Université de Montréal

Méthode bayésienne de détection de rupture et/ou de tendance pour des données temporelles

par

Alexandre Leroux

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en statistique

14 avril 2016

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Méthode bayésienne de détection de rupture et/ou de tendance pour des données temporelles

présenté par

Alexandre Leroux

a été évalué par un jury composé des personnes suivantes :

Mylène Bédard

(président-rapporteur)

Jean-François Angers

(directeur de recherche)

Alejandro Murua

(membre du jury)

Mémoire accepté le

13 avril 2016

SOMMAIRE

Ce mémoire a pour but de déterminer des nouvelles méthodes de détection de rupture et/ou de tendance. Après une brève introduction théorique sur les splines, plusieurs méthodes de détection de rupture existant déjà dans la littérature seront présentées. Puis, de nouvelles méthodes de détection de rupture qui utilisent les splines et la statistique bayésienne seront présentées. De plus, afin de bien comprendre d'où provient la méthode utilisant la statistique bayésienne, une introduction sur la théorie bayésienne sera présentée. À l'aide de simulations, nous effectuerons une comparaison de la puissance de toutes ces méthodes. Toujours en utilisant des simulations, une analyse plus en profondeur de la nouvelle méthode la plus efficace sera effectuée. Ensuite, celle-ci sera appliquée sur des données réelles. Une brève conclusion fera une récapitulation de ce mémoire.

Mots clés : Spline ; Statistique bayésienne ; Rupture abrupte ; Déclenchement de tendance ; Tendance.

SUMMARY

This thesis aims to identify new change-point detection methods and/or trend in temporal data. After a brief theoretical introduction on splines, several existing change-point detection already in the literature will be presented. Then, new change-point detection methods using splines and Bayesian statistics will be presented. Moreover, in order to understand the method using Bayesian statistics, an introduction to Bayesian theory will be presented. Using simulations, we will make a comparison of the power of all these methods. Still using simulations, an analysis of the new most effective method will be performed. Then, this method will be applied to real data. A brief conclusion will make a summary of this thesis.

Keywords : Spline ; Bayesian statistics ; Sudden change ; Continuous change ; Trend.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

En premier lieu, j'aimerais remercier mon directeur de recherche, monsieur Jean-François Angers, pour sa disponibilité, son aide et son soutien tout au long de ce mémoire et surtout même après plus de deux années de rédaction. Il a su m'expliquer clairement les théories avec lesquelles j'étais moins à l'aise et il m'a aidé à m'éclairer l'esprit lorsque j'avais des difficultés de compréhension. Sans lui, je n'aurais certainement pas pu terminer ma rédaction.

Mes remerciements vont aussi à mes parents ainsi qu'à mes frères et soeurs pour leur soutien inconditionnel et leurs encouragements à ne pas abandonner lorsque la motivation n'y était pas. Je tiens aussi à remercier mes amis et mes collègues de travail (qui se reconnaîtront) pour leurs mots d'encouragement ainsi que la confiance qu'ils avaient en ma capacité à terminer ce mémoire.

J'aimerais aussi remercier Rosalie, celle qui a toujours su trouver les mots pour me forcer à me dépasser et, sans qui, je n'aurais pas eu le courage de continuer à persévérer afin de compléter ma rédaction.

Enfin, je tiens à remercier les membres du jury pour leurs commentaires construstifs et éclairés.

INTRODUCTION

Dans la littérature, une rupture est définie comme un changement, dans une série temporelle, dans le comportement des données à partir d'un certain point. La majeure partie du temps, la moyenne des données avant et après la rupture ne seront pas les mêmes. Cela s'appelle une rupture abrupte. Certaines ruptures se passent en deux points où il y a une transition entre ceux-ci. Dans d'autres cas, une tendance apparait après la rupture, d'où le nom de déclenchement de tendance. Dans ce mémoire, nous nous intéressons particulièrement aux ruptures abruptes, aux déclenchements de tendance et aux tendances.

Plusieurs tests de ruptures existent déjà dans la littérature, tous aussi différents les uns des autres. Dans ce mémoire, nous en avons sélectionné quatre qui seront présentés et décrits dans le chapitre 2. Nous voulons découvrir un nouveau test de rupture utilisant les splines et la statistique bayésienne. Les splines seront expliquées dans le chapitre 1 tandis que la statistique bayésienne le sera dans le chapitre 3.

Nous avons donc créé trois tests de rupture qui seront décrits dans le chapitre 3. Ceux-ci seront comparés aux test présentés dans le chapitre 2 à l'aide de simulations de plusieurs types de rupture qui tiendront compte de plusieurs facteurs. Ces simulations ainsi que les résultats de celles-ci sont présentés dans le chapitre 4.

Il n'existe pas, dans la littérature, de méthode de détection de rupture qui utilise les splines et la statistique bayésienne. Nous sommes donc motivés à trouver une méthode utilisant cela qui serait aussi efficace que celles déjà existantes et acceptées dans la littérature.

Notre objectif est qu'au moins un de ces trois tests soit environ égal en efficacité et en fiabilité aux méthodes déjà existantes dans la littérature. Une étude plus poussée sera effectuée dans le chapitre 4 sur le meilleur de ces trois nouveaux tests. Puis, celui-ci sera appliqué sur des données réelles dans le chapitre 5 et sera comparé au meilleur test existant dans la littérature afin de vérifier si leurs résultats sont similaires.

Enfin, un bref retour sur chacun des chapitres ainsi que sur les résultats observés sera effectué dans la conclusion.

Chapitre 1

LES SPLINES

Dans ce chapitre, les splines de régression seront définies et expliquées. Cette partie de ce mémoire est grandement basée sur les travaux de Bennaghmouch (1992).

1.1. Les splines de régression

Supposons que l'on possède un ensemble fini de points (t_i, y_i), $i = 1, ..., n$ où n représente la taille de l'ensemble tel que y_i fait partie de l'intervalle $[a, b]$, où $|a| < \infty$ et $|b| < \infty$, qui est généré par

$$y_i = f(t_i) + \epsilon_i, i = 1, ..., n,$$

où les erreurs sont indépendantes et identiquements distribuées de moyenne 0 et de variance σ^2. Notre but ici est d'approximer la fonction f.

Afin de bien approximer f, il faut choisir une classe P de fonctions qui peut approximer f ainsi que déterminer l'élément de P qui estime le mieux f, c'est-à-dire avoir un bon processus de sélection. Pour ce faire, on va avoir besoin de restreindre la classe P.

Considérons alors la classe des polynômes de degré $k - 1$ définie par :

$$P_k = \{p(x) : p(x) = \sum_{i=1}^{k} c_i x^{i-1}, c_i \in \mathbb{R}\},$$

qui définit f sur tout son domaine. Pour avoir un peu plus de flexibilité pour approximer f, on va séparer le domaine de f, alors f sera approximé par des polynômes sur plusieurs sous-intervalles, des polynômes de degré assez faible. Pour définir cela, supposons une suite croissante $\xi = (\xi_1, ..., \xi_{l+1})$ telle que $a = \xi_1 < ... < \xi_{l+1} = b$, ce qui divise $[a, b]$ en l sous-intervalles $I_i = [\xi_i, \xi_{i+1})$, où $i = 1, ..., l$ et sur chacun on définit un polynôme d'ordre k :

$$P_{k,\xi} = \{f : f(x) = p_i(x)\mathbb{1}_{x \in I_i} \text{ où } p_i \in P_k, i = 1, ..., l\},$$

où $\mathbb{1}_{x \in I_i} = 1$ si x se trouve dans l'intervalle I_i et vaut 0 sinon. Cependant, en utilisant simplement un polynôme $P_{k,\xi}$, il aura de la discontinuité aux points ξ_i.

Afin d'éviter la discontinuité, on considère l'ensemble des fonctions splines défini par :

$$\psi_{k,\xi} = P_{k,\xi} \cap C^{k-2}[a,b],$$

où

$$C^{k-2}[a,b] = \{f : \text{la } j^e \text{dérivée de } f,\ j = 1, ..., k-2 \text{ existe et est continue } \forall t \in [a,b]\}.$$

On appelle $\psi_{k,\xi}$, l'ensemble des fonctions splines d'ordre k avec noeuds de multiplicité 1 aux points $\xi_1, ..., \xi_l$.

Supposons donc une fonction S comprise dans l'ensemble $\psi_{k,\xi}$ qui approxime f, alors Bennaghmouch (1992) stipule que S peut s'écrire d'une façon unique :

$$S(t) = \sum_{j=0}^{k-1} \alpha_j t^j + \sum_{j=1}^{l} \beta_j (t - \xi_j)_+^{k-1},$$

où :

$$u_+ = \begin{cases} u & \text{si } u > 0 \\ 0 & \text{sinon,} \end{cases}$$

et $\alpha = (\alpha_0, ..., \alpha_{k-1})^t$ ainsi que $\beta = (\beta_1, ..., \beta_l)^t$ peuvent être déterminés par la méthode des moindres carrés. L'ensemble $\psi_{k,\xi}$ est un espace linéaire de dimension $k + l$ pour lequel $\{1, t, ..., t^{k-1}, ..., (t - \xi_1)_+^{k-1}, ..., (t - \xi_l)_+^{k-1}\}$ forme une base.

1.2. Les splines polynomiaux

L'ensemble des splines polynomiaux $P_{k,\xi,v}$ est un sous-ensemble de $P_{k,\xi}$ possédant, aux points ξ_i, v_i dérivées continues pour $i = 2, ..., l$ où $v_i \geq 0\ \forall\ i$ et v est défini comme étant le vecteur $(v_2, ..., v_l)^t$. Si f est élément de $P_{k,\xi,v}$, alors tous les polynômes p_i doivent satisfaire :

$$D^{j-1} P_{i-1}(\xi_i) = D^{j-1} P_i(\xi_i),\ j = 1, ..., v_i,\ i = 2, ..., l,$$

où $D^j P_i(\xi_i)$ est la j^e dérivée évaluée en ξ_i et v_i représente le nombre de conditions de continuité en ξ_i. Intuitivement, plus la condition de continuité est élevée, plus la fonction sera lisse aux noeuds. Donc si f est élément de $P_{k,\xi,v}$, alors on peut approximer f par une spline de la forme :

$$S^*(t) = \sum_{j=0}^{k-1} \alpha_j t^j + \sum_{r=2}^{l} \sum_{j=v_r}^{k-1} \beta_{j,r} (t - \xi_r)_+^j.$$

Si $v_i = k-1\ \forall\ i$, alors $S(t)^* = S(t)$ et si $v_i = 0$, alors la spline peut être discontinue en ξ_i. Alors $P_{k,\xi,v}$ est un sous-espace de $P_{k,\xi}$, et ce, pour tout vecteur v. Comme le

définit Bennaghmouch (1992), si f fait partie de l'ensemble $P_{k,\xi,v}$, alors on peut l'écrire de façon unique.

Supposons $\phi_1, \phi_2, \ldots$, des fonctions qui constituent une base de $P_{k,\xi,v}$, alors toute fonction f élément de l'ensemble $P_{k,\xi,v}$ peut s'écrire de façon unique :

$$f = \sum_j a_j \phi_j \text{ où } a_j \in \mathbb{R}.$$

En définissant bien les fonctions ϕ_i, on peut réécrire f sous la forme :

$$f = \sum_{i=1}^{l} \sum_{l=v_i}^{k-1} \alpha_{il} \phi_{il},$$

où $v_1 = 0$ et où :

$$\phi_{ij}(x) = \begin{cases} (x - \xi_1)^j & \text{si } i = 1 \\ (x - \xi_i)_+^j & \text{si } i = 2, ..., l \end{cases}$$

et $j = 0, ..., k-1$. Alors $\{\phi_{ij},\ i = 1, ..., l; j = 0, ..., k-1\}$ est une base de $P_{k,\xi}$ et $\{\phi_{ij},\ i = 1, ..., l; j = v_i, ..., k-1\}$ est une base de $P_{k,\xi,v}$.

1.3. Les B-splines

Tel que défini dans Bennaghmouch (1992), la base des B-splines a été introduite pour contrer certains problèmes de calcul. Cette section est très mathématique et peut être plus difficile à comprendre. Plus intuitivement, séparer la fonction en sections permet de contrer les problèmes d'oscillation de celle-ci. Un simple exemple sera présenté à la fin de la section. Les quelques définitions suivantes seront utiles. Premièrement, la dimension r de $P_{k,\xi,v}$ est égale à :

$$r = k + \sum_{i=2}^{l} (k - v_i) = kl - \sum_{i=2}^{l} v_i.$$

Ensuite, on définit le vecteur de noeuds w tel que $w = (w_1, ..., w_{r+k})^t$ où les paramètres $w_i, i = 1, ..., r+k$ sont définis par :
i) $w_1 \leq w_2 \leq ... \leq w_k \leq \xi_1$ et $\xi_{l+1} \leq w_{r+1} \leq ... \leq w_{r+k}$,
ii) pour $i = 2, ..., l$, ξ_i apparait $(k - v_i)$ fois dans $(w_1, ..., w_{r+k})^t$.
Les w_i permettrent de forcer le lissage. De plus, il faut introduire la notion de différence divisée. La k^e différence divisée d'une fonction f aux points $w_i, ..., w_{i+k}$ est le coefficient de w^k du polynôme $p(w) \in P_{k+1}$ qui coïncide avec f aux points $w_i, ..., w_{i+k}$. On la note par $[w_i, ..., w_{i+k}]f$:

$$[w_i, ..., w_{i+k}]f = \frac{[w_{i+1}, ..., w_{i+k}]f - [w_i, ..., w_{i+k-1}]f}{w_{i+k} - w_i},$$

$$[w_i]f = f(w_i).$$

Donc, la première différence divisée sera :

$$[w_i, w_{i+1}]f = \frac{f(w_{i+1}) - f(w_i)}{w_{i+1} - w_i}.$$

Alors, la i^e B-spline d'ordre k, avec noeuds w, notée $B_{i,k,w}$ est :

$$B_{i,k,w}(x) = (w_{i+k} - w_i)[w_i, ..., w_{i+k}](w_i - x)_+^{k+1},$$

que l'on peut réécrire :

$$B_{i,k,w}(x) = \frac{(x - w_i)}{(w_{i+k-1} - w_i)} B_{i,k-1,w}(x) + \frac{(w_{i+k} - x)}{(w_{i+k} - w_{i+1})} B_{i+1,k-1,w}(x),$$

pour $x \in [w_i, w_{i+k}]$, $i = 1, ..., r$ et $k > 1$.
Si $k = 1$, alors :

$$B_{i,1,w} = \begin{cases} 1 & \text{si } w_i \leq x \leq w_{i+1} \\ 0 & \text{sinon.} \end{cases}$$

Les B-splines possèdent plusieurs propriétés : $B_{i,k,w} = 0$ si $x \notin [w_i, w_{i+k}]$ et $B_{i,k,w} > 0$ si $x \in [w_i, w_{i+k}]$. Les B-splines $B_{1,k,w}, ..., B_{r,k,w}$ d'ordre k et de noeuds w forment une base pour $P_{k,\xi,v}$ de dimension r sur $[w_k, ..., w_{r+1}]$. Alors chaque f du sous-espace $P_{k,\xi,v}$ peut s'écrire de façon unique et exacte sous la forme :

$$f(x) = \sum_{i=1}^{r} \alpha_i B_{i,k,w}(x),$$

où $x \in [w_k, w_{r+1}]$ et $\alpha = (\alpha_1, ..., \alpha_r)^t$ est un vecteur de nombres réels. Il faut noter que les fonctions $S(t)$ définissaient une spline en général, que les fonctions $B(x)$ définissent une B-spline qui utilise les variables explicatives et donc que la fonction $f(x)$ fait partie de sous-espace $P_{k,\xi,v}$. La somme du nombre de contraintes de continuité et du nombre de noeuds en ξ_i est égale à k. Un choix convenable est $w_1 = ... = w_k = \xi_1$ et $w_{r+1} = ... = w_{r+k} = \xi_{l+1}$ avec $v_1 = v_{l+1} = 0$. Enfin, le choix des noeuds est important ; il est recommandé qu'ils correspondent aux points d'observation, qu'il y ait au minimum quatre à cinq observations entre chaque noeud, qu'il n'y ait pas plus d'un extremum ou point d'inflexion entre deux noeuds, l'extremum doit être centré dans l'intervalle et le point d'inflexion doit être près du point noeud. Par exemple, pour des splines polynomiaux, tel que présenté dans la figure 1.1, pour approximer la fonction $f(x) = x^2 + \epsilon$, où ϵ est l'erreur de type normal, il serait important qu'il y a un noeud près de l'observation en $x = 0$. S'il n'y a pas d'observation en $x = 0$, il faudrait que cet extremum soit en milieu d'un intervalle.

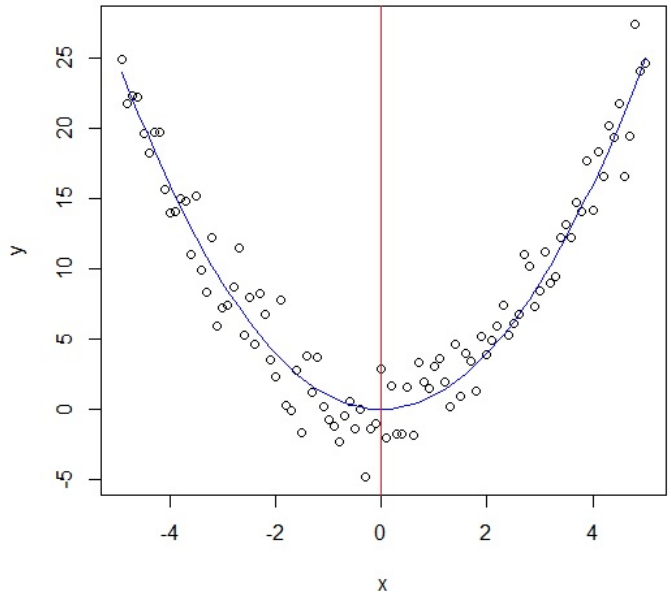

FIGURE 1.1. Exemple de splines polynomiaux simples

Bien que plus complexe, la figure 1.2 illustre un exemple de B-splines simples. Les B-splines ont pour noeuds extérieurs $\{0, 10\}$ et pour noeuds intérieurs $\{1, 5, 8\}$. On remarque donc que la fonction $f(x)$ (en ligne plus épaisse dans la figure) s'écrit de façon unique et exacte comme une combinaison linéaire des B-splines (en pointillés rouges). Leur coefficient respectif α leur est associé dans la figure.

1.4. AUTRE FORMULATION

Il existe de nombreuses formulations pour écrire f en fonction d'une spline. Une formulation qui nous sera utile dans le chapitre 3 est la suivante :

$$y = f(t) + \epsilon = \sum_{i=1}^{k} \alpha_i p_i(t) + \epsilon,$$

où $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)^t$ est le vecteur des coefficients du spline, le vecteur des polynômes qui forme le spline est noté $p(t) = (p_1(t),\ p_2(t),\ ...,\ p_k(t))^t$ et $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^t$ est le vecteur des erreurs.

Bref, les splines permettent d'approximer une fonction en séparant son domaine en plusieurs sous-intervalles et en faisant un test de régression sur chacun d'entre-eux tout en respectant des critères de continuité aux noeuds. C'est donc un bon moyen d'approximer une fonction qui possède plusieurs extremums et points d'inflexion par des polynômes de degré peu élevé.

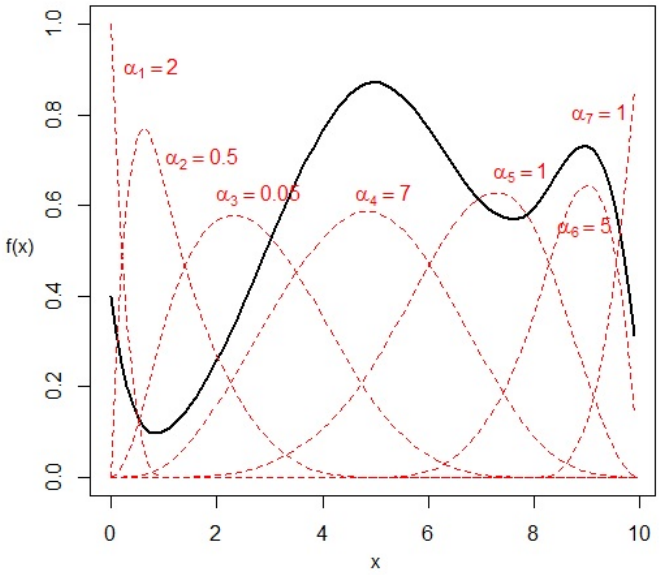

FIGURE 1.2. Exemple de B-splines simples

Chapitre 2

TESTS DE RUPTURE

Dans ce chapitre, quelques tests de rupture déjà présents dans la littérature seront décrits. Ces tests servent, comme leur nom l'indique, à détecter si un jeu de données possède ou non une rupture. Une rupture est un changement dans la moyenne de données qui, sous l'hypothèse nulle, sont identiquement distribuées. Le test de rupture non paramétrique de Lombard (1987) ainsi que les tests paramétriques de Jarušková (1997) et de Reeves *et al.* (2007) seront présentés.

2.1. Tests de rupture non paramétriques

2.1.1. Test de Lombard (1987)

Supposons que l'on a des variables aléatoires $y_1, ..., y_T$ indépendantes qui suivent chacune une distribution continue $F(y, \theta_1), ..., F(y, \theta_T)$. On essait d'étudier s'il y a un changement tel que les paramètres θ_i soient de la forme $\theta_1 = ... = \theta_\tau \neq \theta_{\tau+1} = ... = \theta_T$. Mais on a besoin d'un modèle plus général si les paramètres θ changent graduellement dans un intervalle. Alors, Lombard (1987) a introduit ce modèle :

$$\theta_i = \begin{cases} \rho & \text{si } i = 1, ..., \tau_1; \\ \rho + \frac{i-\tau_1}{\tau_2-\tau_1}\delta & \text{si } i = \tau_1 + 1, ..., \tau_2; \\ \rho + \delta & \text{si } i = \tau_2 + 1, ..., T. \end{cases}$$

pour $\tau_1 < \tau_2$, ρ et δ inconnus . Lombard (1987) utilise l'hypothèse suivante : $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$. Pour commencer, voici un peu de notation :
i) r_i est le rang de la variable aléatoire y_i,
ii) ϕ représente une fonction de score avec :

$$0 < \int_0^1 \phi^2(w)dw < \infty,$$

$$\bar{\phi} = \frac{1}{T}\sum_{i=1}^{T}\phi\left(\frac{i}{T+1}\right),$$
$$A^2 = \frac{1}{T-1}\sum_{i=1}^{T}\left[\phi\left(\frac{i}{T+1}\right) - \bar{\phi}\right]^2.$$

Alors, le score de classement de y_i, noté $s(r_i)$ est :

$$s(r_i) = \frac{1}{A}\left[\phi\left(\frac{r_i}{T+1}\right) - \bar{\phi}\right] \text{ pour } i = 1, ..., T.$$

Donc Lombard (1987) a introduit le test statistique suivant :

$$v_{\tau_1,\tau_2} = \sum_{j=\tau_1+1}^{\tau_2}\sum_{i=1}^{j-1} s(r_i) = (\tau_2 - \tau_1)\sum_{i=1}^{\tau_1} s(r_i) + \sum_{i=\tau_1+1}^{\tau_2-1}(\tau_2 - i)s(r_i),$$

qui est le test de rang localement le plus puissant pour tester l'hypothèse nulle versus une hypothèse alternative unilatérale si τ_1 et τ_2 sont connus. La statistique dépend indirectement des y_i en utilisant le rang de $y_i, i = 1, ..., T$. S'ils sont inconnus, Lombard propose le test suivant :

$$q_{1,T} = \sum_{t_1=1}^{T-1}\sum_{t_2=t_1+1}^{T} v_{t_1,t_2}^2.$$

Si on observe une grande valeur, on rejette l'hypothèse nulle. Lorsque T tend vers l'infini, il est à noter que :

$$\frac{q_{1,T}}{T^5} \to q_1 = \sum_{n=1}^{\infty}\frac{1}{(n\pi)^4}Z_n^2 \text{ sous } H_0,$$

où $Z = (Z_1, Z_2, ...)^t$ est un vecteur de variables aléatoires indépendantes et identiquement distribuées de loi normale standard. Les points critiques pour différentes valeurs de α et pour toute valeur de T sont présentés dans le tableau 2.1.

