

STATISTIQUE ET ANALYSE DES DONNÉES

DARIUSH GHORBANZADEH

DOMINIQUE PICARD

Étude asymptotique et pratique du comportement de deux tests de détection de rupture

Statistique et analyse des données, tome 16, n° 3 (1991), p. 63-84.

http://www.numdam.org/item?id=SAD_1991__16_3_63_0

© Association pour la statistique et ses utilisations, 1991, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/legal.php>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

**ETUDE ASYMPTOTIQUE ET PRATIQUE DU
COMPOTEMENT DE
DEUX TESTS DE DETECTION DE RUPTURE**

Dariussh GHORBANZADEH

U.R.A. C.N.R.S. 1321
Université PARIS VII
U.F.R. de Mathématiques
Tour 45-55-5^e étage
2, Place Jussieu
75251 PARIS CEDEX 05

Dominique PICARD

U.R.A. C.N.R.S. 1321
Université PARIS VII
U.F.R. de Mathématiques
Tour 45-55-5^e étage
2, Place Jussieu
75251 PARIS CEDEX 05

Résumé

Nous considérons l'observation des v.a. indépendantes X_1, \dots, X_n qui sont susceptibles de changer de loi après $[n\tau_0]$ premières observations. L'objet est de détecter l'éventuel changement, pour cela nous étudions le comportement de deux tests spécifiques de type non-paramétrique l'un basé sur la comparaison des moyennes avant et après le changement et l'autre basé sur les rangs séquentiels. Sous certaines hypothèses sur l'amplitude de changement, quand le nombre d'observations devient très grand, la loi des statistiques considérées converge vers celle d'un mouvement Brownien avec drift à partir du temps τ_0 , ce qui permet l'évaluation asymptotique des niveaux et puissances.

Nous étudions la performance des deux tests d'abord par l'étude des puissances et puis par des simulations. Nous proposons une application pratique des deux tests à des données concernant les patients HIV-positifs, pour éventuellement détecter une aggravation biologique de leur état.

Le problème de la taille nécessaire pour que les tests soient opérants est étudié aussi, d'abord par des simulations et ensuite par une évaluation directe.

INTRODUCTION.

Ce travail développe deux procédures de détection de changement de loi sous certaines hypothèses sur la loi initiale. Dans la littérature sur la détection de rupture Chernoff & Zacks (1964), Gardner (1969), Hinkley (1970), Sen & Srivastava (1975), Bhattacharya & Brockwell (1976) et Worsley (1986) ont étudié le changement dans la moyenne d'une loi Normale, Kander & Zacks (1966) ont considéré le changement dans le paramètre d'une loi Exponentielle .

Des méthodes non-paramétriques de détection basées sur les rangs ont été proposées par G.K. Bhattacharya & Johnson (1968), Sen & Srivastava (1975) et P.K. Bhattacharya & Fierson (1981). En général, les types de changement étudiés sont du type : Translation, Dilatation et Contamination [par exemple, Csörgö & Horváth (1988)]. Comme dans P.K. Bhattacharya & Fierson (1981), nous allons restreindre notre spectre d'étude à deux procédures l'une linéaire, l'autre basée sur les rangs séquentiels. Nous présentons les expressions des puissances asymptotiques des tests. La performance de deux procédures dans deux cas particuliers est étudiée.

Au §4 nous appliquons les deux tests à des données concernant les patients HIV-positifs, pour détecter une éventuelle aggravation biologique de leur état. L'étude pratique et la comparaison des deux tests pose le problème du nombre de données nécessaire pour que les tests soient opérants, on donnera à ce problème un début de réponse théorique.

1. MODELE.

Soit \mathcal{F} l'ensemble des fonctions de répartition F continues, admettant une densité f par rapport à une mesure dominante μ , et vérifiant:

- i) $\int_{-\infty}^{+\infty} |x|^j f^k(x) d\mu(x) < \infty$ pour $j = 0,1 ; k = 1,2$
- ii) f est deux fois dérivable et on a $\int_{-\infty}^{+\infty} |f^{(k)}(x)| d\mu(x) < \infty$ pour $k = 1,2$ où $f^{(k)}$ est la dérivée k -ième de f .

Soient X_1, X_2, \dots, X_n des variables aléatoires indépendantes de fonction de répartition $F_i(x)$ ($i=1,2, \dots, n$) qui appartiennent à \mathcal{F} admettant la densité $f_i(x)$.

On se propose de tester les hypothèses suivantes :

$$H_0 : F_i(x) = F_1(x) \quad \forall i=1, \dots, n$$

contre $H_1^{(n)} = \bigcup_{\tau_0 \in]0,1[} H_1^{(n)}(\tau_0)$ avec

$$H_1^{(n)}(\tau_0) : \exists \tau_0 \in]0,1[\text{ tel que } \begin{cases} F_i(x) = F_1(x) & \text{si } 1 \leq i \leq [n\tau_0] \\ F_i(x) = F_{2,n}(x) & \text{si } [n\tau_0] + 1 \leq i \leq n \end{cases} ; F_1 \neq F_{2,n}$$

où $F_1, F_{2,n}$ sont supposées connues .

Pour n assez grand $F_{2,n}$ va être considérée comme une perturbation de F_1 (voisine de F_1) pour cela , on suppose que $F_{2,n}$ est absolument continue par rapport à F_1 et que la condition suivante est vérifiée :

$$\frac{dF_{2,n}}{dF_1}(x) = 1 + g_n(x) + \varepsilon_n(x) \tag{1.1}$$

avec i) $\lim_{n \rightarrow \infty} \sqrt{n} \sup_x |g_n(x)| < \infty$ ii) $\lim_{n \rightarrow \infty} \sqrt{n} \sup_x |\varepsilon_n(x)| = 0$

Posons $a_n = \int F_1 \left(\frac{dF_{2,n}}{dF_1} - 1 \right) dF_1$; on a alors le lemme suivant qui introduit la constante δ qui va se révéler très importante dans l'étude qui suit .

Lemme 1.1.

Si F_1 et $F_{2,n}$ vérifient (1.1) alors, $a_n = O\left(\frac{1}{\sqrt{n}}\right)$.

