle $\mathcal{M}_m$ si et seulement si $s^* \in S_m$, où $S_0, \dots, S_M$ sont des sous-ensembles de S . On note d_m la dimension de S_m ($0 \leq m \leq M$).

Par exemple, on peut considérer une famille de modèles $ARMA(p, q)$ pour $\mathbf{y}$ sans connaître les ordres p et q . Il faut alors construire un critère de sélection de modèle pour choisir ces ordres qui définissent un modèle dans lequel on va travailler. De même, dans un modèle de segmentation, on sélectionne le nombre de ruptures du modèle.

La sélection de modèle peut se faire à l'aide de tests d'hypothèses sur l'appartenance de la loi des observations à un modèle (Rao et Wu, 2001). Le problème est que, au-delà de deux modèles considérés, cela implique de réaliser plusieurs tests successifs, le résultat dépendant de l'ordre dans lequel est réalisée cette succession. Alternativement, on cherche le meilleur modèle parmi une collection, selon un critère à définir. Nous allons aborder cette dernière approche dans deux cadres, fréquentiste et bayésien.

2.5.2 Sélection de modèle par critère pénalisé, cadre fréquentiste

On reprend ici les notations définies dans la section 2.5.1. $\gamma_n : S \rightarrow \mathbb{R}$ est une fonction de contraste empirique, par exemple $-2/n \times$ la log-vraisemblance. Dans chacun des modèles $\mathcal{M}_m$, le paramètre s^* est estimé par minimisation de γ_n sur S_m . Un tel estimateur est noté $\widehat{s}_m$.

Critère pénalisé

La sélection de modèle par critère pénalisé se fait en sélectionnant le modèle $\mathcal{M}_{\widehat{m}_n}$ tel que

$$\widehat{m}_n = \arg \min_{0 \leq m \leq M} \{ \gamma_n(\widehat{s}_m) + \text{pen}_n(m) \} , \quad (2.27)$$

où $\text{pen}_n : \{0, \dots, M\} \rightarrow \mathbb{R}$. pen_n est appelée la *pénalité*. Cette pénalité doit être choisie de façon à répondre à certaines exigences. Au choix :

1. Rendre la sélection de modèle consistante, c'est-à-dire $\widehat{m}_n$ converge en probabilité vers le vrai m .
2. Imiter le comportement de *l'oracle*.

Sélection de modèle consistante

Il faut déterminer quelles conditions asymptotiques sur la suite réelle $(\text{pen}_n(m))_n$ dans (2.27) garantissent la consistance de $\widehat{m}_n$.

Par exemple, on considère une famille finie de modèles $ARMA(p, q)$, ($0 \leq p \leq p_{\max}, 0 \leq q \leq q_{\max}$) gaussiens et centrés contenant le vrai modèle $ARMA(p^*, q^*)$. D'après Azencott et Dacunha-Castelle (1986, chapitre 14, théorème 4.3), si la fonction de contraste γ_n est l'opposé

de la log-vraisemblance et si la pénalité est de la forme $\text{pen}_n(p, q) = \frac{1}{2}(p + q)(1 + p + q)\beta_n$, où $(\beta_n)_n$ est une suite réelle positive, alors

$$\lim_{n \rightarrow \infty} \frac{\beta_n}{n} = 0 \quad \text{et} \quad \lim_{n \rightarrow \infty} \frac{\beta_n}{\log \log n} > 1$$

est une condition suffisante pour que $(\widehat{p}_n, \widehat{q}_n)$ défini par (2.27) converge presque-sûrement vers (p^*, q^*) .

Sélection de modèle et oracle

Pour simplifier l'exposé, prenons l'exemple de la sélection d'un modèle de régression. On dispose d'un vecteur aléatoire $\mathbf{y} = (y_i)_{1 \leq i \leq n}$ tel que $\mathbf{y} = \mathbf{s}^* + \boldsymbol{\epsilon}$, où les composantes ϵ_i de $\boldsymbol{\epsilon}$ sont de loi normale, indépendantes, centrées, de même variance 1 (sans perdre de généralité). S est un sous-espace vectoriel de $\mathbb{R}^n$ et $\|\cdot\|$ est la norme euclidienne sur $\mathbb{R}^n$. Ici,

$$\forall s \in S, \gamma_n(s) = \frac{1}{n} \|\mathbf{y} - s\|^2. \quad (2.28)$$

Idéalement, on aimerait sélectionner le modèle

$$m(s^*) = \arg \min_{0 \leq m \leq M} \mathbb{E} \left[\|\mathbf{s}^* - \widehat{s}_m\|^2 \right]. \quad (2.29)$$

$\widehat{s}_{m(s^*)}$ est appelé *l'oracle*. Ce n'est pas un estimateur de $\mathbf{s}^*$, car il dépend du vrai paramètre $\mathbf{s}^*$. Il faut donc sélectionner le modèle $\mathcal{M}_{\widehat{m}_n}$ de sorte que les propriétés de $\widehat{s}_{\widehat{m}_n}$ s'approchent de celles de l'oracle en un certain sens. L'idée de la sélection de modèle par minimisation du critère C_p de Mallows (1973) est de remplacer l'expression à minimiser dans l'équation (2.29) par une expression approchée ne dépendant pas de $\mathbf{s}^*$. La différence entre les deux critères tend vers 0 quand n tend vers l'infini et l'approximation provient de $\mathbb{E}[\|\widehat{s}_m\|^2] \approx \|\widehat{s}_m\|^2$. Cela revient à prendre une pénalité $\text{pen}_n(m) = \frac{2dm}{n}$. Dans le cas où la fonction de contraste γ_n est $-2/n \times$ la log-vraisemblance, un raisonnement de même type conduit au critère AIC (*Akaike information criterion*, voir Akaike, 1973, 1974), défini par la même pénalité. Toutefois, la construction du critère AIC implique un développement de Taylor à l'ordre deux de la log-vraisemblance. Par conséquent, cette dernière doit être au moins deux fois différentiable.

Dans ce raisonnement, l'approximation $\mathbb{E}[\|\widehat{s}_m\|^2] \approx \|\widehat{s}_m\|^2$ repose sur deux choses. D'une part, il faut que n soit suffisamment grand. Néanmoins, en ce qui concerne le critère AIC, une version adaptée AICc, proposée par Hurvich et Tsai (1989), existe pour les cas où cette condition n'est pas vérifiée. D'autre part, il faut que le nombre de modèles ne dépende pas de n . Ces conditions ne sont pas toujours remplies, ce qui a motivé l'approche non-asymptotique en sélection de modèle (Birgé et Massart, 2001 ; voir Massart, 2007, pour une présentation générale). Ici, plutôt que de compter sur la convergence de $\|\widehat{s}_m\|^2$ vers son espérance, on va contrôler ses variations autour de son espérance à l'aide d'une inégalité de concentration de la mesure, dans le but d'obtenir une inégalité oracle :

$$\exists C_n \geq 1, \mathbb{E}[\|\mathbf{s} - \widehat{s}_{\widehat{m}_n}\|^2] \leq C_n \inf_{0 \leq m \leq M} \mathbb{E}[\|\mathbf{s} - \widehat{s}_m\|^2]. \quad (2.30)$$

2.5.3 Sélection de modèle bayésienne

Introduction

On pourra se référer à Robert (2006, chapitre 7), pour une présentation générale. On reprendra ici des notations usuelles, quoique abusives, de la statistique bayésienne. Si X est une variable aléatoire :

- $P(X = x)$ est la densité de X en x par rapport à une certaine mesure de référence. La fonction de densité pourra être désignée par $P(X = \cdot)$.
- Si la densité de X dépend d'un paramètre $\theta \in \Theta$ et si on considère que ce paramètre est lui-même une variable aléatoire, la densité de X conditionnellement à θ est $P(X = \cdot | \theta)$.
- Toujours dans cette situation, $P(X | \theta)$ est la vraisemblance de X . La fonction de vraisemblance de X en $\vartheta \in \Theta$ est alors $P(X | \theta = \vartheta)$.

Dans cette approche bayésienne :

- Les modèles sont paramétrés, de sorte que dans le modèle $\mathcal{M}_m$, la loi des observations $\mathbf{y}$ dépende d'un paramètre $\theta_m \in \Theta_m$. On note alors $P(\mathbf{y} | m, \theta_m)$ la vraisemblance de $\mathbf{y}$.
- m est une variable aléatoire sur $\{0, \dots, M\}$.
- Le paramètre θ_m est une variable aléatoire, elle suit une loi à priori $P(\theta_m | m)$ conditionnellement à m .

On s'intéresse à la loi à posteriori de m : $P(m = \cdot | \mathbf{y})$. Cette loi est déterminée par la formule de Bayes :

$$P(m = \cdot | \mathbf{y}) = \frac{P(\mathbf{y} | m = \cdot) P(m = \cdot)}{P(\mathbf{y})} .$$

Le principe de la sélection de modèle bayésienne est de sélectionner le modèle qui maximise la probabilité à posteriori du modèle sachant les observations. Cela revient à estimer m par *Maximum à posteriori* :

$$\hat{m} = \arg \max_{0 \leq m' \leq M} P(m = m' | \mathbf{y}) . \quad (2.31)$$

Il est équivalent de maximiser le *facteur de Bayes* des modèles $\mathcal{M}_{m'}$ et $\mathcal{M}_0$:

$$\hat{m} = \arg \max_{0 \leq m' \leq M} \frac{P(m = m' | \mathbf{y})}{P(m = 0 | \mathbf{y})} = \frac{P(\mathbf{y} | m = m') P(m = m')}{P(\mathbf{y} | m = 0) P(m = 0)} = \arg \max_{0 \leq m' \leq M} P(\mathbf{y} | m = m') P(m = m') . \quad (2.32)$$

$P(\mathbf{y} | m)$ est la *vraisemblance intégrée* du modèle $\mathcal{M}_m$:

$$P(\mathbf{y} | m) = \int_{\theta_m \in \Theta_m} P(\mathbf{y} | m, \theta_m) P(\theta_m | m) d\theta_m . \quad (2.33)$$

Le facteur de Bayes des modèles $\mathcal{M}_{m'}$ et $\mathcal{M}_0$ est le rapport des vraisemblances intégrées de ces modèles, corrigées du facteur $\frac{P(m=m')}{P(m=0)}$ représentant l'a priori sur les modèles.

Pour résoudre ce problème de maximisation (2.32), il faut être capable de calculer la vraisemblance intégrée (2.33). Trois cas se présentent :

- Cette intégrale est aisément calculable. On se trouve dans cette situation quand la vraisemblance $P(\mathbf{y}|m, \theta_m)$ et la loi à priori $P(\theta_m = \cdot|m)$ ont de bonnes propriétés réciproques. On peut par exemple se trouver dans cette situation si la loi à priori $P(\theta_m = \cdot|m)$ du paramètre θ_m appartient à une *famille de lois de probabilités conjuguées* par $P(\mathbf{y}|m, \theta_m)$. $P(\theta_m = \cdot|\mathbf{y}, m)$ appartient alors à la même famille de lois de probabilité sur Θ_m . Le lecteur pourra se référer à Robert (2006, section 3.3) pour une présentation générale des lois à priori conjuguées.
- L'intégrale (2.33) est approchée par intégration numérique, faisant éventuellement appel à des méthodes d'échantillonnage. Voir Robert (2006, section 7.3).
- On fournit une approximation de l'intégrale (2.33) valable quand le nombre d'observations n tend vers l'infini. Cette méthode est présentée dans la section suivante.

Critère BIC

Est présenté ici le critère de Schwarz (1978), équivalent au critère appelé BIC pour *Bayesian information criterion*. Ce critère permet d'approcher $P(m = \cdot|\mathbf{y})$ par une log-vraisemblance maximisée pénalisée et, par conséquent, de remplacer le problème de maximisation de $P(m = \cdot|\mathbf{y})$ par le problème approché :

$$\hat{m} = \arg \min_{0 \leq m' \leq M} \left(-2\hat{\mathcal{L}}_{m'} + \text{pen}_n(m') \right) , \quad (2.34)$$

où $\hat{\mathcal{L}}_{m'}$ est le maximum de la log-vraisemblance du modèle $\mathcal{M}_{m'}$.

Nous présentons le critère BIC dans le contexte où variables aléatoires y_i sont i.i.d., suivant la présentation de Kass et Raftery (1995) et de Lebarbier et Mary-Huard (2006). $\mathcal{M}_m$ est ici le modèle auquel appartient la loi de chacune de ces observations. Ici, m est une variable aléatoire sur $\{0, \dots, M\}$, que l'on va supposer de loi uniforme pour simplifier (c'est la loi à priori sur les modèles). D'après (2.32), comme m suit la loi uniforme, il est équivalent de maximiser la loi à posteriori $P(m = \cdot|\mathbf{y})$ ou la vraisemblance intégrée. On peut écrire la vraisemblance intégrée (2.33) sous la forme

$$P(\mathbf{y}|m) = \int_{\Theta_m} e^{nL_n(\vartheta)} d\vartheta , \quad (2.35)$$

où

$$L_n(\vartheta) = \frac{1}{n} \sum_{i=1}^n \log(P(y_i|m, \theta_m = \vartheta)) + \frac{\log P(\theta_m = \vartheta|m)}{n} . \quad (2.36)$$

Proposition 2.5 (Approximation de Laplace). *Soient $d \geq 1$ entier et $L : \mathbb{R}^d \rightarrow \mathbb{R}$ une fonction trois fois différentiable sur $\mathbb{R}^d$. Supposons que L atteint son unique maximum sur $\mathbb{R}^d$ en $\bar{x}$. Alors,*

$$\int_{\mathbb{R}^d} e^{nL(x)} dx = e^{nL(\bar{x})} \left(\frac{2\pi}{n} \right)^{d/2} | -L''(\bar{x}) | + O(n^{-1}) .$$

Nous ne discuterons pas ici des conditions qui permettent d'appliquer ce résultat à la suite de fonctions L_n (voir Lebarbier et Mary-Huard, 2006). Ici, l'approximation de Laplace reste valable, à une modification du terme d'erreur près :

$$P(\mathbf{y}|m) = e^{nL_n(\bar{\theta}_m)} \left(\frac{2\pi}{n} \right)^{d_m/2} | -L_n''(\bar{\theta}_m) | + O_P(n^{-1}) ,$$

où $\bar{\theta}_m$ est l'argument maximum de L_n sur Θ_m . D'après (2.36), le maximum de vraisemblance $\hat{\theta}_m$ maximise $L_n(\vartheta) - \frac{P(\theta=\vartheta|m)}{n}$. Or, $\frac{\log P(\theta=\vartheta|m)}{n}$ est d'ordre $O(n^{-1})$ donc, asymptotiquement, $\bar{\theta}_m$ peut être remplacé par $\hat{\theta}_m$. Finalement, on obtient

$$\log(P(\mathbf{y}|m)) = \hat{\mathcal{L}}_m - \frac{1}{2}d_m \log(n) + O_P(1) , (n \rightarrow \infty) . \quad (2.37)$$

Par conséquent, on propose de sélectionner le modèle suivant :

$$\hat{m} = \arg \min_{0 \leq m' \leq M} BIC(m') , \quad (2.38)$$

$$\text{où } BIC(m') = -2\hat{\mathcal{L}}_{m'} + d_{m'} \log(n) . \quad (2.39)$$

Kass et Raftery (1995) font une distinction entre le *critère de Schwarz* et le critère BIC. Le critère de Schwarz est égal, par définition, à $-1/2 \times$ le critère BIC. C'est ce dernier critère qui approche la probabilité à posteriori du modèle (2.37). Toutefois, la confusion est grande entre ces deux critères qui sélectionnent le même modèle et il n'est pas rare de trouver l'expression « le critère BIC de Schwarz » dans la littérature.

Remarque 2.4. • On retrouve un critère pénalisé, ce qui fait le lien avec le cadre fréquentiste défini dans la section 2.5.2.

- BIC pénalise plus la log-vraisemblance maximisée qu'AIC. Les modèles sélectionnés sont donc plus parcimonieux.
- Comme le fait remarquer Choi (1992, section 3.5), le critère BIC a été introduit à quelques années d'intervalle par des auteurs différents, les familles de modèles considérés et les arguments employés variant avec eux. Akaike (1977, 1978a,b) propose le critère BIC pour différents problèmes de sélection de modèle, dont la sélection des ordres p et q d'une série temporelle $ARMA(p, q)$ gaussienne centrée. Toutefois, l'approximation de Laplace n'est pas encore la justification de ces critères, des lois à priori spécifiques sont définies pour obtenir une identité entre le modèle qui maximise la probabilité à posteriori du modèle et celui qui minimise le critère BIC. Hannan (1980) démontre que les ordres $\hat{p}_n$ et $\hat{q}_n$ ainsi sélectionnés sont consistants.
- La construction du critère BIC est donnée dans le cas d'un échantillon i.i.d., mais ce critère peut s'utiliser dans bien d'autres cas, tant que les lois à priori et la log-vraisemblance sont suffisamment régulières et que la notion de dimension du modèle est bien définie.
- Les hypothèses de la proposition 2.5 imposent des conditions de régularité à la fonction de log-vraisemblance pour que l'approximation (2.37) reste valable. Comme le font

remarquer Berger et Pericchi (2001), l'irrégularité de la vraisemblance pour certains modèles fait que l'approximation (2.37) n'est plus valable. Dans le cas qui nous intéresse, la segmentation, nous verrons dans la section 2.5.4 qu'une autre pénalité nous permettra d'approcher la probabilité à posteriori du modèle.

2.5.4 Sélection du nombre de ruptures dans un modèle de segmentation

Nous allons reprendre chacun des principes de sélection de modèle exposés dans les sections 2.5.2 et 2.5.3 et s'intéresser à leur application éventuelle en segmentation. La difficulté principale dans ce cadre spécifique tient de la nature discrète des instants de rupture. m est ici le nombre de ruptures. C'est pour estimer ce nombre de ruptures que la sélection de modèle est nécessaire.

Sélection consistante du nombre de ruptures

On peut imposer que les variables n et m soient séparées dans la pénalité :

$$\text{pen}_n(m) = \beta_n m .$$

On étudie la convergence du nombre de ruptures estimé $\widehat{m}_n$ vers m en fonction des propriétés asymptotiques de la suite (β_n) . Ce type de résultat est fréquent dans la littérature (C.-B. Lee, 1995 ; Lavielle, 1999, théorème 3.1 ; Lavielle et Moulines, 2000, théorème 9 ; J. V. Braun, R. K. Braun *et al.* 2000, théorème 4 ; Bardet, Kengne et Wintenberger, 2012, théorème 3.1). Les propositions 3.4 et 3.5 sont des résultats de ce type.

Reste le problème de la spécification de (β_n) . En effet, les résultats cités ci-dessus ne donnent que des conditions de convergence de $\widehat{m}_n$ et absolument pas de formule pour (β_n) . On peut fixer une forme de β_n à priori, dépendante de constantes non spécifiées (voir par exemple la remarque 3.4). Les conditions sur (β_n) sont alors traduites en conditions sur ces constantes.

Sélection du nombre de ruptures par imitation d'un oracle

L'approximation à la source du critère du C_p de Mallows n'est plus valable dans le cadre des modèles de segmentation (Massart, 2007, chapitre 4). Cela motive l'utilisation d'un critère non-asymptotique.

On considère $\mathbf{y} = (y_1, \dots, y_n)$, un vecteur gaussien de composantes indépendantes et de même variance 1 (sans perdre de généralité). On cherche à sélectionner le nombre de ruptures m dans l'espérance de $\mathbf{y}$. La fonction de contraste est définie par (2.28).

Pour établir une inégalité oracle, Lebarbier (2005) utilise les résultats de Birgé et Massart (2001) et propose la pénalité suivante :

$$\text{pen}_n(m) = \frac{m+1}{n} \left(c_1 \log \left(\frac{n}{m+1} \right) + c_2 \right) , \quad (2.40)$$

$c_1 > 0$ et $c_2 > 0$ étant deux constantes.

Critères BIC et BIC modifié

Yao (1988) propose l'utilisation du critère BIC défini en (2.39) pour la segmentation de l'espérance d'un processus gaussien à variance inconnue et démontre la consistance du nombre de ruptures ainsi sélectionné. Toutefois, une étude de simulations (Zhang, 2005, section 2.4.) comparant cette méthode à quatre autres met en lumière ses faibles performances.

Contrairement au raisonnement de Schwarz (1978), le critère BIC, dans le cadre de la segmentation, ne correspond pas à une approximation de la probabilité à posteriori du nombre de ruptures sachant les observations. La difficulté d'approcher cette probabilité dans le problème de la segmentation provient de la discontinuité de la fonction de vraisemblance en les paramètres de localisation des ruptures. Or, l'approximation de Laplace (cf. proposition 2.5), permettant de lier le critère BIC à cette probabilité, nécessite des conditions de régularité qui ne sont pas remplies dans le cas présent.

Il est possible de contourner cet obstacle (Zhang et Siegmund, 2007) et d'obtenir une approximation du facteur de Bayes par le résultat qui suit.

Theorème 2.11. *Soit m une v.a. sur $\mathbb{N}$, $\boldsymbol{\mu} = (\mu_0, \dots, \mu_m)$ un vecteur aléatoire réel et σ v.a.r. positive. Conditionnellement à m , $\mathbf{t} = (t_0, \dots, t_{m+1})$ est une v.a. sur*

$$\mathcal{C}_{m,n} = \{(t'_0, \dots, t'_{m+1}) \in \mathbb{R}^{m+2}; 0 = t'_0 < t'_1 < \dots < t'_m < t'_{m+1} = 1\}.$$

Les lois à priori sont spécifiées par Zhang et Siegmund (2007, Théorème 2). Soit $\mathbf{y} = (y_1, \dots, y_n)$ un vecteur aléatoire réel tel que, conditionnellement à m , $\boldsymbol{\mu}$, $\mathbf{t}$ et σ , les y_i sont indépendants et de loi $\mathcal{N}(\mu_k, \sigma^2)$ si $t_k < i \leq t_{k+1}$. Alors,

$$\begin{aligned} \log \frac{P(m|\mathbf{y})}{P(m=0|\mathbf{y})} &= \frac{n+1}{2} \log(SS_{all}) - \frac{n-m+1}{2} \log(SS_{wg}) + \log \left(\frac{\Gamma(\frac{n-m+1}{2})}{\Gamma(\frac{n+1}{2})} \right) \\ &\quad - \frac{1}{2} \sum_{k=0}^m \log \widehat{n}_k + \left(\frac{1}{2} - m \right) \log(n) + O_P(1) \ (n \rightarrow +\infty), \end{aligned} \quad (2.41)$$

où

$$\begin{aligned} SS_{all} &= \min_{\boldsymbol{\mu} \in \mathbb{R}} \sum_{i=1}^n (y_i - \mu)^2, \\ SS_{wg} &= \min_{\mathbf{t} \in \mathcal{C}_m} \min_{\boldsymbol{\mu} \in \mathbb{R}^{m+1}} \sum_{k=0}^m \sum_{i=t_k+1}^{t_{k+1}} (y_i - \mu_k)^2, \\ \widehat{n}_k &= \widehat{t}_{k+1} - \widehat{t}_k, \text{ où} \\ (\widehat{t}_0, \dots, \widehat{t}_{m+1}) &= \arg \min_{\mathbf{t} \in \mathcal{C}_m} \min_{\boldsymbol{\mu} \in \mathbb{R}^{m+1}} \sum_{k=0}^m \sum_{i=t_k+1}^{t_{k+1}} (y_i - \mu_k)^2, \end{aligned}$$

et où Γ est la fonction Gamma d'Euler.

L'approximation du facteur de Bayes (2.41) permet de définir un critère pénalisé de sélection de modèle, qu'on appelle *BIC modifié* (mBIC) de Zhang et Siegmund (2007).

Autres méthodes bayésiennes de sélection du nombre de ruptures

Rigaill *et al.* (2012) proposent l'utilisation de lois à priori conjuguées des paramètres propres à chaque segment dans plusieurs modèles de segmentation où les observations sont indépendantes (conditionnellement aux paramètres) ainsi que certaines lois à priori spécifiques pour les paramètres de localisation des ruptures. Ce choix permet l'obtention d'une forme analytique de la vraisemblance intégrée (2.33) et son calcul. Toutefois, les lois à priori des paramètres dépendent d'hyperparamètres, que l'on retrouve dans la loi à posteriori du nombre de ruptures.

Calibration des constantes et hyperparamètres

Parmi les critères de sélection de modèle pour la segmentation présentés, on trouve des constantes non spécifiées. Lavielle (2005) et Arlot et Massart (2009) proposent des heuristiques de calibration de ces constantes. Dans le cadre bayésien, la présence d'hyperparamètres pose un problème similaire (Rigaill *et al.* 2012). L'avantage du critère BIC modifié est l'absence de telles constantes.

2.6 Problèmes algorithmiques en segmentation

Les questions algorithmiques occupent une place importante en segmentation. La recherche de la segmentation optimale (en un sens à préciser) mène à un problème d'optimisation discrète sur un ensemble fini pouvant être de grande taille, qui ne peut être résolu en un temps raisonnable par l'exploration totale de cet ensemble.

Il s'agit de trouver la segmentation optimale (en un sens à préciser) à m ruptures d'une série temporelle $\mathbf{y} = (y_1, \dots, y_N)$, où $m + 1 \leq N$. On ne fixe pas à priori de restriction sur l'ensemble auquel appartiennent ces observations.

Définissons, pour $1 \leq k + 1 \leq \min(m + 1, n) \leq \max(m + 1, n) \leq N$,

$$C_{k,n} = \min_{\mathbf{t}=(t_0,\dots,t_{k+1}) \in \mathcal{A}_{n,k}} \min_{\boldsymbol{\theta}=(\theta_1,\dots,\theta_{k+1}) \in \Theta^{k+1}} \sum_{l=1}^{k+1} f_{t_{l-1}+1,t_l}(\mathbf{y}, \theta_l), \quad (2.42)$$

où :

- $\mathcal{A}_{n,k} = \{(t_0, \dots, t_{k+1}) \in \mathbb{N}^{k+2}; 0 = t_0 < t_1 < \dots < t_k < t_{k+1} = n\}$. C'est l'ensemble des segmentations à k ruptures d'une série à n observations.
- On ne fixe pas de restriction à priori à l'ensemble Θ . À ce niveau, la seule contrainte sur l'ensemble sur lequel se fait la minimisation en $\boldsymbol{\theta}$ est d'être un produit cartésien Θ^{k+1} .
- Les fonctions $f_{i,j}$ ($1 \leq i \leq j \leq N$) dépendent des observations et d'un paramètre (θ_l pour le l^e terme de la somme). Ces fonctions sont inférieurement bornées. On souhaite que la minimisation de ces fonctions en le paramètre θ_l ne soit pas trop coûteuse (en un sens à préciser).

$C_{k,n}$ est le *coût* de la segmentation optimale à k ruptures de $y_1, \dots, y_n$. Le but est de calculer $C_{m,N}$, ainsi que la segmentation $\mathbf{t}$ et le paramètre $\boldsymbol{\theta}$ qui réalisent ce minimum. Une fois calculée la segmentation optimale $\widehat{\mathbf{t}}$, il est facile de retrouver le paramètre optimal $\widehat{\boldsymbol{\theta}}$ par

$$\widehat{\boldsymbol{\theta}}_l = \arg \min_{\boldsymbol{\theta}_l \in \Theta} f_{\widehat{t}_{l-1}+1, \widehat{t}_l}(\mathbf{y}, \boldsymbol{\theta}_l) .$$

Le problème est donc essentiellement de retrouver la segmentation optimale. Or, l'ensemble $\mathcal{A}_{N,m}$ est de cardinal C_{N-1}^m . Tester chaque segmentation possible peut être très coûteux en temps de calcul. C'est notamment le cas dans certaines applications en génomique où N est d'ordre 10^5 ou 10^6 . L'algorithme de programmation dynamique d'Auger et Lawrence (1989), décrit dans la section 2.6.1, permet de résoudre ce problème. Pour $1 \leq i \leq j \leq N$, on définit :

$$C(i, j) = \min_{\boldsymbol{\theta} \in \Theta} f_{i,j}(\mathbf{y}, \boldsymbol{\theta}) .$$

Ainsi

$$C_{k,n} = \min_{\mathbf{t}=(t_0, \dots, t_{k+1}) \in \mathcal{A}_{n,k}} \sum_{l=1}^{k+1} C(t_{l-1} + 1, t_l) . \quad (2.43)$$

Si la *matrice de coût* $(C(i, j))_{1 \leq i \leq j \leq N}$ est connue, le problème d'optimisation est simplifié.

2.6.1 Programmation dynamique

La programmation dynamique a été introduite par Bellman (1954). Sa philosophie est résumée par le Principe d'optimalité de Bellman :

« Une politique optimale a la propriété que, quels que soient l'état initial et les décisions initiales, les décisions restantes doivent constituer une politique optimale vis-à-vis du résultat des décisions précédentes. » Bellman (1954).

L'utilisation de la programmation dynamique en segmentation a été introduite par Auger et Lawrence (1989).

Pour calculer $C_{m,N}$, on calcule $C_{k,n}$ pour tous k, n tels que $1 \leq k+1 \leq \min(m+1, n) \leq \max(m+1, n) \leq N$. (2.43) donne une relation de récurrence :

$$\begin{aligned} C_{k,n} &= \min_{s=k, \dots, n-1} \min_{(t_0, \dots, t_{k+1}) \in \mathcal{A}_{n,k}, t_k=s} \sum_{l=1}^{k+1} C(t_{l-1} + 1, t_l) \\ &= \min_{s=k, \dots, n-1} \left\{ C(s+1, n) + \min_{(t_0, \dots, t_{k+1}) \in \mathcal{A}_{n,k}, t_k=s} \sum_{l=1}^k C(t_{l-1} + 1, t_l) \right\} \\ &= \min_{s=k, \dots, n-1} \left\{ C(s+1, n) + \min_{(t_0, \dots, t_k) \in \mathcal{A}_{s,k-1}} \sum_{l=1}^k C(t_{l-1} + 1, t_l) \right\} \\ &= \min_{s=k, \dots, n-1} \{ C(s+1, n) + C_{k-1,s} \} . \end{aligned} \quad (2.44)$$

Cette relation permet d'écrire l'algorithme de programmation dynamique pour le calcul de $C_{m,N}$ (cf. algorithme 1).

```

début
  // Calcul de la matrice de coût:
  pour  $i = 1, \dots, N$  faire
    pour  $j = i, \dots, N$  faire
      | Calculer  $C(i, j)$ 
    fin
  fin
  // Initialisation:
  pour  $i = 1, \dots, N$  faire
    |  $C_{0,n} := C(1, n)$ 
  fin
  // Récurrence:
  pour  $k = 1, \dots, m$  faire
    pour  $n = k + 1, \dots, N$  faire
      |  $C_{k,n} := \min_{s=k, \dots, n-1} \{C(s + 1, n) + C_{k-1,s}\}$ 
    fin
  fin
fin

```

Algorithme 1 : Segmentation par programmation dynamique.

Complexité en temps et en espace

La complexité en temps des phases d'initialisation et de récurrence de l'algorithme 1 est en $O(mN^2)$. Si le calcul de la matrice de coût est en $O(N^2)$, la complexité en temps de l'ensemble de l'algorithme est en $O(mN^2)$. Il n'est pas possible d'obtenir une telle borne si le problème est posé en toute généralité comme dans l'équation (2.42). Cette borne est vraie dans un certain nombre de cas comme celui des moindres carrés (cf. exemple 2.5). Pour obtenir une borne de ce type, il suffit que, pour tous $1 \leq i \leq j < N$, $C(i, j + 1)$ se calcule en temps borné (avec une complexité en $O(1)$) à partir de $C(i, j)$ et éventuellement d'un nombre borné d'autres variables calculées en même temps que $C(i, j)$ en temps borné; et que les $C(i, i)$ se calculent aussi en temps borné.

La complexité en espace, à priori en $O(N^2)$ à cause de la matrice de coût, peut être réduite à $O(mN)$ (Guédon, 2013) en calculant les termes de cette matrice au fur et à mesure, le long de l'algorithme, en utilisant une relation de récurrence.

Retour sur trace

Sous cette forme, l'algorithme 1 permet de calculer le coût de la segmentation optimale, mais n'a pas comme sortie la segmentation optimale elle-même. Cet algorithme est complété par une procédure de retour sur trace (*backtracking* en anglais, le retour sur les décisions prises dans l'exécution de l'algorithme) qui permet de reconstruire cette dernière. Par exemple,

Guédon (2013) en donne le pseudo-code, cf. l'algorithme 2.

On calcule $T(k, n)$ pour tous $1 \leq k \leq m, k+1 \leq n \leq N$, défini par

$$T(k, n) = \arg \min_{s=k, \dots, n-1} \{C(s+1, n) + C_{k-1, s}\} .$$

$T(k, n)$ est calculé en même temps que $C_{k,n}$ dans l'algorithme 1. Cette matrice T permet de reconstruire la segmentation optimale à la fin de l'exécution de l'algorithme 1.

début

```

|
|                                     Seg := N
|                                     s := N
|                                     k := m
|
|    tant que k > 0 faire
|    |
|    |                                     s := T(k, s)
|    |                                     Seg := s || Seg
|    |                                     k := k - 1
|    |
|    fin
|    // || est l'opérateur de concaténation.
|    retourner Seg // Sortie: Segmentation optimale.
fin
```

Algorithme 2 : Retour sur trace.

Exemples

Exemple 2.5 (Moindres carrés). Dans cet exemple :

- $y_1, \dots, y_N \in \mathbb{R}$,
- $\Theta = \mathbb{R}$,
- Les fonctions $f_{i,j}$ sont définies par $\forall \mu \in \mathbb{R}, f_{i,j}(\mathbf{y}, \mu) = \sum_{s=i}^j (y_s - \mu)^2$.

Ainsi il s'agit de calculer

$$C_{m,N} = \min_{\mathbf{t}=(t_0, \dots, t_{m+1}) \in \mathcal{A}_{N,m}} \min_{\boldsymbol{\mu}=(\mu_1, \dots, \mu_{m+1}) \in \mathbb{R}^{m+1}} \sum_{k=1}^{m+1} \sum_{i=t_{k-1}+1}^{t_k} (y_i - \mu_k)^2 .$$

Pour tous $1 \leq i \leq j \leq N$,

$$C(i, j) = \sum_{s=i}^j (y_s - \bar{y}_{i:j})^2 ,$$

où $\bar{y}_{i:j} = (j-i+1)^{-1} \sum_{s=i}^j y_s$.

En particulier, $\forall i = 1, \dots, N, C(i, i) = 0$ et pour tous $1 \leq i \leq j < N$,

$$\sum_{s=i}^{j+1} y_s = \left(\sum_{s=i}^j y_s \right) + y_{j+1}, \quad (2.45)$$

$$C(i, j+1) = \frac{j-i+1}{j-i+2} \{C(i, j) + y_{j+1}^2\} - \frac{2}{j-i+2} y_{j+1} \sum_{s=i}^j y_s. \quad (2.46)$$

Si, en même temps que $C(i, j)$, on calcule $\sum_{s=i}^j y_s$, qui se met à jour très simplement en temps borné (cf. équation (2.45)), la formule de récurrence (2.46) permet le calcul en temps borné de $C(i, j+1)$ en fonction de $C(i, j)$ et de $\sum_{s=i}^j y_s$. Cela garantit le calcul de la matrice de coût en un temps $O(N^2)$, ce qui garantit la complexité en temps de l'algorithme 1 en $O(mN^2)$ dans cet exemple.

Exemple 2.6 (Vraisemblance conditionnelle gaussienne $AR(1)$ avec ruptures dans le paramètre d'autorégression). Supposons que $Y_0, \dots, Y_N$ vérifient

$$Y_i = \rho_l Y_{i-1} + \epsilon_i \text{ si } t_{l-1} < i \leq t_l,$$

où les $\rho_l \in \mathbb{R}$ sont inconnus, les ϵ_i sont normales, centrées, de variance σ^2 inconnue, i.i.d. Soient $y_0, \dots, y_N$ des réalisations de $Y_0, \dots, Y_N$. Maximiser la vraisemblance conditionnellement à $Y_0 = y_0$, si l'on considère que cette fonction est déjà maximisée en σ^2 , se ramène au problème de minimisation :

$$\min_{\mathbf{t}=(t_0, \dots, t_{m+1}) \in \mathcal{A}_{N,m}(\rho_1, \dots, \rho_{m+1})} \min_{(\rho_1, \dots, \rho_{m+1}) \in \mathbb{R}^{m+1}} \sum_{l=1}^{m+1} \sum_{i=t_{l-1}+1}^{t_l} (y_i - \rho_l y_{i-1})^2. \quad (2.47)$$

On notera que dans cet exemple, contrairement aux moindres carrés, le l^e terme de la somme de l'équation (2.47) dépend d'une observation y_i avec $i \notin \{t_{l-1} + 1, \dots, t_l\}$.

Les conditions minimales sur les termes de la somme de l'équation (2.42) pour pouvoir utiliser l'algorithme 1 sont, pour tout l , que le l^e terme ne dépende pas :

- des t_r avec $r \notin \{l-1, l\}$,
- des θ_r avec $r \neq l$.

Exemple 2.7 (Vraisemblance gaussienne $AR(1)$ avec ruptures dans l'espérance). Supposons que $Y_0, \dots, Y_N$ vérifient

$$\begin{aligned} Y_i - \mathbb{E}Y_i &= \rho(Y_{i-1} - \mathbb{E}Y_{i-1}) + \epsilon_i \\ \mathbb{E}Y_i &= \mu_l \text{ si } t_{l-1} < i \leq t_l, \end{aligned}$$

où $\rho \in \mathbb{R}$ est connu, les $\mu_l \in \mathbb{R}$ sont inconnus, les ϵ_i sont normales, centrées, de variance σ^2 inconnue, i.i.d. Soient $y_0, \dots, y_N$ des réalisations de $Y_0, \dots, Y_N$. Maximiser la vraisemblance conditionnellement à y_0 , si l'on considère que cette fonction est déjà maximisée en σ^2 , se

ramène à minimiser en $\mathbf{t} = (t_0, \dots, t_{m+1}) \in \mathcal{A}_{N,m}$ et en $(\mu_1, \dots, \mu_{m+1}) \in \mathbb{R}^{m+1}$ l'expression suivante :

$$\sum_{l=1}^{m+1} \left\{ \sum_{i=t_{l-1}+2}^{t_l} (y_i - \rho y_{i-1} - (1-\rho)\mu_l)^2 + (y_{t_l+1} - \mu_l - \rho(y_{t_l} - \mu_{l-1}))^2 \right\}.$$

La programmation dynamique ne permet pas de résoudre ce problème à cause du produit croisé $\mu_l \mu_{l-1}$ qui apparaît dans le développement de l'expression.

De même, cet algorithme ne peut pas être à priori adapté au cas où tous ces termes partagent un même paramètre global. Cette adaptation n'est possible que dans des conditions très particulières où le minimum en ce paramètre global a la propriété de pouvoir s'écrire comme une expression similaire à celle de l'équation (2.42), comme dans l'exemple suivant.

Exemple 2.8 (Vraisemblance gaussienne et moindres carrés). Il s'agit de résoudre le problème d'optimisation suivant :

$$\max_{\mathbf{t}=(t_0,\dots,t_{m+1}) \in \mathcal{A}_{N,m}} \max_{\sigma > 0} \max_{(\mu_1,\dots,\mu_{m+1}) \in \mathbb{R}^{m+1}} \sum_{l=1}^{m+1} \sum_{i=t_{l-1}+1}^{t_l} \log \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp \left(-\frac{1}{2\sigma^2} (y_i - \mu_l)^2 \right) \right\}.$$

Le maximum en σ est atteint en $\sigma = \left(\frac{1}{N} \sum_{l=1}^{m+1} \sum_{i=t_{l-1}+1}^{t_l} (y_i - \mu_l)^2 \right)^{1/2}$. Maximisé en σ , le problème se ramène donc au problème d'optimisation suivant :

$$\max_{\mathbf{t}=(t_0,\dots,t_{m+1}) \in \mathcal{A}_{N,m}} \max_{(\mu_1,\dots,\mu_{m+1}) \in \mathbb{R}^{m+1}} -\frac{N}{2} \left\{ \log(2\pi) + \log \left(\frac{1}{N} \sum_{l=1}^{m+1} \sum_{i=t_{l-1}+1}^{t_l} (y_i - \mu_l)^2 \right) \right\}.$$

Ce problème est équivalent aux moindres carrés.

Exemple 2.9 (Vraisemblance binomiale négative). Ici, il s'agit de résoudre le problème d'optimisation suivant :

$$\min_{\mathbf{t}=(t_0,\dots,t_{m+1}) \in \mathcal{A}_{N,m}} \min_{r > 0} \min_{\boldsymbol{\mu}=(\mu_1,\dots,\mu_{m+1}) \in]0,+\infty[^{m+1}} \sum_{l=1}^{m+1} \sum_{i=t_{l-1}+1}^{t_l} \log \left\{ \frac{(r + \mu_l)^{r+y_i} y_i! \Gamma(r)}{r^r \mu_l^{y_i} \Gamma(r + y_i)} \right\},$$

où Γ est la fonction Gamma d'Euler. La programmation dynamique ne permet pas de résoudre ce problème de minimisation. Mais si le paramètre r est connu, la minimisation en $\boldsymbol{\mu}$ peut se faire à l'aide de l'algorithme 1.

2.6.2 Programmation dynamique élaguée

L'élagage est une technique qui permet d'accélérer des algorithmes basés sur des disjonctions de cas (« Diviser pour régner », programmation dynamique). L'idée est de se donner une variable, facile à mettre à jour durant l'exploration de l'ensemble des possibilités, qui, si elle remplit un certain critère, considère un des sous-cas envisagés comme ne contenant pas l'optimum recherché. Cette variable peut être un nombre, une fonction ou un ensemble.

Rigaill (2010) utilise l'élagage pour accélérer l'algorithme de programmation dynamique pour la segmentation. Cet algorithme a été implémenté pour plusieurs fonctions de coût dans un package R nommé `Segmentor3IsBack` (Cleynen *et al.* 2014).

Cet algorithme peut s'utiliser dans le cas particulier du problème de minimisation de l'équation (2.42), où

$$f_{i,j}(\mathbf{y}, \theta) = \sum_{s=i}^j \gamma(y_s, \theta) . \quad (2.48)$$

Au lieu de calculer directement les $C_{k,n}$, nous nous intéressons à un problème de minimisation fonctionnelle plus complexe :

$$R_{k,n}(\theta) = \min_{\nu=k, \dots, n-1} R_{k,\nu,n}(\theta) ,$$

où

$$R_{k,\nu,n}(\theta) = C_{k-1,\nu} + \sum_{i=\nu+1}^n \gamma(y_i, \theta) .$$

Les $C_{k,n}$ et les $R_{k,n}$ sont liés par la relation $C_{k,n} = \min_{\theta} R_{k,n}(\theta)$.

La relation de récurrence (2.44) permet de calculer $C_{k,n}$ avec une complexité en $O(n)$. La relation de récurrence suivante permettrait de calculer $R_{k,n}(\theta)$ en temps borné :

$$R_{k,n}(\theta) = \min \{ R_{k,n-1}(\theta), C_{k-1,n-1} + \gamma(y_n, \theta) \} . \quad (2.49)$$

Le problème est qu'il est impossible de faire ce calcul pour tous les θ envisagés (possiblement une infinité).

Il est en revanche possible de déterminer l'ensemble $\mathcal{S}_{n,n-1}^k$ des θ tels que $C_{k-1,n-1} + \gamma(y_n, \theta) \leq R_{k,n}(\theta)$, c'est-à-dire l'ensemble des θ tel que c'est la deuxième partie du terme de droite de l'équation (2.49) qui réalise le minimum.

Plus généralement, on définit

$$\mathcal{S}_{n,\nu}^k = \left\{ \theta \in \Theta; C_{k-1,\nu} + \sum_{i=\nu+1}^n \gamma(y_i, \theta) \leq R_{k,n}(\theta) \right\}$$

pour tous $1 \leq k \leq \nu < n \leq N$ et $k \leq m$. C'est l'ensemble des θ , paramètre du dernier segment, tel que ν est la dernière rupture optimale dans la segmentation à k ruptures de $y_1, \dots, y_n$. Cela permet de définir l'ensemble des dernières ruptures optimales pour un certain θ :

$$\tau_n^k = \{ \nu \in \{k, \dots, n-1\}; \mathcal{S}_{n,\nu}^k \neq \emptyset \} .$$

pour tous $1 \leq k \leq \min(m, n) \leq \max(m, n) \leq N$.

Les relations suivantes permettent le calcul par récurrence de ces ensembles.

Proposition 2.6. *Pour tous $1 \leq k \leq m$ et $k \leq n < N$,*

$$\mathcal{S}_{n+1,n}^k = \{ \theta \in \Theta; \forall \nu \in \tau_n^k, C_{k-1,n} \leq R_{k,\nu,n}(\theta) \} , \quad (2.50)$$

$$\forall \nu \in \{k, \dots, n-1\}, \mathcal{S}_{n+1,\nu}^k = \mathcal{S}_{n,\nu}^k \cap \{ \theta \in \Theta; R_{k,\nu,n}(\theta) \leq C_{k-1,n} \} , \quad (2.51)$$

$$\forall \nu \in \{k, \dots, n-1\}, \mathcal{S}_{n,\nu}^k = \emptyset \Rightarrow \forall n' \in \{n, \dots, N\}, \mathcal{S}_{n',\nu}^k = \emptyset , \quad (2.52)$$

$$\tau_{n+1}^k = \{ \nu \in \tau_n^k \cup \{n\}; \mathcal{S}_{n+1,\nu}^k \neq \emptyset \} . \quad (2.53)$$

La proposition 2.6 permet d'écrire l'algorithme de programmation dynamique élagué pour la segmentation. La relation (2.52) fournit un critère d'arrêt de l'exploration de l'ensemble des segmentations.

début

pour $k = 1, \dots, m$ **faire**

$$\tau_{k+2}^k := \{k\}$$

$$R_{k,k+1,k+2}(\theta) := C_{k-1,k+1} + \gamma(y_{k+2}, \theta)$$

pour $n = k + 2, \dots, N - 1$ **faire**

pour $\nu \in \tau_n^k$ **faire**

$$\mathcal{S}_{n+1,\nu}^k := \mathcal{S}_{n,\nu}^k \cap \{\theta \in \Theta; R_{k,\nu,n}(\theta) \leq C_{k-1,n}\}$$

fin

// D'après (2.51)

$$\mathcal{S}_{n+1,n}^k = \{\theta \in \Theta; \forall \nu \in \tau_n^k, C_{k-1,n} \leq R_{k,\nu,n}(\theta)\} \text{ // D'après (2.50)}$$

$$\tau_{n+1}^k = \{\nu \in \tau_n^k \cup \{n\}; \mathcal{S}_{n+1,\nu}^k \neq \emptyset\} \text{ // D'après (2.53)}$$

pour $\nu \in \tau_{n+1}^k$ **faire**

$$R_{k,\nu,n+1}(\theta) := R_{k,\nu,n}(\theta) + \gamma(y_{n+1}, \theta)$$

fin

$$D_{k,n+1} := \min_{\substack{\nu \in \tau_{n+1}^k \\ \theta \in \Theta}} R_{k,\nu,n+1}(\theta)$$

$$I_{k,n+1} := \arg \min_{\substack{\nu \in \tau_{n+1}^k \\ \theta \in \Theta}} R_{k,\nu,n+1}(\theta)$$

// $D_{k,n+1}$ coût optimal de la segmentation à k ruptures de

$y_1, \dots, y_{n+1}$.

// $I_{k,n+1}$ dernière rupture optimale de la segmentation à k

ruptures de $y_1, \dots, y_{n+1}$.

fin

fin

fin

Algorithme 3 : Segmentation par programmation dynamique élaguée.

Complexité

Proposition 2.7. *Si :*

1. Θ est un intervalle de $\mathbb{R}$,
2. pour tous $1 \leq i \leq j \leq N$, $\theta \mapsto \sum_{s=i}^j \gamma(y_s, \theta)$ a un unique extremum local (qui est le minimum global),
3. $\forall k \in \{1, \dots, m\}, \forall n \in \{k+1, \dots, N\}, \forall \nu \in \{k, \dots, n-1\}, R_{k,\nu,n}$ est minimisé en temps borné,
4. $\forall k \in \{1, \dots, m\}, \forall n \in \{k+1, \dots, N\}, \forall \nu \in \{k, \dots, n-1\}$, les racines de $R_{k,\nu,n}(\theta) = C_{k-1,t}$ sont calculées en temps borné,

alors l'algorithme 3 a :

1. une complexité en temps $O(mN^2)$,
2. une complexité en espace $O(mN)$.

La première hypothèse de la proposition 2.7 tient de propriétés des intervalles de $\mathbb{R}$ difficilement généralisables à d'autres espaces.

La proposition 2.7 donne une borne supérieure à la complexité en temps de l'algorithme 3. Toutefois, dans bien des cas, la complexité est inférieure, grâce à l'élagage qui donne des ensembles τ_n^k de cardinal d'ordre inférieur à n , idéalement d'ordre $O(1)$ (qui, dans ce cas, entraîne une complexité en temps de l'algorithme 3 d'ordre $O(mN)$). Les simulations réalisées par Rigai (2010) montrent que, avec la perte quadratique $\gamma(y_i, \mu) = (y_i - \mu)^2$, à m fixé, l'algorithme 3 est plus rapide que l'algorithme 1. Rigai (2010) annonce une complexité en temps « moyenne » en $O(mN \log(N))$.