TABLEAU 2.1. Valeurs critiques de q_1 et $q_{1,T}/T^5$

α	0,1	0,075	0,05	0,025	0,01
q_1	0,0287	0,0334	0,0403	0,0525	0,0690
$q_{1,T}/T^5$	0,0289	0,0334	0,0402	0,0515	0,0662

Si on est intéressé à tester l'hypothèse nulle versus une hypothèse alternative unilatérale ($\delta > 0$ ou $\delta < 0$), Lombard propose trois différents scores :
i) Le score de Wilcoxon : $\phi_1(u) = 2u - 1$, pour tester des changements dans la position (une tendance centrale),

ii) Le score de Mood : $\phi_2(u) = (2u-1)^2$, pour tester les changements d'échelle,
iii) Le score logarithmique : $\phi_3(u) = \log(1-u)$ (les auteurs utilisent $-\log(1-u)$), pour tester des changements d'échelle dans une distribution ayant pour domaine $[0, \infty)$.
Pour le score de Wilcoxon, la fonction de score $s(r_i)$ est équivalente à :

$$s(r_i) = \sqrt{\frac{12}{T(T+1)}} \left(r_i - \left[\frac{T+1}{2} \right] \right).$$

Lorsqu'on veut tester s'il y a l'apparition d'une tendance, c'est-à-dire que $\tau_2 = T$ et que $\tau_1 = \tau$, la statistique du test de Lombard est modifiée tel que $t_2 = T$:

$$q^*_{1,T} = \sum_{t=1}^{T-1} v^2_{t,T},$$

et, lorsque T tend vers l'infini, sous l'hypothèse nulle :

$$\frac{q^*_{1,T}}{T^4} \to q^*_1 = \sum_{n=1}^{\infty} \gamma_n Z^2_n,$$

où $\gamma_1 > \gamma_2 > ... > 0$ sont les solutions positives réelles de $\tan\left(\gamma^{-\frac{1}{4}}\right) + \tanh\left(\gamma^{-\frac{1}{4}}\right) = 0$. Les points critiques pour différentes valeurs de α et pour toute valeur de T sont présentés dans le tableau 2.2.

TABLEAU 2.2. Valeurs critiques de q^*_1 et $q^*_{1,T}/T^4$

α	0,1	0,075	0,05	0,025	0,01
q^*_1	0,0879	0,1027	0,1242	0,162	0,2135
$q^*_{1,T}/T^4$	0,0882	0,1026	0,1241	0,158	0,2035

Lorsqu'on veut tester s'il y a un changement abrupte, c'est-à-dire que $\tau_1 = \tau$ et $\tau_2 = \tau + 1$, toujours en utilisant la fonction de score de Wilcoxon, la statistique de test de Lombard devient :

$$q^0_{1,T} = \sum_{t=1}^{T-1} v^2_{t,t+1} \ = \sum_{t=1}^{T-1} \left[\sum_{i=1}^{t} s(r_i) \right]^2,$$

et, sous l'hypothèse nulle, la distribution limite de $\frac{q^0_{1,T}}{T^2}$ converge vers la distribution limite du test d'adéquation de Cramer von Mises (q^0_1). Une fois de plus, pour toute valeur de T, les points critiques sont présentés dans la tableau 2.3.

TABLEAU 2.3. Valeurs critiques de $q^0_{1,T}$ et $q^0_{1,T}/T^2$

α	0,1	0,075	0,05	0,025	0,01
$q^0_{1,T}$	0,3473	0,3939	0,4614	0,5608	0,7435
$q^0_{1,T}/T^2$	0,3431	0,3870	0,4521	0,5596	0,7022

2.2. TESTS DE RUPTURE PARAMÉTRIQUES

2.2.1. Test de Jarušková (1997)

Le test de Jarušková (1997) est un test paramétrique de type «maximum». Tout comme le test de Lombard (1987), supposons que l'on observe des variables aléatoires $y_1, ..., y_T$. L'hypothèse nulle stipule que $H_0 : \mu_1 = \mu_2$. L'hypothèse alternative stipule qu'il existe un point dans le temps $\tau \in \{1, ..., T-1\}$ qui fait en sorte que le modèle est :

$$y_i \sim \begin{cases} N(\mu_1, \sigma^2) & \text{si } i = 1, ..., \tau; \\ N(\mu_2, \sigma^2) & \text{si } i = \tau, ..., T, \end{cases}$$

où $\mu_1 \neq \mu_2$. En supposant σ^2 inconnu, le test de Jarušková , noté $J(T)$ est le suivant :

$$J(T) = \max_{1 \leq \tau < T} |J_\tau| = \max_{1 \leq \tau < T} \frac{1}{s_\tau} \sqrt{\frac{(T-\tau)\tau}{T}} |\overline{y}_\tau - \overline{y}^*_\tau|,$$

où :

$$\begin{aligned} \overline{y}_\tau &= \frac{\sum_{j=1}^{\tau} y_j}{\tau}, \\ \overline{y}^*_\tau &= \frac{\sum_{j=\tau+1}^{T} y_j}{T-\tau}, \\ s_\tau &= \sqrt{\frac{\sum_{j=1}^{\tau}(y_j - \overline{y}_\tau)^2 + \sum_{j=\tau+1}^{T}(y_j - \overline{y}^*_\tau)^2}{T-2}}. \end{aligned}$$

L'hypothèse nulle est rejetée lorsque la statistique $J(T)$ est plus grande qu'une certaine valeur critique (voir le tableau 2.4). Si l'hypothèse nulle est vraie, alors $\overline{y}_\tau$ et $\overline{y}^*_\tau$ devraient être d'environ la même valeur. Jarušková (1997) obtient celles-ci par simulations. Une statistique tronquée peut aussi être utilisée :

$$J_1(T) = \max_{t_0 T \leq \tau \leq (1-t_0)T} |J_\tau|,$$

où $t_0 \in (0\,;0{,}5)$ (Jarušková utilise $t_0 = 0{,}05$). Jarušková (1997) obtient aussi les valeurs critiques par simulations et sont aussi présentées dans le tableau 2.4.

TABLEAU 2.4. Valeurs critiques de $J(T)$ et $J_1(T)$

	$\alpha = 0{,}05$		$\alpha = 0{,}01$	
T	$J(T)$	$J_1(T)$	$J(T)$	$J_1(T)$
50	3,15	3,08	3,76	3,69
100	3,16	3,06	3,71	3,62
200	3,19	3,07	3,72	3,61
300	3,21	3,08	3,73	3,62
500	3,24	3,09	3,73	3,62

2.2.2. Tests de Reeves *et al.* (2007)

Les tests de Reeves *et al.* (2007) utilisent les sommes au carré des résidus et comparent leur statistique de test à une loi de Fisher dont les degrés de liberté changent dépendamment du test utilisé. Tout comme les autres tests, supposons que l'on a des variables aléatoires $y_1, ..., y_T$ et que celles-ci, sous l'hypothèse nulle, soient indépendantes et identiquement distribuées tandis que, sous l'hypothèse alternative, il existe un point τ tel que la moyenne des variable $y_1, ..., y_\tau$ soit différente des variables $y_{\tau+1}, ..., y_T$. Reeves *et al.* (2007) proposent les tests (méthodes) LR (Modified Lund and Reeves TPR method) et XLW (Modified Wang's TPR method) où la méthode TPR (two-phase regression) avec un point de rupture au point τ est simplement :

$$y_i = \begin{cases} \mu_1 + \beta_1 x_i + \epsilon_i & \text{si } i = 1, ..., \tau \\ \mu_2 + \beta_2 x_i + \epsilon_i & \text{si } i = \tau + 1, ..., T, \end{cases}$$

ainsi que $x_1 \leq ... \leq x_T$, $\epsilon_i \sim N(0, \sigma^2)$ pour $i = 1, ..., T$ et β_1, β_2, μ_1, μ_2, τ sont inconnus. Il y a cependant une contrainte de continuité de la régression au point τ se traduisant par $\mu_2 = \mu_1 + (\beta_1 - \beta_2)x_\tau$. Les deux tests suivants proposent quelques modifications au test TPR, tel que l'égalité de β_1 et β_2.

2.2.2.1. *Test LR*

Lund et Reeves (2007) ont modifié le modèle TPR en cessant d'imposer la contrainte de continuité. Le modèle LR suppose que les x_i sont des entiers. Le modèle LR est donc :

$$y_i = \begin{cases} \mu_1 + \beta_1 t + \epsilon_i & \text{si } i = 1, ..., \tau \\ \mu_2 + \beta_2 t + \epsilon_i & \text{si } i = \tau + 1, ..., T, \end{cases}$$

où on peut alors détecter une rupture dans la moyenne ($\mu_1 \neq \mu_2$) et dans la tendance ($\beta_1 \neq \beta_2$). Les erreurs suivent une loi normale de moyenne nulle et de variance inconnue. Ce qui revient donc à confronter les hypothèses suivantes :

$$H_0 : \mu_1 = \mu_2 \text{ et } \beta_1 = \beta_2,$$
$$H_1 : \mu_1 \neq \mu_2 \text{ et/ou } \beta_1 \neq \beta_2.$$

Si le point de rupture τ est connu et fixé, le test devient simplement :

$$F_\tau = \frac{(SSE_0 - SSE_A)/2}{SSE_A/(T-4)} \sim F_{2,T-4},$$

où SSE_0 et SSE_A sont respectivement les sommes des erreurs au carré sous l'hypothèse nulle et alternative. Sous H_0, la statistique de test devrait suivre une loi de Fisher à $(2, T-4)$ degrés de liberté. On rejette l'hypothèse nulle pour de grandes valeurs de F_τ.
Si le point de rupture τ est inconnu, la statistique devient :

$$F_{max} = \max_{1 \leq \tau \leq T} F_\tau,$$

et on rejette H_0 pour une grande valeur de F_{max}. Comme cette statistique ne suit pas de loi connue, Lund et Reeves (2007) obtiennent les valeurs critiques par simulations et sont présentées dans le tableau 2.5.

2.2.2.2. *Test XLW*

Wang (2007) a apporté une petite modification au test LR car il trouvait que celui-ci expliquait mal les phénomènes climatiques. Son modèle devient donc :

$$y_i = \begin{cases} \mu_1 + \beta t + \epsilon_i & \text{si } i = 1, ..., \tau \\ \mu_2 + \beta t + \epsilon_i & \text{si } i = \tau + 1, ..., T \end{cases}$$

où les termes sont définis précédemment. Les hypothèses deviennent donc :

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2.$$

Si le point de rupture τ est connu et fixé, le test devient simplement :

$$F_\tau = \frac{(SSE_0 - SSE_A)}{SSE_A/(T-3)} \sim F_{1,T-3},$$

où les termes sont aussi définis précédemment. Si τ est inconnu la statistique F_{max} est une fois de plus utilisée. Wang (2007) obtient aussi les valeurs critiques par simulations et sont présentées dans le tableau 2.5.

TABLEAU 2.5. Valeurs critiques avec $\alpha = 0{,}05$ des tests LR et XLW

T	LR (F_{max})	XLW (F_{max})
25	11,67	7,37
50	11,07	6,92
75	11,06	6,88
100	11,09	6,91
200	11,21	7,01
500	11,54	7,24

Cependant, dans Dubé (2011), il est écrit sur le test TPR et ses dérivées que la convergence vers les valeurs critiques obtenues n'est pas fiable et que ces tests ne sont pas robustes par rapport au non-respect du postulat de normalité. Les splines étant plus flexibles qu'une régression linéaire, on s'attend à ce que l'introduction des splines au modèle LR soit une solution envisageable afin de réduire les erreurs. Ceci ne règlera peut-être pas le problème de non-respect du postulat de normalité, mais cela pourra peut-être améliorer les performances du test TPR.

Bref, il existe, dans la littérature, plusieurs tests de détection de rupture. Dans ce mémoire, nous en avons proposé quatre, dont deux modifications du même test initial. Dans le chapitre suivant, nous allons proposer de nouvelles méthodes de détection de rupture n'existant pas dans la littérature et, par la suite, nous comparerons toutes ces méthodes entre-elles à l'aide de simulations.

Chapitre 3

NOUVEAUX TESTS DE RUPTURE

Les tests de Lombard (1987) et de Jarušková (1997) sont très efficaces pour trouver un point de rupture et un déclenchement de tendance, mais tel que stipulé dans le dernier chapitre, le test de Reeves *et al.* (2007) l'est un peu moins. Dans ce chapitre, on va décrire trois nouvelles méthodes de détection des points de rupture, l'une en utilisant les splines, la seconde en utilisant la statistique bayésienne et la troisième combinera les splines et la statistique bayésienne.

3.1. Tests de rupture avec B-splines

Tout en se basant sur les modèles TPR et LR décrits dans la section 2.2.2 et la sous-section 2.2.2.1, on va introduire les splines à ceux-ci, plus précisément les B-splines (section 1.3). Les B-splines sont une alternative à la régression linéaire et, quoi que plus complexes, approximent mieux les données (ici on utilise des B-splines cubiques). Afin d'investiguer l'efficacité de ce nouveau test et de savoir si nous poussons plus loin l'analyse de cette méthode, nous effectuerons des simulations à l'aide du modèle ci-dessous où les ϵ_i sont des erreurs provenant d'une population normale avec une rupture au point τ :

$$y_i = \begin{cases} \mu_1 + \epsilon_i & \text{si } i = 1, ..., \tau \\ \mu_2 + \epsilon_i & \text{si } i = \tau + 1, ..., T, \end{cases}$$

avec les hypothèses :

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2.$$

Puisque le choix des noeuds est important, on fait en sorte qu'il y ait 10 observations entre chaque noeuds ; si la taille de l'échantillon n'est pas un multiple de 10,

alors on inclut moins d'observations entre les deux derniers noeuds. Sous l'hypothèse nulle, la B-spline a pour noeuds extérieurs $\{1, T\}$ et les noeuds intérieurs sont $\{10, 20, ..., T-10\}$.
Sous l'hypothèse alternative, on sépare le modèle en deux B-splines. La première comprend les données $y_1, ..., y_\tau$ et la seconde $y_{\tau+1}, ..., y_T$, où $\{1, \tau\}$ et $\{\tau+1, T\}$ sont respectivement les noeuds externes pour la première et la seconde B-spline. Les noeuds internes restent les mêmes. Si un noeud interne a la même valeur qu'un noeud externe, on le considère comme ce dernier.

3.1.1. Première tentative de test

Si l'on suit le raisonnement de Reeves *et al.*, en supposant le point de rupture τ connu, alors la statistique de test devient :

$$F_\tau = \frac{(SSE_0 - SSE_A)/(df_0 - df_A)}{SSE_A/(df_A)} \sim F_{df_0 - df_A, df_A},$$

où df_0 et df_A sont respectivement les degrés de liberté de SSE_0 et SSE_A. Mais, dans notre cas, on suppose que τ est inconnu et on utilise la statistique :

$$F_{max} = \max_{t_0 \leq \tau \leq T - t_0} F_\tau.$$

Comme il faut qu'il y ait au moins un noeud interne pour utiliser une spline (sinon ce ne serait qu'une régression cubique), on choisit dans notre cas $t_0 = 20$ on fait varier la rupture de 20 à $(T-20)$ et non de 1 à T comme dans le modèle de Reeves *et al.* (2007) afin d'avoir une stabilité dans notre modèle. Nous voulons qu'il y ait au moins une spline de chaque côté de la rupture et c'est pourquoi on choisit $t_0 = 20$. Si la rupture se trouve vraiment au début ou à la fin des données, le test ne la détectera pas. Les valeurs critiques sont encore une fois trouvées par simulation. Nous avons effectué 1000 simulations du modèle présenté à la page précédente avec $T = 100$ et nous avons répété celles-ci pour chaque valeur de i dans $\mu_1 = 0, \mu_2 = i, i = 0, ..., 10$.
Pour les valeurs critiques trouvées au niveau $\alpha = 0{,}05$, la puissance de ce test est beaucoup moins élevée que le test initial de Reeves *et al.* (2007) et encore moins que le test de Lombard (1987). Par exemple, lors d'une simulation de 1000 jeux de données avec $\mu_1 = \mu_2 - 10$, le test avec B-spline détecte une différence dans la moyenne dans seulement 30% des cas, tandis que le test de Lombard détecte une différence dans 100% des cas lorsque $\mu_1 = \mu_2 - 5$. Le problème est que la B-spline approxime trop bien les données. Sa meilleure qualité devient en quelque sorte son défaut pour ce test. On ne développera pas plus ce test. Il faut donc trouver une autre méthode. Il faut bien noter que nous ne supposons pas initialement que

l'utilisation des splines améliorerait à coup sûr le modèle de Reeves *et al.* (2007) mais bien que celles-ci pourraient peut-être, d'une certaine façon, rendre plus performant un modèle déjà existant. La section 3.1.1. a été ajoutée au mémoire afin que tout le raisonnement menant au modèle que nous allons garder plus loin soit exposé.

3.1.2. Seconde tentative de test

On pourrait déterminer une nouvelle méthode en observant les coefficients des B-splines aux noeuds. Pour commencer, si on observe les coefficients des B-splines obtenus sous l'hypothèse alternative, lorsque $\mu_1 \neq \mu_2$, on remarque que les coefficients des B-splines sont sensiblement les mêmes. Trouver une statistique par rapport à cela n'a pas fonctionné, car la valeur critique varie en fonction de la moyenne de l'échantillon (les valeurs critiques pour deux échantillons $(y_1, ..., y_T)$ et $(y_1 + a, ..., y_T + a)$, où a est une constante, n'étaient pas les mêmes). Puis, on a observé les coefficients de la B-spline sous l'hypothèse nulle, c'est-à-dire en supposant que $\mu_1 = \mu_2$. On remarque que la valeur des coefficients après la rupture est beaucoup plus élevée comparativement à celle des coefficients avant la rupture lorsque la différence réelle entre μ_1 et μ_2 est élevée que lors du contraire. On peut donc se demander jusqu'à quel point est-ce que les coefficients diffèrent et si c'est possible de créer une statistique à partir de ceux-ci. On propose donc la statistique suivante :

$$S_\tau = \left| \frac{\sum_{i=1}^{n_1} \alpha_{1i}}{n_1} - \frac{\sum_{i=1}^{n_2} \alpha_{2i}}{n_2} \right|,$$

où $\sum_{i=1}^{n_1} \alpha_{1i}$ et $\sum_{i=1}^{n_2} \alpha_{2i}$ sont respectivement les sommes des coefficients du B-spline avant et après la rupture et où n_1 ainsi que n_2 sont respectivement le nombre de noeuds avant et après la rupture. Comme on suppose que le point de rupture τ est inconnu, on utilise la statistique :

$$S_{\tau,max} = \max_{t_0 \leq \tau \leq T - t_0} S_\tau.$$

Ici on choisit encore $t_0 = 20$ pour les mêmes raisons expliquées précédemment. Si la méthode détecte une rupture, S_τ sera maximale entre les deux noeuds où il y a la rupture. On ne pourra donc pas explicitement dire l'endroit exact de la rupture, mais plutôt entre quels noeuds elle se trouve. Comme cette statistique est nouvelle et suit une loi inconnue, il faut établir ses valeurs critiques au niveau 5% pour différentes tailles d'échantillons, différentes variances et différents emplacements de la rupture. Ce que nous avons fait par simulations, en utilisant des données suivants une loi gaussienne de moyenne 0 et dont les valeurs critiques en

utilisant différentes tailles d'échantillons ainsi que de variances lors de celles-ci sont présentées dans le tableau 3.1.

TABLEAU 3.1. Valeurs critiques à 5% de S_{max} en fonction de T et de l'écart type σ

T	σ					
	1	2	3	4	5	6
50	0,98	1,93	2,94	3,89	4,87	5,85
100	0,83	1,69	2,55	3,36	4,18	5,02
200	0,77	1,57	2,40	3,09	3,84	4,61
400	0,74	1,51	2,29	2,99	3,71	4,45
1000	0,73	1,44	2,27	2,95	3,66	4,40

Les valeurs critiques varient avec la taille de l'échantillon ainsi qu'avec la variance de celui-ci. Lorsque la variance augmente, on remarque que la valeur critique augmente linéairement avec l'écart-type. Pour un échantillon de taille 100 suivant une loi normale d'écart-type $\sigma = 4$, la valeur critique est $0{,}84\sigma$. Comme dans notre cas, la variance est inconnue, on l'estime par :

$$\hat{\sigma^2} = \frac{1}{T-1}\sum_{i=1}^{T}(y_i - \overline{y})^2,$$

où :

$$\overline{y} = \frac{1}{T}\sum_{i=1}^{T} y_i,$$

et alors la valeur critique (toujours pour le même exemple) est $0{,}84\hat{\sigma}$. Nous aurions pu utiliser comme variance l'estimateur défini à la sous-section 2.2.1 (s_τ). Cependant, s_τ peut s'écrire sous la forme suivante :

$$s_\tau = \frac{1}{T-2}\left((T-1)\hat{\sigma^2} - \tau(\overline{y} - \overline{y}_\tau)^2 - (T-\tau)(\overline{y} - \overline{y}^*_\tau)^2\right).$$

Sous l'hypothèse nulle, $\overline{y}_\tau \approx \overline{y}^2_\tau \approx \overline{y}$ et donc, pour un T assez grand, $s_\tau \approx \hat{\sigma^2}$. C'est pourquoi nous avons utilisé $\hat{\sigma^2}$ comme estimateur de la variance pour cette méthode. Afin que les valeurs critiques ne dépendent plus de l'écart-type et afin de simplifier les tableaux, on considère la nouvelle statistique :

$$S^*_\tau = \frac{1}{\hat{\sigma}}\left|\frac{\sum_{i=1}^{n_1}\alpha_{1i}}{n_1} - \frac{\sum_{i=1}^{n_2}\alpha_{2i}}{n_2}\right|.$$

Tout comme S_τ, on suppose que le point de rupture est inconnu et on utilise :

$$S^*_{\tau,max} = \max_{t_0 \le \tau \le T - t_0} S^*_\tau,$$

où t_0 est une fois de plus égal à 20. Les valeurs critiques sont une fois de plus trouvées par simulation (elles correspondent exactement à la première colonne du tableau 3.1). De plus, lorsque la taille de l'échantillon tend vers l'infini, la valeur critique à 5% tend vers 0,71. On remarque que la puissance ne varie pas beaucoup dépendamment de la position de la rupture (voir le tableau 3.2).