Si de plus, $\lim_{n \rightarrow \infty} \sqrt{n} \int_{-\infty}^{\infty} g_n(x) F_1(x) dF_1(x)$ existe et est égale à δ , alors

$$a_n = \frac{\delta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right) ; \delta \in \mathbf{R} \tag{1.2}$$

Exemples.

1) *Modèle paramétrique régulier* : $\frac{dF_{2,n}(x,\theta)}{dF_1(x)} = h\left(x, \frac{\theta}{\sqrt{n}}\right)$ où on a noté $F_1(x) \equiv F_1(x,0)$; $h(x,0) = 1$ et $h'_\theta(x,0), h''_\theta(x,0)$ vérifient :

$$\sup_x |h'_\theta(x,0)| < \infty ; \sup_x \sup_{|v| \leq \frac{c}{\sqrt{n}}} |h''_\theta(x,v)| \leq \sqrt{n} \eta_n$$

où η_n est une suite qui tend vers 0. On a $\delta = \theta \int_{-\infty}^{\infty} h_{\theta}'(x,0) F_1(x) dF_1(x)$.

2) *Translation* : $F_{2,n}(x) = F_1(x - \frac{\theta}{\sqrt{n}})$ avec $\theta > 0$, alors $\delta = \theta \int_{-\infty}^{\infty} f_1^2(x) dx$.

3) *Dilatation* : $F_{2,n}(x) = F_1((1 - \frac{\theta}{\sqrt{n}})x)$ avec $\theta > 0$, on a $\delta = \theta \int_{-\infty}^{\infty} x f_1^2(x) dx$.

4) *Contamination* : $F_{2,n}(x) = (1 - \frac{\theta}{\sqrt{n}}) F_1(x) + \frac{\theta}{\sqrt{n}} G(x)$ avec $\theta \in]0,1[$

$G \in \mathcal{F}$, $E_G[X] > E_{F_1}[X]$ et $\int_{-\infty}^{\infty} F_1(x) dG(x) > \frac{1}{2}$. On a $\delta = \theta (\int_{-\infty}^{\infty} F_1(x) dG(x) - \frac{1}{2})$

2.1. Test linéaire.

Soient $\mu_1 = E_{F_1}[X]$, $\mu_{2,n} = E_{F_{2,n}}[X]$, $\sigma_1^2 = \text{Var}_{F_1}[X]$ et $\sigma_{2,n}^2 = \text{Var}_{F_{2,n}}[X]$.

Supposons que $F_1, F_{2,n}$ vérifient (1.1) et

$$\lim_{n \rightarrow \infty} \frac{\sqrt{n}(\mu_{2,n} - \mu_1)}{\sigma_1} = \delta^*, \quad \lim_{n \rightarrow \infty} \sigma_{2,n} = \sigma_1 \quad (2.1)$$

Exemples .

1) Modèle paramétrique régulier $\frac{dF_{2,n}(x,\theta)}{dF_1(x)} = h(x, \frac{\theta}{\sqrt{n}})$

$$\delta^* = \frac{\theta}{\sigma_1} \int_{-\infty}^{\infty} x h_{\theta}'(x,0) dF_1(x)$$

2) Si $F_{2,n}(x,\theta) = F_1(x - \frac{\theta}{\sqrt{n}})$ alors, $\delta^* = \frac{\theta}{\sigma_1}$.

3) Si $F_{2,n}(x,\theta) = F_1((1 - \frac{\theta}{\sqrt{n}})x)$ alors, $\delta^* = \frac{\theta}{\sigma_1} \mu_1$

4) Si $F_{2,n}(x,\theta) = (1 - \frac{\theta}{\sqrt{n}}) F_1(x) + \frac{\theta}{\sqrt{n}} G(x)$ alors, $\delta^* = \frac{\theta}{\sigma_1} (\mu_G - \mu_1)$ avec

$\mu_G = E_G[X]$.

Construction du Test Linéaire.

On définit la v.a. $Z_i^* = \frac{X_i - \mu_1}{\sigma_1}$ $i = 1, \dots, n$; et soit

$$\lambda_n^*(t) = \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{[nt]} Z_j^* + (nt - [nt]) Z_{[nt]+1}^* \right) \quad (2.2)$$

En utilisant le Théorème 16.1. de Billingsley.P [5] (page 137) on a la proposition suivante :

Proposition 2.1.

- 1) Sous l'hypothèse H_0 , λ_n^* converge en loi vers W .
- 2) Sous l'hypothèse $H_1^{(n)}(\tau_0)$, λ_n^* converge en loi vers λ^* .

Où la convergence a lieu étroitement sur $\mathcal{C}([0,1])$ muni de la topologie uniforme, $\{W(t), t \in [0,1]\}$ est le processus de Wiener et

$$\lambda^*(t) = W(t) + \delta^*(t-\tau_0) \mathbb{1}_{] \tau_0, 1]}(t) \quad (2.3)$$

Démonstration.

Le 1°) est standard, dans le 2°) on peut remplacer Z_i^* par :

$$Y_i = Z_i^* \text{ si } i \leq [n\tau_0], Y_i = Z_i^* - \mu_{2,n} + \mu_{2,n} = \frac{\sigma_{2,n}}{\sigma_1} \frac{X_i - \mu_{2,n}}{\sigma_{2,n}} + \frac{\mu_{2,n} - \mu_1}{\sigma_1} \text{ si } i > [n\tau_0].$$

Alors, le processus $\frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} Z_i^*$ s'écrit sous la forme :

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} Z_i^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{[n\tau_0]} Y_i + \frac{\sigma_{2,n}}{\sigma_1} \frac{1}{\sqrt{n}} \sum_{j=1+[n\tau_0]}^{[nt]} \frac{X_j - \mu_{2,n}}{\sigma_{2,n}} + \frac{\sqrt{n} (\mu_{2,n} - \mu_1) [nt] - [n\tau_0]}{\sigma_1 n}$$

$$\text{or, } \frac{1}{\sqrt{n}} \sum_{i=1}^{[n\tau_0]} Y_i + \frac{\sigma_{2,n}}{\sigma_1} \frac{1}{\sqrt{n}} \sum_{j=1+[n\tau_0]}^{[nt]} \frac{X_j - \mu_{2,n}}{\sigma_{2,n}} \text{ converge étroitement sur } \mathcal{C}([0,1])$$

vers $W(t)$

$$\text{et, } \frac{\sqrt{n} (\mu_{2,n} - \mu_1) [nt] - [n\tau_0]}{\sigma_1 n} \xrightarrow[n \rightarrow \infty]{} \delta^*(t-\tau_0) \mathbb{1}_{] \tau_0, 1]}(t)$$

Conséquences.

