Université de Montréal

Différents procédés statistiques pour détecter la non-stationnarité dans les séries de précipitation

par

Kevin Charette

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

avril 2014

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé

Différents procédés statistiques pour détecter la non-stationnarité dans les séries de précipitation

présenté par

Kevin Charette

a été évalué par un jury composé des personnes suivantes :

Martin Bilodeau
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Alejandro Murua
(membre du jury)

Mémoire accepté le:
Date d'acceptation

SOMMAIRE

Ce mémoire a pour objectif de déterminer si les précipitations convectives estivales simulées par le modèle régional canadien du climat (MRCC) sont stationnaires ou non à travers le temps. Pour répondre à cette question, nous proposons une méthodologie statistique de type fréquentiste et une de type bayésien. Pour l'approche fréquentiste, nous avons utilisé le contrôle de qualité standard ainsi que le CUSUM afin de déterminer si la moyenne a augmenté à travers les années. Pour l'approche bayésienne, nous avons comparé la distribution *a posteriori* des précipitations dans le temps. Pour ce faire, nous avons modélisé la densité *a posteriori* d'une période donnée et nous l'avons comparée à la densité *a posteriori* d'une autre période plus éloignée dans le temps. Pour faire la comparaison, nous avons utilisé une statistique basée sur la distance d'Hellinger, la J-divergence ainsi que la norme L_2. Au cours de ce mémoire, nous avons utilisé l'ARL (longueur moyenne de la séquence) pour calibrer et pour comparer chacun de nos outils. Une grande partie de ce mémoire sera donc dédiée à l'étude de l'ARL. Une fois nos outils bien calibrés, nous avons utilisé les simulations pour les comparer. Finalement, nous avons analysé les données du MRCC pour déterminer si elles sont stationnaires ou non.

Mots clés : Non-stationnarité, bayésien, contrôle de qualité, comparaison densité *a posteriori*, CUSUM, ARL, distance d'Hellinger, J-divergence.

SUMMARY

The main goal of this master's thesis is to find whether the summer convective precipitations simulated by the Canadian Regional Climate Model (CRCM) are stationary over time or not. In order to answer that question, we propose both a frequentist and Bayesian statistical methodology. For the frequentist approach, we used standard quality control and the CUSUM to determine if the mean has increased over the years. For the Bayesian approach, we compared the posterior distributions of the precipitations over time. In order to do the comparison, we used a statistic based on the Hellinger's distance, the J-divergence and the L_2 norm. In this master's thesis, we used the ARL (average run length) to calibrate each of our methods. Therefore, a big part of this thesis is about studying the actual property of the ARL. Once our tools are well calibrated, we used the simulation to compare them together. Finally, we studied the data from the CRCM to decide, whether or not, the data are stationary.

Keywords : Stationarity, bayesian, quality control, posterior distribution comparaison, CUSUM, ARL, Hellinger's distance, J-divergence.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

REMERCIEMENTS

Tout d'abord, j'aimerais remercier le Professeur Jean-François Angers, mon directeur de recherche, qui m'a soutenu tout au long de ma maîtrise. Malgré un horaire plus que chargé, il a toujours su trouver amplement de temps pour m'aider et pour répondre à mes multiples questions. Aussi, maintenant que mes études supérieures se terminent, il continue d'être présent pour moi en m'aidant dans ma recherche d'emplois, ce qui est vraiment apprécié.

Ensuite, j'aimerais souligner l'apport de quelques professeurs qui m'ont particulièrement aidé ou influencé lors de mon cursus universitaire. Premièrement, je pense au Professeur David Haziza, sans qui je ne serais probablement pas en statistique aujourd'hui. De plus, un remerciement spécial au Professeur Christian Léger qui m'a dévoué plusieurs heures afin de répondre à mes nombreuses questions. Finalement, étant un spécialiste de la consultation statistique, Miguel Chagnon a su mettre son expérience à profit en me donnant l'opportunité de travailler avec lui.

De plus, je souhaite remercier le personnel administratif du DMS. Plus particulièrement, j'aimerais souligner la gentillesse et la patience de Émilie Dubois, Églantine Hontanx, Anne-Marie Dupuis ainsi que de Julie Colette.

J'aimerais également remercier mon père et ma mère qui m'ont toujours poussé à me dépasser dans tout ce que j'entreprenais. Sans leur soutien, je n'aurais probablement pas atteint les études supérieures et je leur en serai pour toujours reconnaissant.

Finalement, j'aimerais remercier ma copine Myriam pour tout ce qu'elle a fait pour moi durant mes études. Elle a su m'aider, me motiver ainsi que faire en sorte que tout soit plus simple au quotidien.

Chapitre 1

INTRODUCTION

Plusieurs études récentes ont démontré que nous sommes présentement en pleine phase de changement climatique. Par exemple, il est connu et accepté par la communauté scientifique que nous sommes en train de subir un réchauffement à l'échelle planétaire (voir GIEC, 2007). Plus précisément, des études stipulent que le modèle régissant la température au Québec aurait commencé à changer vers les années 1970 (voir Chaumont *et al.*, 2007). Il est alors justifié de se demander si ces changements climatiques vont entrainer un changement dans les précipitations. Le but de ce mémoire sera donc d'élaborer des méthodes qui peuvent détecter, s'il y a lieu, le changement dans les précipitations. Un changement à travers le temps dans les précipitations indiquerait que le modèle est non stationnaire. Afin de déterminer si les précipitations sont stationnaires ou non, nous utiliserons une approche basée sur la statistique classique et une autre sur la statistique bayésienne.

L'approche par statistique classique traitera la question à l'aide du contrôle de qualité. Le contrôle de qualité sert, principalement, à détecter une variation dans la moyenne ou dans l'écart-type (voir Montgomery, 2007). En général, le contrôle de qualité est substantiellement utilisé dans les usines avec chaînes de montage, mais il est possible de l'utiliser dans plusieurs autres domaines tels que le service à la clientèle ainsi que dans la finance (voir Montgomery, 2007). Plus particulièrement, le fait que nous souhaitons détecter une augmentation de la quantité des précipitations fait de cet outil un bon choix. Pour le contrôle de qualité, nous allons présenter deux outils : les cartes de contrôle standard ainsi que les cartes de sommations cumulatives notées CUSUM.

Pour l'approche bayésienne, nous allons comparer la distribution *a posteriori* des précipitations dans le temps. Pour ce faire, nous allons modéliser la densité *a posteriori* d'une période donnée et nous la comparerons à la densité *a*

posteriori d'une autre période plus éloignée dans le temps. Pour faire la comparaison, nous utiliserons le concept de distance entre deux densités *a posteriori* (voir Ghosh *et al.*, 2007).

Dans le cadre de ce mémoire, nous utiliserons les données de précipitations simulées pour les années 1961 à 2100 par le modèle régional canadien du climat (MRCC) établi par Ouranos. De plus, seules les précipitations convectives estivales sur la région de Montréal seront considérées. D'ailleurs, nous aimerions préciser que la période estivale est définie par la période allant du 1er juin au 31 août inclusivement. Finalement, notons que les données dont nous disposons ont été obtenues uniquement par simulation, mais que nous allons tout de même nous permettre de les appeler par le mot observation au cours de ce mémoire.

Dans ce mémoire, nous débuterons par introduire l'approche classique au chapitre 2. Durant ce chapitre, nous commencerons par faire une brève mise en contexte sur le contrôle de qualité. Ensuite, avant d'élaborer plus sur certains aspects du contrôle de qualité, nous étudierons les données Ouranos dans le but de pouvoir établir des hypothèses plausibles sur les observations. Une fois les hypothèses nécessaires établies, nous approfondirons les notions de base sur le contrôle de qualité tout en présentant un outil qui nous permettra de comparer l'efficacité de nos méthodes. Cet outil sera appelé ARL (longueur moyenne de la séquence) et sera défini plus tard. L'ARL nous permettra , entre autres, de bien calibrer les paramètres du CUSUM pour en maximiser l'efficacité.

Par la suite, au chapitre 3, nous passerons à l'approche bayésienne. Premièrement, nous commencerons par faire une brève mise en contexte de la statistique bayésienne qui motivera l'idée que nous utiliserons ensuite, c'est-à-dire, comparer deux densités *a posteriori* dans le but de détecter un changement dans les précipitations. Deuxièmement, nous allons présenter quelques hypothèses nécessaires à l'application de la statistique bayésienne. Une fois les hypothèses bien établies, nous présenterons les différentes distances avec lesquelles nous allons travailler. Subséquemment, nous nous rendrons compte que toutes les distances sont fonctions d'une statistique que nous noterons W. Ceci nous poussera à étudier cette statistique plus profondément et nous constaterons que nous pouvons trouver la loi de W. Ceci fera en sorte que nos tests ne seront plus basés sur les distances, mais bien sur la statistique W. Enfin, le reste de ce chapitre sera consacré à la mise en place de notre test ainsi qu'à l'étude de l'ARL pour l'approche bayésienne. Tout comme pour l'approche classique, l'ARL servira à calibrer les paramètres de la méthode bayésienne.

Finalement, au chapitre 4, une fois nos outils bien calibrés, nous comparerons les trois outils sur la base de l'ARL en procédant à des simulations. De plus, nous appliquerons ces outils sur le jeu de données simulées Ouranos dans le but de savoir si le modèle régissant les précipitations est stationnaire ou non.

Chapitre 2

APPROCHE CLASSIQUE

Dans ce chapitre, nous discuterons les méthodes que nous utiliserons afin de détecter la non-stationnarité dans les observations dans le cadre de la statistique classique. Ces méthodes seront basées sur le contrôle de qualité. Nous commencerons par une mise en contexte globale du contrôle de qualité afin d'éclairer le lecteur sur l'idée derrière ce processus. Par la suite, une description des données Ouranos sera faite dans le but de justifier le choix de certaines hypothèses de modélisation. Finalement, nous présenterons les deux outils principaux que nous utiliserons pour l'analyse des données.

2.1. MISE EN CONTEXTE

Notre analyse statistique classique repose sur le contrôle de qualité à l'aide de cartes de contrôle. Le contrôle de qualité effectué à l'aide d'une carte est un procédé qui a vu le jour dans les années 1920 et qui a pour objectif de déceler si une chaîne de production industrielle est hors de contrôle. L'inventeur des cartes de contrôle travaillait chez Bell et se nomme Walter A. Shewhart (voir Shewart, 1931). C'est pourquoi, les cartes de contrôle sont souvent appelées *Cartes de contrôle de Shewhart*. L'idée de ce procédé est très simple, il suffit de faire le graphique des statistiques basées sur les données au fur et à mesure que nous les obtenons. L'avantage de ce procédé est qu'il nous donne une idée précise de la variabilité du procédé. Il est connu que chaque processus admet une variabilité intrinsèque et que cette variabilité est aléatoire. Le but des cartes de contrôle est de détecter si une variabilité extrinsèque s'est ajoutée à la variabilité intrinsèque. L'idée est qu'une plus grande variabilité entraîne une augmentation du nombre d'observations *atypiques* et c'est sur cette idée que se basent les cartes de contrôle. En effet, des limites seront fixées et dès qu'une entité dépassera une limite, le système sera considéré comme étant hors de contrôle. Nous allons donc appliquer ce principe aux précipitations pour voir

si elles ont changé en moyenne ou en variance à travers le temps. De plus, un avantage non négligeable du contrôle de qualité est qu'il permet d'avoir une bonne idée du moment à partir duquel le processus générant les précipitations s'est mis à changer.

2.1.1. Présentation des cartes de contrôle

Dans ce chapitre, nous ferons usage de deux types de carte différents. Tout d'abord, nous utiliserons les cartes de contrôle de qualité standard. Cette approche est la plus simple, mais entraîne des désavantages dont nous discuterons ultérieurement. Par la suite, nous utiliserons les cartes de contrôle à sommation cumulative notées CUSUM qui seront beaucoup mieux adaptées à notre problématique. Notons que ces deux outils testent seulement si la vraie moyenne a subi un changement. Finalement, précisons que seule la détection d'une hausse dans la moyenne des précipitations nous intéresse. De ce fait, tous les tests ainsi que les intervalles de confiance présentés seront unilatéraux.

2.1.2. Mesure de performance

Tant dans le cas classique que bayésien, nous allons devoir calibrer les paramètres utilisés dans nos méthodes pour pouvoir en maximiser leur performance. Le critère que nous utiliserons pour mesurer la performance de nos méthodes sera l'ARL : longueur moyenne de la séquence. Cette expression est une traduction de l'acronyme anglophone ARL « *average run length* » et est définie comme le nombre d'essais moyen avant de détecter le changement dans les observations. En tant qu'indice de performance, notons que l'ARL est l'outil principalement utilisé dans le contrôle de qualité (voir Derman et Ross, 1997). Nous pouvons distinguer deux types d'ARL :

(1) l'ARL lorsque le système est en contrôle : nombre de coups moyen avant d'obtenir une fausse alarme. Nous pouvons ici faire une analogie avec l'erreur de premier type ;

(2) l'ARL lorsque le système est hors de contrôle : nombre de coups moyen avant de détecter que le système est hors de contrôle. Nous pouvons ici faire une analogie avec l'erreur de deuxième type.

Il est clair que l'on voudra minimiser l'ARL quand le système est hors de contrôle, mais qu'on voudra également minimiser la fréquence des fausses alarmes. Évidemment, tout comme la puissance, l'ARL dépend de certains paramètres du modèle. On écrira donc $\mathrm{ARL}(\cdot)$, où le point représente les différents paramètres dont l'ARL dépend.

2.1.3. Hypothèses

Afin d'émettre des hypothèses qui sont adaptées aux données d'Ouranos, nous allons faire une analyse descriptive de ces données. Il a été proposé dans Cam et Neyman (1967), que les quantités de précipitation peuvent être ajustées par la loi gamma si on enlevait les journées sans pluie. Nous allons maintenant vérifier la pertinence de cette hypothèse. Pour ce faire, nous avons fait l'histogramme de la quantité quotidienne de pluie pour les dix premières années d'observations (1961-1970). Nous avons sélectionné les dix premières années seulement, car des études suggèrent que les changements climatiques ont commencé vers le début des années 1970 (voir Chaumont *et al.*, 2007). À la figure 2.1, nous pouvons voir que les données semblent suivre une loi gamma. À cet effet, nous voyons que la densité gamma est très bien ajustée à l'histogramme. De plus, en vue de *vérifier* graphiquement cette hypo-

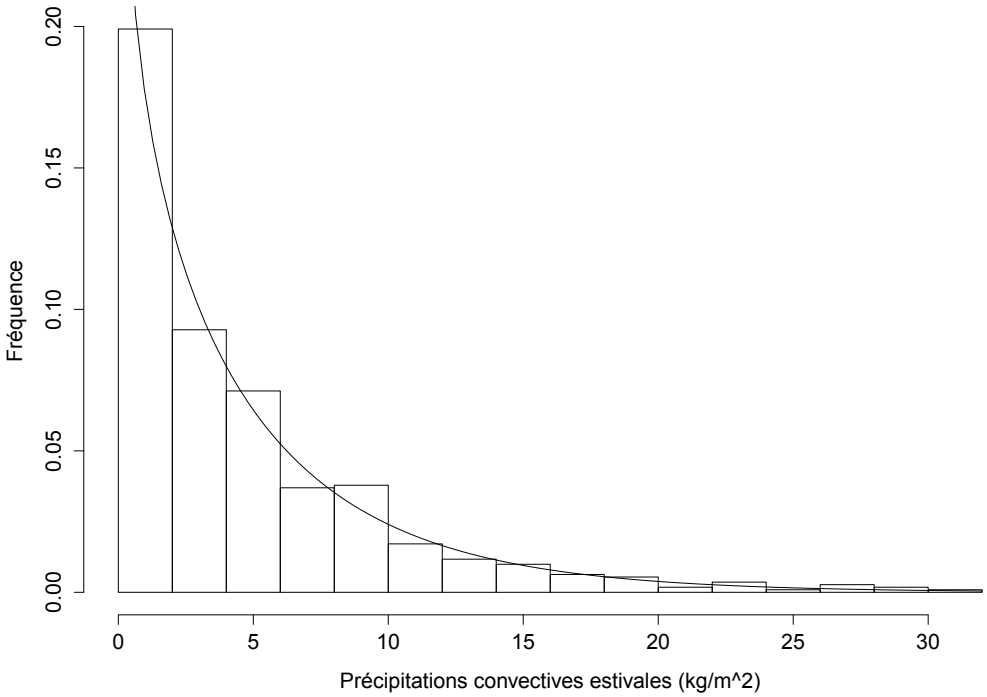

FIGURE 2.1. Histogramme des précipitations convectives estivales.

thèse, regardons le diagramme quantile-quantile à la figure 2.2. Nous pouvons constater que l'hypothèse de la loi gamma est acceptable. Soit X_{ij} les précipitations quotidiennes convectives estivales de la j^e journée de la i^e année, nous supposerons que X_{ij} suit la loi gamma de paramètre α et θ, ce qui sera noté $X_{ij}|(\alpha, \theta) \sim \text{Gamma}(\alpha, \theta)$.

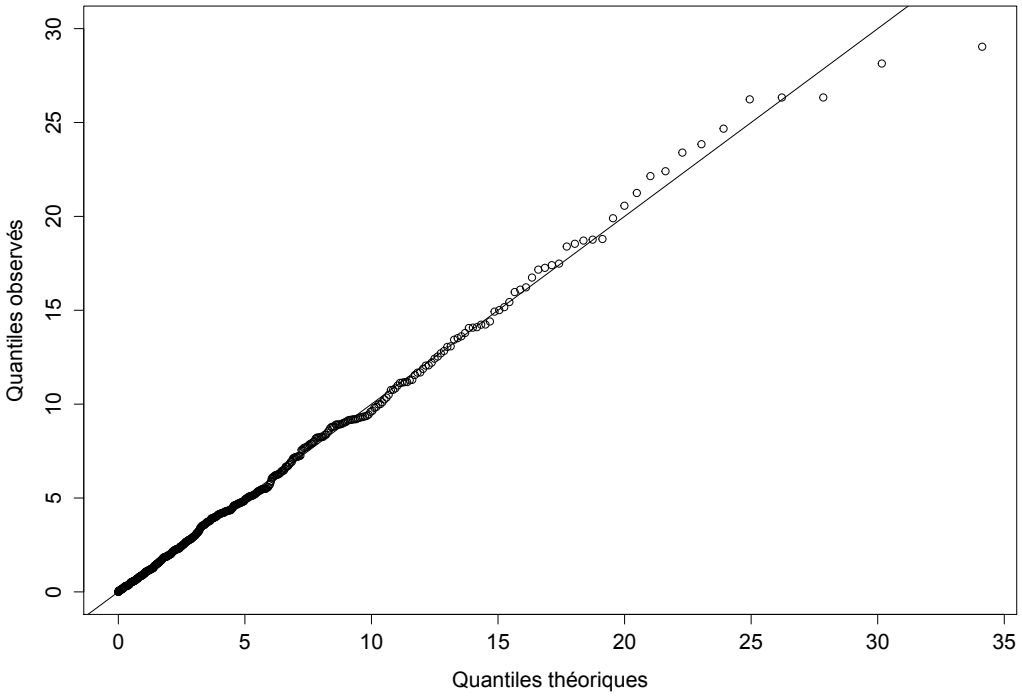

FIGURE 2.2. Diagramme quantile-quantile des données Ouranos pour la loi gamma.

Définition 2.1.1. *Si* X *est de loi gamma de paramètre* α *et* θ*, alors* X *a la densité suivante :*

$$f(x|\theta) = \begin{cases} \frac{\theta^\alpha}{\Gamma(\alpha)} \times x^{\alpha-1} \exp\{-\theta x\} & \textit{si } x > 0, \\ 0 & \textit{sinon.} \end{cases}$$

De plus, nous avons que $E[X] = \alpha/\theta$ *et que* $Var[X] = \alpha/\theta^2$*.*

Lors de notre analyse, nous supposerons α connu. À cette effet, il sera estimé à l'aide de la méthode du maximum de vraisemblance en utilisant les 10 premières années d'observations. Ce faisant, nous avons obtenu $\alpha = 0{,}8$ et cette valeur sera utilisée tout au long de ce mémoire.

Finalement, à titre de dernière hypothèse, nous supposerons que nos observations sont indépendantes.

2.1.4. Statistiques descriptives

Étant donné que nous suspectons un changement dans la moyenne au cours du temps, il ne serait pas d'une grande utilité de fournir les statistiques descriptives pour les 140 années. Toutefois, comme les dix premières années d'observations influenceront notre modélisation, nous nous y attarderons un peu.

En regardant le tableau 2.1, nous pouvons voir qu'il y a en moyenne 55,5 jours

TABLEAU 2.1. Statistiques descriptives pour les dix premières années.

Moyenne	Écart-type	Minimum	Maximum	N
4,72	5,23	$5{,}98 \times 10^{-5}$	30,00	555

de pluie par été pour les dix premières années et que le taux moyen d'accumulation est de 4,72 kg/m^2 par jour.

Afin d'apprécier visuellement le changement dans les précipitations, nous avons fait le diagramme des moyennes annuelles avec les intervalles de confiance associés que nous pouvons apercevoir à la figure 2.3. Notons que les intervalles de confiance sont d'un niveau 95%. Visuellement parlant, le changement dans la moyenne n'est pas si frappant sur cette figure. Toutefois, nous pouvons voir que la variabilité semble augmenter lorsqu'on progresse dans le temps. Afin de vérifier ce fait, nous avons fait le graphique des écarts-types en fonction du temps à la figure 2.4. En regardant cette figure, nous voyons bien que la variabilité semble augmenter avec le temps.