Récemment, Maidstone *et al.* (2014) ont proposé un algorithme de programmation dynamique différemment élagué, l'algorithme SNIP. Les conditions permettant de l'utiliser sont plus faibles que (2.48), cela rend donc cet algorithme particulièrement intéressant quand cette dernière condition n'est pas vérifiée. Toutefois, si (2.48) est vérifiée, l'algorithme SNIP élague toujours moins que l'algorithme de Rigai (2010). Dans les chapitres 3 et 4, nous avons pu nous ramener à un cas où (2.48) est vérifiée. Par conséquent, nous n'avons pas utilisé l'algorithme SNIP dans notre travail.

Chapter 3

A robust approach for estimating change-points in the mean of an AR(1) process

Contenu

3.1	Introduction	52
3.2	Robust estimation of the parameter ρ^*	54
3.3	Change-points and expectations estimation	55
3.4	Selecting the number of change-points	57
3.4.1	Consistent model selection criterion	57
3.4.2	Modified BIC criterion	58
3.5	Numerical experiments	59
3.5.1	Practical implementation	59
3.5.2	Simulation design	60
3.5.3	Results	61
3.5.4	Additional simulation studies	65
3.6	Conclusion	68
3.7	Proofs	68
3.7.1	Proof of Proposition 3.1	68
3.7.2	Hints for (3.7)	70
3.7.3	Proof of Proposition 3.2	71
3.7.4	Proof of Proposition 3.3	77
3.7.5	Proof of Proposition 3.4	79
3.7.6	Proof of Proposition 3.5	81
3.7.7	Proof of Proposition 3.6	82
3.7.8	Proof of Proposition 3.7	84

Dans ce chapitre est présenté un article, co-écrit avec Stéphane Robin, Émilie Lebarbier et Céline Lévy-Leduc. Cet article a été accepté par la revue *Bernoulli*. Des remarques additionnelles sur cet article, non proposées à publication, sont présentées dans la section 5.1.1.

Résumé

Le problème qui nous intéresse est celui de l'estimation de ruptures multiples dans l'espérance d'un processus AR(1) gaussien. La prise en compte de la structure de dépendance nous interdit l'utilisation de l'algorithme de programmation dynamique, qui est le seul algorithme permettant de calculer la segmentation optimale dans le cas indépendant. Nous proposons un estimateur robuste du paramètre de la fonction d'autocorrelation, consistant et asymptotiquement gaussien. Nous proposons alors de suivre l'approche inférentielle classique en injectant cet estimateur dans le critère utilisé pour l'estimation de la localisation des ruptures. Nous démontrons que ces estimateurs partagent les propriétés asymptotiques des estimateurs classiques dans le cadre indépendant. La même approche *plug-in* est alors utilisée pour approcher le critère BIC modifié et sélectionner le nombre de segments. Cette méthode est implémentée dans le package R `AR1seg` et est disponible sur le *Comprehensive R Archive Network* (CRAN). Ce package est utilisé dans la section simulations dans laquelle nous montrons que pour des séries de taille finie la prise en compte de la structure de dépendance améliore les performances statistiques des estimateurs des instants de rupture et du nombre de ruptures sélectionné.

Abstract

We consider the problem of multiple change-point estimation in the mean of a Gaussian AR(1) process. Taking into account the dependence structure does not allow us to use the dynamic programming algorithm, which is the only algorithm giving the optimal solution in the independent case. We propose a robust estimator of the autocorrelation parameter, which is consistent and satisfies a central limit theorem. Then, we propose to follow the classical inference approach, by plugging this estimator in the criterion used for change-points estimation. We show that the asymptotic properties of these estimators are the same as those of the classical estimators in the independent framework. The same plug-in approach is then used to approximate the modified BIC and choose the number of segments. This method is implemented in the R package `AR1seg` and is available from the Comprehensive R Archive Network (CRAN). This package is used in the simulation section in which we show that for finite sample sizes taking into account the dependence structure improves the statistical performance of the change-point estimators and of the selection criterion.

3.1 Introduction

Change-point detection problems arise in many fields, such as genomics (J. V. Braun and Müller, 1998; J. V. Braun, R. K. Braun, et al., 2000; Picard, Robin, et al., 2005), medical imaging (Lavielle, 2005), earth sciences (Williams, 2003; Gazeaux et al., 2013), econometrics (T. L. Lai et al., 2005; Lavielle, 1999) or climate (Mestre, 2000; Lu et al., 2010). In many of these problems, the observations can not be assumed to be independent. Indeed the autocovariance structure of the time series display more complex patterns and might be taken into account in change-point estimation.

An abundant literature exists about the statistical theory of change-point detection. Only speaking about Gaussian processes, various frameworks have been considered ranging from the independent case with changes in the mean (Basseville and Nikiforov, 1993), to more complex structural changes (Bai and Perron, 1998), dependent processes (Lavielle, 1999) or processes with changes in all parameters (Bardet, Kengne, and Wintenberger, 2012).

Lavielle (1999) and Lavielle and Moulines (2000) proved that, if the number of changes is known, the least-squares estimators of the change-point locations and of the parameters of each segment are consistent under very mild conditions on the autocovariance structure of the process with changes in the mean. A quasi-likelihood approach is also proved to provide consistent estimates for the model with changes in all parameters by Bardet, Kengne, and Wintenberger (2012). Many model selection criteria have also been proposed to estimate the number of changes, mostly in the independent case (see for example Yao, 1988; Lavielle, 2005; Lebarbier, 2005; Zhang and Siegmund, 2007).

Change-point detection also raises algorithmic issues as the determination of the optimal set of change-point locations is a discrete optimization problem. The dynamic programming algorithm introduced by Auger and Lawrence (1989) is the only way to recover this optimal segmentation. The computational complexity of this algorithms is quadratic relatively to the length of the series. Only this algorithm and some of its improvements (such as these proposed by Rigail, 2010; or Killick et al., 2012) provide exactly the optimal change-point location estimators.

However, the dynamic programming algorithm only applies when (i) the loss function (e.g. the negative log-likelihood) is additive with respect to the segments and when (ii) no parameter to be estimated is common to several segments. These requirements are met by the least-square criterion (which corresponds to the negative log-likelihood in the Gaussian homoscedastic independent model with changes in the mean) or by the model and criterion considered by Bardet, Kengne, and Wintenberger (2012). In other cases, iterative and stochastic procedures are needed (see Bai and Perron, 2003; or Li and Lund, 2012).

In this paper, we consider the segmentation of an AR(1) process with homogeneous autocorrelation coefficient ρ^* :

$$y_i = \mu_k^* + \eta_i, \quad t_{n,k}^* + 1 \leq i \leq t_{n,k+1}^*, \quad 0 \leq k \leq m^*, \quad 1 \leq i \leq n, \quad (3.1)$$

where $(\eta_i)_{i \in \mathbb{Z}}$ is a zero-mean stationary AR(1) Gaussian process defined as the solution of

$$\eta_i = \rho^* \eta_{i-1} + \epsilon_i, \quad (3.2)$$

where $|\rho^*| < 1$ and the ϵ_i 's are iid zero-mean Gaussian rv's with variance σ^{*2} . We further also assume that y_0 is a Gaussian random variable with mean μ_0^* and variance $\sigma^{*2}/(1 - \rho^{*2})$. Actually, most of the results we provide in this paper hold without the Gaussian assumption.

Note that this model is different from the ones considered by Davis, T. C. M. Lee, et al. (2006) and Bardet, Kengne, and Wintenberger (2012). Indeed, Davis, T. C. M. Lee, et al. (2006) considered the segmentation issue of a non-stationary time series which consists of blocks of different autoregressive processes where all the parameters of the autoregressive processes change from one segment to the other. Bardet, Kengne, and Wintenberger (2012) proposed a methodology for estimating the change-points of a non-stationary time series built from a general class of models having piecewise constant parameters. In this framework, all the parameters may change jointly at each change-point. This differs from our Model (3.1) where the parameters ρ^* and σ^* are not assumed to change from one segment to the other. The direct maximum-likelihood inference for such a process violates both requirements (i) and (ii). Indeed the log-likelihood is not additive with respect to the segments because of the dependence that exists between data from neighbor segments and the unknown coefficient ρ^* needs to be estimated jointly over all segments.

Our aim is to propose a methodology for estimating both the change-point locations $\mathbf{t}_n^* = (t_{n,k}^*)_{1 \leq k \leq m^*}$ and the means $\boldsymbol{\mu}^* = (\mu_k^*)_{0 \leq k \leq m^*}$, accounting for the existence of the autocorrelation ρ^* .

In the sequel, we shall use the following conventions: $t_{n,0}^* = 0, t_{n,m^*+1}^* = n$ and assume that there exists $\boldsymbol{\tau}^* = (\tau_k^*)_{0 \leq k \leq m+1}$ such that, for $0 \leq k \leq m+1$ $t_{n,k}^* = \lfloor n\tau_k^* \rfloor$, $\lfloor x \rfloor$ denoting the integer part of x . Consequently, $\tau_0^* = 0$ and $\tau_{m^*+1}^* = 1$.

If ρ^* was known, the series could be decorrelated and the dynamic programming algorithm then used for the segmentation of this decorrelated series. Here, ρ^* is unknown, but is estimated, and this estimator is then used to decorrelate the series.

To this aim, we borrow techniques from robust estimation (Ma and Genton, 2000). Briefly speaking, we consider the data observed at the change-point locations as outliers and propose an estimate of ρ^* that is robust to the presence of such outliers. We shall prove that the estimate we propose is consistent and satisfies a central limit theorem.

We shall prove that the resulting change-point estimators satisfy the same asymptotic properties as those proposed by Lavielle and Moulines (2000) and Bardet, Kengne, and Wintenberger (2012). Finally, we propose a model selection criterion inspired by the one proposed by Zhang and Siegmund (2007) and prove some asymptotic properties of this criterion.

This method is implemented in the R package `AR1seg` and is available from the Comprehensive R Archive Network (CRAN).

This paper is organized as follows. In Section 3.2, we propose a robust estimator for ρ^* and establish its asymptotic properties. In Section 3.3, we prove that the change-point

estimators defined in (3.11) are consistent in both the Gaussian and the non-Gaussian case. In Section 3.4, we provide a consistent model selection criterion in the non-Gaussian case and derive an approximation of a Gaussian criterion. In Section 3.5, we illustrate by a simulation study the performance of this approach for time series having a finite sample size.

3.2 Robust estimation of the parameter ρ^*

The aim of this section is to provide an estimator of ρ^* which can deal with the presence of change-points in the data. In the absence of change-points ($m^* = 0$ in (3.1)), a consistent estimator of ρ^* could be obtained by using the classical autocorrelation function estimator of $(y_i)_{0 \leq i \leq n}$ computed at lag 1. Since change-points can be seen as outliers in the AR(1) process, we shall propose a robust approach for estimating ρ^* . Ma and Genton (2000) propose a robust estimator of the autocorrelation function of a stationary time series based on the robust scale estimator proposed by Rousseeuw and Croux (1993). More precisely, the approach of Ma and Genton (2000) would result in the following estimate of ρ^* :

$$\widehat{\rho}_{\text{MG}} = \frac{Q_n^2(y^+) - Q_n^2(y^-)}{Q_n^2(y^+) + Q_n^2(y^-)},$$

where $y^+ = (y_{i+1} + y_i)_{0 \leq i \leq n-1}$, $y^- = (y_{i+1} - y_i)_{0 \leq i \leq n-1}$ and Q_n is the scale estimator of Rousseeuw and Croux (1993) which is such that $Q_n(x)$ is proportional to the first quartile of

$$\{|x_i - x_j|; 0 \leq i < j \leq n\}.$$

The asymptotic properties of this estimator are studied by Lévy-Leduc et al. (2011) for Gaussian stationary processes with either short-range or long-range dependence. However, as we shall see in the simulation section we can provide an estimator of ρ^* which is more robust to the presence of change-points than $\widehat{\rho}_{\text{MG}}$. The asymptotic properties of this novel robust estimator are given in Proposition 3.1.

Proposition 3.1. *Let $y_0, \dots, y_n$ be $(n+1)$ observations satisfying (3.1) and let*

$$\widetilde{\rho}_n = \frac{\left(\text{med}_{0 \leq i \leq n-2} |y_{i+2} - y_i| \right)^2}{\left(\text{med}_{0 \leq i \leq n-1} |y_{i+1} - y_i| \right)^2} - 1, \quad (3.3)$$

where $\text{med} x_i$ denotes the median. Then, $\widetilde{\rho}_n$ satisfies the following central limit theorem

$$\sqrt{n}(\widetilde{\rho}_n - \rho^*) \xrightarrow{d} \mathcal{N}(0, \widetilde{\sigma}^2), \text{ as } n \rightarrow \infty, \quad (3.4)$$

where

$$\widetilde{\sigma}^2 = \mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2)^2] + 2 \sum_{k \geq 1} \mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2) \Psi(\eta_k, \eta_{k+1}, \eta_{k+2})], \quad (3.5)$$

and the function Ψ is defined by

$$\Psi : (x_0, x_1, x_2) \mapsto -\frac{(1 + \rho^*)}{\Phi^{-1}(3/4)\varphi(\Phi^{-1}(3/4))} \left[\mathbb{1}_{\{|x_2 - x_0| \leq \sqrt{2\sigma^{*2}}\Phi^{-1}(3/4)\}} - \mathbb{1}_{\{|x_1 - x_0| \leq \sqrt{\frac{2\sigma^{*2}}{1+\rho^*}}\Phi^{-1}(3/4)\}} \right], \quad (3.6)$$

where Φ and φ denote the cumulative distribution function and the probability distribution function of a standard Gaussian rv, respectively.

The proof of Proposition 3.1 is given in Section 3.7.1.

Remark 3.1. Let us now compare the properties of $\tilde{\rho}_n$ with the properties of $\hat{\rho}_n(1)$ where $\hat{\rho}_n(\cdot)$ denotes the classical estimator of the autocorrelation function computed from $y_0, \dots, y_n$ defined in (3.1) with $m^* = 0$. By Brockwell and Davis (1991, Theorem 7.2.1 and Example 7.2.3), we get that

$$\sqrt{n}(\hat{\rho}_n(1) - \rho^*) \xrightarrow{d} \mathcal{N}(0, 1 - \rho^{*2}), \text{ as } n \rightarrow \infty.$$

From this result, we can see that $\tilde{\rho}_n$ converges to ρ^* at the same rate as $\hat{\rho}_n(1)$ except that our result still holds when $m \neq 0$.

Remark 3.2. Note that the asymptotic distribution given in (3.4) allows to define a test of $(H_0) : \rho^* = 0$ as the asymptotic variance $\tilde{\sigma}^2$ does not depend on any unknown parameter under H_0 .

Remark 3.3. Since the estimator (3.3) involves differences of the process (y_i) at different instants, it can only be used in the case of stable distributions as defined by Feller (1971). Among them, we can quote the Cauchy, Lévy and Gaussian distributions, where the Gaussian distribution is the only one to have a finite second order moment. We give some hints in Section 3.7.2 to explain why, in the case of the Cauchy distribution, taking $\tilde{\tilde{\rho}}_n$ defined as follows leads to an accurate estimator of ρ^* :

$$\tilde{\tilde{\rho}}_n = \begin{cases} -1 + \sqrt{1 + \tilde{\rho}_n}, & \text{if } \tilde{\rho}_n \geq 0, \\ -\sqrt{1 - \sqrt{1 + \tilde{\rho}_n}}, & \text{if } \tilde{\rho}_n < 0, \end{cases} \quad (3.7)$$

where $\tilde{\rho}_n$ is defined by (3.3). Some simulations are also provided in Section 3.5.4 to illustrate the finite sample size properties of this estimator.

3.3 Change-points and expectations estimation

In this section, the number of change-points m^* is assumed to be known. In the sequel, for notational simplicity, m^* will be denoted by m . Our goal is to estimate both the change-points and the means in Model (3.1). A first idea consists in using the following criterion which is based on a quasi-likelihood conditioned on y_0 and on the reparametrization $\delta_k = (1 - \rho)\mu_k$, and to minimize it with respect to ρ :

$$\sum_{k=0}^m \sum_{i=t_k+2}^{t_{k+1}} (y_i - \rho y_{i-1} - \delta_k)^2 + \sum_{k=1}^m \left\{ \left(y_{t_{k+1}} - \frac{\delta_k}{1 - \rho} \right) - \rho \left(y_{t_k} - \frac{\delta_{k-1}}{1 - \rho} \right) \right\}^2 + (y_1 - \rho y_0 - \delta_0)^2.$$

Due to the term that involves both δ_{k-1} and δ_k , this criterion cannot be efficiently minimized. Therefore, we propose to use an alternative criterion defined as follows:

$$SS_m(y, \rho, \boldsymbol{\delta}, \mathbf{t}) = \sum_{k=0}^m \sum_{i=t_k+1}^{t_{k+1}} (y_i - \rho y_{i-1} - \delta_k)^2. \quad (3.8)$$

Note that $SS_m(z, \rho, (1-\rho)\boldsymbol{\mu}, \mathbf{t})$ corresponds to $-n/2$ times the log-likelihood of the following model maximized with respect to σ :

$$z_i - \mu_k^* = \rho^* (z_{i-1} - \mu_k^*) + \epsilon_i, \quad t_{n,k}^* + 1 \leq i \leq t_{n,k+1}^*, \quad 0 \leq k \leq m, \quad 1 \leq i \leq n, \quad (3.9)$$

and where z_0 is a Gaussian rv with mean μ_0^* and variance $\sigma^{*2}/(1-\rho^{*2})$. In this model, which is a subset of a model belonging to the class considered in Bardet, Kengne, and Wintenberger (2012), the expectation changes are not abrupt anymore as in Model (3.1).

The parameter ρ , involved in each term of (3.8), is still a problem in order to minimize SS_m with respect to $\rho, \boldsymbol{\delta}$ and $\mathbf{t}$. This minimization problem is a complex discrete and global optimization problem. Dynamic Programming (Auger and Lawrence, 1989) cannot be used in this case. Only iterative methods are suitable to this minimization problem, without any guarantee to converge to the global minimum.

However, if ρ is replaced by an estimator $\bar{\rho}_n$, $SS_m(y, \bar{\rho}_n, \boldsymbol{\delta}, \mathbf{t})$ can be minimized with respect to $\boldsymbol{\delta}$ and $\mathbf{t}$ by Dynamic Programming. Proposition 3.3 gives asymptotic results for the estimators resulting from this method.

Proposition 3.2. *Let $z = (z_0, \dots, z_n)$ be a finite sequence of real-valued rv's satisfying (3.9) and $(\bar{\rho}_n)$ a sequence of real-valued rv's. Let $\widehat{\boldsymbol{\delta}}_n(z, \bar{\rho}_n)$ and $\widehat{\mathbf{t}}_n(z, \bar{\rho}_n)$ be defined by*

$$(\widehat{\boldsymbol{\delta}}_n(z, \bar{\rho}_n), \widehat{\mathbf{t}}_n(z, \bar{\rho}_n)) = \arg \min_{(\boldsymbol{\delta}, \mathbf{t}) \in \mathbb{R}^{m+1} \times \mathcal{A}_{n,m}} SS_m(z, \bar{\rho}_n, \boldsymbol{\delta}, \mathbf{t}), \quad (3.10)$$

$$\widehat{\boldsymbol{\tau}}_n(z, \bar{\rho}_n) = \frac{1}{n} \widehat{\mathbf{t}}_n(z, \bar{\rho}_n), \quad (3.11)$$

where

$$\mathcal{A}_{n,m} = \{(t_0, \dots, t_{m+1}); t_0 = 0 < \dots < t_{m+1} = n, \forall k = 1, \dots, m+1, t_k - t_{k-1} \geq \Delta_n\} \quad (3.12)$$

and where (Δ_n) is a real sequence such that $n^{-1}\Delta_n \xrightarrow{n \rightarrow \infty} 0$ and $n^{-\alpha}\Delta_n \xrightarrow{n \rightarrow \infty} +\infty$ with $\alpha > 0$. Assume that

$$(\bar{\rho}_n - \rho^*) = O_P(n^{-1/2}), \quad (3.13)$$

as n tends to infinity. Then,

$$\|\widehat{\boldsymbol{\tau}}_n(z, \bar{\rho}_n) - \boldsymbol{\tau}^*\| = O_P(n^{-1}), \quad \|\widehat{\boldsymbol{\delta}}_n(z, \bar{\rho}_n) - \boldsymbol{\delta}^*\| = O_P(n^{-1/2}),$$

where $\|\cdot\|$ is the Euclidian norm and

$$\boldsymbol{\delta}^* = (\delta_0^*, \dots, \delta_{m+1}^*) = (1 - \rho^*) (\mu_0^*, \dots, \mu_{m+1}^*).$$

The results still hold if the ϵ_i 's are only assumed to be centered and to have a finite second order moment.

Proposition 3.3. *The results of Proposition 3.2 still hold under the same assumptions when z is replaced with y satisfying (3.1).*

The results still hold if the ϵ_i 's are only assumed to be centered and to have a finite second order moment.

The proofs of Propositions 3.2 and 3.3 are given in Sections 3.7.3 and 3.7.4, respectively. Note that the estimators defined in these propositions have the same asymptotic properties as those of the estimators proposed by Lavielle and Moulines (2000). In the Gaussian framework, the estimator $\tilde{\rho}_n$ defined in Section 3.2 satisfies the same properties as $\bar{\rho}_n$ and can thus be used in the criterion SS_m for providing consistent estimators of the change-points and of the means.

3.4 Selecting the number of change-points

We now consider the selection of the number of change-points. We first propose a penalized contrast criterion, which we prove to be consistent in the non-Gaussian case. The penalty has a general form, which needs to be specified for a practical use. Therefore, we also derive an adaptation of the modified BIC criterion proposed by Zhang and Siegmund (2007) in the Gaussian context. This criterion does not rely on any tuning parameter and has been shown to be efficient in practical cases (see Picard, Lebarbier, et al., 2011).

3.4.1 Consistent model selection criterion

We propose to select the number of change-points m as follows

$$\hat{m} = \arg \min_{0 \leq m \leq m_{\max}} \frac{1}{n} SS_m(z, \bar{\rho}_n) + \beta_n m \quad (3.14)$$

where $m_{\max} \geq m^*$, $(\beta_n)_{n \geq 1}$ is a sequence of positive real numbers, $\bar{\rho}_n$ satisfies the assumptions of Proposition 3.2 and

$$SS_m(z, \rho) = \min_{\delta, \mathbf{t} \in \mathcal{A}_{n,m}} SS_m(z, \rho, \delta, \mathbf{t}), \quad (3.15)$$

$\mathcal{A}_{n,m}$ being defined in (3.12).

Proposition 3.4. *Under the assumptions of Proposition 3.2, and if*

$$\beta_n \xrightarrow{n \rightarrow \infty} 0, \quad n^{1/2} \beta_n \xrightarrow{n \rightarrow \infty} +\infty, \quad \Delta_n \beta_n \xrightarrow{n \rightarrow \infty} +\infty,$$

where Δ_n is defined in Proposition 3.2, $\hat{m}$ defined by (3.14) converges in probability to m^ . The result still holds if the ϵ_i 's are only assumed to be independent, centered and to have a finite second order moment.*

Proposition 3.5. *The result of Proposition 3.4 still holds under the same assumptions when z is replaced by y satisfying (3.1).*

The result still holds if the ϵ_i 's are only assumed to be independent, centered and to have a finite second order moment.

The proofs of Propositions 3.4 and 3.5 are given in Sections 3.7.5 and 3.7.6, respectively.

Remark 3.4. If $\beta_n = n^{-\beta}$, the assumptions of Propositions 3.4 and 3.5 are fulfilled if and only if $0 < \beta < \min(\alpha, 1/2)$, where α is defined in Proposition 3.2. α stands for the usual bound for the control of the minimal segment length (see Lavielle and Moulines, 2000). The $1/2$ bound is the price to pay for the estimation of ρ^* .

3.4.2 Modified BIC criterion

Zhang and Siegmund (2007)¹ proposed a modified Bayesian information criterion (mBIC) to select the number m of change-points in the particular case of segmentation of an independent Gaussian process x . This criterion is defined in a Bayesian context in which a non informative prior is set for the number of segments m . mBIC is derived from an $O_P(1)$ approximation of the Bayes factor between models with m and 0 change-points, respectively. The mBIC selection procedure consists in choosing the number of change-points as:

$$\widehat{m} = \arg \max_m C_m(x, 0) \quad (3.16)$$

where the criterion $C_m(y, \rho)$ is defined for a process y as

$$C_m(y, \rho) = -\frac{n-m+1}{2} \log SS_m(y, \rho) + \log \Gamma\left(\frac{n-m+1}{2}\right) - \frac{1}{2} \sum_{k=0}^m \log n_k(\widehat{\mathbf{t}}(y, \rho)) - m \log n,$$

where Γ is the usual Gamma function. In the latter equation

$$n_k(\widehat{\mathbf{t}}(y, \rho)) = \widehat{t}_{k+1}(y, \rho) - \widehat{t}_k(y, \rho), \quad (3.17)$$

where $\widehat{\mathbf{t}}(y, \rho) = (\widehat{t}_1(y, \rho), \dots, \widehat{t}_m(y, \rho))$ is defined as $\widehat{\mathbf{t}}(y, \rho) = \arg \min_{\mathbf{t} \in \mathcal{A}_{n,m}} \min_{\delta} SS_m(y, \rho, \delta, \mathbf{t})$.

Note that, in Model (3.9), the criterion could be directly applied to the decorrelated series $v^* = (v_i^*)_{1 \leq i \leq n} = (y_i - \rho^* y_{i-1})_{1 \leq i \leq n}$ since

$$C_m(y, \rho^*) = C_m(v^*, 0).$$

We propose to use the same selection criterion, replacing ρ^* by some relevant estimator $\bar{\rho}_n$. The following two propositions show that this plug-in approach result in the same asymptotic properties under both Model (3.9) and (3.1).

Proposition 3.6. *For any positive m , for a process z satisfying (3.9) and under the assumptions of Proposition 3.2, we have*

$$C_m(z, \bar{\rho}_n) = C_m(z, \rho^*) + O_P(1), \text{ as } n \rightarrow \infty.$$

Proposition 3.7. *For any positive m , for a process y satisfying (3.1) and under the assumptions of Proposition 3.3, we have*

$$C_m(y, \bar{\rho}_n) = C_m(y, \rho^*) + O_P(1), \text{ as } n \rightarrow \infty.$$

¹See Theorem 2.11 of this document.

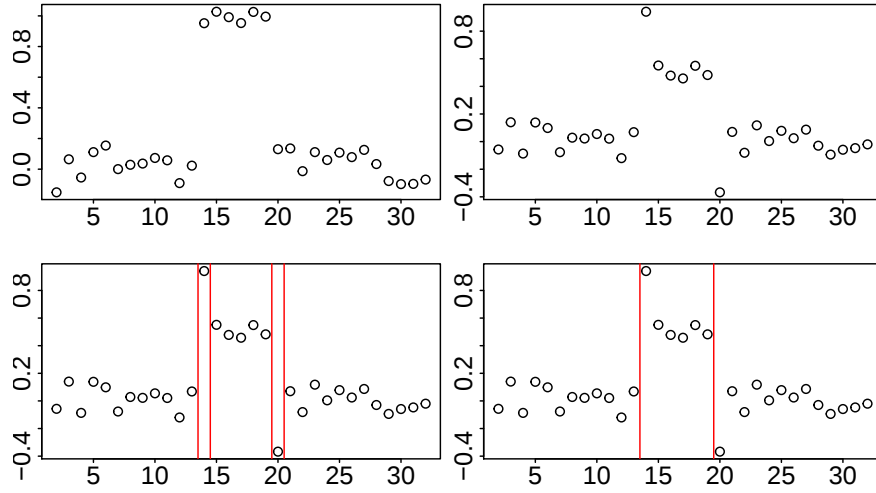


Figure 3.1 – Top left: a series around two changes, with $\rho^* = 0.5$. Top right: the decorrelated series in the same region. Bottom left: before post-processing, two pairs of adjacent change-points are found. Bottom right: post-processing removes the last change-point of each pair of adjacent ones.

The proofs of Propositions 3.6 and 3.7 are given in Sections 3.7.7 and 3.7.8 respectively.

In practice, we propose to take $\bar{\rho}_n = \tilde{\rho}_n$ which satisfies the condition of Proposition 3.7 to estimate the number of segments by

$$\begin{aligned} \hat{m} = \arg \max_m & \left[-\left(\frac{n-m+1}{2} \right) \log SS_m(y, \tilde{\rho}_n) + \log \Gamma \left(\frac{n-m+1}{2} \right) \right. \\ & \left. - \frac{1}{2} \sum_{k=0}^m \log n_k(\hat{\mathbf{t}}(y, \tilde{\rho}_n)) - m \log n \right], \end{aligned} \quad (3.18)$$

where $SS_m(\cdot, \cdot)$ and $n_k(\cdot, \cdot)$ are defined in (3.15) and (3.17), respectively.

Remark 3.5. Since the definition of the original mBIC criterion is intrinsically related to normality, we did not study precisely the quality of our approximation without the normality assumption.

3.5 Numerical experiments

3.5.1 Practical implementation

Our decorrelation procedure introduces spurious change-points in the series, at distance 1 of the true change-points (see Figure 3.1, top). Since these artefacts may affect our procedure, we propose a post-processing to the estimated change-points $\hat{\mathbf{t}}_n$, which consists in removing segments of length 1:

$$PP(\hat{\mathbf{t}}_n) = \{\hat{t}_{n,k} \in \hat{\mathbf{t}}_n\} \setminus \{\hat{t}_{n,i} \text{ such that } \hat{t}_{n,i} = \hat{t}_{n,i-1} + 1 \text{ and } \hat{t}_{n,i+1} \neq \hat{t}_{n,i} + 1\}.$$

This post-processing results in a smaller number of change-points. Figure 3.1 summarizes the whole processing.

In practice, it may also be useful to have some guidance on how to check that the assumptions underpinning our approach are satisfied for a given data set. A possible approach is to subtract the estimated piecewise constant function from the original series. If the model is the expected one, this new series should be a realization of an AR(1) Gaussian process. Hence, the residuals built by decorrelation of this series should be Gaussian and independent. One way to check this is to perform a gaussianity test and a Portmanteau test on this series of residuals.

3.5.2 Simulation design

To assess the performance of the proposed method, we used a simulation design inspired from the one conceived by W. R. Lai et al. (2005). We considered series of length $n \in \{100, 200, 400, 800, 1600\}$ with autocorrelation at lag 1, denoted by ρ^* , ranging from $-.9$ to $.9$ (by steps of $.1$) and residual standard deviation σ^* between $.1$ and $.6$ (by steps of $.1$). All series were affected by $m^* = 6$ change-points located at fractions $1/6 \pm 1/36, 3/6 \pm 2/36, 5/6 \pm 3/36$ of their length. Each combination was replicated $S = 100$ times. The mean within each segment alternates between 0 and 1, starting with $\mu_1 = 0$.

Estimation of ρ^* . For each generated series, two different estimates $\bar{\rho}_n$ of ρ^* were computed: the original estimate $\bar{\rho}_n = \hat{\rho}_{\text{MG}}$ proposed by Ma and Genton (2000) and our revised version $\bar{\rho}_n = \tilde{\rho}_n$. We carried the same study on series with no change-point (centered series).

Estimation of the segmentation parameters. For each generated series, we estimated the change-point locations $\hat{\tau}_n(y, \bar{\rho}_n)$ using Proposition 3.2 for each m from 1 to $m_{\max} = 75$ and with different choices of $\bar{\rho}_n$: $\tilde{\rho}_n$ (our estimator), ρ^* (the true value) and zero (which does not take into account for the autocorrelation). For each choice of $\bar{\rho}_n$, we then selected the number of change-points $\hat{m}$ using (3.18). Actually, the last choice $\bar{\rho}_n = 0$ corresponds to the classical least-squares framework. In addition, we shall also use the post-processing described in Section 3.5.1 for the cases where $\bar{\rho}_n = \tilde{\rho}_n$ and ρ^* .

To study the quality of the proposed model selection criterion, we computed the distribution of $\hat{m}$ for each estimate $\bar{\rho}_n \in \{\tilde{\rho}_n, \rho^*, 0\}$ with post-processing or not for the first two estimates of ρ^* .

In order to assess the performance of the estimation of the change-point locations, we computed the Hausdorff distance defined in the segmentation framework as follows, see Boysen et al. (2009) and Harchaoui and Lévy-Leduc (2010):

$$d(\tau^*, \hat{\tau}_n(y, \bar{\rho}_n)) = \max(d_1(\tau^*, \hat{\tau}_n(y, \bar{\rho}_n)), d_2(\tau^*, \hat{\tau}_n(y, \bar{\rho}_n))), \quad (3.19)$$

where

$$d_1(a, b) = \sup_{b \in \mathbf{b}} \inf_{a \in \mathbf{a}} |a - b|, \quad (3.20)$$

$$\text{and} \quad d_2(a, b) = d_1(b, a). \quad (3.21)$$

d_1 close to zero means that an estimated change-point is likely to be close to a true change-point. A small value of d_2 means that a true change-point is likely to be close to each estimated change-point. A perfect segmentation results in both null d_1 and d_2 . Over-segmentation results in a small d_1 and a large d_2 . Under-segmentation results in a large d_1 and a small d_2 , provided that the estimated change-points are correctly located.

3.5.3 Results

Estimation of ρ^* . In Figure 3.2, we compare the performance of our robust estimator of ρ^* : $\tilde{\rho}_n$ with the ones of the estimator $\hat{\rho}_{MG}$ in the case where there are no change-points in the observations. More precisely, in this case, the observations y are generated under the Model (3.1) with $\mu_k^* = 0$, for all k . We observe that the estimator proposed by Ma and Genton (2000) performs better than our robust estimator. However, it is not the case anymore in the presence of change-points in the data as we can see in Figure 3.3. In the latter case, our robust estimator $\tilde{\rho}_n$ outperforms the estimator $\hat{\rho}_{MG}$ for almost all values of ρ^* .

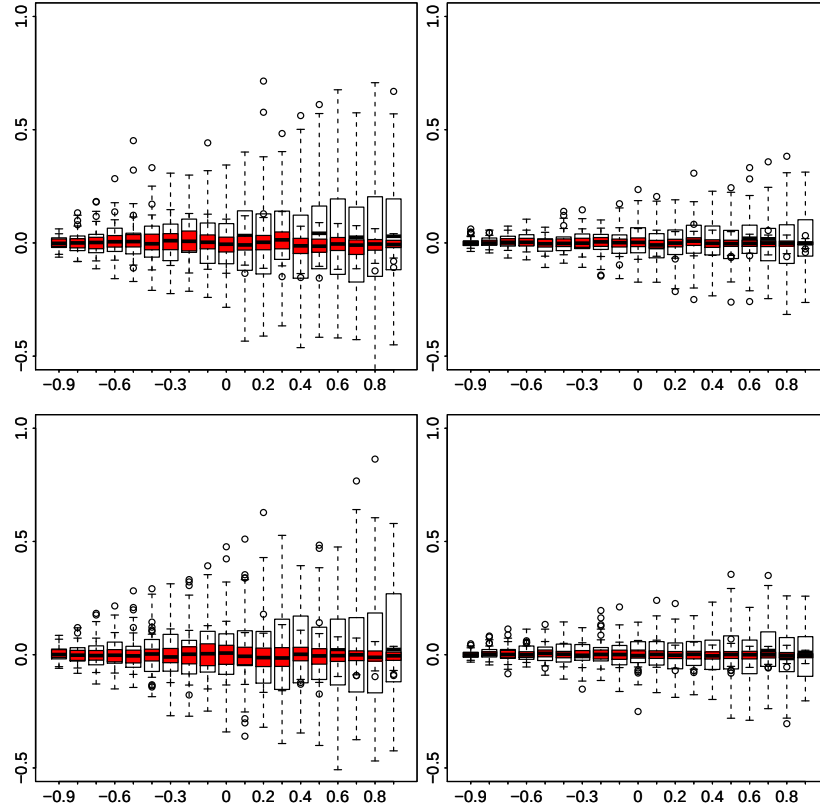


Figure 3.2 – Boxplots of $\hat{\rho}_{MG} - \rho^*$ in red and $\tilde{\rho}_n - \rho^*$ in black for different values of ρ^* in the case where there are no change-points in the data with $n = 400$ (plots on the left), $n = 1600$ (plots on the right), $\sigma^* = 0.2$ (top) and $\sigma^* = 0.6$ (bottom).

Model selection. In Figures 3.4 and 3.5, we compare the estimated number of change-points $\hat{m}$ in two different configurations of signal-to-noise ratio ($\sigma^* = 0.1$ and $\sigma^* = 0.5$) and

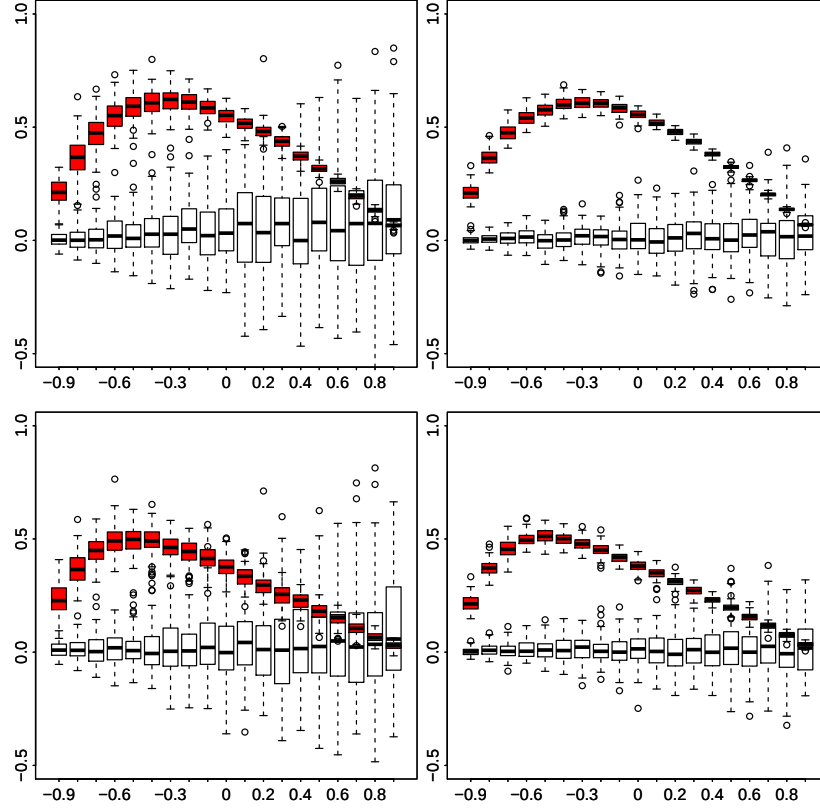


Figure 3.3 – Boxplots of $\widehat{\rho}_{MG} - \rho^*$ in red and $\widetilde{\rho}_n - \rho^*$ in black for different values of ρ^* in the case where there are change-points in the data with $n = 400$ (plots on the left), $n = 1600$ (plots on the right), $\sigma^* = 0.2$ (top) and $\sigma^* = 0.6$ (bottom).

with three different values of ρ^* ($\rho^* = 0.3, 0.6$ and 0.8). In these figures, the notation LS, Robust and Oracle correspond to the cases where $\bar{\rho}_n = 0$, $\bar{\rho}_n = \widetilde{\rho}_n$ and $\bar{\rho}_n = \rho^*$, respectively. Moreover, we use the notation -P when the post-processing described in Section 3.5.1 is used. In the situations where σ^* and ρ^* are small, all the methods provide an accurate estimation of the number of change-points. In the other cases, LS tends to strongly overestimate the number of change-points. Robust and Oracle tend to select twice the true number of change-points due to the artifactual presence of change-points in the decorrelated series as explained in Section 3.5.1. This is corrected by the post-processing and Robust-P provides the correct number of change-points in most of the considered configurations. Moreover, we also observe that the performance of Robust and Robust-P are similar to these of Oracle and Oracle-P: the robust decorrelation procedure we propose performs as well as if ρ^* was known for $n = 1600$. It has to be noted that the post-processing would not improve the performance on LS so we did not consider it.

Change-point locations. In Figures 3.6 and 3.8 are displayed the boxplots of the two parts d_1 and d_2 of the Hausdorff distance defined in (3.20) and (3.21), respectively for different values of ρ^* when $\sigma^* = 0.5$. d_2 is displayed in Figure 3.7 for $\sigma^* = 0.1$; for this value of σ^* , d_1 was found null for all methods and all values of ρ^* .

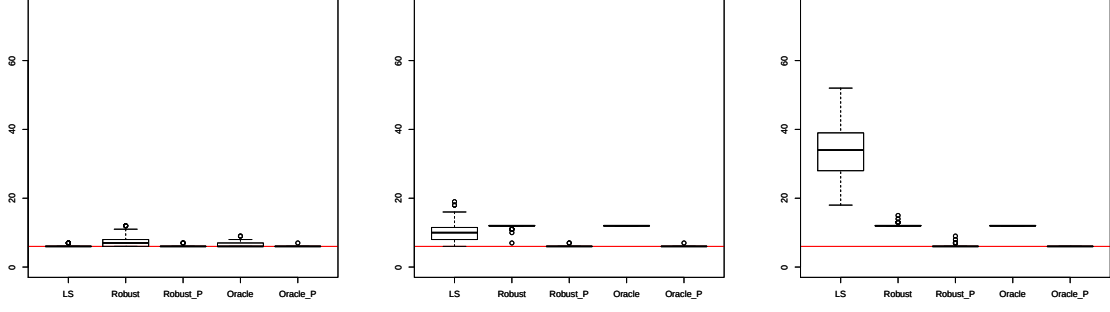


Figure 3.4 – Boxplots for the estimated number of change-points for $n = 1600$ when $\bar{\rho}_n = 0$ (LS), $\bar{\rho}_n = \tilde{\rho}_n$ (Robust and Robust-P with post-processing) and $\bar{\rho}_n = \rho^*$ (Oracle and Oracle-P with post-processing) with $\sigma^* = 0.1$ and $\rho^* = 0.3$ (left), $\rho^* = 0.6$ (middle) and $\rho^* = 0.8$ (right). The true number of change-points is equal to 6 (red horizontal line).

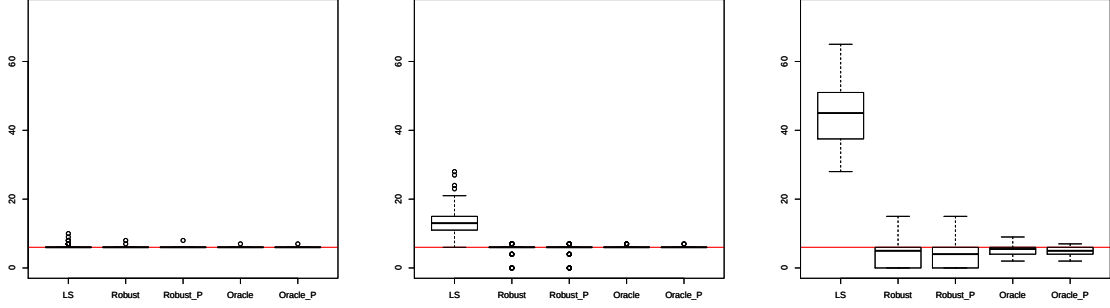


Figure 3.5 – Boxplots for the estimated number of change-points for $n = 1600$ when $\bar{\rho}_n = 0$ (LS), $\bar{\rho}_n = \tilde{\rho}_n$ (Robust and Robust-P with post-processing) and $\bar{\rho}_n = \rho^*$ (Oracle and Oracle-P with post-processing) with $\sigma^* = 0.5$ and $\rho^* = 0.3$ (left), $\rho^* = 0.3$ (middle) and $\rho^* = 0.8$ (right). The true number of change-points is equal to 6 (red horizontal line).

When the noise is small ($\sigma^* = 0.1$), the robust procedure we propose performs well for the whole range of correlation. On the contrary, the performance of LS are deprecated when the correlation increases, whereas these of LS^{*} still provide accurate change-point locations. This shows that the least-square approach only fails because it turns to overestimate the number of change-points. This is all the more true for LS when the variance of the noise is large ($\sigma^* = 0.5$). When the problem gets difficult (both σ^* and ρ^* large), our robust procedure tends to underestimate the number of change-points (which was expected) and the estimated change-points are close to true ones.

An other way to illustrate the performance of the estimation of the change-point locations is the histograms of these estimates. We provide these plots only for LS, Robust-P and Oracle-P, because Post-processing does not change significantly LS estimates, and, furthermore, Robust (respectively Oracle) method's histograms with or without Post-Processing are very similar, see Figures 3.9 and 3.10. These figures illustrate that in case of over-estimation of the number of changes by LS method, the additional change-points seem to be uniformly

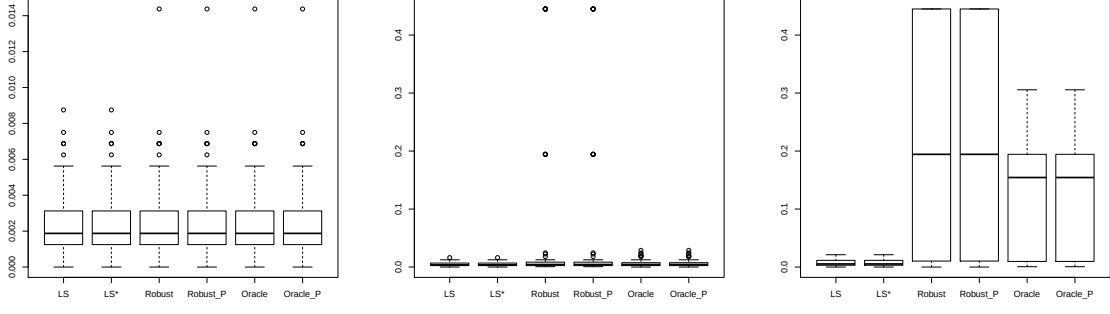


Figure 3.6 – Boxplots for the first part of the Hausdorff distance (d_1) for $n = 1600$ when $\bar{\rho}_n = 0$ (LS and LS* when the true number of change-points is known), $\bar{\rho}_n = \tilde{\rho}_n$ (Robust and Robust-P with post-processing) and $\bar{\rho}_n = \rho^*$ (Oracle and Oracle-P with post-processing) with $\sigma^* = 0.5$ and $\rho^* = 0.3$ (left), $\rho^* = 0.6$ (middle) and $\rho^* = 0.8$ (right).

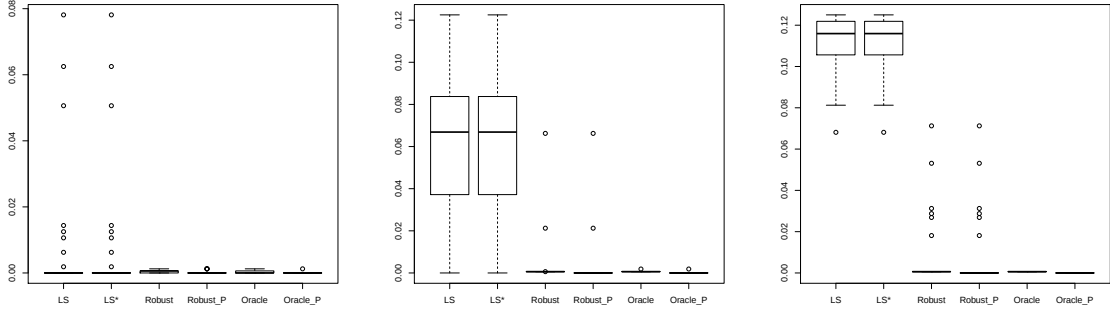


Figure 3.7 – Boxplots for the second part of the Hausdorff distance (d_2) for $n = 1600$ when $\bar{\rho}_n = 0$ (LS and LS* when the true number of change-points is known), $\bar{\rho}_n = \tilde{\rho}_n$ (Robust and Robust-P with post-processing) and $\bar{\rho}_n = \rho^*$ (Oracle and Oracle-P with post-processing) with $\sigma^* = 0.1$ and $\rho^* = 0.3$ (left), $\rho^* = 0.6$ (middle) and $\rho^* = 0.8$ (right).