TABLEAU 3.2. Puissance (en %) du test dépendamment de la position de la rupture ($\mu_1 = \mu_2 - 5$)

σ \ τ	25	40	50	60	75
4	93,8	94,8	95,5	96,6	95,5
5	80,9	78,9	79,8	78,7	77,4

3.2. TEST DE RUPTURE BAYÉSIEN

Dans cette section, la méthode bayésienne de détection de rupture et de tendance sera développé, tout en étant précédée d'une introduction à la densité *a priori-G*, laquelle sera utilisée dans la méthode. Cela permettra de trouver le modèle le plus adéquat pour l'échantillon analysé.

3.2.1. Densité *a priori* et densité *a priori-G*

Supposons que $y = (y_1, ..., y_T)^t$ est un vecteur de variables aléatoires indépendantes suivant une loi de paramètres $\theta = (\theta_1, ..., \theta_p)$. Geinitz (2009) et Robert (2001) écrivent qu'en statistique bayésienne, la fonction de vraisemblance est égale à :

$$l(\theta|y) = f(y|\theta) = \prod_{i=1}^{T} f(y_i|\theta),$$

et par le théorème de Bayes, la densité *a posteriori* $\pi(\theta|y)$ est égale à :

$$\pi(\theta|y) = \frac{l(\theta|y)\pi(\theta)}{m(y)},$$

où $\pi(\theta)$ est la densité *a priori* de θ et $m(y) = \int_\Theta l(\theta|y)\pi(\theta)d\theta$ est simplement une constante de normalisation afin que $\pi(\theta|y)$ soit en effet une densité (pour que $\int_\Theta \pi(\theta|y)d\theta = 1$). Donc,

$$\pi(\theta|y) \propto l(\theta|y)\pi(\theta).$$

Il faut déterminer quelle est la densité *a priori* $\pi(\theta)$. Elle peut-être informative ou non. Parmi les densités *a priori* non informatives les plus utilisée, notons :

$$\pi(\theta) \propto 1,$$

qui signifie que toutes les valeurs de θ sont équiprobables, et :

$$\pi(\theta) \propto I(\theta)^{1/2},$$

où $I(\theta)$ est l'information de Fisher, qui est la densité *a priori* de Jeffrey basée sur le principe de l'invariance (Geinitz, 2009).

Maintenant, supposons que y suit un modèle linéaire. Le but de la statistique bayésienne est d'inclure l'information *a priori* dans l'analyse. On utilise donc la densité *a priori* informative. En supposant que $y = X\beta + \epsilon$ et que la fonction de vraisemblance ainsi que la densité *a priori* suivent les lois suivantes :

$$\begin{aligned} y|\beta, \sigma^2 &\sim N_T(X\beta, \sigma^2 I), \\ \beta|\sigma^2 &\sim N_p(\beta_0, \frac{1}{n_0}\Omega), \end{aligned}$$

où β_0 est la moyenne du modèle utilisé et n_0 est un paramètre qui peut varier de 0 à l'infini qui sert à quantifier la précision de l'information que nous avons sur le modèle. Alors, la densité *a posteriori* sera :

$$\beta|y, \sigma^2 \sim N_p(A^{-1}b, A^{-1}),$$

où

$$\begin{aligned} A &= n_0\Omega^{-1} + \frac{1}{\sigma^2}X^tX, \\ b &= n_0\Omega^{-1}\beta_0 + \frac{1}{\sigma^2}X^ty. \end{aligned}$$

Pour spécifier Ω, on peut procéder comme suit. On sait que l'estimateur du maximum de vraisemblance de β est aymptotiquement normale de moyenne β et de matrice de covariance $\sigma^2(X^tX)^{-1}$. La densité *a priori* de type G consiste à choisir $\Omega = \sigma^2(X^tX)^{-1}$, on arrive aux résultats suivants :

$$A = \frac{1}{\sigma^2}(n_0 X^t X + X^t X),$$
$$b = \frac{1}{\sigma^2}(n_0 X^t X \beta_0 + X^t y).$$

Cette densité *a priori* est connue sous le nom de densité *a priori-G* informative de Zellner (1983), ou tout simplement la densité *a priori-G*.
La densité *a priori-G* nous donne une façon assez intuitive de déterminer jusqu'à quel point la densité *a priori* :

$$\beta|\sigma^2, X \sim N_p\left(\beta_0, \frac{\sigma^2}{n_0}(X^t X)^{-1}\right),$$

contribue à la densité *a posteriori* :

$$\beta|y, \sigma^2, X \sim N_p\left(\frac{n_0}{n_0+1}\left(\beta_0 + \frac{1}{n_0}\hat{\beta}\right), \frac{\sigma^2}{n_0+1}(X^t X)^{-1}\right),$$

en faisant varier n_0 et où $\hat{\beta}$ est l'estimateur du maximum de vraisemblance de β. Plus n_0 est petit, plus la contribution attribuée à la densité *a priori* est atténuée.

La densité marginale *a posteriori* suit donc une loi de Student multidimensionnelle :

$$\beta|y, X \sim t_p\left(T-p, \frac{n_0}{n_0+1}\left(\beta_0 + \frac{1}{n_0}\hat{\beta}\right), \frac{s^2}{T(n_0+1)} + \frac{n_0}{T(n_0+1)^2}(\beta_0 - \hat{\beta})^t X^*\right),$$

où $X^* = X^t X(\beta_0 - \hat{\beta}))(X^t X)^{-1}$ et $s^2 = \left\|y - X\hat{\beta}\right\|^2$(voir Robert, 2001). Alors, si l'on souhaite comparer deux modèles, soit M_0 et M_1, où :

$$\theta = \begin{cases} \theta_0 & \text{dans } M_0, \\ \theta_1 & \text{dans } M_1, \end{cases}$$

on utilise le facteur de Bayes, noté $B_{10}(y)$ (Geinitz, 2009), défini par :

$$B_{10}(y) = \frac{\pi(\theta_1|y)}{\pi(\theta_0|y)} \Big/ \frac{\pi(\theta_0)}{\pi(\theta_1)}.$$

Donc, pour comparer les deux modèles, M_0 et M_1, ayant des hypothèses composés :

$$H_0 : \theta_0 \in \Theta_0,$$
$$H_1 : \theta_1 \in \Theta_1,$$

la densité *a priori* de θ devient :

$$\pi(\theta) = \pi_0 g_0(\theta) + \pi_1 g_1(\theta)$$

où $g_i(\theta)$ est la densité sur Θ_i et où $\pi_0 + \pi_1 = 1$. Donc :

$$g_i(\theta) > 0 \text{ si } \theta \in \Theta_i,$$
$$g_i(\theta) = 0 \text{ si } \theta \notin \Theta_i.$$

Alors :

$$\begin{aligned} m(y) &= \int_\Theta \pi(\theta) f(y|\theta) d\theta, \\ &= \pi_0 m_0(y) + \pi_1 m_1(y). \end{aligned}$$

Avec ce qui est défini précédemment, on trouve que :

$$\pi(\theta|y) = \frac{[\pi_0 m_0(y)]\, \pi_0(\theta|y) + [\pi_1 m_1(y)]\, \pi_1(\theta|y)}{\pi_0 m_0(y) + \pi_1 m_1(y)}.$$

Donc :

$$\begin{aligned} \mathbb{P}(\theta \in \Theta_0|y) &= \int_{\Theta_0} \pi(\theta|y) d\theta, \\ &= \frac{\pi_0 m_0(y)}{\pi_0 m_0(y) + \pi_1 m_1(y)}. \end{aligned}$$

De la même façon, on trouve que :

$$\mathbb{P}(\theta \in \Theta_1|y) = \frac{\pi_1 m_1(y)}{\pi_0 m_0(y) + \pi_1 m_1(y)}$$

Alors :

$$\frac{\mathbb{P}(\theta \in \Theta_1|y)}{\mathbb{P}(\theta \in \Theta_0|y)} = \frac{\pi_1 m_1(y)}{\pi_0 m_0(y)},$$

et le facteur de Bayes peut être réécrit de cette façon :

$$B_{10}(y) = \frac{\pi_1 m_1(y)}{\pi_0 m_0(y)} / \frac{\pi_1}{\pi_0} = \frac{m_1(y)}{m_0(y)} = \frac{\int_{\Theta_1} l(\theta_1|y) g_1(\theta_1) d\theta_1}{\int_{\Theta_0} l(\theta_0|y) g_0(\theta_0) d\theta_0}.$$

Appliqué à la densité *a priori-G* définie précédemment, la densité marginale des observations devient :

$$\begin{aligned} m(y) &= \int_{\mathbb{R}^+} \int_{\mathbb{R}^p} \pi(\beta, \sigma^2) l(\beta, \sigma^2|y) d\beta d\sigma^2 \\ &= \int_{\mathbb{R}^+} \int_{\mathbb{R}^p} \pi_1(\beta|\sigma^2) \pi_2(\sigma^2) l_1(\beta|\sigma^2, y) l_2(\sigma^2|y) d\beta d\sigma^2, \end{aligned}$$

où :

$$\begin{aligned} \beta|\sigma^2 &\sim N_p \left(\beta_0, \frac{\sigma^2}{n_0} (X^t X)^{-1} \right), \\ \sigma^2 &\sim IG \left(\frac{\nu}{2}, \frac{s_*^2}{2} \right), \end{aligned}$$

$$\begin{aligned} l_1(\beta|\sigma^2, y) &\propto \frac{-1}{(\sigma^2)^{\frac{p}{2}}} \exp \left(\frac{-1}{2\sigma^2} (\beta - \hat{\beta})^t X^t X (\beta - \hat{\beta}) \right), \\ l_2(\sigma^2|y) &\propto \frac{1}{(\sigma^2)^{\frac{n-p}{2}}} \exp \left(\frac{-1}{2\sigma^2} (y - X\hat{\beta})^t (y - X\hat{\beta}) \right), \end{aligned}$$

et où :

$$\begin{aligned} s_*^2 &= \frac{1}{T-1} \sum_{i=1}^{T} (y_i - \bar{y})^2, \\ \bar{y} &= \frac{1}{t_0} \sum_{i=1}^{T} (y_i). \end{aligned}$$

Donc :

$$\pi(\beta|\sigma^2, y) \propto \exp \left(\frac{-(n_0+1)}{2\sigma^2} \left(\beta - \left[\frac{\hat{\beta} + n_0 \beta_0}{n_0 + 1} \right] \right)^t X^t X \left(\beta - \left[\frac{\hat{\beta} + n_0 \beta_0}{n_0 + 1} \right] \right) \right),$$

$$\pi(\sigma^2|y) \propto \exp \left(\frac{-1}{2\sigma^2} \left(s_*^2 + (y - X\hat{\beta})^t (y - X\hat{\beta}) + \frac{n_0}{n_0 + 1} (\hat{\beta} - \beta_0)^t X^t X (\hat{\beta} - \beta_0) \right) \right).$$

En résolvant l'intégrale précédente, on arrive au résultat suivant :

$$m(y) = \left(\frac{n_0}{n_0+1}\right)^{\frac{p}{2}} \frac{(s_*/2)^{\frac{\nu}{2}}\Gamma(\frac{\nu+T}{2})}{(2\pi)^{\frac{T}{2}}\Gamma(\frac{\nu}{2})(S^*/2)^{\frac{\nu+T}{2}}},$$

où :

$$S^* = s_*^2 + (y - X\hat{\beta})^t(y - X\hat{\beta}) + \frac{n_0}{n_0+1}(\hat{\beta} - \beta_0)^t X^t X(\hat{\beta} - \beta_0).$$

Donc le facteur de Bayes devient, dans ce cas :

$$B_{10} = \frac{m_1(y)}{m_0(y)} = \left(\frac{n_0}{n_0+1}\right)^{\frac{p_1-p_0}{2}} \left(\frac{S_0^*}{S_1^*}\right)^{\frac{T+\nu}{2}},$$

où :

$$S_i^* = s_*^2 + (y - X_i\hat{\beta})^t(y - X_i\hat{\beta}) + \frac{n_0}{n_0+1}(\hat{\beta} - \beta_i)^t X_i^t X_i(\hat{\beta} - \beta_i),$$

p_i est le nombre de paramètres dans le modèle i et β_i est le vecteur des coefficients *a priori* du modèle i, ici $i = 0, 1$. S^* peut être réécrit (c'est de cette façon qu'on l'a utilisé dans le test présenté plus loin) :

$$S^* = \frac{1}{n_0+1}y^t y - \frac{1}{n_0+1}y^t X(X^t X)^{-1}X^t y + \frac{n_0}{n_0+1}(y - X\beta_i)^t(y - X\beta_i),$$

où β_i correspond à l'espérance *a priori* de β sous le modèle M_i. Il est inconnu et doit être approximé. Notez que s_*^2 a été développé puis combiné à l'intérieur de l'équation.

3.2.2. Test de rupture utilisant la statistique bayésienne

On a donc utilisé les densités *a priori-G* pour trouver des ruptures et/ou des tendances. Cependant, contrairement à ce qui est montré ci-dessus, où il y a simplement deux hypothèses, notre méthode en compare cinq :

H_0 : Aucune rupture ni tendance,

H_1 : Rupture seulement,

H_2 : Tendance seulement,

H_3 : Rupture et tendance,

H_4 : Déclenchement de tendance.

Afin de bien définir ces hypothèses de façon mathématique, voici quelques notations qui seront utilisées :
i) $t = (1, 2, ..., T)^t$, $t_1 = (1, 2, ..., \tau)^t$, $t_2 = (\tau + 1, ..., T)^t$;
ii) $\widehat{\alpha}_0 = \frac{1}{t_0} \sum_{i=1}^{t_0} y_i$, la moyenne des t_0 premières données ;
iii) $\widehat{\alpha}_1 = \frac{1}{t_0} \sum_{i=T-t_0+1}^{T} y_i$, la moyenne des t_0 dernières données ;
iv) $\widehat{\alpha}_2 = \widehat{\alpha}_1 - \widehat{\alpha}_0$;
v) $\widehat{\alpha}_\tau = \frac{1}{t_\tau} \sum_{i=\tau-(t_\tau/2)}^{\tau+(t_\tau/2)} y_i$, la moyenne des t_τ données autour de la rupture ;
vi) $\widehat{\alpha}_3 = (\widehat{\alpha}_\tau - \widehat{\alpha}_0)/(\tau - \left[\frac{t_0+1}{2}\right])$, la pente de la droite reliant α_0 à α_τ ;
vii) $\widehat{\alpha}_4 = (\widehat{\alpha}_2 - \widehat{\alpha}_\tau)/(T - (\left[\frac{t_0-1}{2}\right] + \tau))$, la pente de la droite reliant α_τ à α_2 ;
viii) $\mathbb{1}_{t<\tau}$ et $\mathbb{1}_{t\geq\tau}$ sont les indicatrices indiquant si la donnée t se trouve avant ou après la rupture.
Il y a beaucoup de paramètres à estimer au départ car il y a plus d'une hypothèse à tester. Notez que cette approche permet de limiter le nombre de ceux-ci. Lors de nos simulations, on a utilisé $t_0 = 10$ et $t_\tau = 7$. La matrice X de taille $T \times p$, le vecteur $X\beta_0$ et le nombre de paramètres p varient d'une hypothèse à l'autre. On peut donc réécrire les modèles des hypothèses ci-dessus avec la notation qu'on vient de définir. Pour commencer, on va définir la matrice X de chacune des hypothèses. Sous H_0, comme on suppose qu'il n'y a aucune tendance ni rupture, la matrice ne contient qu'un seul paramètre et peut être écrite sous la forme :

$$X_0 = \begin{pmatrix} 1 \\ 1 \\ ... \\ 1 \end{pmatrix}.$$

Sous H_1, on suppose que le modèle possède une rupture, donc un changement dans la moyenne mais aucune tendance. La matrice X possède alors deux paramètres qui définissent les données qui sont avant la rupture et celles qui sont après la rupture et peut être écrite sous la forme :

$$X_1 = \begin{pmatrix} \mathbb{1}_{1<\tau} & \mathbb{1}_{1\geq\tau} \\ \mathbb{1}_{2<\tau} & \mathbb{1}_{2\geq\tau} \\ ... & ... \\ \mathbb{1}_{T<\tau} & \mathbb{1}_{T\geq\tau} \end{pmatrix}.$$

Sous H_2, on suppose que le modèle suit une tendance mais qu'il n'y a aucune rupture. La matrice X possède alors deux paramètres, soit l'ordonnée à l'origine et la pente des données. Elle peut être écrite sous la forme :

$$X_2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ ... & ... \\ 1 & T \end{pmatrix}.$$

Sous H_3, on suppose que le modèle possède une rupture et deux tendances, soit une avant la rupture et une après la rupture. La matrice X possède alors quatre paramètres, soit l'ordonnée à l'origine et la pente des données avant la rupture, ainsi que l'ordonnée à l'origine et la pente des données après la rupture. Elle peut donc être écrite sous la forme :

$$X_3 = \begin{pmatrix} \mathbb{1}_{1<\tau} & 1\mathbb{1}_{1<\tau} & \mathbb{1}_{1\geq\tau} & 1\mathbb{1}_{1\geq\tau} \\ \mathbb{1}_{2<\tau} & 2\mathbb{1}_{2<\tau} & \mathbb{1}_{2\geq\tau} & 2\mathbb{1}_{2\geq\tau} \\ ... & ... & ... & ... \\ \mathbb{1}_{T<\tau} & T\mathbb{1}_{T<\tau} & \mathbb{1}_{T\geq\tau} & T\mathbb{1}_{T\geq\tau} \end{pmatrix}.$$

Finalement sous H_4, on suppose que le modèle possède un déclenchement de tendance, c'est-à-dire qu'une tendance apparait dans les données seulement après la rupture. La matrice X possède alors deux paramètres, soit l'ordonnée à l'origine et la pente des données après la rupture. Elle peut donc être écrite sous la forme :

$$X_4 = \begin{pmatrix} 1 & 1\mathbb{1}_{1\geq\tau} \\ 1 & 2\mathbb{1}_{2\geq\tau} \\ ... & ... \\ 1 & T\mathbb{1}_{T\geq\tau} \end{pmatrix}.$$

En utilisant ces matrices, on peut donc écrire les cinq modèles sous la forme suivante :

$$\begin{aligned}
H_0 : y = X_0\beta_0 &= \alpha_0 + \epsilon, \\
H_1 : y = X_1\beta_0 &= \alpha_0 + \alpha_2\mathbb{1}_{t\geq\tau} + \epsilon, \\
H_2 : y = X_2\beta_0 &= \alpha_0 + \alpha_3 t + \epsilon, \\
H_3 : y = X_3\beta_0 &= \alpha_0 + \alpha_3 t_1\mathbb{1}_{t<\tau} + \alpha_2\mathbb{1}_{t\geq\tau} + \alpha_4 t_2\mathbb{1}_{t\geq\tau} + \epsilon, \\
&= (\alpha_0 + \alpha_2\mathbb{1}_{t\geq\tau}) + (\alpha_3 t_1\mathbb{1}_{t<\tau} + \alpha_4 t_2\mathbb{1}_{t\geq\tau}) + \epsilon, \\
H_4 : y = X_4\beta_0 &= \alpha_0 + \alpha_4 t_2\mathbb{1}_{t\geq\tau} + \epsilon,
\end{aligned}$$

où α_0, α_2, α_3 et α_4 sont des paramètres inconnus pouvant être approximés respectivement par $\widehat{\alpha}_0$, $\widehat{\alpha}_2$, $\widehat{\alpha}_3$ et $\widehat{\alpha}_4$ qui sont définis ci-dessus. Alors les modèles approximés sous les cinq hypothèses deviennent :

$$\begin{aligned}
\text{sous } H_0 &: \hat{y} = X_0\hat{\beta}_0 = \widehat{\alpha}_0, \\
\text{sous } H_1 &: \hat{y} = X_1\hat{\beta}_1 = \widehat{\alpha}_0 + \widehat{\alpha}_2\mathbb{1}_{t\geq\tau}, \\
\text{sous } H_2 &: \hat{y} = X_2\hat{\beta}_2 = \widehat{\alpha}_0 + \widehat{\alpha}_3 t, \\
\text{sous } H_3 &: \hat{y} = X_3\hat{\beta}_3 = \widehat{\alpha}_0 + \widehat{\alpha}_3 t_1\mathbb{1}_{t<\tau} + \widehat{\alpha}_2\mathbb{1}_{t\geq\tau} + \widehat{\alpha}_4 t_2\mathbb{1}_{t\geq\tau}, \\
&\qquad = (\widehat{\alpha}_0 + \widehat{\alpha}_2\mathbb{1}_{t\geq\tau}) + (\widehat{\alpha}_3 t_1\mathbb{1}_{t<\tau} + \widehat{\alpha}_4 t_2\mathbb{1}_{t\geq\tau}), \\
\text{et sous } H_4 &: \hat{y} = X_4\hat{\beta}_0 = \widehat{\alpha}_0 + \widehat{\alpha}_4 t_2\mathbb{1}_{t\geq\tau}.
\end{aligned}$$

On a donc tout le nécessaire pour approximer les densités *a priori-G* pour chacune des hypothèses. On peut donc estimer les S^* (voir l'équation à la section 3.2.1) par $\hat{S}^*$ pour chacune des hypothèses :

$$\text{sous } H_i : \hat{S}_i^* = \frac{1}{n_0+1}y^t y - \frac{1}{n_0+1}y^t X_i(X_i^t X_i)^{-1}X_i^t y + \frac{n_0}{n_0+1}(y - X_i\hat{\beta}_i)^t(y - X_i\hat{\beta}_i),$$

pour $i = 1, ..., 4$. Alors, on peut donc calculer les quatre rapports de Bayes qui nous intéressent, soit B_{10}, B_{20}, B_{30} et B_{40} (voir section 3.2.1) :

$$B_{i0} = \frac{m_i(y)}{m_0(y)} = \left(\frac{n_0}{n_0+1}\right)^{\frac{p_i - p_0}{2}} \left(\frac{\hat{S}_0^{\,*}}{\hat{S}_i^{\,*}}\right)^{\frac{T+\nu}{2}},$$

pour $i = 1, ..., 4$. De plus, en remplaçant les p_i par leur valeur respective, les rapports de Bayes sont équivalents à :

$$\begin{aligned}
B_{10} &= \sqrt{\frac{n_0}{n_0+1}} \left(\frac{\hat{S}_0^{\,*}}{\hat{S}_1^{\,*}}\right)^{\frac{T+\nu}{2}}, \\
B_{20} &= \sqrt{\frac{n_0}{n_0+1}} \left(\frac{\hat{S}_0^{\,*}}{\hat{S}_2^{\,*}}\right)^{\frac{T+\nu}{2}}, \\
B_{30} &= \left(\frac{n_0}{n_0+1}\right)^{\frac{3}{2}} \left(\frac{\hat{S}_0^{\,*}}{\hat{S}_3^{\,*}}\right)^{\frac{T+\nu}{2}},
\end{aligned}$$

$$B_{40} = \sqrt{\frac{n_0}{n_0+1}} \left(\frac{\hat{S_0}^*}{\hat{S_4}^*} \right)^{\frac{T+\nu}{2}}.$$

Par la suite il faut déterminer quelle hypothèse on doit accepter. Comme le point de rupture τ est inconnu, on utilise la statistique :

$$B_{i0}^* = \max_{t_0 \leq \tau \leq T-t_0} B_{i0},$$

pour $i = 1, 2, 3, 4$. On obtient donc les statistiques B_{10}^*, B_{20}^*, B_{30}^* et B_{40}^*. La statistique la plus grande correspond au rapport de Bayes le plus grand pour un certain jeu de données. On la note :

$$B^* = \max(B_{10}^*, B_{20}^*, B_{30}^*, B_{40}^*).$$

Geinitz (2009) stipule que Jeffreys (1961) a déterminé une échelle pour les rapports de Bayes. Si celui-ci était $< 10^{1/2}$ alors les hypothèses alternatives n'étaient pas significatives et vice-versa. Puisque l'échelle varie légèrement d'article en article (environ entre 2 et 4, Jeffreys (1961), Robert (2001)) alors, pour des fins d'esthétisme et de bons résultats dans les simulations, on a choisi e (nombre d'Euler) comme valeur critique. Alors si $B^* < e$, on accepte H_0 et si $B^* \geq e$, on accepte l'hypothèse correspondante à B^*.