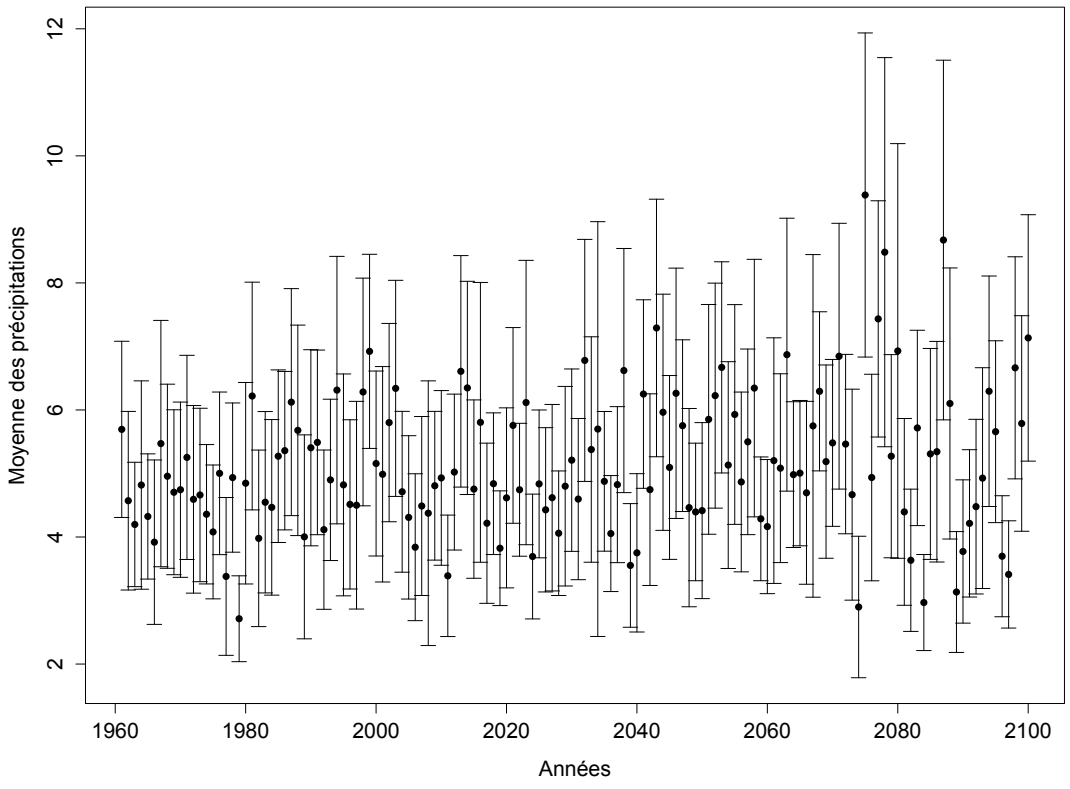

FIGURE 2.3. Diagramme des moyennes annuelles pour les données Ouranos.

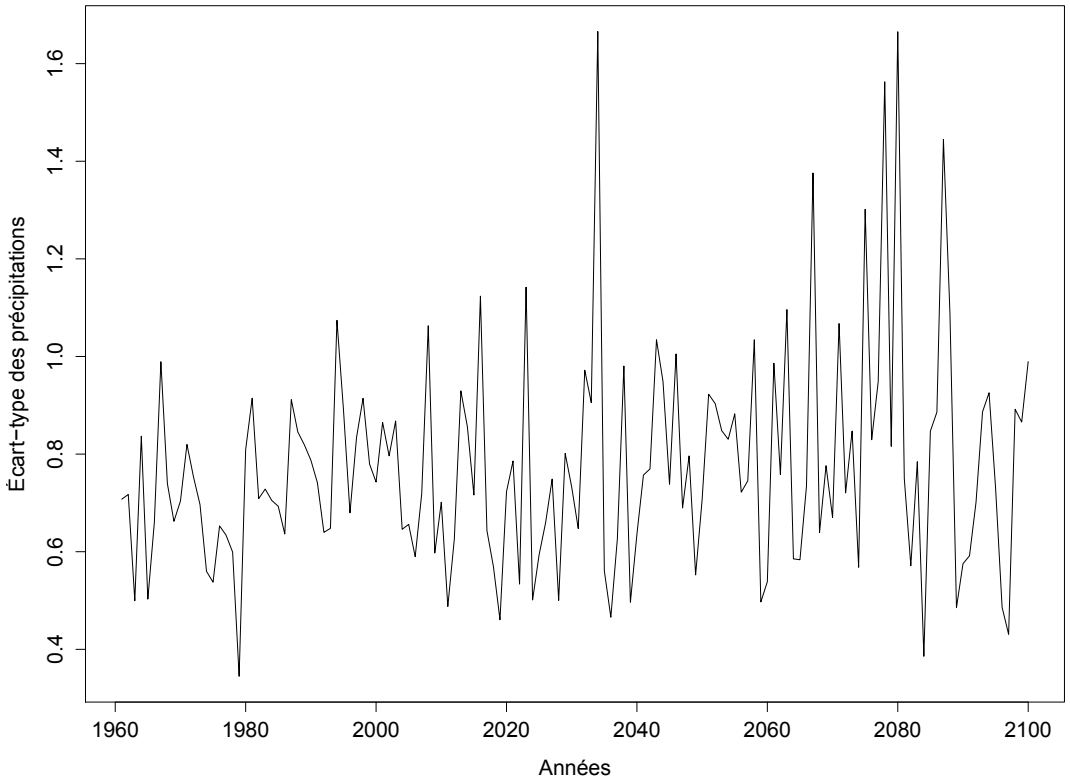

FIGURE 2.4. Écart-type de la moyenne en fonction de l'année.

2.1.5. Différents scénarios

Cette section sera réservée à l'élaboration de deux scénarios sur lesquels nos outils seront utilisés. Étant donné que nous soupçonnons une augmentation des précipitations à travers le temps, nous avons décidé d'utiliser deux scénarios qui prennent en compte un point de rupture, noté t_0, à partir duquel la moyenne des précipitations commence à augmenter.

Plus précisément, dans le premier scénario, noté S1, nous avons qu'à partir de t_0 la moyenne des observations augmente de manière brusque et permanente par un facteur d'augmentation c ($c > 1$). Ceci implique que les observations suivront la loi gamma de paramètre α et θ/c. En effet, nous avons que si $X \sim \text{Gamma}(\alpha, \theta)$ alors $E[X] = \alpha/\theta$. Conséquemment, si $Y \sim \text{Gamma}(\alpha, \theta/c)$ nous avons que $E[Y] = c \times \alpha/\theta = c \times E[X]$. La densité des observations sous le

scénario S1 est donc

$$f(x_{ij}|\theta) = \begin{cases} \frac{\theta^a}{\Gamma(a)} \times x_{ij}^{a-1} \exp\{-\theta x_{ij}\} & \text{si } x_{ij} > 0, i \leq t_0, \\ \frac{\theta^a}{c^a \Gamma(a)} \times x_{ij}^{a-1} \exp\left\{-\frac{\theta}{c} x_{ij}\right\} & \text{si } x_{ij} > 0, i > t_0, \\ 0 & \text{sinon.} \end{cases}$$

De plus, dans le cas où nous allons devoir procéder à des simulations, nous allons fixer la moyenne des observations à μ^* qui est la moyenne des 10 premières années d'observations du jeu de données Ouranos. Lors des simulations, le paramètre θ de la loi gamma sera donc $\theta = a/\mu^*$.

Pour le deuxième scénario, noté S2, nous avons qu'à partir du point de rupture, la moyenne des observations augmente linéairement en fonction des années. Afin de disposer d'un scénario réaliste, nous avons effectué une régression linéaire sur la moyenne de chaque année d'observations du jeu de données Ouranos. Nous avons ainsi enregistré que l'ordonnée à l'origine était de 4,77 et que la pente d'augmentation des moyennes était de $6{,}125 \times 10^{-3}$ par année. De plus, notons que les deux coefficients étaient significatifs. Si on note par μ_i la moyenne de l'année i, nous avons

$$\mu_i = \begin{cases} 4{,}77 & \text{si } i \leq t_0, \\ 4{,}77 + (i - t_0) \times 6{,}125 \times 10^{-3} & \text{sinon.} \end{cases} \tag{2.1.1}$$

Ceci implique que les observations de l'année i suivront une loi gamma de paramètre a et $\theta_i = a/\mu_i$.

Il est évident que plusieurs autres scénarios auraient pu être choisis, mais comme nous essayons de détecter le moment à partir duquel il y a eu un changement dans les précipitations, les modèles avec point de rupture sont particulièrement intéressants.

2.2. CONTRÔLE DE QUALITÉ STANDARD

Dans cette section, nous présenterons le contrôle de qualité standard. Plus précisément, nous commencerons par éclaircir les bases théoriques menant à cet outil. Évidemment, lorsque cela sera nécessaire, la théorie énoncée dans cette section sera établie en fonction du fait que les données suivent la loi gamma. Par la suite, nous présenterons l'estimateur de la moyenne usuel dans ce contexte et expliquerons comment l'outil fonctionne. De plus, comme l'ARL constitue l'indice de performance, nous allons expliquer comment la calculer

de manière théorique pour le scénario S1. Ensuite, nous utiliserons les simulations pour estimer l'ARL pour les deux scénarios climatiques différents. Finalement, nous allons expliquer les désavantages encourus par cette méthode et nous procéderons à un exemple pratique.

2.2.1. Base théorique

Commençons par définir la notation utilisée dans les lignes qui suivent. Soient x_{ij} la j^e observation du i^e groupe et notons qu'ici, les années constitueront les groupes. De plus, posons $\overline{x}_{i\cdot}$ comme étant la moyenne de la i^e année de n_i observations et précisons qu'afin de simplifier la notation, nous supposerons que $n_i = n \; \forall i$. Nous avons donc

$$\overline{x}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} x_{ij}.$$

En faisant l'hypothèse que lorsque le système est en contrôle, la moyenne de la population est μ_0 et que la variance est σ^2, nous savons que

$$\overline{X}_{i\cdot} \sim \text{Gamma}\left(a\,n, \frac{a\,n}{\mu_0}\right).$$

Il est donc possible de trouver des régions de rejet pour chaque $\overline{X}_{i\cdot}$. L'idée du contrôle de qualité standard est de faire le graphique de toutes les $\overline{x}_{i\cdot}$ et de juger le système hors de contrôle dès que l'une des $\overline{x}_{i\cdot}$ dépasse l'une des bornes de la région de rejet pour le niveau de confiance α^* voulu. Étant donné que nous souhaitons seulement détecter une augmentation dans la moyenne des précipitations, nous utiliserons des régions de rejet unilatérales. La limite de contrôle supérieur (LCS) pour $\overline{X}_{i\cdot}$ sera donc de la forme

$$\text{LCS}_{(1-\alpha\%)} = \Gamma_{\alpha^*}(\mu_0), \tag{2.2.1}$$

où $\Gamma_{\alpha^*}(\mu_0) > 0$ est le quantile $\alpha\%$ de la loi $\text{Gamma}(a\,n, a\,n/\mu_0)$ et il est donc choisi en fonction du niveau du test. La valeur par défaut de α^* pour les chaînes de production standard est de $\alpha^* \approx 0,0027$. Toutefois, comme nous allons procéder à des tests d'hypothèses par la suite, nous utiliserons un niveau $\alpha^* = 0{,}05$. Notons qu'un essai constitue le fait de vérifier si $\overline{x}_{i\cdot}$ est à l'intérieur de ses bornes.

2.2.2. Estimation des paramètres

En regardant l'expression de la borne supérieure de l'intervalle de confiance à l'équation (2.2.1), nous nous apercevons que nous devons connaître la valeur

de μ_0 afin de calculer l'intervalle de confiance. Cette valeur est le paramètre cible que le système devrait normalement utiliser. En industrie, nous connaissons cette valeur qui est prédéterminée par la compagnie. Par contre, pour notre véritable analyse, nous ne connaissons pas la valeur de μ_0. Il nous faudra donc l'estimer à l'aide des données. Évidemment, pour ce faire nous devrons tenter d'utiliser des données qui sont saines. Pour estimer μ_0, nous utiliserons k sous-groupes et l'estimateur de μ_0 sera

$$\widehat{\mu}_0 = \frac{1}{k}\sum_{i=1}^{k} \overline{x}_{i\cdot}. \tag{2.2.2}$$

En d'autres mots, cet estimateur, proposé dans Derman et Ross (1997) est la moyenne des k moyennes, ce qui revient à la grande moyenne si la taille des groupes est égale. Comme nous l'avons mentionné précédemment, l'estimateur de μ_0 a un sens seulement s'il est estimé à l'aide de k groupes sains. Afin de vérifier cette hypothèse (voir Derman et Ross, 1997), nous commençons par calculer la borne de la région de rejet basée sur l'estimation de μ_0 et on vérifie si

$$\overline{x}_{i\cdot} \leq \Gamma_{\alpha^*}(\widehat{\mu}_0), \quad \forall i \in \{1, ..., k\}.$$

Dans ce cas, nous supposons que les paramètres ont bien été estimés par des observations saines.

2.2.3. Calcul de l'$\mathrm{ARL}_{\mathrm{std}}$

L'objectif de cette section sera de calculer l'ARL dans le cadre du contrôle de qualité standard pour les deux scénarios. Dans ce cas particulier, l'ARL sera notée $\mathrm{ARL}^{\ell}_{\mathrm{std}}(\cdot)$, où le point représente les différents paramètres dont dépend l'ARL et où $\ell = 1$ pour le scénario S1 et $\ell = 2$ pour le scénario S2. Pour le premier scénario, l'ARL sera calculée de manière théorique ainsi que par simulation alors que pour le scénario S2, seule la simulation sera utilisée.

De plus, comme nous avions déjà effectué les simulations ainsi que les calculs théorique dans le cas où les observations suivent la loi exponentielle, nous supposerons que $a = 1$, c'est-à-dire, que les observations sont de loi exponentielle.

2.2.3.1. *Scénario S1*

Cette section sera consacrée au calcul de l'ARL lorsque les observations sont de moyenne $\mu_1 = \mu_0 + \delta\sigma/\sqrt{n}$. En outre, comme nous supposons la loi

Gamma$(a, a/\mu_0)$, nous avons que $\sigma = \mu/\sqrt{a}$ et donc

$$\begin{aligned}\mu_1 =& \mu_0 \left[1 + \frac{\delta}{\sqrt{a\,n}}\right] \\ =& \mu_0 \times c(\delta),\end{aligned} \tag{2.2.3}$$

où $c(\delta) = 1+\delta/\sqrt{a\,n}$. Nous sommes donc bien sous S1, c'est-à-dire, la moyenne des observations est multipliée par une constante c. Notons que ceci implique que $\overline{X}_{i\cdot} \sim \text{Gamma}\,(a\,n, a\,n/\mu_1)$. Nous avons donc que $\text{ARL}^1_{\text{std}}(\delta)$ est le nombre moyen d'essais avant de détecter que le système est hors de contrôle si la moyenne est décalée de $\delta/\sqrt{a\,n}$ écarts-types. Par conséquent, $\text{ARL}^1_{\text{std}}(0)$ est le nombre d'essais avant d'obtenir une fausse alarme. Afin de pouvoir calculer l'$\text{ARL}^1_{\text{std}}(\delta)$, nous allons tout d'abord trouver la probabilité de détecter que le système est hors de contrôle. Ceci va comme comme suit

$$\begin{aligned}&\mathbb{P}\,(\text{Détecter que le système est hors de contrôle si la moyenne est } \mu_1) \\ &= \mathbb{P}\left(\overline{X}_{i\cdot} \geq \Gamma_{\alpha^*}(\mu_0)\right) \\ &= p_{\text{std}},\end{aligned}$$

où $\overline{X}_{i\cdot} \sim \text{Gamma}\,(a\,n, a\,n/\mu_1)$.

Définition 2.2.1. *Soit* G *une variable aléatoire de loi géométrique de paramètre* p*, notée* $G \sim \text{Geo}\,(p)$*. Alors, la fonction de masse de G est :*

$$\mathbb{P}(G = k) = (1-p)^{k-1}p, \quad k \in \mathbb{N}^+.$$

De plus,

$$\mathbb{E}[G] = \frac{1}{p}$$

et

$$\text{Var}[G] = \frac{1-p}{p^2}.$$

Soit G défini comme étant le nombre d'essais avant de détecter la défaillance du système. Nous avons donc que $G \sim \text{Geo}\,(p_{\text{std}})$. Comme $\text{ARL}^1_{\text{std}}(\delta) = \mathbb{E}\,[G]$, nous avons que

$$\text{ARL}^1_{\text{std}}\,(\delta, \alpha^*) = \frac{1}{p_{\text{std}}}. \tag{2.2.4}$$

Nous pouvons voir, au tableau 2.2, l'$\text{ARL}^1_{\text{std}}(\delta)$ théorique et simulée pour quelques valeurs de δ et pour $\alpha^* = 0{,}05$. À la ligne théorique, nous avons utilisé l'équation (2.2.4) et à la ligne simulation, nous avons procédé à des simulations. De plus, rappelons que nous avons fixé $a = 1$ pour l'étude de l'ARL.

TABLEAU 2.2. ARL^1_{std} en fonction de δ et pour $\alpha^* = 0{,}05$, $a = 1$.

δ	0	0,1	0,25	0,5	0,75	1	2	2,5	3
Théorique	20,00	16,14	11,99	7,76	5,36	3,93	1,75	1,41	1,22
Simulation	27,38	21,56	15,48	9,48	6,32	4,47	1,84	1,46	1,25

En bref, pour la partie simulation, à chaque itération et pour chaque valeur de δ, nous avons procédé aux étapes suivantes :

(1) Nous avons créé 10 années d'observations saines de moyenne égale à μ^*.

(2) Nous avons calculé $\widehat{\mu}_0$ en utilisant l'équation (2.2.2) et en utilisant les 10 années d'observations saines ($k = 10$).

(3) Nous avons créé 100 années d'observations contenant chacune 55 observations qui proviennent d'une loi Gamma(a, θ) de moyenne μ_1, où μ_1 est défini comme à l'équation (2.2.3) et où $a = 1$. Le nombre d'années d'observations pouvant être ajusté au besoin.

(4) Nous avons appliqué la méthode du contrôle de qualité standard, c'est-à-dire, nous avons enregistré le nombre d'essais avant que nous détections un $\overline{x}_{i.}$ dans la région de rejet construite comme à l'équation (2.2.1).

Nous avons répété le processus 50 millions de fois pour chaque δ et nous avons fait la moyenne du nombre d'années nécessaire avant d'avoir une alarme. Ceci constitue la ligne simulation du tableau 2.2.

En temps normal, si toutes les hypothèses étaient respectées, les valeurs de la partie simulation devraient converger vers les valeurs théoriques lorsque le nombre d'itérations augmente, mais ce n'est pas le cas ici. En effet, nous avons recommencé la simulation à plusieurs reprises et les valeurs de l'ARL^1_{std} semblent avoir convergé. Toutefois, nous pouvons observer que les résultats obtenus pour la partie simulation ne sont pas très éloignés des résultats théoriques. De plus, nous pouvons remarquer que $1/27{,}38 \approx 0{,}037$, ce qui n'est pas loin du niveau espérer de 0,05. Nous allons donc tout de même considérer que nos tests se font à un niveau de 5%.

En approfondissant un peu, nous nous sommes rendus compte que le problème venait de l'estimation de μ_0. En effet, à chaque itération, nous estimons μ_0 et son estimation varie relativement beaucoup. En fait, nous avons refait les tests en supposant μ_0 connu et la pratique convergeait vers la théorie très rapidement. Comme nous ne disposons pas de μ_0 dans notre cas concret, nous allons nous fier aux résultats de simulation.

2.2.3.2. *Scénario S2*

Comme nous l'avons spécifié plus tôt, pour le deuxième scénario, seules les simulations seront utilisées pour déterminer l'ARL^2_{std}. Les étapes concernant cette simulation sont très similaires à celles de la simulation pour le premier scénario. En effet, seul le modèle de génération de données change. À cet effet, la moyenne des observations sera régie par le modèle des moyennes présenté à l'équation (2.1.1) avec $t_0 = 10$. En d'autres mots, les dix premières années seront les années considérées comme saines. Après 50 millions d'itérations, nous avons obtenu l'$ARL^2_{std} = 17{,}16$, c'est-à-dire, nous avons eu besoin de 17,16 coups avant de détecter un changement dans les données.

2.2.4. Exemple pratique pour carte de contrôle standard

Cette section a pour objectif de procéder à une application du contrôle de qualité standard sur un jeu de données simple. Ce jeu de données comprendra 21 années de 55 observations chacune. Pour les 10 premières années, les observations ont été générées avec une loi Gamma$(1, 1/\mu_i)$ de moyenne $\mu_0 = 10$ alors que les 11 dernières avaient une moyenne de $\mu_1 = \mu_0 + 0{,}5\mu_0/\sqrt{n} = 10{,}67$. Notons que ceci représente une augmentation de 0,5 écart-type. Les moyennes des 21 années sont présentées au tableau 2.3. Les dix premières années d'observations seront les années indemnes et les autres seront considérées comme corrompues. Nous utiliserons donc les dix premières années d'observations pour estimer μ_0 à l'aide de l'équation (2.2.2). Plus précisément, en utilisant les

TABLEAU 2.3. Moyennes pour les 21 années d'observations.

ℓ	$\overline{X}_{\ell\cdot}$	ℓ	$\overline{X}_{\ell\cdot}$
1	9,96	11	7,90
2	7,72	12	11,63
3	11,26	13	11,00
4	9,06	14	11,48
5	10,42	15	10,54
6	7,91	16	9,69
7	11,31	17	9,88
8	8,00	18	14,51
9	9,57	19	9,44
10	10,02	20	9,56
		21	9,37

données présentées au tableau 2.3, nous obtenons $\widehat{\mu}_0 = 9{,}52$. De plus, en utilisant l'expression (2.2.1), nous avons que la borne de la région de rejet est

$$\Gamma_{0{,}05}\left(\widehat{\mu}_0\right) = 11{,}73.$$

Nous devons maintenant vérifier si les dix premières années sont considérées comme saines. Pour ce faire, nous avons inclus les années indemnes avec les corrompues sur la carte de contrôle présentée à la figure 2.5. Notons que la ligne rouge représente la borne supérieure de l'intervalle de confiance et que la ligne bleue représente la démarcation entre les données indemnes et corrompues. Nous devons donc regarder si la moyenne des années indemnes (à gauche de la ligne bleue) sont toutes inférieures à la ligne rouge (11,73), ce qui est le cas ici. Passons maintenant aux données corrompues. Nous pouvons voir que nous avons besoin de huit essais avant de pouvoir détecter l'augmentation de 0,5 écart-type dans la moyenne. De plus, remarquons qu'il n'y a pas de tendance particulière qui indique clairement que la moyenne a augmenté, ce qui n'est pas très rassurant.