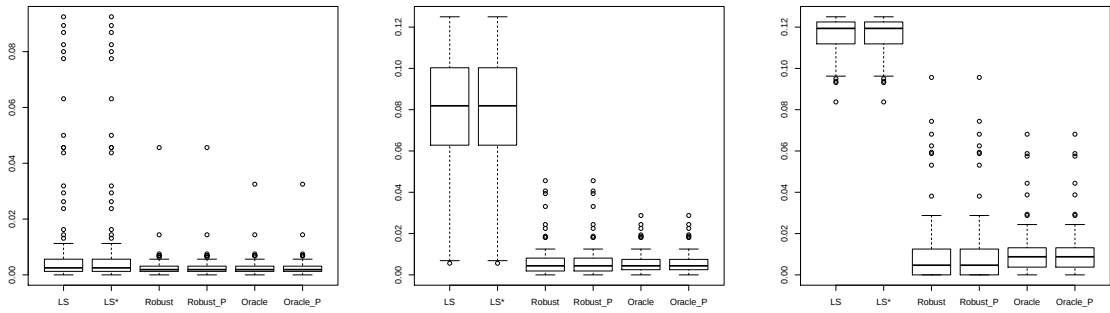


Figure 3.8 – Boxplots for the second part of the Hausdorff distance (d_2) when $\bar{\rho}_n = 0$ (LS and LS* when the true number of change-points is known), $\bar{\rho}_n = \tilde{\rho}_n$ (Robust and Robust-P with post-processing) and $\bar{\rho}_n = \rho^*$ (Oracle and Oracle-P with post-processing) with $\sigma^* = 0.5$ and $\rho^* = 0.3$ (left), $\rho^* = 0.6$ (middle) and $\rho^* = 0.8$ (right).

distributed.

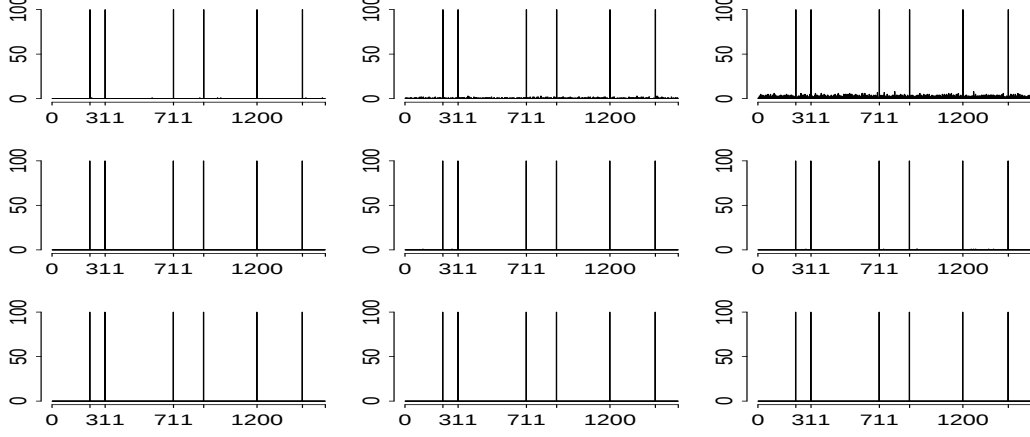


Figure 3.9 – Frequencies of each possible change-point estimator, with $\sigma^* = 0.1$ and $n = 1600$. Tick-marks on bottom-side axis represent the true change-point locations. $\bar{\rho}_n = 0$ (LS, top line), $\bar{\rho}_n = \tilde{\rho}_n$ (Robust-P, middle line) and $\bar{\rho}_n = \rho^*$ (Oracle-P, bottom line) with $\rho^* = 0.3$ (left), $\rho^* = 0.6$ (middle) and $\rho^* = 0.8$ (right).

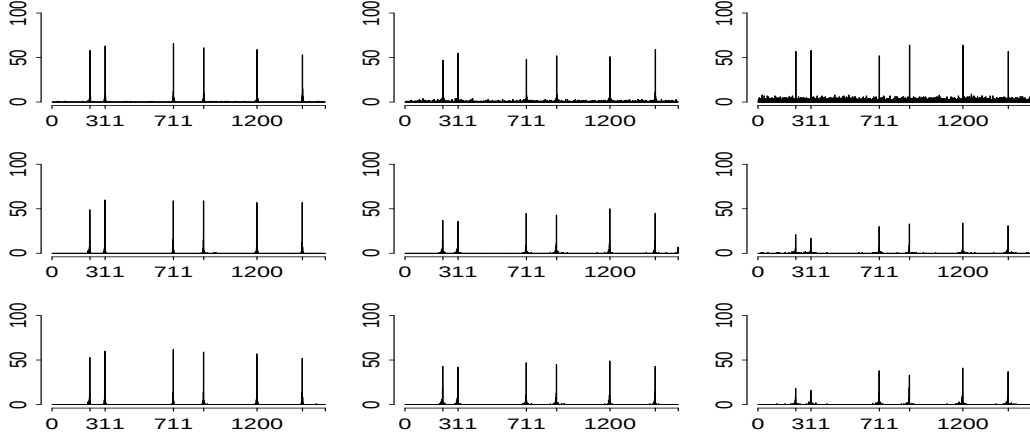


Figure 3.10 – Frequencies of each possible change-point estimator, with $\sigma^* = 0.5$ and $n = 1600$. Tick-marks on bottom-side axis represent the true change-point locations. $\bar{\rho}_n = 0$ (LS, top line), $\bar{\rho}_n = \tilde{\rho}_n$ (Robust-P, middle line) and $\bar{\rho}_n = \rho^*$ (Oracle-P, bottom line) with $\rho^* = 0.3$ (left), $\rho^* = 0.6$ (middle) and $\rho^* = 0.8$ (right).

3.5.4 Additional simulation studies

Comparison with Bardet, Kengne, and Wintenberger (2012)

The quasi-maximum likelihood method proposed by Bardet, Kengne, and Wintenberger (2012), when applied to a Gaussian AR(1) process with changes in the mean $(y_0, \dots, y_n)$, consists in the minimization with respect to $\boldsymbol{\rho} = (\rho_0, \dots, \rho_m)$, $\boldsymbol{\sigma} = (\sigma_0, \dots, \sigma_m)$, $\boldsymbol{\delta} = (\delta_0, \dots, \delta_m)$ and $\boldsymbol{t} = (t_0, \dots, t_m)$ of the following function:

$$(\boldsymbol{\rho}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \boldsymbol{t}) \mapsto \sum_{k=0}^m \left\{ (t_{k+1} - t_k) \log(\sigma_k^2) + \frac{1}{\sigma_k^2} \sum_{i=t_k+1}^{t_{k+1}} (y_i - \rho_k y_{i-1} - \delta_k)^2 \right\}. \quad (3.22)$$

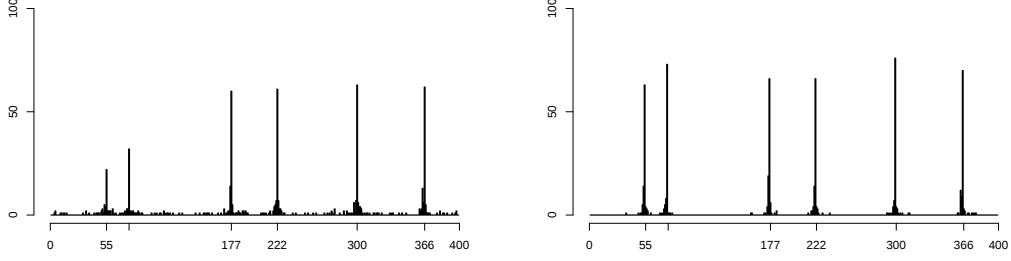


Figure 3.11 – Frequencies of each possible change-point location estimate. Tick-marks on bottom-side axis represent the true change-point locations. Left: Estimation by the minimization of (3.22). Right: Our method.

Indeed, in the class of models considered in Bardet, Kengne, and Wintenberger (2012), changes in all the parameters are possible at each change-point. Using this method to estimate the change-point locations for data satisfying Model (3.1) or (3.9) boils down to ignore the stationarity of $(\eta_i)_{i \geq 0}$ as defined in (3.2). It can lead to a poor estimation of change-point locations, especially when there are many changes close to each other. To illustrate this fact, we compared our estimator of change-point locations to the estimates given by the minimization of (3.22). We generated 100 series of length 400, under Model (3.1), with $\rho^* = 0.3$ and $\sigma^* = 0.4$. The number of change-points, their locations and the means within segments are the same as in Section 3.5.2. The number of changes is assumed to be known and we did not post-process the estimates. Simulations show that using the method of Bardet, Kengne, and Wintenberger (2012) in this case can lead to a poor estimation of close change-points, while our method is less affected by the length of segments (see Figure 3.11). For example, the boundaries of the smallest segment are recovered in less than half of the simulations when minimizing (3.22).

Robustness to model mis-specification

In this section, we study the behaviour of our proposed robust procedure (Robust-P) when the signal is corrupted by an AR(2) Gaussian process, e.g. in Model 3.1, η_i is a zero-mean stationary process such that

$$\eta_i = \phi_1^* \eta_{i-1} + \phi_2^* \eta_{i-2} + \varepsilon_i,$$

where $|\phi_2| < 1$, $\phi_1 + \phi_2 < 1$ and $\phi_2 - \phi_1 < 1$. We considered series of fixed length $n = 1600$, a residual standard deviation $\sigma^* = 0.1$, $\phi_1^* = 0.3$ and ϕ_2^* in $\{-0.9, -0.8, -0.7, \dots, 0.5, 0.6\}$. We used the same segmentation design as in Subsection 3.5.1. Each combination was replicated 100 times. All the results are displayed in Figure 3.12.

The procedure performs well when ϕ_2^* belongs to the interval $[-0.5, 0.2]$ as expected (similar to the case of AR(1)): the estimated segmentation is close to the true one. When $\phi_2^* > 0.2$, it tends to over-estimate the number of change-points. The true change-points are detected (d_1 is close to zero, e.g. the decorrelation procedure with the obtained negative estimation

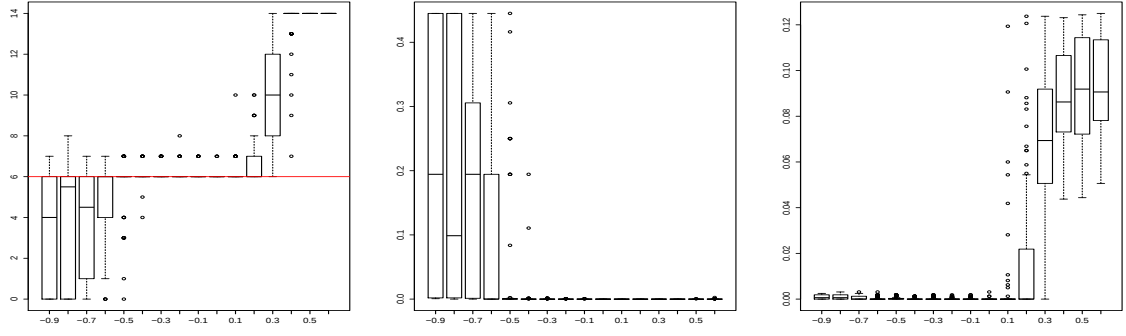


Figure 3.12 – Left: Boxplots for the estimated number of change-points. Center and right: Boxplots for the first part of the Hausdorff distance (d_1) and for the second part of the Hausdorff distance (d_2) with $n = 1600$, $\sigma^* = 0.1$ and $\phi_1^* = 0.3$ with respect to different values of ϕ_2^* .

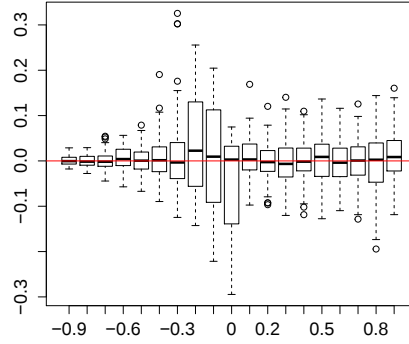


Figure 3.13 – Boxplots of $\tilde{\rho}_n - \rho^*$ for different values of ρ^* when $n = 1600$ and $\sigma^* = 0.1$.

of ρ^* leads to an increasing in the mean differences) but false change-points are added (large d_2). When $\phi_2^* < -0.5$, under-segmentation is observed: the decorrelation procedure with a large estimated value of ρ^* leads to a difficult segmentation problem.

Estimator of ρ^* in the case of the Cauchy distribution

In Section 3.2, an analogous estimator of ρ^* in the case of Cauchy distributed observations is proposed. We follow the simulation design described in Subsection 3.5.2, where the Gaussian rv's are replaced by Cauchy rv's. More precisely, the expectation parameters are replaced by the location parameters of the Cauchy distribution and σ^* is replaced by the scale parameter of the Cauchy distribution. We can see from Figure 3.13 that $\tilde{\rho}_n$ is an accurate estimator of ρ^* except when ρ^* is close to zero. When this estimator of ρ^* is used in our change-point estimation method, it leads to poor estimations of the change-points since the Cauchy distribution does not have finite second order moment (simulations not shown).

3.6 Conclusion

In this paper, we propose a novel approach for estimating multiple change-points in the mean of a Gaussian AR(1) process. Our approach is based on two main stages. The first one consists in building a robust estimator of the autocorrelation parameter which is used for whitening the original series. In the second stage, we apply the inference approach commonly used to estimate change-points in the mean of independent random variables. In the course of this study, we have shown that our approach, which is implemented in the R package **AR1seg**, is a very efficient technique both on a theoretical and practical point of view. More precisely, it has two main features which make it very attractive. Firstly, the estimators that we propose have the same asymptotic properties as the classical estimators in the independent framework which means that the performances of our estimators are not affected by the dependence assumption. Secondly, from a practical point of view, **AR1seg** is computationally efficient and exhibits better performance on finite sample size data than existing approaches which do not take into account the dependence structure of the observations.

3.7 Proofs

3.7.1 Proof of Proposition 3.1

Let F_1 and F_2 denote the cumulative distribution functions (cdf) of $(|y_{i+1} - y_i|)$ for $i \neq t_{n,1}^*, \dots, t_{n,m^*}^*$ and $(|y_{i+2} - y_i|)$ for $i \neq t_{n,1}^* - 1, \dots, t_{n,m^*}^* - 1$, respectively. By (3.1), $(y_i - \mathbb{E}(y_i))_{0 \leq i \leq n}$ are $(n+1)$ observations of a AR(1) stationary Gaussian process thus for any $i \neq t_{n,1}^*, \dots, t_{n,m^*}^*$, $(y_{i+1} - y_i)$ and for any $i \neq t_{n,1}^* - 1, \dots, t_{n,m^*}^* - 1$, $(y_{i+2} - y_i)$ are zero-mean Gaussian rv's with variances equal to $2\sigma^{*2}/(1+\rho^*)$ and $2\sigma^{*2}$, respectively. Hence, for all t in $\mathbb{R}$,

$$F_1 : t \mapsto 2\Phi\left(t\sqrt{\frac{1+\rho^*}{2\sigma^{*2}}}\right) - 1 \text{ and } F_2 : t \mapsto 2\Phi\left(t\sqrt{\frac{1}{2\sigma^{*2}}}\right) - 1, \quad (3.23)$$

where Φ denotes the cumulative distribution function of a standard Gaussian rv.

Let also denote by $F_{1,n}$ and $F_{2,n-1}$ the empirical cumulative distribution functions of $(|y_{i+1} - y_i|)_{0 \leq i \leq n-1}$ and $(|y_{i+2} - y_i|)_{0 \leq i \leq n-2}$, respectively. Observe that for all t in $\mathbb{R}$,

$$\begin{aligned} \sqrt{n}(F_{1,n}(t) - F_1(t)) &= \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} (\mathbb{1}_{\{|y_{i+1} - y_i| \leq t\}} - F_1(t)) \\ &= \frac{1}{\sqrt{n}} \sum_{i \in \{t_{n,1}^*, \dots, t_{n,m^*}^*\}} (\mathbb{1}_{\{|y_{i+1} - y_i| \leq t\}} - F_1(t)) + \frac{1}{\sqrt{n}} \sum_{\substack{0 \leq i \leq n-1 \\ i \notin \{t_{n,1}^*, \dots, t_{n,m^*}^*\}}} (\mathbb{1}_{\{|y_{i+1} - y_i| \leq t\}} - F_1(t)) \\ &= \frac{1}{\sqrt{n}} \sum_{0 \leq i \leq n-1} (\mathbb{1}_{\{|z_i| \leq t\}} - F_1(t)) + R_n(t), \end{aligned} \quad (3.24)$$

where $\sup_{t \in \mathbb{R}} |R_n(t)| = o_p(1)$, the $z_i = y_{i+1} - y_i$ except for $i = t_{n,1}^*, \dots, t_{n,m^*}^*$, where $z_i = \eta_{i+1} - \eta_i$, (η_i) being defined in (3.2).

Thus, by using the theorem of Csörgő and Mielniczuk (1996)², we obtain that the first term in the right-hand side of (3.24) converges in distribution to a zero-mean Gaussian process G in the space of càdlàg functions equipped with the uniform norm. Since the second term in the right-hand side tends uniformly to zero in probability, we get that $\sqrt{n}(F_{1,n} - F_1)$ converges in distribution to a zero-mean Gaussian process in the space of càdlàg functions equipped with the uniform norm and that the same holds for $\sqrt{n-1}(F_{2,n-1} - F_2)$.

By Lemma 21.3 of Van der Vaart (2000) the quantile function $T : F \mapsto F^{-1}(1/2)$ is Hadamard differentiable at F tangentially to the set of càdlàg functions h that are continuous at $F^{-1}(1/2)$ with derivative $T'_F(h) = -h(F^{-1}(1/2))/F'(F^{-1}(1/2))$. By applying the functional delta method (Van der Vaart, 2000, Theorem 20.8)³, we get that $\sqrt{n}(T(F_{1,n}) - T(F_1))$ converges in distribution to $T'_{F_1}(G)$. Moreover, by the Continuous mapping theorem, it is the same for $T'_{F_1}\{\sqrt{n}(F_{1,n} - F_1)\}$. Thus,

$$\begin{aligned} \sqrt{n}(F_{1,n}^{-1}(1/2) - F_1^{-1}(1/2)) &= T'_{F_1}\{\sqrt{n}(F_{1,n} - F_1)\} + o_p(1) \\ &= -\frac{1}{\sqrt{n}} \frac{\sum_{i=0}^{n-1} \left(\mathbb{1}_{\{|y_{i+1}-y_i| \leq F_1^{-1}(1/2)\}} - 1/2 \right)}{F'_1(F_1^{-1}(1/2))} + o_p(1). \end{aligned}$$

In the same way,

$$\begin{aligned} \sqrt{n-1}(F_{2,n-1}^{-1}(1/2) - F_2^{-1}(1/2)) &= \\ &= -\frac{1}{\sqrt{n-1}} \frac{\sum_{i=0}^{n-2} \left(\mathbb{1}_{\{|y_{i+2}-y_i| \leq F_2^{-1}(1/2)\}} - 1/2 \right)}{F'_2(F_2^{-1}(1/2))} + o_p(1), \end{aligned}$$

By applying the Delta method (Van der Vaart, 2000, Theorem 3.1)⁴ with the transformation $f(x) = x^2$, we get

$$\begin{aligned} \sqrt{n}(F_{1,n}^{-1}(1/2)^2 - F_1^{-1}(1/2)^2) &= \\ &= -\frac{2F_1^{-1}(1/2)}{\sqrt{n}} \frac{\sum_{i=0}^{n-1} \left(\mathbb{1}_{\{|y_{i+1}-y_i| \leq F_1^{-1}(1/2)\}} - 1/2 \right)}{F'_1(F_1^{-1}(1/2))} + o_p(1), \end{aligned}$$

$$\begin{aligned} \sqrt{n-1}(F_{2,n-1}^{-1}(1/2)^2 - F_2^{-1}(1/2)^2) &= \\ &= -\frac{2F_2^{-1}(1/2)}{\sqrt{n-1}} \frac{\sum_{i=0}^{n-2} \left(\mathbb{1}_{\{|y_{i+2}-y_i| \leq F_2^{-1}(1/2)\}} - 1/2 \right)}{F'_2(F_2^{-1}(1/2))} + o_p(1), \end{aligned}$$

Note that by (3.23), we obtain that

$$F_1^{-1}(1/2) = \sqrt{\frac{2\sigma^{*2}}{1+\rho^*}} \Phi^{-1}(3/4) \text{ and } F_2^{-1}(1/2) = \sqrt{2\sigma^{*2}} \Phi^{-1}(3/4). \quad (3.25)$$

Moreover,

$$F'_1(F_1^{-1}(1/2)) = 2\sqrt{\frac{1+\rho^*}{2\sigma^{*2}}} \varphi(\Phi^{-1}(3/4)) \text{ and } F'_2(F_2^{-1}(1/2)) = 2\sqrt{\frac{1}{2\sigma^{*2}}} \varphi(\Phi^{-1}(3/4)), \quad (3.26)$$

²See Theorem 2.4 of this document.

³See Theorem 2.6 of this document.

⁴See Theorem 2.5 of this document.

where φ denotes the probability distribution function of a standard Gaussian rv.

Observe that $\sqrt{n}(\tilde{\rho}_n - \rho^*)$ can be rewritten as follows:

$$\begin{aligned}\sqrt{n}(\tilde{\rho}_n - \rho^*) &= \sqrt{n} \frac{F_{2,n}^{-1}(1/2)^2 - (1 + \rho^*)F_{1,n}^{-1}(1/2)^2}{F_{1,n}^{-1}(1/2)^2} \\ &= \sqrt{n} \frac{(F_{2,n-1}^{-1}(1/2)^2 - F_2^{-1}(1/2)^2) - (1 + \rho^*)(F_{1,n}^{-1}(1/2)^2 - F_1^{-1}(1/2)^2)}{F_{1,n}^{-1}(1/2)^2} \\ &\quad + \sqrt{n} \frac{F_2^{-1}(1/2)^2 - (1 + \rho^*)F_1^{-1}(1/2)^2}{F_{1,n}^{-1}(1/2)^2} .\end{aligned}\quad (3.27)$$

By (3.25) the last term in the right-hand side of (3.27) is equal to zero. Thus,

$$\begin{aligned}F_{1,n}^{-1}(1/2)^2 \sqrt{n}(\tilde{\rho}_n - \rho^*) &= \\ \frac{1}{\sqrt{n-1}} \sum_{i=0}^{n-2} \left\{ a_2 \left(\mathbb{1}_{\{|y_{i+2}-y_i| \leq F_2^{-1}(1/2)\}} - 1/2 \right) - a_1 (1 + \rho^*) \left(\mathbb{1}_{\{|y_{i+1}-y_i| \leq F_1^{-1}(1/2)\}} - 1/2 \right) \right\} + o_p(1) ,\end{aligned}$$

where, by (3.26),

$$\begin{aligned}a_2 &= -\frac{2F_2^{-1}(1/2)}{F_2'(F_2^{-1}(1/2))} = -2\sigma^{*2} \frac{\Phi^{-1}(3/4)}{\varphi(\Phi^{-1}(3/4))} \\ \text{and } a_1 &= -\frac{2F_1^{-1}(1/2)}{F_1'(F_1^{-1}(1/2))} = -\frac{2\sigma^{*2}}{1 + \rho^*} \frac{\Phi^{-1}(3/4)}{\varphi(\Phi^{-1}(3/4))} .\end{aligned}$$

By (3.25), $\sqrt{n}(\tilde{\rho}_n - \rho^*)$ can thus be rewritten as follows:

$$F_{1,n}^{-1}(1/2)^2 \sqrt{n}(\tilde{\rho}_n - \rho^*) = \frac{2\sigma^{*2}\Phi^{-1}(3/4)^2}{1 + \rho^*} \frac{1}{\sqrt{n-1}} \sum_{0 \leq i \leq n-2} \Psi(\eta_i, \eta_{i+1}, \eta_{i+2}) + o_p(1) .$$

where Ψ is defined in (3.6) and (η_i) is defined in (3.2). Since Ψ is a function on $\mathbb{R}^3$ with Hermite rank greater than 1 and $(\eta_i)_{i \geq 0}$ is a stationary AR(1) Gaussian process, and since $F_{1,n}^{-1}(1/2)^2$ converges in probability to $F_1^{-1}(1/2)^2$, (3.4) follows by applying Arcones (1994, Theorem 4)⁵, Slutsky's lemma and Equation (3.25).

3.7.2 Hints for (3.7)

Note that if X has a Cauchy(x_0, γ) distribution then the characteristic function φ_X of X can be written as $\varphi_X(t) = e^{ix_0 t - \gamma|t|}$. Moreover, the cdf F_X of X is such that $F_X^{-1}(3/4) = x_0 + \gamma$. Thus, $\eta_i = \sum_{k \geq 0} (\rho^*)^k \varepsilon_{i-k}$ has a Cauchy $\left(\frac{x_0}{1-\rho^*}, \frac{\gamma}{1-|\rho^*|}\right)$ distribution and $(\rho^* - 1)\eta_i$ has a Cauchy $\left(-x_0, \frac{\gamma|\rho^*-1|}{1-|\rho^*|}\right)$ distribution. Since $\eta_{i+1} - \eta_i = (\rho^* - 1)\eta_i + \varepsilon_i$ is a sum of two independent Cauchy rv's, it is distributed as a Cauchy $\left(0, \gamma \left(1 + \left|\frac{\rho^*-1}{1-|\rho^*|}\right|\right)\right)$ distribution. In the same way, $\eta_{i+2} - \eta_i = (\rho^{*2} - 1)\eta_i + \rho^* \varepsilon_i + \varepsilon_{i+2}$ is a sum of three independent Cauchy rv's and has thus a Cauchy $(0, 2\gamma(1 + |\rho^*|))$. Let F_1 and F_2 denote the cdf of $(\eta_{i+1} - \eta_i)$ and $(\eta_{i+2} - \eta_i)$, respectively. By using the properties of the cdf of a Cauchy distribution, we get, on the one hand, that $F_2^{-1}(3/4) = 2\gamma(1 + |\rho^*|)$ and, on the other hand, that

$$F_1^{-1}(3/4) = \begin{cases} 2\gamma, & \text{if } \rho^* > 0, \\ \frac{2\gamma}{1+\rho^*}, & \text{if } \rho^* < 0. \end{cases}$$

⁵See Theorem 2.7 of this document.

From this we get that

$$\left(\frac{F_2^{-1}(3/4)}{F_1^{-1}(3/4)} \right)^2 - 1 = \begin{cases} \rho^*(2 + \rho^*), & \text{if } \rho^* > 0, \\ \rho^{*2}(\rho^{*2} - 2), & \text{if } \rho^* < 0. \end{cases}$$

The definition of $\widetilde{\rho}_n$ comes by inverting these last two functions.

3.7.3 Proof of Proposition 3.2

In the sequel, we need the following definitions, notations and remarks. Observe that (3.9) can be rewritten as follows:

$$z = \rho^* Bz + T(\mathbf{t}_n^*) \boldsymbol{\delta}^* + \epsilon, \quad (3.28)$$

where

$$z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}, \quad Bz = \begin{pmatrix} z_0 \\ \vdots \\ z_{n-1} \end{pmatrix}, \quad \boldsymbol{\delta}^* = \begin{pmatrix} \delta_0^* \\ \vdots \\ \delta_m^* \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad (3.29)$$

where $\delta_k^* = (1 - \rho^*)\mu_k^*$, for $0 \leq k \leq m$, and $T(\mathbf{t})$ is an $n \times (m+1)$ matrix where the k th column is $(\underbrace{0, \dots, 0}_{t_{k-1}}, \underbrace{1, \dots, 1}_{t_k - t_{k-1}}, \underbrace{0, \dots, 0}_{n - t_k})^T$.

Let us define the exact and estimated decorrelated series by

$$w^* = z - \rho^* Bz, \quad (3.30)$$

$$\bar{w} = z - \bar{\rho}_n Bz. \quad (3.31)$$

For any vector subspace E of $\mathbb{R}^n$, let π_E denote the orthogonal projection of $\mathbb{R}^n$ on E . Let also $\|\cdot\|$ be the Euclidian norm on $\mathbb{R}^n$, $\langle \cdot, \cdot \rangle$ the canonical scalar product on $\mathbb{R}^n$ and $\|\cdot\|_\infty$ the sup norm. For x a vector of $\mathbb{R}^n$ and $\mathbf{t} \in \mathcal{A}_{n,m}$, let

$$J_{n,m}(x, \mathbf{t}) = \frac{1}{n} \left(\|\pi_{E_{\mathbf{t}_n^*}}(x)\|^2 - \|\pi_{E_{\mathbf{t}}}(x)\|^2 \right), \quad (3.32)$$

written $J_n(x, \mathbf{t})$ in the sequel for notational simplicity. In (3.32), $E_{\mathbf{t}_n^*}$ and $E_{\mathbf{t}}$ correspond to the linear subspaces of $\mathbb{R}^n$ generated by the columns of $T(\mathbf{t}_n^*)$ and $T(\mathbf{t})$, respectively. We shall use the same decomposition as the one introduced in Lavielle and Moulines (2000):

$$J_n(x, \mathbf{t}) = K_n(x, \mathbf{t}) + V_n(x, \mathbf{t}) + W_n(x, \mathbf{t}), \quad (3.33)$$

where

$$K_n(x, \mathbf{t}) = \frac{1}{n} \left\| \left(\pi_{E_{\mathbf{t}_n^*}} - \pi_{E_{\mathbf{t}}} \right) \mathbb{E}x \right\|^2, \quad (3.34)$$

$$V_n(x, \mathbf{t}) = \frac{1}{n} \left(\left\| \pi_{E_{\mathbf{t}_n^*}}(x - \mathbb{E}x) \right\|^2 - \left\| \pi_{E_{\mathbf{t}}}(x - \mathbb{E}x) \right\|^2 \right), \quad (3.35)$$

$$W_n(x, \mathbf{t}) = \frac{2}{n} \left(\left\langle \pi_{E_{\mathbf{t}_n^*}}(x - \mathbb{E}x), \pi_{E_{\mathbf{t}_n^*}}(\mathbb{E}x) \right\rangle - \left\langle \pi_{E_{\mathbf{t}}}(x - \mathbb{E}x), \pi_{E_{\mathbf{t}}}(\mathbb{E}x) \right\rangle \right). \quad (3.36)$$

We shall also use the following notations:

$$\underline{\lambda} = \min_{1 \leq k \leq m} |\delta_k^* - \delta_{k-1}^*| , \quad (3.37)$$

$$\bar{\lambda} = \max_{1 \leq k \leq m} |\delta_k^* - \delta_{k-1}^*| , \quad (3.38)$$

$$\Delta_{\tau^*} = \min_{1 \leq k \leq m+1} (\tau_k^* - \tau_{k-1}^*) , \quad (3.39)$$

$$\mathcal{C}_{\nu, \gamma, n, m} = \{ \mathbf{t} \in \mathcal{A}_{n, m}; \nu \underline{\lambda}^{-2} \leq \|\mathbf{t} - \mathbf{t}_n^*\| \leq n\gamma \Delta_{\tau^*} \} , \quad (3.40)$$

$$\mathcal{C}'_{\nu, \gamma, n, m} = \mathcal{C}_{\nu, \gamma, n, m} \cap \{ \mathbf{t} \in \mathcal{A}_{n, m}; \forall k = 1, \dots, m, t_k \geq t_{n, k}^* \} , \quad (3.41)$$

$$\begin{aligned} \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I}) = & \{ \mathbf{t} \in \mathcal{C}'_{\nu, \gamma, n, m}; \\ & \forall k \in \mathcal{I}, \nu \underline{\lambda}^{-2} \leq t_k - t_{n, k}^* \leq n\gamma \Delta_{\tau^*} \text{ and } \forall k \notin \mathcal{I}, t_k - t_{n, k}^* < \nu \underline{\lambda}^{-2} \} , \end{aligned} \quad (3.42)$$

for any $\nu > 0$, $0 < \gamma < 1/2$ and $\mathcal{I} \subset \{1, \dots, m\}$. We shall also need the following lemmas in order to prove Proposition 3.2 which are proved below.

Lemma 3.1. *Let $(z_0, \dots, z_n)$ be defined by (3.1) or (3.9), then*

$$\|Bz\| = O_P(n^{1/2}) , \quad (3.43)$$

$$\|z\| = O_P(n^{1/2}) , \quad (3.44)$$

as n tends to infinity, where Bz and z are defined in (3.29).

Lemma 3.2. *Let $(z_0, \dots, z_n)$ be defined by (3.1) or (3.9) then, for all $\mathbf{t} \in \mathcal{A}_{n, m}$,*

$$|J_n(\bar{w}, \mathbf{t}) - J_n(w^*, \mathbf{t})| \leq 2 \frac{|\rho^* - \bar{\rho}|}{n} \|Bz\| (|\rho^* + \bar{\rho}| \|Bz\| + 2 \|z\|) = O_P(n^{-1/2}) = o_P(1) ,$$

as $n \rightarrow \infty$, where J_n is defined in (3.32), Bz and z are defined in (3.29), w^* is defined in (3.30) and $\bar{w}$ is defined in (3.31).

Lemma 3.3. *Under the assumptions of Proposition 3.2, $\|\bar{\tau}_n - \tau^*\|_\infty$ converges in probability to 0, as n tends to infinity.*

Lemma 3.4. *Under the assumptions of Proposition 3.2 and for any $\nu > 0$, $0 < \gamma < 1/2$ and $\mathcal{I} \subset \{1, \dots, m\}$,*

$$P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})} \left(\frac{1}{2} K_n(w^*, \mathbf{t}) + V_n(w^*, \mathbf{t}) + W_n(w^*, \mathbf{t})\right) \leq 0\right) \longrightarrow 0 , \text{ as } n \rightarrow \infty ,$$

where $\mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})$ is defined in (3.42) and w^* is defined in (3.30).

Lemma 3.5. *Under the assumptions of Proposition 3.2 and for any $\nu > 0$, $0 < \gamma < 1/2$ and $\mathcal{I} \subset \{1, \dots, m\}$,*

$$P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})} J_n(\bar{w}, \mathbf{t}) \leq 0\right) \longrightarrow 0 , \text{ as } n \rightarrow \infty ,$$

where $\mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})$ is defined in (3.42) and $\bar{w}$ is defined in (3.31).

Lemma 3.6. *Under the assumptions of Proposition 3.2,*

$$\|\hat{\tau}_n(z, \bar{\rho}_n) - \tau^*\|_\infty = O_P(n^{-1}) .$$

Proof of Lemma 3.1. Without loss of generality, assume $(z_0, \dots, z_n)$ is defined by (3.9). $\|z\|^2 - \|Bz\|^2 = z_n^2 - z_0^2 = O_P(1)$ thus we only need to prove (3.43). $\|Bz\|^2 = \sum_{i=0}^{n-1} z_i^2$ then Markov inequality implies that $\|Bz\|^2 = O_P(n)$. $\square$

Proof of Lemma 3.2. By (3.30), $\bar{w} = w^* + (\rho^* - \bar{\rho}_n)Bz$. Thus, by (3.32), we get

$$\begin{aligned} J_n(\bar{w}, \mathbf{t}) - J_n(w^*, \mathbf{t}) = & \frac{(\rho^* - \bar{\rho}_n)^2}{n} \|\pi_{E_{t^*}}(Bz)\|^2 + \frac{2(\rho^* - \bar{\rho}_n)}{n} \langle \pi_{E_{t^*}}(z - \rho^* Bz), \pi_{E_{t^*}}(Bz) \rangle \\ & - \frac{(\rho^* - \bar{\rho}_n)^2}{n} \|\pi_{E_t}(Bz)\|^2 - \frac{2(\rho^* - \bar{\rho}_n)}{n} \langle \pi_{E_t}(z - \rho^* Bz), \pi_{E_t}(Bz) \rangle. \end{aligned} \quad (3.45)$$

Observe that the sum of the first two term in the right-hand side of (3.45) can be rewritten as follows:

$$\begin{aligned} \frac{1}{n}(\rho^* - \bar{\rho}_n) \langle \pi_{E_{t^*}}(Bz), (\rho^* - \bar{\rho}_n)\pi_{E_{t^*}}(Bz) + 2\pi_{E_{t^*}}(z - \rho^* Bz) \rangle \\ = \frac{1}{n}(\rho^* - \bar{\rho}_n) \langle \pi_{E_{t^*}}(Bz), \pi_{E_{t^*}}(2z - (\rho^* + \bar{\rho}_n)Bz) \rangle. \end{aligned}$$

Since the same can be done for the last two terms in the right-hand side of (3.45), the Cauchy-Schwarz inequality and the 1-Lipschitz property of projections give

$$|J_n(\bar{w}, \mathbf{t}) - J_n(w^*, \mathbf{t})| \leq 2 \frac{|\rho^* - \bar{\rho}_n|}{n} \|Bz\| (|\rho^* + \bar{\rho}_n| \|Bz\| + 2\|z\|).$$

The conclusion follows from (3.13) and Lemma 3.1. $\square$

Proof of Lemma 3.3. Lavielle and Moulines (2000, proof of Theorem 3) give the following bounds for any $\mathbf{t} \in \mathcal{A}_{n,m}$:

$$K_n(w^*, \mathbf{t}) \geq \underline{\lambda}^2 \min \left(\frac{1}{n} \max_{1 \leq k \leq m} |t_k - t_{n,k}^*|, \Delta_{\tau^*} \right), \quad (3.46)$$

$$V_n(w^*, \mathbf{t}) \geq -\frac{2(m+1)}{n\Delta_n} \left(\max_{1 \leq s \leq n} \left(\sum_{i=1}^s \epsilon_i \right)^2 + \max_{1 \leq s \leq n} \left(\sum_{i=n-s}^n \epsilon_i \right)^2 \right), \quad (3.47)$$

$$|W_n(w^*, \mathbf{t})| \leq \frac{3(m+1)^2 \bar{\lambda}}{n} \left(\max_{1 \leq s \leq n} \left| \sum_{i=1}^s \epsilon_i \right| + \max_{1 \leq s \leq n} \left| \sum_{i=n-s}^n \epsilon_i \right| \right), \quad (3.48)$$

where $\underline{\lambda}$, $\bar{\lambda}$ and Δ_{τ^*} are defined in (3.37)–(3.39). For any $\nu > 0$, define, as in Lavielle and Moulines (2000, proof of Theorem 3),

$$\mathcal{C}_{n,m,\nu} = \{\mathbf{t} \in \mathcal{A}_{n,m}; \|\mathbf{t} - \mathbf{t}_n^*\|_\infty \geq n\nu\}. \quad (3.49)$$

For $0 < \nu < \Delta_{\tau^*}$, we have:

$$\begin{aligned}
P\left(\|\widehat{\mathbf{t}}_n(z, \bar{\rho}_n) - \mathbf{t}_n^*\|_\infty \geq n\nu\right) &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} J_n(\bar{\mathbf{w}}, \mathbf{t}) \leq 0\right) \\
&\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} (J_n(\bar{\mathbf{w}}, \mathbf{t}) - J_n(w^*, \mathbf{t})) \leq -\nu\lambda^2\right) \\
&+ P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} (V_n(w^*, \mathbf{t}) + W_n(w^*, \mathbf{t})) \leq -\nu\lambda^2\right) \\
&\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} (J_n(\bar{\mathbf{w}}, \mathbf{t}) - J_n(w^*, \mathbf{t})) \leq -\nu\lambda^2\right) \\
&+ P\left(\max_{1 \leq s \leq n} \left(\sum_{i=1}^s \epsilon_i\right)^2 + \max_{1 \leq s \leq n} \left(\sum_{i=n-s}^n \epsilon_i\right)^2 \geq c\lambda^2 n \Delta_n \nu\right) \\
&+ P\left(\max_{1 \leq s \leq n} \left|\sum_{i=1}^s \epsilon_i\right| + \max_{1 \leq s \leq n} \left|\sum_{i=n-s}^n \epsilon_i\right| \geq c\lambda^2 n \nu \bar{\lambda}^{-1}\right)
\end{aligned} \tag{3.50}$$

for some positive constant c . The last two terms of this sum go to 0 when n goes to infinity (see Lavielle and Moulines, 2000, proof of Theorem 3). To show that the first term shares the same property, it suffices to show that $J_n(\bar{\mathbf{w}}, \mathbf{t}) - J_n(w^*, \mathbf{t})$ is bounded uniformly in $\mathbf{t}$ by a sequence of rv's which converges to 0 in probability. This result holds by Lemma 3.2. $\square$

Proof of Lemma 3.4. Using Equations (64)–(66) of Lavielle and Moulines (2000), one can show the bound (73) of Lavielle and Moulines (2000) on

$$P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} (K_n(w^*, \mathbf{t}) + V_n(w^*, \mathbf{t}) + W_n(w^*, \mathbf{t})) \leq 0\right).$$

Using the same arguments, we have the same bound on

$$P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} \left(\frac{1}{2}K_n(w^*, \mathbf{t}) + V_n(w^*, \mathbf{t}) + W_n(w^*, \mathbf{t})\right) \leq 0\right).$$

We conclude using Equations (67)–(71) of Lavielle and Moulines (2000). $\square$

Proof of Lemma 3.5. By (3.33),

$$\begin{aligned}
&P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} J_n(\bar{\mathbf{w}}, \mathbf{t}) \leq 0\right) \\
&\leq P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} \left(J_n(\bar{\mathbf{w}}, \mathbf{t}) - J_n(w^*, \mathbf{t}) + \frac{1}{2}K_n(w^*, \mathbf{t})\right) \leq 0\right) \\
&\quad + P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} \left(\frac{1}{2}K_n(w^*, \mathbf{t}) + V_n(w^*, \mathbf{t}) + W_n(w^*, \mathbf{t})\right) \leq 0\right).
\end{aligned}$$

By Lemma 3.4, the conclusion thus follows if

$$P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} \left(J_n(\bar{\mathbf{w}}, \mathbf{t}) - J_n(w^*, \mathbf{t}) + \frac{1}{2}K_n(w^*, \mathbf{t})\right) \leq 0\right) \longrightarrow 0, \text{ as } n \rightarrow \infty.$$

Since $\min_{t \in \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})} K_n(w^*, t) \geq (1 - \gamma) \Delta_{\tau^*} \nu$ (see Lavielle and Moulines, 2000, Equation (65)),

$$\begin{aligned} & P \left(\min_{t \in \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})} \left(J_n(\bar{w}, t) - J_n(w^*, t) + \frac{1}{2} K_n(w^*, t) \right) \leq 0 \right) \\ & \leq P \left(\min_{t \in \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})} (J_n(\bar{w}, t) - J_n(w^*, t)) \leq \frac{1}{2} (\gamma - 1) \Delta_{\tau^*} \nu \right), \end{aligned}$$

and we conclude by Lemma 3.2. $\square$

Proof of Lemma 3.6. For notational simplicity, $\widehat{\mathbf{t}}_n(z, \bar{\rho}_n)$ will be replaced by $\bar{\mathbf{t}}_n$ in this proof. Since for any $\nu > 0$,

$$P \left(\|\bar{\mathbf{t}}_n - \mathbf{t}_n^*\|_\infty < \nu \underline{\lambda}^{-2} \right) = P(\|\bar{\mathbf{t}}_n - \mathbf{t}_n^*\|_\infty \leq n\gamma \Delta_{\tau^*}) - P(\bar{\mathbf{t}}_n \in \mathcal{C}_{\nu, \gamma, n, m}),$$

it is enough, by Lemma 3.3, to prove that

$$P(\bar{\mathbf{t}}_n \in \mathcal{C}_{\nu, \gamma, n, m}) \longrightarrow 0, \text{ as } n \rightarrow \infty,$$

for all $\nu > 0$ and $0 < \gamma < 1/2$. Since

$$\mathcal{C}_{\nu, \gamma, n, m} = \bigcup_{\mathcal{I} \subset \{1, \dots, m\}} \mathcal{C}_{\nu, \gamma, n, m} \cap \{t \in \mathcal{A}_{n, m}; \forall k \in \mathcal{I}, t_k \geq t_{n, k}^*\},$$

we shall only study one set in the union without loss of generality and prove that

$$P(\bar{\mathbf{t}}_n \in \mathcal{C}'_{\nu, \gamma, n, m}) \longrightarrow 0, \text{ as } n \rightarrow \infty,$$

where $\mathcal{C}'_{\nu, \gamma, n, m}$ is defined in (3.41). Since $\mathcal{C}'_{\nu, \gamma, n, m} = \bigcup_{\mathcal{I} \subset \{1, \dots, m\}} \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})$, we shall only study one set in the union without loss of generality and prove that

$$P(\bar{\mathbf{t}}_n \in \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})) \longrightarrow 0, \text{ as } n \rightarrow \infty.$$

Since

$$P(\bar{\mathbf{t}}_n \in \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})) \leq P \left(\min_{t \in \mathcal{C}'_{\nu, \gamma, n, m}(\mathcal{I})} J_n(\bar{w}, t) \leq 0 \right),$$

the proof is complete by Lemma 3.5. $\square$

Proof of Proposition 3.2. For notational simplicity, $\widehat{\boldsymbol{\delta}}_n(z, \bar{\rho}_n)$ will be replaced by $\bar{\boldsymbol{\delta}}_n$ in this proof. By Lemma 3.6, the last result to show is

$$\|\bar{\boldsymbol{\delta}}_n - \boldsymbol{\delta}^*\| = O_P(n^{-1/2}),$$

that is, for all k , $\bar{\delta}_{n, k} - \delta_k^* = O_P(n^{-1/2})$. By (3.30) and (3.31),

$$\begin{aligned} \bar{\delta}_{n, k} &= \frac{1}{\bar{t}_{n, k+1} - \bar{t}_{n, k}} \sum_{i=\bar{t}_{n, k}+1}^{\bar{t}_{n, k+1}} \bar{w}_i \\ &= \frac{1}{n(\bar{\tau}_{n, k+1} - \bar{\tau}_{n, k})} \left(\sum_{i=\bar{t}_{n, k}+1}^{\bar{t}_{n, k+1}} w_i^* + (\rho^* - \bar{\rho}_n) \sum_{i=\bar{t}_{n, k}+1}^{\bar{t}_{n, k+1}} z_{i-1} \right). \end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\left| \sum_{i=\bar{t}_{n,k+1}}^{\bar{t}_{n,k+1}} z_{i-1} \right| \leq (\bar{t}_{n,k+1} - \bar{t}_{n,k})^{1/2} \left(z_{\bar{t}_{n,k}}^2 + \cdots + z_{\bar{t}_{n,k+1}-1}^2 \right)^{1/2} \leq n^{1/2} \|Bz\| = O_P(n),$$

where the last equality comes from Lemma 3.1. Hence by (3.13) and Lemma 3.6,

$$\begin{aligned} \bar{\delta}_{n,k} &= \frac{1}{n(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})} \sum_{i=\bar{t}_{n,k+1}}^{\bar{t}_{n,k+1}} w_i^* + O_P(n^{-1/2}) \\ &= \frac{1}{n(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})} \left(\sum_{i=\bar{t}_{n,k+1}}^{\bar{t}_{n,k+1}} \mathbb{E} w_i^* + \sum_{i=\bar{t}_{n,k+1}}^{\bar{t}_{n,k+1}} \epsilon_i \right) + O_P(n^{-1/2}), \end{aligned}$$

where the last equality comes from (3.28) and (3.30).

Let us now prove that

$$\frac{1}{n(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})} \sum_{i=\bar{t}_{n,k+1}}^{\bar{t}_{n,k+1}} \epsilon_i = O_P(n^{-1/2}). \quad (3.51)$$

By Lemma 3.3, $n^{-1}(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})^{-1} = O_P(n^{-1})$. Moreover,

$$\sum_{i=\bar{t}_{n,k+1}}^{\bar{t}_{n,k+1}} \epsilon_i = \sum_{i=t_{n,k}^*+1}^{t_{n,k+1}^*} \epsilon_i \pm \sum_{i=\bar{t}_{n,k+1}}^{t_{n,k}^*} \epsilon_i \pm \sum_{i=t_{n,k+1}^*}^{\bar{t}_{n,k+1}+1} \epsilon_i. \quad (3.52)$$

By the Chebyshev inequality, the first term in the right-hand side of (3.52) is $O_P(n^{1/2})$.

By using the Cauchy-Schwarz inequality, we get that the second term of (3.52) satisfies: $|\sum_{i=\bar{t}_{n,k+1}}^{t_{n,k}^*} \epsilon_i| \leq |t_{n,k}^* - \bar{t}_{n,k}|^{1/2} (\sum_{i=1}^n \epsilon_i^2)^{1/2} = O_P(1) O_P(n^{1/2}) = O_P(n^{1/2})$, by Lemma 3.6. The same holds for the last term in the right-hand side of (3.52), which gives (3.51). Hence,

$$\begin{aligned} \bar{\delta}_{n,k} - \delta_k^* &= \frac{1}{n(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})} \sum_{i=\bar{t}_{n,k+1}}^{\bar{t}_{n,k+1}} (\mathbb{E} w_i^* - \delta_k^*) + O_P(n^{-1/2}) \\ &= \frac{1}{n(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})} \sum_{i \in \{\bar{t}_{n,k}+1, \dots, \bar{t}_{n,k+1}\} \setminus \{t_{n,k}^*+1, \dots, t_{n,k+1}^*\}} (\mathbb{E} w_i^* - \delta_k^*) + O_P(n^{-1/2}), \end{aligned}$$

and then

$$\begin{aligned} |\bar{\delta}_{n,k} - \delta_k^*| &\leq \frac{1}{n(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})} \# \{ \bar{t}_{n,k} + 1, \dots, \bar{t}_{n,k+1} \} \setminus \{ t_{n,k}^* + 1, \dots, t_{n,k+1}^* \} \max_{l=0, \dots, m} |\delta_l^* - \delta_k^*| \\ &\quad + O_P(n^{-1/2}). \end{aligned}$$

We conclude by using Lemma 3.6 to get $\# \{ \bar{t}_{n,k} + 1, \dots, \bar{t}_{n,k+1} \} \setminus \{ t_{n,k}^* + 1, \dots, t_{n,k+1}^* \} = O_P(1)$ and Lemma 3.3 to get $(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})^{-1} = O_P(1)$.