Par le principe de réflexion, pour tout $\eta \geq 0$ sous H_0 :

$$P\left(\sup_{t \in [0,1]} \lambda_n^*(t) \geq \eta\right) \xrightarrow{n \rightarrow \infty} P\left(\sup_{t \in [0,1]} W(t) \geq \eta\right) = 2P(\mathcal{N}(0,1) \geq \eta)$$

On pourra donc, prendre comme région de rejet $\mathcal{R}_{\text{Lin}}^{(n)} = \left\{ \sup_{t \in [0,1]} \lambda_n^*(t) \geq \eta \right\}$.

Pour un $\alpha \in [0,1]$ donné, η_α est déterminé par $P\left(\sup_{t \in [0,1]} W(t) \geq \eta_\alpha\right) \leq \alpha$ où

$\eta_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ et la puissance asymptotique de test est donnée par :

$$\beta_{\text{Lin}}(\tau_0) = P\left(\sup_{t \in [0,1]} \lambda_n^*(t) \geq \eta_\alpha\right) \text{ avec } \beta_{\text{Lin}}(\tau_0) = \lim_{n \rightarrow \infty} P_{H_1^{(n)}}^{(\tau_0)}(\mathcal{R}_{\text{Lin}}^{(n)}).$$

Remarque : Nous nous sommes intéressé à ce test qui correspond au fait que l'on s'attend à une croissance de moyenne. Bien entendu si l'on ne sait rien, il convient de prendre $|\lambda_n^*|$ comme statistique de test et de faire la même théorie.

2.2. Test logarithmique (Test de rang séquentiel).**Définition.**

On définit le rang séquentiel de X_i parmi les X_1, \dots, X_i par :

$$R_i = 1 + \sum_{k=1}^{i-1} \mathbf{1}_{\{X_k - X_i \leq 0\}}$$

On remarque, que sous H_0 les R_i sont de loi uniforme sur $\{1, \dots, i\}$ et les vecteurs aléatoires (R_1, \dots, R_i) et (R_{i+1}, \dots, R_n) sont indépendants.

Construction du Test logarithmique.

Considérons les v.a. Z_i définies par $Z_i = \frac{1}{i} \left(R_i - \frac{i+1}{2}\right)$ et soit $S_k = \sum_{i=1}^k Z_i$.

A partir de $S_k = \sum_{i=1}^k Z_i$, $1 \leq k \leq n$, on considère la suite de processus :

$$\lambda_n(t) = \sqrt{\frac{12}{n}} (S_{[nt]} + (nt - [nt]) Z_{[nt]+1}) \quad (2.4)$$

Par le Lemme 2.2. de Bhattacharya.P.K.[2] on a la proposition suivante :

Proposition 2.2.

- 1) Sous l'hypothèse H_0 , λ_n converge en loi vers W .
- 2) Sous l'hypothèse $H_1^{(n)}(\tau_0)$, λ_n converge en loi vers λ .

Où la convergence a lieu étroitement sur $\mathcal{C}([0,1])$ muni de la topologie uniforme, $\{W(t), t \in [0,1]\}$ est le processus de Wiener et

$$\lambda(t) = W(t) + \sqrt{12} \tau_0 \delta \text{Log} \frac{1}{\tau_0} \mathbb{1}_{] \tau_0, 1]}(t) \quad (2.5)$$

Conséquences.

Par le principe de réflexion, pour tout $\eta \geq 0$ sous H_0 :

$$P\left(\sup_{t \in [0,1]} \lambda_n(t) \geq \eta\right) \xrightarrow{n \rightarrow \infty} P\left(\sup_{t \in [0,1]} W(t) \geq \eta\right) = 2P(\mathcal{N}(0,1) \geq \eta)$$

On pourra donc, prendre comme région de rejet $\mathcal{R}_{\text{Log}}^{(n)} = \left\{ \sup_{t \in [0,1]} \lambda_n(t) \geq \eta \right\}$.

Pour un $\alpha \in [0,1]$ donné, η_α est déterminé par $P\left(\sup_{t \in [0,1]} W(t) \geq \eta_\alpha\right) \leq \alpha$ où

$\eta_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ et la puissance asymptotique de test est donnée par :

$$\beta_{\text{Log}}(\tau_0) = P\left(\sup_{t \in [0,1]} \lambda(t) \geq \eta_\alpha\right) \text{ avec } \beta_{\text{Log}}(\tau_0) = \lim_{n \rightarrow \infty} P_{H_1^{(n)}(\tau_0)}(\mathcal{R}_{\text{Log}}^{(n)}).$$

On remarque donc, que ces deux tests ont été construit de façon à avoir le même niveau asymptotique .

3. ETUDE ASYMPTOTIQUE DES PUISSANCES.

Dans cette partie, à l'aide du "principe de réflexion" [3] et du théorème de Girsanov [24], nous allons évaluer les puissances asymptotiques des deux tests.

Théorème 3.1.

Soit $\{X(t), t \in [0,1]\}$ le processus défini par

$$X(t) = \begin{cases} W(t) & \text{si } t < \tau_0 \\ W(t) + \mu t + \gamma & \text{si } t > \tau_0 \end{cases}$$

où $\tau_0 \in [0,1]$ est fixé et $\{W(t), t \in [0,1]\}$ est le processus de Wiener.

Alors, pour tout $\eta \geq 0$ on a :

$$P\left(\sup_{t \in [0,1]} X(t) \geq \eta\right) = 1 + Q_1(\eta, \mu, \gamma, \tau_0) - Q_1(-\eta, \mu, \gamma, \tau_0)$$

où,

$$\begin{aligned} Q_1(\eta, \mu, \gamma, \tau_0) = & -\Phi\left(\frac{\eta - \gamma - \mu\tau_0}{\sqrt{\tau_0}}\right) + \int_{-\infty}^{\frac{\eta - \gamma - \mu\tau_0}{\sqrt{\tau_0}}} \phi(\zeta) \Phi\left(\sqrt{\frac{\tau_0}{1-\tau_0}} \zeta - \frac{\eta - \gamma - \mu}{\sqrt{1-\tau_0}}\right) d\zeta \\ & + \exp\{2\mu(\eta - \gamma)\} \int_{-\infty}^{\frac{\eta - \gamma + \mu\tau_0}{\sqrt{\tau_0}}} \phi(\zeta) \Phi\left(\sqrt{\frac{\tau_0}{1-\tau_0}} \zeta - \frac{\eta - \gamma + \mu}{\sqrt{1-\tau_0}}\right) d\zeta \end{aligned} \quad (3.1)$$

où ϕ et Φ désignent la densité et la fonction de répartition de $\mathcal{N}(0,1)$.