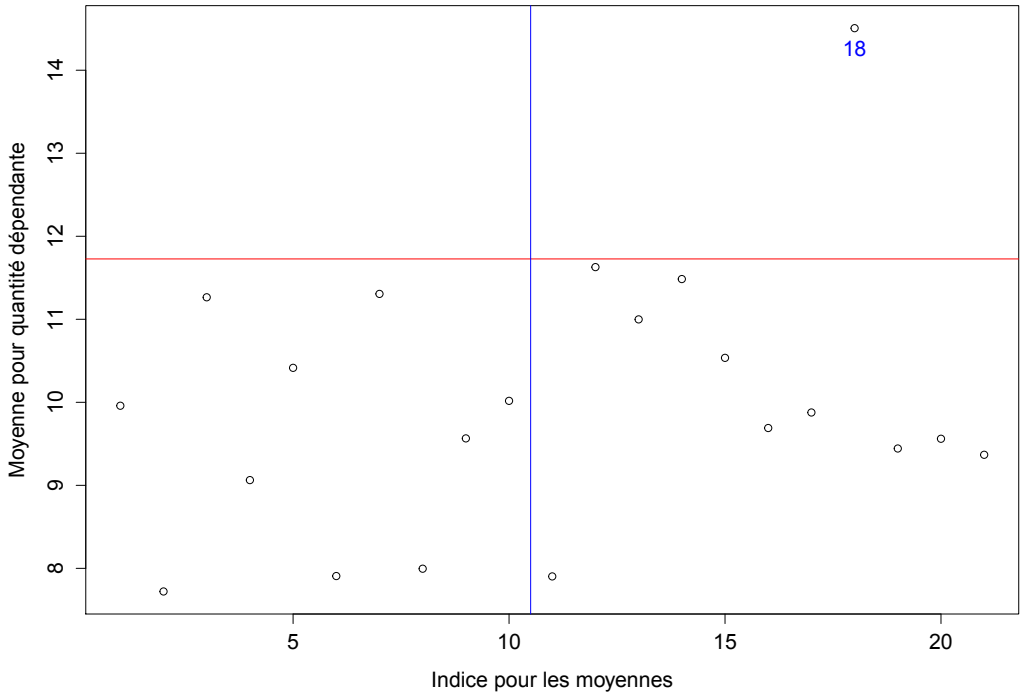

FIGURE 2.5. Carte de contrôle standard.

2.2.5. Désavantage majeur

Bien que cette approche soit très efficace pour détecter des changements de grande amplitude, elle n'est pas très efficace pour repérer un petit changement définitif. En fait, les cartes de contrôle de Shewhart sont vraiment performantes lorsque la grandeur de l'accroissement est supérieure à 1,5 écarts-types, mais pas très efficace pour des accroissements inférieurs à un écart-type (voir Montgomery, 2007). Par exemple, pour un α^* fixé, si un changement permanent se fait au niveau de la moyenne μ_0, mais que ce changement est trop petit, le système ne sera pas jugé hors de contrôle rapidement. Quelqu'un pourrait suggérer d'augmenter α^* pour pouvoir détecter plus vite une faible variation, mais le problème est que plus nous augmentons α^*, plus la fréquence de fausses alertes augmente. Ce désavantage est dû, en grande partie, au fait que nous n'utilisons pas l'information disponible sur les observations passées, car nous traitons indépendamment chaque $\overline{X}_i$.. Il serait alors plus efficace de trouver une méthode qui utilise les déviations des observations précédentes.

2.3. Carte de contrôle CUSUM

Dans cette section, nous commencerons par énoncer les bases théoriques inhérentes au CUSUM. Par la suite, nous expliquerons comment procéder à la sélection des paramètres optimaux et nous terminerons avec un exemple d'application sur jeu de données.

2.3.1. Base théorique

Depuis son introduction par Page (1954), la carte de contrôle de sommation cumulative (traduction de « *Cumulative sum control charts* ») a été énormément utilisée dans le domaine de l'industrie afin de détecter les changements dans la moyenne. En effet, les cartes CUSUM sont utilisées dans plusieurs domaines tels que l'économie (voir Saniga *et al.*, 2006) et la médecine (voir Grigg *et al.*, 2003). Un avantage majeur lié à cette technique est qu'elle prend en compte les données précédentes qui s'écartent du modèle. Cette carte est probablement l'une des plus puissantes pour détecter un petit changement qui persiste dans le temps (voir Derman et Ross, 1997) et semble donc très bien adaptée aux données climatiques que nous étudierons.

Afin d'aider à la compréhension du modèle CUSUM, nous avons décidé de commencer par définir le modèle des sommes partielles. Voici le modèle permettant de détecter une augmentation de la moyenne. Premièrement, on

choisit des constantes k et $h > 0$ et ensuite nous posons :

$$\begin{aligned} Y_i &= \overline{X}_{i\cdot} - \mu_0 - k\sigma^* \quad i \geq 1, \\ &= \overline{X}_{i\cdot} - \mu_0 \left(1 + \frac{k}{\sqrt{a\,n}}\right) \quad i \geq 1, \end{aligned} \tag{2.3.1}$$

où $\sigma^* = \sigma/\sqrt{n} = \mu_0/\sqrt{a\,n}$ dans le cas de la loi gamma. Notons que si le système est en contrôle, nous avons que $\mathbb{E}[Y_i] = -k\sigma^* < 0$. De plus, posons :

$$P_\ell = \sum_{i=1}^{\ell} Y_i,$$

qui aurait pu être écrit

$$P_0 = 0$$

et

$$P_{\ell+1} = P_\ell + Y_{\ell+1}, \quad \ell \geq 1.$$

Étant donné que l'espérance de Y_i est négative, nous pouvons nous attendre à ce que P_ℓ soit négatif aussi. Par contre, si l'espérance des $\overline{X}_{i\cdot}$ augmente suffisamment, l'espérance des Y_i deviendra positive. Conséquemment, l'idée de cette carte est de considérer le système hors de contrôle lorsque $P_\ell \geq h\sigma^*$. Le problème avec cette méthode est que si le processus a été en contrôle pour une très longue durée, il est fort probable que la somme partielle soit très négative. De ce fait, même si l'espérance augmente de manière significative, cela risque de prendre beaucoup de temps avant de juger le système comme étant hors de contrôle.

Afin de pallier ce problème, Page (1954) a eu l'idée de réinitialiser la somme partielle à zéro à chaque fois qu'elle devenait négative. De cette manière, le fait que le processus ait été en contrôle très longtemps n'influencera pas sur le nombre d'essais avant de pouvoir détecter une augmentation de la moyenne. Le modèle s'écrit comme suit :

$$C_0 = 0$$

et

$$C_\ell = \max\left(C_{\ell-1} + Y_\ell, 0\right), \quad \ell \geq 1.$$

Le procédé du contrôle de qualité CUSUM consiste à faire le graphique de chaque C_ℓ et de considérer que le système est hors de contrôle (et donc que l'espérance des $\overline{X}_{i\cdot}$ a augmenté) dès que $C_\ell > h\sigma^*$.

2.3.2. Calcul de l'ARL_{CUSUM}

Tout comme dans le contrôle de qualité standard, il est d'usage d'utiliser l'ARL comme mesure de performance. Aussi, comme le modèle dépend de k et de h, nous tenterons de trouver la combinaison optimale de ces paramètres. En général, ce choix est fait de manière à maximiser la performance de l'outil (voir Montgomery, 2007), ce qui veut dire que ce choix sera étroitement lié à l'ARL. Cette section sera donc consacrée à l'évaluation de l'ARL pour les deux scénarios. Tout comme dans le contrôle de qualité standard, nous discuterons de l'ARL théorique seulement dans le cas du premier scénario, mais nous simulerons l'ARL pour les deux scénarios. Dans le cas du CUSUM, l'ARL sera notée $ARL^{\ell}_{CUSUM}(\cdot)$, où $\ell = 1$ pour le premier scénario et où $\ell = 2$ dans le cas du deuxième.

De plus, comme nous avions déjà effectué les simulations ainsi que les calculs théorique dans le cas où les observations suivent la loi exponentielle, nous supposerons que $a = 1$, c'est-à-dire, que les observations sont de loi exponentielle.

2.3.2.1. *Scénario S1*

Dans le cas du premier scénario, l'ARL sera notée $ARL^{1}_{CUSUM}(\delta, k, h)$ et est définie comme étant le nombre d'essais moyen avant que $C_\ell \geq h\sigma^*$ si les observations sont de moyenne μ_1, où μ_1 est défini comme à l'équation (2.2.3). Notons que ceci implique que $\overline{X}_{i\cdot} \sim N\left(\mu_1, \mu_1^2/n\right)$ par le théorème central limite. De plus, remarquons que lorsque $\delta = 0$, ceci revient au nombre d'observations moyen avant d'avoir une fausse alerte. Il convient alors de regarder le choix optimal pour k et h, c'est-à-dire, trouver les valeurs de k et h qui augmentent le plus possible le nombre d'essais avant une fausse alerte ($\delta = 0$), et qui le minimisent quand on est en présence d'un réel changement ($\delta > 0$). Pour approximer l'$ARL_{CUSUM}(\delta, k, h)$, nous utiliserons l'approximation de Siegmund présentée dans Siegmund (1985), c'est-à-dire,

$$ARL_{CUSUM}(\delta, k, h) \approx \begin{cases} \frac{\exp\{-2\Delta b\}+2\Delta b-1}{2\Delta^2} & \text{si } \Delta \neq 0, \\ b^2 & \text{si } \Delta = 0, \end{cases} \tag{2.3.2}$$

où $\Delta = \delta - k$ et $b = h + 1{,}166$. Toutefois, cette approximation est valide seulement si la statistique d'intérêt, ici la moyenne, suit la loi normale. Dans le cas particulier de la moyenne, le théorème central limite nous assure que la distribution de la moyenne tend vers la loi normale lorsque $n \to \infty$. Par contre, comme nous l'avons souligné à la section 2.1.4, il y a environ 55 jours de pluie

en moyenne par été et donc $n \approx 55$. Ceci n'est pas énorme dans le cas de la loi gamma (ou exponentielle) qui est très asymétrique et il se peut donc que l'approximation soit un peu moins précise. De plus, l'approximation de Siegmund est connue pour être moins précise lorsque $k \geq 1$ (voir Rogerson, 2006) et ce, particulièrement pour $\delta = 0$ (fausse alarme).

En regardant le modèle du CUSUM, nous pouvons noter que le modèle n'utilise pas explicitement le niveau α^*. En fait, pour pouvoir comparer les deux approches du contrôle de qualité ainsi que l'approche bayésienne, qui sera présentée sous peu, nous devons fixer le même niveau pour chaque approche. En effet, il serait inapproprié de comparer la performance de différents outils si les niveaux α^* ne sont pas semblables. Pour ce qui est du niveau du CUSUM, il sera approximé à l'aide de l'expression suivante :

$$\alpha^*_{\text{CUSUM}}(k, h) \approx \frac{1}{\text{ARL}_{\text{CUSUM}}(0, k, h)}. \tag{2.3.3}$$

Comme nous pouvons le voir à l'équation (2.3.3), le niveau sera fonction de k et h. Il faudra donc choisir k et h de façon à respecter le niveau α^* et aussi de façon à minimiser $\text{ARL}_{\text{CUSUM}}(\delta, k, h)$ pour $\delta > 0$.

L'approche usuelle pour résoudre ce problème serait de regarder dans les tables déjà produites à cet effet. Toutefois, le problème est que les valeurs de k et h disponibles dans la littérature ont été générées en fonction de l'approche six sigma, c'est-à-dire, qu'on juge le système hors de contrôle seulement si la statistique d'intérêt s'écarte à plus de trois écarts-types de la vraie moyenne. Ceci correspond à des $\text{ARL}_{\text{CUSUM}}(0, k, h)$ de 370 dans le cas bilatéral et de 740 dans le cas unilatéral et à un niveau de $\alpha^* = 0{,}0027$ dans le cas bilatéral et de $\alpha^* = 0{,}00135$ dans le cas unilatéral. Évidemment, ces niveaux sont très petits et il serait intéressant de pouvoir déterminer les valeurs de k et h sans se restreindre à ces niveaux.

En premier lieu, pour le choix de k, il est d'usage de choisir k en fonction de

$$\delta^* = \frac{\mu_1 - \mu_0}{\sigma^*},$$

où δ^* est l'accroissement qu'on aimerait être capable de détecter si nous écrivons l'accroissement de la moyenne comme étant $\mu_1 = \mu_0 + \delta^*\sigma^*$. Notons qu'on ne doit pas confondre δ^* avec le δ défini plus tôt comme étant le vrai accroissement de la moyenne. En théorie, la valeur $k = \delta^*/2$ est très proche de minimiser l'$\text{ARL}_{\text{CUSUM}}(\delta, k, h)$ pour une $\text{ARL}_{\text{CUSUM}}(0, k, h)$ fixée (voir Montgomery, 2007). Dans notre cas, il serait intéressant d'au moins détecter une augmentation de la moyenne de l'ordre des 10%. De plus, rappelons que les données

suivent la loi gamma et donc que $\sigma = \mu/\sqrt{a}$. Nous avons donc que

$$\begin{aligned} k &= \frac{\delta^*}{2} \\ &= \frac{\mu_1 - \mu_0}{2\sigma^*} \\ &= \frac{(1{,}1\mu_0 - \mu_0) \times \sqrt{a\,n}}{2\mu_0} \\ &= 0{,}10 \times \sqrt{a\,n}. \end{aligned} \tag{2.3.4}$$

Comme nous avons pu le voir au tableau 2.1, il y a environ 55 jours de pluie par été, ce qui nous donne un k proche de 0,7 en utilisant l'équation (2.3.4). Nous allons donc, pour la suite, étudier les propriétés de l'ARL pour des k se rapprochant de cette valeur.

Maintenant que nous avons une bonne idée de l'étendue des valeurs possibles de k, il nous reste simplement à trouver h pour l'$\mathrm{ARL}_{\mathrm{CUSUM}}(0, k, h)$ voulue, c'est-à-dire, 20 dans notre cas. À cet effet, plusieurs chercheurs se sont penchés sur la question, c'est-à-dire, trouver un moyen de trouver les paramètres pour une ARL en contrôle fixée. Entre autre, les ouvrages Hanif *et al*. (2012); Ryu *et al*. (2010); Rogerson (2006) traitent de cette question. Comme il n'est pas possible d'isoler h dans l'approximation de Siegmund établie à l'équation (2.3.2), nous devrons utiliser des méthodes numériques ou d'autres approximations. Dans l'article Rogerson (2006), il est proposé d'utiliser l'approximation suivante

$$b \approx \left(\frac{2k^2 \times \mathrm{ARL}_0 + 2}{2k^2 \times \mathrm{ARL}_0 + 1}\right) \times \frac{\log\left(1 + 2k^2 \times \mathrm{ARL}_0\right)}{2k}, \tag{2.3.5}$$

où $\mathrm{ARL}_0 = \mathrm{ARL}_{\mathrm{CUSUM}}(0, k, h)$ et où $b = 1 + h$. Donc, en obtenant une estimation pour b nous obtenons automatiquement une estimation pour h. Notons que cette approximation découle directement de l'approximation de Siegmund présentée à l'équation (2.3.2) et aura donc les mêmes limitations. L'auteur suggère que l'approximation est plus précise pour $1/\sqrt{\mathrm{ARL}_0} < k \leq 1$ et nous allons donc rester dans cet intervalle. Plus précisément, pour un $\alpha^* = 0{,}05$, nous devons prendre $k \in [0{,}22; 1]$. Comme nous pouvons le voir au tableau 2.4, les ARL_0 sont très près des valeurs attendues pour chaque valeur de k et de h. L'algorithme (2.3.5) semble donc fonctionner correctement. Maintenant que nous avons obtenu des valeurs de k et de h qui nous donnent les ARL_0 désirées, nous devons choisir la combinaison qui minimise l'$\mathrm{ARL}_{\mathrm{CUSUM}}(\delta, k, h)$ pour $\delta > 0$. Toutefois, avant de choisir les paramètres k et h sur la base de l'ARL, nous allons calculer l'ARL par la simulation pour voir à quel point la

TABLEAU 2.4. Valeurs théoriques de $ARL^1_{CUSUM}(\delta, k, h)$ pour $\alpha^* = 0{,}05$.

δ	k=0,3 h= 1,93	k= 0,4 h= 1,67	k= 0,5 h= 1,45	k= 0,6 h= 1,26	k= 0,7 h= 1,1	k= 0,8 h= 0,96	k= 0,9 h= 0,84	k= 1 h= 0,74
0	19,73	20,00	20,13	20,09	20,09	20,01	19,99	20,21
0,1	15,15	15,45	15,67	15,78	15,90	15,96	16,06	16,34
0,25	10,66	10,90	11,13	11,29	11,47	11,62	11,79	12,08
0,5	6,60	6,72	6,84	6,96	7,11	7,25	7,41	7,64
0,75	4,56	4,58	4,63	4,68	4,77	4,86	4,97	5,12
1	3,42	3,38	3,38	3,39	3,42	3,47	3,54	3,63
2	1,65	1,58	1,52	1,48	1,45	1,43	1,42	1,42
2,5	1,30	1,24	1,18	1,14	1,10	1,08	1,06	1,05
3	1,08	1,02	0,97	0,92	0,89	0,86	0,84	0,83

pratique rejoint la théorie. Pour la simulation, nous avons utilisé la même approche que pour le contrôle de qualité standard, mais nous devions en plus fixer un k. Après avoir fixé un k, nous devions trouver h à l'aide de l'équation (2.3.5) pour une valeur de ARL_0 voulue (ici 20).

De nouveau, nous pouvons voir que la pratique ne converge pas totalement vers la théorie. Par contre, encore une fois, nous pouvons voir que le niveau pratique est très près du niveau théorique. Par exemple, nous avons que $1/33 \approx 0{,}03$ et donc, nous allons considérer que le niveau a été conservé dans la partie pratique étant donné que le taux de fausses alarmes est très proche de 5%. De plus, comme la théorie semble relativement éloignée de la pratique,

TABLEAU 2.5. Valeurs simulées de $ARL^1_{CUSUM}(\delta, k, h)$ pour $\alpha^* = 0{,}05$.

δ	k=0,3 h= 1,93	k= 0,4 h= 1,67	k= 0,5 h= 1,45	k= 0,6 h= 1,26	k= 0,7 h= 1,1	k= 0,8 h= 0,96	k= 0,9 h= 0,84	k= 1 h= 0,74
0	33,77	30,97	28,83	27,21	26,31	25,83	25,91	26,70
0,1	24,14	22,71	21,57	20,66	20,17	19,96	20,13	20,79
0,25	15,47	14,97	14,53	14,17	14,03	14,01	14,21	14,71
0,5	8,52	8,44	8,37	8,32	8,34	8,42	8,59	8,90
0,75	5,48	5,44	5,43	5,43	5,47	5,54	5,66	5,87
1	3,94	3,90	3,88	3,87	3,90	3,95	4,03	4,16
2	1,91	1,84	1,80	1,77	1,75	1,75	1,76	1,78
2,5	1,57	1,51	1,47	1,44	1,43	1,42	1,42	1,43
3	1,36	1,31	1,28	1,26	1,25	1,24	1,24	1,24

nous allons encore une fois utiliser les résultats de simulations comme étant les résultats les plus représentatifs de notre situation. Finalement, comme les choix de k et h donnent des résultats similaires, nous allons garder la première valeur de k que nous avons proposée, c'est-à-dire, nous choisissons $k = 0{,}7$ et $h = 1{,}1$.

2.3.2.2. *Scénario S2*

Comme nous l'avons spécifié plus tôt, pour le deuxième scénario, seules les simulations seront utilisées pour déterminer l'$ARL^2_{CUSUM}(k, h)$. Les étapes concernant cette simulation sont très similaires à celles de la simulation pour le premier scénario. En effet, seul le modèle de génération de données change. À cet effet, la moyenne des observations sera régie par le modèle des moyennes présenté à l'équation (2.1.1) avec $t_0 = 10$. En d'autres mots, les dix premières années seront les années considérées comme saines. Après 50 millions d'itérations, nous pouvons voir ce que nous avons obtenu au tableau 2.6.

TABLEAU 2.6. Valeurs simulées de $ARL^2_{CUSUM}(k, h)$ pour $\alpha^* = 0{,}05$.

k=0,3 h= 1,93	k= 0,4 h= 1,67	k= 0,5 h= 1,45	k= 0,6 h= 1,26	k= 0,7 h= 1,1	k= 0,8 h= 0,96	k= 0,9 h= 0,84	k= 1 h= 0,74
16,72	16,49	16,26	16,03	15,95	15,96	16,14	16,58

Encore une fois, le choix de $k = 0{,}7$ et $h = 1,1$ semble être un choix astucieux. En fait, avec ce choix de paramètres, nous avons eu besoin de 15,95 années, en moyenne, afin de détecter un changement. De plus, étant donné que les deux scénarios mènent au même choix de k et de h et que de surcroit, l'impact de ce choix n'a pas une grande influence sur l'ARL, nous utiliserons $k = 0{,}7$ et $h = 1,1$ pour le reste de cet exposé.

2.3.3. Exemple pratique pour CUSUM

Cette section a pour objectif de montrer une application des cartes de contrôle de qualité CUSUM sur un jeu de données. Nous utiliserons le même jeu de données que pour le contrôle de qualité standard, où les moyennes ont été présentées au tableau 2.3. Afin de pouvoir tracer la carte CUSUM, nous devons calculer les quantités $C_\ell = \max(C_{\ell-1} + Y_\ell, 0), \ell \geq 1$. Pour ce faire, nous devons avoir calculé préalablement les quantités Y_ℓ définies à l'équation (2.3.1). Nous pouvons voir le résumé de ces entités numériques au tableau 2.7. En plus de ces données, nous aurons besoin de calculer la borne supérieure pour C_ℓ qui est

$$\begin{aligned} h\sigma^* &= 1,1 \times \frac{\widehat{\mu}_0}{\sqrt{a\,n}} \\ &= 1,1 \times \frac{9{,}52}{\sqrt{55}} \\ &= 1{,}41, \end{aligned}$$

où cette valeur correspond à la ligne rouge horizontale sur la figure 2.6.

TABLEAU 2.7. Valeurs pour la carte de contrôle CUSUM.