□

3.7.4 Proof of Proposition 3.3

The connection between models (3.1) and (3.9) is made by the following lemmas.

Lemma 3.7. *Let $(y_0, \dots, y_n)$ be defined by (3.1) and let*

$$v_i^* = y_i - \rho^* y_{i-1}, \quad (3.53)$$

$$\Delta_i^* = \begin{cases} -\rho^* (\mu_k^* - \mu_{k-1}^*) & \text{if } i = t_{n,k}^* + 1, \\ 0, & \text{otherwise,} \end{cases} \quad (3.54)$$

where the μ_k^* 's are defined in (3.1), then the process

$$w_i^* = v_i^* + \Delta_i^* \quad (3.55)$$

has the same distribution as $z_i - \rho^* z_{i-1}$ where $(z_0, \dots, z_n)$ is defined by (3.9). Such a process $(z_0, \dots, z_n)$ can be constructed recursively as

$$\begin{cases} z_0 &= y_0 \\ z_i &= w_i^* + \rho^* z_{i-1} \text{ for } i > 0. \end{cases} \quad (3.56)$$

Lemma 3.8. *Let $(y_0, \dots, y_n)$ be defined by (3.1) and let z be defined by (3.53– 3.56). Then*

$$\bar{w}_i = \bar{v}_i + \bar{\Delta}_i \quad (3.57)$$

where

$$\bar{v}_i = y_i - \bar{\rho}_n y_{i-1} \quad (3.58)$$

$$\bar{w}_i = z_i - \bar{\rho}_n z_{i-1} \quad (3.59)$$

$$\bar{\Delta}_i = \Delta_i^* + (\rho^* - \bar{\rho}_n) (z_{i-1} - y_{i-1}) . \quad (3.60)$$

Lemma 3.9. *Let $\bar{\Delta} = (\bar{\Delta}_i)_{0 \leq i \leq n}$ as defined in (3.60). Then $\|\bar{\Delta}\| = O_P(1)$.*

Proof of Lemma 3.7. Let z being defined by (3.56). Using (3.55), we get, for all $0 \leq k \leq m$, $t_{n,k}^* < i \leq t_{n,k+1}^*$,

$$\begin{aligned} (z_i - \mu_k^*) - \rho^* (z_{i-1} - \mu_k^*) &= (y_i - \mu_k^*) - \rho^* (y_{i-1} - \mu_k^*) + \Delta_i^* \\ &= \begin{cases} (y_i - \mu_k^*) - \rho^* (y_{i-1} - \mu_{k-1}^*) & \text{if } i = t_{n,k}^* + 1, \\ (y_i - \mu_k^*) - \rho^* (y_{i-1} - \mu_k^*) & \text{otherwise.} \end{cases} \end{aligned}$$

This expression equals $(y_i - \mathbb{E}(y_i)) - \rho^* (y_{i-1} - \mathbb{E}(y_{i-1})) = \eta_i - \rho^* \eta_{i-1} = \epsilon_i$ by (3.1) and (3.2). Then z satisfies (3.9). $\square$

Proof of Lemma 3.8. The proof of Lemma 3.8 is straightforward. $\square$

Proof of Lemma 3.9. (3.60) can be written as

$$\overline{\Delta} = \Delta^* + (\rho^* - \bar{\rho}_n)(By - Bz)$$

where $\Delta^* = (\Delta_i^*)_{1 \leq i \leq n}$, $By = (y_{i-1})_{1 \leq i \leq n}$ and Bz is defined in (3.29). By the triangle inequality,

$$\|\overline{\Delta}\| \leq \|\Delta^*\| + |\rho^* - \bar{\rho}_n|(\|By\| + \|Bz\|). \quad (3.61)$$

Since $\|\Delta^*\|$ is constant it is bounded. The conclusion follows from (3.61), (3.13) and Lemma 3.1. $\square$

Proof of Proposition 3.3. Let $y, z, \bar{v}, \bar{w}$ and $\overline{\Delta}$ be defined in Lemma 3.8.

Using (3.32) and Lemma 3.8, we get

$$J_n(\bar{v}, \mathbf{t}) = J_n(\bar{w}, \mathbf{t}) + J_n(\overline{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right).$$

By the Cauchy-Schwarz inequality and the 1-Lipschitz property of projections, we have

$$\begin{aligned} |J_n(\overline{\Delta}, \mathbf{t})| &\leq \frac{2}{n} \|\overline{\Delta}\|^2 \\ \left| \left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right| &\leq 2 \|\overline{\Delta}\| \|\bar{w}\|. \end{aligned}$$

Note that $\bar{w} = z - \bar{\rho}_n Bz$ thus by the triangle inequality

$$\|\bar{w}\| \leq \|z\| + |\bar{\rho}_n| \|Bz\|.$$

Since $|\bar{\rho}_n| = O_P(1)$, we deduce from Lemma 3.1 that $\|\bar{w}\| = O_P(n^{1/2})$. Since, by Lemma 3.9, $\|\overline{\Delta}\| = O_P(1)$, we obtain that

$$\sup_{\mathbf{t}} \left| J_n(\overline{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right) \right| = O_P(n^{-1/2}). \quad (3.62)$$

For $0 < \nu < \Delta_{\tau^*}$, using (3.33) and (3.49), we get:

$$\begin{aligned} P\left(\|\bar{\mathbf{t}}_n - \mathbf{t}^*\|_{\infty} \geq \nu\right) &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} J_n(\bar{v}, \mathbf{t}) \leq 0\right) \\ &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} \left\{ J_n(\bar{w}, \mathbf{t}) + J_n(\overline{\Delta}, \mathbf{t}) \right. \right. \\ &\quad \left. \left. - \frac{2}{n} \left(\left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right) \right\} \leq 0\right) \\ &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} \left\{ K_n(\bar{w}, \mathbf{t}) + V_n(\bar{w}, \mathbf{t}) + W_n(\bar{w}, \mathbf{t}) + J_n(\overline{\Delta}, \mathbf{t}) \right. \right. \\ &\quad \left. \left. - \frac{2}{n} \left(\left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right) \right\} \leq 0\right) \\ &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} \left\{ \frac{1}{2} K_n(\bar{w}, \mathbf{t}) + V_n(\bar{w}, \mathbf{t}) + W_n(\bar{w}, \mathbf{t}) \right\} \leq 0\right) \\ &\quad + P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,m,\nu}} \left\{ \frac{1}{2} K_n(\bar{w}, \mathbf{t}) + J_n(\overline{\Delta}, \mathbf{t}) \right. \right. \\ &\quad \left. \left. - \frac{2}{n} \left(\left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right) \right\} \leq 0\right). \end{aligned}$$

Following the proof of Lemma 3.3, one can prove that

$$P\left(\min_{t \in \mathcal{C}_{n,m,\nu}} \left\{ \frac{1}{2} K_n(\bar{w}, t) + V_n(\bar{w}, t) + W_n(\bar{w}, t) \right\} \leq 0\right) \xrightarrow{n \rightarrow \infty} 0. \quad (3.63)$$

Using (3.46), we get that

$$\begin{aligned} P\left(\min_{t \in \mathcal{C}_{n,m,\nu}} \left\{ \frac{1}{2} K_n(\bar{w}, t) + J_n(\bar{\Delta}, t) - \frac{2}{n} \left(\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\bar{\Delta}) \rangle \right) \right\} \leq 0\right) \\ \leq P\left(\frac{1}{2} \lambda^2 \nu + \min_{t \in \mathcal{C}_{n,m,\nu}} \left\{ J_n(\bar{\Delta}, t) - \frac{2}{n} \left(\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\bar{\Delta}) \rangle \right) \right\} \leq 0\right) \end{aligned} \quad (3.64)$$

which goes to zero when n goes to infinity by (3.62).

Then Lemma 3.3 still holds if y is defined by (3.1). To show the rate of convergence, we use the same decomposition. As in the proof of Lemma 3.6, $P\left(\min_{t \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} J_n(\bar{v}, t) \leq 0\right) \xrightarrow{n \rightarrow \infty} 0$ for all $\nu > 0$ and $0 < \gamma < 1/2$ is a sufficient condition for proving that $P(\hat{\mathbf{t}}_n(y, \bar{\rho}_n) \in \mathcal{C}_{\nu,\gamma,n,m}) \xrightarrow{n \rightarrow \infty} 0$, which allows us to conclude on the rate of convergence of the estimated change-points. Note that

$$\begin{aligned} P\left(\min_{t \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} J_n(\bar{v}, t) \leq 0\right) &\leq P\left(\min_{t \in \mathcal{C}'_{\nu,\gamma,n,m}(\mathcal{I})} \left\{ \frac{1}{2} K_n(\bar{w}, t) + V_n(\bar{w}, t) + W_n(\bar{w}, t) \right\} \leq 0\right) \\ &+ P\left(\frac{1}{2} \lambda^2 \nu + J_n(\bar{\Delta}, t) - \frac{2}{n} \left(\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\bar{\Delta}) \rangle \right) \leq 0\right). \end{aligned}$$

In the latter equation, the second term of the right-hand side goes to zero as n goes to infinity by (3.62).

The first term of right-hand side goes to zero when n goes to infinity by following the same line of reasoning as the one of Lemma 3.5. This concludes the proof of Proposition 3.3. $\square$

3.7.5 Proof of Proposition 3.4

We shall use in this section the notations introduced in Sections 3.7.3 and 3.4.1. The result derives directly from Lemmas 3.10 and 3.11.

Lemma 3.10. *Under the assumptions of Proposition 3.4, $P(\hat{m} = m) \xrightarrow{n \rightarrow \infty} 0$ if $m < m^*$.*

Lemma 3.11. *Under the assumptions of Proposition 3.4, $P(\hat{m} = m) \xrightarrow{n \rightarrow \infty} 0$ if $m > m^*$.*

Proof of Lemma 3.10. If $\hat{m} = m < m^*$, then

$$\frac{1}{n} SS_m(z, \bar{\rho}_n) + \beta_n m \leq \frac{1}{n} SS_{m^*}(z, \bar{\rho}_n) + \beta_n m^*,$$

where SS_m is defined in (3.15). In particular, there exists $t \in \mathcal{A}_{n,m}$ such that

$$\frac{1}{n} \min_{\delta} SS_m(z, \bar{\rho}_n, \delta, t) + \beta_n m \leq \frac{1}{n} \min_{\delta} SS_m(z, \bar{\rho}_n, \delta, t_n^*) + \beta_n m^*.$$

From (3.32), we get

$$J_n(\bar{w}, \mathbf{t}) \leq \beta_n(m^* - m) .$$

Since (β_n) converges to zero, for any $\varepsilon > 0$, $\beta_n(m^* - m) \leq \varepsilon$ for a large enough n , and so

$$J_n(\bar{w}, \mathbf{t}) \leq \varepsilon .$$

One can check that there exist $0 < \nu < \Delta_{\tau^*}$ such that, for a large enough n , there exists $\mathbf{t}' \in \mathcal{C}_{n, m^*, \nu}$ such that $E_{\mathbf{t}} \subset E_{\mathbf{t}'}$ (that is the change-points of $\mathbf{t}$ are change-points of $\mathbf{t}'$) for all $\mathbf{t} \in \mathcal{A}_{n, m}$, where $\mathcal{C}_{n, m^*, \nu}$ is defined in (3.49). From (3.32) and $E_{\mathbf{t}} \subset E_{\mathbf{t}'}$, we get $J_n(\bar{w}, \mathbf{t}') \leq J_n(\bar{w}, \mathbf{t})$. Then, the following inequality holds for all $\varepsilon > 0$ and any large enough n :

$$P(\widehat{m} = m) \leq P(\exists \mathbf{t}' \in \mathcal{C}_{n, m^*, \nu}, J_n(\bar{w}, \mathbf{t}') \leq \varepsilon) . \quad (3.65)$$

We then follow the steps of (3.50), $-\nu\lambda^2$ being replaced by $\varepsilon - \nu\lambda^2$. The convergence of $P(\exists \mathbf{t}' \in \mathcal{C}_{n, m^*, \nu}, J_n(\bar{w}, \mathbf{t}') \leq \varepsilon)$ to zero holds with $\varepsilon < \nu\lambda^2$. We can conclude with (3.65). $\square$

Proof of Lemma 3.11. Following the proof of Lemma 3.10, if $\widehat{m} = m > m^*$, there exists $\mathbf{t} \in \mathcal{A}_{n, m}$ such that $J_n(\bar{w}, \mathbf{t}) \leq \beta_n(m^* - m)$ and then $J_n(\bar{w}, \mathbf{t}) + \beta_n \leq 0$ since $m > m^*$. Then

$$P(\widehat{m} = m) \leq P(\exists \mathbf{t} \in \mathcal{A}_{n, m}, J_n(\bar{w}, \mathbf{t}) + \beta_n \leq 0) . \quad (3.66)$$

Adding the change-points of $\mathbf{t}_n^*$ to those of such a $\mathbf{t}$, one can get $\mathbf{t}' \in \mathcal{A}_{n, m'}$ with $m^* < m \leq m' \leq m + m^*$ such that $E_{\mathbf{t}} \cup E_{\mathbf{t}_n^*} \subset E_{\mathbf{t}'}$, provided that $(m + m^*)[\Delta_n] \leq n$, where $[\cdot]$ is the ceiling function, this condition being fulfilled for any sufficiently large n under the assumptions of Proposition 3.4 since $n^{-1}\Delta_n$ converges to zero. Since $E_{\mathbf{t}} \subset E_{\mathbf{t}'}$, we derive $J_n(\bar{w}, \mathbf{t}') + \beta_n \leq J_n(\bar{w}, \mathbf{t}) + \beta_n$ from (3.32). Then, from (3.66), we get

$$\forall m' > m^*, P(\exists \mathbf{t}' \in \mathcal{A}_{n, m'}, E_{\mathbf{t}_n^*} \subset E_{\mathbf{t}'}, J_n(\bar{w}, \mathbf{t}') + \beta_n \leq 0) \xrightarrow{n \rightarrow \infty} 0 \quad (3.67)$$

is a sufficient condition to prove the lemma. Let us prove (3.67). Let $m' > m^*$ and such a $\mathbf{t}'$. We compare $J_n(\bar{w}, \mathbf{t}')$ to $J_n(w^*, \mathbf{t}')$. Since $\mathbb{E}w^* \in E_{\mathbf{t}_n^*} \subset E_{\mathbf{t}'}$, $K_n(w^*, \mathbf{t}') = 0$ by (3.34). By (3.36) and $\mathbb{E}w^* \in E_{\mathbf{t}_n^*} \subset E_{\mathbf{t}'}$,

$$\begin{aligned} W_n(w^*, \mathbf{t}') &= \frac{2}{n} \left(\left\langle \pi_{E_{\mathbf{t}_n^*}}(w^* - \mathbb{E}w^*), \pi_{E_{\mathbf{t}_n^*}}(\mathbb{E}w^*) \right\rangle - \left\langle \pi_{E_{\mathbf{t}'}}(w^* - \mathbb{E}w^*), \pi_{E_{\mathbf{t}'}}(\mathbb{E}w^*) \right\rangle \right) \\ &= \frac{2}{n} \left\langle \pi_{E_{\mathbf{t}_n^*}}(w^* - \mathbb{E}w^*) - \pi_{E_{\mathbf{t}'}}(w^* - \mathbb{E}w^*), \pi_{E_{\mathbf{t}_n^*}}(\mathbb{E}w^*) \right\rangle \\ &= -\frac{2}{n} \left\langle \pi_{E_{\mathbf{t}_n^*}^\perp} \pi_{E_{\mathbf{t}'}}(w^* - \mathbb{E}w^*), \pi_{E_{\mathbf{t}_n^*}}(\mathbb{E}w^*) \right\rangle \\ &= 0 , \end{aligned}$$

where $E^\perp$ is the (Euclidian) orthogonal complement of the vector subspace E . Then $J_n(w^*, \mathbf{t}') = V_n(w^*, \mathbf{t}')$ and

$$J_n(\bar{w}, \mathbf{t}') = V_n(w^*, \mathbf{t}') + (J_n(\bar{w}, \mathbf{t}') - J_n(w^*, \mathbf{t}')) . \quad (3.68)$$

Using (3.47), $V_n(w^*, \mathbf{t}) \geq -\frac{2(m'+1)}{n\Delta_n} M_n$, where

$$\begin{aligned} M_n &= M_{n,1} + M_{n,2}, \\ M_{n,1} &= \max_{1 \leq s \leq n} \left(\sum_{i=1}^s \epsilon_i \right)^2, \\ M_{n,2} &= \max_{1 \leq s \leq n} \left(\sum_{i=n-s}^n \epsilon_i \right)^2. \end{aligned}$$

We define $D_n = \sup_{\mathbf{t}' \in \mathcal{A}_{n,m'}} |J_n(\bar{w}, \mathbf{t}') - J_n(w^*, \mathbf{t}')|$. Then, using (3.68),

$$J_n(\bar{w}, \mathbf{t}') \geq -\frac{2(m+1)}{n\Delta_n} M_n - D_n,$$

which implies

$$\begin{aligned} P(\exists \mathbf{t}' \in \mathcal{A}_{n,m'}, E_{\mathbf{t}'}^* \subset E_{\mathbf{t}'}, J_n(\bar{w}, \mathbf{t}') + \beta_n \leq 0) &\leq P\left(-\frac{2(m'+1)}{n\Delta_n} M_n - D_n + \beta_n \leq 0\right) \\ &\leq P\left(\frac{2(m'+1)}{n\Delta_n} M_n \geq \frac{\beta_n}{2}\right) + P\left(D_n \geq \frac{\beta_n}{2}\right). \end{aligned}$$

By Lemma 3.2, $D_n = O_P(n^{-1/2})$ and then $P(D_n \geq \frac{\beta_n}{2})$ tends to zero as n tends to infinity since $n^{1/2}\beta_n \xrightarrow{n \rightarrow \infty} +\infty$. Let us now prove that $P\left(\frac{2(m'+1)}{n\Delta_n} M_n \geq \frac{\beta_n}{2}\right)$ tends to zero as n tends to infinity, which concludes the proof. Note that

$$P\left(\frac{2(m'+1)}{n\Delta_n} M_n \geq \frac{\beta_n}{2}\right) \leq P\left(M_{n,1} \geq \frac{n\Delta_n\beta_n}{8(m'+1)}\right) + P\left(M_{n,2} \geq \frac{n\Delta_n\beta_n}{8(m'+1)}\right).$$

We prove the convergence for each term in the right-hand side of the above equation. We shall prove it for the first term in the right-hand side since the arguments for the other term are the same. From Kolmogorov's maximal inequality (see for example Durrett, 2010, Theorem 2.5.2.)⁶, since $(\epsilon_i)_{i \geq 0}$ is a sequence of independent rv's with zero-mean and finite variance σ^{*2} ,

$$\forall \delta > 0, P(M_{n,1} \geq \delta^2) \leq \frac{n\sigma^{*2}}{\delta^2}. \quad (3.69)$$

Letting $\delta^2 = \frac{n\Delta_n\beta_n}{8(m'+1)}$ in (3.69), we get

$$P\left(M_{n,1} \geq \frac{n\Delta_n\beta_n}{8(m'+1)}\right) \leq \frac{8(m'+1)\sigma^{*2}}{\Delta_n\beta_n},$$

which goes to 0 as n tends to infinity because $\Delta_n\beta_n \xrightarrow{n \rightarrow \infty} +\infty$. The proof of the convergence of $P\left(M_{n,2} \geq \frac{n\Delta_n\beta_n}{8(m'+1)}\right)$ follows the same lines. $\square$

3.7.6 Proof of Proposition 3.5

Lemma 3.12. *Under the assumptions of Proposition 3.5, $P(\widehat{m} = m) \xrightarrow{n \rightarrow \infty} 0$ if $m < m^*$.*

Lemma 3.13. *Under the assumptions of Proposition 3.5, $P(\widehat{m} = m) \xrightarrow{n \rightarrow \infty} 0$ if $m > m^*$.*

⁶See Theorem 2.9 of this document.

Proof of Lemma 3.12. Following the proof of Lemma 3.10 and replacing $\bar{w}$ by $\bar{v}$, we get, for any $\varepsilon > 0$,

$$P(\widehat{m} = m) \leq P(\exists \mathbf{t}' \in \mathcal{C}_{n,m^*,\nu}, J_n(\bar{v}, \mathbf{t}') \leq \varepsilon) \quad (3.70)$$

$$\begin{aligned} &\leq P\left(\exists \mathbf{t}' \in \mathcal{C}_{n,m^*,\nu}, \frac{1}{2}K_n(\bar{w}, \mathbf{t}') + V_n(\bar{w}, \mathbf{t}') + W_n(\bar{w}, \mathbf{t}') \leq \frac{\varepsilon}{2}\right) \\ &+ P\left(\exists \mathbf{t}' \in \mathcal{C}_{n,m^*,\nu}, \frac{1}{2}K_n(\bar{w}, \mathbf{t}') + J_n(\bar{v}, \mathbf{t}') - J_n(\bar{w}, \mathbf{t}') \leq \frac{\varepsilon}{2}\right), \end{aligned} \quad (3.71)$$

since

$$J_n(\bar{v}, \mathbf{t}') = \frac{1}{2}K_n(\bar{w}, \mathbf{t}') + V_n(\bar{w}, \mathbf{t}') + W_n(\bar{w}, \mathbf{t}') + \frac{1}{2}K_n(\bar{w}, \mathbf{t}') + J_n(\bar{v}, \mathbf{t}') - J_n(\bar{w}, \mathbf{t}').$$

From (3.63) and (3.71), it suffices to prove that

$$P\left(\exists \mathbf{t}' \in \mathcal{C}_{n,m^*,\nu}, \frac{1}{2}K_n(\bar{w}, \mathbf{t}') + J_n(\bar{v}, \mathbf{t}') - J_n(\bar{w}, \mathbf{t}') \leq \frac{\varepsilon}{2}\right) \xrightarrow{n \rightarrow \infty} 0$$

to conclude the proof. It follows from (3.62) and (3.64), $\frac{1}{2}\underline{\lambda}^2\nu$ being replaced by $\frac{1}{2}(\underline{\lambda}^2\nu - \varepsilon)$, which is positive if $\varepsilon < \underline{\lambda}^2\nu$. $\square$

Proof of Lemma 3.13. As in the proof of Lemma 3.11, it suffices to show that

$$P(\exists \mathbf{t} \in \mathcal{A}_{n,m}, J_n(\bar{v}, \mathbf{t}) + \beta_n \leq 0) \xrightarrow{n \rightarrow \infty} 0.$$

Since

$$J_n(\bar{v}, \mathbf{t}) \geq J_n(\bar{w}, \mathbf{t}) - \sup_t |J_n(\bar{v}, \mathbf{t}) - J_n(\bar{w}, \mathbf{t})|,$$

the result follows from

$$P\left(\exists \mathbf{t} \in \mathcal{A}_{n,m}, J_n(\bar{w}, \mathbf{t}) + \frac{1}{2}\beta_n \leq 0\right) \xrightarrow{n \rightarrow \infty} 0, \quad (3.72)$$

$$P\left(\sup_t |J_n(\bar{v}, \mathbf{t}) - J_n(\bar{w}, \mathbf{t})| \geq \frac{1}{2}\beta_n\right) \xrightarrow{n \rightarrow \infty} 0. \quad (3.73)$$

Equation (3.72) follows from the Proof of Lemma 3.11, replacing β_n by $\frac{1}{2}\beta_n$. Equation (3.73) follows from (3.62) and from $n^{1/2}\beta_n \xrightarrow{n \rightarrow \infty} +\infty$. $\square$

3.7.7 Proof of Proposition 3.6

We first give some lemmas which are useful for the proof of Proposition 3.6.

Lemma 3.14. *Under the assumptions of Proposition 3.6 with SS_m given by (3.15), we have, for any positive m ,*

$$SS_m(z, \bar{\rho}_n) = SS_m(z, \rho^*) + O_P(1), \text{ as } n \rightarrow \infty.$$

Lemma 3.15. *Under the assumptions of Proposition 3.6 with SS_m given by (3.15), we have, for any positive m ,*

$$SS_m(z, \rho^*)^{-1} = O_P(n^{-1}), \text{ as } n \rightarrow \infty.$$

Proof of Lemma 3.14. The proof of this Lemma follows exactly this of Lemma 3.16. The difference is that, in (3.9), the term Δ^* appearing in the decomposition (3.77) vanishes. $\square$

Proof of Lemma 3.15. We first define

$$SS_m(z, \rho, \mathbf{t}) = \arg \min_{\delta} SS_m(z, \rho, \delta, \mathbf{t}).$$

We have, for any positive M ,

$$\begin{aligned} P\left(\frac{n}{SS_m(z, \rho^*)} > M\right) &\leq P\left(\left\{\frac{SS_m(z, \rho^*)}{SS_m(z, \rho^*, \mathbf{t}^*)} > 1\right\} \cap \left\{\frac{n}{SS_m(z, \rho^*)} > M\right\}\right) \\ &\quad + P\left(\left\{\frac{SS_m(z, \rho^*)}{SS_m(z, \rho^*, \mathbf{t}^*)} < 1\right\} \cap \left\{\frac{n}{SS_m(z, \rho^*)} > M\right\}\right) \\ &\leq P\left(\frac{n}{SS_m(z, \rho^*, \mathbf{t}^*)} > M\right) + P\left(\frac{SS_m(z, \rho^*)}{SS_m(z, \rho^*, \mathbf{t}^*)} < 1\right). \end{aligned}$$

Under the assumptions of Proposition 3.2, a by product of the proof of Theorem 3 in Lavielle and Moulines (2000) is that

$$P\left(\frac{SS_m(z, \rho^*)}{SS_m(z, \rho^*, \mathbf{t}^*)} < 1\right) = P(SS_m(z, \rho^*) - SS_m(z, \rho^*, \mathbf{t}^*) < 0) \leq \kappa n^{-\alpha},$$

where κ is a positive constant depending on δ^* and $\mathbf{t}^*$, and α is a positive constant. Furthermore, as $\sigma^{*-2}SS_m(z, \rho^*, \mathbf{t}^*)$ has a χ_{n-m-1}^2 distribution, $n^{-1}SS_m(z, \rho^*, \mathbf{t}^*) = \sigma^{*2} + o_P(1)$ and thus $n^{-1}SS_m(z, \rho^*, \mathbf{t}^*) = O_P(1)$, which concludes the proof. $\square$

Proof of Proposition 3.6. We have to prove that, for a given positive m , $C_m(z, \rho^*) - C_m(z, \bar{\rho}_n) = O_P(1)$. Observe that, since $\widehat{\tau}_k(z, \rho) = \widehat{t}_k(z, \rho)/n$,

$$\begin{aligned} &\sum_{k=0}^m \log n_k(\widehat{t}(z, \bar{\rho}_n)) - \sum_{k=0}^m \log n_k(\widehat{t}(z, \rho^*)) \\ &= \sum_{k=0}^m \log(\widehat{\tau}_{k+1}(z, \bar{\rho}_n) - \widehat{\tau}_k(z, \bar{\rho}_n)) - \sum_{k=0}^m \log(\widehat{\tau}_{k+1}(z, \rho^*) - \widehat{\tau}_k(z, \rho^*)). \end{aligned} \quad (3.74)$$

By Proposition 3.2, both quantities of the previous equation converge in probability to

$$\sum_{k=0}^m \log(\tau_{k+1}^* - \tau_k^*)$$

thus

$$\sum_{k=0}^m \log n_k(\widehat{t}(z, \bar{\rho}_n)) - \sum_{k=0}^m \log n_k(\widehat{t}(z, \rho^*)) = O_P(1). \quad (3.75)$$

Further note that

$$\log SS_m(z, \bar{\rho}_n) - \log SS_m(z, \rho^*) = \log\left(\frac{SS_m(z, \bar{\rho}_n)}{SS_m(z, \rho^*)}\right) = R\left(\frac{SS_m(z, \bar{\rho}_n) - SS_m(z, \rho^*)}{SS_m(z, \rho^*)}\right),$$

where $R(x) = \log(1 + x)$. Lemma 3.14 states that $SS_m(z, \bar{\rho}_n) - SS_m(z, \rho^*) = O_P(1)$ and Lemma 3.15 that $[SS_m(z, \rho^*)]^{-1} = O_P(n^{-1})$ so, by Van der Vaart (2000, Lemma 2.12), we get that

$$\log SS_m(z, \bar{\rho}_n) - \log SS_m(z, \rho^*) = O_P(n^{-1}).$$

Hence

$$\frac{n-m+1}{2} \log SS_m(z, \bar{\rho}_n) - \frac{n-m+1}{2} \log SS_m(z, \rho^*) = O_P(1),$$

which with (3.75) concludes the proof of Proposition 3.6. $\square$

3.7.8 Proof of Proposition 3.7

We first give some lemmas which are useful for the proof of Proposition 3.7.

Lemma 3.16. *Under the assumptions of Proposition 3.6 with SS_m given by (3.15), we have, for any positive m ,*

$$SS_m(y, \bar{\rho}_n) = SS_m(y, \rho^*) + O_P(1), \text{ as } n \rightarrow \infty.$$

Lemma 3.17. *If $(y_0, \dots, y_n)$ is defined by (3.1) and $(z_0, \dots, z_n)$ is defined as in Lemma 3.7, then*

$$SS_m(y, \rho^*) = SS_m(z, \rho^*) + O_P(1), \text{ as } n \rightarrow \infty.$$

Lemma 3.18. *Let (X_n) and (Y_n) be two sequences of rv's such that $X_n - Y_n = O_P(1)$. If $Y_n^{-1} = O_P(n^{-1})$ then $X_n^{-1} = O_P(n^{-1})$.*

Proof of Lemma 3.16. Using the matrix notations from the proof of Lemma 3.9, we have

$$SS_m(y, \rho^*) = \min_{T, \delta} \|y - \rho^* B y - T \delta\|^2, \quad SS_m(y, \bar{\rho}_n) = \min_{T, \delta} \|y - \bar{\rho}_n B y - T \delta\|^2,$$

where all minimizations are achieved over all segmentations with m change points belonging to $\mathcal{A}_{n,m}$. Let us define $(\hat{T}^*, \hat{\delta}^*)$ and $(\bar{T}, \bar{\delta})$ by

$$(\hat{T}^*, \hat{\delta}^*) = \arg \min_{T, \delta} \|y - \rho^* B y - T \delta\|, \quad (\bar{T}, \bar{\delta}) = \arg \min_{T, \delta} \|y - \bar{\rho}_n B y - T \delta\|.$$

Note that $\hat{T}^*$ and $\bar{T}$ refer to $\hat{t}(y, \rho^*)$ and $\hat{t}(y, \bar{\rho}_n)$, respectively. We have

$$\begin{aligned} |SS_m(y, \bar{\rho}_n) - SS_m(y, \rho^*)| &= \left| \min_{T, \delta} \|y - \bar{\rho}_n B y - T \delta\|^2 - \min_{T, \delta} \|y - \rho^* B y - T \delta\|^2 \right| \\ &\leq \max \left(\|y - \bar{\rho}_n B y - \hat{T}^* \hat{\delta}^*\|^2 - \|y - \rho^* B y - \hat{T}^* \hat{\delta}^*\|^2, \right. \\ &\quad \left. \|y - \bar{\rho}_n B y - \bar{T} \bar{\delta}\|^2 - \|y - \rho^* B y - \bar{T} \bar{\delta}\|^2 \right). \end{aligned} \quad (3.76)$$

We now have to prove that this upper bound is $O_P(1)$. We first prove it for the second term of in the right-hand side of (3.76). To do so, observe that $\|y - \bar{\rho}_n B y - \bar{T} \bar{\delta}\|^2 = \|y - \rho^* B y - \bar{T} \bar{\delta} + (\rho^* - \bar{\rho}_n) B y\|^2$. Thus,

$$\|y - \bar{\rho}_n B y - \bar{T} \bar{\delta}\|^2 - \|y - \rho^* B y - \bar{T} \bar{\delta}\|^2 = (\bar{\rho}_n - \rho^*)^2 \|B y\|^2 + 2(\rho^* - \bar{\rho}_n) \langle B y, y - \rho^* B y - \bar{T} \bar{\delta} \rangle.$$

Since, by (3.28) and Lemma 3.7, $y - \rho^* B y - \bar{T} \bar{\delta} = \epsilon - \Delta^* + (T^* \delta^* - \bar{T} \bar{\delta}) = \epsilon - \Delta^* + T^* (\delta^* - \bar{\delta}) + (T^* - \bar{T}) \bar{\delta}$, where Δ^* is the n -dimensional vector with entries Δ_i^* , we get

$$\begin{aligned} &\|y - \bar{\rho}_n B y - \bar{T} \bar{\delta}\|^2 - \|y - \rho^* B y - \bar{T} \bar{\delta}\|^2 \\ &= (\bar{\rho}_n - \rho^*)^2 \|B y\|^2 + 2(\rho^* - \bar{\rho}_n) \left(\langle B y, \epsilon \rangle + \langle B y, T^* (\delta^* - \bar{\delta}) \rangle + \langle B y, (T^* - \bar{T}) \bar{\delta} \rangle - \langle B y, \Delta^* \rangle \right). \end{aligned} \quad (3.77)$$

Let us now prove that each term in the right-hand side of (3.77) is $O_P(1)$.

(a) Let us study the first term of (3.77). Using Lemma 3.1 and (3.13) we get that

$$(\bar{\rho}_n - \rho^*)^2 \|By\|^2 = O_P(1). \quad (3.78)$$

(b) Let us now study the second term of (3.77). Observe that $\langle By, \epsilon \rangle = \sum_{i=1}^n y_{i-1} \epsilon_i = \sum_{i=1}^n (y_{i-1} - \mathbb{E}(y_{i-1})) \epsilon_i + \sum_{i=1}^n \mathbb{E}(y_{i-1}) \epsilon_i$. By using the Central limit theorem for iid rv's and since there is a finite number of change-points, the second term is $O_P(\sqrt{n})$. As for the first term, since $(y_{i-1} - \mathbb{E}(y_{i-1}))$ is a causal AR(1) process, then by using the beginning of the proof of Proposition 8.10.1 of Brockwell and Davis (1991), we get that $\sum_{i=1}^n (y_{i-1} - \mathbb{E}(y_{i-1})) \epsilon_i = O_P(\sqrt{n})$. Thus,

$$\langle By, \epsilon \rangle = O_P(\sqrt{n}). \quad (3.79)$$

Furthermore, we have $\|T^*(\delta^* - \bar{\delta})\|^2 = \sum_{k=0}^m (t_{k+1}^* - t_k^*) (\delta_k^* - \bar{\delta}_k)^2$ where each term of the sum is $O_P(1)$, thanks to Proposition 3.3, and so is the sum. Now using Lemma 3.1 and the Cauchy-Schwarz inequality, we get

$$\langle By, T^*(\delta^* - \bar{\delta}) \rangle = O_P(\sqrt{n}). \quad (3.80)$$

The convergence rate of $\widehat{t}(y, \bar{\rho}_n)$ given in Proposition 3.3 ensures that, for any $\varepsilon > 0$ there exists a positive M such that each column of $(T^* - \bar{T})$ has at most M non-zero coefficients with probability greater than $1 - \varepsilon$. By using Proposition 3.3, we obtain that with probability greater than $1 - \varepsilon$

$$\|(T^* - \bar{T})\bar{\delta}\|^2 \leq M \sum_k \bar{\delta}_k^2 = 2M \sum_k (\bar{\delta}_k - \delta_k^*)^2 + 2M \sum_k \delta_k^{*2} \leq MM', \quad (3.81)$$

where M' is a positive constant. By the Cauchy-Schwarz inequality, (3.81) and Lemma 3.1, we get

$$\langle By, (T^* - \bar{T})\bar{\delta} \rangle = O_P(\sqrt{n}). \quad (3.82)$$

As Δ^* has only m non-zero entries, $\langle By, \Delta^* \rangle$ is the sum of m Gaussian rv's and is therefore $O_P(1)$.

Thus, combining (3.79), (3.80) and (3.82) with (3.13), we get

$$(\rho^* - \bar{\rho}_n) (\langle By, \epsilon \rangle + \langle By, T^*(\delta^* - \bar{\delta}) \rangle + \langle By, (T^* - \bar{T})\bar{\delta} \rangle - \langle By, \Delta^* \rangle) = O_P(1).$$

To complete the proof, we need to consider the first term of (3.76). As ρ^* satisfies the same assumptions as $\bar{\rho}_n$, using the same line of reasoning as for the second term holds so we get

$$\|y - \bar{\rho}_n By - \widehat{T}^* \widehat{\delta}^*\|^2 - \|y - \rho^* By - \widehat{T}^* \widehat{\delta}^*\|^2 = O_P(1).$$

□

Proof of Lemma 3.17. The proof follows the same line of reasoning as the proof of Lemma 3.16.

Let us define $(\widehat{T}^y, \widehat{\delta}^y)$ and $(\widehat{T}^z, \widehat{\delta}^z)$ by

$$(\widehat{T}^y, \widehat{\delta}^y) = \arg \min_{T, \delta} \|y - \rho^* B y - T \delta\|^2, \quad (\widehat{T}^z, \widehat{\delta}^z) = \arg \min_{T, \delta} \|z - \rho^* B z - T \delta\|^2.$$

We have

$$\begin{aligned} \|SS_m(y, \rho^*) - SS_m(z, \rho^*)\| &\leq \max \left(\|y - \rho^* B y - \widehat{T}^y \widehat{\delta}^y\|^2 - \|z - \rho^* B z - \widehat{T}^y \widehat{\delta}^y\|^2, \right. \\ &\quad \left. \|y - \rho^* B y - \widehat{T}^z \widehat{\delta}^z\|^2 - \|z - \rho^* B z - \widehat{T}^z \widehat{\delta}^z\|^2 \right). \end{aligned}$$

According to Lemma 3.7, we have $y - \rho^* B y = z - \rho^* B z - \Delta^*$ where $\Delta^* = (\Delta_i^*)$. As for the first term

$$\begin{aligned} \|y - \rho^* B y - \widehat{T}^y \widehat{\delta}^y\|^2 - \|z - \rho^* B z - \widehat{T}^y \widehat{\delta}^y\|^2 \\ = \|\Delta^*\|^2 - 2 \left(\langle \Delta^*, \epsilon \rangle + \langle \Delta^*, T^*(\delta^* - \widehat{\delta}^y) \rangle + \langle \Delta^*, (T^* - \widehat{T}^y) \widehat{\delta}^y \rangle \right), \end{aligned}$$

the first term of which is a constant and all other terms being $O_P(1)$, which can be proved following the same line as the proof of Lemma 3.16. The control of $\|y - \rho^* B y - \widehat{T}^z \widehat{\delta}^z\|^2 - \|z - \rho^* B z - \widehat{T}^z \widehat{\delta}^z\|^2$ follows the same lines. $\square$

Proof of Lemma 3.18. Observe that

$$X_n^{-1} = (Y_n + (X_n - Y_n))^{-1} = Y_n^{-1} (1 + Y_n^{-1}(X_n - Y_n))^{-1}.$$

Since, by assumption, $Y_n^{-1}(X_n - Y_n) = O_P(n^{-1})$, the terms inside the parentheses converges in probability to one. Thus, $(1 + Y_n^{-1}(X_n - Y_n))^{-1}$ is in particular $O_P(1)$ which concludes the proof. $\square$

Proof of Proposition 3.7. As for the proof of Proposition 3.6, denoting $\widehat{\tau}_k(y, \rho) = \widehat{t}_k(y, \rho)/n$, the decomposition (3.74) still holds, replacing z with y . Then, by Proposition 3.3, we have

$$\sum_{k=0}^m \log n_k(\widehat{t}(y, \bar{\rho}_n)) - \sum_{k=0}^m \log n_k(\widehat{t}(y, \rho^*)) = O_P(1).$$

For a process y under Model (3.1), we construct a process z under Model (3.9) using Lemma 3.7. The proof relies on the fact that y inherits some properties of z . Again, we note that

$$\log SS_m(y, \bar{\rho}_n) - \log SS_m(y, \rho^*) = R \left(\frac{SS_m(y, \bar{\rho}_n) - SS_m(y, \rho^*)}{SS_m(y, \rho^*)} \right).$$

Lemma 3.16 states that $SS_m(y, \bar{\rho}_n) - SS_m(y, \rho^*) = O_P(1)$. To conclude the proof we need to further show that $[SS_m(y, \rho^*)]^{-1} = O_P(n^{-1})$. We first show that $[SS_m(y, \rho^*) - SS_m(z, \rho^*)] = O_P(1)$ in Lemma 3.17 and, because $[SS_m(z, \rho^*)]^{-1} = O_P(n^{-1})$, we conclude using Lemma 3.18. $\square$

Chapter 4

A robust approach for estimating change-points in the mean of an AR(p) process

Contenu

4.1	Introduction	90
4.2	Robust estimation of the autoregression coefficients	92
4.3	Change-points and expectations estimation	93
4.4	Selecting the number of change-points	94
4.4.1	Criterion to select the number of change-points, the order of the autoregression being fixed	94
4.4.2	Heuristic to select both the number of changes and the autoregression order	95
4.5	Numerical experiments	96
4.5.1	Practical implementation	96
4.5.2	Simulation design	96
4.5.3	Quality criteria	97
4.5.4	Results	98
4.6	Proofs	99
4.6.1	Proof of Proposition 4.1	99
4.6.2	Proof of Proposition 4.2	100
4.6.3	Proof of Proposition 4.3	106
4.7	Tables and figures	109

Cet article a été soumis à une revue pour publication. Des remarques additionnelles à cet article sont présentées dans la section 5.1.2.

Résumé

Le problème qui nous intéresse est celui de l'estimation de ruptures multiples dans l'espérance d'un processus $AR(p)$. La prise en compte de la structure de dépendance nous interdit l'approche inférentielle du cas indépendant. En particulier, l'algorithme de programmation dynamique, qui est le seul algorithme permettant de calculer la segmentation optimale dans le cas indépendant, ne peut plus être utilisé. Nous proposons une méthode en deux étapes, basée sur l'estimation préliminaire des paramètres d'autorégression. Cette dernière est basée sur des techniques de statistique robuste car notre estimateur doit être robuste aux ruptures si nous souhaitons ne pas avoir à les estimer en premier lieu. Nous proposons alors de suivre l'approche inférentielle classique en injectant cet estimateur dans le critère utilisé pour l'estimation de la localisation des ruptures, ce qui est équivalent à décorréler la série à l'aide des paramètres d'autorégression estimés. Nous démontrons que ces estimateurs des instants de rupture et des espérances partagent les propriétés asymptotiques des estimateurs classiques dans le cadre indépendant. La même approche *plug-in* est alors utilisée pour approcher le critère BIC modifié, sélectionner le nombre de segments et pour proposer un critère BIC heuristique pour sélectionner simultanément le nombre de ruptures et l'ordre de l'autorégression. Finalement, nous montrons, dans la section simulations, que pour des séries de taille finie la prise en compte de la structure de dépendance améliore les performances statistiques des estimateurs des lieux de rupture et du nombre de ruptures sélectionné.

Abstract

We consider the problem of multiple change-point estimation in the mean of an $AR(p)$ process. Taking into account the dependence structure does not allow us to use the inference approach of the independent case. Especially, the dynamic programming algorithm giving the optimal solution in the independent case cannot be used anymore. We propose a two-step method, based on the preliminary estimation of the autoregression parameters. It is based on robust statistics techniques, since our estimator has to be robust to the change-points if we do not want to estimate them before. Then, we propose to follow the classical inference approach, by plugging this estimator in the criterion used for change-point estimation, which is equivalent to decorrelate the series using the estimated autoregression parameters. We show that the asymptotic properties of these change-point location and mean estimators are the same as those of the classical estimators in the independent framework. The same plug-in approach is then used to approximate the modified BIC and choose the number of segments, and to derive a heuristic BIC criterion to select both the number of changes and the order of the autoregression. Finally, we show, in the simulation section, that for finite sample size taking into account the dependence structure improves the statistical performance of the change-point estimators and of the selection criterion.

4.1 Introduction

Change-point detection problems arise in many fields, such as genomics (J. V. Braun and Müller, 1998; J. V. Braun, R. K. Braun, et al., 2000; Picard, Robin, et al., 2005), medical imaging (Lavielle, 2005), earth sciences (Williams, 2003; Gazeaux et al., 2013) or climate (Mestre, 2000; Lu et al., 2010). In many of these problems, the observations cannot be assumed to be independent.

In the literature, in the frequentist framework, there is two ways to deal with the dependence structure of series affected by multiple change-points:

- Apply the methodology of the independent case, and prove that asymptotic results are not affected by dependence under some conditions (Lavielle and Moulines, 2000; Lavielle, 1999).
- Consider that all the parameters of the model can change at each change-point (Bardet, Kengne, and Wintenberger, 2012). In fact, inference is performed by considering also possible changes in the spectrum of the process.

The parameters of the dependence structure are here assumed to be global (i.e. not depending on the segments) nuisance parameters that have to be estimated to perform the change-points inference.

In this paper, we consider the segmentation of an $\text{AR}(p)$ process with homogeneous autoregression coefficients:

$$y_i = \mu_k^* + \eta_i, \quad t_{n,k}^* + 1 \leq i \leq t_{n,k+1}^*, \quad 0 \leq k \leq m^*, \quad 1 \leq i \leq n, \quad (4.1)$$

where $(\eta_i)_{i \in \mathbb{Z}}$ is a zero-mean stationary $\text{AR}(p)$ process. That is, it is a stationary solution of

$$\forall i \in \mathbb{Z}, \eta_i - \sum_{r=1}^p \phi_r^* \eta_{i-r} = \epsilon_i, \quad (4.2)$$

where the ϵ_i 's are uncorrelated zero-mean rv's with variance σ^2 and $\Phi^* = (\phi_r^*)_{1 \leq r \leq p} \in \mathbb{R}^p$ is such that a stationary solution to (4.2) exists. In (4.1), we use the following conventions: $t_{n,0}^* = 0, t_{n,m^*+1}^* = n$.

Our aim is to propose a methodology for estimating both the change-point locations $\mathbf{t}_n^* = (t_{n,k}^*)_{1 \leq k \leq m^*}$ and the means $\boldsymbol{\mu}^* = (\mu_k^*)_{0 \leq k \leq m^*}$, accounting for the existence of Φ^* . Moreover, we propose a criterion to choose the number of change-points m^* .

In the sequel, we shall assume that there exists $\boldsymbol{\tau}^* = (\tau_k^*)_{0 \leq k \leq m+1}$ such that, for $0 \leq k \leq m+1$ $t_{n,k}^* = \lfloor n\tau_k^* \rfloor$, $\lfloor x \rfloor$ denoting the integer part of x . Consequently, $\tau_0^* = 0$ and $\tau_{m^*+1}^* = 1$.

A first idea is to use the maximum-likelihood approach to estimate the parameters of the model. However, maximizing the likelihood function, especially in the change-point location parameters $\boldsymbol{\tau}$, leads to a complex discrete optimization problem in an algorithmic point of view.

When the observations are independent, the optimal segmentation (e.g. in the maximum-likelihood sense) can be recovered via the dynamic programming (DP) algorithm introduced by Auger and Lawrence (1989). The computational complexity of this algorithm is quadratic relatively to the length of the series. This algorithm and some of its improvements (such as these proposed by Rigaiil, 2010; or Killick et al., 2012) are the only one that provide exactly the optimal change-point location estimators. However, the DP algorithm only applies when (i) the loss function (e.g. the negative log-likelihood) is additive with respect to the segments and when (ii) no parameter to be estimated is common to several segments.

In the autoregressive case, the likelihood function violates both requirements (i) and (ii). Indeed the log-likelihood is not additive with respect to the segments because of the dependence that exists between data from neighbor segments and the unknown coefficients Φ^* needs to be estimated jointly over all segments. Even if Φ^* was known, the DP principle would not apply to the log-likelihood of Model (4.1) as it will still not be additive. We introduce an alternative criterion, based on the quasi-likelihood described by Bardet, Kengne, and Wintenberger (2012). This criterion is equivalent to the classic least-squares applied to a decorrelated version of the series, computed with an estimated $\widehat{\Phi}$. To achieve this decorrelation, we shall provide an estimator of Φ^* .

We shall prove that, under mild asymptotic assumptions on the estimator of Φ^* , the resulting change-point estimators satisfy the same rate of convergence as those proposed by Lavielle and Moulines (2000) and Bardet, Kengne, and Wintenberger (2012).