Démontrons d'abord le lemme suivant :

Lemme 3.1.

Soit $\{W(t), t \in [0,T]\}$ le processus de Wiener, alors la densité de la loi de $M = \sup_{t \in [0,T]} \{W(t) + \mu t + \gamma\}$ est donnée par :

$$\left\{ \frac{2}{\sqrt{2\pi T}} \exp\left\{-\frac{1}{2T}(y - \gamma - \mu T)^2\right\} - 2\mu e^{2\mu(y - \gamma)} \left(1 - \Phi\left(\frac{y - \gamma + \mu T}{\sqrt{T}}\right)\right) \right\} \mathbf{1}_{\{y \geq \gamma\}} \quad (3.2)$$

Démonstration.

Par le "Principe de réflexion", [cf. Bhattacharya Rabi N. [3], page 432], la densité de la loi du couple $(W(T), \sup_{t \in [0, T]} W(t))$ est donnée par :

$$\frac{2}{\sqrt{2\pi}} \frac{2y-x}{T^{3/2}} \exp \left\{ -\frac{1}{2T} (2y-x)^2 \right\} \mathbb{1}_{\{x \leq y, y \geq 0\}}$$

On en déduit la densité de la loi du couple $(W(T), \sup_{t \in [0, T]} W(t) + \gamma)$:

$$\frac{2}{\sqrt{2\pi}} \frac{2(y-\gamma) - x}{T^{3/2}} \exp \left\{ -\frac{1}{2T} (2(y-\gamma) - x)^2 \right\} \mathbb{1}_{\{x \leq y-\gamma, y \geq \gamma\}} \quad (3.3)$$

Soit $\mathcal{F}_T = \sigma\{W(s), 0 \leq s \leq T\}$, par le théorème de Girsanov, [cf. Williams [24]] on a :

$$\frac{dP_\mu}{dP_0} \Big|_{\mathcal{F}_T} = \exp \left\{ \mu \int_0^T dW(s) - \frac{\mu^2}{2} \int_0^T ds \right\}$$

si P_μ est loi du processus de $\mathcal{C}([0, 1]) \quad t \rightarrow W(t) + \mu t$.

On en déduit la densité de la loi du couple $(W(T), M)$:

$$\frac{2}{\sqrt{2\pi}} \frac{2(y-\gamma) - x}{T^{3/2}} \exp \left\{ -\frac{1}{2T} (2(y-\gamma) - x)^2 + \mu x - \frac{\mu^2}{2} T \right\} \mathbb{1}_{\{x \leq y-\gamma, y \geq \gamma\}} \quad (3.4)$$

Démonstration de Théorème 3.1.

Posons $M_1 = \sup_{t \in [0, \tau_0]} W(t)$ et $M_2 = \sup_{t \in [\tau_0, 1]} W(t)$. Pour tout η on a :

$$P\left(\sup_{t \in [0, 1]} X(t) \geq \eta \right) = 1 - P\left(\sup_{t \in [0, 1]} X(t) \leq \eta \right) = 1 - P(M_1 \leq \eta, M_2 \leq \eta)$$

$$= P(M_1 \geq \eta) + P(M_2 \geq \eta) - P(M_1 \geq \eta, M_2 \geq \eta).$$

En appliquant la propriété de Markov forte, conditionnellement à $W(\tau_0)$, $M_2 - W(\tau_0)$ est indépendants de \mathcal{F}_{τ_0} , et on a :

$$\begin{aligned} P(M_1 \geq \eta, M_2 \geq \eta) &= \int P(M_1 \geq \eta, M_2 - W(\tau_0) \geq \eta - z \mid W(\tau_0) = z) dP(z) \\ &= \int P(M_1 \geq \eta \mid W(\tau_0) = z) P(M_2 - W(\tau_0) \geq \eta - z \mid W(\tau_0) = z) dP(z) \end{aligned}$$

D'autre part, conditionnellement à $W(\tau_0)$ on a :

$$\begin{aligned} P(M_2 - W(\tau_0) \geq \eta - z \mid W(\tau_0) = z) &= P\left(\sup_{t \in [\tau_0, 1]} \{W(t) - W(\tau_0) + \mu t + \gamma\} \geq \eta - z \mid W(\tau_0) = z\right) \\ &= P\left(\sup_{t \in [0, 1 - \tau_0]} \{W_1(t) + \mu t + \gamma + \mu \tau_0\} \geq \eta - z\right) \end{aligned}$$

où $W_1(t) = W(t + \tau_0) - W(\tau_0)$ est un processus de Wiener indépendant de W .
On déduit, par des calculs élémentaires, le résultat annoncé.

3.1. Puissance du Test Linéaire.

D'après le Théorème 3.1. on a :

$$P\left(\sup_{t \in [0, 1]} \lambda^*(t) \geq \eta\right) = 2\left(1 - \Phi\left(\frac{\eta}{\sqrt{\tau_0}}\right)\right) + Q_1(\eta, \delta^*, \tau_0) - Q_1(-\eta, \delta^*, \tau_0) \quad (3.5)$$

Avec

$$\begin{aligned} Q_1(\eta, \delta^*, \tau_0) &= \int_{-\infty}^{\frac{\eta}{\sqrt{\tau_0}}} \phi(\zeta) \Phi\left(\sqrt{\frac{\tau_0}{1-\tau_0}} \zeta - \frac{\eta - \delta^*(1-\tau_0)}{\sqrt{1-\tau_0}}\right) d\zeta \\ &+ \exp\{2\delta^{*2} \tau_0 + 2\eta \delta^*\} \int_{-\infty}^{\frac{\eta + 2\delta^* \tau_0}{\sqrt{\tau_0}}} \phi(\zeta) \Phi\left(\sqrt{\frac{\tau_0}{1-\tau_0}} \zeta - \frac{\eta + \delta^*(1+\tau_0)}{\sqrt{1-\tau_0}}\right) d\zeta \end{aligned}$$

3.2. Puissance du Test Logarithmique (Majoration asymptotique).

On ne peut pas comme dans le paragraphe précédent évaluer la puissance asymptotique du test Logarithmique.