ℓ	$\overline{X}_{\ell\cdot}$	Y_ℓ	C_ℓ	ℓ	$\overline{X}_{\ell\cdot}$	Y_ℓ	C_ℓ
1	9,96	-0,46	0,00	11	7,90	-2,52	0,00
2	7,72	-2,70	0,00	12	11,63	1,21	1,21
3	11,26	0,84	0,84	13	11,00	0,58	1,79
4	9,06	-1,36	0,00	14	11,48	1,06	2,85
5	10,42	-0,01	0,00	15	10,54	0,12	2,97
6	7,91	-2,51	0,00	16	9,69	-0,73	2,24
7	11,31	0,89	0,89	17	9,88	-0,54	1,69
8	8,00	-2,42	0,00	18	14,51	4,09	5,78
9	9,57	-0,85	0,00	19	9,44	-0,98	4,80
10	10,02	-0,40	0,00	20	9,56	-0,86	3,94
				21	9,37	-1,05	2,89

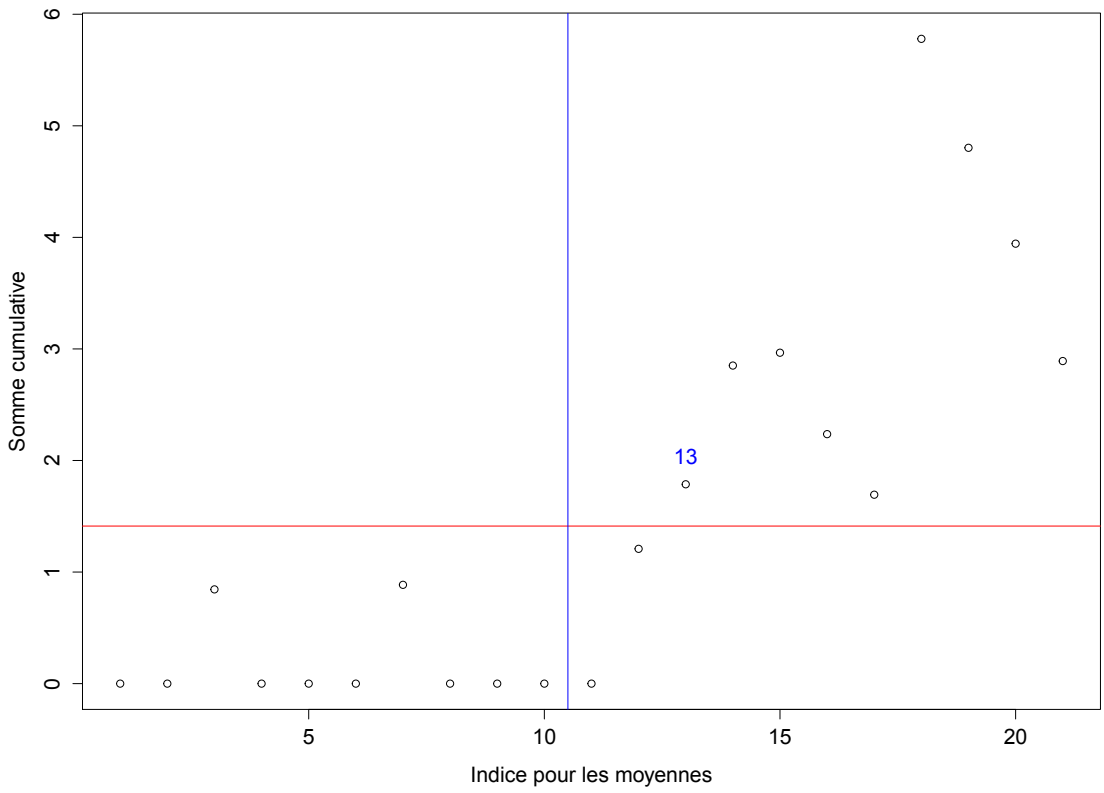

FIGURE 2.6. Carte de contrôle à sommation cumulative.

Comme nous pouvons le voir à la figure 2.6, la carte de contrôle CUSUM commence à accumuler réellement les déviations seulement pour les années qui sont corrompues. De plus, notons que cet outil détecte l'augmentation de moyenne en seulement trois essais. Finalement, nous aimerions préciser qu'il

est clair ici que nous sommes en présence d'une augmentation de la moyenne puisque C_ℓ est clairement au-dessus du seuil pour plusieurs essais consécutifs, ce qui est très rassurant.

De plus, nous aimerions prendre le temps de souligner une des particularités de cette carte. En regardant plus attentivement le tableau 2.7, pour $\ell = 12$ nous ne détectons pas une augmentation de moyenne alors que pour $\ell = 13$, nous la détectons. Pourtant, nous avons que $\overline{X}_{12} > \overline{X}_{13}$. Ceci peut paraître bizarre, mais la carte CUSUM est bâtie de manière à accumuler les déviations et vérifier si l'accumulation des déviations dépasse le seuil $h\sigma^*$.

Chapitre 3

APPROCHE BAYÉSIENNE

Dans ce chapitre, nous discuterons la méthode que nous utiliserons afin de détecter la non-stationnarité dans les observations, et ce, dans le cadre de la statistique bayésienne. Cette méthode sera basée sur le concept de distance entre les densités *a posteriori*. Nous commencerons par motiver le choix d'utiliser des distances sur les densités *a posteriori*. Par la suite, une grande partie du chapitre consistera à expliquer et à approfondir les différentes distances qui seront utilisées. Sur la base de l'étude de ces distances, nous trouverons une statistique de test, notée W, qui résumera l'information qui était apportée par les distances. Après avoir trouvé la densité prédictive de W et ainsi d'avoir conçu un test, nous étudierons l'ARL de notre méthode afin de pouvoir en calibrer les différents paramètres. Finalement, nous ferons un exemple de l'utilisation de l'outil sur un jeu de données.

3.1. MOTIVATION

Dans un contexte bayésien, nous disposons de deux sources d'information. Premièrement, nous avons l'information apportée par les données observées $x_1, x_2, ..., x_{N_x}$ qui proviennent d'une population de densité $f(\underset{\sim}{x}|\theta)$ où $\theta \in \Theta$ et où $\underset{\sim}{x} = (x_1, x_2, ..., x_{N_x})$. Deuxièmement, nous supposons que θ est une variable aléatoire et que nous disposons d'information *a priori* sur θ qui peut être représentée à l'aide d'une densité. Cette densité est notée $\pi(\theta)$ et on la désignera en tant que densité *a priori*. La densité *a posteriori* est la combinaison de ces deux sources d'information et elle est définie par

$$\pi(\theta|\underset{\sim}{x}) = \frac{\pi(\theta) \times \prod_{i=1}^{N_x} f(x_i|\theta)}{\int_\Theta \pi(\theta) \times \left[\prod_{i=1}^{N_x} f(x_i|\theta)\right] d\theta}.$$

De plus, la quantité $\int_\Theta \pi(\theta) \times \left[\prod_{i=1}^{N_x} f(x_i|\theta)\right] d\theta$ est généralement notée $m(\underset{\sim}{x})$, ce qui nous donne

$$\pi(\theta|\underset{\sim}{x}) = \frac{\pi(\theta) \times \prod_{i=1}^{N_x} f(x_i|\theta)}{m(\underset{\sim}{x})}.$$

Comme nous pouvons le remarquer, si la loi des observations change, la loi *a posteriori* change aussi. Donc, pour détecter la non-stationnarité des précipitations à travers le temps, dans un cadre bayésien, nous modéliserons la densité *a posteriori* d'une période donnée et nous la comparerons à la densité *a posteriori* d'une autre période plus éloignée dans le temps. Si la ressemblance entre les deux densités *a posteriori* est faible, nous pourrons alors soupçonner la non-stationnarité des observations. Afin de comparer les densités *a posteriori*, nous utiliserons une statistique basée sur les différentes distances qui seront définies plus tard. De plus, il sera possible d'obtenir la loi de cette statistique, et d'ainsi, obtenir un test statistique basé sur celle-ci. Une différence significative entre deux densités *a posteriori* nous indiquerait qu'il y a eu un changement dans la distribution des précipitations à travers les années. Comme l'expression des distances change pour chaque combinaison de densités *a posteriori* modélisées, nous allons devoir spécifier des hypothèses pour qu'on puisse étudier plus précisément les distances.

3.1.1. Hypothèse bayésienne

Évidemment, dans un cadre bayésien, nous devons définir une densité *a priori* pour θ. Afin de simplifier le problème, nous supposerons que $\theta \sim \text{Gamma}(\alpha, \beta)$. En fait, comme la loi $\text{Gamma}(\alpha, \beta)$ est conjuguée avec la loi $\text{Gamma}(a, \theta)$ lorsque a est connu, nous savons que la densité *a posteriori* résultante sera sous une forme analytique. Ceci nous donne que la densité *a priori* pour θ est

$$\pi(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \times \theta^{\alpha-1} \exp\{-\beta\theta\} & \text{si } \theta > 0, \\ 0 & \text{sinon,} \end{cases}$$

ce qui nous mène à la densité *a posteriori* suivante :

$$\begin{aligned} \pi(\theta|\underset{\sim}{x}) &\propto \pi(\theta) \times \prod_{i=1}^{N_x} f(x_i|\theta) \\ &\propto \theta^{\alpha-1} \times \exp\{-\beta\theta\} \times \prod_{i=1}^{N_x} \frac{\theta^a}{\Gamma(a)} \times x_i^{a-1} \times \exp\{-\theta x_i\} \\ &\propto \theta^{\alpha+aN_x-1} \times \exp\{-\theta(\beta + N_x\bar{x})\}, \end{aligned}$$

où N_x est le nombre total d'observations. Nous avons donc que

$$\theta|\underset{\sim}{x} \sim \text{Gamma}\left(\alpha + aN_x, \beta + N_x\overline{x}\right).$$

En résumé, le modèle de base utilisé est

$$M_0 := \begin{cases} \theta \sim \text{Gamma}\left(\alpha, \beta\right), \\ X_i|\theta \sim \text{Gamma}\left(a, \theta\right). \end{cases} \tag{3.1.1}$$

3.2. PRÉSENTATION DES DISTANCES UTILISÉES

Pour ce travail, nous allons tenter de trouver la distance entre les densités $\pi\left(\theta|\underset{\sim}{x}\right)$ et $\pi\left(\theta|\underset{\sim}{x}, \underset{\sim}{y}\right)$. En résumé, nous allons trouver la distance entre deux densités *a posteriori* qui se distinguent seulement par l'ajout de N_y nouvelles observations, représentées par le vecteur $\underset{\sim}{y} = (y_1, y_2, ..., y_{N_y})$, au vecteur d'observations $\underset{\sim}{x}$ de taille N_x. Dans le cadre de ce mémoire, nous utiliserons trois distances. Tout d'abord, nous travaillerons avec la distance d'Hellinger.

Définition 3.2.1. *La distance d'Hellinger est définie comme étant*

$$H(f_1, f_2) = \int_\chi \left[\sqrt{f_1(x)} - \sqrt{f_2(x)}\right]^2 dx.$$

Ensuite, nous travaillerons avec la J-divergence qui utilise la *divergence de Kullback-Leibler.*

Définition 3.2.2. *La divergence de Kullback-Leibler se définit comme étant*

$$D_{KL}(f_1, f_2) = \int_\chi f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx.$$

Définition 3.2.3. *La J-divergence est définie comme ceci*

$$J(f_1, f_2) = D_{KL}\left(f_1(x), f_2(x)\right) + D_{KL}\left(f_2(x), f_1(x)\right).$$

Finalement, nous utiliserons la norme L_2.

Définition 3.2.4. *La norme* L_p *(cas général) est définie comme étant*

$$L_p\left(f_1, f_2\right) = \left[\int_\chi \left(f_1(x) - f_2(x)\right)^p dx\right]^{1/p}.$$

Dans notre cas, f_1 et f_2 seront les deux densités *a posteriori* que nous voudrons comparer. Comme nous allons étudier les propriétés des différentes distances, il est important d'expliciter ces distances à leur forme la plus simple possible. Ceci sera l'objectif de cette section.

De plus, comme nous allons le voir sous peu, les distances peuvent toutes s'exprimer en fonction de l'expression

$$W = \frac{N_y \overline{y}}{\beta + N_x \overline{x}}. \tag{3.2.1}$$

3.2.1. Distance d'Hellinger

Commençons par trouver la forme la plus simple pour la distance d'Hellinger.

Théorème 3.2.1. *Sous le modèle M_0, la distance d'Hellinger peut s'écrire de la manière suivante :*

$$\begin{aligned} &H\left(\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})\right) \\ &= 2\left[1 - \frac{\Gamma\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{\sqrt{\Gamma(\alpha + aN_x + aN_y)\Gamma(\alpha + aN_x)}} \times \frac{(1+W)^{\frac{\alpha + aN_x + aN_y}{2}}}{\left(1 + \frac{W}{2}\right)^{\alpha + aN_x + aN_y/2}}\right]. \end{aligned}$$

Avant de procéder à la démonstration, nous allons énoncer quelques lemmes qui faciliteront la compréhension de la preuve.

Lemme 3.2.1. *Lorsqu'on compare deux densités a posteriori avec la distance d'Hellinger, nous pouvons simplifier l'écriture de la distance d'Hellinger à l'expression suivante :*

$$\begin{aligned} H(\underset{\sim}{y}) &= H\left(\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})\right) \\ &= 2\left[1 - \sqrt{\frac{m(\underset{\sim}{x})}{m(\underset{\sim}{x}, \underset{\sim}{y})}} \times \int_\Theta \pi(\theta|\underset{\sim}{x}) \times \sqrt{\prod_{i=1}^{N_y} f(y_i|\theta)}\, d\theta\right]. \end{aligned}$$

DÉMONSTRATION.

$$\begin{aligned} &H(\underset{\sim}{y}) = H\left(\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})\right) \\ &= \int_\Theta \left[\sqrt{\frac{\pi(\theta)\left[\prod_{i=1}^{N_x} f(x_i|\theta)\right]\left[\prod_{i=1}^{N_y} f(y_i|\theta)\right]}{m(\underset{\sim}{x}, \underset{\sim}{y})}} - \sqrt{\frac{\pi(\theta)\left[\prod_{i=1}^{N_x} f(x_i|\theta)\right]}{m(\underset{\sim}{x})}}\right]^2 d\theta \\ &= \int_\Theta \left[\sqrt{\frac{m(\underset{\sim}{x})\left[\prod_{i=1}^{N_y} f(y_i|\theta)\right]}{m(\underset{\sim}{x}, \underset{\sim}{y})}} - 1\right]^2 \times \frac{\pi(\theta)\left[\prod_{i=1}^{N_x} f(x_i|\theta)\right]}{m(\underset{\sim}{x})} d\theta \end{aligned}$$

$$= \int_\Theta \left[\frac{m(\underset{\sim}{x}) \left[\prod_{i=1}^{N_y} f(y_i|\theta)\right]}{m(\underset{\sim}{x}, \underset{\sim}{y})} - 2\sqrt{\frac{m(\underset{\sim}{x}) \left[\prod_{i=1}^{N_y} f(y_i|\theta)\right]}{m(\underset{\sim}{x}, \underset{\sim}{y})}} + 1 \right]$$
$$\times \frac{\pi(\theta) \left[\prod_{i=1}^{N_x} f(x_i|\theta)\right]}{m(\underset{\sim}{x})} d\theta$$
$$= \int_\Theta \left[\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}) - 2\pi(\theta|\underset{\sim}{x}) \sqrt{\frac{m(\underset{\sim}{x}) \left[\prod_{i=1}^{N_y} f(y_i|\theta)\right]}{m(\underset{\sim}{x}, \underset{\sim}{y})}} + \pi(\theta|\underset{\sim}{x}) \right] d\theta$$
$$= 2\left[1 - \sqrt{\frac{m(\underset{\sim}{x})}{m(\underset{\sim}{x}, \underset{\sim}{y})}} \int_\Theta \pi(\theta|\underset{\sim}{x}) \sqrt{\prod_{i=1}^{N_y} f(y_i|\theta)} d\theta \right].$$

□

Enchaînons avec le deuxième lemme.

Lemme 3.2.2. *Sous les différentes hypothèses du modèle, nous avons que*

$$m(\underset{\sim}{x}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\left(\prod_{i=1}^{N_x} x_i\right)^{a-1}}{[\Gamma(a)]^{N_x}} \times \frac{\Gamma(\alpha + aN_x)}{(\beta + N_x\overline{x})^{\alpha+aN_x}},$$

et que

$$m(\underset{\sim}{x}, \underset{\sim}{y}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\left(\prod_{i=1}^{N_x} x_i\right)^{a-1} \left(\prod_{i=1}^{N_y} y_i\right)^{a-1}}{[\Gamma(a)]^{N_x+N_y}} \times \frac{\Gamma(\alpha + aN_x)}{(\beta + N_x\overline{x} + N_y\overline{y})^{\alpha+aN_x+aN_y}}.$$

DÉMONSTRATION.

$$m(\underset{\sim}{x}) = \int_\Theta \pi(\theta) \left[\prod_{i=1}^{N_x} f(x_i|\theta)\right] d\theta$$
$$= \int_\Theta \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp\{-\beta\theta\} \left[\frac{\theta^{aN_x}}{[\Gamma(a)]^{N_x}} \left(\prod_{i=1}^{N_x} x_i\right)^{a-1} \exp\{-\theta \times N_x\overline{x}\} \right] d\theta$$
$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\left(\prod_{i=1}^{N_x} x_i\right)^{a-1}}{[\Gamma(a)]^{N_x}} \int_\Theta \theta^{\alpha+aN_x-1} \exp\{-(\beta + N_x\overline{x})\theta\} d\theta$$
$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\left(\prod_{i=1}^{N_x} x_i\right)^{a-1}}{[\Gamma(a)]^{N_x}} \times \frac{\Gamma(\alpha + aN_x)}{(\beta + N_x\overline{x})^{\alpha+aN_x}}.$$

La deuxième partie du lemme s'obtient de façon similaire. □

Nous allons maintenant procéder à la démonstration du théorème 3.2.1.

DÉMONSTRATION. Grâce aux lemmes 3.2.1 et 3.2.2, nous avons que

$$H(\underset{\sim}{y}) = 2\left[1 - \sqrt{\frac{m(\underset{\sim}{x})}{m(\underset{\sim}{x},\underset{\sim}{y})}} \times \int_\Theta \pi(\theta|\underset{\sim}{x}) \sqrt{\prod_{i=1}^{N_y} f(y_i|\theta)}\, d\theta\right]$$

et que

$$\begin{aligned}
&\frac{m(\underset{\sim}{x})}{m(\underset{\sim}{x},\underset{\sim}{y})} \\
&= \frac{\Gamma(\alpha + aN_x)}{\Gamma(\alpha + aN_x + aN_y)} \times \frac{[\Gamma(a)]^{N_y}}{\left(\prod_{i=1}^{N_y} y_i\right)^{a-1}} \times \frac{(\beta + N_x\overline{x} + N_y\overline{y})^{\alpha+aN_x+aN_y}}{(\beta + N_x\overline{x})^{\alpha+aN_x}}.
\end{aligned}$$

De plus, en utilisant les hypothèses du modèle, nous trouvons que

$$\begin{aligned}
&\int_\Theta \pi(\theta|\underset{\sim}{x}) \sqrt{\prod_{i=1}^{N_y} f(y_i|\theta)}\, d\theta \\
&= \int_\Theta \frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{\Gamma(\alpha + aN_x)} \times \theta^{\alpha+aN_x-1} \exp\{-(\beta + N_x\overline{x})\,\theta\} \\
&\qquad \times \left[\frac{\theta^{aN_y}}{[\Gamma(a)]^{N_y}} \times \left(\prod_{i=1}^{N_y} y_i\right)^{a-1} \exp\{-\theta N_y\overline{y}\}\right]^{1/2} d\theta \\
&= \frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{\Gamma(\alpha + aN_x)} \times \frac{\left(\prod_{i=1}^{N_y} y_i\right)^{\frac{a-1}{2}}}{[\Gamma(a)]^{N_y/2}} \\
&\qquad \times \int_\Theta \theta^{\left(\alpha+aN_x+\frac{aN_y}{2}\right)-1} \exp\left\{-\theta\left(\beta + N_x\overline{x} + \frac{N_y\overline{y}}{2}\right)\right\} d\theta \\
&= \frac{\Gamma\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{\Gamma(\alpha + aN_x)} \times \frac{\left(\prod_{i=1}^{N_y} y_i\right)^{\frac{a-1}{2}}}{[\Gamma(a)]^{N_y/2}} \times \frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{(\beta + N_x\overline{x} + \frac{N_y\overline{y}}{2})^{\alpha+aN_x+\frac{aN_y}{2}}}.
\end{aligned}$$

Avec tous ces éléments mis ensemble, nous obtenons

$$\begin{aligned}
&\sqrt{\frac{m(\underset{\sim}{x})}{m(\underset{\sim}{x},\underset{\sim}{y})}} \times \int_\Theta \pi(\theta|\underset{\sim}{x}) \sqrt{\left[\prod_{i=1}^{N_y} f(y_i|\theta)\right]}\, d\theta \\
&= \left[\frac{\Gamma(\alpha + aN_x)}{\Gamma(\alpha + aN_x + aN_y)} \times \frac{[\Gamma(a)]^{N_y}}{\left(\prod_{i=1}^{N_y} y_i\right)^{a-1}} \times \frac{(\beta + N_x\overline{x} + N_y\overline{y})^{\alpha+aN_x+aN_y}}{(\beta + N_x\overline{x})^{\alpha+aN_x}}\right]^{1/2}
\end{aligned}$$

$$\times \frac{\Gamma\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{\Gamma(\alpha + aN_x)} \times \frac{\left(\prod_{i=1}^{N_y} y_i\right)^{\frac{a-1}{2}}}{[\Gamma(a)]^{N_y/2}} \times \frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{(\beta + N_x\overline{x} + \frac{N_y\overline{y}}{2})^{\alpha+aN_x+\frac{aN_y}{2}}}$$

$$= \frac{\Gamma\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{\sqrt{\Gamma(\alpha + aN_x + aN_y)\Gamma(\alpha + aN_x)}} \times \frac{\left[1 + \frac{N_y\overline{y}}{\beta+N_x\overline{x}}\right]^{\frac{\alpha+aN_x+aN_y}{2}}}{\left[1 + \frac{N_y\overline{y}}{2(\beta+N_x\overline{x})}\right]^{\alpha+aN_x+\frac{aN_y}{2}}},$$

ce qui nous amène à la forme que nous utiliserons. □

3.2.2. J-divergence

Par définition, nous avons que la J-divergence est définie comme étant

$$J\left(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})\right) = D_{KL}\left(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})\right) + D_{KL}\left(\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})\right).$$

Le théorème suivant, donne la forme simplifiée de la J-divergence sous le modèle M_0.