We show that such an estimator of Φ^* exists and can be computed before segmenting the series. In order to estimate Φ^* , we first differentiate the series of observations (y_i) and work on $x_i = y_i - y_{i-1}$ which satisfies

$$x_i = \nu_i \text{ except for } i = t_{n,k}^* + 1, \text{ where } x_{t_{n,k}^*+1} = (\mu_k^* - \mu_{k-1}^*) + \nu_{t_{n,k}^*+1}, \quad 0 \leq k \leq m^*, \quad (4.3)$$

where $\nu_i = \eta_i - \eta_{i-1}$ is an ARMA($p,1$) defined from (η_i) by

$$\nu_i - \sum_{r=1}^p \phi_r^* \nu_{i-r} = \epsilon_i - \epsilon_{i-1} . \quad (4.4)$$

To this aim, we borrow techniques from robust estimation (Ma and Genton, 2000). Briefly speaking, we consider the data observed at the change-point locations as outliers and propose an estimator of Φ^* that is robust to the presence of such outliers. We shall prove that the estimator that we propose is consistent and asymptotically Gaussian. Moreover, we propose a model selection criterion on the number of changes, the order of the autoregression being fixed, inspired by the one proposed by Zhang and Siegmund (2007) and prove some asymptotic properties of this criterion. Finally, we discuss the problem of selecting jointly the number of changes and the order of the autoregression and propose a practical solution to this problem based on a Bayesian heuristic.

4.2 Robust estimation of the autoregression coefficients

The aim of this section is to provide an estimator of Φ^* which can deal with the presence of change-points in the data. In the absence of change-points ($m^* = 0$ in (4.1)), the estimation of Φ^* is a well-known issue (see Brockwell and Davis, 1991, for a comprehensive introduction) and a consistent, asymptotically Gaussian estimator is given by the Yule-Walker method. We aim to adapt this method to our framework.

Since change-points can be seen as outliers in the $\text{AR}(p)$ process, we shall propose a robust approach for estimating Φ^* . Our approach is based on the estimator of the autocorrelation function of a stationary time series proposed by Ma and Genton (2000) which is based on the robust scale estimator of Rousseeuw and Croux (1993). More precisely, let us define $\tilde{\Phi}_n^{(p)}$ by

$$\tilde{\Phi}_n^{(p)} = \tilde{R}_{n,p}^{-1} \tilde{\boldsymbol{\rho}}_{n,2:(p+1)}^T, \quad (4.5)$$

where for $i < j$ integers,

$$\tilde{\boldsymbol{\rho}}_{n,i:j} = (\tilde{\rho}_n(h))_{i \leq h \leq j}, \quad (4.6)$$

is an estimator of $\boldsymbol{\rho}_{i:j}$ defined by

$$\boldsymbol{\rho}_{i:j} = (\rho(h))_{i \leq h \leq j}, \quad (4.7)$$

where $\rho(h)$ denotes the autocorrelation of the process (x_i) defined in (4.3) at lag h , and $\cdot^T$ state for the transpose. In (4.5), $\tilde{R}_{n,p}$ denotes the following matrix

$$\tilde{R}_{n,p} = (\tilde{\rho}_n(j-i-1))_{1 \leq i,j \leq p}, \quad (4.8)$$

which is an estimator of

$$R_p = (\rho(j-i-1))_{1 \leq i,j \leq p}. \quad (4.9)$$

Moreover, for all h in $\mathbb{Z}$,

$$\tilde{\rho}_n(h) = \frac{Q_n^2(x_h^+) - Q_n^2(x_h^-)}{Q_n^2(x_h^+) + Q_n^2(x_h^-)}, \quad (4.10)$$

where $x_h^+ = (x_{i+h} + x_i)_{0 \leq i \leq n-h}$, $x_h^- = (x_{i+h} - x_i)_{0 \leq i \leq n-h}$ and Q_n is the scale estimator of Rousseeuw and Croux (1993) which is such that $Q_n(z)$ is proportional to the first quartile of $\{|z_i - z_j|; 0 \leq i < j \leq n\}$.

Proposition 4.1. *Let $\tilde{\boldsymbol{\rho}}_{n,i:j} = (\tilde{\rho}_n(h))_{i \leq h \leq j}$ and $\boldsymbol{\rho}_{i:j} = (\rho(h))_{i \leq h \leq j}$ be defined in (4.6), (4.7) and (4.10). Under the assumption that ν_i in (4.3) is Gaussian, we have*

$$n^{1/2} (\tilde{\boldsymbol{\rho}}_{n,1:(p+1)} - \boldsymbol{\rho}_{1:(p+1)}) \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, V) \quad (4.11)$$

in distribution, where $\mathcal{N}(0, V)$ is the $(p+1)$ -dimensional centered Gaussian distribution with covariance matrix V and

$$n^{1/2} (\tilde{\Phi}_n^{(p)} - \Phi^*) \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, M^T (R_p^{-1})^T V R_p^{-1} M), \quad (4.12)$$

in distribution, where

$$M = \left(\phi_{i-j+1}^* \mathbf{1}_{i \geq j} + \phi_{i+j+1}^* \mathbf{1}_{i+j \leq p-1} - \mathbf{1}_{j=i+1} \right)_{1 \leq i \leq p, 1 \leq j \leq p+1} . \quad (4.13)$$

The proof of Proposition 4.1 is given in Section 4.6.

Remark 4.1. Proposition 4.1 argues for the possibility to estimate robustly Φ^* . Equations (4.12) and (4.13) show that the variance of $\widehat{\Phi}$ depends on Φ^* . This variance may be large for some Φ^* . One can deal with this problem by preferring a regularized version of this estimator, that is

$$\left(\widetilde{R}_{n,p}^T \widetilde{R}_{n,p} + S_{n,p} \right)^{-1} \widetilde{R}_{n,p}^T \widetilde{\rho}_{n,2:(p+1)}^T , \quad (4.14)$$

where $S_{n,p}$ is a $p \times p$ positive semi-definite symmetric real matrix. If $S_{n,p} = o_P(n^{-1/2})$, then Proposition 4.1 remains true with $\widetilde{\Phi}_n^{(p)}$ being replaced by the expression of Equation (4.14).

4.3 Change-points and expectations estimation

In this section, the number of change-points m^* is assumed to be known. In the sequel, for notational simplicity, m^* will be denoted by m . Our goal is to estimate both the change-points and the means in Model (4.1). A first idea consists in maximizing the Gaussian quasi-likelihood conditioned on $y_0, \dots, y_{1-p}$ and to minimize it with respect to Φ . Due to a term that involves both μ_{k-1} and μ_k , this criterion cannot be efficiently minimized. Therefore, we propose to use an alternative criterion defined as follows:

$$SS_m(y, \Phi, \delta, t) = \sum_{k=0}^m \sum_{i=t_k+1}^{t_{k+1}} \left(y_i - \sum_{r=1}^p \phi_r y_{i-r} - \delta_k \right)^2 .$$

Note that we used the reparametrization $\delta_k = (1 - \sum_{r=1}^p \phi_r) \mu_k$ and that minimizing

$$SS_m \left(z, \Phi, \left(1 - \sum_{r=1}^p \phi_r \right) \mu, t \right)$$

is equivalent to maximizing the Gaussian log-likelihood, conditioned on $z_0, \dots, z_{1-p}$, of the following model maximized with respect to σ :

$$z_i - \mu_k^* = \sum_{r=1}^p \phi_r (z_{i-r} - \mu_k^*) + \epsilon_i, \quad t_{n,k}^* + 1 \leq i \leq t_{n,k+1}^*, \quad 0 \leq k \leq m, \quad 1 \leq i \leq n, \quad (4.15)$$

where the ϵ_i 's are defined as in Model (4.1). In this model, as in the models considered in Bardet, Kengne, and Wintenberger (2012), the expectation changes are not abrupt anymore as in Model (4.1).

Proposition 4.2. *Let $(\overline{\Phi}_n)$ be a sequence of rv's on $\mathbb{R}^p$ and $z = (z_{1-p}, \dots, z_n)$ a finite sequence of real-valued rv's satisfying (4.15). Let $\widehat{\delta}_n(z, \overline{\Phi}_n)$ and $\widehat{t}_n(z, \overline{\Phi}_n)$ be defined by*

$$\left(\widehat{\delta}_n(z, \overline{\Phi}_n), \widehat{t}_n(z, \overline{\Phi}_n) \right) = \arg \min_{(\delta, t) \in \mathbb{R}^{m+1} \times \mathcal{A}_{n,m}} SS_m(z, \overline{\Phi}_n, \delta, t) , \quad (4.16)$$

$$\widehat{\tau}_n(z, \overline{\Phi}_n) = \frac{1}{n} \widehat{t}_n(z, \overline{\Phi}_n), \quad (4.17)$$

where

$$\mathcal{A}_{n,m} = \{(t_0, \dots, t_{m+1}); t_0 = 0 < t_1 < \dots < t_m < t_{m+1} = n, \forall k = 1, \dots, m+1, t_k - t_{k-1} \geq \Delta_n\} \quad (4.18)$$

and where (Δ_n) is a real sequence such that $n^{-\alpha} \Delta_n \rightarrow \infty$, as $n \rightarrow \infty$ and $\alpha > 0$. Assume that

$$(\bar{\Phi}_n - \Phi^*) = O_P(n^{-1/2}), \quad (4.19)$$

as n tends to infinity. Then,

$$\|\widehat{\tau}_n(z, \bar{\Phi}_n) - \tau^*\| = O_P(n^{-1}), \quad \|\widehat{\delta}_n(z, \bar{\Phi}_n) - \delta^*\| = O_P(n^{-1/2}),$$

where $\|\cdot\|$ is the Euclidian norm and

$$\delta^* = (\delta_0^*, \dots, \delta_{m+1}^*) = \left(1 - \sum_{r=1}^p \phi_r\right) (\mu_0^*, \dots, \mu_{m+1}^*) .$$

Proposition 4.3. *The result of Proposition 4.2 still holds under the same assumptions when z is replaced with y satisfying (4.1).*

The proofs of Propositions 4.2 and 4.3 are given in Section 4.6.2 and 4.6.3, respectively. Note that the estimators defined in these propositions have the same asymptotic properties as those of the estimators proposed by Lavielle and Moulines (2000). Also, the estimator $\widetilde{\Phi}_n^{(p)}$ defined in Section 4.2 satisfies the same properties as $\bar{\Phi}_n$ and can thus be used in the criterion SS_m for providing consistent estimators of the change-points and of the means.

4.4 Selecting the number of change-points

4.4.1 Criterion to select the number of change-points, the order of the autoregression being fixed

In this section, we propose to adapt the modified Bayesian information criterion (mBIC, Zhang and Siegmund, 2007) to our autoregressive noise framework. mBIC was proposed to select the number m of change-points in the mean in the particular case of segmentation of an independent Gaussian process x . This criterion is derived from an $O_P(1)$ approximation of the Bayes factor between models with m and 0 change-points, respectively. Its performance for model selection is assessed by simulation studies (Zhang, 2005; Frick et al., 2014).

The mBIC selection procedure consists in choosing the number of change-points as:

$$\widehat{m} = \arg \max_m C_m(x, 0), \quad (4.20)$$

where the criterion $C_m(y, \Phi)$ is defined for a process y as

$$C_m(y, \Phi) = -\left(\frac{n-m+1}{2}\right) \log SS_m(y, \Phi) + \log \Gamma\left(\frac{n-m+1}{2}\right) - \frac{1}{2} \sum_{k=0}^m \log n_k(\widehat{\mathbf{t}}(y, \Phi)) - m \log n. \quad (4.21)$$

In the latter equation

$$SS_m(y, \Phi) = \min_{\delta, \mathbf{t}} SS_m(y, \Phi, \delta, \mathbf{t}), \quad (4.22)$$

where the minimization with respect to $\mathbf{t}$ is performed in $\mathcal{A}_{n,m}$ defined in (4.18), and

$$n_k(\widehat{\mathbf{t}}(y, \Phi)) = \widehat{t}_{k+1}(y, \Phi) - \widehat{t}_k(y, \Phi),$$

where $\widehat{\mathbf{t}}(y, \Phi) = (\widehat{t}_1(y, \Phi), \dots, \widehat{t}_m(y, \Phi))$ is defined as $\widehat{\mathbf{t}}(y, \Phi) = \arg \min_{\mathbf{t} \in \mathcal{A}_{n,m}} \min_{\delta} SS_m(y, \Phi, \delta, \mathbf{t})$.

Note that, in Model (4.15), the criterion could be directly applied to the decorrelated series $v^* = (v_i^*)_{1 \leq i \leq n} = (y_i - \sum_{r=1}^p \phi_r^* y_{i-r})_{1 \leq i \leq n}$ since

$$C_m(y, \Phi^*) = C_m(v^*, 0).$$

We propose to use the same selection criterion, replacing Φ^* by some relevant estimator $\overline{\Phi}_n$. The following two propositions show that this plug-in approach results in the same asymptotic properties under both Models (4.15) and (4.1).

Proposition 4.4. *For any positive m , for a process z satisfying (4.15) and under the assumptions of Proposition 4.2, we have*

$$C_m(z, \overline{\Phi}_n) = C_m(z, \Phi^*) + O_P(1), \text{ as } n \rightarrow \infty.$$

Proposition 4.5. *For any positive m , for a process y satisfying (4.1) and under the assumptions of Proposition 4.3, we have*

$$C_m(y, \overline{\Phi}_n) = C_m(y, \Phi^*) + O_P(1), \text{ as } n \rightarrow \infty.$$

The proofs of Propositions 4.4 and 4.5 are similar to Chakar et al. (2014, Propositions 6 and 7)¹.

In practice, we propose to take $\overline{\Phi}_n = \widetilde{\Phi}_n^{(p)}$ as defined in Section 4.2, which satisfies the conditions of Proposition 4.5 to estimate the number of segments by

$$\widehat{m} = \arg \max_{0 \leq m \leq m_{\max}} C_m(y, \widetilde{\Phi}_n^{(p)}), \quad (4.23)$$

where C_m is defined in (4.21), and for a given maximum number of changes $m_{\max}$.

4.4.2 Heuristic to select both the number of changes and the autoregression order

Applying Zhang (2005, Theorem 2.2)² to the series $v^* = (v_i^*)_{1 \leq i \leq n} = (z_i - \sum_{r=1}^p \phi_r^* z_{i-r})_{1 \leq i \leq n}$, for a process z satisfying (4.15) and with the corresponding priors, gives:

$$\log P(m|v^*) = C_m(v^*, 0) - C_0(v^*, 0) + \log P(m=0|v^*) + O_P(1),$$

¹See Propositions 3.6 and 3.7 of this document.

²See Theorem 2.11 of this document.

that is

$$P(m|z, p, \Phi) = C_m(z, \Phi) - C_0(v^*, 0) + \log P(m = 0|v^*) + O_P(1) ,$$

From Schwarz (1978), we approximate then $\log P(m, p|z)$, up to a constant, by

$$\log P\left(m \middle| z, p, \widehat{\Phi}_{ML}^{(p)}\right) - \frac{p}{2} \log n ,$$

and then, up to a constant, by

$$C_m\left(z, \widehat{\Phi}_{ML}^{(p)}\right) - \frac{p}{2} \log n ,$$

where $\widehat{\Phi}_{ML}^{(p)}$ is the maximum likelihood estimator of Φ^* in the model with m changes and autoregression order p . Replacing $\widehat{\Phi}_{ML}^{(p)}$ by $\widetilde{\Phi}_n^{(p)}$, and z by y satisfying (4.1), we propose to select m and p by

$$(\widehat{m}', \widehat{p}') = \arg \max_{0 \leq m \leq m_{\max}, 0 \leq p \leq p_{\max}} \left\{ C_m\left(y, \widetilde{\Phi}_n^{(p)}\right) - \frac{p}{2} \log n \right\} , \quad (4.24)$$

with given $m_{\max}$ and $p_{\max}$. Even if we do not aim to estimate the order p of the autoregression, this criterion is interesting by being more flexible than (4.23). Indeed, if at the true order p , $\widetilde{\Phi}_n^{(p)}$ provides a poor estimate of Φ^* , a different order (e.g. $p+1$) can lead to a better fitting of the model and then a better estimate of the number of changes.

4.5 Numerical experiments

4.5.1 Practical implementation

Our decorrelation procedure introduces spurious change-points in the series, at distance $\leq p$ of the true change-points. When the length of the series tends to infinity, the effect of these artefacts on estimates vanishes, but with a finite length series our procedure may be affected. To fix this point, we propose a post-processing to the estimated change-points $\widehat{\mathbf{t}}_n$, which consists in removing segments of length smaller than p :

$$PP(\widehat{\mathbf{t}}_n, p) = \left\{ \widehat{t}_{n,k} \in \widehat{\mathbf{t}}_n \right\} \setminus \left\{ \widehat{t}_{n,i} \text{ such that } \exists j, \widehat{t}_{n,i} - p \leq \widehat{t}_{n,j} < \widehat{t}_{n,i} \text{ and } (j = 1 \text{ or } \widehat{t}_{n,j-1} < \widehat{t}_{n,j} - p) \right\} .$$

4.5.2 Simulation design

We simulated Gaussian $AR(p)$ series $y_1, \dots, y_n$ of length $n = 7200$ and $n = 14400$, with $p = 2$ or $p = 5$. Additionally, the observations $y_{-19}, \dots, y_0$, are simulated and used for conditioning the quasi-likelihood. All series were affected by $m^* = 6$ change-points located at fractions $1/6 \pm 1/36, 3/6 \pm 2/36, 5/6 \pm 3/36$ of their length. The mean within each segment alternates between 0 and 1, starting with $\mu_1 = 0$. We considered autoregression parameters that verify the assumptions to get a stationary causal process, see Brockwell and Davis (1991, Theorem

3.1.1)³. We focused our attention on some parameters for which the computed estimators have a typical behavior and are interesting to illustrate our method.

For $p = 2$ the parameters are the following

$$(\phi_1^*, \phi_2^*, \sigma^*) \in \{(-1.2, -0.4, 0.4), (1.6, -0.8, 0.4), (0.2, 0.2, 0.4), (0.2, 0.6, 0.4), (0.4, 0.2, 0.2)\} .$$

For $p = 5$ the parameters are the following

$$(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) \in \{(0.5, 0, 0, 0.5, -0.5, 0.4), (0.5, 0, 0, 0, -0.5, 0.4)\} .$$

Each combination was replicated $S = 100$ times.

4.5.3 Quality criteria

To assess the quality of the estimation of the autoregression parameters, we used the Root-Mean-Square Errors (RMSE) of $\widehat{\Phi}_n^{(p)}$ defined in (4.5).

To study the performances of our proposed model selection criteria, we computed and compared:

- $\widehat{m}^0 = \arg \max_m C_m(y, 0)$, the criterion C_m without any decorrelation procedure,
- $\widehat{m}_Y^0$, the estimated number of changes derived from the BIC-type penalized criterion defined by Yao (1988),
- $\widehat{m}^* = \arg \max_m C_m(y, \Phi^*)$, the criterion C_m where the series is exactly decorrelated,
- $\widehat{m}_{PP}^*$, the post-processed number of changes, that is the number of changes of $PP(\widehat{\mathbf{t}}_n, p)$ if $\widehat{\mathbf{t}}_n$ is the vector of the estimated change-point locations obtained by the minimization (4.22) with $\Phi = \Phi^*$,
- $\widehat{m} = \arg \max_m C_m(y, \widehat{\Phi}_n^{(p)})$, and the result of post-processing $\widehat{m}_{PP}$,
- $(\widehat{m}', \widehat{p}')$ defined in (4.24). The post-processing, giving $\widehat{m}'_{PP}$, is performed with $PP(\cdot, \widehat{p}')$,

where C_m is defined in (4.21). We were particularly interested in the comparison of:

- $\widehat{m}_Y^0$ and $\widehat{m}^0$ to the other estimates to illustrate how much our method improves the estimation of the number of changes,
- $\widehat{m}^*$ and $\widehat{m}_{PP}^*$ are compared to $\widehat{m}$ and $\widehat{m}_{PP}$ to identify the errors coming from the estimation of Φ^* by $\widehat{\Phi}_n$,
- the post-processed estimates are compared to the non post-processed estimates to assess the usefulness of this finite-sample correction of our method.

³See Theorem 2.1 of this document.

In order to measure the performance of change-point location estimators, we plotted their frequencies. In particular we are interested in the following change-point estimates:

- $\hat{t}_n^0$, the minimizer on $\mathcal{A}_{n,\hat{m}^0}$ of $\min_{\delta} SS_{\hat{m}^0}(\mathbf{y}, 0, \delta, \cdot)$.
- $\hat{t}_{n,PP} = PP(\hat{t}_n, p)$, where $\hat{t}_n$ minimizes $\min_{\delta} SS_{\hat{m}}(\mathbf{y}, \tilde{\Phi}_n^{(p)}, \delta, \cdot)$ on $\mathcal{A}_{n,\hat{m}}$.
- $\hat{t}_{n,PP} = PP(\hat{t}_n, p')$, where $\hat{t}_n$ minimizes $\min_{\delta} SS_{\hat{m}'}(\mathbf{y}, \tilde{\Phi}_n^{(p')}, \delta, \cdot)$ on $\mathcal{A}_{n,\hat{m}'}$.

To highlight the peaks, we plotted also the Gaussian kernel density estimate. The bandwidth is selected following the method of Sheather and Jones (1991).

4.5.4 Results

For $p = 2$ the simulation results suggest that the decorrelation procedure is not necessary in all cases. If ϕ_1^* and ϕ_2^* are negative, $\hat{m}^0 = 6$ almost all the times (see Table 4.1 and Figures 4.1 and 4.2). If only one of the parameters is negative, problems can arise without decorrelation if the other parameter is positively large (see Table 4.2 and Figures 4.3 and 4.4). The core of the problems with $\hat{m}^0$ is located at $\phi_1^* > 0, \phi_2^* > 0$. In these cases, our decorrelation procedure is required (see Tables 4.3, 4.4, and 4.5, and Figures 4.5, 4.6, 4.7, 4.8, 4.9, and 4.10). $\hat{m}$ can provide poor estimates of m because of a poor preliminary estimate of Φ^* . However, in this case, $\hat{m}'$ can provide better estimates, thanks to an overestimation of p by $\hat{p}'$, as noticed in Section 4.4.2 (see Tables 4.4, 4.5, and Figures 4.6, 4.7, 4.8, 4.9, 4.10).

Tables 4.4 and 4.5 illustrate the usefulness of post-processing.

For $p = 5$ we can see from Tables 4.6 and 4.7 that our method, when p is known, has the same performance as the methodology which would have access to the autoregression parameters Φ^* , see the lines $\hat{m}$ and $\hat{m}^*$. These performances are not altered by the choice of p , even if it is overestimated by our model selection approach, see the lines $\hat{m}'$ and $\hat{p}'$. Moreover, our method outperforms a methodology which would ignore the dependence structure of the process, see the line $\hat{m}^0$ of these tables. In addition, our method is not only able to select the true number of change-points, whatever p , but also the true change-point positions, as displayed in Figures 4.12 and 4.14.

4.6 Proofs

4.6.1 Proof of Proposition 4.1

Since there is only a finite number of atypical values in the process (x_i) , Theorem 4 of Lévy-Leduc et al. (2011)⁴ still holds and gives that for all fixed $h \geq 1$:

$$\sqrt{n-h} \left(\frac{Q_n^2(x_h^+) - Q_n^2(x_h^-)}{4} - \gamma(h) \right) = \frac{1}{\sqrt{n-h}} \sum_{i=1}^{n-h} \psi(\nu_i, \nu_{i+h}) + o_P(1),$$

where γ denotes the autocovariance of (ν_i) and where for all x and y ,

$$\psi : (x, y) \mapsto \left\{ (\gamma(0) + \gamma(h)) \text{IF} \left(\frac{x+y}{\sqrt{2(\gamma(0) + \gamma(h))}}, Q, \Phi \right) - (\gamma(0) - \gamma(h)) \text{IF} \left(\frac{x-y}{\sqrt{2(\gamma(0) - \gamma(h))}}, Q, \Phi \right) \right\}.$$

In this equation IF is defined by

$$\text{IF}(x, Q, \Phi) = c(\Phi) \left(\frac{1/4 - \Phi(x + 1/c(\Phi)) + \Phi(x - 1/c(\Phi))}{\int_{\mathbb{R}} \phi(y) \phi(y + 1/c(\Phi)) dy} \right),$$

where $c(\Phi) = 1/(\sqrt{2}\Phi^{-1}(5/8)) \approx 2.21914$, and Φ is here the cumulative distribution function of the standard normal distribution. By Theorem 2 of Lévy-Leduc et al. (2011)⁵, we obtain that

$$\frac{Q_n^2(x_h^+) + Q_n^2(x_h^-)}{4} - \gamma(0) = o_p(1).$$

Let $\widehat{\gamma}(0) = (Q_n^2(x_h^+) + Q_n^2(x_h^-))/4$ then

$$\sqrt{n-h} (\widehat{\rho}(h) - \rho(h)) = \frac{\widehat{\gamma}(0)^{-1}}{\sqrt{n-h}} \sum_{i=1}^{n-h} \psi(\nu_i, \nu_{i+h}) + o_P(1),$$

In order to prove a central limit theorem for $\widehat{\rho}_{1:(p+1)}$, it is enough to prove by the Cramér-Wold device (Billingsley, 1995, Theorem 29.4), that for any a_k in $\mathbb{R}$:

$$\frac{\widehat{\gamma}(0)^{-1}}{\sqrt{n}} \sum_{i=1}^{n-(p+1)} \sum_{k=1}^{p+1} a_k \psi(\nu_i, \nu_{i+k})$$

converges in distribution to a centered Gaussian rv. By Lemma 13 of Lévy-Leduc et al. (2011), ψ is of Hermite rank⁶ 2. Hence, the Hermite rank of the linear combination of the ψ function is larger than 2. Thus, Condition (2.40) of Theorem 4 in Arcones (1994)⁷ is satisfied and the quantity $n^{-1/2} \sum_{i=1}^{n-(p+1)} \sum_{k=1}^{p+1} a_k \psi(\nu_i, \nu_{i+k})$ converges in distribution to a centered Gaussian rv with variance $\widetilde{\sigma}^2$ given by

$$\widetilde{\sigma}^2 = \mathbb{E} \left[\left(\sum_{k=1}^{p+1} a_k \psi(\nu_1, \nu_{k+1}) \right)^2 \right] + 2 \sum_{\ell \geq 1} \mathbb{E} \left[\left(\sum_{k=1}^{p+1} a_k \psi(\nu_1, \nu_{k+1}) \right) \left(\sum_{k=1}^{p+1} a_k \psi(\nu_{\ell+1}, \nu_{k+\ell+1}) \right) \right].$$

⁴See Theorem 2.8 of this document.

⁵See Theorem 2.8 of this document.

⁶See Definition 2.15 of this document.

⁷See Theorem 2.7 of this document.

By Slutsky's lemma, $\sqrt{n-h}(\widehat{\rho}(h) - \rho(h))$ converges in distribution to a centered Gaussian rv with variance $\gamma(0)\widetilde{\sigma}^2$. Since $(x_i - \mathbb{E}x_i)$ is an ARMA($p,1$) process, with autoregressive parameters $\phi_1^*, \dots, \phi_p^*$, we get, by Azencott and Dacunha-Castelle (1986, Chapter 11, Paragraph 2)⁸, that R_p , as defined in (4.9), is invertible, and

$$\Phi^* = R_p^{-1} \boldsymbol{\rho}_{2:(p+1)}^T,$$

where $\boldsymbol{\rho}_{2:(p+1)}$ is defined in (4.7).

Let $g : \mathcal{D} \subset \mathbb{R}^{p+1} \rightarrow \mathbb{R}^p$ defined by

$$g(u) = \left(u_{|j-i-1|} \mathbf{1}_{j-i-1 \neq 0} + \mathbf{1}_{j-i-1=0} \right)_{1 \leq i, j \leq p}^{-1} (u_2, \dots, u_{p+1})^T. \quad (4.25)$$

Hence,

$$\begin{aligned} \Phi^* &= g(\boldsymbol{\rho}_{1:(p+1)}), \\ \widetilde{\Phi}_n^{(p)} &= g(\widetilde{\boldsymbol{\rho}}_{n1:(p+1)}). \end{aligned}$$

By (4.11) and the delta method:

$$\sqrt{n} \left(\widetilde{\Phi}_n^{(p)} - \Phi^* \right) \xrightarrow{n \rightarrow \infty} \mathcal{N} \left(0, \nabla g(\boldsymbol{\rho}_{1:(p+1)})^T V \nabla g(\boldsymbol{\rho}_{1:(p+1)}) \right),$$

$\nabla g(\boldsymbol{\rho}_{1:(p+1)})$ being the Jacobian matrix of g in $\boldsymbol{\rho}_{1:(p+1)}$. Let us determine $\nabla g(u) \cdot h$ for all $u \in \mathcal{D}, h \in \mathbb{R}^{p+1}$. Using (4.25), we get

$$\left(h_{|j-i-1|} \mathbf{1}_{j-i-1 \neq 0} \right)_{1 \leq i, j \leq p} g(u) + \left(u_{|j-i-1|} \mathbf{1}_{j-i-1 \neq 0} + \mathbf{1}_{j-i-1=0} \right)_{1 \leq i, j \leq p} \nabla g(u) \cdot h = (h_2, \dots, h_{p+1})^T.$$

Applied to $u = \boldsymbol{\rho}_{1:(p+1)}$, we have

$$R_p \nabla g(\boldsymbol{\rho}_{1:(p+1)}) \cdot h = \left(h_{j+1} - \sum_{j=1}^i \phi_{i+1-j}^* h_j + \sum_{j=1}^{p-i-1} \phi_{i+j+1}^* h_j \right)_{1 \leq j \leq p},$$

and then $\nabla g(\boldsymbol{\rho}_{1:(p+1)}) = R_p^{-1} M$, where M is defined in (4.13).

4.6.2 Proof of Proposition 4.2

In the sequel, we need the following definitions, notations and remarks. Observe that (4.15) can be rewritten as follows:

$$z = \sum_{r=1}^p \phi_r^* B^r z + T(\mathbf{t}_n^*) \boldsymbol{\delta}^* + \epsilon, \quad (4.26)$$

where

$$z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}, \quad B^r z = \begin{pmatrix} z_{1-r} \\ \vdots \\ z_{n-r} \end{pmatrix}, \quad \boldsymbol{\delta}^* = \begin{pmatrix} \delta_0^* \\ \vdots \\ \delta_m^* \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad (4.27)$$

⁸See Proposition 2.3 of this document.

where $\delta_k^* = (1 - \sum_{r=1}^p \phi_r^*) \mu_k^*$, for $0 \leq k \leq m$, and $T(\mathbf{t})$ is an $n \times (m+1)$ matrix where the k th column is $(\underbrace{0, \dots, 0}_{t_{k-1}}, \underbrace{1, \dots, 1}_{t_k - t_{k-1}}, \underbrace{0, \dots, 0}_{n - t_k})^T$.

Let us define the exact and estimated decorrelated series by

$$w^* = z - \sum_{r=1}^p \phi_r^* B^r z, \quad (4.28)$$

$$\bar{w} = z - \sum_{r=1}^p \bar{\phi}_{r,n} B^r z. \quad (4.29)$$

where $\bar{\Phi}_n = (\phi_{r,n})_{1 \leq r \leq p}$.

For any vector subspace E of $\mathbb{R}^n$, let π_E denote the orthogonal projection of $\mathbb{R}^n$ on E . Let also $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^n$, $\langle \cdot, \cdot \rangle$ the canonical scalar product on $\mathbb{R}^n$ and $\|\cdot\|_\infty$ the sup norm.

For x a vector of $\mathbb{R}^n$ and $\mathbf{t} \in \mathcal{A}_{n,m}$, let

$$J_{n,m}(x, \mathbf{t}) = \frac{1}{n} \left(\|\pi_{E_{t_n^*}}(x)\|^2 - \|\pi_{E_t}(x)\|^2 \right), \quad (4.30)$$

written $J_n(x, \mathbf{t})$ in the sequel for notational simplicity. In (4.30), $E_{t_n^*}$ and E_t correspond to the linear subspaces of $\mathbb{R}^n$ generated by the columns of $T(\mathbf{t}_n^*)$ and $T(\mathbf{t})$, respectively. We shall use the same decomposition as the one introduced in Lavielle and Moulines (2000):

$$J_n(x, \mathbf{t}) = K_n(x, \mathbf{t}) + V_n(x, \mathbf{t}) + W_n(x, \mathbf{t}), \quad (4.31)$$

where

$$\begin{aligned} K_n(x, \mathbf{t}) &= \frac{1}{n} \left\| \left(\pi_{E_{t_n^*}} - \pi_{E_t} \right) \mathbb{E}x \right\|^2, \\ V_n(x, \mathbf{t}) &= \frac{1}{n} \left(\left\| \pi_{E_{t_n^*}}(x - \mathbb{E}x) \right\|^2 - \left\| \pi_{E_t}(x - \mathbb{E}x) \right\|^2 \right), \\ W_n(x, \mathbf{t}) &= \frac{2}{n} \left(\left\langle \pi_{E_{t_n^*}}(x - \mathbb{E}x), \pi_{E_{t_n^*}}(\mathbb{E}x) \right\rangle - \left\langle \pi_{E_t}(x - \mathbb{E}x), \pi_{E_t}(\mathbb{E}x) \right\rangle \right). \end{aligned}$$

We shall also use the following notations:

$$\underline{\lambda} = \min_{1 \leq k \leq m} |\delta_k^* - \delta_{k-1}^*|, \quad (4.32)$$

$$\bar{\lambda} = \max_{1 \leq k \leq m} |\delta_k^* - \delta_{k-1}^*|, \quad (4.33)$$

$$\Delta_{\tau^*} = \min_{1 \leq k \leq m+1} (\tau_k^* - \tau_{k-1}^*), \quad (4.34)$$

$$\mathcal{C}_{\alpha, \gamma, n} = \left\{ \mathbf{t} \in \mathcal{A}_{n,m}; \alpha \underline{\lambda}^{-2} \leq \|\mathbf{t} - \mathbf{t}_n^*\| \leq n \gamma \Delta_{\tau^*} \right\}, \quad (4.35)$$

$$\mathcal{C}'_{\alpha, \gamma, n} = \mathcal{C}_{\alpha, \gamma, n} \cap \left\{ \mathbf{t} \in \mathcal{A}_{n,m}; \forall k = 1, \dots, m, t_k \geq t_{n,k}^* \right\}, \quad (4.36)$$

$$\begin{aligned} \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I}) &= \left\{ \mathbf{t} \in \mathcal{C}'_{\alpha, \gamma, n}; \right. \\ &\quad \left. \forall k \in \mathcal{I}, \alpha \underline{\lambda}^{-2} \leq t_k - t_{n,k}^* \leq n \gamma \Delta_{\tau^*} \text{ and } \forall k \notin \mathcal{I}, t_k - t_{n,k}^* < \alpha \underline{\lambda}^{-2} \right\}, \end{aligned} \quad (4.37)$$

for any $\alpha > 0$, $0 < \gamma < 1/2$ and $\mathcal{I} \subset \{1, \dots, m\}$. We shall also need the following lemmas in order to prove Proposition 4.2 which are proved below.

Lemma 4.1. Let $(z_{1-p}, \dots, z_n)$ be defined by (4.1) or (4.15), then, for all $r = 0, \dots, p$:

$$\|B^r z\| = O_P(n^{1/2}) ,$$

as n tends to infinity, where $B^r z$ is defined in (4.27).

Lemma 4.2. Let $(z_{1-p}, \dots, z_n)$ be defined by (4.1) or (4.15) then, for all $\mathbf{t} \in \mathcal{A}_{n,m}$,

$$|J_n(\bar{w}, \mathbf{t}) - J_n(w^*, \mathbf{t})| \leq \frac{2}{n} \sum_{r=1}^p |\phi_r^* - \bar{\phi}_{r,n}| \|B^r z\| (p|\phi_r^* - \bar{\phi}_{r,n}| \|B^r z\| + 2\|w^*\|) = O_P(n^{-1/2}) ,$$

where J_n is defined in (4.30), Bz and z are defined in (4.27), w^* is defined in (4.28) and $\bar{w}$ is defined in (4.29).

Lemma 4.3. Under the assumptions of Proposition 4.2, $\|\bar{\tau}_n - \tau^*\|_\infty$ converges in probability to 0, as n tends to infinity.

Lemma 4.4. Under the assumptions of Proposition 4.2 and for any $\alpha > 0$, $0 < \gamma < 1/2$ and $\mathcal{I} \subset \{1, \dots, m\}$,

$$P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} \left(\frac{1}{2}K_n(w^*, \mathbf{t}) + V_n(w^*, \mathbf{t}) + W_n(w^*, \mathbf{t})\right) \leq 0\right) \rightarrow 0, \text{ as } n \rightarrow \infty ,$$

where $\mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})$ is defined in (4.37) and w^* is defined in (4.28).

Lemma 4.5. Under the assumptions of Proposition 4.2 and for any $\alpha > 0$, $0 < \gamma < 1/2$ and $\mathcal{I} \subset \{1, \dots, m\}$,

$$P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} J_n(\bar{w}, \mathbf{t}) \leq 0\right) \rightarrow 0, \text{ as } n \rightarrow \infty ,$$

where $\mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})$ is defined in (4.37) and $\bar{w}$ is defined in (4.29).

Lemma 4.6. Under the assumptions of Proposition 4.2,

$$\|\hat{\tau}_n(z, \bar{\Phi}_n) - \tau^*\|_\infty = O_P(n^{-1}) .$$

Proof of Lemma 4.1. Without loss of generality, assume $(z_{1-p}, \dots, z_n)$ is defined by (4.15). $\|B^r z\|^2 = \sum_{i=1}^{n-r} z_i^2$ then Markov inequality implies that $\|B^r z\|^2 = O_P(n)$. $\square$

Proof of Lemma 4.2. By (4.28), $\bar{w} = w^* + \sum_{r=1}^p (\phi_r^* - \bar{\phi}_{r,n}) B^r z$. We get

$$\begin{aligned} \|\pi_{E_t}(\bar{w})\|^2 - \|\pi_{E_t}(w^*)\|^2 &= \left\| \pi_{E_t}(w^*) + \sum_{r=1}^p (\phi_r^* - \bar{\phi}_{r,n}) \pi_{E_t}(B^r z) \right\|^2 - \|\pi_{E_t}(w^*)\|^2 \\ &= \left\| \sum_{r=1}^p (\phi_r^* - \bar{\phi}_{r,n}) \pi_{E_t}(B^r z) \right\|^2 + 2 \sum_{r=1}^p (\phi_r^* - \bar{\phi}_{r,n}) \langle \pi_{E_t}(w^*), \pi_{E_t}(B^r z) \rangle \\ &\leq \sum_{r=1}^p p(\phi_r^* - \bar{\phi}_{r,n})^2 \|\pi_{E_t}(B^r z)\|^2 + 2 \sum_{r=1}^p (\phi_r^* - \bar{\phi}_{r,n}) \langle \pi_{E_t}(w^*), \pi_{E_t}(B^r z) \rangle \\ &\leq \sum_{r=1}^p (\phi_r^* - \bar{\phi}_{r,n}) (p(\phi_r^* - \bar{\phi}_{r,n}) \|\pi_{E_t}(B^r z)\|^2 + 2 \langle \pi_{E_t}(w^*), \pi_{E_t}(B^r z) \rangle) \\ &\leq \sum_{r=1}^p (\phi_r^* - \bar{\phi}_{r,n}) \langle \pi_{E_t}(B^r z), p(\phi_r^* - \bar{\phi}_{r,n}) \pi_{E_t}(B^r z) + 2\pi_{E_t}(w^*) \rangle . \end{aligned}$$

The Cauchy-Schwarz inequality and the 1-Lipschitz property of projections give

$$\left| \|\pi_{E_t}(\bar{w})\|^2 - \|\pi_{E_t}(w^*)\|^2 \right| \leq \sum_{r=1}^p |\phi_r^* - \bar{\phi}_{r,n}| \|B^r z\| (p |\phi_r^* - \bar{\phi}_{r,n}| \|B^r z\| + 2 \|w^*\|)$$

The conclusion follows from (4.30), (4.19) and Lemma 4.1. $\square$

Proof of Lemma 4.3. Lavielle and Moulines (2000, proof of Theorem 3) give the following bounds for any $\mathbf{t} \in \mathcal{A}_{n,m}$:

$$K_n(w^*, \mathbf{t}) \geq \underline{\lambda}^2 \min \left(\frac{1}{n} \max_{1 \leq k \leq m} |t_k - t_{n,k}^*|, \Delta_{\tau^*} \right), \quad (4.38)$$

$$V_n(w^*, \mathbf{t}) \geq -\frac{2(m+1)}{n\Delta_n} \left(\max_{1 \leq s \leq n} \left(\sum_{i=1}^s \epsilon_i \right)^2 + \max_{1 \leq s \leq n} \left(\sum_{i=n-s}^n \epsilon_i \right)^2 \right), \quad (4.39)$$

$$|W_n(w^*, \mathbf{t})| \leq \frac{3(m+1)^2 \bar{\lambda}}{n} \left(\max_{1 \leq s \leq n} \left| \sum_{i=1}^s \epsilon_i \right| + \max_{1 \leq s \leq n} \left| \sum_{i=n-s}^n \epsilon_i \right| \right), \quad (4.40)$$

where $\underline{\lambda}$, $\bar{\lambda}$ and Δ_{τ^*} are defined in (4.32–4.34). For any $\alpha > 0$, define, as in the proof of Theorem 3 of Lavielle and Moulines (2000),

$$\mathcal{C}_{n,\alpha} = \{\mathbf{t} \in \mathcal{A}_{n,m}; \|\mathbf{t} - \mathbf{t}_n^*\|_\infty \geq n\alpha\}. \quad (4.41)$$

For $0 < \alpha < \Delta_{\tau^*}$, we have:

$$\begin{aligned} P(\|\widehat{\mathbf{t}}_n(z, \bar{\Phi}_n) - \mathbf{t}_n^*\|_\infty \geq n\alpha) &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} J_n(\bar{w}, \mathbf{t}) \leq 0\right) \\ &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} (J_n(\bar{w}, \mathbf{t}) - J_n(w^*, \mathbf{t})) \leq -\alpha \underline{\lambda}^2\right) \\ &+ P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} (V_n(w^*, \mathbf{t}) + W_n(w^*, \mathbf{t})) \leq -\alpha \underline{\lambda}^2\right) \\ &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} (J_n(\bar{w}, \mathbf{t}) - J_n(w^*, \mathbf{t})) \leq -\alpha \underline{\lambda}^2\right) \\ &+ P\left(\max_{1 \leq s \leq n} \left(\sum_{i=1}^s \epsilon_i \right)^2 + \max_{1 \leq s \leq n} \left(\sum_{i=n-s}^n \epsilon_i \right)^2 \geq c \underline{\lambda}^2 n \Delta_n \alpha\right) \\ &+ P\left(\max_{1 \leq s \leq n} \left| \sum_{i=1}^s \epsilon_i \right| + \max_{1 \leq s \leq n} \left| \sum_{i=n-s}^n \epsilon_i \right| \geq c \underline{\lambda}^2 n \alpha \bar{\lambda}^{-1}\right) \end{aligned}$$

for some positive constant c . The last two terms of this sum go to 0 when n goes to infinity (Lavielle and Moulines, 2000, proof of Theorem 3). To show that the first term shares the same property, it suffices to show that $J_n(\bar{w}, \mathbf{t}) - J_n(w^*, \mathbf{t})$ is bounded uniformly in $\mathbf{t}$ by a sequence of rv's which converges to 0 in probability. This result holds by Lemma 4.2. $\square$

Proof of Lemma 4.4. Using Lavielle and Moulines (2000, Equations (64–66)), one can show the bound (73) of Lavielle and Moulines (2000) on

$$P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\alpha, \gamma, n}(I)} (K_n(w^*, \mathbf{t}) + V_n(w^*, \mathbf{t}) + W_n(w^*, \mathbf{t})) \leq 0\right).$$

Using the same arguments, we have the same bound on

$$P\left(\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} \left(\frac{1}{2}K_n(w^*, t) + V_n(w^*, t) + W_n(w^*, t)\right) \leq 0\right).$$

We conclude using Lavielle and Moulines (2000, Equations (67–71)). $\square$

Proof of Lemma 4.5. By (4.31),

$$\begin{aligned} P\left(\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} J_n(\bar{w}, t) \leq 0\right) &\leq P\left(\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} \left(J_n(\bar{w}, t) - J_n(w^*, t) + \frac{1}{2}K_n(w^*, t)\right) \leq 0\right) \\ &\quad + P\left(\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} \left(\frac{1}{2}K_n(w^*, t) + V_n(w^*, t) + W_n(w^*, t)\right) \leq 0\right). \end{aligned}$$

By Lemma 4.4, the conclusion thus follows if

$$P\left(\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} \left(J_n(\bar{w}, t) - J_n(w^*, t) + \frac{1}{2}K_n(w^*, t)\right) \leq 0\right) \longrightarrow 0, \text{ as } n \rightarrow \infty.$$

Since $\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} K_n(w^*, t) \geq (1 - \gamma) \Delta_{\tau^*} \alpha$ (Lavielle and Moulines, 2000, Equation (65)),

$$\begin{aligned} &P\left(\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} \left(J_n(\bar{w}, t) - J_n(w^*, t) + \frac{1}{2}K_n(w^*, t)\right) \leq 0\right) \\ &\leq P\left(\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} (J_n(\bar{w}, t) - J_n(w^*, t)) \leq \frac{1}{2}(\gamma - 1) \Delta_{\tau^*} \alpha\right), \end{aligned}$$

and we conclude by Lemma 4.2. $\square$

Proof of Lemma 4.6. For notational simplicity, $\widehat{\mathbf{t}}_n(z, \bar{\Phi}_n)$ will be replaced by $\bar{\mathbf{t}}_n$ in this proof. Since for any $\alpha > 0$,

$$P\left(\|\bar{\mathbf{t}}_n - \mathbf{t}_n^*\|_\infty < \alpha \lambda^{-2}\right) = P(\|\bar{\mathbf{t}}_n - \mathbf{t}_n^*\|_\infty \leq n\gamma \Delta_{\tau^*}) - P(\bar{\mathbf{t}}_n \in C_{\alpha, \gamma, n}),$$

it is enough, by Lemma 4.3, to prove that

$$P\left(\bar{\mathbf{t}}_n \in C_{\alpha, \gamma, n}\right) \longrightarrow 0, \text{ as } n \rightarrow \infty,$$

for all $\alpha > 0$ and $0 < \gamma < 1/2$. Since $\mathcal{C}_{\alpha, \gamma, n} = \bigcup_{\mathcal{I} \subset \{1, \dots, m\}} \mathcal{C}_{\alpha, \gamma, n} \cap \{\mathbf{t} \in \mathcal{A}_{n, m}; \forall k \in \mathcal{I}, t_k \geq t_{n, k}^*\}$, we shall only study one set in the union without loss of generality and prove that

$$P\left(\bar{\mathbf{t}}_n \in \mathcal{C}'_{\alpha, \gamma, n}\right) \longrightarrow 0, \text{ as } n \rightarrow \infty,$$

where $\mathcal{C}'_{\alpha, \gamma, n}$ is defined in (4.36). Since $\mathcal{C}'_{\alpha, \gamma, n} = \bigcup_{\mathcal{I} \subset \{1, \dots, m\}} \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})$, we shall only study one set in the union without loss of generality and prove that

$$P\left(\bar{\mathbf{t}}_n \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})\right) \longrightarrow 0, \text{ as } n \rightarrow \infty.$$

Since

$$P\left(\bar{\mathbf{t}}_n \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})\right) \leq P\left(\min_{t \in \mathcal{C}'_{\alpha, \gamma, n}(\mathcal{I})} J_n(\bar{w}, t) \leq 0\right),$$

the proof is complete by Lemma 4.5. $\square$

Proof of Proposition 4.2. For notational simplicity, $\widehat{\delta}_n(z, \overline{\Phi}_n)$ will be replaced by $\overline{\delta}_n$ in this proof. By Lemma 4.6, the last result to show is

$$\|\overline{\delta}_n - \delta^\star\| = O_P(n^{-1/2}),$$

that is, for all k , $\overline{\delta}_{n,k} - \delta_k^\star = O_P(n^{-1/2})$. By (4.28) and (4.29),

$$\overline{\delta}_{n,k} = \frac{1}{\overline{t}_{n,k+1} - \overline{t}_{n,k}} \sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} \overline{w}_i = \frac{1}{n(\overline{\tau}_{n,k+1} - \overline{\tau}_{n,k})} \left(\sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} w_i^\star + \sum_{r=1}^p (\phi_r^\star - \overline{\phi}_{r,n}) \sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} z_{i-r} \right).$$

By the Cauchy-Schwarz inequality,

$$\left| \sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} z_{i-r} \right| \leq (\overline{t}_{n,k+1} - \overline{t}_{n,k})^{1/2} (z_{\overline{t}_{n,k+1}-r}^2 + \dots + z_{\overline{t}_{n,k+1}-r}^2)^{1/2} \leq n^{1/2} \|Bz\| = O_P(n),$$

where the last equality comes from Lemma 4.1. Hence by (4.19) and Lemma 4.6,

$$\begin{aligned} \overline{\delta}_{n,k} &= \frac{1}{n(\overline{\tau}_{n,k+1} - \overline{\tau}_{n,k})} \sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} w_i^\star + O_P(n^{-1/2}) \\ &= \frac{1}{n(\overline{\tau}_{n,k+1} - \overline{\tau}_{n,k})} \left(\sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} \mathbb{E} w_i^\star + \sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} \epsilon_i \right) + O_P(n^{-1/2}), \end{aligned}$$

where the last equality comes from (4.26) and (4.28).