Il ne nous reste plus qu'une valeur à déterminer, soit n_0. Tel que mentionné dans la section 3.2.1, n_0 appartient à l'ensemble des réels positifs. On l'a choisi de telle sorte que, si H_0 était vraie, alors le test décrit ci-dessus accepterait H_0 dans 95% des cas. Ce qui nous a mené à $n_0 = 0{,}002$. C'est donc dire que la densité *a priori* est plutôt non-informative. Pour arriver à ce résultat, on a simulé 1000 échantillons de bruit blanc normal. On a ensuite appliqué le test à ces échantillons en utilisant différentes valeurs de n_0 et on a choisit celle pour laquelle l'hypothèse nulle fut acceptée dans 95% des cas.

3.2.3. Test de rupture utilisant les B-splines et la statistique bayésienne

Dans la section 3.2.1, on a défini une méthode qui nous permettait de déceler une rupture en utilisant les coefficients d'une B-spline appliquée aux données. Dans la section 3.2.2, on a utilisé la densité *a priori-G* afin de déterminer si un certain jeu de données possédait une rupture et/ou une tendance. Dans cette section, on va combiner les deux sections précédentes afin de définir un nouveau test. Celui-ci sera composé d'une partie utilisant les B-splines et d'une autre

utilisant la statistique bayésienne. On a utilisé les données suivantes pour ce test :

$$y_i = \begin{cases} \mu_1 + \epsilon_i & \text{si } i = 1, ..., \tau \\ \mu_2 + \epsilon_i & \text{si } i = \tau + 1, ..., T, \end{cases}$$

avec les hypothèses :

$$H_0 : \mu_1 = \mu_2, \text{le jeu de données ne possède pas de rupture,}$$
$$H_1 : \mu_1 \neq \mu_2, \text{le jeu de données possède une rupture.}$$

Ces hypothèses peuvent être réécrites sous cette forme :

H_0 : Le jeu de données est mieux approximé par une B-spline;

H_1 : Le jeu de données est mieux approximé par deux B-splines.

Sous H_0, on approxime donc le jeu de données par une B-spline ayant pour noeuds extérieurs $\{1, T\}$ et les noeuds intérieurs sont $\{10, 20, ..., T - 10\}$. On se rappelle qu'à la section 1.4, on a formulé $y = (y_1, y_2, ..., y_T)$ de cette façon :

$$y = \sum_{i=1}^{k} \alpha_{0i} p_i(t) + \epsilon,$$

qui, sous forme matricielle, peut être réécrit de cette façon :

$$y = X_0 \alpha_0 + \epsilon,$$

où $\alpha_0 = (\alpha_{01}, \alpha_{02}, ..., \alpha_{0k})^t$ est le vecteur des coefficients de la spline, $p(t) = (p_1(t),\ p_2(t),\ ...,\ p_k(t))^t$ est le vecteur des polynômes qui forme la spline et $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_T)^t$ est le vecteur des erreurs. La matrice X_0 est, pour sa part, définie comme suit :

$$X_0 = \begin{pmatrix} p_1(1) & p_2(1) & ... & p_k(1) \\ p_1(2) & p_2(2) & ... & p_k(2) \\ ... & ... & ... & ... \\ p_1(T) & p_2(T) & ... & p_k(T) \end{pmatrix},$$

où $p_i(t)$ est le polynôme i de la B-spline évaluée au point i. Comme les coefficients α_0 sont inconnus, ils sont estimés par $\widehat{\alpha}_0 = (\widehat{\alpha}_{01}, \widehat{\alpha}_{02}, ..., \widehat{\alpha}_{0k})^t$.

Sous H_1, on sépare le jeu de données en deux jeux de données, soit $y_1^* = (y_1, y_2, ..., y_\tau)$ et $y_2^* = (y_{\tau+1}, y_{\tau+2}, ..., y_T)$ qui sont chacun approximé par une B-spline ayant respectivement $\{1, \tau\}$ et $\{\tau + 1, T\}$ comme noeuds extérieurs. Les noeuds intérieurs sont les mêmes que sous H_0, mais à l'intérieur de leur B-spline respective. Les données suivent donc le modèle suivant :

$$y_1^* = \sum_{i=1}^{k_1} \alpha_{1i} p_{1i}(t) + \epsilon_1,$$
$$y_2^* = \sum_{i=1}^{k_2} \alpha_{2i} p_{2i}(t) + \epsilon_2,$$

qui, sous forme matricielle, peut être réécrit de cette façon :

$$y_1^* = X_1 \alpha_1 + \epsilon_1,$$
$$y_2^* = X_2 \alpha_2 + \epsilon_2,$$

où $\alpha_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ik_i})^t$ est le vecteur des coefficients de la i^e spline, $p_i(t) = (p_{i1}(t),\ p_{i2}(t),\ ...,\ p_{ik_i}(t))^t$ est le vecteur des polynômes qui forme la i^e spline et $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{iT_i})^t$ est le vecteur des erreurs, $i = 1, 2$. Les matrices X_i est, pour sa part, définie comme suit :

$$X_i = \begin{pmatrix} p_{i1}(t_i) & p_{i2}(t_i) & ... & p_{ik_i}(t_i) \\ p_{i1}(t_i + 1) & p_{i2}(t_i + 1) & ... & p_{ik_i}(t_i + 1) \\ ... & ... & ... & ... \\ p_{i1}(T_i) & p_{i2}(T_i) & ... & p_{ik_i}(T_i) \end{pmatrix},$$

où $p_{ij}(t)$ est le polynôme j de la B-spline i évaluée au point i, $i = 1, 2$ et $j = 1, ..., k_i$. De plus, $t_1 = 1, t_2 = \tau + 1, T_1 = \tau$ et $T_2 = T$. Comme les coefficients α_i sont inconnus, ils sont estimés par $\widehat{\alpha}_i = (\widehat{\alpha}_{i1}, \widehat{\alpha}_{i2}, ..., \widehat{\alpha}_{ik_i})^t$, $i = 1, 2$. Alors les hypothèses peuvent être réécrites de la façon suivante :

$$H_0 : y = X_0 \alpha_0 + \epsilon,$$
$$H_1 : y = X_1^* \alpha_1^* + \epsilon,$$

où $X_1^* = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$ et $\alpha_1^* = (\alpha_1, \alpha_2)$. Comme les coefficients α_i sont inconnus, ils sont estimés par $\widehat{\alpha}_1^* = (\widehat{\alpha}_1, \widehat{\alpha}_2)$ qui est défini ci-dessus. Alors les modèles approximés sous les deux hypothèses deviennent :

$$\text{sous } H_0 : \hat{y} = X_0\widehat{\alpha}_0,$$
$$\text{et sous } H_1 : \hat{y} = X_1^*\widehat{\alpha}_1^*.$$

C'est donc le moment d'introduire les densités *a priori-G*. En utilisant la formule de la variance déterminée à la section 3.2.2 et la formule pour la densité *a priori-G* définie dans la section 3.2.1, on peut estimer les S^* :

$$H_0 : \widehat{S}_0^* = \frac{1}{n_0+1}y^t y - \frac{1}{n_0+1}y^t X_0(X_0^t X_0)^{-1}X_0^t y + \frac{n_0}{n_0+1}(y - X_0\widehat{\alpha}_0)^t(y - X_0\widehat{\alpha}_0),$$

$$H_1 : \widehat{S}_1^* = \frac{1}{n_0+1}y^t y - \frac{1}{n_0+1}y^t X_1^*(X_1^{*t}X_1^*)^{-1}X_1^{*t} y + \frac{n_0}{n_0+1}(y - X_1^*\hat{\alpha}_1^*)^t(y - X_1^*\hat{\alpha}_1^*).$$

Alors, on peut donc calculer le rapport de Bayes qui nous intéresse :

$$B_{10} = \left(\frac{n_0}{n_0+1}\right)^{\frac{p_1-p_0}{2}} \left(\frac{\widehat{S}_0^*}{\widehat{S}_1^*}\right)^{\frac{T+\nu}{2}}.$$

Comme le point de rupture τ est inconnu, on le fait varier et on choisit celui qui maximise B_{10}^* où :

$$B_{10}^* = \max_{t_0 \leq \tau \leq T - t_0} B_{10}.$$

Comme dans la section 3.1.2, on a choisit $t_0 = 20$ pour les mêmes raisons et, comme à la section 3.2.2, on a choisi e comme valeur critique. Alors si $B_{10}^* < e$, on accepte H_0 et si $B_{10}^* \geq e$, on accepte l'hypothèse H_1. Il ne nous reste plus qu'une valeur à déterminer, soit n_0. Tel que mentionné dans la section 3.2.1, n_0 appartient à l'ensemble des réels positifs. On l'a choisi de tel sorte que, si H_0 était vraie, alors le test décrit ci-dessus accepterait H_0 dans 95% des cas, ce qui nous a mené à $n_0 = 0{,}3$. La densité *a priori* est donc plus informative que celle à la section précédente, mais reste tout de même assez vague. Cela est dû au fait que les splines apportent plus de stabilité dans l'estimation des courbes. Les simulations ont été effectuées de la même façon que lors du choix de n_0 dans la section 3.2.2.. Bref, en utilisant la théorie des splines ainsi que celle de la statistique bayésienne, nous pouvons proposer différentes méthodes de détection de point de rupture.

Nous avons donc défini un test qui utilise seulement les splines, un qui utilise seulement la statistique bayésienne et un autre qui combine les deux. Maintenant, on va passer à la partie simulation afin de comparer ces nouvelles méthodes à celles déjà existantes. Dans les prochains chapitres, la nomenclature pour le test utilisant les splines sera le test TS, celui utilisant la statistique bayésienne sera le test TBG et celui combinant les deux sera appelé le test $TSBG$.

Chapitre 4

COMPARAISON DES DIFFÉRENTS TESTS

Dans ce chapitre, nous allons comparer les trois nouvelles méthodes définies dans le chapitre 3 avec les quatre méthodes existantes qui ont été définies dans le chapitre 2. En premier lieu, on va expliquer de quelle façon les simulations ont été effectuées, puis on va exposer les résultats par la suite.

4.1. Simulations

Les simulations ont été effectuées sur des échantillons de taille T suivant le modèle suivant, exception faite des simulations de tendances (voir section 4.4) :

$$y_i = \begin{cases} \mu_1 + \epsilon_i & \text{si } i = 1, 2, ..., \tau \\ \mu_2 + \beta\left(\frac{t_i - \tau}{T - \tau}\right) + \epsilon_i & \text{si } i = \tau + 1, \tau + 2, ..., T, \end{cases}$$

où :

$$\epsilon_i \sim \begin{cases} N(0, \sigma^2) & \text{si } i = 1, 2, ..., \tau \\ N(0, \omega\sigma^2) & \text{si } i = \tau + 1, \tau + 2, ..., T, \end{cases}$$

où ω et β seront définis par la suite. Afin de simplifier les choses, $\mu_1 = 0$ dans toutes les simulations. Chaque simulation sera formée de 1000 échantillons indépendants de taille $T = 100$ en premier lieu et de $T = 200$ par la suite. Comme beaucoup de paramètres peuvent varier : $\tau, |\mu_1 - \mu_2|, \beta, \sigma^2, \omega$ et T, il y aura beaucoup de simulations. De plus, le type de rupture peut varier. Afin de ne pas trop exagérer le nombre de simulations, on s'est basé sur les simulations exécutées dans les travaux de Dubé (2011). On fera donc des simulations en utilisant trois types de données, correspondant à trois des quatre hypothèses alternatives du test TBG, soit une rupture abrupte, une tendance ainsi qu'un déclenchement de tendance. Les simulations utilisant le type de données correspondant à la quatrième hypothèse, soit une rupture abrupte avec tendance ont été omises. On a aussi effectué des simulations de données avec un déclenchement de tendance et

un changement dans la variance afin de tester la robustesse des méthodes face à une variation dans la variance à partir de la rupture.

Lors des simulations avec des échantillons de taille $T = 200$, on a comparé la nouvelle méthode la plus efficace (le test TBG) aux deux méthodes existantes dans la littérature qui étaient les plus efficaces (Lombard, 1987 et Jarušková, 1997).

4.2. Simulations de rupture abrupte

Afin de simuler une rupture abrupte, on pose, dans le modèle présenté dans la section 4.1, $\beta = 0$, $\omega = 1$ et on fait varier μ_2 dépendamment de la taille de la rupture que l'on souhaite. Ici, on a effectué des simulations pour $\mu_2 \in \{0{,}1; 0{,}2; ...; 1\}$. On a répété ce processus pour toutes les combinaisons de $\tau \in \{0{,}25T; 0{,}50T; 0{,}75T\}$, $\sigma \in \{1, 2\}$ et $T \in \{100, 200\}$.

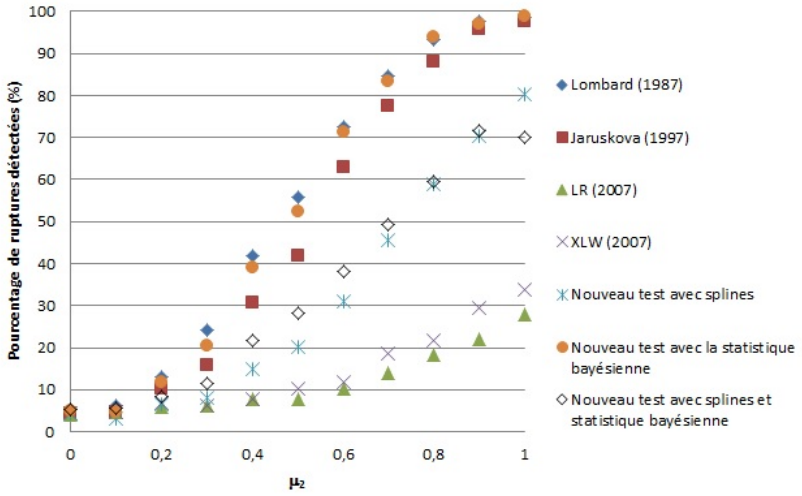

Figure 4.1. Simulations des différents tests en présence d'une rupture abrupte où $T = 100, \sigma = 1$ et $\tau = 0{,}5T$

En examinant les figures 4.1 à 4.12, comme les courbes de pourcentage de ruptures détectées ne se croisent pas trop, on remarque que la variance (σ) et la taille de l'échantillon (T) n'ont pas d'influence sur le classement des tests. Cependant, le pourcentage de détection de rupture des tests dépend de la variance et de la taille de l'échantillon. En effet, lorsque la variance augmente, le pourcentage de détection de rupture diminue et lorsque la taille de l'échantillon augmente, le pourcentage de détection de rupture fait de même. Donc, seul l'emplacement de la rupture a une influence sur le classement des tests.

Lorsque l'emplacement de la rupture est au centre des données ($\tau = 0{,}50T$), le test de Lombard (1987) est le plus puissant dans tous les cas, suivi du test

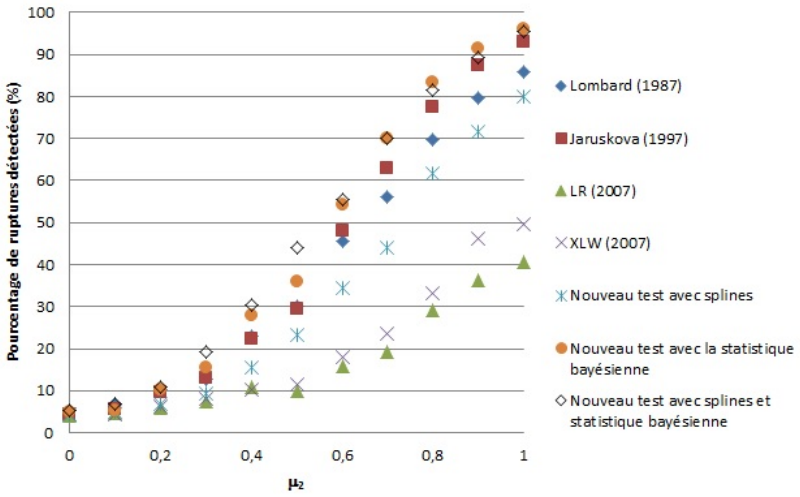

FIGURE 4.2. Simulations des différents tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 100, \sigma = 1$ et $\tau = 0{,}25T$

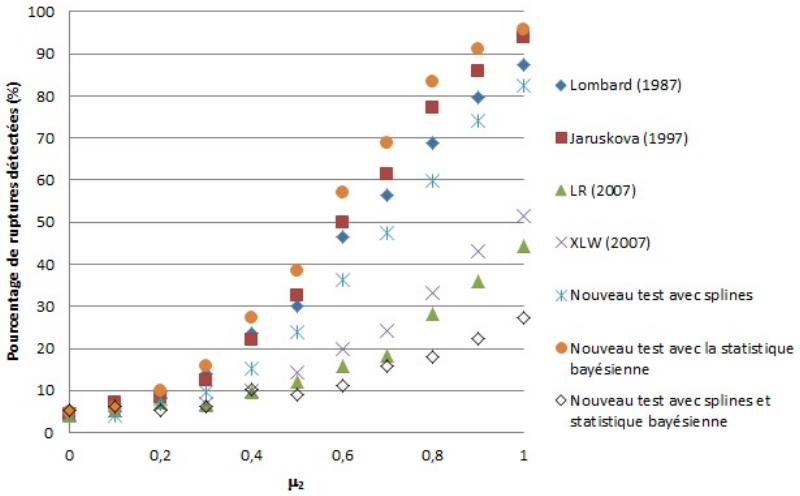

FIGURE 4.3. Simulations des différents tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 100, \sigma = 1$ et $\tau = 0{,}75T$

TBG, puis du test de Jarušková (1997). Les nouveaux tests, plus précisément les tests TS et $TSBG$, arrivent respectivement en quatrième et cinquième position. Les tests LR et XLW ferment la marche. Dans toutes les conditions de test pour une rupture abrupte, ces deux tests ne sont pas performants, tout comme il l'est indiqué dans Dubé (2011).

Donc, le test de Lombard (1987) est excellent pour détecter les ruptures abruptes et le test le plus performant lorsque celle-ci se trouve vers le milieu des données. Lorsque celle-ci se situe plus vers les extrémités, il se fait surclasser

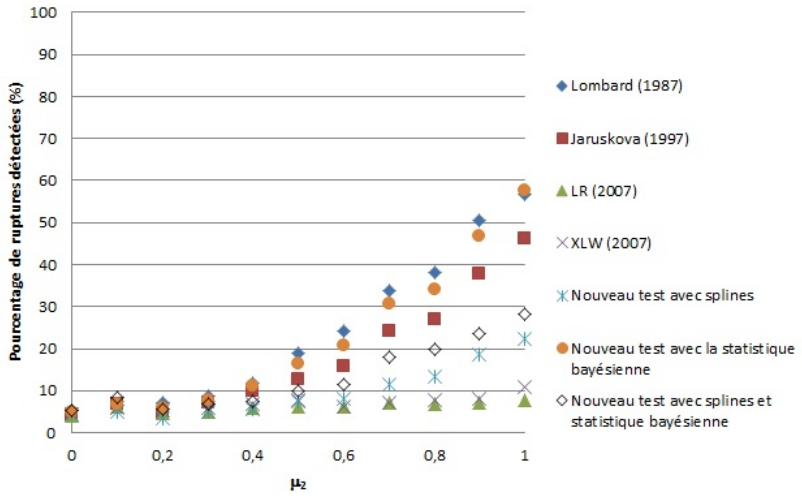

FIGURE 4.4. Simulations des différents tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 100, \sigma = 2$ et $\tau = 0{,}5T$

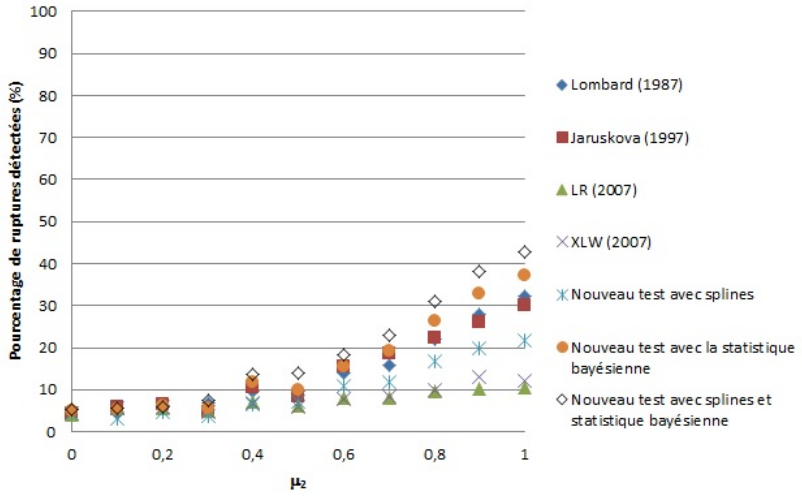

FIGURE 4.5. Simulations des différents tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 100, \sigma = 2$ et $\tau = 0{,}25T$

par le test de Jarušková (1997) et le test TBG. Le test de Jarušková (1997) est toujours performant, mais il n'est jamais celui qui se démarque le plus. Comme mentionné ci-haut, les tests LR et XLW sont les tests les moins puissants. Le test TS n'est pas mauvais, mais plusieurs autres tests sont plus performants que lui. Le test TBG est excellent, le meilleur test lorsque la rupture s'éloigne du centre des données, seul le test de Lombard (2007) est plus performant que lui lorsque la rupture est située au centre des données. Le test $TSBG$ est excellent lorsque la rupture est au début des données, mais sa puissance diminue plus la

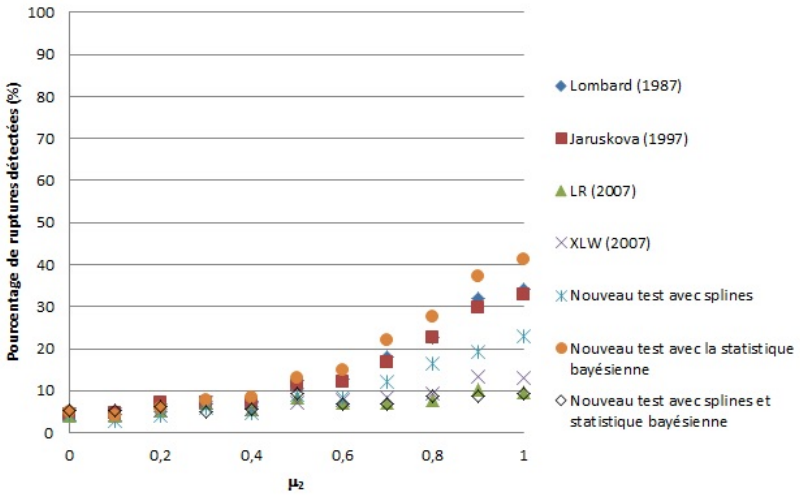

FIGURE 4.6. Simulations des différents tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 100, \sigma = 2$ et $\tau = 0{,}75T$

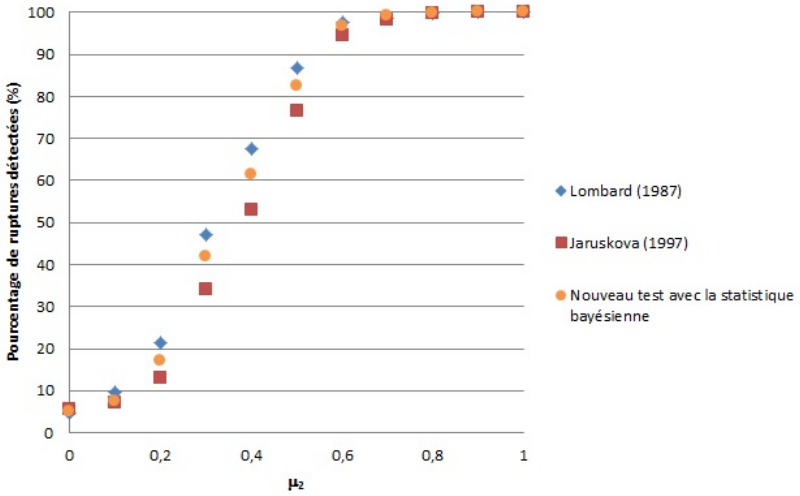

FIGURE 4.7. Simulations des trois meilleurs tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 200, \sigma = 1$ et $\tau = 0{,}5T$

rupture est vers la fin des données ; il devient même le test qui détecte le plus petit pourcentage de rupture de tous ceux étudiés.