Théorème 3.2.2. *Sous les différentes hypothèses du modèle M_0, l'expression de la J-divergence peut s'écrire comme*

$$\begin{aligned} &J(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})) \\ &= aN_y\left\{\psi(\alpha + aN_x + aN_y) - \psi(\alpha + aN_x) - \log(1 + W)\right\} \\ &\quad + \frac{W}{1+W}\left[W(\alpha + aN_x) - aN_y\right]. \end{aligned}$$

Afin de procéder à la démonstration de ce théorème, nous allons énoncer quelques résultats qui en faciliteront la démonstration.

Lemme 3.2.3. *Sous le modèle M_0, la divergence de Kullback-Leibler est de la forme suivante :*

$$\begin{aligned} D_{KL}(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})) &= \log\left(\frac{m(\underset{\sim}{x}, \underset{\sim}{y})}{m(\underset{\sim}{x})}\right) \\ &\quad - \left\{aN_y\left[\psi(\alpha + aN_x) - \log(\beta + N_x\overline{x})\right] - N_y\overline{y}\left[\frac{\alpha + aN_x}{\beta + N_x\overline{x}}\right]\right\}, \end{aligned}$$

et de manière similaire, nous avons

$$\begin{aligned} &D_{KL}(\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})) \\ &= \log\left(\frac{m(\underset{\sim}{x})}{m(\underset{\sim}{x}, \underset{\sim}{y})}\right) \\ &+ \left\{aN_y\left[\psi(\alpha + aN_x + aN_y) - \log(\beta + N_x\overline{x} + N_y\overline{y})\right] - N_y\overline{y}\left[\frac{\alpha + aN_x + aN_y}{\beta + N_x\overline{x} + N_y\overline{y}}\right]\right\}, \end{aligned}$$

où la fonction digamma est définie par

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Pour procéder à la démonstration du lemme 3.2.3, nous aurons besoin de la proposition suivante (voir équation 4.352.1 p573, Gradshteyn et Ryzhik (2007)).

Proposition 3.2.1.

$$\int_\Theta \log(\theta) \times \theta^{\mu-1} \exp\{-\nu\theta\} \ \ d\theta = \frac{1}{\nu^\mu}\Gamma(\mu)\left[\psi(\mu) - \log(\nu)\right].$$

Passons maintenant à la démonstration du lemme 3.2.3.

DÉMONSTRATION.

$$\begin{aligned}
&D_{KL}\left(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})\right) = \int_\Theta \pi(\theta|\underset{\sim}{x}) \times \log\left(\frac{\pi(\theta|\underset{\sim}{x})}{\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})}\right) d\theta \\
&= \int_\Theta \pi(\theta|\underset{\sim}{x}) \times \log\left(\frac{\pi(\theta)\left[\prod_{i=1}^{N_x} f(x_i|\theta)\right] m(\underset{\sim}{x}, \underset{\sim}{y})}{m(\underset{\sim}{x})\,\pi(\theta)\left[\prod_{i=1}^{N_x} f(x_i|\theta)\right]\left[\prod_{i=1}^{N_y} f(y_i|\theta)\right]}\right) d\theta \\
&= \int_\Theta \pi(\theta|\underset{\sim}{x}) \left[\log\left(\frac{m(\underset{\sim}{x}, \underset{\sim}{y})}{m(\underset{\sim}{x})}\right) - \log\left(\prod_{i=1}^{N_y} f(y_i|\theta)\right)\right] d\theta \\
&= \int_\Theta \pi(\theta|\underset{\sim}{x}) \left[\log\left(\frac{m(\underset{\sim}{x}, \underset{\sim}{y})}{m(\underset{\sim}{x})}\right) - \log\left(\frac{\theta^{aN_y}}{[\Gamma(a)]^{N_y}}\left[\prod_{i=1}^{N_y} y_i\right]^{a-1} \exp\{-\theta N_y \overline{y}\}\right)\right] d\theta \\
&= \log\left(\frac{m(\underset{\sim}{x}, \underset{\sim}{y})}{m(\underset{\sim}{x})}\right) - \log\left(\frac{\left[\prod_{i=1}^{N_y} y_i\right]^{a-1}}{[\Gamma(a)]^{N_y}}\right) - aN_y E^{\pi(\theta|\underset{\sim}{x})}\left[\log(\theta)\right] + N_y \overline{y} E^{\pi(\theta|\underset{\sim}{x})}\left[\theta\right].
\end{aligned}$$

Pour pouvoir résoudre cette intégrale, nous utiliserons la proposition 3.2.1. En posant $\mu = \alpha + aN_x$ et $\nu = \beta + N_x\overline{x}$, nous obtenons :

$$\begin{aligned}
E^{\pi(\theta|\underset{\sim}{x})}\left[\log(\theta)\right] &= \frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{\Gamma(\alpha + aN_x)} \int_0^\infty \log(\theta)\theta^{\alpha+aN_x-1} \exp\{-\theta\left[\beta + N_x\overline{x}\right]\} \ \ d\theta \\
&= \frac{\nu^\mu}{\Gamma(\mu)} \int_0^\infty \log(\theta)\theta^{\mu-1} \exp\{-\nu\theta\} \ \ d\theta \\
&= \frac{\nu^\mu}{\Gamma(\mu)} \times \frac{\Gamma(\mu)}{\nu^\mu} \times \left[\psi(\nu) - \log(\mu)\right] \\
&= \psi(\nu) - \log(\mu) \\
&= \psi(\alpha + aN_x) - \log(\beta + N_x\overline{x}).
\end{aligned}$$

Nous avons donc que

$$E^{\pi(\theta|\underset{\sim}{x})}\left[aN_y \log(\theta) - \theta N_y \overline{y}\right] = aN_y\left[\psi(\alpha + aN_x) - \log(\beta + N_x\overline{x})\right] - N_y\overline{y}\left[\frac{\alpha + aN_x}{\beta + N_x\overline{x}}\right].$$

Nous pouvons maintenant obtenir la divergence de Kullback-Leibler :

$$\begin{aligned} &D_{KL}(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})) \\ &= \log\left(\frac{m(\underset{\sim}{x}, \underset{\sim}{y})}{m(\underset{\sim}{x})}\right) - \left[aN_y\left[\psi(\alpha + N_x) - \log(\beta + N_x\overline{x})\right] - N_y\overline{y}\left\{\frac{\alpha + aN_x}{\beta + N_x\overline{x}}\right\}\right]. \end{aligned}$$

La preuve pour la deuxième partie du lemme est similaire à la première partie. □

Passons maintenant à la démonstration du théorème 3.2.2.

DÉMONSTRATION. Par définition, nous savons que

$$\begin{aligned} J(\underset{\sim}{y}) &= J(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})) \\ &= D_{KL}(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})) + D_{KL}(\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})). \end{aligned}$$

À l'aide du lemme 3.2.3, nous obtenons

$$\begin{aligned} &J(\underset{\sim}{y}) \\ &= \left\{aN_y\left[\psi(\alpha + aN_x + aN_y) - \log(\beta + N_x\overline{x} + N_y\overline{y})\right] - N_y\overline{y}\left[\frac{\alpha + aN_x + aN_y}{\beta + N_x\overline{x} + N_y\overline{y}}\right]\right\} \\ &\quad - \left\{aN_y\left[\psi(\alpha + aN_x) - \log(\beta + N_x\overline{x})\right] - N_y\overline{y}\left[\frac{\alpha + aN_x}{\beta + N_x\overline{x}}\right]\right\} \\ &= aN_y\left[\psi(\alpha + aN_x + aN_y) - \psi(\alpha + aN_x) + \log(\beta + N_x\overline{x})\right. \\ &\quad \left. - \log(\beta + N_x\overline{x} + N_y\overline{y})\right] + N_y\overline{y}\left[\frac{\alpha + aN_x}{\beta + N_x\overline{x}} - \frac{\alpha + aN_x + aN_y}{\beta + N_x\overline{x} + N_y\overline{y}}\right] \\ &= aN_y\left\{\psi(\alpha + aN_x + aN_y) - \psi(\alpha + aN_x) - \log(1 + W)\right\} \\ &\quad + \frac{W}{1 + W}\left[W(\alpha + aN_x) - aN_y\right]. \end{aligned}$$

□

3.2.3. Norme L_2

Par définition, nous avons que la norme L_2 est définie comme étant

$$L_2(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})) = \sqrt{\int_\Theta \left[\pi(\theta|\underset{\sim}{x}) - \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})\right]^2 d\theta}.$$

Dans le but de simplifier l'écriture, nous allons étudier l'expression de L_2^2. Le théorème suivant, donne la forme simplifiée de l'expression L_2^2 sous les hypothèses du modèle M_0.

Théorème 3.2.3. *Sous le modèle M_0, voici le résultat simplifié de la norme L_2^2*

$$\begin{aligned}
&L_2^2(\pi(\theta|\underset{\sim}{x}), \pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})) \\
&= \frac{(\beta + N_x\overline{x})}{2^{2\alpha+2aN_x-1}} \left[\frac{\Gamma(2\alpha + 2aN_x - 1)}{[\Gamma(\alpha + aN_x)]^2} - \frac{\Gamma(2\alpha + 2aN_x + aN_y - 1)}{2^{aN_y-1}\Gamma(\alpha + aN_x)\Gamma(\alpha + aN_x + aN_y)} \right. \\
&\quad \left. \times \frac{(1+W)^{\alpha+aN_x+aN_y}}{\left(1+\frac{W}{2}\right)^{2\alpha+2aN_x+aN_y-1}} + \frac{\Gamma(2\alpha + 2aN_x + 2aN_y - 1)}{2^{2aN_y}\left[\Gamma(\alpha + aN_x + aN_y)\right]^2} (1+W) \right].
\end{aligned}$$

Avant de passer à la démonstration du théorème 3.2.3, voici trois lemmes qui simplifieront grandement l'aspect algébrique de la preuve.

Lemme 3.2.4. *Sous les différentes hypothèses du modèle M_0, nous avons que*

$$\begin{aligned}
&\int_\Theta \pi(\theta|\underset{\sim}{x})\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}) \ d\theta \\
&= \frac{\Gamma\left(2\alpha + 2aN_x + aN_y - 1\right)}{\Gamma\left(\alpha + aN_x\right)\Gamma\left(\alpha + aN_x + aN_y\right)} \times \frac{(\beta + N_x\overline{x})}{2^{2\alpha+2aN_x+aN_y-1}} \times \frac{(1+W)^{\alpha+aN_x+aN_y}}{\left(1+\frac{W}{2}\right)^{2\alpha+2aN_x+aN_y-1}}.
\end{aligned}$$

DÉMONSTRATION.

$$\begin{aligned}
&\int_\Theta \pi(\theta|\underset{\sim}{x})\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}) \ d\theta \\
&= \int_\Theta \left(\frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{\Gamma(\alpha + aN_x)} \right) \left(\frac{(\beta + N_x\overline{x} + N_y\overline{y})^{\alpha+aN_x+aN_y}}{\Gamma(\alpha + aN_x + aN_y)} \right) \theta^{2\alpha+2aN_x+aN_y-2} \\
&\quad \times \exp\{-\theta(2\beta + 2N_x\overline{x} + N_y\overline{y})\} \ d\theta \\
&= \left(\frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{\Gamma(\alpha + aN_x)} \right) \left(\frac{(\beta + N_x\overline{x} + N_y\overline{y})^{\alpha+aN_x+aN_y}}{\Gamma(\alpha + aN_x + aN_y)} \right) \\
&\quad \times \frac{\Gamma(2\alpha + 2aN_x + aN_y - 1)}{(2\beta + 2N_x\overline{x} + N_y\overline{y})^{2\alpha+2aN_x+aN_y-1}} \\
&= \frac{\Gamma\left(2\alpha + 2aN_x + aN_y - 1\right)}{\Gamma\left(\alpha + aN_x\right)\Gamma\left(\alpha + aN_x + aN_y\right)} \times \frac{(\beta + N_x\overline{x})}{2^{2\alpha+2aN_x+aN_y-1}} \times \frac{(1+W)^{\alpha+aN_x+aN_y}}{\left(1+\frac{W}{2}\right)^{2\alpha+2aN_x+aN_y-1}}.
\end{aligned}$$

□

Lemme 3.2.5. *Sous les différentes hypothèses du modèle M_0, nous avons que*

$$\int_\Theta \pi^2(\theta|\underset{\sim}{x})d\theta = \frac{\Gamma(2\alpha + 2aN_x - 1)}{[\Gamma(\alpha + aN_x)]^2} \times \frac{(\beta + N_x\overline{x})}{2^{2\alpha+2aN_x-1}}.$$

DÉMONSTRATION. Il suffit de poser $N_y = 0$ dans le lemme 3.2.4 et nous obtenons le résultat. □

Lemme 3.2.6. *Sous les différentes hypothèses du modèle M_0, nous avons que*

$$\int_\Theta \pi^2(\theta|\underset{\sim}{x},\underset{\sim}{y})\,d\theta = \frac{\Gamma(2\alpha+2aN_x+2aN_y-1)}{[\Gamma(\alpha+aN_x+aN_y)]^2}\times\frac{(\beta+N_x\overline{x})}{2^{2\alpha+2aN_x+2aN_y-1}}\times(1+W)\,.$$

DÉMONSTRATION. La preuve de ce lemme découle directement du résultat énoncé dans le lemme 3.2.5. □

Passons maintenant à la démonstration du théorème 3.2.3.

DÉMONSTRATION. Par définition, nous savons que la norme L_2^2 s'écrit comme ceci :

$$\begin{aligned}
L_2^2(\pi(\theta|\underset{\sim}{x}),\pi(\theta|\underset{\sim}{x},\underset{\sim}{y})) &= \int_\Theta \left[\pi(\theta|\underset{\sim}{x})-\pi(\theta|\underset{\sim}{x},\underset{\sim}{y})\right]^2 d\theta\\
&= \int_\Theta \pi^2(\theta|\underset{\sim}{x}) - 2\pi(\theta|\underset{\sim}{x})\pi(\theta|\underset{\sim}{x},\underset{\sim}{y}) + \pi^2(\theta|\underset{\sim}{x},\underset{\sim}{y})\,d\theta\\
&= \int_\Theta \pi^2(\theta|\underset{\sim}{x})\,d\theta - 2\int_\Theta \pi(\theta|\underset{\sim}{x})\pi(\theta|\underset{\sim}{x},\underset{\sim}{y})\,d\theta + \int_\Theta \pi^2(\theta|\underset{\sim}{x},\underset{\sim}{y})\,d\theta.
\end{aligned}$$

Remarquons que chaque intégrale est déjà résolue par l'un des lemmes précédents. De ce fait, nous avons que

$$\begin{aligned}
&L_2^2(\pi(\theta|\underset{\sim}{x}),\pi(\theta|\underset{\sim}{x},\underset{\sim}{y})) =\\
&\frac{\Gamma(2\alpha+2aN_x-1)}{[\Gamma(\alpha+aN_x)]^2}\times\frac{(\beta+N_x\overline{x})}{2^{2\alpha+2aN_x-1}}\\
&\quad-\frac{2\Gamma\left(2\alpha+2aN_x+aN_y-1\right)}{\Gamma\left(\alpha+aN_x\right)\Gamma\left(\alpha+aN_x+aN_y\right)}\times\frac{(\beta+N_x\overline{x})}{2^{2\alpha+2aN_x+aN_y-1}}\times\frac{(1+W)^{\alpha+aN_x+aN_y}}{\left(1+\frac{W}{2}\right)^{2\alpha+2aN_x+aN_y-1}}\\
&\quad+\frac{\Gamma(2\alpha+2aN_x+2aN_y-1)}{[\Gamma(\alpha+aN_x+aN_y)]^2}\times\frac{(\beta+N_x\overline{x})}{2^{2\alpha+2aN_x+2aN_y-1}}\times(1+W)\\
&=\frac{(\beta+N_x\overline{x})}{2^{2\alpha+2aN_x-1}}\left[\frac{\Gamma(2\alpha+2aN_x-1)}{[\Gamma(\alpha+aN_x)]^2}-\frac{\Gamma(2\alpha+2aN_x+aN_y-1)}{2^{aN_y-1}\Gamma(\alpha+aN_x)\Gamma(\alpha+aN_x+aN_y)}\right.\\
&\quad\left.\times\frac{(1+W)^{\alpha+aN_x+aN_y}}{\left(1+\frac{W}{2}\right)^{2\alpha+2aN_x+aN_y-1}}+\frac{\Gamma(2\alpha+2aN_x+2aN_y-1)}{2^{2aN_y}\left[\Gamma(\alpha+aN_x+aN_y)\right]^2}(1+W)\right].
\end{aligned}$$

□

3.3. DENSITÉ PRÉDICTIVE

Comme nous avons pu le remarquer, chacune des distances peut s'exprimer en fonction de W. Dans cette section, nous allons donc étudier cette variable avec un peu plus d'attention. Plus précisément, nous allons trouver la densité prédictive de W. Afin de trouver cette densité, nous devons commencer par trouver la densité prédictive conjointe des $\underset{\sim}{y}$ et ensuite utiliser l'approche par changement de variable pour trouver la densité prédictive de la variable W.

Voici un lemme qui simplifiera la démarche menant à la densité prédictive.

Lemme 3.3.1. *Sous les différentes hypothèses du modèle* M_0 *et en supposant que* $t_0 = 0$*, nous avons que*

$$\int_0^{t_{N_y}} \int_0^{t_{N_y-1}} \dots \int_0^{t_3} \int_0^{t_2} \left[\prod_{i=1}^{N_y} t_i - t_{i-1} \right]^{a-1} \times \prod_{i=1}^{N_y-1} dt_i = \frac{t_{N_y}^{aN_y-1} \left[\Gamma(a)\right]^{N_y}}{\Gamma(aN_y)}.$$

DÉMONSTRATION. En posant

$$U_i = \frac{T_i}{T_{i+1}},$$

nous avons que

$$T_\ell = T_{N_y} \times \prod_{j=\ell}^{N_y-1} U_j$$

et que

$$\begin{aligned} T_\ell - T_{\ell-1} &= T_\ell - T_\ell U_{\ell-1} \\ &= T_\ell [1 - U_{\ell-1}] \\ &= T_{N_y} [1 - U_{\ell-1}] \prod_{j=\ell}^{N_y-1} U_j. \end{aligned}$$

De plus, nous obtenons que le jacobien de cette transformation est

$$\begin{aligned} |J|^{-1} &= \prod_{i=2}^{N_y} T_i \\ &= \prod_{i=2}^{N_y} T_{N_y} \prod_{j=i}^{N_y-1} U_j \\ &= T_{N_y}^{N_y-1} \prod_{i=2}^{N_y} \prod_{j=i}^{N_y-1} U_j, \end{aligned}$$

où $\prod_{i=N_y}^{N_y-1} u_i = 1$, selon notre notation. Ceci implique que

$$\begin{aligned}
&\int_0^{t_{N_y}} \int_0^{t_{N_y-1}} \dots \int_0^{t_3} \int_0^{t_2} \left[\prod_{i=1}^{N_y} t_i - t_{i-1}\right]^{a-1} \times \prod_{i=1}^{N_y-1} dt_i \\
&= \int_0^1 \int_0^1 \dots \int_0^1 \int_0^1 t_{N_y}^{aN_y-N_y} \left[\prod_{i=1}^{N_y} \prod_{j=i}^{N_y-1} u_j(1-u_{i-1})\right]^{a-1} \times \left[t_{N_y}^{N_y-1} \prod_{i=2}^{N_y} \prod_{j=i}^{N_y-1} u_j\right] \prod_{i=1}^{N_y-1} du_i \\
&= \int_0^1 \int_0^1 \dots \int_0^1 \int_0^1 t_{N_y}^{aN_y-1} \left[\prod_{i=1}^{N_y} \prod_{j=i}^{N_y-1} u_j^{a-1}(1-u_{i-1})^{a-1}\right] \times \left[\frac{\prod_{i=1}^{N_y} \prod_{j=i}^{N_y-1} u_j}{\prod_{j=1}^{N_y-1} u_j}\right] \prod_{i=1}^{N_y-1} du_i \\
&= t_{N_y}^{aN_y-1} \int_0^1 \int_0^1 \dots \int_0^1 \int_0^1 \frac{\prod_{i=1}^{N_y} \prod_{j=i}^{N_y-1} u_j^a (1-u_{i-1})^{a-1}}{\prod_{j=1}^{N_y-1} u_j} \prod_{i=1}^{N_y-1} du_i \\
&= t_{N_y}^{aN_y-1} \int_0^1 \int_0^1 \dots \int_0^1 \int_0^1 \frac{u_1^a u_2^{2a} u_3^{3a} \cdots u_{N_y-1}^{a(N_y-1)}}{u_1 u_2 u_3 \cdots u_{N_y-1}} \\
&\qquad \times (1-u_1)^{a-1}(1-u_2)^{a-1}(1-u_3)^{a-1} \cdots (1-u_{N_y-1})^{a-1} \prod_{i=1}^{N_y-1} du_i \\
&= t_{N_y}^{aN_y-1} \times \frac{\Gamma(a)\Gamma(a)}{\Gamma(2a)} \times \frac{\Gamma(2a)\Gamma(a)}{\Gamma(3a)} \times \dots \times \frac{\Gamma(aN_y-2a)\Gamma(a)}{\Gamma(aN_y-a)} \times \frac{\Gamma(aN_y-a)\Gamma(a)}{\Gamma(aN_y)} \\
&= t_{N_y}^{aN_y-1} \times \frac{[\Gamma(a)]^{N_y}}{\Gamma(aN_y)}.
\end{aligned}$$

□

Voici la densité prédictive conjointe des N_y observations regroupées dans le vecteur $\underset{\sim}{y}$.