Let us now prove that

$$\frac{1}{n(\overline{\tau}_{n,k+1} - \overline{\tau}_{n,k})} \sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} \epsilon_i = O_P(n^{-1/2}). \quad (4.42)$$

By Lemma 4.3, $n^{-1}(\overline{\tau}_{n,k+1} - \overline{\tau}_{n,k})^{-1} = O_P(n^{-1})$. Moreover,

$$\sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} \epsilon_i = \sum_{i=t_{n,k}^\star+1}^{t_{n,k+1}^\star} \epsilon_i \pm \sum_{i=\overline{t}_{n,k+1}}^{t_{n,k}^\star} \epsilon_i \pm \sum_{i=t_{n,k+1}^\star+1}^{\overline{t}_{n,k+1}+1} \epsilon_i. \quad (4.43)$$

By the Central limit theorem, the first term in the right-hand side of (4.43) is $O_P(n^{1/2})$.

By using the Cauchy-Schwarz inequality, we get that the second term of (4.43) satisfies: $|\sum_{i=\overline{t}_{n,k+1}}^{t_{n,k}^\star} \epsilon_i| \leq |t_{n,k}^\star - \overline{t}_{n,k}|^{1/2} (\sum_{i=1}^n \epsilon_i^2)^{1/2} = O_P(1) O_P(n^{1/2}) = O_P(n^{1/2})$, by Lemma 4.6. The same holds for the last term in the right-hand side of (4.43), which gives (4.42).

Hence,

$$\begin{aligned} \overline{\delta}_{n,k} - \delta_k^\star &= \frac{1}{n(\overline{\tau}_{n,k+1} - \overline{\tau}_{n,k})} \sum_{i=\overline{t}_{n,k+1}}^{\overline{t}_{n,k+1}} (\mathbb{E} w_i^\star - \delta_k^\star) + O_P(n^{-1/2}) \\ &= \frac{1}{n(\overline{\tau}_{n,k+1} - \overline{\tau}_{n,k})} \sum_{i \in \{\overline{t}_{n,k+1}, \dots, \overline{t}_{n,k+1}\} \setminus \{t_{n,k}^\star+1, \dots, t_{n,k+1}^\star\}} (\mathbb{E} w_i^\star - \delta_k^\star) + O_P(n^{-1/2}), \end{aligned}$$

and then

$$\begin{aligned} |\overline{\delta}_{n,k} - \delta_k^\star| &\leq \frac{1}{n(\overline{\tau}_{n,k+1} - \overline{\tau}_{n,k})} \# \{ \overline{t}_{n,k+1}, \dots, \overline{t}_{n,k+1} \} \setminus \{ t_{n,k}^\star+1, \dots, t_{n,k+1}^\star \} \max_{l=0, \dots, m} |\delta_l^\star - \delta_k^\star| \\ &\quad + O_P(n^{-1/2}). \end{aligned}$$

We conclude by using Lemma 4.6 to get $\#\{\bar{t}_{n,k} + 1, \dots, \bar{t}_{n,k+1}\} \setminus \{t_{n,k}^* + 1, \dots, t_{n,k+1}^*\} = O_P(1)$ and Lemma 4.3 to get $(\bar{\tau}_{n,k+1} - \bar{\tau}_{n,k})^{-1} = O_P(1)$. $\square$

4.6.3 Proof of Proposition 4.3

The connection between Models (4.1) and (4.15) is made by the following lemmas.

Lemma 4.7. *Let $(y_0, \dots, y_n)$ be defined by (4.1) and let*

$$v_i^* = y_i - \sum_{r=1}^p \phi_r^* y_{i-r}, \quad (4.44)$$

$$\Delta_i^* = \begin{cases} -(\mu_k^* - \mu_{k-1}^*) \sum_{s=r}^p \phi_s^* & \text{if } i = t_{n,k}^* + r \text{ and } 1 \leq r \leq p \\ 0, & \text{otherwise,} \end{cases} \quad (4.45)$$

where the μ_k^* 's are defined in (4.1), then the process

$$w_i^* = v_i^* + \Delta_i^* \quad (4.46)$$

equals $z_i - \sum_{r=1}^p z_{i-r}$ where $(z_{1-p}, \dots, z_n)$ verify (4.15). Such a process $(z_{1-p}, \dots, z_n)$ can be constructed recursively as

$$\begin{cases} z_i = y_i & \text{for } 1-p \leq i \leq 0 \\ z_i = w_i^* + \sum_{r=1}^p \phi_r^* z_{i-r} & \text{for } i > 0. \end{cases} \quad (4.47)$$

Lemma 4.8. *Let $(y_{1-p}, \dots, y_n)$ be defined by (4.1) and let z be defined by (4.44–4.47). Then*

$$\bar{w}_i = \bar{v}_i + \bar{\Delta}_i \quad (4.48)$$

where

$$\bar{v}_i = y_i - \sum_{r=1}^p \bar{\phi}_{r,n} y_{i-r} \quad (4.49)$$

$$\bar{w}_i = z_i - \sum_{r=1}^p \bar{\phi}_{r,n} z_{i-r} \quad (4.50)$$

$$\bar{\Delta}_i = \Delta_i^* + \sum_{r=1}^p (\phi_r^* - \bar{\phi}_{r,n}) (z_{i-r} - y_{i-r}). \quad (4.51)$$

Lemma 4.9. *Let $\bar{\Delta} = (\bar{\Delta}_i)_{0 \leq i \leq n}$ as defined in (4.51). Then $\|\bar{\Delta}\| = O_P(1)$.*

Proof of Lemma 4.7. Let z being defined by (4.47). Using (4.46), we get, for all $0 \leq k \leq m, t_{n,k}^* < i \leq t_{n,k+1}^*$

$$\begin{aligned} (z_i - \mu_k^*) - \sum_{r=1}^p \phi_r^* (z_{i-r} - \mu_k^*) &= (y_i - \mu_k^*) - \sum_{r=1}^p \phi_r^* (y_{i-r} - \mu_k^*) + \Delta_i^* \\ &= (y_i - \mu_k^*) - \sum_{r=1}^p \phi_r^* \left(y_{i-1} - (\mu_{k-1}^* \mathbf{1}_{r \geq i-t_{n,k}^*} + \mu_k^* \mathbf{1}_{r < i-t_{n,k}^*}) \right) \end{aligned}$$

This expression equals $(y_i - \mathbb{E}(y_i)) - \sum_{r=1}^p \phi_r^* (y_{i-r} - \mathbb{E}(y_{i-r})) = \eta_i - \sum_{r=1}^p \phi_r^* \eta_{i-r} = \epsilon_i$ by (4.1) and (4.2). Then z satisfies (4.15). $\square$

The proof of Lemma 4.8 is straightforward.

Proof of Lemma 4.9. (4.51) can be written as

$$\overline{\Delta} = \Delta^* + \sum_{r=1}^p (\phi_r^* - \overline{\phi}_{r,n}) (B^r y - B^r z)$$

where $\Delta^* = (\Delta_i^*)_{1 \leq i \leq n}$, $B^r y = (y_{i-r})_{1 \leq i \leq n}$ and $B^r z$ is defined in (4.27). By the triangle inequality,

$$\|\overline{\Delta}\| \leq \|\Delta^*\| + \sum_{r=1}^p |\phi_r^* - \overline{\phi}_{r,n}| (\|B^r y\| + \|B^r z\|). \quad (4.52)$$

Since $\|\Delta^*\|$ is constant in n it is bounded. The conclusion follows from (4.52), (4.19) and Lemma 4.1. $\square$

Proof of Proposition 4.3. Let $y, z, \bar{v}, \bar{w}$ and $\overline{\Delta}$ be defined in Lemma 4.8.

Using (4.30) and Lemma 4.8, we get

$$J_n(\bar{v}, \mathbf{t}) = J_n(\bar{w}, \mathbf{t}) + J_n(\overline{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right).$$

By the Cauchy-Schwarz inequality and the 1-Lipschitz property of projections, we have

$$\begin{aligned} |J_n(\overline{\Delta}, \mathbf{t})| &\leq 2\|\overline{\Delta}\|^2, \\ \left| \left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right| &\leq 2\|\overline{\Delta}\|\|\bar{w}\|. \end{aligned}$$

Note that $\bar{w} = z - \sum_{r=1}^p \overline{\phi}_{r,n} B^r z$ thus by the triangle inequality

$$\|\bar{w}\| \leq \|z\| + \sum_{r=1}^p |\overline{\phi}_{r,n}| \|B^r z\|.$$

Since $|\overline{\phi}_{r,n}| = O_P(1)$ for all $1 \leq r \leq p$, we deduce from Lemma 4.1 that $\|\bar{w}\| = O_P(n^{1/2})$.

Since, by Lemma 4.9, $\|\overline{\Delta}\| = O_P(1)$, we obtain that

$$\sup_{\mathbf{t}} \left\{ J_n(\overline{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\left\langle \pi_{E_{t_n}^*}(\bar{w}), \pi_{E_{t_n}^*}(\overline{\Delta}) \right\rangle - \left\langle \pi_{E_t}(\bar{w}), \pi_{E_t}(\overline{\Delta}) \right\rangle \right) \right\} = O_P(n^{-1/2}). \quad (4.53)$$

For $0 < \alpha < \Delta_{\tau^*}$, using (4.31) and (4.41), we get:

$$\begin{aligned}
P\left(\|\bar{\mathbf{t}}_n - \mathbf{t}^*\|_\infty \geq \alpha\right) &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} J_n(\bar{\mathbf{w}}, \mathbf{t}) \leq 0\right) \\
&\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} \left\{J_n(\bar{\mathbf{w}}, \mathbf{t}) + J_n(\bar{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\langle \pi_{E_{t_n^*}}(\bar{\mathbf{w}}), \pi_{E_{t_n^*}}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{\mathbf{w}}), \pi_{E_t}(\bar{\Delta}) \rangle\right)\right\} \leq 0\right) \\
&\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} \left\{K_n(\bar{\mathbf{w}}, \mathbf{t}) + V_n(\bar{\mathbf{w}}, \mathbf{t}) + W_n(\bar{\mathbf{w}}, \mathbf{t}) + J_n(\bar{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\langle \pi_{E_{t_n^*}}(\bar{\mathbf{w}}), \pi_{E_{t_n^*}}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{\mathbf{w}}), \pi_{E_t}(\bar{\Delta}) \rangle\right)\right\} \leq 0\right) \\
&\leq P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} \left\{\frac{1}{2}K_n(\bar{\mathbf{w}}, \mathbf{t}) + V_n(\bar{\mathbf{w}}, \mathbf{t}) + W_n(\bar{\mathbf{w}}, \mathbf{t})\right\} \leq 0\right) \\
&\quad + P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} \left\{\frac{1}{2}K_n(\bar{\mathbf{w}}, \mathbf{t}) + J_n(\bar{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\langle \pi_{E_{t_n^*}}(\bar{\mathbf{w}}), \pi_{E_{t_n^*}}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{\mathbf{w}}), \pi_{E_t}(\bar{\Delta}) \rangle\right)\right\} \leq 0\right).
\end{aligned}$$

Following the proof of Lemma 4.3, one can prove that

$$P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} \left\{\frac{1}{2}K_n(\bar{\mathbf{w}}, \mathbf{t}) + V_n(\bar{\mathbf{w}}, \mathbf{t}) + W_n(\bar{\mathbf{w}}, \mathbf{t})\right\} \leq 0\right) \xrightarrow{n \rightarrow \infty} 0.$$

Using (4.38), we get that

$$\begin{aligned}
&P\left(\min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} \left\{\frac{1}{2}K_n(\bar{\mathbf{w}}, \mathbf{t}) + J_n(\bar{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\langle \pi_{E_{t_n^*}}(\bar{\mathbf{w}}), \pi_{E_{t_n^*}}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{\mathbf{w}}), \pi_{E_t}(\bar{\Delta}) \rangle\right)\right\} \leq 0\right) \\
&\leq P\left(\frac{1}{2}\lambda^2\alpha + \min_{\mathbf{t} \in \mathcal{C}_{n,\alpha}} \left\{J_n(\bar{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\langle \pi_{E_{t_n^*}}(\bar{\mathbf{w}}), \pi_{E_{t_n^*}}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{\mathbf{w}}), \pi_{E_t}(\bar{\Delta}) \rangle\right)\right\} \leq 0\right)
\end{aligned}$$

which goes to zero when n goes to infinity by (4.53).

Then Lemma 4.3 still holds if y is defined by (4.1). To show the rate of convergence, we use the same decomposition. As in the proof of Lemma 4.6, $P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\alpha,\gamma,n}(\mathcal{I})} J_n(\bar{\mathbf{v}}, \mathbf{t}) \leq 0\right) \xrightarrow{n \rightarrow \infty} 0$ for all $\alpha > 0$ and $0 < \gamma < 1/2$ is a sufficient condition for proving that $P(\widehat{\mathbf{t}}_n(y, \bar{\rho}_n) \in \mathcal{C}_{\alpha,\gamma,n}) \xrightarrow{n \rightarrow \infty} 0$, which allows us to conclude on the rate of convergence of the estimated change-points. Note that

$$\begin{aligned}
P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\alpha,\gamma,n}(\mathcal{I})} J_n(\bar{\mathbf{v}}, \mathbf{t}) \leq 0\right) &\leq P\left(\min_{\mathbf{t} \in \mathcal{C}'_{\alpha,\gamma,n}} \left\{\frac{1}{2}K_n(\bar{\mathbf{w}}, \mathbf{t}) + V_n(\bar{\mathbf{w}}, \mathbf{t}) + W_n(\bar{\mathbf{w}}, \mathbf{t})\right\} \leq 0\right) \\
&\quad + P\left(\frac{1}{2}\lambda^2\alpha + J_n(\bar{\Delta}, \mathbf{t}) - \frac{2}{n} \left(\langle \pi_{E_{t_n^*}}(\bar{\mathbf{w}}), \pi_{E_{t_n^*}}(\bar{\Delta}) \rangle - \langle \pi_{E_t}(\bar{\mathbf{w}}), \pi_{E_t}(\bar{\Delta}) \rangle\right) \leq 0\right).
\end{aligned}$$

In the latter equation, the second term of the right-hand side goes to zero as n goes to infinity by (4.53). The first term of the right-hand side goes to zero when n goes to infinity by following the same line of reasoning as the one of Lemma 4.5. This concludes the proof of Proposition 4.3. $\square$

4.7 Tables and figures

n	7200					14400				
estimate \ number of changes	< 5	5	6	7	> 7	< 5	5	6	7	> 7
$\widehat{m}_Y^0$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}^0$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}$	0	0	97	3	0	0	0	100	0	0
$\widehat{m}_{PP}$	0	0	99	1	0	0	0	100	0	0
$\widehat{m}^*$	0	0	97	3	0	0	0	100	0	0
$\widehat{m}_{PP}^*$	0	0	99	1	0	0	0	100	0	0
$\widehat{m}'$	0	0	97	3	0	0	0	100	0	0
$\widehat{m}'_{PP}$	0	0	99	1	0	0	0	100	0	0
estimate \ order of the autoregression	0	1	2	3	> 3	0	1	2	3	> 3
$\widehat{p}'$	0	0	96	4	0	0	0	97	1	2
$\widetilde{\phi}_{n,1}^{(2)}$ RMSE	$1.99 \cdot 10^{-2}$					$1.64 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,2}^{(2)}$ RMSE	$1.80 \cdot 10^{-2}$					$1.54 \cdot 10^{-2}$				

Table 4.1 – Estimates of the number of changes, of the order of the autoregression, and RMSEs of the estimates of the autoregression parameters, for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (-1.2, -0.4, 0.4)$.

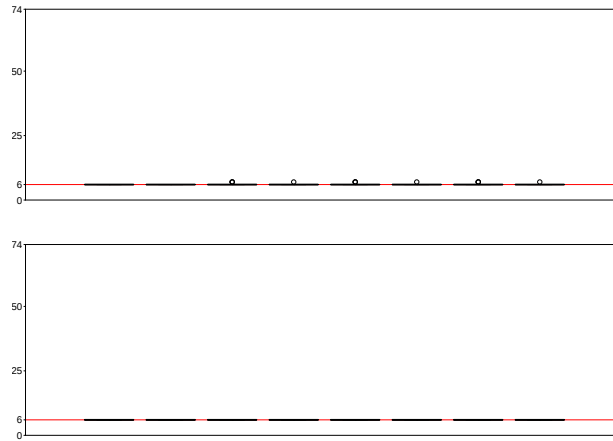


Figure 4.1 – Boxplots of the estimates of the number of changes for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (-1.2, -0.4, 0.4)$. $n = 7200$ (top) or 14400 (bottom). In each plot, the estimates boxplots are in the following order (from left to right): $\widehat{m}_Y^0$, $\widehat{m}^0$, $\widehat{m}$, $\widehat{m}_{PP}$, $\widehat{m}^*$, $\widehat{m}_{PP}^*$, $\widehat{m}'$, $\widehat{m}'_{PP}$. The true number of changes is equal to 6 (red horizontal line).

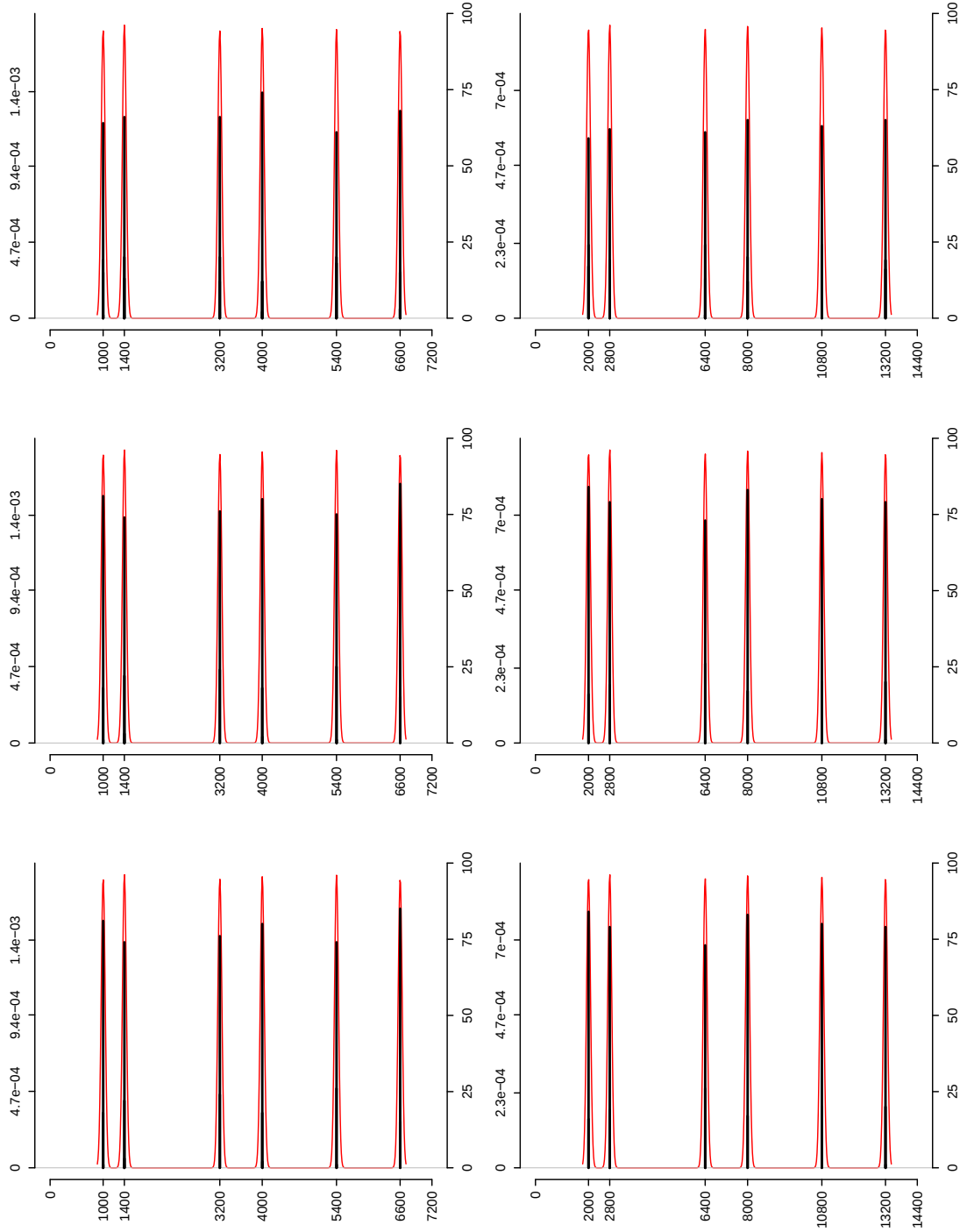


Figure 4.2 – Frequency plots of the change-point location estimates for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (-1.2, -0.4, 0.4)$. $n = 7200$ (left) or 14400 (right). Estimates: $\hat{t}_n^0$ (top), $\hat{t}_{n,PP}$ (middle), $\hat{t}_{n,PP}$ (bottom). The black line represents the absolute frequency of each location between 1 and n in estimates (scale on right axis). The red line represents the Gaussian kernel density estimate of this dataset (scale on left axis).

n	7200					14400				
estimate \ number of changes	< 5	5	6	7	> 7	< 5	5	6	7	> 7
$\widehat{m}_Y^0$	0	0	0	0	100	0	0	0	0	100
$\widehat{m}^0$	0	0	0	0	100	0	0	0	1	99
$\widehat{m}$	3	0	91	6	0	0	0	99	1	0
$\widehat{m}_{PP}$	3	0	97	0	0	0	0	100	0	0
$\widehat{m}^*$	0	0	94	6	0	0	0	99	1	0
$\widehat{m}_{PP}^*$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}'$	4	0	90	6	0	0	0	98	2	0
$\widehat{m}'_{PP}$	4	0	96	0	0	0	0	99	1	0
estimate \ order of the autoregression	0	1	2	3	> 3	0	1	2	3	> 3
$\widehat{p}'$	0	0	46	24	30	0	0	55	15	30
$\widetilde{\phi}_{n,1}^{(2)}$ RMSE	$4.93 \cdot 10^{-2}$					$3.46 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,2}^{(2)}$ RMSE	$3.13 \cdot 10^{-2}$					$2.16 \cdot 10^{-2}$				

Table 4.2 – Estimates of the number of changes, of the order of the autoregression, and RMSEs of the estimates of the autoregression parameters, for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (1.6, -0.8, 0.4)$.

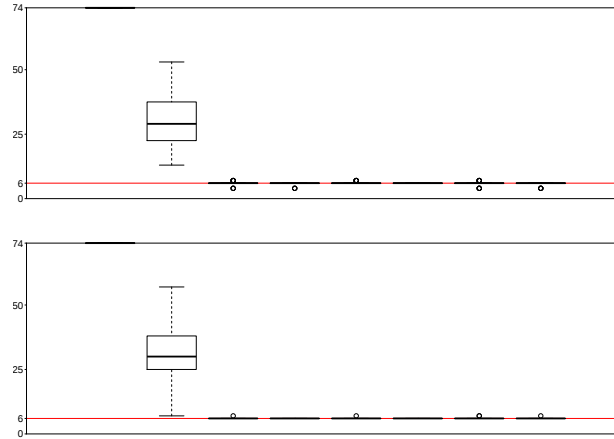


Figure 4.3 – Boxplots of the estimates of the number of changes for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (1.6, -0.8, 0.4)$. $n = 7200$ (top) or 14400 (bottom). In each plot, the estimates boxplots are in the following order (from left to right): $\widehat{m}_Y^0$, $\widehat{m}^0$, $\widehat{m}$, $\widehat{m}_{PP}$, $\widehat{m}^*$, $\widehat{m}_{PP}^*$, $\widehat{m}'$, $\widehat{m}'_{PP}$. The true number of changes is equal to 6 (red horizontal line).

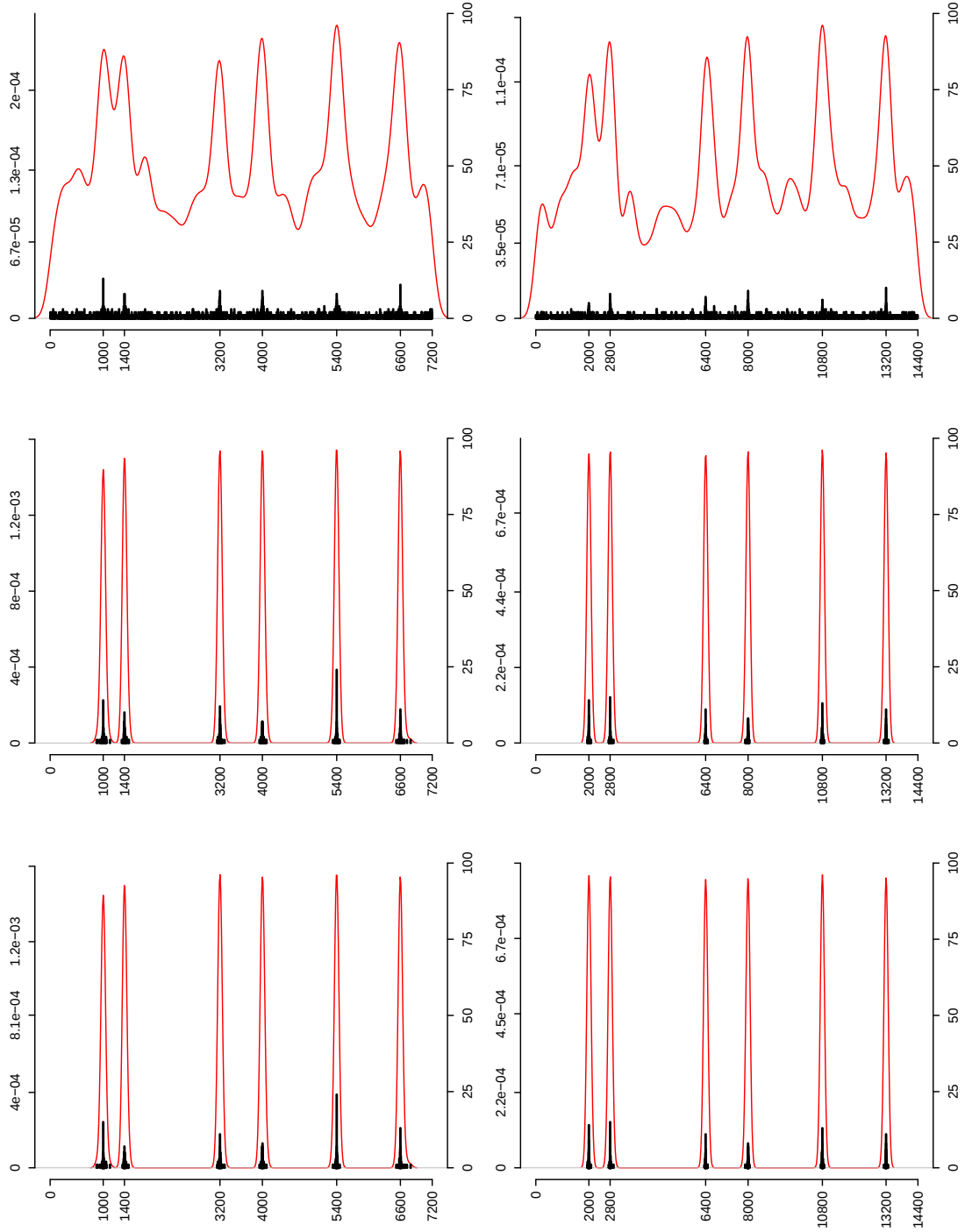


Figure 4.4 – Frequency plots of the change-point location estimates for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (1.6, -0.8, 0.4)$. $n = 7200$ (left) or 14400 (right). Estimates: $\hat{t}_n^0$ (top), $\hat{t}_{n,PP}$ (middle), $\hat{t}_{n,PP}$ (bottom). The black line represents the absolute frequency of each location between 1 and n in estimates (scale on right axis). The red line represents the Gaussian kernel density estimate of this dataset (scale on left axis).

n	7200					14400				
estimate \ number of changes	< 5	5	6	7	> 7	< 5	5	6	7	> 7
$\hat{m}_Y^0$	0	0	0	0	100	0	0	0	0	100
$\hat{m}^0$	0	0	51	20	29	0	0	60	17	23
$\hat{m}$	3	0	96	3	1	0	0	98	2	0
$\hat{m}_{PP}$	3	0	97	2	1	0	0	98	2	0
$\hat{m}^*$	0	0	97	3	0	0	0	99	1	0
$\hat{m}_{PP}^*$	0	0	98	2	0	0	0	99	1	0
$\hat{m}'$	0	0	97	3	0	0	0	99	1	0
$\hat{m}'_{PP}$	0	0	98	2	0	0	0	99	1	0
estimate \ order of the autoregression	0	1	2	3	> 3	0	1	2	3	> 3
$\hat{p}'$	0	0	59	25	16	0	0	45	30	25
$\tilde{\phi}_{n,1}^{(2)}$ RMSE	$7.00 \cdot 10^{-2}$					$6.44 \cdot 10^{-2}$				
$\tilde{\phi}_{n,2}^{(2)}$ RMSE	$4.20 \cdot 10^{-2}$					$3.68 \cdot 10^{-2}$				

Table 4.3 – Estimates of the number of changes, of the order of the autoregression, and RMSEs of the estimates of the autoregression parameters, for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.2, 0.4)$.

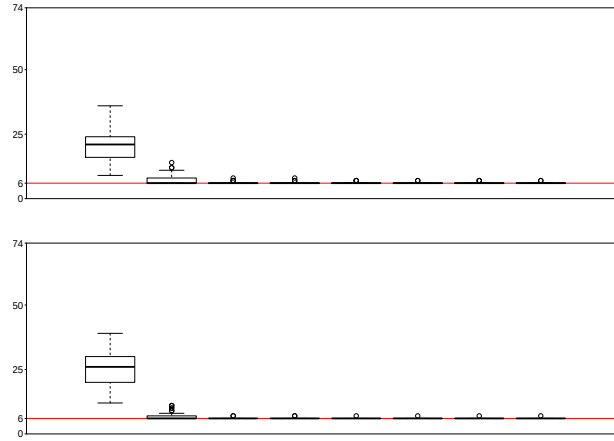


Figure 4.5 – Boxplots of the estimates of the number of changes for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.2, 0.4)$. $n = 7200$ (top) or 14400 (bottom). In each plot, the estimates boxplots are in the following order (from left to right): $\hat{m}_Y^0$, $\hat{m}^0$, $\hat{m}$, $\hat{m}_{PP}$, $\hat{m}^*$, $\hat{m}_{PP}^*$, $\hat{m}'$, $\hat{m}'_{PP}$. The true number of changes is equal to 6 (red horizontal line).

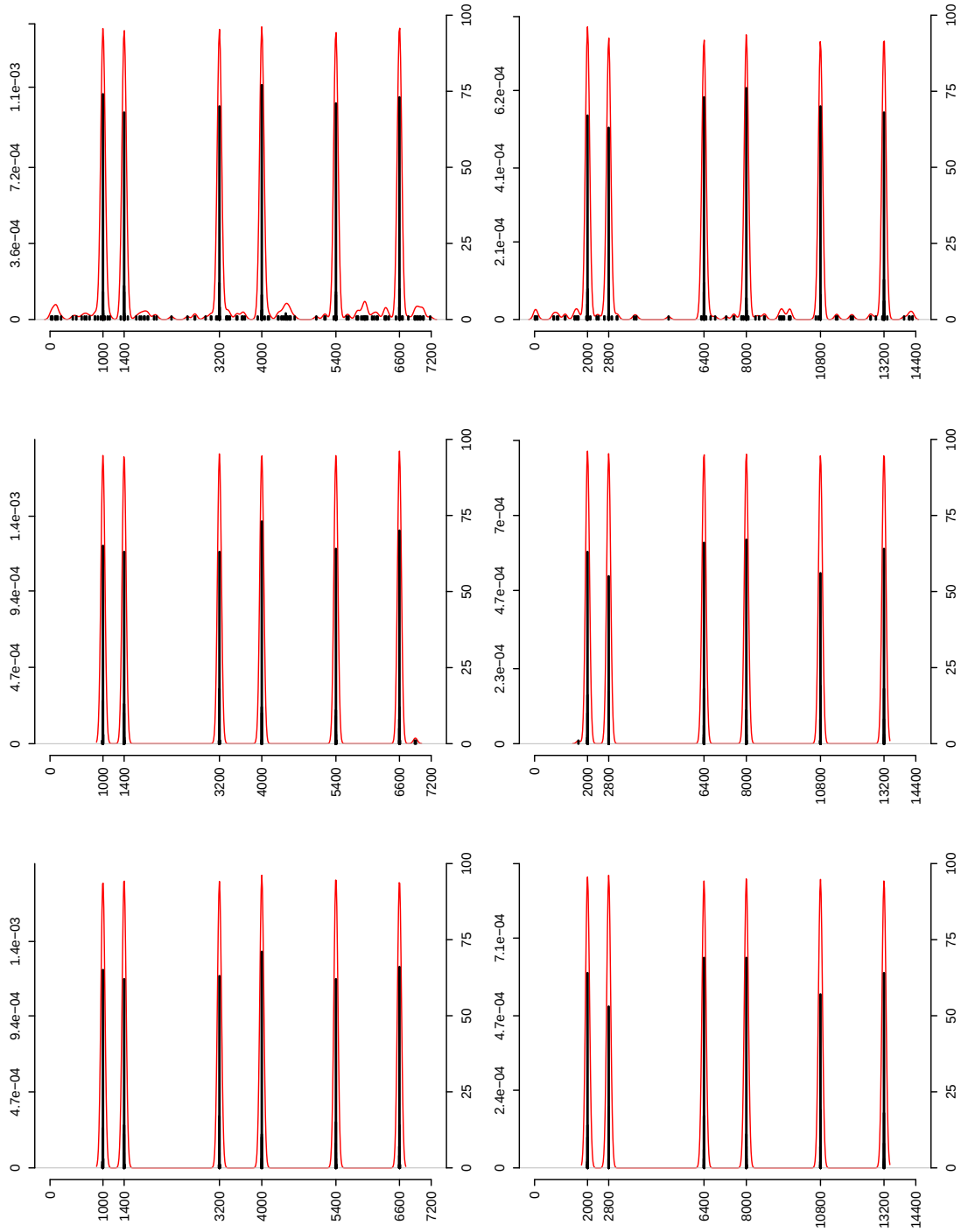


Figure 4.6 – Frequency plots of the change-point location estimates for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.2, 0.4)$. $n = 7200$ (left) or 14400 (right). Estimates: $\hat{t}_n^0$ (top), $\hat{t}_{n,PP}$ (middle), $\hat{\tau}_{n,PP}$ (bottom). The black line represents the absolute frequency of each location between 1 and n in estimates (scale on right axis). The red line represents the Gaussian kernel density estimate of this dataset (scale on left axis).

n	7200					14400				
estimate \ number of changes	< 5	5	6	7	> 7	< 5	5	6	7	> 7
$\hat{m}_Y^0$	0	0	0	0	100	0	0	0	0	100
$\hat{m}^0$	0	0	0	0	100	0	0	0	0	100
$\hat{m}$	13	0	27	4	56	25	0	29	9	37
$\hat{m}_{PP}$	13	0	28	3	56	25	0	33	5	37
$\hat{m}^*$	0	1	80	18	1	0	0	88	11	1
$\hat{m}_{PP}^*$	1	0	91	7	1	0	0	98	2	0
$\hat{m}'$	22	0	54	15	9	6	0	80	11	3
$\hat{m}'_{PP}$	22	0	66	4	8	6	0	90	3	1
estimate \ order of the autoregression	0	1	2	3	> 3	0	1	2	3	> 3
$\hat{p}'$	0	0	18	12	70	0	0	14	12	74
$\tilde{\phi}_{n,1}^{(2)}$ RMSE	$3.44 \cdot 10^{-1}$					$2.40 \cdot 10^{-1}$				
$\tilde{\phi}_{n,2}^{(2)}$ RMSE	$2.41 \cdot 10^{-1}$					$1.71 \cdot 10^{-1}$				

Table 4.4 – Estimates of the number of changes, of the order of the autoregression, and RMSEs of the estimates of the autoregression parameters, for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.6, 0.4)$.

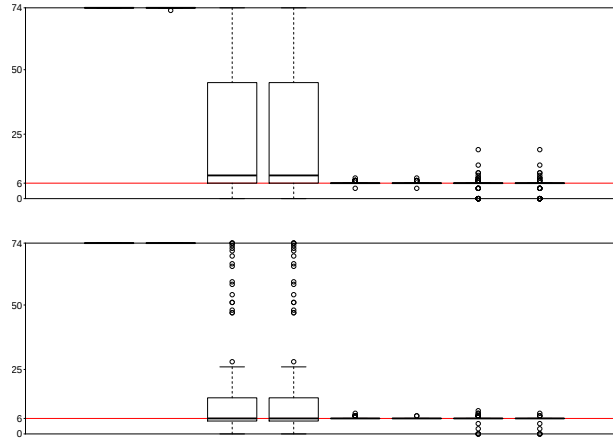


Figure 4.7 – Boxplots of the estimates of the number of changes for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.6, 0.4)$. $n = 7200$ (top) or 14400 (bottom). In each plot, the estimates boxplots are in the following order (from left to right): $\hat{m}_Y^0$, $\hat{m}^0$, $\hat{m}$, $\hat{m}_{PP}$, $\hat{m}^*$, $\hat{m}_{PP}^*$, $\hat{m}'$, $\hat{m}'_{PP}$. The true number of changes is equal to 6 (red horizontal line).

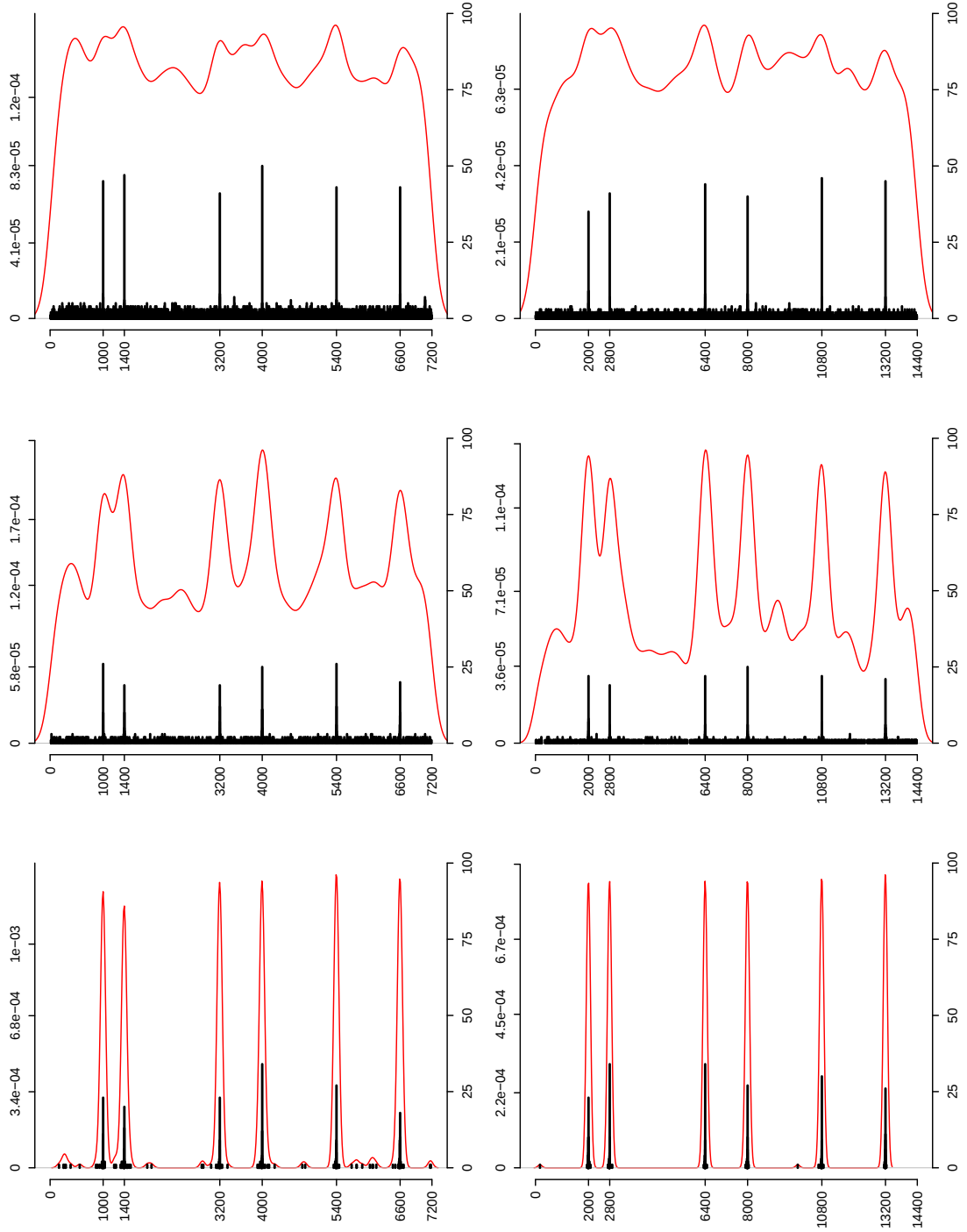


Figure 4.8 – Frequency plots of the change-point location estimates for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.6, 0.4)$. $n = 7200$ (left) or 14400 (right). Estimates: $\hat{t}_n^0$ (top), $\hat{t}_{n,PP}$ (middle), $\hat{t}_{n,PP}$ (bottom). The black line represents the absolute frequency of each location between 1 and n in estimates (scale on right axis). The red line represents the Gaussian kernel density estimate of this dataset (scale on left axis).

n	7200					14400				
estimate \ number of changes	< 5	5	6	7	> 7	< 5	5	6	7	> 7
$\widehat{m}_Y^0$	0	0	0	0	100	0	0	0	0	100
$\widehat{m}^0$	0	0	0	0	100	0	0	0	0	100
$\widehat{m}$	2	0	52	28	18	1	0	59	28	12
$\widehat{m}_{PP}$	3	0	85	10	2	1	0	96	3	0
$\widehat{m}^*$	0	0	65	28	7	0	0	70	29	1
$\widehat{m}_{PP}^*$	0	0	98	2	0	0	0	98	2	0
$\widehat{m}'$	0	0	63	28	9	0	0	74	24	2
$\widehat{m}'_{PP}$	0	0	100	0	0	0	0	99	1	0
estimate \ order of the autoregression	0	1	2	3	> 3	0	1	2	3	> 3
$\widehat{p}'$	0	0	36	20	44	0	0	36	21	43
$\widetilde{\phi}_{n,1}^{(2)}$ RMSE	$1.11 \cdot 10^{-1}$					$8.17 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,2}^{(2)}$ RMSE	$5.16 \cdot 10^{-2}$					$3.76 \cdot 10^{-2}$				

Table 4.5 – Estimates of the number of changes, of the order of the autoregression, and RMSEs of the estimates of the autoregression parameters, for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.4, 0.2, 0.2)$.

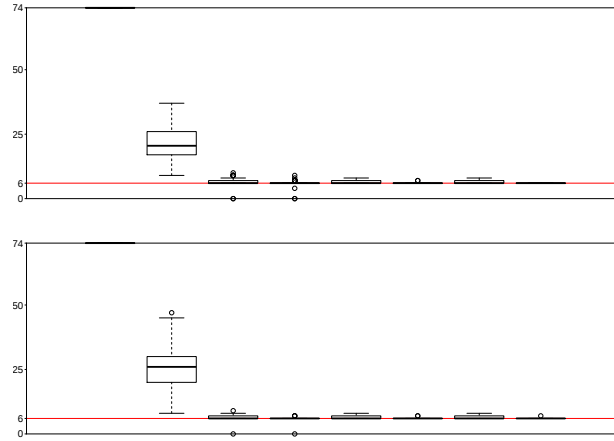


Figure 4.9 – Boxplots of the estimates of the number of changes for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.4, 0.2, 0.2)$. $n = 7200$ (top) or 14400 (bottom). In each plot, the estimates boxplots are in the following order (from left to right): $\widehat{m}_Y^0$, $\widehat{m}^0$, $\widehat{m}$, $\widehat{m}_{PP}$, $\widehat{m}^*$, $\widehat{m}_{PP}^*$, $\widehat{m}'$, $\widehat{m}'_{PP}$. The true number of changes is equal to 6 (red horizontal line).

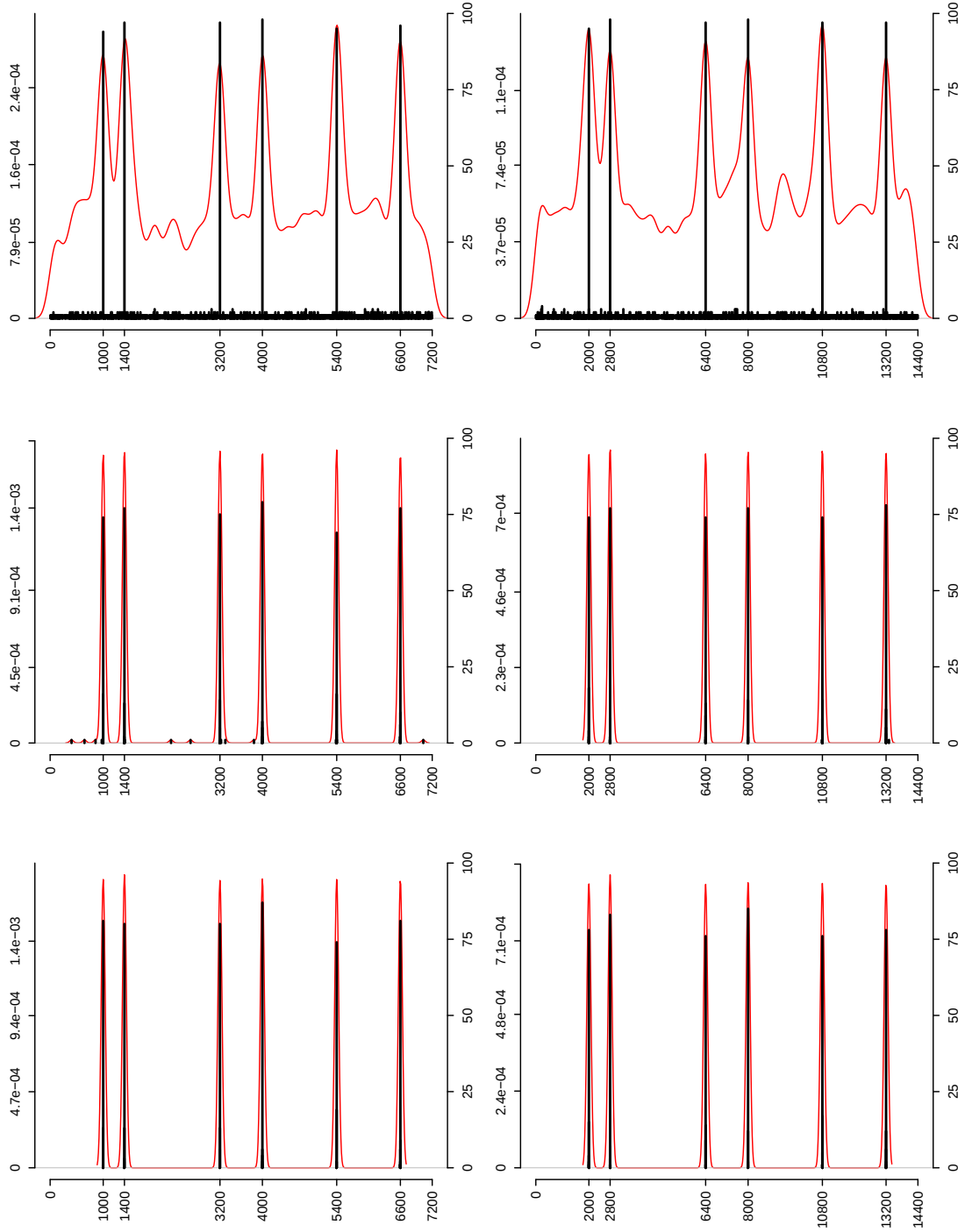


Figure 4.10 – Frequency plots of the change-point location estimates for 100 AR(2) series with the parameters $(\phi_1^*, \phi_2^*, \sigma^*) = (0.4, 0.2, 0.2)$. $n = 7200$ (left) or 14400 (right). Estimates: $\hat{t}_n^0$ (top), $\hat{t}_{n,PP}$ (middle), $\hat{t}_{n,PP}^*$ (bottom). The black line represents the absolute frequency of each location between 1 and n in estimates (scale on right axis). The red line represents the Gaussian kernel density estimate of this dataset (scale on left axis).

n	7200					14400				
estimate \ number of changes	< 5	5	6	7	> 7	< 5	5	6	7	> 7
$\widehat{m}_Y^0$	0	0	0	0	100	0	0	0	0	100
$\widehat{m}^0$	0	0	31	16	53	0	0	31	14	55
$\widehat{m}$	2	0	90	4	4	0	0	100	0	0
$\widehat{m}_{PP}$	2	0	92	2	4	0	0	100	0	0
$\widehat{m}^*$	0	0	100	0	0	0	0	99	1	0
$\widehat{m}_{PP}^*$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}'$	0	0	99	1	0	0	0	99	1	0
$\widehat{m}'_{PP}$	0	0	99	1	0	0	0	100	0	0
estimate \ order of the autoregression	< 4	4	5	6	> 6	< 4	4	5	6	> 6
$\widehat{p}'$	0	0	41	25	34	0	0	45	21	34
$\widetilde{\phi}_{n,1}^{(5)}$ RMSE	$1.01 \cdot 10^{-1}$					$6.92 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,2}^{(5)}$ RMSE	$4.36 \cdot 10^{-2}$					$3.19 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,3}^{(5)}$ RMSE	$3.54 \cdot 10^{-2}$					$2.45 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,4}^{(5)}$ RMSE	$2.48 \cdot 10^{-2}$					$1.84 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,5}^{(5)}$ RMSE	$3.72 \cdot 10^{-2}$					$2.35 \cdot 10^{-2}$				

Table 4.6 – Estimates of the number of changes, of the order of the autoregression, and RMSEs of the estimates of the autoregression parameters, for 100 AR(5) series with the parameters $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0.5, -0.5, 0.4)$.