Par la concavité du Logarithme on a une majoration du type suivant :

$$\lambda(t) \leq W(t) + \sqrt{12} \tau_0 \delta \left(\frac{t}{\tau_0} - 1 \right) \mathbb{1}_{]t_0, 1]}(t).$$

On obtient donc une majoration asymptotique de la puissance du test logarithmique en changeant δ^* en $\sqrt{12} \delta$ dans la relation (3.5).

Cette majoration est une approximation pour τ_0 au voisinage de 1 .

Appelons :

$$\beta_{\text{Lin}} = 2 \left(1 - \Phi \left(\frac{\eta}{\sqrt{\tau_0}} \right) \right) + Q_1(\eta, \delta^*, \tau_0) - Q_1(-\eta, \delta^*, \tau_0)$$

$$\beta_{\text{Log}} = 2 \left(1 - \Phi \left(\frac{\eta}{\sqrt{\tau_0}} \right) \right) + Q_1(\eta, \sqrt{12} \delta, \tau_0) - Q_1(-\eta, \sqrt{12} \delta, \tau_0)$$

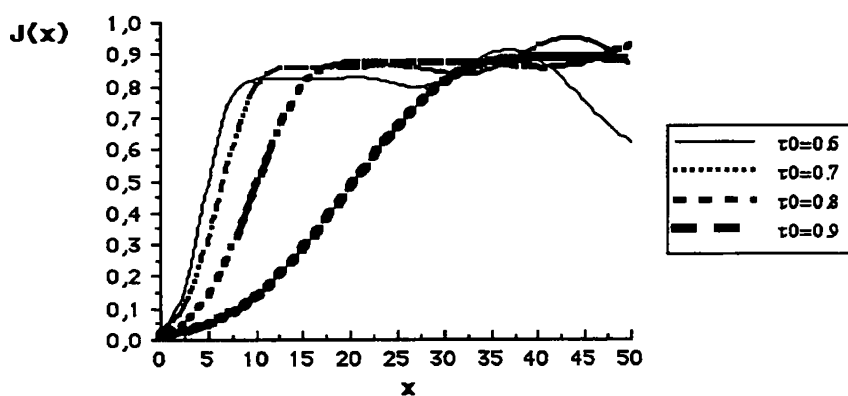
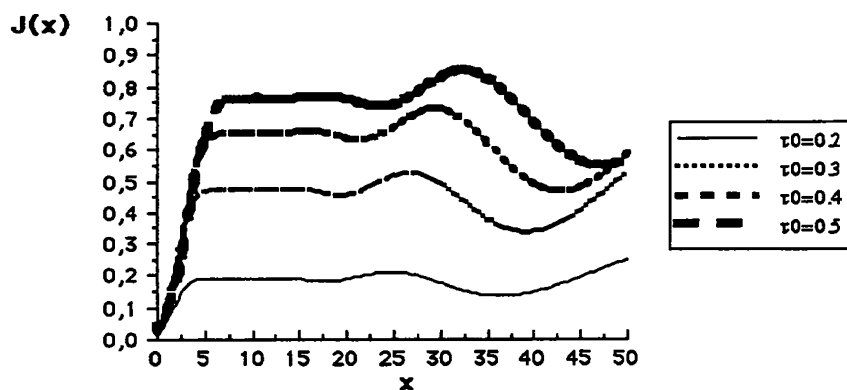
Le paragraphe suivant fait une comparaison de β_{Lin} et β_{Log} .

3.3. Discussion et Exemples.

On a

$$\beta_{\text{Lin}} - \beta_{\text{Log}} = Q_1(\eta, \delta^*, \tau_0) - Q_1(-\eta, \delta^*, \tau_0) - Q_1(\eta, \sqrt{12} \delta, \tau_0) + Q_1(-\eta, \sqrt{12} \delta, \tau_0)$$

Pour $x > 0$, considérons la fonction $J(x) = Q_1(\eta, x, \tau_0) - Q_1(-\eta, x, \tau_0)$ cette fonction est difficile à étudier; faisons en une étude systématique par des simulations. Les graphes suivants présentent $J(x)$ pour les valeurs particulières $\eta = 1.96$ et $\tau_0 = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$



Les graphes montrent que la comparaison de deux tests dépend de la position de τ_0 dans l'intervalle d'observations. On observe toutefois que J semble toujours être croissante pour $x \leq 15$. En tout cas, J est croissante au voisinage de 0. Donc pour δ^* et $\sqrt{12} \delta$ "raisonnablement petits" la comparaison des deux tests revient à la comparaison de δ^* et $\sqrt{12} \delta$.

Observons ce phénomène dans deux exemples particuliers, l'un dans un cadre proche de la loi normale, pour lequel le test Linéaire est optimal, l'autre dans un cadre loin de la loi normale, pour lequel le test Logarithmique est meilleur.

Dans le cas de Translation : $F_{2,n}(x, \theta) = F_1(x - \frac{\theta}{\sqrt{n}})$.

Exemple 1.

Soit F_1 avec la densité $f_1(x) = 2\phi(x) \Phi(ax)$, $a \in \mathbb{R}$.

(ϕ et Φ désignent la densité et la fonction de répartition de $\mathcal{N}(0,1)$).