$$\begin{aligned}
m(\underset{\sim}{y}|\underset{\sim}{x}) &= E^{\pi(\theta|\underset{\sim}{x})} \left[\prod_{i=1}^{N_y} f(y_i|\theta, \underset{\sim}{x})\right] \\
&= \int_\Theta \left[\prod_{i=1}^{N_y} \frac{\theta^a}{\Gamma(a)} \times y_i^{a-1} \exp\{-\theta y_i\}\right] \\
&\quad \times \left[\frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{\Gamma(\alpha+aN_x)} \times \theta^{\alpha+aN_x-1} \times \exp\{-\theta(\beta+N_x\overline{x})\}\right] d\theta \\
&= \left[\prod_{i=1}^{N_y} y_i\right]^{a-1} \frac{(\beta+N_x\overline{x})^{\alpha+aN_x}}{[\Gamma(a)]^{N_y}\,\Gamma(\alpha+aN_x)} \\
&\quad \times \int_\Theta \theta^{\alpha+aN_x+aN_y-1} \times \exp\{-\theta(\beta+N_x\overline{x}+N_y\overline{y})\}\, d\theta
\end{aligned}$$

$$= \left[\prod_{i=1}^{N_y} y_i\right]^{a-1} \frac{(\beta + N_x\overline{x})^{\alpha+aN_x}}{[\Gamma(a)]^{N_y}\,\Gamma(\alpha+aN_x)} \times \frac{\Gamma(\alpha+aN_x+aN_y)}{(\beta+N_x\overline{x}+N_y\overline{y})^{\alpha+aN_x+aN_y}}.$$

Procédons aux changements de variable suivants :

$$T_i = \sum_{j=1}^{i} Y_j \quad , i = 1, ..., N_y.$$

Il est facile de vérifier que le jacobien de cette transformation est 1. Nous avons donc

$$f(\underset{\sim}{t}|\underset{\sim}{x}) = \frac{\Gamma(\alpha+aN_x+aN_y)}{[\Gamma(a)]^{N_y}\,\Gamma(\alpha+aN_x)} \times \frac{(\beta+N_x\overline{x})^{\alpha+aN_x}}{(\beta+N_x\overline{x}+t_{N_y})^{\alpha+aN_x+aN_y}} \times \left[\prod_{i=1}^{N_y}(t_i - t_{i-1})\right]^{a-1},$$

où $t_0 = 0$. Grâce à l'équation ci-haut ainsi qu'au lemme 3.3.1, nous pouvons trouver la densité prédictive de T_{N_y} :

$$\begin{aligned}
f(t_{N_y}|\underset{\sim}{x}) &= \int_0^{t_{N_y}} \int_0^{t_{N_y-1}} ... \int_0^{t_3}\int_0^{t_2} f(\underset{\sim}{t}|\underset{\sim}{x}) \times \prod_{i=1}^{N_y-1} dt_i \\
&= K\int_0^{t_{N_y}} \int_0^{t_{N_y-1}} ... \int_0^{t_3}\int_0^{t_2} \left[\prod_{i=1}^{N_y}(t_i - t_{i-1})\right]^{a-1} \times \prod_{i=1}^{N_y-1} dt_i \\
&= K \times t_{N_y}^{aN_y-1} \times \frac{[\Gamma(a)]^{N_y}}{\Gamma(aN_y)} \\
&= \frac{\Gamma(\alpha+aN_x+aN_y)}{[\Gamma(a)]^{N_y}\,\Gamma(\alpha+aN_x)} \times \frac{(\beta+N_x\overline{x})^{\alpha+aN_x}}{(\beta+N_x\overline{x}+t_{N_y})^{\alpha+aN_x+aN_y}} \times t_{N_y}^{aN_y-1} \times \frac{[\Gamma(a)]^{N_y}}{\Gamma(aN_y)} \\
&= \frac{\Gamma(\alpha+aN_x+aN_y)}{\Gamma(\alpha+aN_x)\Gamma(aN_y)} \times \frac{(\beta+N_x\overline{x})^{\alpha+aN_x}}{(\beta+N_x\overline{x}+t_{N_y})^{\alpha+aN_x+aN_y}} \times t_{N_y}^{aN_y-1} \\
&= \frac{\Gamma(\alpha+aN_x+aN_y)}{\Gamma(\alpha+aN_x)\Gamma(aN_y)} \times \frac{1}{(\beta+N_x\overline{x})} \times \frac{\left(\frac{t_{N_y}}{\beta+N_x\overline{x}}\right)^{aN_y-1}}{\left(1+\frac{t_{N_y}}{\beta+N_x\overline{x}}\right)^{\alpha+aN_x+aN_y}},
\end{aligned}$$

où

$$K = \frac{\Gamma(\alpha+aN_x+aN_y)}{[\Gamma(a)]^{N_y}\,\Gamma(\alpha+aN_x)} \times \frac{(\beta+N_x\overline{x})^{\alpha+aN_x}}{(\beta+N_x\overline{x}+t_{N_y})^{\alpha+aN_x+aN_y}}.$$

Faisons un dernier changement de variable et posons

$$W = \frac{T_{N_y}}{\beta+N_x\overline{x}},$$

ceci nous donne

$$dW = \frac{dT_{N_y}}{\beta+N_x\overline{x}},$$

et la densité prédictive de W sera alors

$$f(w|\underset{\sim}{x}) = \frac{\Gamma(\alpha + aN_x + aN_y)}{\Gamma(\alpha + aN_x)\,\Gamma(aN_y)} \times \frac{w^{aN_y-1}}{(1+w)^{\alpha+aN_x+aN_y}}.$$

Définition 3.3.1. *Si X est de loi bêta prime généralisée de paramètres α, β, p, q, notée Beta'(α, β, p, q), alors X a la densité suivante*

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{p}{q} \times \frac{\left(\frac{x}{q}\right)^{\alpha p-1}}{\left(1+\left(\frac{x}{q}\right)^{p}\right)^{\alpha+\beta}} & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases}$$

De plus, dans le cas où $p = 1$ et $q = 1$, cette loi se nomme bêta prime standard et elle est tout simplement notée par Beta'(α, β). Dans ce cas, si $X \sim$ Beta'(α, β), la densité de X peut se simplifier à

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{x^{\alpha-1}}{(1+x)^{\alpha+\beta}} & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases}$$

Finalement, dans le cas particulier de la loi bêta prime standard, nous avons que $E[X] = \frac{\alpha}{\beta-1}$ pour $\beta > 1$ et $Var[X] = \frac{\alpha(\alpha+\beta-1)}{(\beta-2)(\beta-1)^2}$ pour $\beta > 2$.

Nous obtenons donc que la variable $W|\underset{\sim}{x} = \frac{\sum_{i=1}^{N_y} Y_i}{\beta+N_x\bar{x}} \sim \text{Beta}'(aN_y, \alpha + aN_x)$. Ce résultat est très intéressant, car si les distances sont monotones croissantes en fonction de W, nous aurons que W est une variable qui résume l'information apportée par les distances. De ce fait, nous pourrions laisser tomber l'utilisation des distances et simplement travailler avec la variable W dont la densité est sous une forme analytique.

3.4. Étude de monotonicité des différentes distances en fonction de W

Cette section sera consacrée à l'étude de la monotonicité des différentes distances en fonction de W.

3.4.1. Distance d'Hellinger

Pour la distance d'Hellinger, si nous posons

$$A = \frac{\alpha + aN_x + aN_y}{2},$$

$$B = \alpha + aN_x + \frac{aN_y}{2},$$

$$C = \frac{\Gamma\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{\sqrt{\Gamma(\alpha + aN_x + aN_y)\,\Gamma(\alpha + aN_x))}},$$

nous obtenons, à l'aide du théorème 3.2.1, l'expression suivante :

$$H(W) = 2\left[1 - \frac{(1+W)^A}{\left(1+\frac{W}{2}\right)^B} \times C\right].$$

Nous allons maintenant étudier la croissance de la fonction $H(\cdot)$ en fonction de W.

$$\begin{aligned} H'(W) &= -2C \times \frac{\left[A\,(1+W)^{A-1}\left(1+\frac{W}{2}\right)^B - (1+W)^A\left(1+\frac{W}{2}\right)^{B-1} \times \frac{B}{2}\right]}{\left[1+\frac{W}{2}\right]^{2B}} \\ &= -2C \times \frac{(1+W)^{A-1}}{\left[1+\frac{W}{2}\right]^{B+1}} \times \left[A\left(1+\frac{W}{2}\right) - \frac{B}{2}(1+W)\right]. \end{aligned}$$

En posant

$$g(W) = \left[A\left(1+\frac{W}{2}\right) - \frac{B}{2}(1+W)\right],$$

nous avons que

$$\begin{aligned} H'(W) > 0 &\iff g(W) < 0 \\ &\iff A + \frac{AW}{2} - \frac{B}{2} - \frac{BW}{2} < 0 \\ &\iff W\left(\frac{A}{2} - \frac{B}{2}\right) + \left(A - \frac{B}{2}\right) < 0 \\ &\iff \frac{-(\alpha + aN_x)\,W}{4} + \frac{aN_y}{4} < 0 \\ &\iff W > \frac{aN_y}{\alpha + aN_x}, \end{aligned}$$

car

$$\begin{aligned} A - B &= \frac{\alpha + aN_x + aN_y}{2} - \left(\alpha + aN_x + \frac{aN_y}{2}\right) \\ &= -\frac{(\alpha + N_x)}{2} \end{aligned}$$

et que

$$\begin{aligned} A - \frac{B}{2} &= \frac{\alpha + aN_x + aN_y}{2} - \frac{\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{2} \\ &= \frac{aN_y}{4}. \end{aligned}$$

Finalement, nous avons que la distance d'Hellinger est croissante lorsque $W > \frac{aN_y}{\alpha + aN_x}$. Passons maintenant à la J-divergence.

3.4.2. J-divergence

Pour la J-divergence, si nous posons

$$A = \alpha + aN_x,$$
$$B = \alpha + aN_x + aN_y,$$

nous obtenons, à l'aide du théorème 3.2.2, l'expression suivante :

$$J(W) = (B - A)\left[\psi(B) - \psi(A) - \log(1 + W)\right] + W\left[A - \frac{B}{1 + W}\right].$$

Nous allons maintenant étudier la croissance de la fonction $J(\cdot)$ en fonction de W.

$$\begin{aligned} J'(W) = \frac{A - B}{1 + W} + \left(A - \frac{B}{1 + W}\right) + W\left(\frac{B}{(1 + W)^2}\right) > 0 \\ \iff (A - B)(1 + W) + A(1 + W)^2 - B(1 + W) + WB > 0 \\ \iff W^2[A] + W[A - B + 2A - B + B] + A - B + A - B > 0 \\ \iff W^2[A] + W[3A - B] + 2[A - B] > 0. \end{aligned}$$

Nous devons donc résoudre l'inégalité pour un polynôme de degré deux.

$$\begin{aligned} W &= \frac{-(3A - B) \pm \sqrt{(3A - B)^2 - 4A \times 2(A - B)}}{2A} \\ &= \frac{-(3A - B) \pm \sqrt{(A + B)^2}}{2A} \\ &= \frac{-3A + B) \pm (A + B)}{2A}. \end{aligned}$$

Nous avons les solutions

$$\begin{aligned} W_1 &= \frac{-A + B}{A} \\ &= \frac{aN_y}{\alpha + aN_x} \end{aligned}$$

et

$$\begin{aligned} W_2 &= \frac{-4A}{2A} \\ &= -2. \end{aligned}$$

Notons que la solution W_2 est à rejeter, car elle est négative, ce qui n'est pas possible vu la forme de W. Finalement, nous avons que la J-divergence est

croissante lorsque $W > \frac{aN_y}{\alpha+aN_x}$. Remarquons que nous avons obtenu le même résultat que pour la distance d'Hellinger. Finalement, passons à la norme L_2.

3.4.3. Norme L_2

Dans le cas de la norme L_2, si nous posons

$$\begin{aligned} A &= \alpha + aN_x, \\ B &= \beta + N_x\overline{x}, \\ c_1 &= \frac{\Gamma(2A + aN_y - 1)}{\Gamma(A)\,\Gamma(A + aN_y)\,2^{aN_y-1}} \end{aligned}$$

et

$$c_2 = \frac{\Gamma(2A + aN_y - 1)}{[\Gamma(A + aN_y)]^2\,2^{2aN_y}},$$

nous obtenons, à l'aide du théorème 3.2.3, l'expression suivante :

$$L_2^2(W) = \frac{B}{2^{2A-1}}\left[\frac{\Gamma(2A-1)}{[\Gamma(A)]^2} - c_1\frac{(1+W)^{A+aN_y}}{\left(1+\frac{W}{2}\right)^{2A+aN_y-1}} + c_2(1+W)\right].$$

Nous avons donc que la dérivé est

$$\begin{aligned} L_2^2{}'(W) =& \frac{B}{2^{2A-1}}\left[\frac{-c_1}{\left(1+\frac{W}{2}\right)^{2(2A+aN_y-1)}}\left\{(A+aN_y)(1+W)^{A+aN_y-1}\left(1+\frac{W}{2}\right)^{2A+aN_y-1}\right.\right. \\ &\left.\left. - \left(\frac{2A+aN_y-1}{2}\right)\left(1+\frac{W}{2}\right)^{2A+aN_y-2}(1+W)^{A+aN_y}\right\} + c_2\right] \\ =& \frac{B}{2^{2A-1}}\left[\frac{-c_1(1+W)^{A+aN_y-1}\left(1+\frac{W}{2}\right)^{2A+aN_y-2}}{\left(1+\frac{W}{2}\right)^{2(2A+aN_y-1)}}\right. \\ &\left. \times\left\{(A+aN_y)\left(1+\frac{W}{2}\right) - \frac{(2A+aN_y-1)}{2}(1+W)\right\} + c_2\right] \\ =& \frac{B}{2^{2A-1}}\left[\frac{-c_1(1+W)^{A+aN_y-1}}{\left(1+\frac{W}{2}\right)^{2A+aN_y}} \times \frac{1}{2}\{aN_y + 1 + W(1-A)\} + c_2\right]. \end{aligned}$$

La norme est croissante en fonction de W

$$\begin{aligned} &\iff L_2^2{}'(W) > 0 \\ &\iff \frac{-c_1(1+W)^{A+aN_y-1}}{\left(1+\frac{W}{2}\right)^{2A+aN_y}} \times \frac{1}{2}\{aN_y + 1 + W(1-A)\} + c_2 > 0 \\ &\iff \frac{(1+W)^{A+aN_y-1}}{\left(1+\frac{W}{2}\right)^{2A+aN_y}} \times \frac{1}{2}\{aN_y + 1 + W(1-A)\} < \frac{c_2}{c_1} \end{aligned}$$

$$\iff \frac{(1+W)^{A+aN_y-1}}{\left(1+\frac{W}{2}\right)^{2A+aN_y}} \times \frac{1}{2}\{aN_y+1+W(1-A)\} < \frac{\Gamma(A)}{2^{aN_y+1}\Gamma(A+aN_y)}.$$

De plus, étant donné que

$$\frac{c_2}{c_1} = \frac{\Gamma(A)}{2^{aN_y+1}\Gamma(A+aN_y)} \approx 0$$

et que $1-A<0$, nous avons que

$$\begin{aligned} & L_2^2{}'(W) > 0 \\ \iff & aN_y+1+W(1-A) < 0 \\ \iff & W > \frac{aN_y+1}{A-1} \\ \iff & W > \frac{aN_y+1}{\alpha+aN_x-1}. \end{aligned}$$

Afin d'illustrer le fait que

$$\begin{aligned} \frac{c_2}{c_1} &= \frac{\Gamma(A)}{2^{aN_y+1}\Gamma(A+aN_y)} \\ &= \frac{\Gamma(\alpha+aN_x)}{2^{aN_y+1}\Gamma(\alpha+aN_x+aN_y)} \\ &\approx 0, \end{aligned}$$

notons que $N_y \geq 55$. En effet, nous allons toujours au moins prendre une année d'observations de comparaison ce qui constitue environ 55 observations.

Finalement, nous avons que $L_2^2(W)$ est croissant en $W \iff W > \frac{aN_y+1}{\alpha+aN_x-1}$. Ceci est très proche des résultats trouvés pour les deux autres distances.

3.4.4. Retour sur la croissance des distances en fonction de W

Nous avons précédemment trouvé les intervalles de croissance des différentes distances. Pour ce projet, nous sommes intéressés à détecter des dis-

TABLEAU 3.1. Intervalles de croissance des différentes distances en fonction de W.

Distance	Intervalle de croissance
Hellinger	$W > \frac{aN_y}{\alpha+aN_x}$
J-divergence	$W > \frac{aN_y}{\alpha+aN_x}$
Norme L_2^2	$W > \frac{aN_y+1}{\alpha+aN_x-1}$

tances qui sont grandes. Conséquemment, nous utiliserons des tests unilatéraux. La zone de rejet pour W sera donc de la forme $]W_0, \infty[$, où W_0 est telle

que

$$\mathbb{P}(W > W_0) = \alpha^*,$$

et où α^* est le niveau du test. Dans le cas où W_0 sera dans l'intervalle de croissance pour les différentes distances, nous pourrons utiliser W comme un résumé de l'information qu'apporteraient les distances. Comme $\mathbb{E}[W] = \frac{aN_y}{\alpha+aN_x-1}$ (voir la définition 3.3.1), nous pouvons aisément faire l'hypothèse que W_0 est dans l'intervalle de croissance de chaque fonction et nous pouvons donc conclure que W résume l'information apportée par les distances. Nous utiliserons donc W comme statistique de test pour le reste du mémoire.

3.5. Fenêtre mobile

Dans cette section, nous verrons que si nous comparons deux densités *a posteriori* qui sont modélisées avec un très grand nombre d'observations, alors les deux densités seront asymptotiquement identiques. En d'autres mots, les densités *a posteriori* deviennent trop stables et les distances tendent vers 0. Il devient alors impossible de détecter quoi que ce soit. L'objectif de cette section, sera de présenter et de justifier l'utilisation de fenêtres mobiles en tant que notre choix de modélisation pour faire en sorte que le nombre d'observations soit limité.

3.5.1. Mise en contexte et notations

Afin d'aider à visualiser ce que sont les fenêtres mobiles, nous pouvons aisément les comparer à l'utilisation de moyennes mobiles. En effet, lorsque nous utilisons les moyennes mobiles, nous recalculons continuellement la moyenne, qui est une statistique quelconque, pour différents échantillons contigus dans le temps. Généralement, le sous-ensemble progresse dans le temps, c'est-à-dire, l'entité la plus vieille est remplacée au profit d'une plus récente. Les fenêtres mobiles sont dans ce cas, le sous-ensemble de données qui progresse dans le temps. Donc, l'utilisation des fenêtres mobiles avec la statistique W revient à l'utilisation des moyennes mobiles sauf qu'ici, la statistique n'est pas la moyenne, mais bien W. Dans le cadre de notre mémoire, nous avons décidé d'utiliser les fenêtres mobiles afin que le nombre d'observations soit limité.

Dans les sections précédentes, nous avons développé des outils qui peuvent comparer des densités *a posteriori* de la forme $\pi(\theta|\underset{\sim}{x})$ avec $\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y})$. Notons que dans le cadre de ce projet, $\underset{\sim}{y}$ sera le vecteur d'observations qui suit le vecteur d'observations $\underset{\sim}{x}$ dans le temps. À partir de ces présupposés, plusieurs options

s'offraient à nous quant au choix de comment choisir $\underset{\sim}{x}$ et $\underset{\sim}{y}$. Nous avons décidé de balayer, à l'aide d'un petit intervalle I_x, l'étendue des observations dont nous disposons. Cet intervalle contiendra P_x années d'observations pour un total de N_x observations. Ces observations, constitueront le vecteur d'observations $\underset{\sim}{x}$. En fait, si nous notons par n_{x_i} le nombre d'observations pour l'année i, nous avons que $N_x = \sum_{i=1}^{P_x} n_{x_i}$. De plus, soit I_y un intervalle subséquent dans le temps à I_x contenant P_y années d'observations pour un total de N_y observations en tout. Ces observations, constitueront le vecteur d'observations $\underset{\sim}{y}$. Le terme fenêtre mobile désignera l'union de I_x et de I_y.

3.5.2. Procédure

Notre procédure consiste à prendre un intervalle de temps I_x de P_x années d'observations et d'évaluer l'effet de l'ajout de l'intervalle I_y de P_y années d'observations sur la densité *a posteriori*. De plus, nous avancerons la fenêtre mobile d'une année à la fois. Ceci implique que l'année la plus ancienne de I_x sera laissée de côté au profit de la plus vieille année de I_y et que la plus ancienne année de I_y sera remplacée par une nouvelle année d'observations.

Par exemple, supposons que $P_x = 10$ et que $P_y = 1$. On pourrait comparer la densité *a posteriori* qui inclut les observations des années 1961 à 1970 avec la densité *a posteriori* des années 1961 à 1971. Dans le cas de non-détection, nous continuerons en comparant la densité *a posteriori* des années 1962 à 1971 avec la densité *a posteriori* des années 1962 à 1972 et ainsi de suite.

3.5.3. Pourquoi utiliser les fenêtres mobiles.

Une question légitime serait de demander pourquoi vouloir utiliser les fenêtres mobiles alors que nous aurions pu utiliser toutes les observations. Cette question prend tout son sens, car généralement, en statistique, plus nous avons d'observations disponibles, plus nous sommes heureux. Premièrement, nous savons que P_y (le nombre d'années d'observations de prévision) devra être petit, car nous voulons pouvoir détecter rapidement la non-stationnarité. En fait, c'est le nombre d'années d'observations P_x qui n'est pas restreint. Supposons que nous avons N années d'observations disponibles, quelqu'un aurait pu proposer d'utiliser toutes les N années et de comparer les densités *a posteriori* avec P_x ($P_x >> P_y$) et $P_x + P_y$ années, où $P_x + P_y = N$. De cette façon, toutes les années sont utilisées. Toutefois, le problème est que si P_x devient trop grand comparer à P_y (qui doit être minimisé), les densités *a posteriori* deviennent trop semblables. En fait, lorsque $N_x \to \infty$ (N_y fixe), les deux densités *a posteriori* deviennent asymptotiquement égales. En effet, les N_x observations prennent

toute la place et écrasent l'information amenée par les N_y observations venant de la fenêtre de prévision. Ceci se traduira par le fait que les distances tendront vers 0 lorsque le nombre d'observations tend vers l'infini. Ceci nous mène à la proposition suivante.