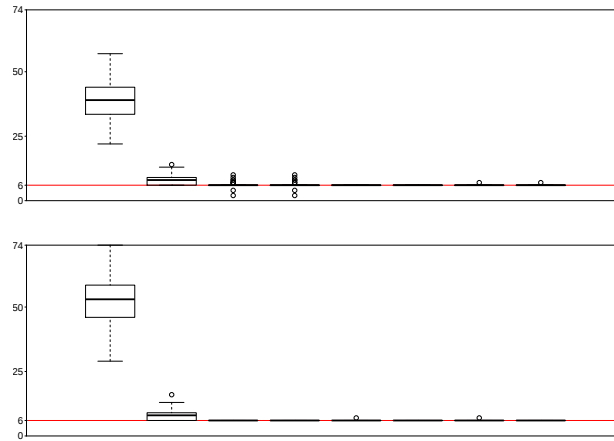


Figure 4.11 – Boxplots of the estimates of the number of changes for 100 AR(5) series with the parameters $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0.5, -0.5, 0.4)$. $n = 7200$ (top) or 14400 (bottom). In each plot, the estimates boxplots are in the following order (from left to right): $\widehat{m}_Y^0$, $\widehat{m}^0$, $\widehat{m}$, $\widehat{m}_{PP}$, $\widehat{m}^*$, $\widehat{m}_{PP}^*$, $\widehat{m}'$, $\widehat{m}'_{PP}$. The true number of changes is equal to 6 (red horizontal line).

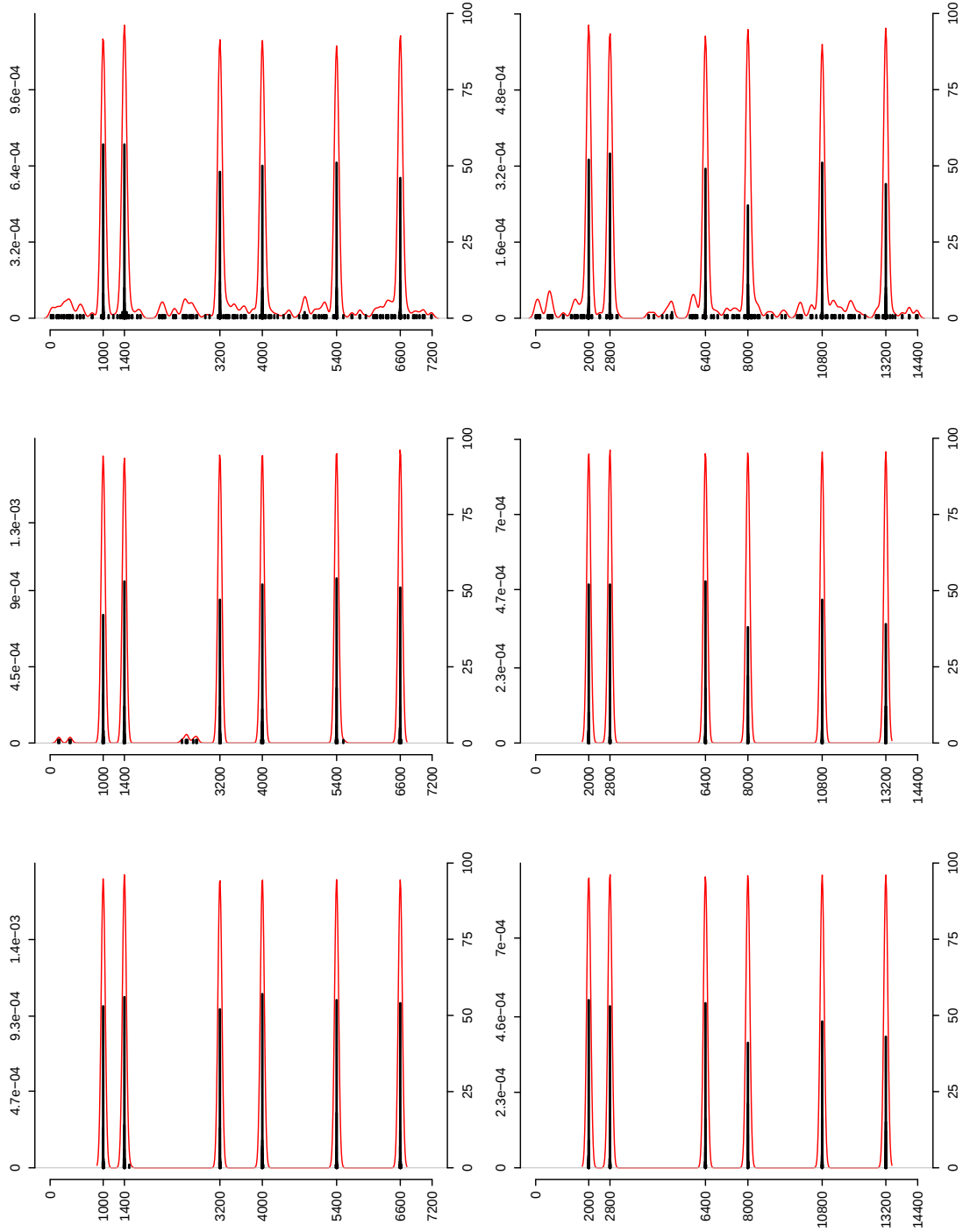


Figure 4.12 – Frequency plots of the change-point location estimates for 100 AR(5) series with the parameters $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0.5, -0.5, 0.4)$. $n = 7200$ (left) or 14400 (right). Estimates: $\hat{t}_n^0$ (top), $\hat{t}_{n,PP}$ (middle), $\hat{t}_{n,PP}$ (bottom). The black line represents the absolute frequency of each location between 1 and n in estimates (scale on right axis). The red line represents the Gaussian kernel density estimate of this dataset (scale on left axis).

n	7200					14400				
estimate \ number of changes	< 5	5	6	7	> 7	< 5	5	6	7	> 7
$\widehat{m}_Y^0$	0	0	0	0	100	0	0	0	0	100
$\widehat{m}^0$	0	0	53	5	42	0	0	54	2	44
$\widehat{m}$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}_{PP}$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}^*$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}_{PP}^*$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}'$	0	0	100	0	0	0	0	100	0	0
$\widehat{m}'_{PP}$	0	0	100	0	0	0	0	100	0	0
estimate \ order of the autoregression	< 4	4	5	6	> 6	< 4	4	5	6	> 6
$\widehat{p}'$	0	0	72	17	11	0	0	83	12	5
$\widetilde{\phi}_{n,1}^{(5)}$ RMSE	$2.99 \cdot 10^{-2}$					$1.77 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,2}^{(5)}$ RMSE	$1.24 \cdot 10^{-2}$					$1.05 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,3}^{(5)}$ RMSE	$1.25 \cdot 10^{-2}$					$1.03 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,4}^{(5)}$ RMSE	$1.28 \cdot 10^{-2}$					$1.01 \cdot 10^{-2}$				
$\widetilde{\phi}_{n,5}^{(5)}$ RMSE	$1.29 \cdot 10^{-2}$					$9.47 \cdot 10^{-3}$				

Table 4.7 – Estimates of the number of changes, of the order of the autoregression, and RMSEs of the estimates of the autoregression parameters, for 100 AR(5) series with the parameters $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0, -0.5, 0.4)$.

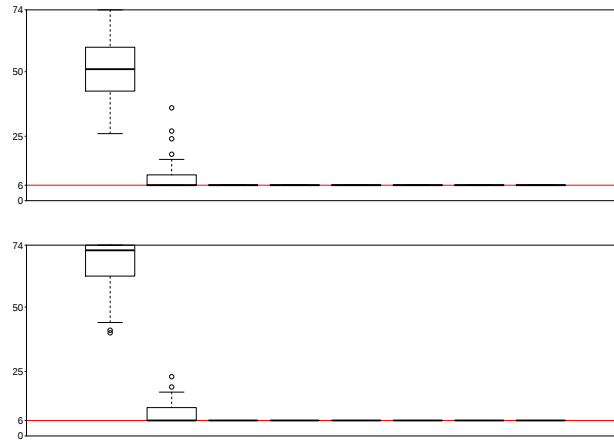


Figure 4.13 – Boxplots of the estimates of the number of changes for 100 AR(5) series with the parameters $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0, -0.5, 0.4)$. $n = 7200$ (top) or 14400 (bottom). In each plot, the estimates boxplots are in the following order (from left to right): $\widehat{m}_Y^0$, $\widehat{m}^0$, $\widehat{m}$, $\widehat{m}_{PP}$, $\widehat{m}^*$, $\widehat{m}_{PP}^*$, $\widehat{m}'$, $\widehat{m}'_{PP}$. The true number of changes is equal to 6 (red horizontal line).

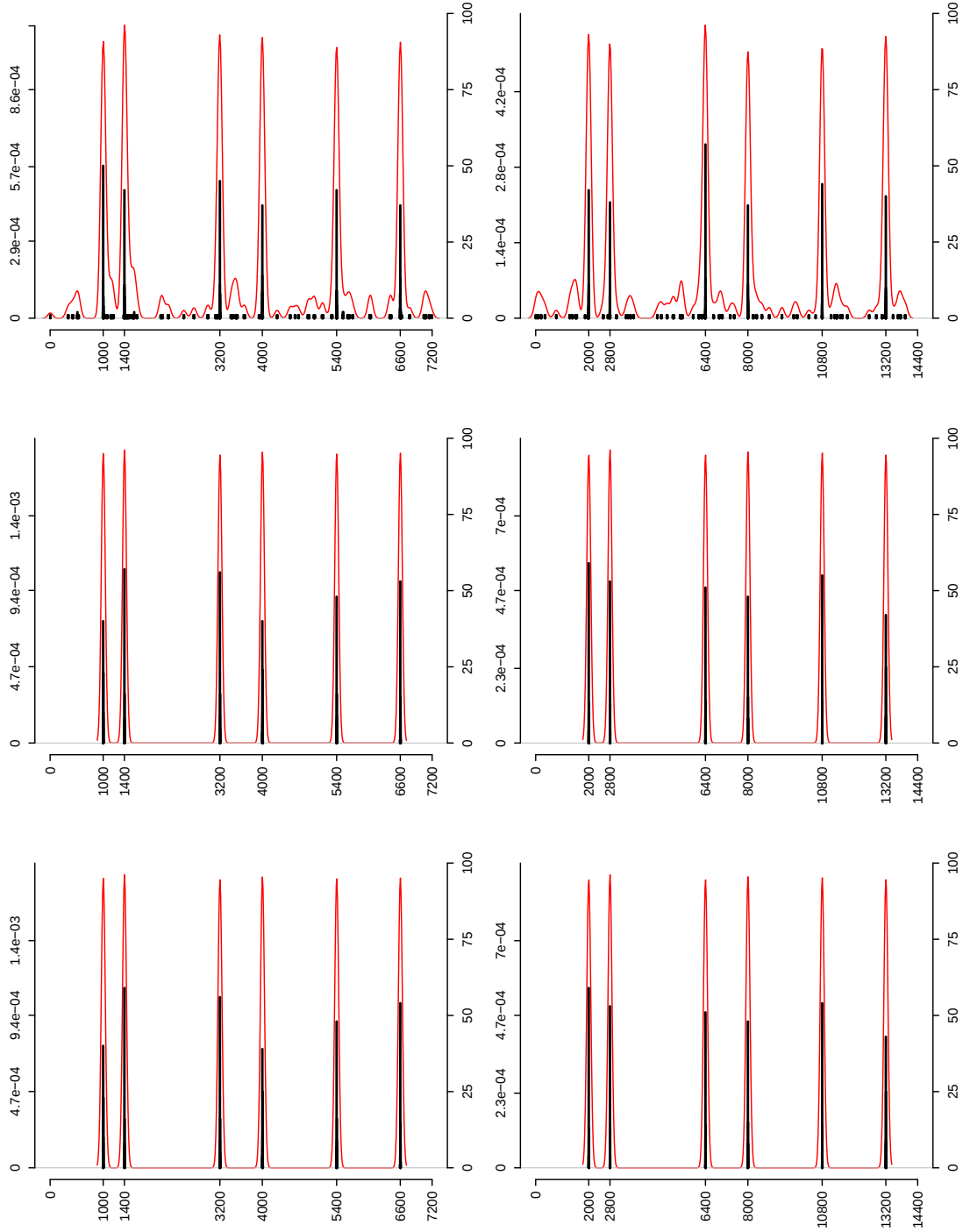


Figure 4.14 – Frequency plots of the change-point location estimates for 100 AR(5) series with the parameters $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0, -0.5, 0.4)$. $n = 7200$ (left) or 14400 (right). Estimates: $\hat{t}_n^0$ (top), $\hat{t}_{n,PP}$ (middle), $\hat{t}_{n,PP}$ (bottom). The black line represents the absolute frequency of each location between 1 and n in estimates (scale on right axis). The red line represents the Gaussian kernel density estimate of this dataset (scale on left axis).

Chapitre 5

Commentaires et discussion

Contenu

5.1	Commentaires	125
5.1.1	Commentaires sur le chapitre 3	125
5.1.2	Commentaires sur le chapitre 4	130
5.2	Limites de l'approche adoptée	131
5.2.1	Programmation dynamique et dépendance	131
5.2.2	Cadre asymptotique	133
5.3	Perspectives	134
5.3.1	Estimation de la fonction d'autocovariance et d'autocorrélation . . .	135
5.3.2	Structure de dépendance non-paramétrique	137
5.3.3	Méthodologie basée sur l'estimation robuste préliminaire d'un para- mètre global dans un modèle de segmentation	138

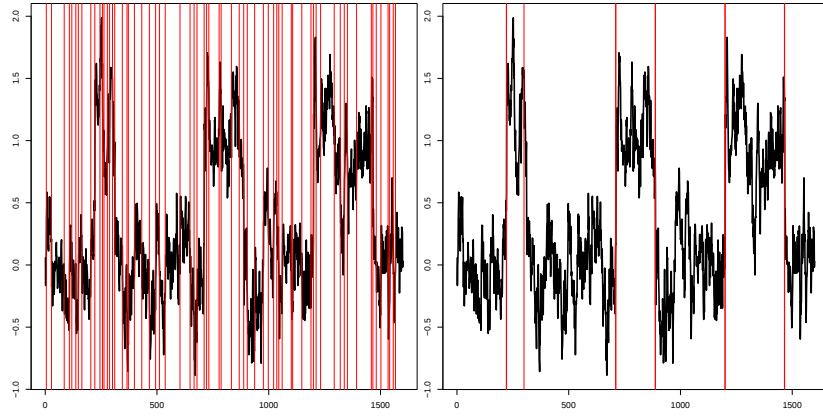


FIGURE 5.1 – Exemple d’une des séries simulées dans l’étude réalisée dans la section 3.5, affectée de six ruptures. Ici, $n = 1600$, $\rho^* = 0.9$, $\sigma^* = 0.3$. Les droites verticales marquent la localisation des ruptures estimées. On constate la sur-segmentation de la série si on utilise le critère mBIC sans décorrélation (gauche), contrairement au résultat donné par le critère mBIC après décorrélation (droite).

5.1 Commentaires

Cette section contient des commentaires sur les articles présentés dans les chapitres 3 et 4.

5.1.1 Commentaires sur le chapitre 3

L’article présenté au chapitre 3 a pour objet principal la segmentation, c’est sur ce point que porte sa contribution principale, à savoir la justification théorique de la procédure consistant à décorréler les séries temporelles étudiées avant de les segmenter. C’est sur ce point, notamment la possibilité offerte de sélectionner le nombre de ruptures par le critère mBIC de Zhang et Siegmund (2007) sans sur-segmentation (cf. figure 5.1), que la méthodologie proposée est évaluée. Toutefois, les commentaires porteront principalement sur l’estimation du paramètre d’autorégression ρ^* . On notera que l’estimation de ce paramètre n’est qu’un outil de la méthode de segmentation proposée, et absolument pas un but en soi. Seront évoqués la construction de l’estimateur utilisé, un autre estimateur, leurs propriétés asymptotiques respectives et leur utilisation dans une statistique de test.

Construction de l’estimateur $\tilde{\rho}_n$

Les estimateurs S_n et Q_n de Rousseeuw et Croux (1993) d’un paramètre d’échelle d’un échantillon $\mathbf{x} = (x_1, \dots, x_n)$, définis resp. en (2.15) et en (2.19), font intervenir les quantiles des différences $|x_i - x_j|$, ($1 \leq i < j \leq n$). Plaçons-nous dans le modèle des valeurs aberrantes additives (*additive outliers*, AO), où un nombre k d’observations de cet échantillon est affecté par une perturbation déterministe : $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{v}$ où $\mathbf{v} \in \mathbb{R}^n$ n’a que k composantes non-nulles. La définition du modèle AO est ici simplifiée pour la clarté de l’exposé (cf. Maronna *et al.* 2006, section 8.1.2).

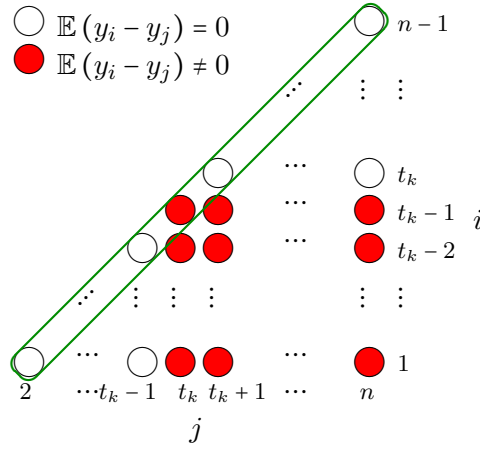


FIGURE 5.2 – Illustration de la surcontamination des différences entre observations $y_i - y_j$, ($1 \leq i < j \leq n$) dans un modèle de segmentation par rapport au modèle AO. Le problème disparaît si on impose $j = i + 1$.

Parmi les $n(n-1)/2$ différences $v_i - v_j$, ($1 \leq i < j \leq n$), au plus $k(n-1)$ différences sont non nulles. Donc parmi les $n(n-1)/2$ différences $\tilde{x}_i - \tilde{x}_j = (x_i - x_j) + (v_i - v_j)$, ($1 \leq i < j \leq n$), au plus $k(n-1)$ différences sont différentes de $x_i - x_j$, on dira qu'elles sont contaminées. C'est ce qui permet d'obtenir un point d'effondrement (cf. définition 2.16) tendant vers $1/2$ pour Q_n .

Supposons maintenant que l'on observe $(y_i)_{1 \leq i \leq n}$ défini en l'équation (3.1). Les seules différences $y_i - y_j$, ($1 \leq i < j \leq n$) dont on est sûr qu'elles soient égales à $\eta_i - \eta_j$, et donc centrées, sont les différences intra-segment :

$$y_i - y_j, (t_{k-1}^* + 1 \leq i < j \leq t_k^*, 1 \leq k \leq m^* + 1).$$

Il n'y en a que $\sum_{k=1}^{m^*+1} (t_k^* - t_{k-1}^*) (t_k^* - t_{k-1}^* - 1)/2$. S'il n'y a qu'une seule rupture, la borne supérieure de cette quantité, quand la rupture varie, est $n(n-2)/4$; soit une proportion $(n-2)/(2n-2) < 1/2$ des différences $y_i - y_j$, ($1 \leq i < j \leq n$). De plus, cette borne supérieure décroît avec le nombre de ruptures. Par conséquent, même avec une seule rupture, on a possiblement un taux de contamination strictement supérieur à $1/2$ des différences $y_i - y_j$, ($1 \leq i < j \leq n$) (cf. figure 5.2).

Pour construire un estimateur de ρ^* dans ce cadre, nous ne considérons donc pas toutes les différences $y_i - y_j$, ($1 \leq i < j \leq n$), mais seulement une partie d'entre elles. En effet, parmi les $n-1$ différences $y_{i+1} - y_i$, ($1 \leq i \leq n-1$), seules m^* différences sont contaminées. De même, parmi les $n-2$ différences $y_{i+2} - y_i$, ($1 \leq i \leq n-2$), au plus $2m^*$ différences sont contaminées. D'où l'estimateur proposé, défini en (3.3), faisant intervenir les médianes de ces deux familles de différences entre observations.

Un autre estimateur

L'estimateur $\tilde{\rho}_n$, défini en (3.3), permet de justifier, par ses propriétés et sa rapidité de calcul, la démarche suivie dans le reste de l'article (complexité en temps en $O(n)$ grâce à un

algorithme de sélection, voir par exemple Blum *et al.* 1973). Mais on pourrait trouver bien d'autres estimateurs.

L'estimateur que l'on va construire ici a, au prix d'une complexité en temps de son calcul plus grande (en $O(n \log n)$, comme l'estimateur Q_n de Rousseeuw et Croux, 1993), une variance asymptotique qui semble, d'après les simulations (cf. section 5.1.1), plus petite que $\tilde{\rho}_n$. Supposons que l'on observe $(y_i)_{1 \leq i \leq n}$ défini en l'équation (3.1). Soit $\mathbf{x} = (x_i)_{1 \leq i \leq n-1} = (y_{i+1} - y_i)_{1 \leq i \leq n-1}$. Soit $\hat{\rho}_{MG,n}(h, \cdot)$ l'estimateur de Ma et Genton (2000) de l'autocorrélation à l'ordre $h \in \mathbb{Z}$, défini en (2.21). On considère l'estimateur suivant de ρ^* :

$$\tilde{\rho}'_n = 1 + 2\hat{\rho}_{MG,n}(1, \mathbf{x}) . \quad (5.1)$$

Proposition 5.1. *Soit $(y_i)_{1 \leq i \leq n}$ défini en l'équation (3.1) et $(\eta_i)_{i \in \mathbb{Z}}$ défini en (3.2). On note $\mathbf{x} = (x_i)_{1 \leq i \leq n-1} = (y_{i+1} - y_i)_{1 \leq i \leq n-1}$, et $(\nu_i)_{i \in \mathbb{Z}} = (\eta_{i+1} - \eta_i)_{i \in \mathbb{Z}}$. Soit $\tilde{\rho}'_n$ défini en (5.1). Alors*

$$\sqrt{n}(\tilde{\rho}'_n - \rho^{*2}) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \tilde{\sigma}'^2) ,$$

où

$$\tilde{\sigma}'^2 = 4\mathbb{E}[\psi^2(\nu_0, \nu_1)] + 8 \sum_{k=1}^{\infty} \mathbb{E}[\psi(\nu_0, \nu_1) \psi(\nu_k, \nu_{k+1})] , \quad (5.2)$$

$$\psi : (x, y) \mapsto \left\{ \frac{1 + \rho^*}{2} \text{IF} \left(\frac{x + y}{\sqrt{\sigma^{*2}(1 - \rho^*)}} \right) - \frac{3 - \rho^*}{2} \text{IF} \left(\frac{(x - y)\sqrt{1 + \rho^*}}{\sqrt{\sigma^{*2}(3 - \rho^*)(1 - \rho^*)}} \right) \right\} , \quad (5.3)$$

$$\text{IF} : x \mapsto c_{Q, \Phi} \left(\frac{1/4 - \Phi(x + 1/c_{Q, \Phi}) + \Phi(x - 1/c_{Q, \Phi})}{\int_{\mathbb{R}} \phi(y) \phi(y + 1/c_{Q, \Phi}) dy} \right) , \quad (5.4)$$

Φ est la fonction de répartition de la loi normale centrée réduite, $\phi = \Phi'$ et $c_{Q, \Phi}$ est défini par (2.18).

Démonstration. Le résultat découle directement de :

- le corollaire 2.1 appliqué à $(\nu_i)_{i \in \mathbb{Z}}$.
- $\forall h \in \mathbb{Z}, \gamma_\nu(h) = 2\gamma_\eta(h) - \gamma_\eta(h - 1) - \gamma_\eta(h + 1)$.
- $\gamma_\eta(0) = \frac{\sigma^{*2}}{1 - \rho^{*2}}$, en particulier $\rho^* = 1 + 2\rho_\nu(1)$.

□

Remarque 5.1. $\tilde{\sigma}'^2$ ne dépend que de ρ^* et non de σ^* . À l'instar de la variance asymptotique $\tilde{\sigma}'^2$ de l'estimateur $\tilde{\rho}_n$ donnée par l'équation (3.5), l'expression (5.2) n'est pas aisément calculable en fonction de ρ^* . On propose ci-dessous un calcul par la méthode de Monte-Carlo.

Efficacité asymptotique relative

Pour chaque paramètre d'autorégression $\rho^* \in \{-0.9, -0.8, \dots, 0.9\}$, nous avons simulé 5000 séries AR(1) centrées, gaussiennes, de variance 1, de longueur 10000. Le but n'étant pas de mesurer la robustesse de ces estimateurs à la présence de ruptures, ces séries ne sont pas

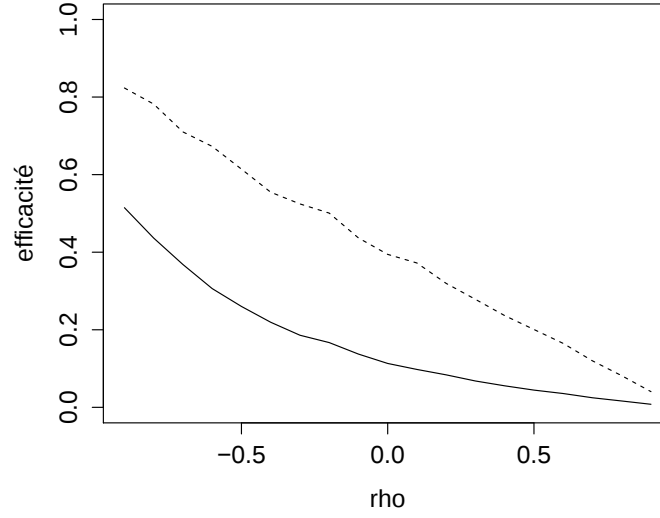


FIGURE 5.3 – Efficacité asymptotique relative estimée de $\tilde{\rho}_n$ et $\hat{\rho}_n(1)$ (ligne pleine) et de $\tilde{\rho}'_n$ et $\hat{\rho}'_n(1)$ (en pointillés), en fonction de ρ^* .

affectées par des ruptures. Les variances asymptotiques des deux estimateurs $\tilde{\rho}_n$ et $\tilde{\rho}'_n$, respectivement $\tilde{\sigma}^2$ et $\tilde{\sigma}'^2$, ont été estimées. Cela a permis d'estimer l'efficacité asymptotique relative respective des deux estimateurs $\tilde{\rho}_n$ et $\tilde{\rho}'_n$ relativement à l'estimateur des moments $\hat{\rho}_n(1)$ (cf. remarque 3.1), c'est-à-dire

$$\frac{1 - \rho^{*2}}{\tilde{\sigma}^2}, \quad \frac{1 - \rho^{*2}}{\tilde{\sigma}'^2}.$$

La figure 5.3 représente les résultats. $\tilde{\rho}'_n$ est meilleur que $\tilde{\rho}_n$ et les performances de ces deux estimateurs décroissent fortement quand ρ^* tend vers 1. L'influence de l'erreur d'estimation de ρ^* dans les performances de la segmentation était visible dans l'étude de simulation de la section 3.5, voir par exemple la figure 3.6. La figure 3.5 semble elle suggérer que pour sélectionner le nombre de ruptures à l'aide du critère mBIC, il vaut mieux décorréler avec un estimateur peu performant que ne pas décorréler du tout quand ρ^* est proche de 1.

Pour résumer, l'estimateur $\tilde{\rho}'_n$ semble avoir de meilleures performances que $\tilde{\rho}_n$. Ces dernières se dégradent fortement quand ρ^* est positivement grand. Toutefois, malgré cette perte d'efficacité des estimateurs du paramètre d'autorégression, il reste préférable de décorréler la série avant de sélectionner le nombre de ruptures par mBIC.

Tester l'hypothèse « $\rho^* = 0$ »

Comme annoncé dans la remarque 3.2, la normalité asymptotique de $\hat{\rho}_n$ permet de tester $H_0 : \rho^* = 0$ contre $H_1 : \rho^* \neq 0$ ou $H'_1 : \rho^* > 0$. La méthode de segmentation après décorrélation introduite au chapitre 3 n'apporte pas d'amélioration substantielle à la segmentation dans le cas $\rho^* < 0$. Le test le plus judicieux pour savoir s'il faut décorréler avant de segmenter semble

être H_0 contre H'_1 .

Proposition 5.2. *Sous les mêmes hypothèses et avec les mêmes notations que la proposition 3.1, si $\rho^* = 0$ alors $\tilde{\sigma}^2 = \tilde{\sigma}_0^2$, où*

$$\tilde{\sigma}_0^2 = \frac{3 - 8P\left(\max(|Y - X|, |Z - X|) \leq \sqrt{2}\Phi^{-1}(3/4)\right)}{2\left[\Phi^{-1}(3/4)\phi\left(\Phi^{-1}(3/4)\right)\right]^2},$$

où X, Y, Z sont trois v.a.r. indépendantes de loi normale centrée réduite, Φ est la fonction de répartition de la loi normale centrée réduite et $\phi = \Phi'$.

Démonstration. $\rho^* = 0$ donc les v.a.r. gaussiennes η_i sont indépendantes. On supposera, sans perdre de généralité, que $\sigma^* = 1$. On remarque que $\forall k \in \mathbb{N}, \mathbb{E}[\Psi(\eta_k, \eta_{k+1}, \eta_{k+2})] = 0$, où Ψ est défini en (3.6). Or les v.a.r. η_i sont indépendantes, donc

$$\forall k \geq 3, \mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2) \Psi(\eta_k, \eta_{k+1}, \eta_{k+2})] = 0.$$

Par conséquent, d'après (3.5),

$$\tilde{\sigma}^2 = \mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2)^2] + 2\mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2) \Psi(\eta_1, \eta_2, \eta_3)] + \mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2) \Psi(\eta_2, \eta_3, \eta_4)].$$

Notons :

$$\begin{aligned} \lambda &= \left[\Phi^{-1}(3/4)\phi\left(\Phi^{-1}(3/4)\right)\right]^{-2}, \\ c &= \sqrt{2}\Phi^{-1}(3/4), \\ p &= P\left(\max(|Y - X|, |Z - X|) \leq \sqrt{2}\Phi^{-1}(3/4)\right). \end{aligned}$$

On remarque que, pour tous $i, j, k, l \in \mathbb{Z}$ tels que $i \neq j$ et $k \neq l$:

- si $\{i, j\} \cap \{k, l\} = \emptyset$, alors $P(|\eta_i - \eta_j| \leq c, |\eta_k - \eta_l| \leq c) = P(|\eta_i - \eta_j| \leq c)P(|\eta_k - \eta_l| \leq c) = 1/4$,
- si $\#(\{i, j\} \cap \{k, l\}) = 1$, alors $P(|\eta_i - \eta_j| \leq c, |\eta_k - \eta_l| \leq c) = p$,
- si $\{i, j\} = \{k, l\}$, alors $P(|\eta_i - \eta_j| \leq c, |\eta_k - \eta_l| \leq c) = P(|\eta_i - \eta_j| \leq c) = 1/2$.

Cela permet de calculer :

$$\begin{aligned} \mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2)^2] &= \lambda(1 - 2p), \\ 2\mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2) \Psi(\eta_1, \eta_2, \eta_3)] &= \lambda\left(\frac{1}{2} - 2p\right), \\ \mathbb{E}[\Psi(\eta_0, \eta_1, \eta_2) \Psi(\eta_2, \eta_3, \eta_4)] &= 0. \end{aligned}$$

D'où $\tilde{\sigma}_0^2 = \lambda\left(\frac{3}{2} - 4p\right)$ et le résultat. □

Remarque 5.2. Le calcul numérique de $\tilde{\sigma}_0^2$ donne 8.567488.

La p-valeur asymptotique du test de H_0 contre H_1 est donc $2 - 2\Phi\left(\sqrt{\frac{n}{\tilde{\sigma}_0^2}}|\tilde{\rho}_n|\right)$. La p-valeur asymptotique du test de H_0 contre H'_1 est $1 - \Phi\left(\sqrt{\frac{n}{\tilde{\sigma}_0^2}}\tilde{\rho}_n\right)$.

On peut faire la même chose avec $\tilde{\rho}'_n$. On a estimé la valeur de $\tilde{\sigma}'^2$ quand $\rho^\star = 0$ par Monte-Carlo (cf. section 5.1.1). La valeur trouvée était 2.53566. Ces estimations, par la méthode de Monte-Carlo, semblent indiquer que, pour tout $\rho^\star$, la variance asymptotique de $\tilde{\rho}'_n$ est plus petite que celle de $\tilde{\rho}_n$. Par conséquent, le test de H_0 contre H_1 , resp. H'_1 , de p-valeur asymptotique $2 - 2\Phi\left(\sqrt{\frac{n}{\tilde{\sigma}'^2}}|\tilde{\rho}'_n|\right)$, resp. $1 - \Phi\left(\sqrt{\frac{n}{\tilde{\sigma}'^2}}\tilde{\rho}'_n\right)$, semble plus puissant que le test qui utilise $\tilde{\rho}_n$ au lieu de $\tilde{\rho}'_n$.

Pour résumer, pour déterminer s'il faut ou non décorréler avant de segmenter une série AR(1), nous recommanderions de tester l'hypothèse $H_0 : \rho^\star = 0$ contre $H'_1 : \rho^\star > 0$, en utilisant le test de p-valeur asymptotique $1 - \Phi\left(\sqrt{\frac{n}{\tilde{\sigma}'^2}}\tilde{\rho}'_n\right)$.

5.1.2 Commentaires sur le chapitre 4

Les commentaires qui suivent sur l'article présenté dans le chapitre 4 portent sur l'estimation des paramètres de l'autorégression, notamment sur quelques idées pour réduire la variance des estimateurs, ainsi que sur le post-traitement à l'ordre $p > 1$ des instants de rupture estimés, qui pose des problèmes sensiblement différents de l'ordre 1.

À propos de l'estimation des paramètres de l'autorégression

Les performances de l'estimateur $\tilde{\Phi}_n^{(p)}$ varient beaucoup sur l'ensemble des paramètres. Cela peut être vu comme un problème de conditionnement de la matrice R_p , définie en (4.9), à inverser. Voilà pourquoi l'emploi d'une version régularisée de l'estimateur est suggérée dans la remarque 4.1. Toutefois, la sélection conjointe du nombre de ruptures et de l'ordre de l'autorégression parvient à éviter en partie ces configurations qui induisent de mauvaises estimations des paramètres de l'autorégression et des instants de rupture (voir le tableau 4.4 et les figures 4.7 et 4.8).

À propos du post-traitement

Le post-traitement des ruptures estimées (cf. section 4.5.1), est de même nature dans le cas des séries AR(p) ou AR(1). La décorrélation induit possiblement p ruptures artéfactuelles adjacentes à droite de la « vraie » rupture estimée. On les élimine donc. Des problèmes peuvent apparaître dans le cas où l'on a deux vraies ruptures à distance inférieure ou égale à p . Dans ce cas, la « vraie » seconde rupture ne peut pas être distinguée des ruptures artéfactuelles engendrées par la première rupture. Cette seconde rupture sera éliminée par le post-traitement, mais le premier des artéfacts qu'elle a engendré se trouvant à une distance supérieure à p de la première rupture, s'il est détecté, sera conservé (voir figure 5.4). Pour éviter ce problème, on peut interdire à des ruptures estimées d'être à distance inférieure ou égale à p , ce qui est cohérent avec le fait que dans les propositions 4.2 et 4.3, l'optimisation du critère se fait

sous contrainte d'une distance entre ruptures qui tend vers l'infini avec la taille de la série (équation (4.18)).

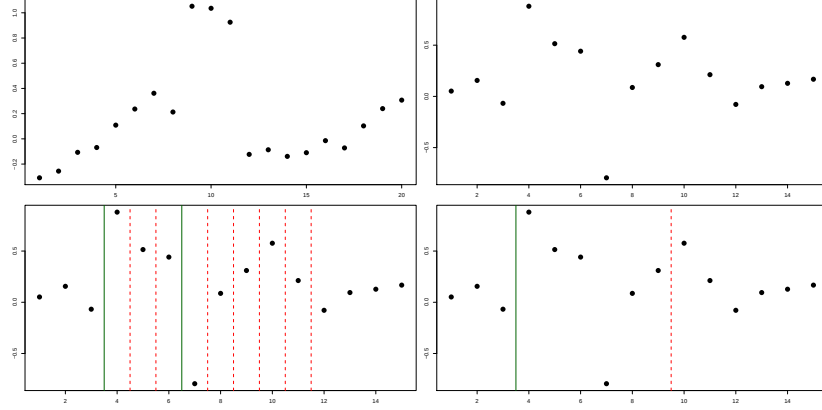


FIGURE 5.4 – En haut à gauche : une série AR(5) affectée de deux ruptures dans l'espérance, à distance 3 l'une de l'autre. En haut à droite : la série décorrelée. En bas à gauche : avant le post-traitement, on détecte des ruptures, qu'il s'agisse de vraies ruptures (vert, trait plein) ou d'artéfacts dûs à la décorrélation (rouge, trait pointillé). En bas à droite : dans cette configuration, le post-traitement élimine les artéfacts de la première rupture, mais aussi la vraie seconde rupture, alors qu'un des artéfacts qui lui est associé est conservé.

5.2 Limites de l'approche adoptée

L'approche proposée n'est pas exempte de certaines limites. Nous les formulerons concernant deux aspects : les limitations algorithmiques et la nature des résultats proposés.

5.2.1 Programmation dynamique et dépendance

Dans cette section, nous discutons la spécificité de la méthode employée au cas d'un bruit autorégressif.

Nous avons proposé une méthode pour la segmentation de l'espérance d'un processus autorégressif. La justification de l'approche adoptée repose en grande partie sur la volonté de conserver l'utilisation de l'algorithme de programmation dynamique pour obtenir la segmentation optimale. Cet algorithme permet de minimiser un critère additif sur les segments (voir section 2.6). Toutefois, cette démarche semble difficilement généralisable. Plaçons-nous dans le modèle général suivant :

$$\mathbf{y} = T(\mathbf{t}_n^*) \boldsymbol{\mu}^* + \sigma \boldsymbol{\eta},$$

où $\sigma > 0$ inconnu, $\mathbf{y}$ et $\boldsymbol{\eta}$ sont des matrices $n \times 1$ aléatoires ; $\boldsymbol{\eta}^T$ est centrée, de carré intégrable et de matrice de variance/covariance Σ ; $\boldsymbol{\mu}^*$ est une matrice $(m+1) \times 1$; $\mathbf{t}_n^* \in \mathcal{A}_{n,m} = \{\mathbf{t} = (t_0, \dots, t_{m+1}) \in \mathbb{R}^{m+2}; 0 = t_0 < t_1 < \dots < t_m < t_{m+1} = n\}$ et pour $\mathbf{t} \in \mathcal{A}_{n,m}$, $T(\mathbf{t})$ est une matrice $n \times (m+1)$ dont la k^e colonne est $\underbrace{(0, \dots, 0)}_{t_{k-1}}, \underbrace{1, \dots, 1}_{t_k - t_{k-1}}, \underbrace{0, \dots, 0}_{n - t_k}^T$.

On cherche à estimer $\mathbf{t}_n^*$ à m et Σ connus. On a :

$$\Sigma^{-1/2} \mathbf{y} = \Sigma^{-1/2} T(\mathbf{t}_n^*) \boldsymbol{\mu}^* + \sigma \boldsymbol{\epsilon} ,$$

où $\boldsymbol{\epsilon}^T$ est une v.a. sur $\mathbb{R}^n$ centrée, de carré intégrable, dont la matrice de variance/covariance est l'identité. On souhaite estimer $\mathbf{t}_n^*$ et $\boldsymbol{\mu}^*$ en minimisant sur $\mathcal{A}_{n,m}$ et l'ensemble des matrices réelles $n \times 1$ le critère des moindres carrés :

$$C(\mathbf{t}, \boldsymbol{\mu}) = \left\| \Sigma^{-1/2} \mathbf{y} - \Sigma^{-1/2} T(\mathbf{t}) \boldsymbol{\mu} \right\|^2 .$$

À $\mathbf{t}$ fixé, on a :

$$\min_{\boldsymbol{\mu}} C(\mathbf{t}, \boldsymbol{\mu}) = \left\| \Sigma^{-1/2} \mathbf{y} - \pi_{\Sigma^{-1/2} T(\mathbf{t})} \Sigma^{-1/2} \mathbf{y} \right\|^2 ,$$

où $\pi_{\Sigma^{-1/2} T(\mathbf{t})}$ est la matrice dans la base canonique de la projection orthogonale sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $\Sigma^{-1/2} T(\mathbf{t})$. Pour que ce critère soit additif sur les segments, nous devons disposer d'une famille de vecteurs $(f(i, j))_{1 \leq i < j \leq n}$ de $\mathbb{R}^n$ telle que pour tout $\mathbf{t}$, $(f(t_{k-1} + 1, t_k))_{1 \leq k \leq m+1}$ est une base orthonormée du sous-espace vectoriel engendré par les colonnes de $\Sigma^{-1/2} T(\mathbf{t})$. Dans ce cas, on a :

$$\min_{\boldsymbol{\mu}} C(\mathbf{t}, \boldsymbol{\mu}) = \left\| \Sigma^{-1/2} \mathbf{y} \right\|^2 - \sum_{k=1}^{m+1} \left\| \pi_{f(t_{k-1}+1, t_k)} \Sigma^{-1/2} \mathbf{y} \right\|^2 ,$$

où $\pi_{f(t_{k-1}+1, t_k)}$ est la matrice dans la base canonique de la projection orthogonale sur la droite engendrée par $f(t_{k-1} + 1, t_k)$, et le critère est bien additif sur les segments. Si Σ est l'identité, l'existence d'une telle famille de vecteurs est triviale. Mais en toute généralité, on ne peut pas obtenir une telle famille de vecteurs pour un Σ symétrique défini positif, éventuellement de Toeplitz, donné. Dans le cas des processus autorégressifs d'ordre fini, la solution proposée à ce problème de dépendance inter-segments fut, *modulo* le conditionnement aux premières observations, d'approcher le sous-espace vectoriel engendré par $\Sigma^{-1/2} T(\mathbf{t})$ par le sous-espace vectoriel engendré par $T(\mathbf{t}')$, où $\mathbf{t}'$ contient des ruptures artéfactuelles, puis de post-traiter la segmentation obtenue. Cette démarche est impossible à généraliser à des processus stationnaires autres que l'AR(p), car c'est seulement dans ce cas-là que les coefficients (i, j) de $\Sigma^{-1/2}$ sont nuls si $|i - j| > p$.

Un autre choix pourrait être d'approcher Σ par $\Sigma(\mathbf{t})$ dépendant de la segmentation $\mathbf{t}$, composée de blocs diagonaux de la taille des segments. Cela revient à ne pas prendre en compte la dépendance inter-segments dans l'inférence. Dans ce cas, les colonnes de $\Sigma^{-1/2}(\mathbf{t}) T(\mathbf{t})$ sont bien orthogonales et le critère est additif sur les segments. Cette additivité est suffisante pour minimiser le critère par l'algorithme de programmation dynamique, mais insuffisante pour permettre l'utilisation de l'algorithme de programmation dynamique élaguée qui requiert des conditions supplémentaires (voir section 2.6.2). Notre approche de segmentation après décorrélation nous a permis de préserver la possibilité d'utiliser la programmation dynamique élaguée. Elle ne peut cependant pas être généralisée à des structures de dépendance plus complexes.

5.2.2 Cadre asymptotique

Le cadre asymptotique permettant d'énoncer les résultats de notre travail peut être critiqué à plusieurs niveaux. Nous nous sommes placés dans le cadre asymptotique désormais classique où la localisation des ruptures dépend linéairement de la longueur des séries :

$$t_{k,n}^* = \lfloor n\tau_k^* \rfloor \quad (1 \leq k \leq m), \quad (5.5)$$

où $t_{1,n}^*, \dots, t_{m,n}^*$ sont les instants des ruptures de la série $y_1, \dots, y_n$. Dans le cas où les observations sont indépendantes, ce cadre est naturel et indolore, nous l'avons déjà évoqué dans la section 1.1.4. Plusieurs résultats (Bai et Perron, 1998 ; Lavielle, 1999 ; Lavielle et Moulines, 2000) sont venus étendre les propriétés des estimateurs de la localisation des ruptures dans le cas où une dépendance sérielle est présente. Si l'on excepte Bai et Perron (1998), qui se placent dans le cadre plus général des modèles linéaires à changements structuraux où certaines formes de dépendance sérielle peuvent être prises en compte dans l'inférence (sans que cela ne soit une obligation), ces résultats impliquent qu'asymptotiquement, des conditions relativement permissives sur cette dépendance sérielle font que cette dernière n'affecte ni la consistance ni la vitesse de convergence des estimateurs des instants de rupture. Si l'on se limite aux processus stationnaires du second ordre, on suppose qu'il existe un tel processus $(\eta_i)_{i \in \mathbb{Z}}$ centré tel que :

$$\forall k \in \{1, \dots, m^* + 1\}, \forall i \in \{t_{k-1,n}^* + 1, \dots, t_{k,n}^*\}, y_i = \mu_k^* + \eta_i, \quad (5.6)$$

où $\mu_1^*, \dots, \mu_{m^*+1}^* \in \mathbb{R}$. Des conditions sont alors imposées à la fonction d'autocovariance γ_η du processus $(\eta_i)_{i \in \mathbb{Z}}$ pour permettre l'établissement des résultats asymptotiques ; on a en particulier $\gamma_\eta(h) \xrightarrow{h \rightarrow \infty} 0$. Par exemple, pour tout $k \in \{1, \dots, m^* + 1\}$,

$$\text{Cov}\left(y_{t_{k-1,n}^*+1}^*, y_{t_{k,n}^*}^*\right) = \gamma_\eta\left(t_{k,n}^* - t_{k-1,n}^* - 1\right) = \gamma_\eta\left(\lfloor n\tau_k^* \rfloor - \lfloor n\tau_{k-1}^* \rfloor - 1\right) \xrightarrow{n \rightarrow \infty} 0.$$

On remarque que γ_η ne dépend pas de n , contrairement aux instants de rupture. Bien qu'il ait permis d'établir des résultats importants, deux critiques peuvent être formulées à ce modèle :

- la première concerne son réalisme pour certains problèmes. En effet, si n représente le nombre de mesures réalisées sur un intervalle de temps fixe, il n'y a aucune raison pour que prendre de nouvelles mesures entre deux précédentes réduise la dépendance entre ces dernières. Par exemple, dans le cas de la météorologie, cela revient à supposer qu'en prenant de nombreuses mesures (de température, etc.) on réduit la portée temporelle de la dépendance météorologique.
- la seconde concerne ce qu'on pourrait appeler un déséquilibre dans les rôles respectifs des ruptures et de la dépendance sérielle. Supposons par exemple que (5.5) ne soit plus vraie et qu'il n'y a qu'une rupture à l'instant t . L'estimateur des moments de l'autocovariance à l'ordre h fixé ne sera pas affecté asymptotiquement par la présence de cette rupture ; cela revient de ce point de vue à considérer t valeurs aberrantes parmi n observations.

Dans le modèle (5.6), on fait le contraire ; on efface partiellement l'effet asymptotique de la dépendance sérielle sur les estimateurs des instants de rupture.

Un choix alternatif consisterait en le modèle où (5.5) reste vrai et on observe $y_1, \dots, y_n$, tels que :

$$\forall k \in \{1, \dots, m^* + 1\}, \forall i \in \{t_{k-1,n}^* + 1, \dots, t_{k,n}^*\}, y_i \stackrel{d}{=} \mu_k^* + X_{i/n}, \quad (5.7)$$

où $(X_t)_{0 < t \leq 1}$ est un processus stochastique centré. Par exemple, on pourrait considérer que $(X_t)_{t \geq 0}$ est un processus d'Ornstein-Uhlenbeck centré et stationnaire, c'est-à-dire un processus gaussien centré de fonction de covariance Γ définie par :

$$\forall (s, t) \in (0, 1], \Gamma(s, t) = v e^{-\alpha|s-t|}, \quad (5.8)$$

où $v, \alpha > 0$, voir par exemple Breiman (1968). C'est l'analogue en temps continu de l'AR(1), dont les discrétisations régulières $(X_{i/n})_{i \in \mathbb{N}}$ sont des projections sur $\mathbb{R}^{\mathbb{N}}$ de processus AR(1) de paramètre d'autorégression $v e^{-\alpha/n}$. À ma connaissance, rien ne permet d'affirmer que l'estimateur des moindres carrés des ruptures converge dans ce cas, et encore moins à la vitesse $1/n$. La figure 5.5 illustre le fait que, contrairement au modèle (5.6) avec $\gamma_\eta(h) \xrightarrow{h \rightarrow \infty} 0$, dans le modèle défini par (5.7), la dépendance sérielle reste visible quand n devient grand, ce qui affecte la visibilité à l'œil nu de la rupture.