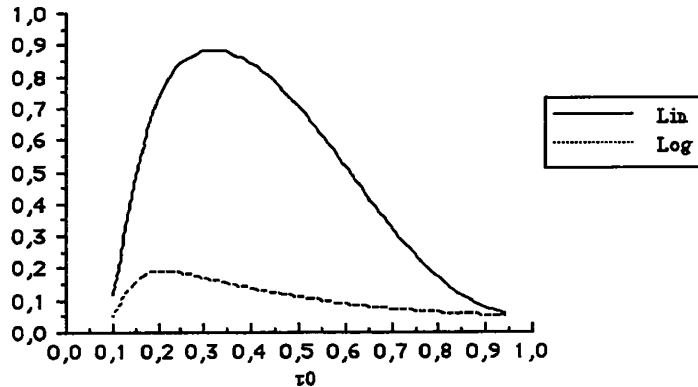
Cet exemple a l'intérêt de présenter une famille qui peut se rapprocher de la loi normale ($a=0$) et dans lequel le test Linéaire est meilleur.

on a :

$$\delta^*(a) = \theta \sqrt{1 - \frac{2a^2}{\pi(1+a^2)}}, \quad \delta(a) = \theta \left[\frac{1}{2\sqrt{\pi}} + \frac{\text{Arctg}(\sqrt{1+a^2}) - \frac{\pi}{4}}{2\pi^{3/2}} \right]$$

Le graphe ci-dessous présente β_{Lin} et β_{Log} en fonction de τ_0 pour les valeurs particulières $a=1$ et $\theta=3$.

Pour les différentes valeurs de a le graphe a toujours la même allure.



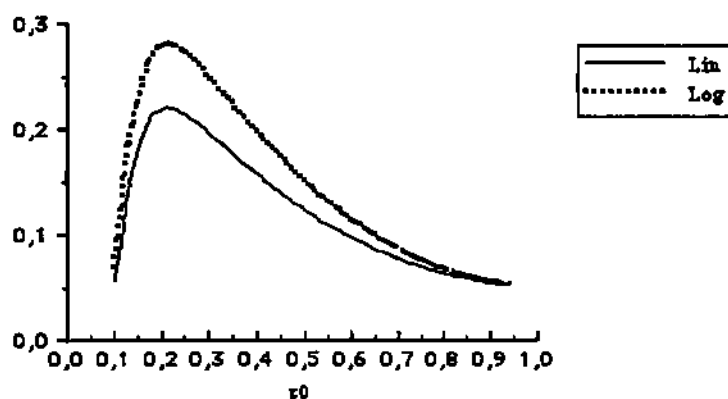
Exemple 2.

$F_1 \equiv \chi^2$ à m d.l. avec la densité $f_1(x) = \frac{1}{2^{m/2} \Gamma(\frac{m}{2})} x^{m/2-1} e^{-x/2}$ on a

$$\delta^*(m) = \frac{\theta}{\sqrt{2m}}, \quad \delta(m) = \frac{\theta \Gamma(m-1)}{2^m \Gamma(\frac{m}{2})^2}$$

- si $m > 30$, $\sqrt{12} \delta \leq \delta^*$
- si $2 \leq m \leq 30$, $\sqrt{12} \delta \leq \delta^*$.

Le graphe ci-dessous présente les puissances de deux tests en fonction de τ_0 pour les valeurs particulières $m=5$ et $\theta=3$.



Dans les deux exemples, on vient de voir que pour toute valeur de θ , la comparaison est basée sur la quantité $\frac{\sqrt{12} \delta}{\delta^*}$, ce qui revient à comparer les dérivées des fonctions :

$$\psi^*(t) = \delta^*(1 - \tau_0) \mathbb{1}_{t \geq \tau_0}, \quad \psi(t) = \sqrt{12} \delta \tau_0 \text{Log} \frac{t}{\tau_0} \mathbb{1}_{t \geq \tau_0}.$$

Il est intéressant de noter qu'on retrouve ici les résultats de comparaison de Bhattacharya & Fierson [2] obtenus en évaluant par simulations les lois des temps d'arrêt : $T = \inf \{t: W(t) \geq \Psi(t)\}$, $T^* = \inf \{t: W(t) \geq \Psi^*(t)\}$.

Ces résultats sont à prendre avec une certaine précaution dans la mesure, où comme il a été remarqué au § 3.2., on n'a qu'une majoration de la puissance pour le test logarithmique. Toutefois cette majoration est très bonne au voisinage de τ_0 , donc ici pour les τ_0 petits.

4. APPLICATION AUX DONNEES BIOLOGIQUES.

Dans cette partie, nous considérons une application des deux tests décrits précédemment à des données concernant les patients HIV-positifs du service de professeur W. ROZENBAUM (Hôpital Rotschild, Paris).

Les données concernent 457 patients qui ont tous plusieurs relevés de 8 variables biologiques à des temps de suivi différents :

- Globules Blancs (GB)

- Lymphocytes (LY)
- Lymphocytes T4 (T4)
- Lymphocytes T8 (T8)
- b2-Microglobulines (b2)
- Immunoglobulines G , A et M (IGG , IGA , IGM) .

Chaque patient a, en moyenne, 5 temps de suivi.

Le but est de déterminer chez un patient le moment le plus judicieux pour lui donner des traitements qui sont parfois difficiles à supporter et qu'il ne faut pas donner trop tôt, mais, pour des raisons d'efficacité, qui ne doivent pas être attribués trop tardivement. non plus.

Il est donc nécessaire d'établir un "critère" permettant de prévoir une aggravation de l'état de santé, qui soit une fonction de relevés de biologiques et dont l'évolution montrerait une rupture lors de l'aggravation, ou mieux précéderait cette aggravation.

Nous allons d'abord faire un travail préliminaire pour extraire des données une variable unidimensionnelle pertinente sur l'évolution clinique de la maladie et que l'on utilisera ensuite pour établir un test de rupture. Vu le petit nombre d'observations par malade, il ne serait pas raisonnable de traiter le problème de rupture de façon multidimensionnelle. A cette fin nous utiliserons le nombre de lymphocytes T4 comme la " variable prédictive "; en effet le nombre de T4 décroît au fur et à mesure de l'évolution de la maladie .

4.1. Analyse Discriminante.

Les coordonnées des 8 variables pour les 457 patients sont classées en quatre groupes selon le relevé du nombre de T4 sur le patient à la date de l'examen. Le classement a été établi à partir des groupes définis par Longini I.M. [18] :

- groupe 1 : $T4 < 200$
- groupe 2 : $200 \leq T4 < 500$
- groupe 3 : $500 \leq T4 < 700$
- groupe 4 : $T4 \geq 700$

L'analyse discriminante est faite par rapport aux quatre groupes .

4.1.1. Pourcentages de classement.

Nous avons examiné les différents pourcentages de bonnes ou de mauvaises classifications. Pour $i, j \in \{1, 2, 3, 4\}$ notons p_{ij} le pourcentage de classification du groupe i vers le groupe j .