En résumé, le test TBG est le meilleur test, dans ceux étudiés, pour détecter une rupture abrupte, suivi du test de Lombard (1987) et de Jarušková (1997). Les autres tests sont moins puissants, et c'est pour cette raison qu'on a omis les simulations de taille $T = 200$ pour ceux-ci.

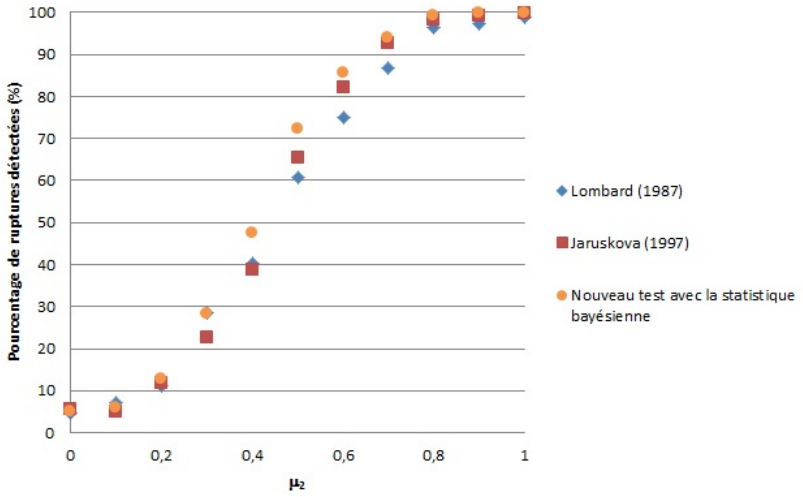

FIGURE 4.8. Simulations des trois meilleurs tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 200, \sigma = 1$ et $\tau = 0{,}25T$

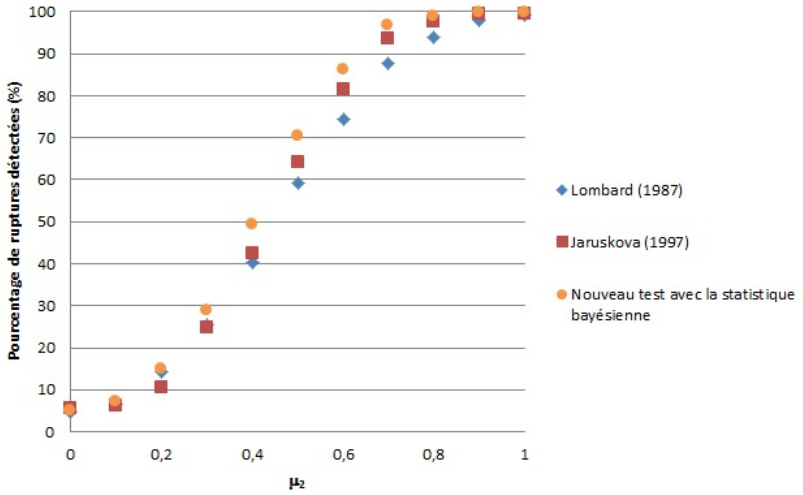

FIGURE 4.9. Simulations des trois meilleurs tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 200, \sigma = 1$ et $\tau = 0{,}75T$

4.3. SIMULATIONS DE DÉCLENCHEMENT DE TENDANCE

Afin de simuler un déclenchement de tendance, on pose $\mu_2 = 0$, $\omega = 1$ et on fait varier β dépendamment de la taille de la tendance de l'on souhaite. Ici, on a effectué des simulations pour $\beta \in \{0{,}1; 0{,}2; ...; 1\}$. On a répété ce processus pour toutes les combinaisons de $\tau \in \{0{,}25T; 0{,}50T; 0{,}75T\}$, $\sigma \in \{1, 2\}$ et $T \in \{100, 200\}$.

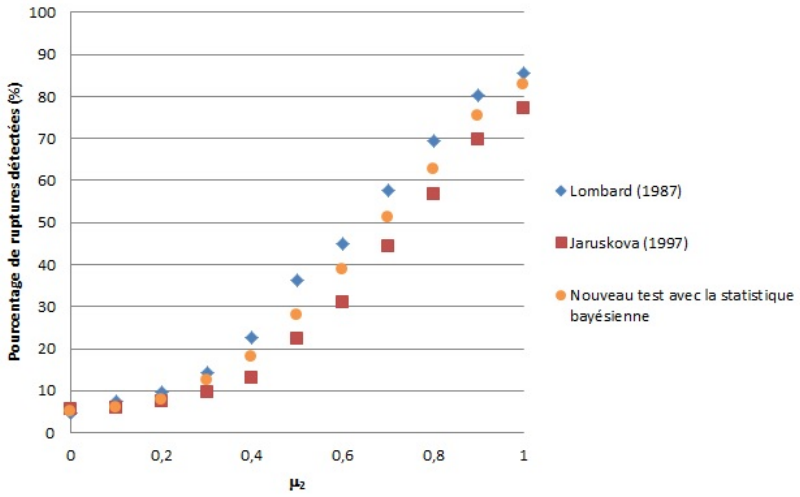

FIGURE 4.10. Simulations des trois meilleurs tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 200, \sigma = 2$ et $\tau = 0{,}5T$

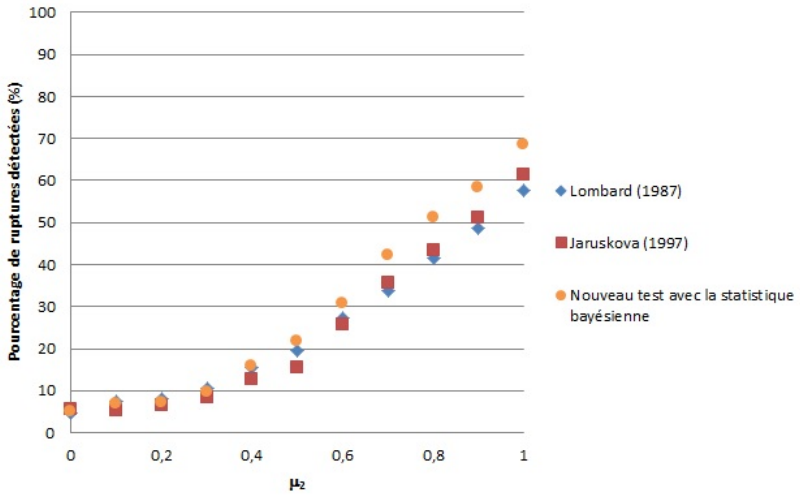

FIGURE 4.11. Simulations des trois meilleurs tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 200, \sigma = 2$ et $\tau = 0{,}25T$

En examinant les figures 4.13 à 4.24, on remarque que, tout comme c'était le cas pour les ruptures abruptes (section 4.2), la variance (σ) et la taille de l'échantillon (T) n'ont pas d'influence sur le classement des tests, mais seulement sur le pourcentage de détection de déclenchement de tendance des tests.

Lorsque l'emplacement du déclenchement de tendance est au centre des données ($\tau = 0{,}50T$), le test de Lombard (1987) et le test TBG sont les plus puissants, la différence entre ceux-ci étant minime, on ne peut déterminer lequel est le plus

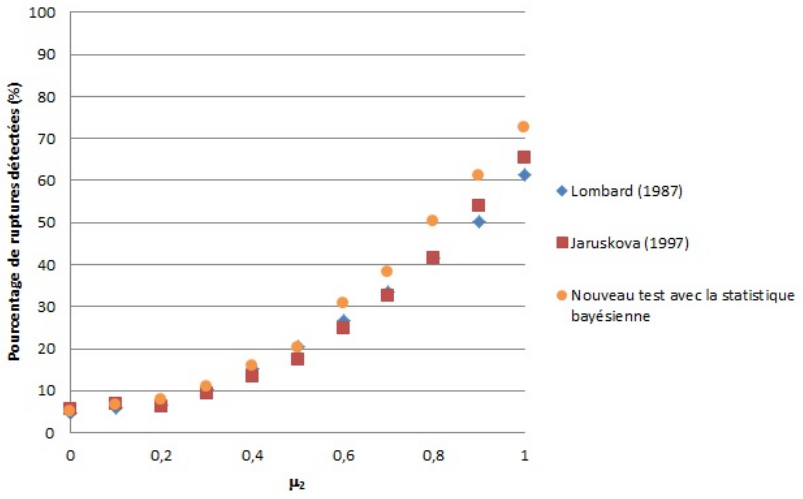

FIGURE 4.12. Simulations des trois meilleurs tests pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 200, \sigma = 2$ et $\tau = 0{,}75T$

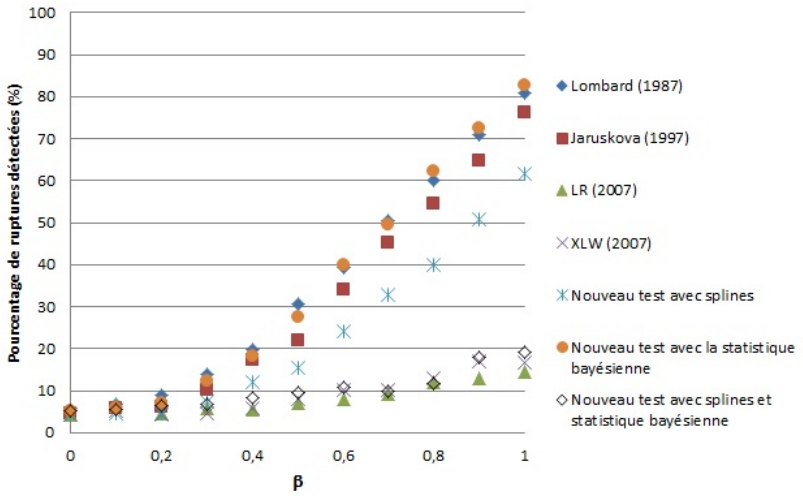

FIGURE 4.13. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 100, \sigma = 1$ et $\tau = 0{,}5T$

puissant. Le test de Jarušková (1997) les suit de près. Les nouveaux tests, c'est-à-dire les tests TS et $TSBG$, arrivent respectivement encore en quatrième et cinquième position. Les tests LR et XLW (2007) sont les moins performants dans toutes les conditions de test pour un déclenchement de tendance, tout comme c'était le cas lors d'une rupture abrupte.

Donc, le test de Lombard (1987) est excellent pour détecter les déclenchements de tendance et est le test le plus performant lorsque ceux-ci se trouvent vers le début et le milieu des données. Lorsque ceux-ci se situent plus vers la fin de

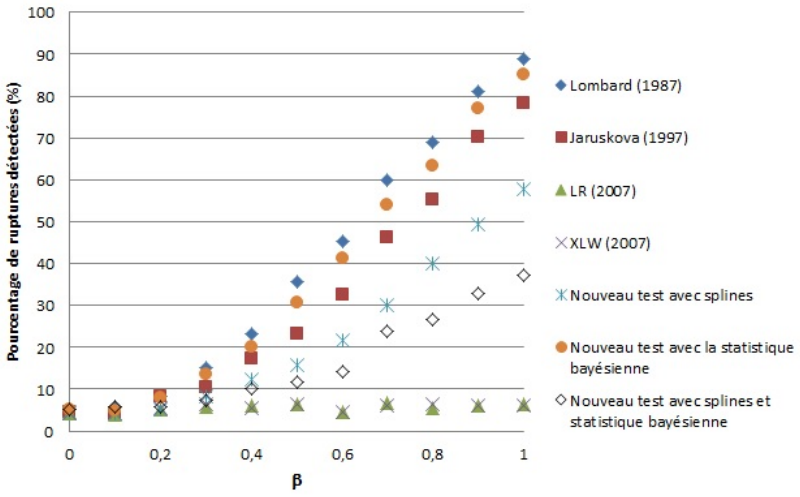

FIGURE 4.14. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 100, \sigma = 1$ et $\tau = 0{,}25T$

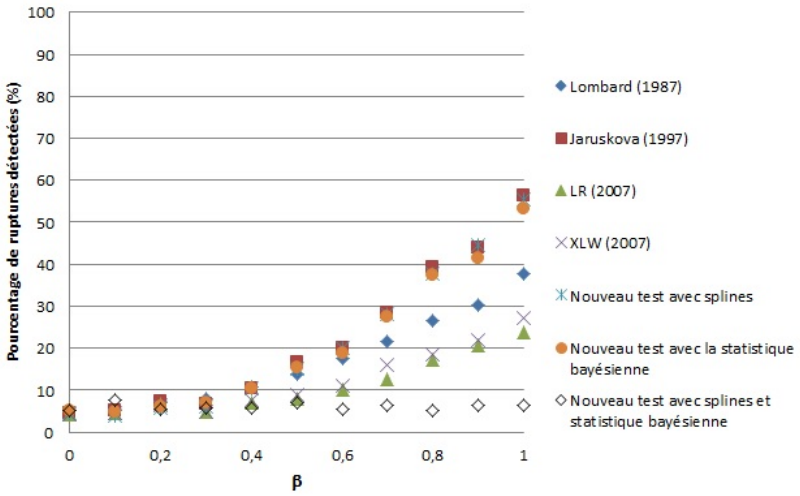

FIGURE 4.15. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 100, \sigma = 1$ et $\tau = 0{,}75T$

l'échantillon, il se fait surclasser par le test de Jarušková (1997) et le test TBG. Le test de Jarušková (1997) est toujours performant, il est le test le plus efficace si le déclenchement de tendance est à la fin des données, autant que le test TBG. Comme mentionné ci-haut, les tests LR et XLW sont les tests les moins puissants. Le test TS détecte bien les déclenchements de tendance en fin d'échantillon, mais il est moins performant pour les autres emplacements. Le test TBG est excellent, il est toujours dans les deux tests les plus performants. Seul le test de Lombard (1987) l'égalise lorsque le déclenchement de tendance est au milieu des données.

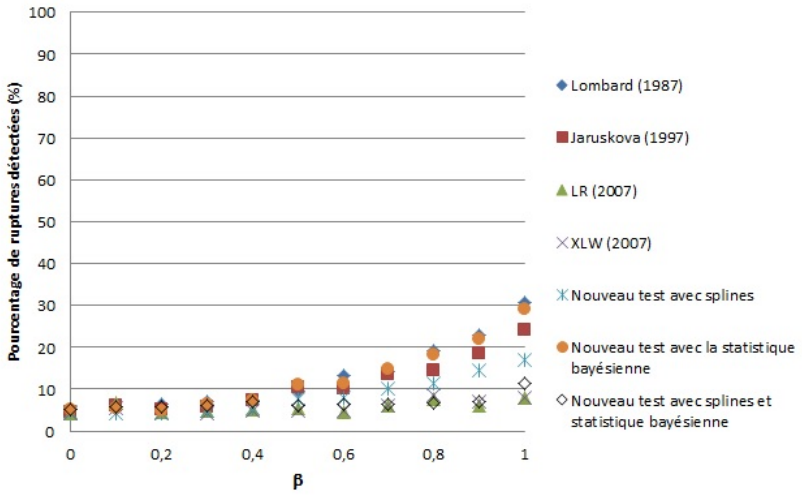

FIGURE 4.16. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 100, \sigma = 2$ et $\tau = 0{,}5T$

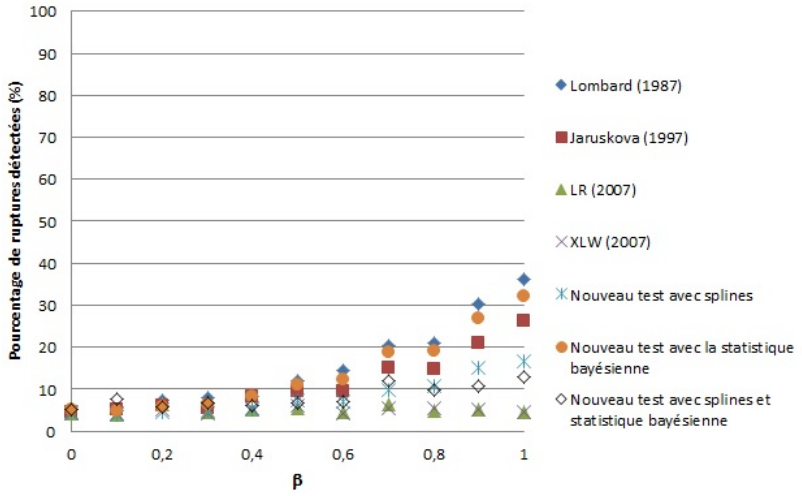

FIGURE 4.17. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 100, \sigma = 2$ et $\tau = 0{,}25T$

Le test $TSBG$ n'est pas performant, il devient même le test qui a la plus faible puissance de tous ceux étudiés lorsque le déclenchement de tendance est en fin d'échantillon.

En résumé, le test de Lombard (1987) et le test TBG sont les meilleurs tests, dans ceux étudiés, pour détecter un déclenchement de tendance, suivi du test Jarušková (1997). Les autres tests sont moins puissants, et c'est pour cette raison qu'on a omis les simulations de taille $T = 200$ pour ceux-ci.

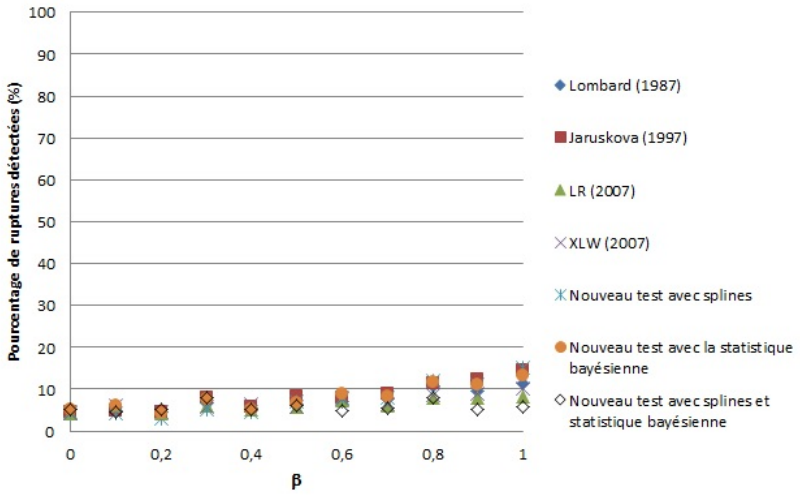

FIGURE 4.18. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 100, \sigma = 2$ et $\tau = 0{,}75T$

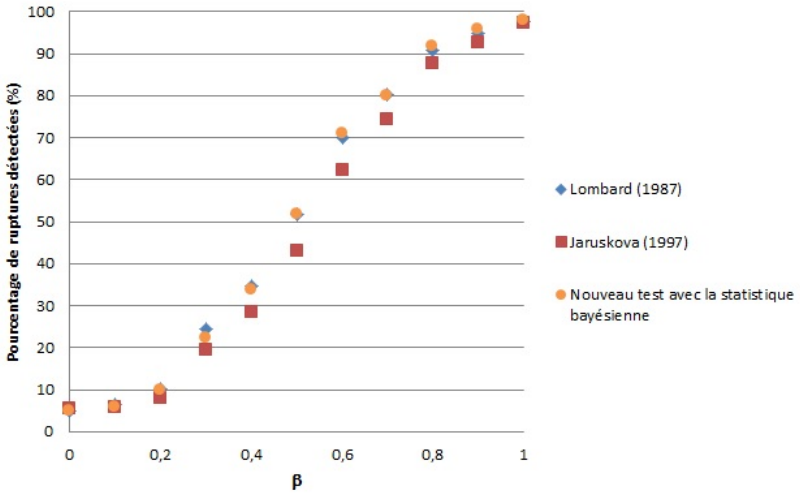

FIGURE 4.19. Simulations trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 200, \sigma = 1$ et $\tau = 0{,}5T$

4.4. SIMULATIONS DE TENDANCES

Afin de simuler une tendance, on utilise le modèle suivant :

$$y_i = \alpha t_i + \epsilon_i, i = 1, 2, ..., T,$$

où :

$$\epsilon_i \sim N(0, \sigma^2), i = 1, 2, ..., T,$$

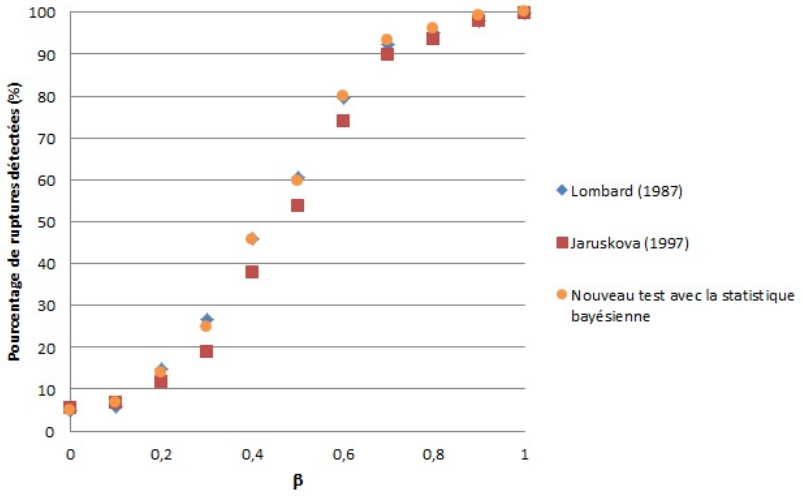

FIGURE 4.20. Simulations trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 200, \sigma = 1$ et $\tau = 0{,}25T$

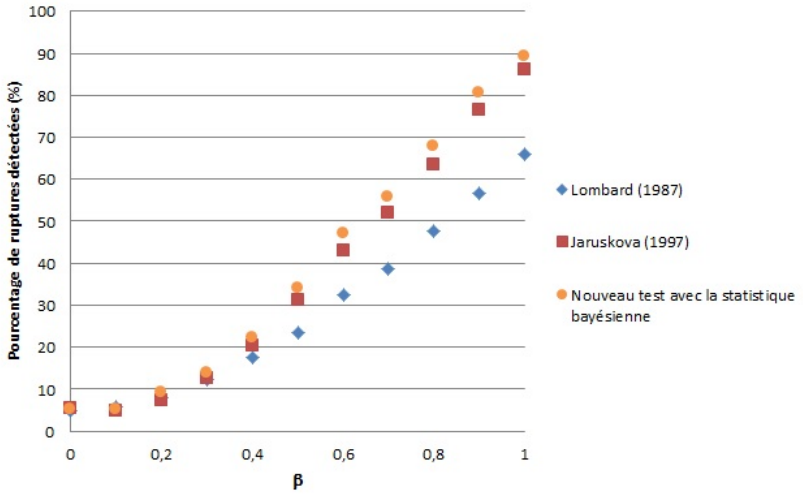

FIGURE 4.21. Simulations trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 200, \sigma = 1$ et $\tau = 0{,}75T$

et $t_i = (1, 2, ..., T)^T$. On fait varier α dépendamment de la taille de la tendance de l'on souhaite. Ici, on a effectué des simulations pour $\alpha \in \{0{,}01; 0{,}02; ...; 0{,}05\}$. On a répété ce processus pour toutes les combinaisons de $\sigma \in \{1, 2\}$ et $T \in \{100, 200\}$.