Pour qui voudrait montrer la nécessité de prendre en compte la dépendance sérielle dans l'estimation des instants de rupture, il faudrait plutôt se placer dans le modèle (5.7) que dans le modèle (5.6), d'éventuelles conditions permettant la convergence des estimateurs pourraient être plus discriminantes. Nous n'avons pas fait cela. Nous avons proposé une solution pratique et réaliste d'un point de vue calculatoire à la prise en compte de la dépendance sérielle dans l'estimation des instants de ruptures avec un bruit autorégressif, et avons exhibé une classe d'estimateurs des paramètres de l'autorégression dont les propriétés asymptotiques pour des modèles de type (5.6) justifient notre approche.

5.3 Perspectives

Dans cette section sont tracées quelques perspectives possibles à ce travail. En premier lieu, dans la section 5.3.1, on propose l'exploration d'une voie permettant peut-être de lever certaines hypothèses de normalité des observations présentes dans les chapitres 3 et 4. Ensuite, dans la section 5.3.2, est discutée la possibilité de se servir de notre méthode à priori spécifique à la segmentation dans l'espérance de processus avec un bruit autorégressif dans le cas où la structure de dépendance est plus générale et non-paramétrique. Enfin, dans la section 5.3.3, un parallèle est établi avec des problèmes connexes au notre qui pourtant n'impliquent pas de dépendance entre observations.

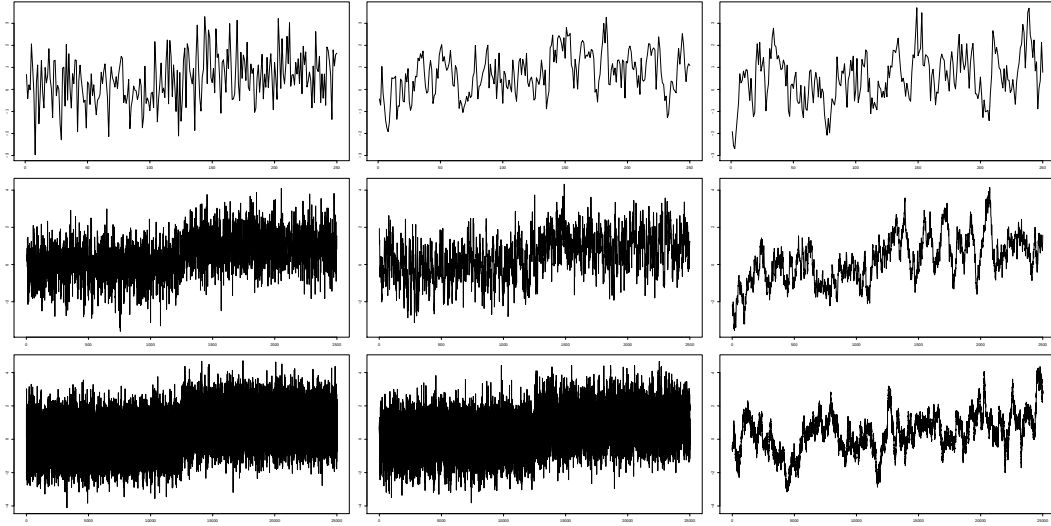


FIGURE 5.5 – On observe $(y_i)_{1 \leq i \leq n} = (\mu_i + \eta_i)_{1 \leq i \leq n}$ où, d'une part, $\mu_i = 0$ si $i \leq n/2$ et $\mu_i = 1$ sinon et, d'autre part, $(\eta_i)_{1 \leq i \leq n}$ est centré, gaussien et de variances marginales 1. Gauche : les η_i sont indépendants. Centre : $\eta_1, \dots, \eta_n$, sont n observations successives d'un processus AR(1) d'autocorrélation 0.7 à l'ordre 1. Droite : $(\eta_i)_{1 \leq i \leq n} = (X_{i/n})_{1 \leq i \leq n}$, où $(X_t)_{t \geq 0}$ est un processus d'Ornstein-Uhlenbeck stationnaire de fonction d'autocovariance donnée par (5.8), avec $v = 1$ et $\alpha = -250 \log(0.7)$. $n = 250$ (haut), $n = 2500$ (milieu) ou $n = 25000$ (bas).

5.3.1 Estimation de la fonction d'autocovariance et d'autocorrélation

Les propriétés asymptotiques des estimateurs des paramètres de l'autorégression proposés n'ont été étudiées que dans le cas gaussien. Dans le chapitre 4 et la section 5.1, les propriétés asymptotiques des estimateurs proposés $\widehat{\Phi}_n^{(p)}$ (cas AR(p)) et $\widetilde{\rho}_n$ (estimateur alternatif dans le cas AR(1)) reposent sur les propriétés asymptotiques des estimateurs de Ma et Genton (2000) de l'autocovariance. Ces propriétés ont été établies dans le cas gaussien par Lévy-Leduc *et al.* (2011). Ainsi, pour étendre les propriétés des estimateurs proposés au cas non-gaussien, il faudrait étendre les propriétés asymptotiques des estimateurs de Ma et Genton (2000). Lévy-Leduc *et al.* (2011) utilisent des résultats de convergence du processus empirique (cf. section 2.2.1) pour déterminer les propriétés asymptotiques des quantiles intervenant dans les estimateurs de Ma et Genton (2000). Un autre type de résultats avec des implications concernant les propriétés asymptotiques des quantiles est les représentations de Bahadur. La spécificité du cas qui nous intéresse est que les quantiles intervenant dans l'estimateur de Ma et Genton (2000) sont des *U-quantiles*.

Représentation de Bahadur pour les U-quantiles

Wendler (2011) établit une représentation de Bahadur pour les U-quantiles de processus strictement stationnaires dépendants réels $\mathbf{X} = (X_i)_{i \in \mathbb{Z}}$, sous certaines conditions de régularité sur la dépendance entre observations. Une représentation de Bahadur est un développement asymptotique des quantiles empiriques sur la fonction de répartition empirique (Kiefer, 1967).

On notera F la fonction de répartition de la loi des X_i . Pour $h \in \mathbb{Z}$, on considèrera aussi les deux processus $\mathbf{X}^{h\pm} = (X_i \pm X_{i+h})_{i \in \mathbb{Z}}$, de loi marginale de fonction de répartition $F^{h\pm}$.

Ces résultats portent sur les U-quantiles, que l'on définit à partir des U-fonctions de répartition empirique. On considèrera des U-fonctions de répartition empirique qui sont des fonctions aléatoires réelles de variable réelle U_n , définies par :

$$\forall t \in \mathbb{R}, U_n(t) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} g(X_i, X_j, t), \quad (5.9)$$

où $g : \mathbb{R}^3 \rightarrow [0, 1]$ est une fonction symétrique en les deux premières variables et croissante en la dernière. La U-fonction de répartition est alors $U : \mathbb{R} \rightarrow [0, 1]$ où, pour tout $t \in \mathbb{R}$, $U(t) = \mathbb{E}[g(A, B, t)]$ avec A et B des v.a.r. indépendantes de même loi que les X_i . Pour $p \in (0, 1)$, le p -U-quantile et le p -U-quantile empirique sont alors respectivement $U^{-1}(p)$ et $U_n^{-1}(p)$. Par exemple, l'estimateur Q_n de Rousseeuw et Croux (1993) est :

$$Q_n = c_{Q,F} U_n^{-1}(1/4), \quad (5.10)$$

où U_n est défini en (5.9) avec

$$\forall (x, y, t) \in \mathbb{R}^3, g(x, y, t) = \mathbb{1}_{|x-y| \leq t}, \quad (5.11)$$

et $c_{Q,F} > 0$ est une constante, dépendante de F . On considèrera aussi l'estimateur Q_n des processus $\mathbf{X}^{h\pm}$ qu'on notera $Q_n^{(h\pm)}$. Les notations $U^{(h\pm)}$, $U_n^{(h\pm)}$ s'en déduisent.

Soit $p \in (0, 1)$. Sous certaines conditions de régularité sur g —remplies si (5.11) est vérifiée—sur U et sur la dépendance sérielle, Wendler (2011) établit une représentation de Bahadur :

$$U_n^{-1}(p) - U^{-1}(p) = \frac{U_n(U^{-1}(p)) - p}{U'(U^{-1}(p))} + R_n. \quad (5.12)$$

Nous ne détaillerons pas ici en quel sens le reste R_n est « petit ». Wendler (2011) en déduit ensuite la convergence en loi vers une loi normale de $\sqrt{n}(U_n^{-1}(p) - U^{-1}(p))$ sous les mêmes hypothèses.

On obtient la convergence et la normalité asymptotique de Q_n . Si X_1 est de carré intégrable, pour que la limite de Q_n soit bien $\sqrt{\text{Var}(X_1)}$, il faut avoir :

$$c_{Q,F} = \frac{\sqrt{\text{Var}(X_1)}}{U^{-1}(1/4)}. \quad (5.13)$$

Il n'est pas nécessaire de connaître F pour déterminer une telle constante, il suffit de connaître un modèle de localisation-dispersion auquel il appartient :

$$F \in \left\{ t \mapsto G\left(\frac{t - \mu}{\sigma}\right); \mu \in \mathbb{R}, \sigma > 0 \right\},$$

où G est une fonction de répartition.

Les estimateurs de Ma et Genton (2000) de l'autocovariance et de l'autocorrélation à l'ordre $h \in \mathbb{Z}$ de $\mathbf{X}$, sont notés respectivement $\widehat{\gamma}_n(h)$ et $\widehat{\rho}_n(h)$ (cf. équations (2.20) et (2.21)).

Comme conséquence de la consistance et de la normalité asymptotique de $Q_n^{h\pm}$, sous les hypothèses du théorème 2 de Wendler (2011) sur les processus $\mathbf{X}^{h\pm}$, on a la consistance et la convergence en $O_P(n^{-1/2})$ de $\widehat{\gamma}_n(h)$ et de $\widehat{\rho}_n(h)$. On remarquera que, dans le cas AR(1), cette propriété suffit à faire converger l'estimateur $\widehat{\rho}_n$ du paramètre d'autorégression ρ^* , défini en (5.1), en $O_P(n^{-1/2})$. Or, il s'agit de la condition énoncée dans la proposition 3.3 pour décorrélérer les séries à segmenter à l'aide de cet estimateur.

Pour établir la normalité asymptotique de $\widehat{\gamma}_n(h)$ et de $\widehat{\rho}_n(h)$, il faudrait établir la normalité asymptotique conjointe des estimateurs $Q_{n-h}^{(h\pm)}$. On peut montrer que cette question se réduit à celle de la normalité asymptotique de U-statistiques sur le processus $(X_i, X_{i+h})_{i \in \mathbb{Z}}$ par l'intermédiaire de l'équation (5.12) et du théorème de Cramér-Wold (Billingsley, 1995, théorème 29.4).

Le problème des constantes

La convergence de Q_n^2 vers la variance du processus dépend du choix de la constante $c_{Q,F}$, qu'il faut fixer comme dans l'équation (5.13). Pour fixer ainsi une telle constante, il faut connaître un modèle de localisation-dispersion auquel appartient F . Par conséquent, la consistance de $\widehat{\gamma}_n(h)$ dépend des constantes $c_{Q,F^{(h\pm)}}$ des estimateurs $Q_{n-h}^{(h\pm)}$, où $F^{(h\pm)}$ est la loi marginale de $\mathbf{X}^{(h\pm)}$. Pour pouvoir les choisir de façon adéquate, il faut donc connaître les modèles de localisation-dispersion auxquels appartiennent respectivement $F^{(h+)}$ et $F^{(h-)}$. Il ne suffit pas de connaître F , il faut avoir un minimum d'information sur la loi des couples (X_i, X_{i+h}) , ce qui n'est pas toujours le cas quand l'on souhaite précisément estimer la covariance de ce couple. Si $\mathbf{X}$ est un processus gaussien, on sait que $F^{(h+)}$ et $F^{(h-)}$ appartiennent au même modèle de localisation-dispersion, celui des lois normales. En dehors de ce cas, cela reste un problème. La consistance de $\widehat{\rho}_n(h)$ ne dépend, pour sa part, que de $c_{Q,F^{(h+)}}/c_{Q,F^{(h-)}}$. Si $\mathbf{X}$ est un processus gaussien, ce rapport vaut 1. Il serait précieux d'être capable d'établir des conditions sous lesquelles ce rapport vaut 1.

5.3.2 Structure de dépendance non-paramétrique

Dans les chapitres 3 et 4, nous avons supposé que le modèle autorégressif considéré était le vrai modèle. Dans la section 4.4.2, en sélectionnant l'ordre de l'autorégression p en même temps que le nombre de ruptures m , nous n'avons supposé vrai qu'un des modèles autorégressifs. De façon plus générale, nous ne sommes pas contraints de supposer qu'un des modèles autorégressifs est le vrai. On peut voir l'estimation dans un modèle autorégressif comme une façon commode d'estimer des coefficients permettant de décorrélérer la série.

En particulier, il faudrait refaire les démonstrations des propositions de la section 4.3 avec des hypothèses plus faibles sur le processus d'innovation $(\epsilon_i)_{i \in \mathbb{Z}}$. C'est possible et cela repose sur les résultats de Lavielle et Moulines (2000) appliqués à la série décorrélée, qui n'exigent pas tout ce que nous avons exigé de $(\epsilon_i)_{i \in \mathbb{Z}}$.

De plus, l'ordre p de l'autorégression à sélectionner ne doit pas nécessairement être le

« bon » ordre, qui d'ailleurs dans ce cas n'existe pas, mais celui qui fournit la meilleure décorrélation. Cette idée est en contradiction avec la philosophie générale du critère BIC, qui cherche à identifier le modèle le plus proche du vrai modèle.

5.3.3 Méthodologie basée sur l'estimation robuste préliminaire d'un paramètre global dans un modèle de segmentation

Le travail proposé est fondé sur une stratégie en deux temps qui a montré son efficacité et peut sans aucun doute être généralisée à d'autres cas. Un exemple est le cas d'observations indépendantes où les observations suivent une loi dépendant de deux paramètres : l'un local, dépendant de chaque segment ; l'autre global, commun aux observations de toute la série. Nous avons évoqué dans la section 2.6 l'impossibilité en général d'utiliser la programmation dynamique pour optimiser un critère de segmentation dans le cas où un paramètre global est présent. L'exemple 2.9 sur la loi binomiale négative permettant d'illustrer ce fait avec des ruptures dans le paramètre d'espérance et un paramètre de surdispersion constant. Il s'agit ici du problème de minimisation suivant :

$$\min_{t_0 < t_1 < \dots < t_{m+1} = n} \min_{r > 0} \min_{\mu = (\mu_1, \dots, \mu_{m+1}) \in]0, +\infty[^{m+1}} \sum_{l=1}^{m+1} \sum_{i=t_{l-1}+1}^{t_l} \log \left\{ \frac{(r + \mu_l)^{r+y_i} y_i! \Gamma(r)}{r^r \mu_l^{y_i} \Gamma(r + y_i)} \right\},$$

où Γ est la fonction Gamma d'Euler. Pour de plus amples développements sur la loi binomiale négative en elle-même, voir Johnson *et al.* (2005). Ce problème est lié à la segmentation de données génomiques de comptage issues du séquençage à haut-débit (*Next-generation sequencing*, NGS). Cleynen (2013) adopte une approche en deux temps : estimation préalable du paramètre de surdispersion r , puis segmentation à r fixé. Ci-dessous, certaines de ses considérations.

Même en l'absence de ruptures, il n'y a pas d'expression explicite de l'estimateur du maximum de vraisemblance du paramètre de surdispersion. L'estimateur de la méthode des moments s'obtient dans ce cas par :

$$\hat{r} = \frac{\bar{y}^2}{S^2 - \bar{y}}, \quad (5.14)$$

avec les notations usuelles de la moyenne et de la variance empiriques. Si on dispose de plusieurs échantillons indépendants de la loi binomiale négative, avec un paramètre de surdispersion r commun, Bliss et Owen (1958) proposent :

- pour chaque échantillon, estimer r par la méthode des moments (5.14),
- estimer r par une moyenne pondérée de ces estimateurs propres à chaque échantillon, les poids privilégiant les estimateurs de petite variance,
- recalculer les poids en fonction de cette estimation de r et recommencer.

Dans un modèle de segmentation, le problème est qu'on ne sait pas où termine l'échantillon d'une loi et débute l'échantillon d'une autre loi. Cleynen (2013) s'inspire de l'approche de Bliss

et Owen (1958) en considérant des fenêtres glissantes sur les données segmentées au lieu des échantillons différents. Au lieu d'une moyenne pondérée, une médiane est utilisée pour rendre robuste l'estimateur à l'effet des fenêtres contenant les instants de rupture. Toutefois, pour le moment, les propriétés d'un estimateur de ce type n'ont pas été établies.

Bibliographie

- Aguilar, E., I. Auer, M. Brunet, T. C. Peterson et J. Wieringa (2003). *Guidelines on climate metadata and homogenization*. Rapp. tech. WMO/TD 53. Genève : World Meteorological Organization (cf. p. 7).
- Akaike, H. (1973). « Information theory and an extension of the maximum likelihood principle ». In : *Proceedings of the 2nd International Symposium on Information Theory*. Sous la dir. de B. N. Petrov et F. Csáki. Budapest : Akadémiai Kiadó, p. 267–281 (cf. p. 34).
- (1974). « A new look at the statistical model identification ». In : *Automatic Control, IEEE Transactions on* 19.6, p. 716–723 (cf. p. 34).
- (1977). « On entropy maximization principle ». In : *Applications of statistics : proceedings of the symposium held at Wright State University, Dayton, Ohio, 14-18 June 1976*. Sous la dir. de P. R. Krishnaiah. North-Holland, p. 27–41 (cf. p. 37).
- (1978a). « A Bayesian analysis of the minimum AIC procedure ». In : *Annals of the Institute of Statistical Mathematics* 30.1, p. 9–14 (cf. p. 37).
- (1978b). « Time series analysis and control through parametric models ». In : *Applied Time Series Analysis : Proceedings of the First Applied Time Series Symposium Held in Tulsa, Oklahoma, May 14-15, 1976*. Sous la dir. de D. F. Findley. New York : Academic Press, p. 1–23 (cf. p. 37).
- Antoch, J., M. Hušková et Z. Prášková (1997). « Effect of dependence on statistics for determination of change ». In : *Journal of Statistical Planning and Inference* 60.2, p. 291–310 (cf. p. 15).
- Arcones, M. A. (1994). « Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors ». In : *The Annals of Probability*, p. 2242–2274 (cf. p. 26, 70, 99).
- Arlot, S. et P. Massart (2009). « Data-driven calibration of penalties for least-squares regression ». In : *The Journal of Machine Learning Research* 10, p. 245–279 (cf. p. 40).
- Auger, I. E. et C. E. Lawrence (1989). « Algorithms for the optimal identification of segment neighborhoods ». In : *Bulletin of Mathematical Biology* 51.1, p. 39–54 (cf. p. 13, 41, 52, 56, 91).
- Azencott, R. et D. Dacunha-Castelle (1986). *Series of Irregular Observations. Forecasting and model building*. Réd. par J. Gani et C. C. Heyde. T. 2. Applied Probability. New York : Springer New York (cf. p. 22, 33, 100).

- Bai, J. (1994). « Least squares estimation of a shift in linear processes ». In : *Journal of Time Series Analysis* 15.5, p. 453–472 (cf. p. 11, 32).
- (1997). « Estimating multiple breaks one at a time ». In : *Econometric Theory* 13.03, p. 315–352 (cf. p. 13).
- Bai, J. et P. Perron (1998). « Estimating and testing linear models with multiple structural changes ». In : *Econometrica*, p. 47–78 (cf. p. 8, 11, 12, 14, 15, 52, 133).
- (2003). « Computation and analysis of multiple structural change models ». In : *Journal of Applied Econometrics* 18.1, p. 1–22 (cf. p. 7, 13, 52).
- Bardet, J.-M. et W. C. Kengne (2014). « Monitoring procedure for parameter change in causal time series ». In : *Journal of Multivariate Analysis* 125, p. 204–221 (cf. p. 32).
- Bardet, J.-M., W. C. Kengne et O. Wintenberger (2012). « Detecting multiple change-points in general causal time series using penalized quasi-likelihood ». In : *Electronic Journal of Statistics*, p. 1–50 (cf. p. 11, 15, 16, 38, 52, 53, 56, 65, 66, 90, 91, 93).
- Barry, D. et J. A. Hartigan (1992). « Product Partition Models for Change Point Problems ». In : *The Annals of Statistics* 20.1, p. 260–279 (cf. p. 10).
- Basseville, M. et I. V. Nikiforov (1993). *The Detection of abrupt changes - Theory and applications*. Prentice-Hall : Information et System sciences series (cf. p. 8, 52).
- Beaulieu, C., J. Chen et J. L. Sarmiento (2012). « Change-point analysis as a tool to detect abrupt climate variations ». In : *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences* 370.1962, p. 1228–1249 (cf. p. 7).
- Beaulieu, C., T. B. M. J. Ouarda et O. Seidou (2007). « Synthèse des techniques d’homogénéisation des séries climatiques et analyse d’applicabilité aux séries de précipitations ». In : *Hydrological sciences journal* 52.1, p. 18–37 (cf. p. 7).
- Bellman, R. (1954). « The theory of dynamic programming ». In : *Bulletin of the American Mathematical Society* 60.6, p. 503–515 (cf. p. 41).
- Beran, J., Y. Feng, S. Ghosh et R. Kulik (2013). *Long-Memory Processes*. Springer-Verlag Berlin Heidelberg (cf. p. 32).
- Berger, J. O. et L. R. Pericchi (2001). « Objective Bayesian Methods for Model Selection : Introduction and Comparison ». In : *Model selection*. Sous la dir. de P. Lahiri. T. Volume 38. Lecture Notes–Monograph Series. Beachwood, Ohio : Institute of Mathematical Statistics, p. 135–207 (cf. p. 38).
- Billingsley, P. (1995). *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley (cf. p. 99, 137).
- Birgé, L. et P. Massart (2001). « Gaussian model selection ». In : *Journal of the European Mathematical Society* 3.3, p. 203–268 (cf. p. 34, 38).
- Bliss, C. I. et A. R. G. Owen (1958). « Negative binomial distributions with a common k ». In : *Biometrika* 45, p. 37–58 (cf. p. 138).
- Blum, M., R. W. Floyd, V. Pratt, R. L. Rivest et R. E. Tarjan (1973). « Time bounds for selection ». In : *Journal of computer and system sciences* 7.4, p. 448–461 (cf. p. 127).

- Boysen, L., A. Kempe, A. Munk, V. Liebscher et O. Wittich (2009). « Consistencies and Rates of Convergence of Jump Penalized Least Squares Estimators ». In : *The Annals of Statistics* 37.1, p. 157–183 (cf. p. 60).
- Braun, J. V., R. K. Braun et H.-G. Müller (2000). « Multiple changepoint fitting via quasilielihood, with application to DNA sequence segmentation ». In : *Biometrika* 87.2, p. 301–314 (cf. p. 7, 38, 52, 90).
- Braun, J. V. et H.-G. Müller (1998). « Statistical methods for DNA sequence segmentation ». In : *Statistical Science*, p. 142–162 (cf. p. 7, 52, 90).
- Breiman, L. (1968). *Probability*. Classics in Applied Mathematics. Society for Industrial et Applied Mathematics (cf. p. 134).
- Brockwell, P. J. et R. A. Davis (1991). *Time Series : Theory and Methods*. Springer series in statistics. Springer-Verlag (cf. p. 19, 23, 55, 85, 92, 96).
- Chakar, S., É. Lebarbier, C. Lévy-Leduc et S. Robin (2014). « A robust approach for estimating change-points in the mean of an AR(1) process ». In : *arXiv eprint 1403.1958* (cf. p. 95).
- Chen, H. et N. R. Zhang (2015). « Graph-based change-point detection ». In : *The Annals of Statistics* 43.1, p. 139–176 (cf. p. 7).
- Chernoff, H. et S. Zacks (1964). « Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time ». In : *Ann. Math. Statist.* 35.3, p. 999–1018 (cf. p. 10).
- Choi, B. (1992). *ARMA Model Identification*. Réd. par J. Gani, C. C. Heyde, J. Berger, S. Fienberg, K. Krickeberg, I. Olkin et B. Singer. Springer Series in Statistics. New York : Springer US (cf. p. 37).
- Cleynen, A. (2013). « Approches statistiques en segmentation : application à la ré-annotation de génome ». Thèse de doct. Université Paris Sud-Paris XI (cf. p. 138).
- Cleynen, A., M. Koskas, É. Lebarbier, G. Rigai et S. Robin (2014). « Segmentor3IsBack : an R package for the fast and exact segmentation of Seq-data. » In : *Algorithms for Molecular Biology* 9, p. 6 (cf. p. 45).
- Cobb, G. W. (1978). « The problem of the Nile : conditional solution to a changepoint problem ». In : *Biometrika* 65.2, p. 243–251 (cf. p. 7).
- Croux, C. et P. J. Rousseeuw (1992). « Time-Efficient Algorithms for Two Highly Robust Estimators of Scale ». In : *Computational Statistics*. Sous la dir. d'Y. Dodge et J. Whittaker. Physica-Verlag HD, p. 411–428 (cf. p. 28).
- Csörgő, S. et J. Mielniczuk (1996). « The empirical process of a short-range dependent stationary sequence under Gaussian subordination ». In : *Probability Theory and Related Fields* 104 (1), p. 15–25 (cf. p. 24, 69).
- Davis, R. A., S. Hancock et Y.-C. Yao (2010). « Consistency of Minimum Description Length model selection for piecewise autoregressions ». In : *Preprint* (cf. p. 16).

- Davis, R. A., T. C. M. Lee et G. A. Rodriguez-Yam (2006). « Structural break estimation for nonstationary time series models ». In : *Journal of The American Statistical Association* 101.473, p. 223–239 (cf. p. 13, 15, 16, 53).
- Davis, R. A. et C. Y. Yau (2013). « Consistency of minimum description length model selection for piecewise stationary time series models ». In : *Electronic Journal of Statistics* 7, p. 381–411 (cf. p. 16).
- Durrett, R. (2010). *Probability : theory and examples*. Cambridge university press (cf. p. 31, 81).
- Eckley, I. A., P. Fearnhead et R. Killick (2011). « Analysis of changepoint models. » In : *Bayesian time series models*. Sous la dir. de D. Barber, A. T. Cemgil et S. Chiappa. Cambridge University Press, p. 205–224 (cf. p. 13).
- Ehsanzadeh, E., T. B. M. J. Ouarda et H. M. Saley (2011). « A simultaneous analysis of gradual and abrupt changes in Canadian low streamflows ». In : *Hydrological Processes* 25.5, p. 727–739 (cf. p. 7).
- Fearnhead, P. (2006). « Exact and efficient Bayesian inference for multiple changepoint problems ». In : *Statistics and computing* 16.2, p. 203–213 (cf. p. 10).
- Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. 2*. second. New York : Wiley (cf. p. 55).
- Frick, K., A. Munk et H. Sieling (2014). « Multiscale change point inference ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 76.3, p. 495–580 (cf. p. 12, 94).
- Fryzlewicz, P. (2014). « Wild Binary Segmentation for multiple change-point detection ». In : *The Annals of Statistics* 42.6, p. 2243–2281 (cf. p. 13).
- Gazeaux, J., S. Williams, M. King, M. Bos, R. Dach, M. Deo, A. W. Moore, L. Ostini, E. Petrie, M. Roggero, F. N. Teferle, G. Olivares et F. H. Webb (2013). « Detecting offsets in GPS time series : First results from the detection of offsets in GPS experiment ». In : *Journal of Geophysical Research (Solid Earth)* 118.5 (cf. p. 7, 52, 90).
- Geweke, J. et Y. Jiang (2011). « Inference and prediction in a multiple-structural-break model ». In : *Journal of Econometrics* 163.2, p. 172–185 (cf. p. 7).
- Guédon, Y. (2013). « Exploring the latent segmentation space for the assessment of multiple change-point models ». In : *Computational Statistics* 28.6, p. 2641–2678 (cf. p. 42, 43).
- Hájek, J. et A. Rényi (1955). « Generalization of an inequality of Kolmogorov ». In : *Acta Mathematica Hungarica* 6.3, p. 281–283 (cf. p. 31).
- Hannan, E. J. (1980). « The estimation of the order of an ARMA process ». In : *The Annals of Statistics*, p. 1071–1081 (cf. p. 37).
- Harchaoui, Z. et C. Lévy-Leduc (2010). « Multiple change-point estimation with a total variation penalty ». In : *Journal of The American Statistical Association* 105.492 (cf. p. 8, 13, 60).

- He, H. et T. A. Severini (2010). « Asymptotic properties of maximum likelihood estimators in models with multiple change points ». In : *Bernoulli* 16.3, p. 759–779 (cf. p. 11, 12).
- Hocking, T. D., G. Schleiermacher, I. Janoueix-Lerosey, V. Boeva, J. Cappel, O. Delattre, F. Bach et J.-P. Vert (2013). « Learning smoothing models of copy number profiles using breakpoint annotations ». In : *BMC bioinformatics* 14.1, p. 164 (cf. p. 7).
- Hurvich, C. M. et C.-L. Tsai (1989). « Regression and time series model selection in small samples ». In : *Biometrika* 76.2, p. 297–307 (cf. p. 34).
- Jackson, B., J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumoussis, E. Gwin, P. Sangtrakulcharoen, L. Tan et T. T. Tsai (2005). « An algorithm for optimal partitioning of data on an interval ». In : *Signal Processing Letters, IEEE* 12.2, p. 105–108 (cf. p. 13).
- Jandhyala, V., S. Fotopoulos, I. MacNeill et P. Liu (2013). « Inference for single and multiple change-points in time series ». In : *Journal of Time Series Analysis*, n/a–n/a (cf. p. 13).
- Jarušková, D. (1997). « Some Problems with Application of Change-Point Detection Methods to Environmental Data ». In : *Environmetrics* 8.5, p. 469–483 (cf. p. 7).
- Johnson, N. L., A. W. Kemp et S. Kotz (2005). *Univariate discrete distributions*. T. 444. John Wiley & Sons (cf. p. 138).
- Kaplan, A. Y. et S. . Shishkin (2000). « Application of the change-point analysis to the investigation of the brain's electrical activity ». In : *Non-Parametric Statistical Diagnosis*. Sous la dir. de B. Brodsky et B. S. Darkhovsky. T. 509. Mathematics and Its Applications. Springer Netherlands, p. 333–388 (cf. p. 7).
- Kass, R. E. et A. E. Raftery (1995). « Bayes factors ». In : *Journal of The American Statistical Association* 90.430, p. 773–795 (cf. p. 36, 37).
- Kenyeres, A. et C. Bruyninx (2004). « EPN coordinate time series monitoring for reference frame maintenance ». In : *GPS Solutions* 8.4, p. 200–209 (cf. p. 7).
- Kiefer, J. (1967). « On Bahadur's Representation of Sample Quantiles ». In : *The Annals of Mathematical Statistics* 38.5, p. 1323–1342 (cf. p. 135).
- Killick, R., P. Fearnhead et I. A. Eckley (2012). « Optimal detection of changepoints with a linear computational cost ». In : *Journal of The American Statistical Association* 107.500, p. 1590–1598 (cf. p. 13, 52, 91).
- Kolmogorov, A. (1928). « Über die Summen durch den Zufall bestimmter unabhängiger Grössen ». In : *Mathematische Annalen* 99.1, p. 309–319 (cf. p. 31).
- Lai, T. L., H. Liu et H. Xing (2005). « Autoregressive models with piecewise constant volatility and regression parameters ». In : *Statistica Sinica* 15, p. 279–301 (cf. p. 52).
- Lai, W. R., M. D. Johnson, R. Kucherlapati et P. J. Park (2005). « Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data ». In : *Bioinformatics* 21.19, p. 3763 (cf. p. 60).
- Lavielle, M. (1999). « Detection of multiple changes in a sequence of dependent variables ». In : *Stochastic Processes and their Applications* 83.1, p. 79–102 (cf. p. 12, 14, 15, 32, 38, 52, 90, 133).

- Lavielle, M. (2005). « Using penalized contrasts for the change-point problem ». In : *Signal Processing* 85.8, p. 1501–1510 (cf. p. 7, 40, 52, 90).
- Lavielle, M. et É. Lebarbier (2001). « An application of MCMC methods for the multiple change-points problem ». In : *Signal Processing* 81.1, p. 39–53 (cf. p. 8, 10).
- Lavielle, M. et É. Moulines (2000). « Least-squares Estimation of an Unknown Number of Shifts in a Time Series ». In : *Journal of Time Series Analysis* 21.1, p. 33–59 (cf. p. 11, 12, 14–16, 32, 38, 52, 53, 57, 58, 71, 73–75, 83, 90, 91, 94, 101, 103, 104, 133, 137).
- Lebarbier, É. (2005). « Detecting multiple change-points in the mean of Gaussian process by model selection ». In : *Signal processing* 85.4, p. 717–736 (cf. p. 12, 38, 52).
- Lebarbier, É. et T. Mary-Huard (2006). « Une introduction au critere BIC : fondements théoriques et interprétation ». In : *Journal de la Société Française de Statistique* 147.1, p. 39–57 (cf. p. 36, 37).
- Lee, C.-B. (1995). « Estimating the number of change points in a sequence of independent normal random variables ». In : *Statistics & Probability Letters* 25.3, p. 241–248 (cf. p. 12, 38).
- Lévy-Leduc, C., H. Boistard, É. Moulines, M. S. Taqqu et V. A. Reisen (2011). « Robust estimation of the scale and of the autocovariance function of Gaussian short-and long-range dependent processes ». In : *Journal of Time Series Analysis* 32.2, p. 135–156 (cf. p. 28, 29, 54, 99, 135).
- Li, S. et R. Lund (2012). « Multiple changepoint detection via genetic algorithms ». In : *Journal of Climate* 25.2, p. 674–686 (cf. p. 52).
- Lu, Q., R. Lund et T. C. M. Lee (2010). « An MDL approach to the climate segmentation problem ». In : *The Annals of Applied Statistics* 4.1, p. 299–319 (cf. p. 52, 90).
- Ma, Y. et M. G. Genton (2000). « Highly Robust Estimation of the Autocovariance Function ». In : *Journal of Time Series Analysis* 21.6, p. 663–684 (cf. p. 28, 29, 53, 54, 60, 61, 91, 92, 127, 135, 136).
- Maidstone, R., T. Hocking, G. Rigaiil et P. Fearnhead (2014). « On Optimal Multiple Changepoint Algorithms for Large Data ». In : *arXiv eprint 1409.1842* (cf. p. 13, 48).
- Mallows, C. L. (1973). « Some comments on C_p ». In : *Technometrics* 15.4, p. 661–675 (cf. p. 34).
- Maronna, R. A., D. Martin et V. J. Yohai (2006). *Robust statistics*. John Wiley & Sons, Chichester (cf. p. 27, 125).
- Massart, P. (2007). *Concentration inequalities and model selection*. T. 10. Springer (cf. p. 34, 38).
- Mestre, O. (2000). « Méthodes statistiques pour l’homogénéisation de longues séries climatiques ». Thèse de doct. Université Toulouse 3 (cf. p. 52, 90).
- Olshen, A. B., E. S. Venkatraman, R. Lucito et M. Wigler (2004). « Circular binary segmentation for the analysis of array-based DNA copy number data ». In : *Biostatistics* 5.4, p. 557–572 (cf. p. 13).

- Page, E. S. (1954). « Continuous inspection schemes ». In : *Biometrika*, p. 100–115 (cf. p. 8).
- (1955). « A test for a change in a parameter occurring at an unknown point ». In : *Biometrika*, p. 523–527 (cf. p. 8).
- Perfetti, N. (2006). « Detection of station coordinate discontinuities within the Italian GPS Fiducial Network ». In : *Journal of Geodesy* 80.7, p. 381–396 (cf. p. 7).
- Perreault, L., J. Bernier, B. Bobée et É. Parent (2000). « Bayesian change-point analysis in hydrometeorological time series. » In : *Journal of Hydrology* 235.3, p. 221–263 (cf. p. 7).
- Pesaran, H., D. Pettenuzzo et A. Timmermann (2006). « Forecasting time series subject to multiple structural breaks ». In : *The Review of Economic Studies* 73.4, p. 1057–1084 (cf. p. 7).
- Peterson, T. C., D. R. Easterling, T. R. Karl, P. Groisman, N. Nicholls, N. Plummer, S. Torok, I. Auer, R. Boehm et D. Gullett (1998). « Homogeneity adjustments of in situ atmospheric climate data : a review ». In : *International Journal of Climatology* 18.13, p. 1493–1517 (cf. p. 7).
- Picard, F. (2007). *An Introduction to process segmentation*. Research Report 6. Statistics for Systems Biology (cf. p. 13).
- Picard, F., É. Lebarbier, M. Hoebeker, G. Rigai, B. Thiam et S. Robin (2011). « Joint segmentation, calling, and normalization of multiple CGH profiles ». In : *Biostatistics* 12.3, p. 413–428 (cf. p. 12, 57).
- Picard, F., S. Robin, M. Lavielle, C. Vaisse et J.-J. Daudin (2005). « A statistical approach for array CGH data analysis ». In : *BMC bioinformatics* 6.1, p. 27 (cf. p. 7, 52, 90).
- Rao, C. R. et Y. Wu (2001). « On model selection ». In : *Model selection*. Sous la dir. de P. Lahiri. T. Volume 38. Lecture Notes–Monograph Series. Beachwood, Ohio : Institute of Mathematical Statistics, p. 1–64 (cf. p. 33).
- Rigai, G. (2010). « Pruned dynamic programming for optimal multiple change-point detection ». In : *arXiv eprint 1004.0887* (cf. p. 13, 45, 48, 52, 91).
- Rigai, G., É. Lebarbier et S. Robin (2012). « Exact posterior distributions and model selection criteria for multiple change-point detection problems ». In : *Statistics and Computing* 22.4, p. 917–929 (cf. p. 10, 12, 40).
- Robbins, M., C. Gallagher, R. Lund et A. Aue (2011). « Mean shift testing in correlated data ». In : *Journal of Time Series Analysis* 32.5, p. 498–511 (cf. p. 15).
- Robert, C. (2006). *Le choix bayésien : Principes et pratique*. Springer Science & Business Media (cf. p. 35, 36).
- Rousseeuw, P. J. et C. Croux (1993). « Alternatives to the Median Absolute Deviation ». In : *Journal of The American Statistical Association* 88.424, p. 1273–1283 (cf. p. 28, 54, 92, 125, 127, 136).
- Schwarz, G. (1978). « Estimating the dimension of a model ». In : *The Annals of Statistics*, p. 461–464 (cf. p. 10, 36, 39, 96).

- Scott, A. J. et M. Knott (1974). « A cluster analysis method for grouping means in the analysis of variance ». In : *Biometrics*, p. 507–512 (cf. p. 13).
- Sheather, S. J. et M. C. Jones (1991). « A reliable data-based bandwidth selection method for kernel density estimation ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, p. 683–690 (cf. p. 98).
- Shevlyakov, G., N. Lyubomishchenko et P. Smirnov (2014). « A Few Remarks on Robust Estimation of Power Spectra ». In : *Austrian Journal of Statistics* 43.4, p. 237–245 (cf. p. 23).
- Stock, J. H. et M. W. Watson (1996). « Evidence on structural instability in macroeconomic time series relations ». In : *Journal of Business & Economic Statistics* 14.1, p. 11–30 (cf. p. 7).
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. 3. Cambridge Univ Pr (cf. p. 23, 24, 69, 83).
- Venkatraman, E. S. (1992). « Consistency results in multiple change-point problems ». Thèse de doct. Stanford University (cf. p. 13).
- Wendler, M. (2011). « Bahadur representation for U-quantiles of dependent data ». In : *Journal of Multivariate Analysis* 102.6, p. 1064–1079 (cf. p. 135–137).
- Williams, S. (2003). « Offsets in Global Positioning System time series ». In : *Journal of Geophysical Research (Solid Earth)* 108 (cf. p. 7, 52, 90).
- Yao, Y.-C. (1988). « Estimating the number of change-points via Schwarz' criterion ». In : *Statistics & Probability Letters* 6.3, p. 181–189 (cf. p. 12, 39, 52, 97).
- Yao, Y.-C. et S. T. Au (1989). « Least-squares estimation of a step function ». In : *Sankhyā : The Indian Journal of Statistics, Series A*, p. 370–381 (cf. p. 12).
- Zhang, N. R. (2005). « Change-point detection and sequence alignment : statistical problems of genomics ». Thèse de doct. Stanford University (cf. p. 39, 94, 95).
- Zhang, N. R. et D. O. Siegmund (2007). « A modified Bayes Information Criterion with applications to the analysis of Comparative Genomic Hybridization data ». In : *Biometrics* 63.1, p. 22–32 (cf. p. 10, 12, 16, 39, 52, 53, 57, 58, 91, 94, 125).

Liste des illustrations

1.1	Séries dans différents modèles de segmentation gaussiens.	9
3.1	Post-traitement des instants de rupture estimés, cas AR(1).	59
3.2	Boîtes à moustaches de $\tilde{\rho}_n - \rho^*$ et de $\hat{\rho}_{MG} - \rho^*$, sans ruptures dans la série. . .	61
3.3	Boîtes à moustaches de $\tilde{\rho}_n - \rho^*$ et de $\hat{\rho}_{MG} - \rho^*$, avec ruptures dans la série. . .	62
3.4	Boîtes à moustaches du nombre estimé de ruptures, $\sigma^* = 0.1$	63
3.5	Boîtes à moustaches du nombre estimé de ruptures, $\sigma^* = 0.5$	63
3.6	Boîtes à moustaches de la première partie de la distance de Hausdorff, $\sigma^* = 0.5$. . .	64
3.7	Boîtes à moustaches de la seconde partie de la distance de Hausdorff, $\sigma^* = 0.1$. . .	64
3.8	Boîtes à moustaches de la seconde partie de la distance de Hausdorff, $\sigma^* = 0.5$. . .	64
3.9	Fréquence des instants de rupture estimés, $\sigma^* = 0.1$	65
3.10	Fréquence des instants de rupture estimés, $\sigma^* = 0.5$	65
3.11	Fréquence des instants de rupture estimés, comparaison avec Bardet, Kengne, and Wintenberger (2012).	66
3.12	Résultats de la méthode pour l'AR(1) dans le cas d'un bruit AR(2).	67
3.13	Boîtes à moustaches pour l'estimateur de l'autocorrélation à l'ordre 1, cas de la loi de Cauchy.	67
4.1	Boîtes à moustaches du nombre de ruptures estimé, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) =$ $(-1.2, -0.4, 0.4)$	109
4.2	Fréquence des instants de rupture estimés, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) = (-1.2, -0.4, 0.4)$. . .	110
4.3	Boîtes à moustaches du nombre de ruptures estimé, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) =$ $(1.6, -0.8, 0.4)$	111
4.4	Fréquence des instants de rupture estimés, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) = (1.6, -0.8, 0.4)$. . .	112
4.5	Boîtes à moustaches du nombre de ruptures estimé, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) =$ $(0.2, 0.2, 0.4)$	113
4.6	Fréquence des instants de rupture estimés, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.2, 0.4)$. . .	114
4.7	Boîtes à moustaches du nombre de ruptures estimé, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) =$ $(0.2, 0.6, 0.4)$	115
4.8	Fréquence des instants de rupture estimés, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.6, 0.4)$. . .	116

4.9	Boîtes à moustaches du nombre de ruptures estimé, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) = (0.4, 0.2, 0.2)$	117
4.10	Fréquence des instants de rupture estimés, cas AR(2), $(\phi_1^*, \phi_2^*, \sigma^*) = (0.4, 0.2, 0.2)$	118
4.11	Boîtes à moustaches du nombre de ruptures estimé, cas AR(5), $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0.5, -0.5, 0.4)$	119
4.12	Fréquence des instants de rupture estimés, cas AR(5), $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0.5, -0.5, 0.4)$	120
4.13	Boîtes à moustaches du nombre de ruptures estimé, cas AR(5), $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0, -0.5, 0.4)$	121
4.14	Fréquence des instants de rupture estimés, cas AR(5), $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0, -0.5, 0.4)$	122
5.1	Exemple de série où la décorrélation permet d'éviter la sur-segmentation.	125
5.2	Surcontamination des différences entre observations dans un modèle de segmentation.	126
5.3	Efficacité asymptotique relative des estimateurs du paramètre d'autorégression.	128
5.4	Post-traitement, cas AR(p).	131
5.5	Comparaison des cas où le bruit est un bruit blanc, est AR(1) et est la discrétisation d'un processus d'Ornstein-Uhlenbeck pour des séries avec rupture.	135

Liste des tableaux

4.1	Résultats dans le cas $\text{AR}(2)$, $(\phi_1^*, \phi_2^*, \sigma^*) = (-1.2, -0.4, 0.4)$	109
4.2	Résultats dans le cas $\text{AR}(2)$, $(\phi_1^*, \phi_2^*, \sigma^*) = (1.6, -0.8, 0.4)$	111
4.3	Résultats dans le cas $\text{AR}(2)$, $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.2, 0.4)$	113
4.4	Résultats dans le cas $\text{AR}(2)$, $(\phi_1^*, \phi_2^*, \sigma^*) = (0.2, 0.6, 0.4)$	115
4.5	Résultats dans le cas $\text{AR}(2)$, $(\phi_1^*, \phi_2^*, \sigma^*) = (0.4, 0.2, 0.2)$	117
4.6	Résultats dans le cas $\text{AR}(5)$, $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0.5, -0.5, 0.4)$. .	119
4.7	Résultats dans le cas $\text{AR}(5)$, $(\phi_1^*, \phi_2^*, \phi_3^*, \phi_4^*, \phi_5^*, \sigma^*) = (0.5, 0, 0, 0, -0.5, 0.4)$. . .	121

Notations

Abréviations latines usuelles en français ou en anglais : *cf.*, *e.g.*, *et al.*, *i.e.*

Abréviations en français

càdlàg	continu à droite et limité à gauche
i.i.d.	indépendants et identiquement distribués
resp.	respectivement
v.a.	variable aléatoire
v.a.r.	variable aléatoire réelle

Abréviations en anglais

AIC	Akaike information criterion
AICc	Akaike information criterion corrected
AO	Additive outliers
ARMA	Autoregressive-moving-average
BIC	Bayesian information criterion
cdf	cumulative distribution function
DP	Dynamic programming
iid	independent and identically distributed
IQR	Interquartile range
Lasso	Least absolute shrinkage and selection operator
MAD	Median absolute deviation
mBIC	Modified Bayesian information criterion
MDL	Minimum description length
NGS	Next-generation sequencing
rv	random variable
SNIP	Segment neighborhood with inequality pruning

Symboles

Dans tout le document, les variables aléatoires ont pour espace de départ un même espace probabilisé $(\Omega, \mathcal{A}, P)$.

$\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$	ensembles usuels
$\delta_{x,y}$	Symbole égal à 1 si et seulement si $x = y$, égal à 0 sinon.
$\#$	cardinal
C_n^k	coefficient binomial k parmi n
$[a, b)$	intervalle $\{x \in \mathbb{R}; a \leq x < b\}$
$\arg \min, \arg \max$	argument resp. minimum et maximum
$\log$	logarithme népérien
$[\cdot], \lceil \cdot \rceil$	partie entière resp. inférieure et supérieure
$\ \cdot \ $	Norme, supposée euclidienne sauf mention du contraire.
π_E	projection orthogonale sur le sous-espace vectoriel E
A^T	transposée de la matrice A
$\lim$	limite
$\liminf$	limite inférieure
$u_n = O(v_n), u_n = o(v_n) \ (n \rightarrow \infty)$	(u_n) est resp. dominée par (v_n) et négligeable devant (v_n) en $+\infty$.
$\mathbb{E}$	espérance
Var	variance
Cov	covariance
$\mathcal{N}(m, \sigma^2)$	loi normale d'espérance m et de variance σ^2
$\mathcal{N}(\mathbf{m}, \Sigma)$	loi normale multivariée d'espérance $\mathbf{m}$ et de matrice de variance/covariance Σ
$\chi^2(d)$	loi du chi-deux à d degrés de liberté
$\stackrel{p.s.}{=}$	égalité presque-sûre
$\stackrel{d}{=}$	égalité en loi
$\xrightarrow[n \rightarrow \infty]{d}$	convergence en loi
$\xrightarrow[n \rightarrow \infty]{P}$	convergence en probabilité
$X_n = O_P(Y_n), o_P(Y_n) \ (n \rightarrow \infty)$	(X_n/Y_n) est resp. bornée en probabilité et convergente en probabilité vers 0 en $+\infty$.