Le tableau suivant donne les pourcentages p_{ij} :

Du groupe	Vers le groupe			
	1	2	3	4
1	84.10	13.80	1.11	0
2	19.15	64.13	14.44	2.28
3	0	21.12	59.36	19.52
4	0	3.45	27.16	69.40

En tenant compte du fait que les pourcentages de bonnes classifications sont souvent optimistes, les pourcentages :

$$p_{11} = 85\% , \quad p_{22} = 64\% , \quad p_{33} \approx 60\% \text{ et } p_{44} \approx 70\%$$

montrent que le modèle considéré est assez bien adapté au problème.

4.1.2. Coefficients des variables sur les axes canoniques.

Le tableau suivant résume la description des axes canoniques :

Variable	CAN1	CAN2	CAN3
GB	0.000026	0.000378	0.000451
b2	-0.044366	-0.275021	0.210746
LY	0.004056	-0.001894	-0.000048
T8	-0.004114	0.002459	-0.000958
IGG	-0.010455	0.056537	0.072693
IGA	-0.065777	-0.202473	0.224286
IGM	-0.028101	-0.029856	0.223212

Les résultats ci-dessus montrent qu'il y a une opposition sur les axes canoniques, CAN1 et CAN2, entre les variables LY et T8.

Les variables biologiques étant pondérées par les coefficients de l'axe canonique (CAN1, CAN2 ou CAN3) résultant de l'analyse discriminante, pour un patient j observé à plusieurs temps successifs, si l'on note C_i la valeur du coefficient sur l'axe canonique pour la variable biologique V_i , alors pour le patient j on $X_{pj} = \sum_{i=1}^7 C_i V_{i,p}^j$ où X_{pj} est la valeur de la $p^{\text{ème}}$ coordonnée pour le patient j .

Par exemple, la variable considérée à partir des coefficients de l'axe canonique "CAN1" est donnée par :

$$0.000026 * GB - 0.044366 * b2 + 0.004056 * LY - 0.004114 * T8 \\ - 0.010455 * IGG - 0.065777 * IGA + 0.028101 * IGM$$

Dans notre fichier, nous avons sélectionné, 251 patients qui ont au moins 6 temps de suivi. Il y a 195 patients qui ont changé de groupes et 56 patients n'ont pas changé de groupes .

Parmi les 56 patients qui n'ont pas changé de groupe, 29 appartiennent au groupe 1 (nombre de T4 < 200) ce qui signifie qu'ils ont déjà subi une ou plusieurs ruptures avant d'entrer dans le fichier; d'autre part il y a 26 patients qui appartiennent au groupe 2 ($200 \leq T4 < 500$) chez ces patients il faudrait attendre à une baisse éventuelle de nombre de T4 (inconnue pour notre fichier) .

4.3. Résultats sous H_0 .

Nous avons appliqué les deux tests (Linéaire , Logarithmique) à la variable construite à partir des coefficient de l'axe canonique CAN1 .

Le tableau suivant présente les résultats sous H_0 obtenus pour les deux tests:

	Proportion de bonnes détections		Proportion de mauvaises détections	
α %	Test linéaire	Test logarithmique	Test linéaire	Test logarithmique

1	34/195 [12% , 23%]*	0/195	0/56	0/56
2	47/195 [18% , 30%]*	1/195	1/56	0/56
5	98/195 [43% , 57%]*	2/195	7/56	0/56
10	125/195 [57% , 71%]*	4/195	16/56	2/56

*: Intervalle de confiance à 5% pour la proportion

Le tableau ci-dessus montre que sur nos données le test Linéaire est plus opérant que le test Logarithmique; d'autre part, pour un niveau de 10%, la puissance du test Linéaire est comprise entre 0.57 et 0.71 .

4.4. Discussion.

On peut expliquer les performances moyennes des deux tests par le faible nombre de données par patient (Minimum 6, Maximum 18), comme nos procédures sont asymptotiques, il est normal qu'en présence d'un nombre faible de données les évaluations des niveaux et puissances en soient largement affectées. Evaluer un niveau exact normé par simulation est difficile, néanmoins, nous allons essayer de donner une idée de la taille nécessaire pour une détection de rupture par les deux tests.

Introduisons les temps d'arrêt suivants :

$$T^* = \inf\{t : \lambda^*(t) \geq \eta_\alpha\}, \quad T = \inf\{t : \lambda(t) \geq \eta_\alpha\}$$

où $\eta_\alpha = \Phi^{-1}(1-\alpha/2)$ et λ^*, λ sont définis précédemment .

Nous allons procéder par deux méthodes :

1°) Simulations 2°) Evaluation directe de $E[T^*]$ et $E[T]$.

4.4.1. Simulations.

Dans le cas du modèle de translation, avec $F_1 \equiv \mathcal{N}(0,1)$, nous avons simulé 100-échantillons de tailles n ($n=6,18,50,1000$) de loi $\mathcal{N}(0,1)$ avec $\theta = \sqrt{n}/2$, $\tau_0 = 0.33$ et $\alpha = 0.05$. Si l'on note :

p^* = proportion d'échantillons pour lesquels $T^* \leq 1$

p = proportion d'échantillons pour lesquels $T \leq 1$

nous avons obtenu :

n	p^*	$E[T^*]$	p	$E[T]$
6	0.17	0.882	0.08	0.804
18	0.29	0.791	0.17	0.805
50	0.74	0.784	0.23	0.715
1000	0.92	0.383	0.86	0.357

Le tableau ci-dessus montre que pour les échantillons de tailles $n=6,18$ les deux temps d'arrêt sont concentrés à la fin de l'intervalle d'observations et d'autre part, plus la taille augmente plus la moyenne des temps d'arrêt se concentrent au tour de τ_0 .

4.4.2. Evaluation de $E[T^*]$ et $E[T]$.

Soit $\tau = \inf \{ t : W(t) \geq \gamma + \mu t \}$ avec $\mu < 0$. Par un résultat classique [cf Billingsley [6], Lemme.2. page 456] on a :

$\forall \lambda \leq 0$, $\exp\{\lambda W(t) - \frac{\lambda^2}{2} t\}$ est une martingale bornée, on a donc,

$$E[\exp\{\lambda W(t) - \frac{\lambda^2}{2} t\}] = 1 \text{ et donc, } E[\exp\{\lambda(\gamma + \mu\tau) - \frac{\lambda^2}{2} \tau\}] = 1 .$$

En différentiant par rapport à λ on obtient :

$$E[\tau(\mu - \lambda) \exp\{(\lambda\mu - \frac{\lambda^2}{2}) \tau\}] = -\gamma e^{-\gamma\lambda}$$

donc, en particulier pour $\lambda = 0$, on a $E[\tau] = -\frac{\gamma}{\mu}$.