En examinant les figures 4.25 à 4.28, on remarque qu'en plus d'avoir un effet sur le pourcentage de tendances détectées, la taille de l'échantillon a un effet sur le classement des tests. En effet, pour le test de Lombard (1987), le pourcentage de détection de tendance augmente plus rapidement que celui des autres tests,

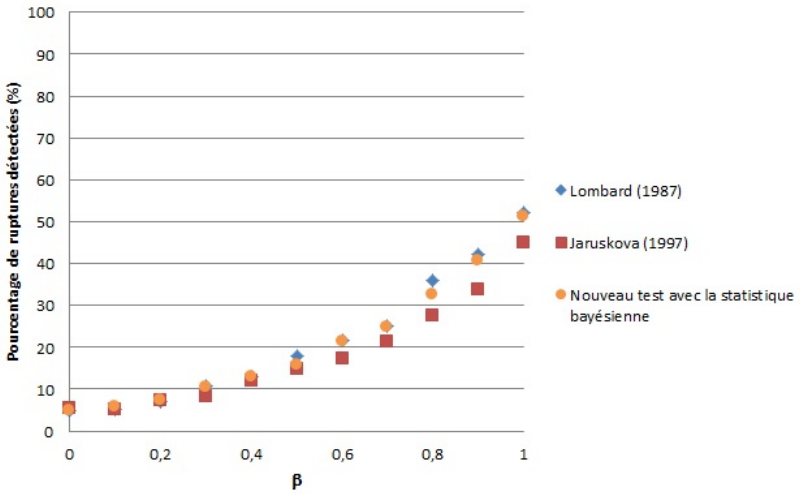

FIGURE 4.22. Simulations trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 200, \sigma = 2$ et $\tau = 0{,}5T$

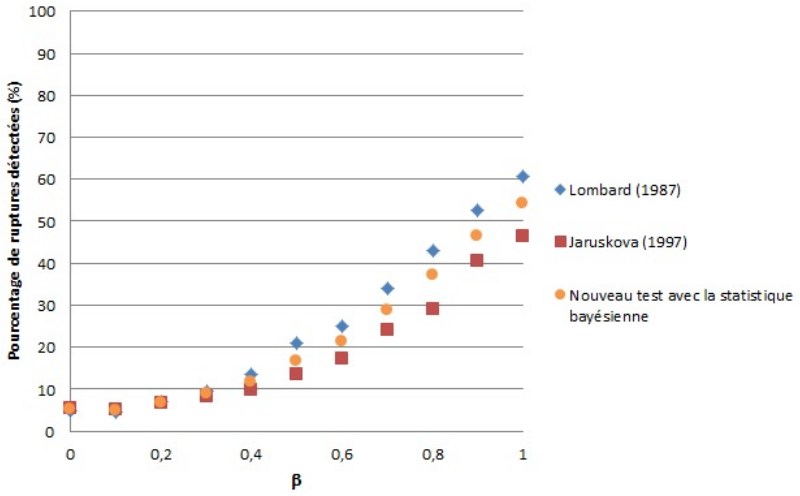

FIGURE 4.23. Simulations trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 200, \sigma = 2$ et $\tau = 0{,}25T$

passant du cinquième meilleur test lorsque la taille de l'échantillon est de 100 au meilleur lorsque celle-ci double.

On remarque que le test TS a la plus faible puissance de toutes les méthodes utilisées dans les simulations. Cependant, contrairement à la détection de rupture, les tests LR et XLW (2007) détectent très bien les tendances, n'étant surpassés que par le test TBG. Mais puisqu'ils ne sont pas efficaces pour détecter les ruptures, ils ne sont pas considérés comme de bonnes méthodes.

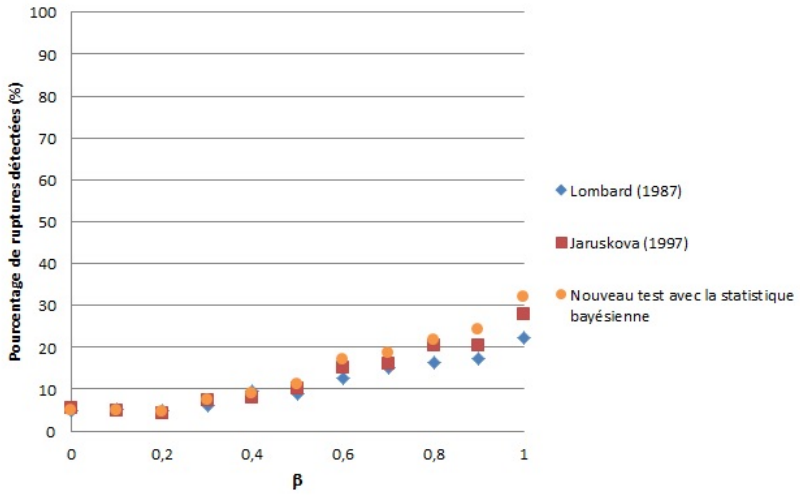

FIGURE 4.24. Simulations trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 200, \sigma = 2$ et $\tau = 0{,}75T$

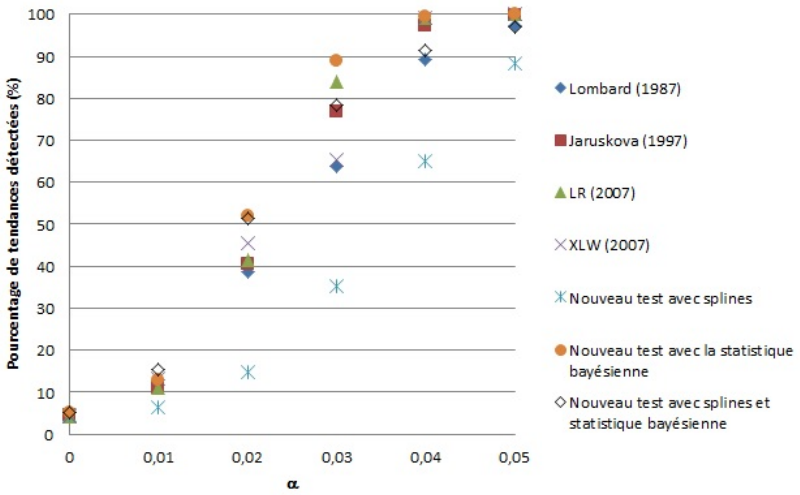

FIGURE 4.25. Simulations des différents tests pour différentes valeurs de α en présence d'une tendance où $T = 100$ et $\sigma = 1$

Le test $TSBG$ ne se classe pas comme le meilleur ni comme le plus faible des tests, tout comme c'est le cas dans la détection de rupture.

Donc, si on se restreint aux trois tests les plus efficaces pour détecter les ruptures, soit Lombard (1987), Jarušková (1997) et le test TBG, tests qui sont tout aussi efficaces pour détecter les tendances, Lombard (1987) est le plus efficace, suivi par le test TBG. En utilisant la statistique bayésienne, l'avantage est qu'il est possible de tout tester d'un coup, alors il suffit d'un seul test qui détectera tous les types de rupture.

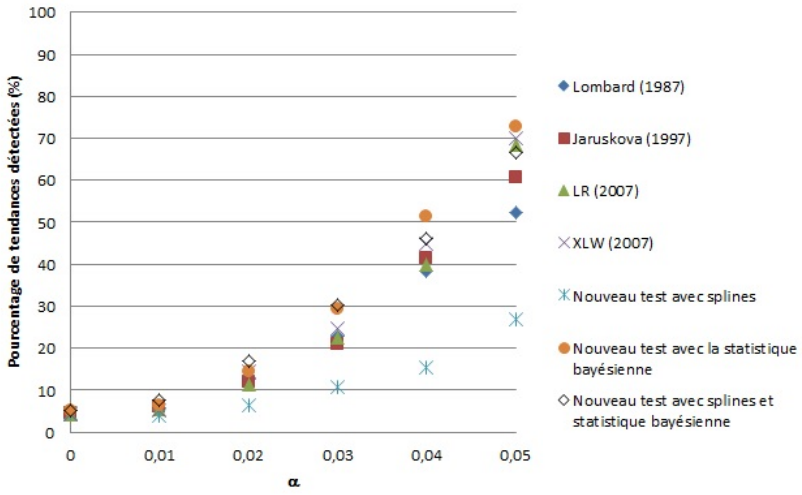

FIGURE 4.26. Simulations des différents tests pour différentes valeurs de α en présence d'une tendance où $T = 100$ et $\sigma = 2$

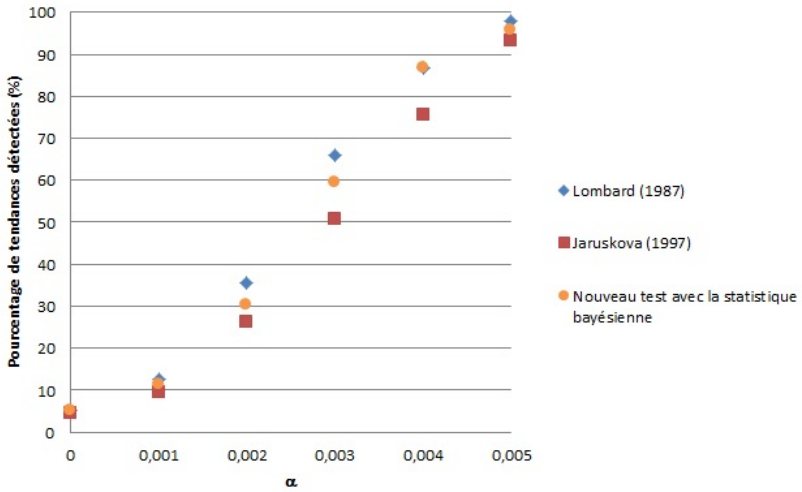

FIGURE 4.27. Simulations des trois meilleurs tests pour différentes valeurs de α en présence d'une tendance où $T = 200$ et $\sigma = 1$

Noter que la pente utilisée dans les échantillons de taille $T = 200$ est 10 fois plus petite que celle utilisée dans les échantillons de taille $T = 100$ (par exemple 0,01 pour $T = 100$ versus 0,001 pour $T = 200$) afin d'éviter que toutes les trois méthodes aient une puissance de 100%. Par conséquent, nous n'aurions pas été capable de les classer.

4.5. SIMULATIONS AVEC UN SAUT DANS LA VARIANCE

Afin de simuler un déclenchement de tendance avec un saut dans la variance, en utilisant le modèle présenté dans la section 4.1, on pose $\mu_2 = 0$ et on fait varier

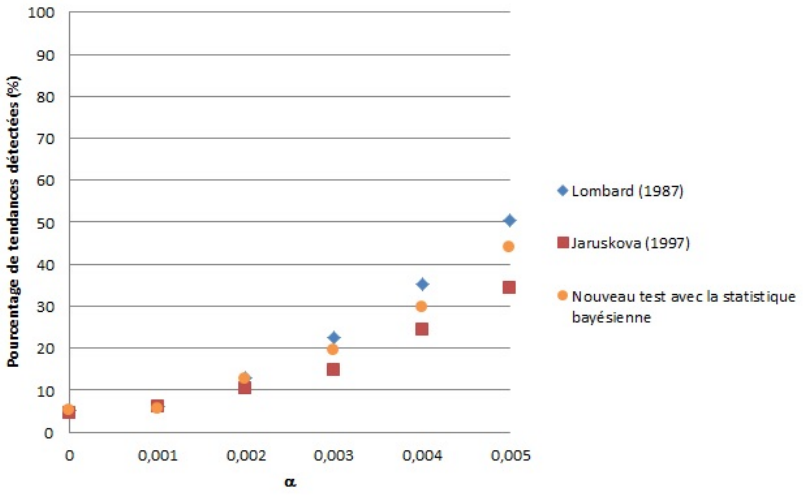

FIGURE 4.28. Simulations des trois meilleurs tests pour différentes valeurs de α en présence d'une tendance où $T = 200$ et $\sigma = 2$

β dépendamment de la taille de la tendance de l'on souhaite. Ici, on a effectué des simulations pour $\beta \in \{0{,}1; 0{,}2; ...; 1\}$. On a répété ce processus pour toutes les combinaisons de $\tau \in \{0{,}25T; 0{,}50T; 0{,}75T\}$, $\omega \in \{2, 3\}$ et $T \in \{100, 200\}$.

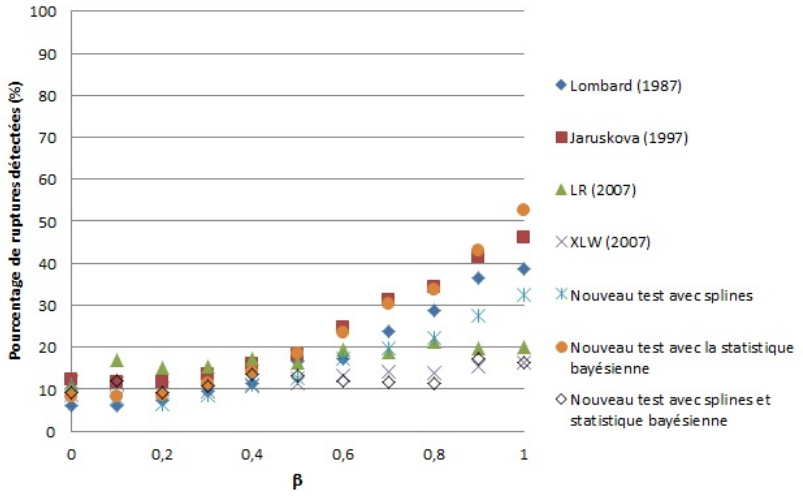

FIGURE 4.29. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 100, \omega = 2$ et $\tau = 0{,}5T$

Les simulations de déclenchement de tendance avec un saut dans la variance ont pour but de déterminer les méthodes qui sont le moins affectées par un saut dans la variance. En examinant les figures 4.29 à 4.40, on remarque que beaucoup d'entre-elles sont extrêmement sensibles à un changement brusque de variance, ce qui affecte leur capacité à détecter une rupture dans les données.

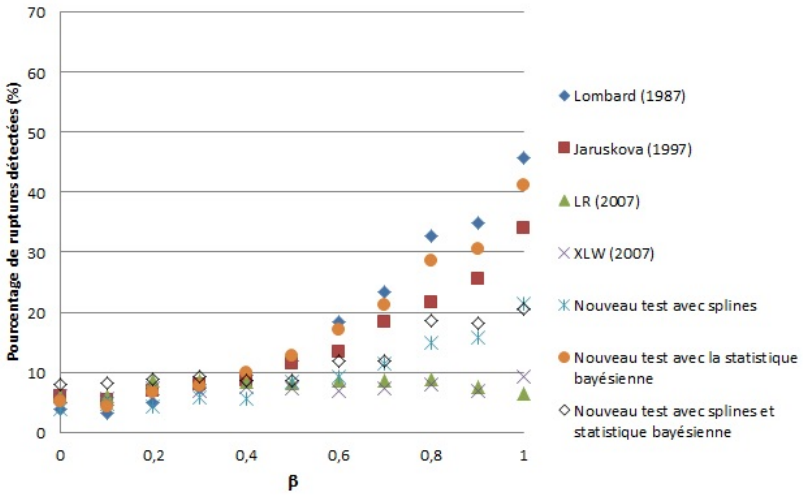

FIGURE 4.30. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 100, \omega = 2$ et $\tau = 0{,}25T$

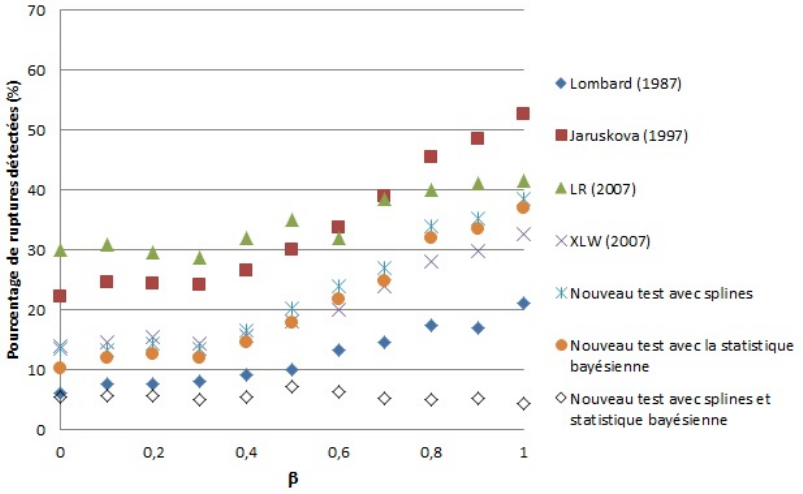

FIGURE 4.31. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 100, \omega = 2$ et $\tau = 0{,}75T$

On remarque que plus le déclenchement de tendance avec un saut dans la variance se produit vers la fin de l'échantillon, plus les méthodes utilisées sont sensibles au changement de variance. Les méthodes les plus affectées sont celles de Jarušková (1997) ainsi que les tests LR et XLW (2007). Dubé (2011) avait mentionné les lacunes de ces tests à un saut dans la variance. Les tests les moins sensibles à un saut dans la variance sont ceux de Lombard (1987) et le test TBG, celui de Lombard (1987) n'étant que très peu affecté.

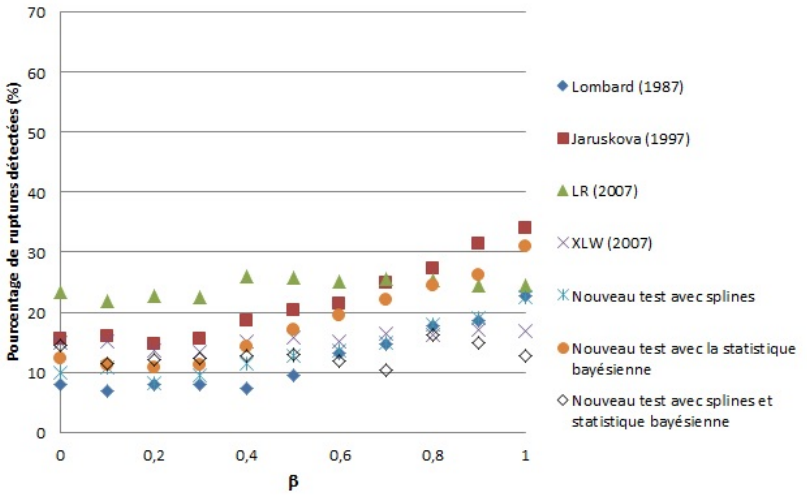

FIGURE 4.32. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 100, \omega = 3$ et $\tau = 0{,}5T$

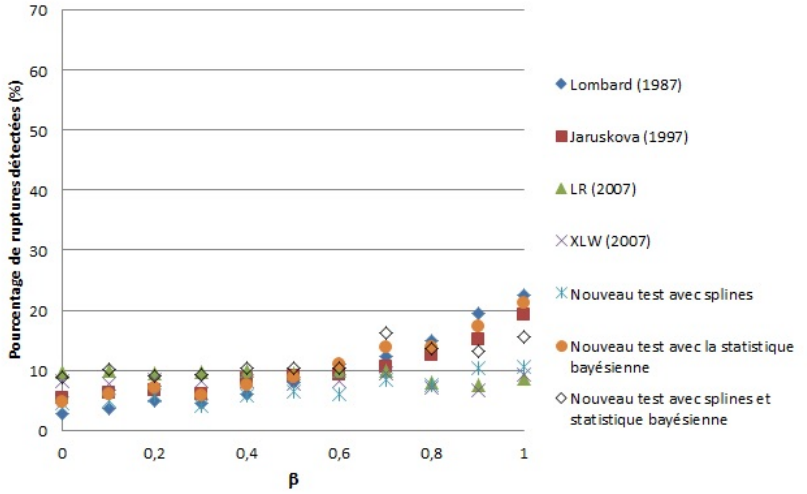

FIGURE 4.33. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 100, \omega = 3$ et $\tau = 0{,}25T$

Lorsqu'il y a un saut dans la variance, on remarque que la puissance des tests est aussi affectée ; plus le changement de variance est élevé, plus les tests perdent de la puissance. Cependant, il faut faire attention, car les tests les plus affectés par ce changement de variance semblent devenir plus puissants, mais ceux-ci sont biaisés car ils détectent une rupture là où la plupart du temps il n'y en a pas.