Avant de passer aux différentes propositions, nous supposerons que

$$\begin{cases} E[X_i] = \mu_x < \infty \,, \\ Var[X_i] = \sigma_x^2 < \infty \,, \end{cases} \tag{3.5.1}$$

et ce $\forall i$.

Proposition 3.5.1. *Si* N_y *est fixe et sous les hypothèses* (3.5.1)*, nous avons que*

$$\lim_{N_x \to \infty} H\left(\pi(\theta|\underset{\sim}{x}, \underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})\right) = 0.$$

Afin de clarifier la preuve de la proposition 3.5.1, nous allons énoncer quelques lemmes.

Lemme 3.5.1. *Si* N_y *est maintenu fixe et sous les hypothèses* (3.5.1)*, nous avons*

$$\lim_{N_x \to \infty} \left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{\frac{\alpha + aN_x + aN_y}{2}} = \exp\left\{\frac{aN_y \overline{y}}{2\mu_x}\right\}$$

et que

$$\lim_{N_x \to \infty} \left[1 + \frac{N_y \overline{y}}{2\left(\beta + N_x \overline{x}\right)}\right]^{\alpha + aN_x + aN_y/2} = \exp\left\{\frac{aN_y \overline{y}}{2\mu_x}\right\},$$

où $\mu_x = E[X]$.

DÉMONSTRATION.

$$\begin{aligned}
&\lim_{N_x \to \infty} \left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{\frac{\alpha + aN_x + aN_y}{2}} \\
&= \lim_{N_x \to \infty} \left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{\frac{aN_x}{2}} \times \left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{\frac{\alpha + aN_y}{2}} \\
&= \lim_{N_x \to \infty} \left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{\frac{aN_x}{2}} \\
&= \lim_{N_x \to \infty} \left(\left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{N_x \overline{x}}\right)^{\frac{a}{2\overline{x}}} \times \left(\left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{\beta}\right)^{\frac{a}{2\overline{x}}} \times \left(\left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{-\beta}\right)^{\frac{a}{2\overline{x}}} \\
&= \lim_{N_x \to \infty} \left(\left[1 + \frac{N_y \overline{y}}{\beta + N_x \overline{x}}\right]^{\beta + N_x \overline{x}}\right)^{\frac{a}{2\overline{x}}} \\
&= \exp\left\{\frac{aN_y \overline{y}}{2\mu_x}\right\}.
\end{aligned}$$

La deuxième partie de la preuve est similaire. □

Lemme 3.5.2. *Si* N_y *est maintenu fixe, nous avons*

$$\lim_{N_x\to\infty} \frac{\Gamma\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{\sqrt{\Gamma\left(\alpha + aN_x + aN_y\right)\Gamma\left(\alpha + aN_x\right)}} = 1.$$

DÉMONSTRATION. Ceci s'obtient en utilisant l'approximation de Stirling ainsi que les propriétés des limites. □

Passons maintenant à la démonstration de la proposition 3.5.1

DÉMONSTRATION. Tout d'abord, grâce au théorème 3.2.1, nous savons que

$$\begin{aligned}
&H\left(\pi(\theta|\underset{\sim}{x},\underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})\right)\\
&= 2\left[1 - \frac{\Gamma\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{\sqrt{\Gamma(\alpha + aN_x + aN_y)\Gamma(\alpha + aN_x)}} \times \frac{\left(1 + \frac{N_y\overline{y}}{\beta+N_x\overline{x}}\right)^{\frac{\alpha+aN_x+aN_y}{2}}}{\left(1 + \frac{N_y\overline{y}}{2(\beta+N_x\overline{x})}\right)^{\alpha+aN_x+aN_y/2}}\right].
\end{aligned}$$

Pour prouver la proposition 3.5.1, nous devons montrer que

$$\lim_{N_x\to\infty} 2\left[1 - \frac{\Gamma\left(\alpha + aN_x + \frac{aN_y}{2}\right)}{\sqrt{\Gamma(\alpha + aN_x + aN_y)\Gamma(\alpha + aN_x)}} \times \frac{\left(1 + \frac{N_y\overline{y}}{\beta+N_x\overline{x}}\right)^{\frac{\alpha+aN_x+aN_y}{2}}}{\left(1 + \frac{N_y\overline{y}}{2(\beta+N_x\overline{x})}\right)^{\alpha+aN_x+aN_y/2}}\right] = 0.$$

En utilisant les lemmes 3.5.1 et 3.5.2 ainsi que les propriétés des limites, le résultat est immédiat. □

Regardons maintenant ce qui se passe avec la J-divergence.

Proposition 3.5.2. *Si* N_y *est fixe et sous les hypothèses* (3.5.1)*, nous avons que*

$$\lim_{N_x\to\infty} J\left(\pi(\theta|\underset{\sim}{x},\underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})\right) = 0.$$

DÉMONSTRATION.

$$\begin{aligned}
&\lim_{N_x\to\infty} J\left(\pi(\theta|\underset{\sim}{x},\underset{\sim}{y}), \pi(\theta|\underset{\sim}{x})\right)\\
&= \lim_{N_x\to\infty} \frac{\left(\frac{N_y\overline{y}}{\beta+N_x\overline{x}}\right)}{1 + \left(\frac{N_y\overline{y}}{\beta+N_x\overline{x}}\right)} \times \left[\left(\frac{N_y\overline{y}}{\beta + N_x\overline{x}}\right) \times (\alpha + aN_x) - aN_y\right]\\
&= 0,
\end{aligned}$$

car nous avons que

$$\lim_{N_x\to\infty} \frac{\left(\frac{N_y\overline{y}}{\beta+N_x\overline{x}}\right)}{1 + \left(\frac{N_y\overline{y}}{\beta+N_x\overline{x}}\right)} = 0$$

et que

$$\lim_{N_x\to\infty}\left[\left(\frac{N_y\overline{y}}{\beta+N_x\overline{x}}\right)\times(\alpha+aN_x)-aN_y\right]=\lim_{N_x\to\infty}\frac{aN_xN_y\overline{y}}{\beta+N_x\overline{x}}-aN_y$$
$$=\frac{aN_y\overline{y}}{\mu_x}-aN_y.$$

□

Finalement, regardons la norme L_2^2.

Proposition 3.5.3. *Si* N_y *est fixe et sous les hypothèses* (3.5.1)*, nous avons que*

$$\lim_{N_x\to\infty} L_2^2\left(\pi(\theta|\underset{\sim}{x},\underset{\sim}{y}),\pi(\theta|\underset{\sim}{x})\right)=0.$$

Dû à la forme algébrique complexe de L_2^2, nous avons fait un graphique pour montrer le comportement de la distance lorsque N_x augmente. À la figure 3.1, nous voyons bien que la distance tend vers zéro lorsque N_x augmente.

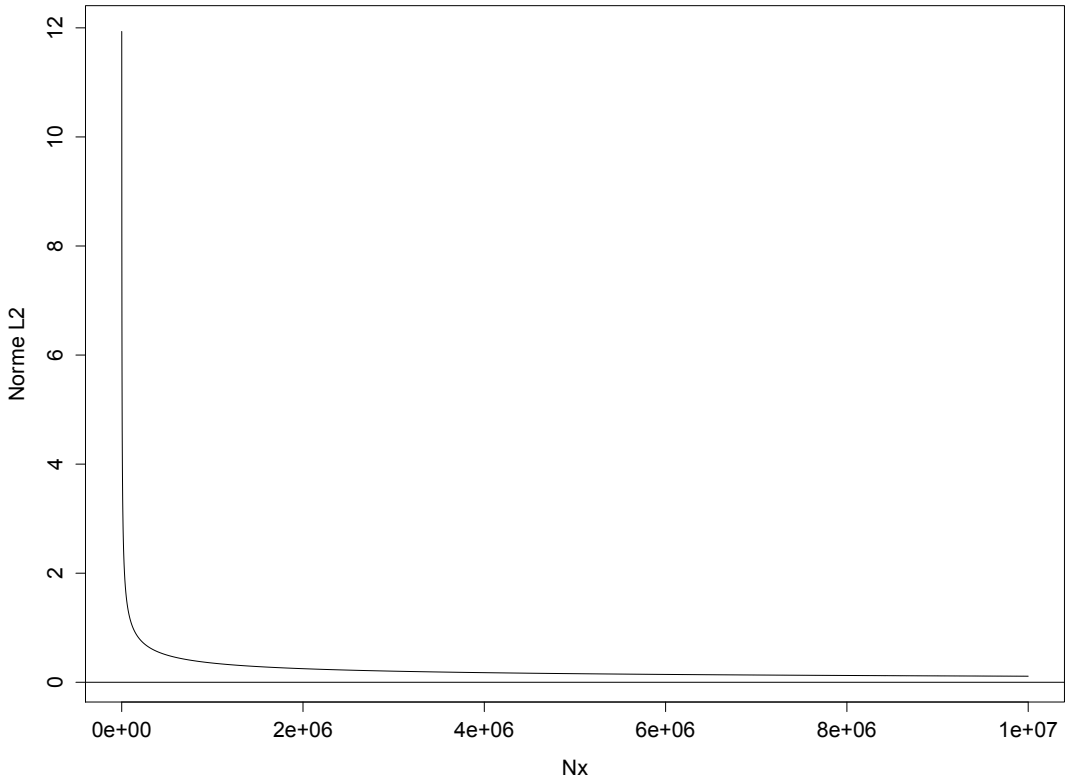

FIGURE 3.1. Graphique de la norme L_2^2 lorsque $N_x\to\infty$, N_y fixe.

Finalement, après avoir montré que les deux densités *a posteriori* sont indiscernables lorsque $N_x\to\infty$ (N_y fixe), nous pouvons aussi regarder le comportement de W. En fait, à la section précédente, nous avons montré que $W\sim$

$\text{Beta}'(aN_y, \alpha + aN_x)$, ce qui implique que

$$E[W] = \frac{aN_y}{\alpha + aN_x - 1},$$
$$Var[W] = \frac{aN_y(aN_y + \alpha + aN_x - 1)}{(\alpha + aN_x - 2)(\alpha + aN_x - 1)^2}.$$

Ainsi, nous avons que l'espérance ainsi que la variance tendent vers 0 lorsque $N_x \to \infty$ (N_y fixe), ce qui est cohérent avec ce que nous avons trouvé précédemment.

En résumé, comme nous venons de le voir, lorsque $N_x \to \infty$ (N_y fixe), les différentes distances entre deux densités *a posteriori* tendent vers 0. À l'opposé, si N_x (ou P_x) est trop petit, les densités *a posteriori* seront instables et le processus mènera à de fausses alertes. L'objectif sera donc de trouver un choix raisonnable de P_x et P_y.

Aussi, un autre argument non négligeable en faveur des fenêtres mobiles, est que si nous utilisons toutes les observations disponibles et que nous détectons de la non-stationnarité, nous ne saurons pas à partir de quel moment le système générant les observations sera devenu instable. En fait, le fait de balayer l'étendue des observations en petits sous-intervalles permet de détecter à partir de quel moment la densité des observations a changé.

3.6. Estimation des hyper-paramètres

Dans le cadre de notre analyse bayésienne, nous avons décidé d'utiliser la méthode des moments bayésiens afin d'estimer les hyper-paramètres α et β de la densité *a priori* que nous avons choisie. En fait, cette méthode consiste à égaler les moments empiriques avec les moments théoriques de la densité marginale $m(\underset{\sim}{x})$. Le théorème suivant nous donne l'expression de l'espérance et de la variance des observations selon la densité marginale.

Théorème 3.6.1. *Posons* $\mu(\theta) = E[X|\theta]$ *et* $\sigma^2(\theta) = Var[X|\theta]$. *Si* μ_m *et* σ_m^2 *représentent l'espérance et la variance de X selon la densité marginale* $m(x)$ *respectivement, alors :*

$$\mu_m = E^{\pi(\theta)}[\mu(\theta)],$$
$$\sigma_m^2 = E^{\pi(\theta)}\left[\sigma^2(\theta)\right] + E^{\pi(\theta)}\left[(\mu(\theta) - \mu_m)^2)\right].$$

Démonstration.

$$\mu_m = \int_{\mathbb{R}} x \times m(x)dx = \int_{\mathbb{R}} x \left[\int_{\Theta} \pi(\theta) f(x|\theta) d\theta\right] dx$$

$$
\begin{aligned}
&= \int_{\Theta} \left[\int_{\mathbb{R}} x \times f(x|\theta) dx \right] \pi(\theta) d\theta \\
&= \int_{\Theta} \mu(\theta) \pi(\theta) d\theta \\
&= E^{\pi(\theta)} \left[\mu(\theta) \right].
\end{aligned}
$$

De plus, nous avons

$$
\begin{aligned}
\sigma_m^2 &= \int_{\mathbb{R}} (x - \mu_m)^2 \, m(x) dx \\
&= \int_{\Theta} \left[\int_{\mathbb{R}} \left([x - \mu(\theta)] + [\mu(\theta) - \mu_m] \right)^2 f\left(x|\theta\right) dx \right] \pi(\theta) d\theta \\
&= \int_{\Theta} \left[\int_{\mathbb{R}} \left[(x - \mu(\theta))^2 + (\mu(\theta) - \mu_m)^2 + 2 \left(x - \mu(\theta)\right) \left(\mu(\theta) - \mu_m\right) \right] \right. \\
&\qquad \left. \times f\left(x|\theta\right) dx \right] \pi(\theta) d\theta \\
&= \int_{\Theta} \left[\sigma^2(\theta) + (\mu(\theta) - \mu_m)^2 + 0 \right] \pi(\theta) d\theta \\
&= E^{\pi(\theta)} \left[\sigma^2(\theta) + (\mu(\theta) - \mu_m)^2 \right].
\end{aligned}
$$

□

L'estimation par la méthode des moments bayésiens consiste à résoudre les équations : $\mu_m = \bar{x}$ et $\sigma_m^2 = s^2$. Dans le cas particulier où $X \sim \text{Gamma}(a, \theta)$ et $\theta \sim \text{Gamma}(a, \theta)$, nous avons

$$\mu(\theta) = E\left[X|\theta\right] = \frac{a}{\theta}$$

et que

$$\sigma^2(\theta) = Var\left[X|\theta\right] = \frac{a}{\theta^2}.$$

Ceci implique que

$$
\begin{aligned}
\mu_m &= E^{\pi(\theta)} \left[\mu(\theta) \right] \\
&= E^{\pi(\theta)} \left[\frac{a}{\theta} \right] \\
&= \frac{a\beta^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} \frac{1}{\theta} \times \theta^{\alpha - 1} \exp\{-\theta\beta\} \\
&= \frac{a\beta^{\alpha}}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha - 1)}{\beta^{\alpha - 1}} \\
&= \frac{a\beta}{(\alpha - 1)}
\end{aligned}
$$

et que

$$\begin{aligned}
\sigma_m^2 &= E^{\pi(\theta)}\left[\sigma^2(\theta)\right] + E^{\pi(\theta)}\left[(\mu(\theta)-\mu_m)^2)\right] \\
&= E^{\pi(\theta)}\left[\frac{a}{\theta^2}\right] + E^{\pi(\theta)}\left[\left(\frac{a}{\theta}-\frac{a\beta}{\alpha-1}\right)^2\right] \\
&= \frac{a\beta^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha-2)}{\beta^{\alpha-2}} + E\pi(\theta)\left[\frac{a^2}{\theta^2}\right] - 2\frac{a^2\beta}{(\alpha-1)}E^{\pi(\theta)}\left[\frac{1}{\theta}\right] + \frac{a^2\beta^2}{(\alpha-1)^2} \\
&= \frac{a\beta^2}{(\alpha-1)(\alpha-2)} + \frac{a^2\beta^2}{(\alpha-1)(\alpha-2)} - \frac{a^2\beta^2}{(\alpha-1)^2} \\
&= \frac{a\beta^2}{(\alpha-1)}\left[\frac{1}{\alpha-2} + \frac{a}{(\alpha-2)} - \frac{a}{\alpha-1}\right] \\
&= \frac{a\beta^2}{(\alpha-1)^2(\alpha-2)}\left[a+\alpha-1\right]
\end{aligned}$$

Si nous estimons μ_m par $\overline{x}$ et σ_m^2 par s^2, nous obtenons

$$\alpha = \frac{2as^2 + a\overline{x}^2 - \overline{x}^2}{as^2 - \overline{x}^2}$$

et

$$\beta = \frac{\overline{x}\,(\alpha-1)}{a}.$$

Toutefois, afin de nous assurer que μ_m et σ_m^2 soient définis, nous allons utiliser

$$\begin{aligned}
\widehat{\alpha} &= \max\left\{\frac{2as^2 + a\overline{x}^2 - \overline{x}^2}{as^2 - \overline{x}^2}, 2\right\}, \\
\widehat{\beta} &= \frac{\overline{x} \times (\widehat{\alpha}-1)}{a}.
\end{aligned}$$

Ces estimateurs seront ceux que nous utiliserons lorsqu'il sera nécessaire d'estimer le couple d'hyper-paramètres (α, β).

3.7. CHOIX DES PARAMÈTRES OPTIMAUX

Comme nous l'avons vu précédemment, l'utilisation des fenêtres mobiles nécessite le besoin de définir différents paramètres. L'objectif de cette section sera de trouver les paramètres les mieux adaptés qui maximisent la performance de notre outil. Encore une fois et pour être cohérent, nous utiliserons l'ARL comme critère de performance. Le but principal de cette section sera donc de déterminer l'ARL pour les différents scénarios ainsi que pour les différents paramètres possibles, et ce, afin de sélectionner les meilleurs paramètres.

3.7.1. Présentation des paramètres

Dans le cadre de la partie bayésienne, nous devons considérer deux paramètres :

(1) P_x : nombre d'années que contient l'intervalle I_x, appelé intervalle d'observations.

(2) P_y : nombre d'années que contient l'intervalle I_y, appelé intervalle de *prévision*. Idéalement, le nombre d'années dans la fenêtre de *prévision* devra être minimisé, car il est intéressant de savoir rapidement lorsqu'on quitte la stationnarité.

Voici quelques précisions supplémentaires pour bien mettre en place le canevas :

- Nous allons faire l'hypothèse que pour notre premier essai, toutes les observations de l'intervalle I_x sont régulières et que toutes les observations dans la fenêtre de prévision (I_y) sont non régulières.
- Nous allons, par la suite, faire *avancer* la fenêtre mobile d'une année à la fois.

Au cours de cette section, nous commencerons par montrer qu'il est possible d'estimer l'ARL de manière théorique pour le premier scénario, mais que ceci présente beaucoup de difficultés et, par la suite, nous estimerons l'ARL pratique (par simulation) pour les deux scénarios.

De plus, tout comme dans le chapitre 2, comme nous avions déjà effectué les simulations ainsi que les calculs théorique dans le cas où les observations suivent la loi exponentielle, nous supposerons que $a = 1$, c'est-à-dire, que les observations sont de loi exponentielle.

3.7.2. ARL théorique

Dans cette section, nous tenterons d'estimer l'ARL de manière théorique pour le premier scénario seulement, le deuxième menant à une situation trop complexe. Comme l'ARL dépendra de P_x, P_y et c, nous la noterons $ARL_{Theo}(P_x, P_y, c)$.

Afin de simplifier l'écriture mathématique des prochaines lignes, nous allons introduire quelques notations supplémentaires. Soit R : le nombre d'années corrompues que nous avons dans l'intervalle I_x, où $R \in [0, P_x]$. Lorsque R = P_x, cela signifie que l'ensemble des observations que contient la fenêtre mobile est corrompu. De plus, définissons S_{x_i} (respectivement S_{y_i}) comme étant la somme des observations de la i^e année de l'intervalle I_x (respectivement I_y), $i = 1..., P_x$ (respectivement $i = 1..., P_y$). En plus, notons que x_{ij} (respectivement

y_{ij}) sera la j^e observation de la i^e année de I_x (respectivement I_y). Finalement, notons que $\underset{\sim}{x}$ (respectivement $\underset{\sim}{y}$) constituera toutes les observations de I_x (respectivement I_y). Nous aimerions souligner qu'à chaque essai, nous allons faire avancer la fenêtre mobile et donc que les observations de I_x et de I_y changeront. Ceci implique que tous les vecteurs définis précédemment changeront aussi.

3.7.2.1. *Première tentative*

Afin de trouver l'ARL de manière théorique, nous allons calculer la puissance à chaque essai, c'est-à-dire, la probabilité de rejeter le modèle sachant que la vraie moyenne a été multipliée par un facteur c. Typiquement, la puissance devrait diminuer en fonction du nombre d'essais, car les densités *a posteriori* contiendront de moins en moins d'observations saines au fur et à mesure que la fenêtre mobile avancera. Nous allons maintenant procéder au calcul de la puissance à chaque essai.