On en déduit, $E[T^*] = \tau_0 + \frac{\eta_\alpha}{\delta^*}$ et $E[T] = \tau_0 + \frac{\eta_\alpha}{\sqrt{12} \delta}$.

Dans le cas du modèle de translation, avec $F_1 \equiv \mathcal{N}(0,1)$, avec $\theta/\sqrt{n}=0.5$ on a :

$$E[T^*] = \tau_0 + \frac{2\eta_\alpha}{\sqrt{n}}, \quad E[T] = \tau_0 + \frac{2\sqrt{\pi}\eta_\alpha}{\sqrt{3n}}$$

d'où $E[T^*] \leq 1$ et $E[T] \leq 1$ impliquent :

$$n_{\text{Lin}} \geq \left(\frac{2\eta_\alpha}{1 - \tau_0} \right)^2, \quad n_{\text{Log}} \geq \left(\frac{2\sqrt{\pi}\eta_\alpha}{\sqrt{3}(1 - \tau_0)} \right)^2$$

En particulier pour $\tau_0=0.33$ et $\eta_\alpha=1.96$ on a : $n_{\text{Lin}} \geq 35$ et $n_{\text{Log}} \geq 36$.

Dans ce cas on voit donc que le nombre de données par patient est toujours trop faible mais, les deux tests sont pénalisés de la même façon.

APPENDICE (SIMULATION DU PROCESSUS DE WIENER).

Soient X_1 et X_2 deux v.a. de loi uniforme sur $[0,1]$, alors

$$Y_1 = \sqrt{-2 \text{Log } X_1} \cos 2\pi X_2$$

$$Y_2 = \sqrt{-2 \text{Log } X_1} \sin 2\pi X_2$$

sont indépendantes et de loi $\mathcal{N}(0,1)$.

Soit $\{W(t), t \in [0,1]\}$ le processus de Wiener alors par la construction de Donsker [cf.

Billingsley [5]] on a $\frac{1}{\sqrt{n}} \sum_{j=1}^{[nt]} Y_j$ converge étroitement sur $\mathcal{C}([0,1])$ vers W ; où

Y_1, \dots, Y_n sont des variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$.

BIBLIOGRAPHIE

- [1] Barndorff-Nielsen, O. (1963). On the limit behavior of extreme order statistics. Ann of Stat. 34 pp 992-1002.

- [2] **Bhattacharya, P.K & Fierson.D, JR** (1981). A non parametric control chart for detecting disorders. *Ann of Stat.* **9**. pp 544-554
- [3] **Bhattacharya, R.N. & Waymire, E.C.** (1990) *Stochastic Processes with Applications* .Wiley .
- [4] **Bhattacharya, G.K. & Johnson, R.A.**(1968). Non parametric tests for shifts at an unknown time point. *Ann. Math.Stat.* **39** pp 1731-1743
- [5] **Billingsley, P** (1968). *Convergence of Probability Measures*. Wiley, New York.
- [6] **Billingsley, P** (1979). *Probability and Measure*. Wiley, New York.
- [7] **Chernoff, H. & Zacks, S** (1964). Estimating the current mean of a normal distribution which is subjected to change in time. *Ann. Math. Stat.* **35** pp 999-1018.
- [9] **Cobb, G.W.** (1978). The problem of the Nile: Conditional solution to a changepoint problem. *Biometrika*, **65**, 2, pp. 243-251.
- [10] **Csörgö, M. & Horvath, L** (1988). Nonparametric Methods for changepoint problems. *Handbook of Statistic*, Vol.7. pp 403-425.
- [11] **Deshayes, J. & Picard, D.** (1986) *Off line Statistical analysis of changepoint models using non parametric and likelihood methods* . *Lecture Notes on Information and Control*. Springer verlag.
- [12] **Gardner, L.A. Jr.** (1969). On detecting changes in the mean of normal variates. *Ann.Math.Stat.* **40**. pp 114-115.
- [13] **Ghorbanzadeh, D.** (1991). Ruptures dans les modèles non paramétriques *Applications et Tests. Séminaire 90-91. PARIS V*. pp 195-210.
- [14] **Hajek, J.** (1968) Asymptotic normality of simple linear rank statistics under alternatives. *Ann Math Stat.* **39**. pp 325-346 .

- [15] Hajek, J & Sidak, Z. (1967). Theory of Rank. Academic, New York .
- [16] Hinkley, D.V. & Hinkley, E.A. (1970) Inference about the change point in a sequence of binomial variables. *Biometrika*, 57, 2, pp.477-488.
- [17] Kander, A. & Zacks, S. (1966). Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points. *Ann Math Stat.* 37. pp 1196-1210 .
- [18] Longini, I.M. (1990). Modeling the Decline of CD4⁺ T-Lymphocyte Counts in HIV-Infected Individuals. *Journal of Acquired Immune Deficiency Syndromes*, Vol.3, N° 9.
- [19] Marion, R. & Reynolds, JR. (1975). A sequential Signed Rank Test for symmetry. *Ann of Stat* 2. pp 382-400.
- [20] Picard, D. (1985). Testing and estimating change-point in time series. *Advances in Appl. Prob.* 17. pp 841-867 .
- [21] Roussas, G. (1972). Contiguity of Probability measure. Combrige University press.
- [22] Sen, A. & Srivastava, M.S. (1975). On tests for detecting change in mean. *Ann of Stat.* 3 pp 90-108.
- [23] Shorac, G.R. & Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics* . Wiley, New York.
- [24] Williams, D. (1979). *Diffusions, Markov processes and Martingales*. Vol.1, Wiley, New York .
- [25] Worsley, K.J. (1986). Confidence regions and tests for a change point in a sequence of exponential family variables. *Biometrika*, 73, pp.91-104.
- [26] Zacks, S. (1982). Fixed sample and sequential procedures. *Statistique et Analyse des Données*. pp.48-81.