Donc, les simulations de déclenchement de variance avec un saut dans la variance permettent de montrer la robustesse du test de Lombard (1987) par rapport

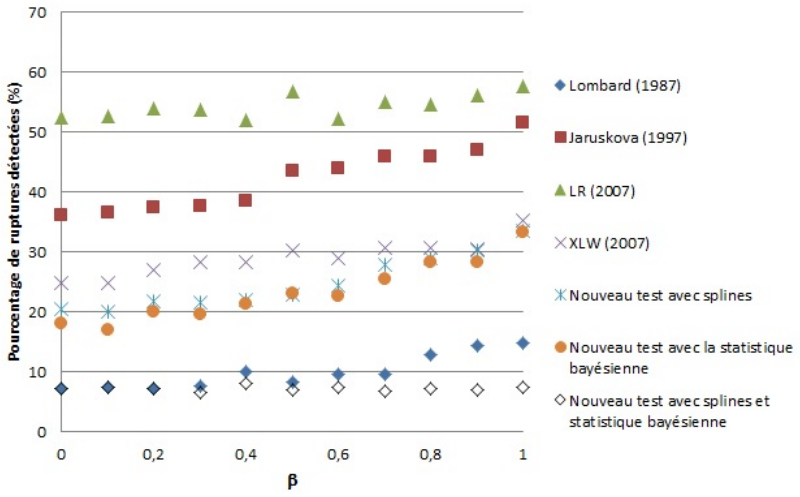

FIGURE 4.34. Simulations des différents tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 100, \omega = 3$ et $\tau = 0{,}75T$

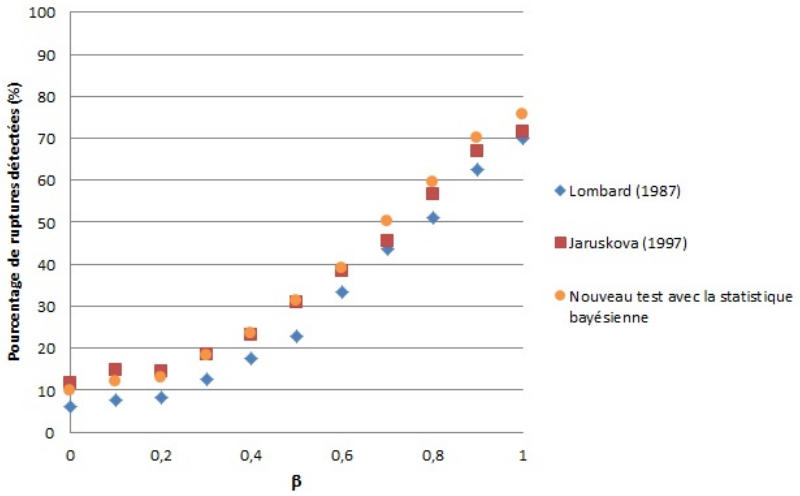

FIGURE 4.35. Simulations des trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 200, \omega = 2$ et $\tau = 0{,}5T$

aux autres tests du même type. Le test TBG est aussi assez robuste, mais n'égale jamais le test de Lombard (1987).

Le niveau du test $TSBG$ est assez stable en présence d'un saut dans la variance, mais la puissance de ce test en est beaucoup affectée. Ce qui en fait un test inefficace.

Bref, la plupart des tests sont extrêmement sensibles à un saut dans la variance, ce qui affecte leur capacité à bien détecter les déclenchements de tendance et les ruptures. Seul le test de Lombard (1987) reste robuste à celui-ci.

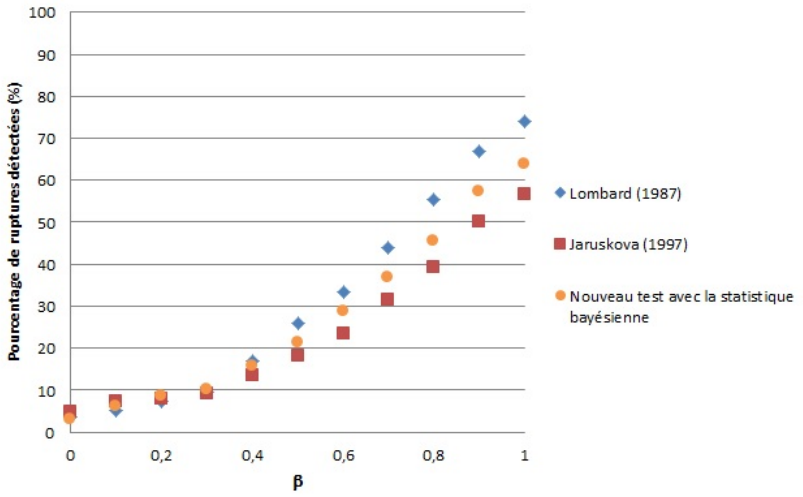

FIGURE 4.36. Simulations des trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 200, \omega = 2$ et $\tau = 0{,}25T$

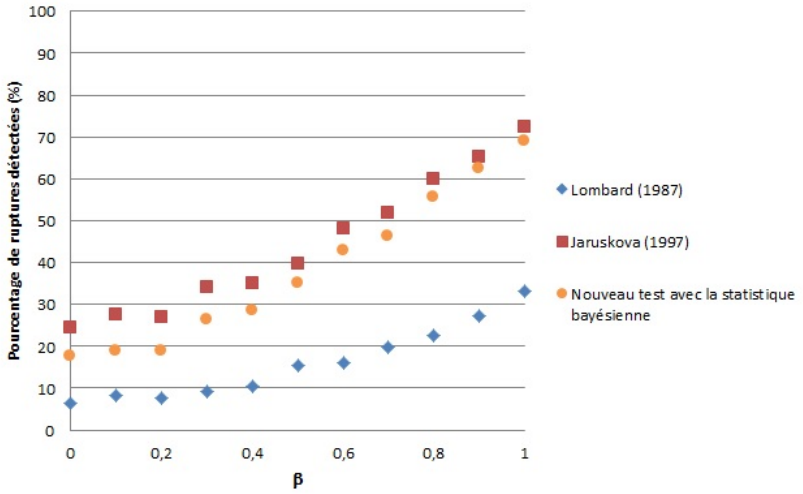

FIGURE 4.37. Simulations des trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 200, \omega = 2$ et $\tau = 0{,}75T$

Pour faire un court résumé des sections 4.2 à 4.5, il y a trois méthodes qui sont nettement plus efficaces que les autres pour détecter les ruptures et les tendances, soit les tests de Lombard (1987), de Jarušková (1997) et le test TBG. De ces trois tests, Lombard (1987) et le test TBG sont nettement supérieurs. Le test de Lombard (1987) est le plus robuste à un changement dans la variance, il est le meilleur lorsque la rupture abrupte se trouve au milieu de l'échantillon et si le déclenchement de tendance se produit au début de l'échantillon. Le test TBG est supérieur au test de Lombard (1987) lorsque la rupture abrupte est aux

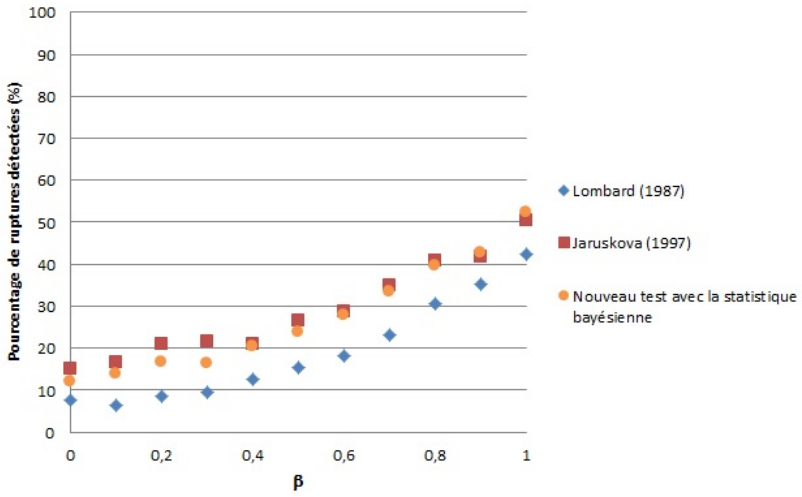

FIGURE 4.38. Simulations des trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 200, \omega = 3$ et $\tau = 0{,}5T$

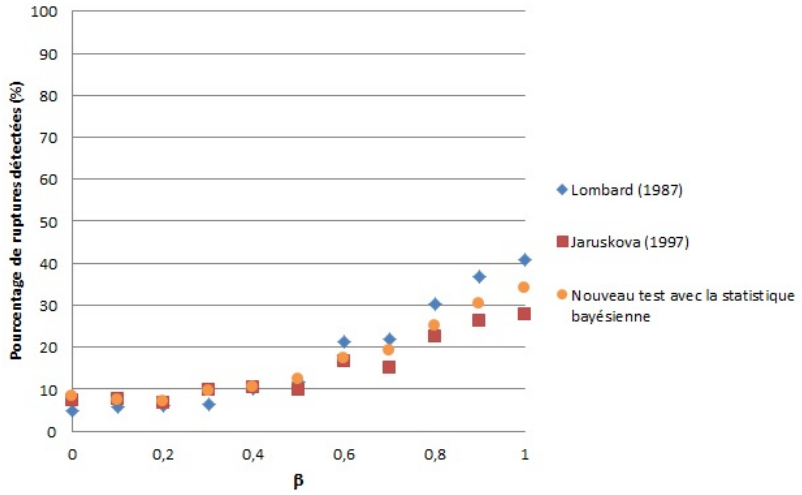

FIGURE 4.39. Simulations des trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 200, \omega = 3$ et $\tau = 0{,}25T$

extrémités de l'échantillon ou si le déclenchement de tendance est plus vers la fin de l'échantillon. Il est cependant un peu moins robuste à un changement de variance.

4.6. TEST UTILISANT LA STATISTIQUE BAYÉSIENNE (TBG)

Dans cette section, une étude individuelle du test TBG sera effectué. Dubé (2011) a déjà effectué une étude individuelle des tests de Lombard (1987) et de

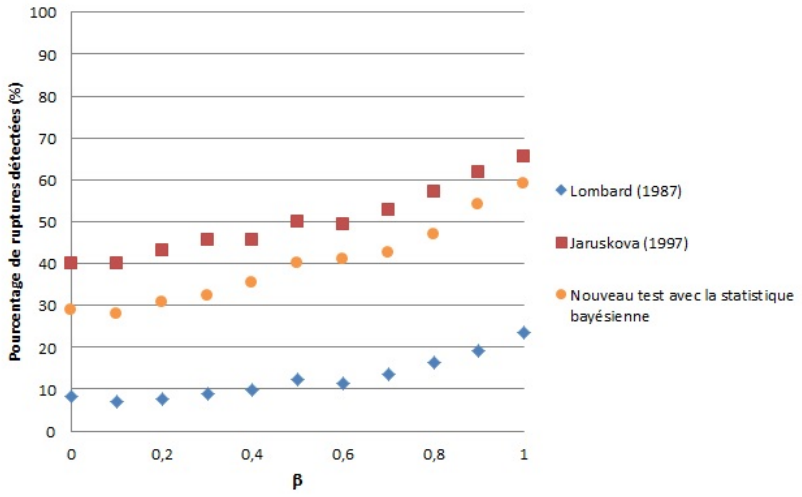

FIGURE 4.40. Simulations des trois meilleurs tests pour différentes valeurs de β en présence d'un déclenchement de tendance avec un saut dans la variance où $T = 200, \omega = 3$ et $\tau = 0{,}75T$

Jarušková (1997) dans son compte rendu, il n'est donc pas nécessaire de le refaire ici. Les figures 4.41 à 4.46 regroupent tous les résultats des simulations du test TBG classés par type de rupture (abrupte, déclenchement de tendance et déclenchement de tendance avec un saut dans la variance) et par taille d'échantillon ($T = 100$ et $T = 200$).

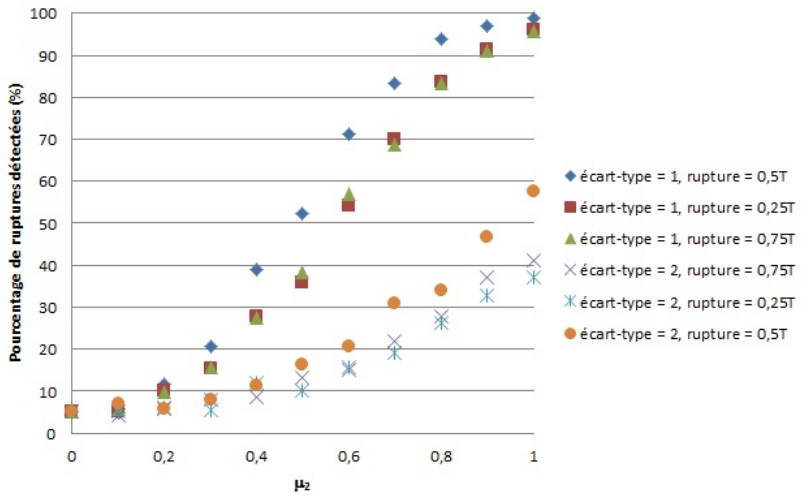

FIGURE 4.41. Simulations du test TBG pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 100$

En étudiant individuellement le test TBG, on remarque que si l'on fixe le type de rupture, la variance de l'échantillon et l'emplacement de la rupture, le pourcentage de détection de rupture augmente proportionnellement avec la taille

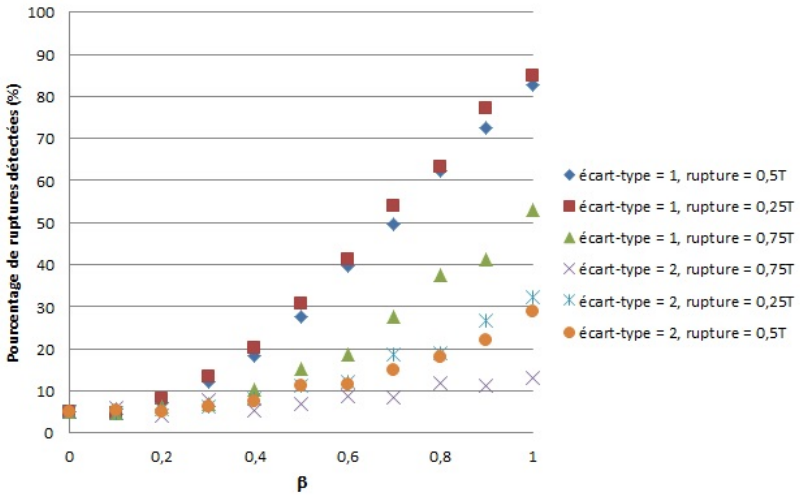

FIGURE 4.42. Simulations du test TBG pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 100$

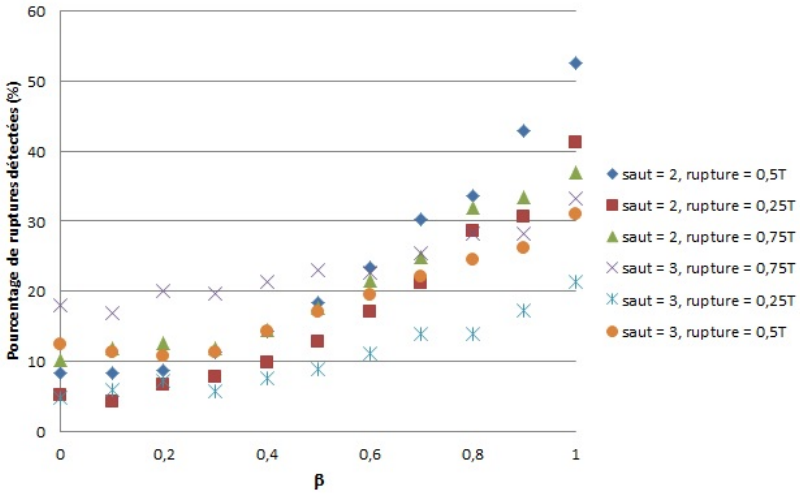

FIGURE 4.43. Simulations du test TBG pour différentes valeurs de β en présence d'un déclenchement de tendance avec saut dans la variance où $T = 100$

de l'échantillon. De plus, celui-ci diminue proportionnellement avec la variance lorsque le type et l'emplacement de la rupture ainsi que la taille de l'échantillon ne changent pas. Aussi, toutes choses étant égales par ailleurs, on remarque que le test TBG est plus efficace à détecter une rupture abrupte qu'un déclenchement de tendance et ce, peu importe l'emplacement de la rupture et la variance de l'échantillon.

Le cas le plus intéressant est lorsque l'emplacement de la rupture change, toutes choses étant égales par ailleurs. En premier lieu, lorsqu'on est en présence d'une rupture abrupte, le test TBG est plus puissant lorsque la rupture se produit

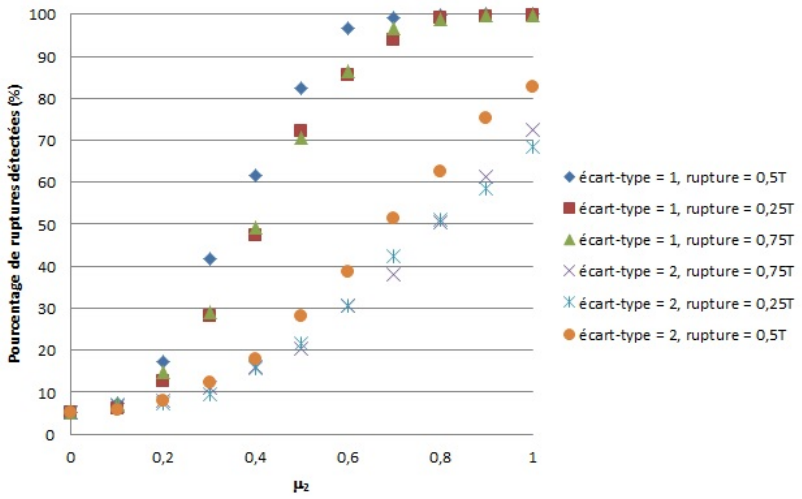

FIGURE 4.44. Simulations du test TBG pour différentes valeurs de μ_2 en présence d'une rupture abrupte où $T = 200$

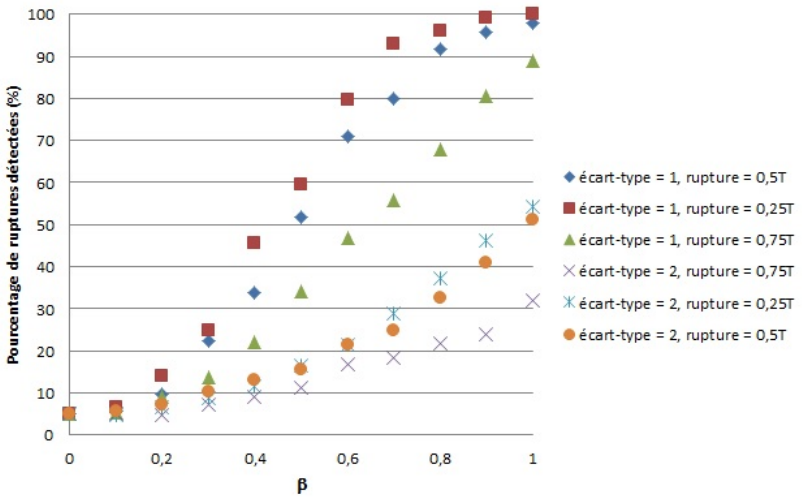

FIGURE 4.45. Simulations du test TBG pour différentes valeurs de β en présence d'un déclenchement de tendance où $T = 200$

en milieu d'échantillon et l'est legèrement moins lorsque celle-ci se produit aux extrémités de l'échantillon. En présence d'un déclenchement de tendance, le test détecte plus de ruptures lorsque celles-ci sont au début de l'échantillon et la puissance du test diminue graduellement lorsque le déclenchement de tendance se déplace vers la fin de l'échantillon.

Lorsqu'on est en présence d'un déclenchement de tendance avec un saut dans la variance, on remarque que le niveau du test en est proportionnellement affecté, surtout lorsque la rupture se produit vers la fin de l'échantillon. Si le déclenchement de tendance se produit au début de l'échantillon, le niveau du test ne sera

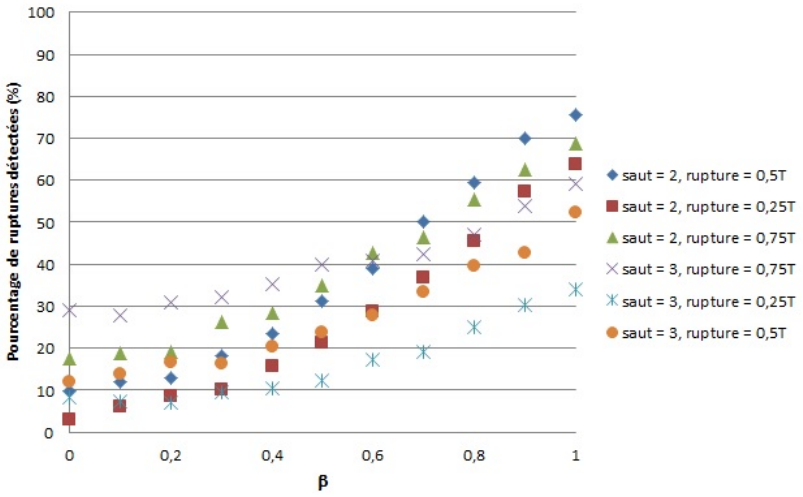

FIGURE 4.46. Simulations du test TBG pour différentes valeurs de β en présence d'un déclenchement de tendance avec saut dans la variance où $T = 200$

presque pas affecté par un saut dans la variance. Dans tous les cas, la puissance du test diminue en présence d'un saut dans la variance.

4.7. POURCENTAGE DE DÉTECTION DE LA RUPTURE

Dans cette section, nous allons rapidement analyser le pourcentage de détection de rupture du test TBG. Bien que le test soit efficace pour détecter les ruptures abruptes et les déclenchement de tendance, nous voulons vérifier si celui-ci est capable d'estimer l'emplacement réel de la rupture ; si tel est le cas, le test pourra donc être utilisé aussi à cette fin.

On a donc fait une simulation de 1000 échantillons pour chaque type de rupture (abrupte et déclenchement de tendance), d'emplacement $\{0{,}25T; 0{,}5T; 0{,}75T\}$ et de taille d'échantillon $\{100, 200\}$ qu'on avait simulé dans la section précédente (voir le modèle présenté à la section 4.1) et tel que Dubé (2011) l'avait fait pour Lombard (1987) et Jarušková (1997), on va présenter les résultats généraux pour une détection de rupture dans un intervalle de $\pm$ 5% et $\pm$ 10% de la vraie position du point de rupture spécifié dans la simulation. Par exemple, en supposant que dans un échantillon de 100 données, la rupture réelle se situe entre l'observation $t = 50$ et $t = 51$, que l'hypothèse alternative du test est acceptée et que la rupture détectée par le test se situe à quelque part entre les observations $t = 45$ et $t = 55$, cela équivaut à une détection à $\pm$ 5%. Le test détecte une rupture près de l'emplacement réel de la rupture, emplacement qui est inconnu en présence de données réelles. En présence d'une rupture abrupte, en utilisant $\beta = 1$, on arrive aux résultats présentés dans les tableaux 4.1 et 4.2.

TABLEAU 4.1. Pourcentage de détection de rupture abrupte ($\sigma = 1$)

	± 5%			± 10%		
Emplacement	$0{,}25T$	$0{,}50T$	$0{,}75T$	$0{,}25T$	$0{,}50T$	$0{,}75T$
Pourcentage	84,6%	76,8%	82,7%	94,5%	89,9%	89,8%

TABLEAU 4.2. Pourcentage de détection de rupture abrupte ($\sigma = 2$)

	± 5%			± 10%		
Emplacement	$0{,}25T$	$0{,}50T$	$0{,}75T$	$0{,}25T$	$0{,}50T$	$0{,}75T$
Pourcentage	61,2%	45,5%	57,1%	69,9%	61,5%	65,3%

En présence d'un déclenchement de tendance, en utilisant $\beta = 2$, on arrive aux résultats présentés dans les tableaux 4.3 et 4.4.

TABLEAU 4.3. Pourcentage de détection de déclenchement de tendance ($\sigma = 1$)

	± 5%			± 10%		
Emplacement	$0{,}25T$	$0{,}50T$	$0{,}75T$	$0{,}25T$	$0{,}50T$	$0{,}75T$
Pourcentage	13,3%	13,3%	82,7%	23,3%	31,4%	89,8%

TABLEAU 4.4. Pourcentage de détection de déclenchement de tendance ($\sigma = 2$)

	± 5%			± 10%		
Emplacement	$0{,}25T$	$0{,}50T$	$0{,}75T$	$0{,}25T$	$0{,}50T$	$0{,}75T$
Pourcentage	11,8%	11,0%	39,7%	34,1%	24,1%	43,7%

Comme on le remarque, le pourcentage de détection de rupture est beaucoup plus petit en présence d'un déclenchement de tendance qu'une rupture, comme son nom l'indique. Ce qui était prévisible, puisque la rupture lors d'un déclenchement de tendance est moins prononcée qu'une rupture abrupte. Avant celle-ci, pour les cas de déclenchement de tendance, le test n'estime presqu'aucun emplacement de rupture et en détecte graduellement après celle-ci. C'est pourquoi, lorsque l'emplacement réel de la rupture est vers la fin de l'échantillon, le test détecte plus de rupture près de l'emplacement réel de celle-ci. Dans le cas d'une rupture abrupte, le test va détecter une rupture également avant et après la position réelle de la rupture.

4.8. TYPE DE RUPTURE

Dans cette section, nous allons rapidement analyser si le test TBG détecte le bon type de rupture. À l'instar des autres méthodes utilisées dans ce mémoire, ce

test peut déterminer quel type de rupture est présent dans l'échantillon observé. Pour effectuer cela, les mêmes échantillons qu'à la section précédente seront utilisées soit en utilisant $\beta = 1$ pour une rupture abrupte et en utilisant $\beta = 2$ pour un déclenchement de tendance.

TABLEAU 4.5. Type de rupture détectée en présence d'une rupture abrupte

	$\sigma = 1$			$\sigma = 2$		
Emplacement	$0{,}25T$	$0{,}50T$	$0{,}75T$	$0{,}25T$	$0{,}50T$	$0{,}75T$
Rupture abrupte	95,5%	77,7%	65,1%	95,5%	85,4%	83,5%
Déclenclement de tendance	3,6%	21,4%	34,0%	3,8%	14,0%	15,8%
Tendance	0,5%	0,5%	0,5%	0,1%	0,3%	0,1%
Ne peut le déterminer	0,4%	0,4%	0,4%	0,6%	0,3%	0,6%

TABLEAU 4.6. Type de rupture détectée en présence d'un déclenchement de tendance

	$\sigma = 1$			$\sigma = 2$		
Emplacement	$0{,}25T$	$0{,}50T$	$0{,}75T$	$0{,}25T$	$0{,}50T$	$0{,}75T$
Rupture abrupte	12,1%	9,9%	4,1%	56,8%	47,9%	47,9%
Déclenclement de tendance	77,8%	86,9%	87,9%	41,5%	49,8%	49,8%
Tendance	9,8%	2,5%	0,5%	1,4%	1,1%	1,1%
Ne peut le déterminer	0,3%	0,7%	8,3%	0,3%	1,2%	1,2%

En regardant les tableaux 4.5 et 4.6, on remarque que le type de rupture détectée par le test TBG est, dans la plupart des cas, précis lorsqu'il s'agit d'une rupture abrupte mais il l'est moins lorsqu'il s'agit d'un déclenchement de tendance, surtout lorsque la variance augmente. Plus la variance augmente, plus il est difficile de détecter une rupture, mais le lien n'est pas linéaire. Bref, le but premier du test était de détecter si un échantillon possédait ou non une rupture, ce que celui-ci fait aussi bien que les méthodes déjà existantes. Les buts secondaires étaient de déterminer si celui-ci pouvait estimer l'emplacement de la rupture et déterminer le type de celle-ci. Nous pouvons affirmer, que le test est capable de le faire, la marge d'erreur étant plus élevée dans le second but cependant.

Bref, après avoir comparé les trois méthodes proposées au chapitre 3 avec celles déjà acceptées dans la littérature (présentées au chapitre 2), nous avons observé que la méthode TBG est la plus efficace des trois et est même aussi performante dans certaines situations que les méthodes déjà présentes dans la littérature. En approfondissant un peu plus l'analyse de cette méthode, on observe qu'elle peut détecter l'emplacement et le type de la rupture avec une marge d'erreur assez respectable. Dans le chapitre suivant, nous allons appliquer cette nouvelle méthode sur des données réelles tout en comparant les résultats observés avec

ceux du test de Lombard (1987) qui est le test accepté dans la littérature le plus efficace présenté dans ce mémoire.

Chapitre 5

APPLICATION SUR DONNÉES RÉELLES

Dans ce chapitre, le test TBG sera appliqué sur des données réelles. Le but est de déterminer si le test peut aussi bien faire que les autres tests dans les cas où le type de bruit est inconnu. Pour ce faire, la méthode de Lombard (1987) sera aussi appliquée aux données réelles. Comme celle-ci est déjà acceptée dans la littérature et est la meilleure des méthodes déjà existantes traitées dans ce mémoire, les résultats attendus par le test TBG devraient en principe ressembler aux résultats obtenus en utilisant le test de Lombard (1987).

5.1. Différents types de données réelles

Le test de Lombard et le test TBG ont donc été appliqués sur plusieurs échantillons de données tous aussi différents les uns des autres. Ces données proviennent des fichiers de données déjà présentes et en utilisation libre du logiciel R version 2.14.2. Le premier échantillon (figure 5.1) est la profondeur du lac Huron entre les années 1875 et 1972. Les deux méthodes indiquent qu'il y a une rupture. Le test TBG indique qu'il s'agit d'une rupture abrupte vers l'année 1887.

La figure 5.2 représente le nombre de décès, par mois, entre les années 1974 et 1980, du cancer du poumon en Grande-Bretagne. Cette fois-ci, le test de Lombard (1987) indique qu'il y a une rupture tandis que le test TBG n'en détecte pas. Il est intéressant ici de noter que la statistique du test calculée par Lombard (1987) est de 0,0583 comparativement à sa valeur p qui est de 0,0402. Donc, le test passe très proche de ne pas détecter de rupture statistiquement significative.

La figure 5.3 représente le nombre de lynx chassés au Canada entre les années 1821 et 1934. Le test de Lombard (1987) et le test TBG ne détectent aucune rupture dans cet échantillon.

La figure 5.4 illustre la température moyenne qu'il faisait, par année, à New Haven entre les années 1912 à 1971. Dans ce cas-ci, les deux méthodes détectent

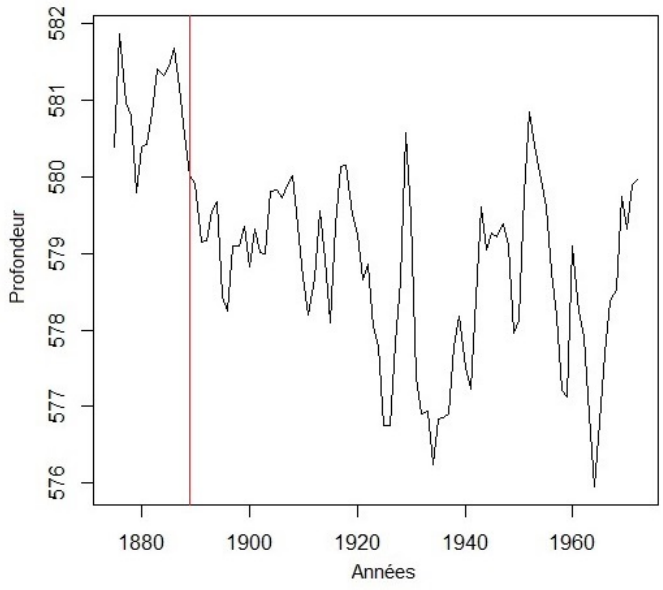

FIGURE 5.1. Profondeur du Lac Huron

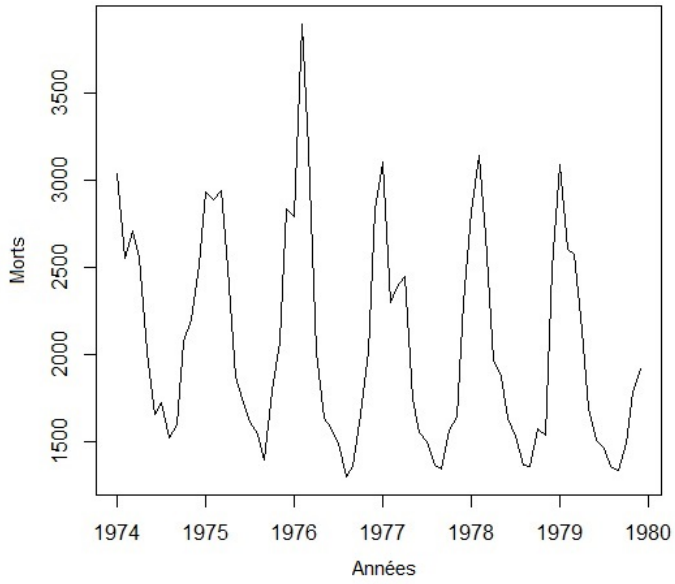

FIGURE 5.2. Nombre de décès causés par le cancer du poumon en Grande-Bretagne

une rupture. Le test TBG estime qu'il s'agit d'une rupture abrupte vers l'année 1944.

La figure 5.5 illustre la température moyenne, par mois, qu'il faisait à Nottingham entre les années 1920 à 1940. Contrairement à la température moyenne de New Haven, les deux méthodes ne détectent pas rupture.

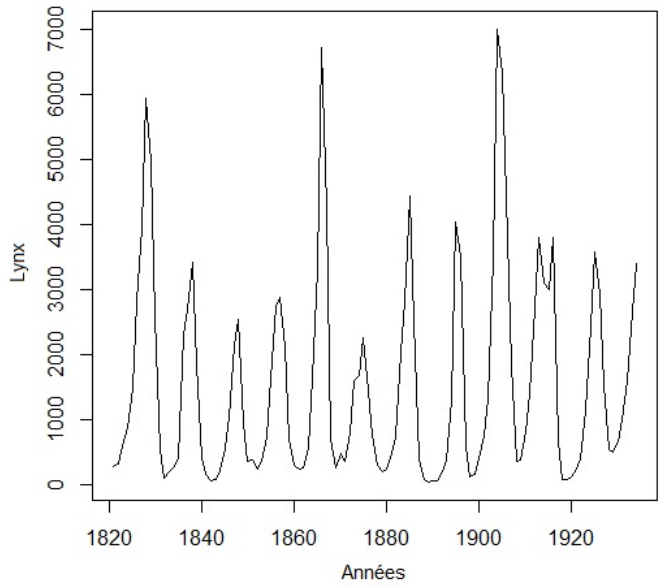

FIGURE 5.3. Nombre de lynx chassés au Canada

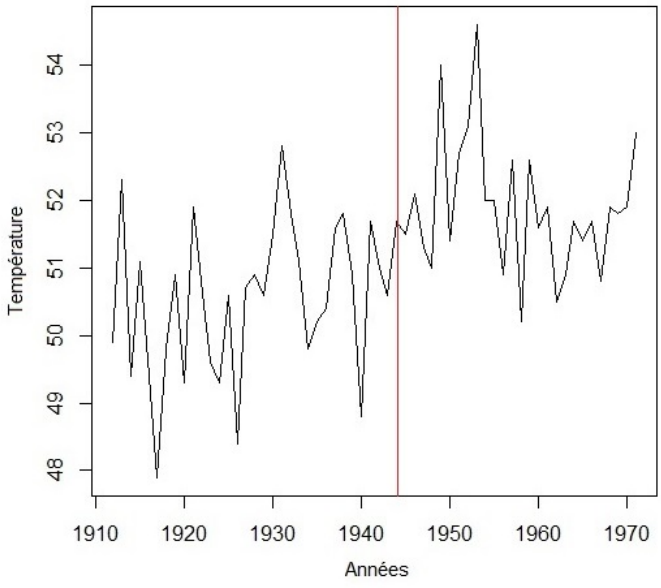

FIGURE 5.4. Température moyenne à New Haven

Nous voulions aussi déterminer si le test TBG est capable de déterminer que la rupture est une tendance lorsque celle-ci est vraiment flagrante. La figure 5.6 représente l'évolution de la population aux États Unis de 1790 à 1970. La tendance dans cet échantillon est flagrante. Les deux méthodes détectent bien sûr qu'il y a une rupture. Le test TBG estime qu'il s'agit d'une tendance, exactement ce que nous voulions montrer.

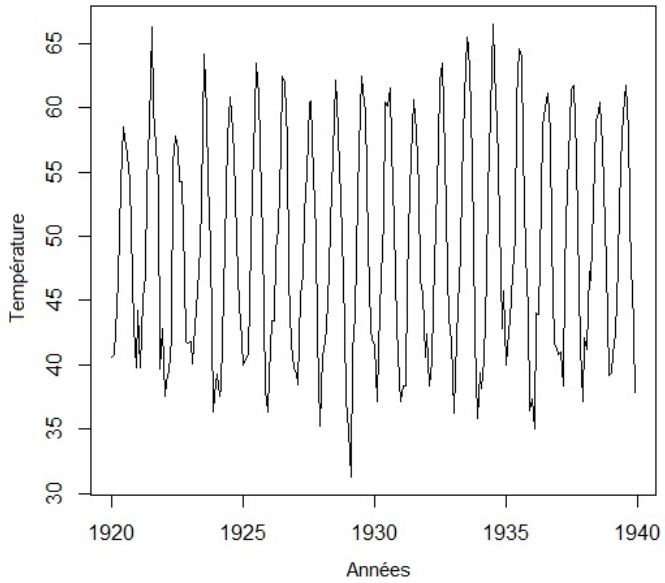

FIGURE 5.5. Température moyenne à Nottingham

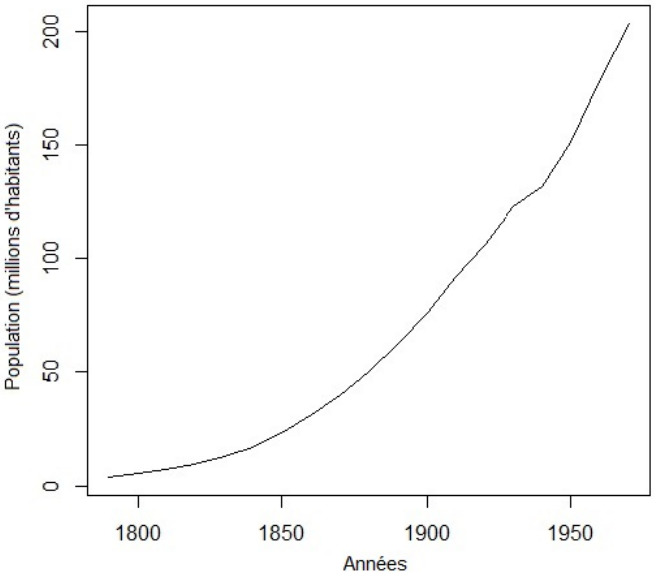

FIGURE 5.6. Population aux États-Unis

La figure 5.7 illustre le nombre conducteurs tués lors d'un accident de la route en Grande-Bretagne, par mois, entre les années 1969 et 1984 inclusivement. Les méthodes estiment qu'il y a une rupture dans cet échantillon. Le test TBG estime qu'il s'agit d'une rupture abrupte, environ en février 1973. En Grande-Bretagne, une nouvelle loi a été instaurée en février 1983 et nous voulons voir si cela a eu un effet sur le nombre de conducteurs tués lors d'un accident de la route. Le

test TBG détecte la rupture qui possède la statistique de test la plus élevée. En restreignant notre échantillon aux dates après février 1973, soit de mars 1973 jusqu'à décembre 1984, le test TBG détecte une rupture, quoi qu'il ne peut pas en déterminer le type, en février 1983, soit le même mois que l'instauration de la nouvelle loi.

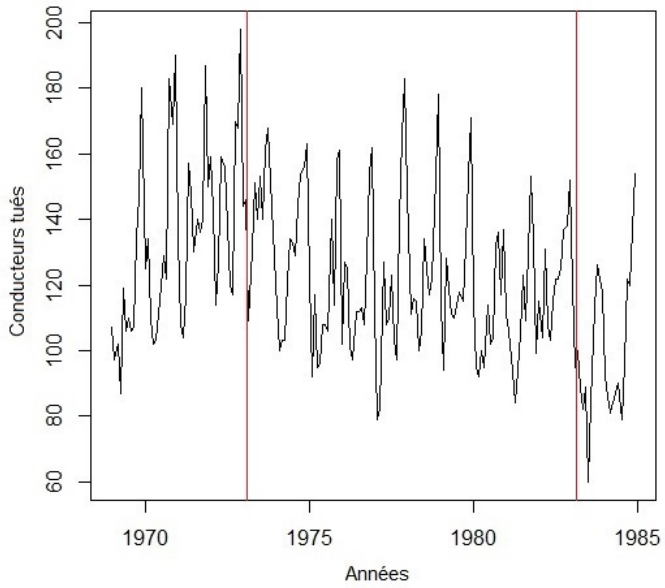

FIGURE 5.7. Conducteurs tués lors d'un accident de la route en Grande-Bretagne

Bref, lors de l'analyse de données réelles, le test TBG est à considérer si on cherche à montrer qu'il existe une rupture dans ces données. On propose tout de même d'utiliser le test de Lombard (1987) afin de voir si la conclusion des deux tests est la même. Si tel est le cas, on pourra avoir plus d'assurance dans notre conclusion. Si la conclusion des deux tests diffèrent, il faut se poser plus de questions et pousser l'analyse des données.

Chapitre 6

CONCLUSION

Pour faire un bref retour sur les chapitres de ce mémoire, nous avons vu, au chapitre 1, une définition sommaire des splines de régression, des splines polynomiaux et des B-splines ainsi que la façon dont elles sont calculées. Dans le chapitre 2, plusieurs méthodes de détection de rupture existant déjà dans la littérature ont été présentées. Le test de Lombard (1987), qui utilise une approche non paramétrique et une fonction de score, est extrêmement efficace pour détecter une rupture. Le test de Jarušková (1997) est un test paramétrique de type «maximum» qui suppose que les variables de l'échantillon suivent une loi normale de moyenne différente avant et après la rupture. Le tests de Reeves *et al.* (2007) utilisent les sommes au carré des résidus et comparent leur statistique de test à une loi de Fischer dont les degrés de liberté changent dépendamment du test utilisé. Il y a cependant une contrainte de continuité dans ces tests. Lund et Reeves (2007) ont modifié le test TPR en cessant d'imposer la contrainte de continuité. Wang (2007) a apporté une petite modification au test de Lund et Reeves (2007) car il trouvait que celui-ci expliquait mal les phénomènes climatiques. Les test de Lund et Reeves (2007) et Wang (2007) ont été utilisés dans ce mémoire mais la convergence vers les valeurs critiques obtenues n'est pas fiable et ces tests ne sont pas robustes à un non-respect du postulat de normalité. Il s'avère qu'ils sont les deux tests les moins efficaces de tous ceux comparés dans ce mémoire.

Dans le chapitre 3, nous avons incorporé les B-splines au test de Reeves *et al.* (2007) afin d'en améliorer l'efficacité et la fiabilité (le test TS). De plus, nous avons créé un test de rupture utilisant la statistique bayésienne (en utilisant la densité *a priori-G*) afin de comparer une méthode bayésienne aux méthodes classiques (le test TBG). Une brève description de la statistique bayésienne et de la densité *a priori-G* y sont aussi mentionnées. Nous avons aussi incorporé les B-splines à la statistique bayésienne afin de créer un test de rupture qui combine les deux nouvelles méthodes (le test $TSBG$).

Dans le chapitre 4, nous avons comparé les sept tests décrits dans les chapitres 2 et 3 à l'aide de plusieurs simulations. Nous avons utilisé diverses combinaisons de type de rupture, d'emplacement de rupture, de taille de variance et de taille d'échantillon afin de déterminer quelles méthodes se démarquaient des autres. Les trois qui ressortaient du lot sont le test de Lombard (2007), le test TBG et le test de Jarušková (1997). Comme un des trois nouveaux tests décrits au chapitre 3 fait partie de ces trois tests, nous voulions savoir jusqu'à quel point il était efficace. Une étude individuelle de ce test à été effectuée (celles des tests de Lombard (2007) et de Jarušková (1997) l'ayant déjà été dans Dubé (2011)). Nous avons vérifié le pourcentage de détection de rupture, c'est-à-dire si le test était capable d'estimer l'emplacement de la rupture, et si le test était aussi capable de déterminer le type de rupture. Car il faut se rappeler que le test TBG teste plusieurs hypothèses qui correspondent chacune à un type de rupture.

Dans le chapitre 5, nous avons testé cette nouvelle méthode sur des données réelles. Nous avons aussi testé la méthode de Lombard (2007) car, selon la littérature, elle serait une des meilleures méthodes de détection de rupture et nous voulions vérifier si le test TBG donnait les mêmes résultats que le test déjà existant.

En conclusion, nous pouvons affirmer que le test TBG est fiable, détecte très bien les ruptures abruptes et l'emplacement de celles-ci. Il est légèrement moins fiable que le test de Lombard (2007) pour détecter les déclenchements de tendance et est un peu plus affecté que celui-ci par un changement de variance, mais est tout de même meilleur que les autres tests définis dans ce mémoire. Les tests avec les splines ne sont pas efficaces afin de détecter une rupture comparativement au test TBG.

Pour finir, il serait intéressant de pousser encore plus loin l'étude sur le test TBG et peut-être le modifier légèrement afin qu'il puisse devenir encore plus fiable que le test de Lombard (2007) pour détecter des ruptures.

BIBLIOGRAPHIE

Bennaghmouch, Zouhair (1992), *Estimation bayésienne d'une fonction avec contraintes*, mémoire de maîtrise, Université de Montréal

Dubé (2011), *Analyse des extrêmes de précipitation simulés par le Modèle régional canadien du climat*, mémoire de maîtrise, Université Laval

Geinitz (2009), Prior Covariance Choices and the g Prior, *Seminar on Bayesian Linear Model*, University Zurich

Jarušková (1997), Some Problems With Application Of Change-point Detection Methods To Environmental Data, Environmetrics, *Vol. 8*, pp. 469-483

Jeffreys (1961), The Theory of Probability, Third Edition, Oxford Classic Texts in the Physical Sciences

Lombard (1987), Rank Tests for Changepoint Problems, Biometrika, *Vol. 74*, No. 3, pp. 615-624

Reeves, Chen, Wang, Lund and Lu (2007), A Review and Comparison of Changepoint Detection Techniques for Climate Data, Journal Of Applied Meteorology And Climatology, *Vol. 46*, pp. 900-915

Robert (2001), The Bayesian Choice, Second Edition, Springer Texts in Statistics

Zellner (1983), Applications of Bayesian analysis in econometrics, The Statistician, *Vol. 32*, pp. 23–34