– **Essai #1** : R=0. Nous avons les hypothèses suivantes :

$$X_{ij} \sim \text{Gamma}(a, \theta), \quad i = 1, ..., P_x \text{ et } j = 1, ..., n_{x_i},$$
$$Y_{ij} \sim \text{Gamma}\left(a, \frac{\theta}{c}\right), \quad i = 1, ..., P_y \text{ et } j = 1, ..., n_{y_i}$$

et

$$\theta|\underset{\sim}{x} \sim \text{Gamma}\left(\alpha + aN_x, \beta + \sum_{i=1}^{P_x} S_{x_i}\right).$$

Trouvons maintenant la densité prédictive dans ce cas.

$$\begin{aligned} m(\underset{\sim}{y}|\underset{\sim}{x}) &= \int_\Theta \left[\prod_{i=1}^{P_y}\prod_{j=1}^{n_{y_i}} \frac{\left(\frac{\theta}{c}\right)^a}{\Gamma(a)} \times y_{ij}^{a-1} \exp\left\{-\frac{\theta}{c} y_{ij}\right\}\right] \times \left[\frac{\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)^{\alpha+aN_x}}{\Gamma(\alpha + aN_x)}\right. \\ &\quad \left. \times \theta^{\alpha+aN_x-1} \exp\left\{-\theta\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)\right\}\right] d\theta \\ &= \frac{\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)^{\alpha+aN_x} \left[\prod_{i=1}^{P_y}\prod_{j=1}^{n_{y_i}} y_{ij}\right]^{a-1}}{c^{aN_y}\left[\Gamma(a)\right]^{N_y}\Gamma(\alpha + aN_x)} \times \int_\Theta \theta^{\alpha+aN_x+aN_y-1} \\ &\quad \times \exp\left\{-\theta\left(\beta + \sum_{i=1}^{P_x} S_{x_i} + \frac{1}{c}\sum_{i=1}^{P_y} S_{y_i}\right)\right\} d\theta \\ &= \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\left[\Gamma(a)\right]^{N_y}} \times \left[\prod_{i=1}^{P_y}\prod_{j=1}^{n_{y_i}} y_{ij}\right]^{a-1} \end{aligned}$$

$$\times \frac{\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)^{\alpha+aN_x}}{\left(\beta + \sum_{i=1}^{P_x} S_{x_i} + \frac{1}{c}\sum_{i=1}^{P_y} S_{y_i}\right)^{\alpha+aN_x+aN_y}}.$$

Posons $\underset{\sim}{\mathcal{Y}} = \left(Y_{11}, Y_{12}, ..., Y_{1n_{y_1}}, ..., Y_{P_y1}, Y_{P_y2}, ..., Y_{P_y n_{y_{P_y}}}\right)$, où $\mathcal{Y}_\ell$ est la ℓ^e composante du vecteur $\underset{\sim}{\mathcal{Y}}$. En procédant aux changements de variable

$$T_i = \sum_{\ell=1}^{i} \mathcal{Y}_\ell \quad , i = 1, ..., N_y, \tag{3.7.1}$$

nous pouvons facilement montrer que le jacobien de cette transformation donne 1. Ceci fait en sorte que

$$f(\underset{\sim}{t}|\underset{\sim}{x}) = \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\left[\Gamma(a)\right]^{N_y}} \times \left[\prod_{i=1}^{N_y}(t_i - t_{i-1})\right]^{a-1} \times \frac{\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)^{\alpha+aN_x}}{\left(\beta + \sum_{i=1}^{P_x} S_{x_i} + \frac{1}{c}t_{N_y}\right)^{\alpha+aN_x+aN_y}}$$

De plus, à partir de ce résultat et du lemme 3.3.1, il est aisé de trouver la loi de T_{N_y}.

$$\begin{aligned}
&f(t_{N_y}|\underset{\sim}{x}) \\
&= \int_0^{t_{N_y}} \int_0^{t_{N_y-1}} ... \int_0^{t_3} \int_0^{t_2} f(\underset{\sim}{t}|\underset{\sim}{x}) \times \prod_{i=1}^{N_y-1} dt_i \\
&= \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\left[\Gamma(a)\right]^{N_y}} \times \frac{\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)^{\alpha+aN_x}}{\left(\beta + \sum_{i=1}^{P_x} S_{x_i} + \frac{1}{c}t_{N_y}\right)^{\alpha+aN_x+aN_y}} \\
&\quad \times \int_0^{t_{N_y}} \int_0^{t_{N_y-1}} ... \int_0^{t_3} \int_0^{t_2} \left[\prod_{i=1}^{N_y}(t_i - t_{i-1})\right]^{a-1} \times \prod_{i=1}^{N_y-1} dt_i \\
&= \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\left[\Gamma(a)\right]^{N_y}} \times \frac{\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)^{\alpha+aN_x}}{\left(\beta + \sum_{i=1}^{P_x} S_{x_i} + \frac{1}{c}t_{N_y}\right)^{\alpha+aN_x+aN_y}} \times \frac{t_{N_y}^{aN_y-1}\left[\Gamma(a)\right]^{N_y}}{\Gamma(aN_y)} \\
&= \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\Gamma(aN_y)} \times \frac{\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)^{\alpha+aN_x}}{\left(\beta + \sum_{i=1}^{P_x} S_{x_i} + \frac{1}{c}t_{N_y}\right)^{\alpha+aN_x+aN_y}} \times t_{N_y}^{aN_y-1}.
\end{aligned}$$

Finalement, en posant $U_1 = \frac{T_{N_y}}{c\left(\beta+\sum_{i=1}^{P_x} S_{x_i}\right)}$, nous avons

$$dU_1 = \frac{dT_{N_y}}{c\left(\beta + \sum_{i=1}^{P_x} S_{x_i}\right)}$$

et nous obtenons la densité suivante pour $U_1|\underset{\sim}{x}$

$$f(u_1|\underset{\sim}{x}) = \frac{\Gamma(\alpha + aN_x + aN_y)}{\Gamma(\alpha + aN_x)\Gamma(aN_y)} \times \frac{u_1^{aN_y-1}}{(1+u_1)^{\alpha+aN_x+aN_y}}.$$

Ceci implique que $U_1|\underset{\sim}{x} \sim Beta'(aN_y, \alpha+aN_x)$. Pour trouver la puissance, il suffit de calculer

$$\begin{aligned}\mathbb{P}(W_1 \geq W_0) &= \mathbb{P}(c \times U_1 \geq W_0)\\ &= \mathbb{P}\left(U_1 \geq \frac{W_0}{c}\right)\\ &= \mathbb{P}(\text{Rejeter}|R=0).\end{aligned}$$

En supposant que nous n'ayons pas pu détecter la non-stationnarité à l'essai #1, nous allons devoir faire progresser la fenêtre mobile d'une année.

– **essai #2** : R=1. Pour cet essai, nous avons qu'une seule année d'observations est corrompue dans I_x. Nous avons donc les hypothèses suivantes.

$$X_{ij} \sim \begin{cases}\text{Gamma}(a,\theta) & \text{pour} \quad i = 1,...,P_x-1 \text{ et } j = 1,...,n_{x_i},\\ \text{Gamma}\left(a,\frac{\theta}{c}\right) & \text{pour} \quad i = P_x \text{ et } j = 1,...,n_{x_i},\end{cases}$$

$$Y_{ij} \sim \text{Gamma}\left(\frac{\theta}{c}\right), \quad i = 1,...,P_y \text{ et } j = 1,...,n_{y_i}$$

et

$$\theta|\underset{\sim}{x} \sim \text{Gamma}\left(\alpha + aN_x, \beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right).$$

Trouvons maintenant la densité prédictive dans ce cas.

$$\begin{aligned}m(\underset{\sim}{y}|\underset{\sim}{x}) &= \int_\Theta \left[\prod_{i=1}^{P_y}\prod_{j=1}^{n_{y_i}} \frac{\left(\frac{\theta}{c}\right)^a}{\Gamma(a)} \times y_{ij}^{a-1}\exp\left\{-\frac{\theta}{c}y_{ij}\right\}\right] \times \left[\frac{\left(\beta+\sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right)^{\alpha+aN_x}}{\Gamma(\alpha+aN_x)}\right.\\ &\qquad \left.\times\theta^{\alpha+aN_x-1}\exp\left\{-\theta\left(\beta+\sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right)\right\}\right] d\theta\\ &= \frac{\left(\beta+\sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right)^{\alpha+aN_x}\left[\prod_{i=1}^{P_y}\prod_{j=1}^{n_{y_i}} y_{ij}\right]^{a-1}}{c^{aN_y}\left[\Gamma(a)\right]^{N_y}\Gamma(\alpha+aN_x)} \times \int_\Theta \theta^{\alpha+aN_x+aN_y-1}\end{aligned}$$

$$\times \exp\left\{-\theta\left(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}} + \frac{1}{c}\sum_{i=1}^{P_y} S_{y_i}\right)\right\} d\theta$$

$$= \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\left[\Gamma(a)\right]^{N_y}} \times \left[\prod_{i=1}^{P_y}\prod_{j=1}^{n_{y_i}} y_{ij}\right]^{a-1}$$

$$\times \frac{\left(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right)^{\alpha+aN_x}}{\left(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}} + \frac{1}{c}\sum_{i=1}^{P_y} S_{y_i}\right)^{\alpha+aN_x+aN_y}}.$$

En procédant au changement de variable (3.7.1) , nous obtenons

$$f(\underset{\sim}{t}|\underset{\sim}{x}) = \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\left[\Gamma(a)\right]^{N_y}} \times \left[\prod_{i=1}^{N_y}(t_i - t_{i-1})\right]^{a-1}$$

$$\times \frac{(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}})^{\alpha+aN_x}}{(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}(S_{x_{P_x}} + t_{N_y}))^{\alpha+aN_x+aN_y}}.$$

De plus, à partir de ce résultat, il est aisé de trouver la loi de T_{N_y}.

$$f(t_{N_y}|\underset{\sim}{x}) = \int_0^{t_{N_y}}\int_0^{t_{N_y-1}} \dots \int_0^{t_3}\int_0^{t_2} f(\underset{\sim}{t}|\underset{\sim}{x}) \times \prod_{i=1}^{N_y-1} dt_i$$

$$= \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\Gamma(aN_y)} \times \frac{(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}})^{\alpha+N_x}}{(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}(S_{x_{P_x}} + t_{N_y}))^{\alpha+N_x+N_y}}$$

$$\times t_{N_y}^{N_y-1}$$

$$= \frac{\Gamma(\alpha + aN_x + aN_y)}{\Gamma(\alpha + aN_x)\Gamma(aN_y)} \times \frac{1}{c\left(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right)}$$

$$\times \frac{\left(\frac{t_{N_y}}{c\left(\beta+\sum_{i=1}^{P_x-1} S_{x_i}+\frac{1}{c}S_{x_{P_x}}\right)}\right)^{aN_y-1}}{\left(1 + \frac{t_{N_y}}{c\left(\beta+\sum_{i=1}^{P_x-1} S_{x_i}+\frac{1}{c}S_{x_{P_x}}\right)}\right)^{\alpha+aN_x+aN_y}}.$$

Finalement, en posant $U_2 = \frac{T_{N_y}}{c\left(\beta+\sum_{i=1}^{P_x-1} S_{x_i}+\frac{1}{c}S_{x_{P_x}}\right)}$ nous avons

$$dU_2 = \frac{dT_{N_y}}{c\left(\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right)}$$

et nous obtenons la densité suivante pour $U_2|\underset{\sim}{x}$

$$f(u_2|\underset{\sim}{x}) = \frac{\Gamma(\alpha + aN_x + aN_y)}{\Gamma(\alpha + aN_x)\Gamma(aN_y)} \times \frac{u_2^{aN_y-1}}{(1+u_2)^{\alpha+aN_x+aN_y}}.$$

Ceci implique que $U_2|\underset{\sim}{x} \sim \text{Beta}'(aN_y, \alpha + aN_x)$. Pour trouver la puissance, il suffit de calculer

$$\begin{aligned}\mathbb{P}(W_2 \geq W_0) &= \mathbb{P}\left(U_2 \times \frac{c\left[\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right]}{\beta + \sum_{i=1}^{P_x} S_{x_i}} \geq W_0\right) \\ &= \mathbb{P}\left(U_2 \geq W_0 \times \frac{\beta + \sum_{i=1}^{P_x} S_{x_i}}{c\left[\beta + \sum_{i=1}^{P_x-1} S_{x_i} + \frac{1}{c}S_{x_{P_x}}\right]}\right) \\ &= \mathbb{P}(\text{Rejeter}|R = 1).\end{aligned}$$

– **Essai #k** : R = k-1 (Généralisation) : Nous avons les hypothèses suivantes :

$$X_{ij} \sim \begin{cases}\text{Gamma}(a, \theta) & \text{pour} \quad i = 1, ..., P_x - R \text{ et } j = 1, ..., n_{x_i}, \\ \text{Gamma}\left(a, \frac{\theta}{c}\right) & \text{pour} \quad i = P_x - R + 1, ..., P_x \text{ et } j = 1, ..., n_{x_i},\end{cases}$$

$$Y_{ij} \sim \text{Gamma}\left(a, \frac{\theta}{c}\right) \quad \text{pour} \quad i = 1, ..., P_y \text{ et } j = 1, ..., n_{y_i}$$

et

$$\theta|\underset{\sim}{x} \sim \text{Gamma}\left(\alpha + aN_x, \beta + \sum_{i=1}^{P_x-R} S_{x_i} + \frac{1}{c}\sum_{i=Px-R+1}^{Px} S_{x_i}\right).$$

Trouvons maintenant la densité prédictive dans ce cas.

$$\begin{aligned}m(\underset{\sim}{y}|\underset{\sim}{x}) = \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\left[\Gamma(a)\right]^{N_y}} \times \left[\prod_{i=1}^{P_y}\prod_{j=1}^{n_{y_i}} y_{ij}\right]^{a-1} \\ \times \frac{\left(\beta + \sum_{i=1}^{P_x-R} S_{x_i} + \frac{1}{c}\sum_{i=P_x-R+1}^{P_x} S_{x_i}\right)^{\alpha+aN_x}}{\left[\beta + \sum_{i=1}^{P_x-R} S_{x_i} + \frac{1}{c}\left(\sum_{i=P_x-R+1}^{P_x} S_{x_i} + \sum_{i=1}^{P_y} S_{y_i}\right)\right]^{\alpha+aN_x+aN_y}}.\end{aligned}$$

En procédant au changement de variable (3.7.1), nous obtenons

$$\begin{aligned}f(\underset{\sim}{t}|\underset{\sim}{x}) = \frac{\Gamma(\alpha + aN_x + aN_y)}{c^{aN_y}\Gamma(\alpha + aN_x)\left[\Gamma(a)\right]^{N_y}} \times \left[\prod_{i=1}^{N_y}(t_i - t_{i-1})\right]^{a-1} \\ \times \frac{(\beta + \sum_{i=1}^{P_x-R} S_{x_i} + \frac{1}{c}\sum_{i=P_x-R+1}^{P_x} S_{x_i})^{\alpha+aN_x}}{(\beta + \sum_{i=1}^{P_x-R} S_{x_i} + \frac{1}{c}\left(\sum_{i=P_x-R+1}^{P_x} S_{x_i} + t_{N_y}\right))^{\alpha+aN_x+aN_y}}.\end{aligned}$$

De plus, à partir de ce résultat, nous trouvons la loi de T_{N_y}.

$$f(t_{N_y}|\underset{\sim}{x}) = \frac{\Gamma(\alpha + aN_x + aN_y)}{\Gamma(\alpha + aN_x)\Gamma(aN_y)} \times \frac{1}{c\left(\beta + \sum_{i=1}^{P_x-R} S_{x_i} + \frac{1}{c}\sum_{i=P_x-R+1}^{P_x} S_{x_i}\right)}$$

$$\times \frac{\left(\frac{t_{N_y}}{c\left(\beta+\sum_{i=1}^{P_x-R} S_{x_i}+\frac{1}{c}\sum_{i=P_x-R+1}^{P_x} S_{x_i}\right)}\right)^{aN_y-1}}{\left(1+\frac{t_{N_y}}{c\left(\beta+\sum_{i=1}^{P_x-R} S_{x_i}+\frac{1}{c}\sum_{i=P_x-R+1}^{P_x} S_{x_i}\right)}\right)^{\alpha+aN_x+aN_y}}.$$

Finalement, en posant $U_k = \frac{T_{N_y}}{c\left(\beta+\sum_{i=1}^{P_x-R} S_{x_i}+\frac{1}{c}\sum_{i=P_x-R+1}^{P_x} S_{x_i}\right)}$, nous obtenons la densité suivante pour $U_k|\underset{\sim}{x}$:

$$f(u_k|\underset{\sim}{x}) = \frac{\Gamma(\alpha+aN_x+aN_y)}{\Gamma(\alpha+aN_x)\Gamma(aN_y)} \times \frac{u_k^{aN_y-1}}{(1+u_k)^{\alpha+aN_x+aN_y}}.$$

Ceci implique que $U_k|\underset{\sim}{x} \sim \text{Beta}'(aN_y, \alpha+aN_x)$. Pour trouver la puissance, il suffit de calculer

$$\begin{aligned}
\mathbb{P}(W_k \geq W_0) &= \mathbb{P}\left(U_k \times \frac{c\left[\beta+\sum_{i=1}^{P_x-R} S_{x_i}+\frac{1}{c}\sum_{i=P_x-R+1}^{P_x} S_{x_i}\right]}{\beta+\sum_{i=1}^{P_x} S_{x_i}} \geq W_0\right) \\
&= \mathbb{P}\left(U_k \geq W_0 \times \frac{\beta+\sum_{i=1}^{P_x} S_{x_i}}{c\left[\beta+\sum_{i=1}^{P_x-R} S_{x_i}+\frac{1}{c}\sum_{i=P_x-R+1}^{P_x} S_{x_i}\right]}\right) \\
&= \mathbb{P}(\text{Rejeter}|R = k-1).
\end{aligned}$$

Si nous définissons G comme étant le nombre d'essais avant de détecter un changement, nous venons de voir que nous pouvons établir sa distribution de probabilité. En effet, nous avons que

$$\begin{aligned}
&\mathbb{P}(G=1) = \mathbb{P}(W_1 \geq W_0), \\
&\mathbb{P}(G=2) = (1-\mathbb{P}(G=1)) \times \mathbb{P}(W_2 \geq W_0), \\
&\mathbb{P}(G=3) = (1-\mathbb{P}(G=1)) \times (1-\mathbb{P}(G=2)) \times \mathbb{P}(W_3 \geq W_0), \\
&\vdots \\
&\mathbb{P}(G=k) = \left(\prod_{i=1}^{k-1}[1-\mathbb{P}(G=i)]\right) \times \mathbb{P}(W_k \geq W_0).
\end{aligned}$$

De ce fait, il est possible de trouver l'ARL qui n'est rien d'autre que l'espérance de G. En fait, G est comme une variable aléatoire de distribution géométrique, mais à probabilité variable d'un essai à l'autre.

3.7.2.2. *Problème de dépendance*

Bien que cette approche semble fonctionner au premier coup d'oeil, ce n'est pas le cas. En fait, pour pouvoir trouver la puissance comme présentée ci-dessus, nous devons faire l'hypothèse que les essais sont indépendants entre

eux, or ceci n'est pas le cas. En effet, supposons que nous sommes très loin de rejeter à l'essai k, ceci implique que les P_y années de l'intervalle I_y sont très similaires aux P_x années de l'intervalle I_x. Ceci signifie qu'à l'essai $k+1$, l'impact qu'aura le fait de remplacer la dernière année de I_y par une nouvelle année d'observations corrompues sera, en quelque sorte, absorbé par les $P_y - 1$ années qui n'étaient pas très différentes de celle de I_x. Afin de mieux visualiser ce qui se passe, nous allons introduire $\mathcal{S}_i$ la somme des observations de la i^e année d'observations, c'est-à-dire, que l'indice n'est plus par rapport à un ensemble (I_x ou I_y), mais représente les années de calendrier. À titre d'exemple, si nous prenons $P_x = 3$ et $P_y = 1$ nous avons

$$W_1 = \frac{S_{y_1}}{\beta + S_{x_1} + S_{x_2} + S_{x_3}} = \frac{\mathcal{S}_4}{\beta + \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3},$$
$$W_2 = \frac{S_{y_1}}{\beta + S_{x_1} + S_{x_2} + S_{x_3}} = \frac{\mathcal{S}_5}{\beta + \mathcal{S}_2 + \mathcal{S}_3 + \mathcal{S}_4},$$
$$W_3 = \frac{S_{y_1}}{\beta + S_{x_1} + S_{x_2} + S_{x_3}} = \frac{\mathcal{S}_6}{\beta + \mathcal{S}_3 + \mathcal{S}_4 + \mathcal{S}_5}.$$

Comme nous pouvons le voir, les W_k consécutifs ont des termes en commun et sont donc dépendants. De plus, l'histoire se complique largement si $P_y > 1$, car il y aura des termes communs aux numérateurs aussi. Par exemple, si nous prenons $P_x = 10$ et $P_y = 5$, nous avons

$$W_1 = \frac{S_{y_1} + \ldots + S_{y_5}}{\beta + S_{x_1} + \ldots + S_{x_{10}}} = \frac{\mathcal{S}_{11} + \mathcal{S}_{12} + \mathcal{S}_{13} + \mathcal{S}_{14} + \mathcal{S}_{15}}{\beta + \mathcal{S}_1 + \ldots + \mathcal{S}_{10}},$$
$$W_2 = \frac{S_{y_1} + \ldots + S_{y_5}}{\beta + S_{x_1} + \ldots + S_{x_{10}}} = \frac{\mathcal{S}_{12} + \mathcal{S}_{13} + \mathcal{S}_{14} + \mathcal{S}_{15} + \mathcal{S}_{16}}{\beta + \mathcal{S}_2 + \ldots + \mathcal{S}_{10} + \mathcal{S}_{11}},$$
$$W_3 = \frac{S_{y_1} + \ldots + S_{y_5}}{\beta + S_{x_1} + \ldots + S_{x_{10}}} = \frac{\mathcal{S}_{13} + \mathcal{S}_{14} + \mathcal{S}_{15} + \mathcal{S}_{16} + \mathcal{S}_{17}}{\beta + \mathcal{S}_3 + \ldots + \mathcal{S}_{10} + \mathcal{S}_{11} + \mathcal{S}_{12}},$$
$$W_4 = \frac{S_{y_1} + \ldots + S_{y_5}}{\beta + S_{x_1} + \ldots + S_{x_{10}}} = \frac{\mathcal{S}_{14} + \mathcal{S}_{15} + \mathcal{S}_{16} + \mathcal{S}_{17} + \mathcal{S}_{18}}{\beta + \mathcal{S}_4 + \ldots + \mathcal{S}_{10} + \mathcal{S}_{11} + \mathcal{S}_{12} + \mathcal{S}_{13}},$$
$$W_5 = \frac{S_{y_1} + \ldots + S_{y_5}}{\beta + S_{x_1} + \ldots + S_{x_{10}}} = \frac{\mathcal{S}_{15} + \mathcal{S}_{16} + \mathcal{S}_{17} + \mathcal{S}_{18} + \mathcal{S}_{19}}{\beta + \mathcal{S}_3 + \ldots + \mathcal{S}_{10} + \mathcal{S}_{11} + \mathcal{S}_{12} + \mathcal{S}_{13} + \mathcal{S}_{14}}.$$

Comme nous pouvons le voir, plus la fenêtre mobile avance dans le temps, moins W_1 a de termes communs avec W_k, $k > 1$. En fait, il faudra faire avancer la fenêtre de $P_x + P_y + 1$ coups avant qu'il n'y ait plus de termes en commun.

Afin de souligner à quel point les W_k sont dépendants, nous avons effectué une petite simulation pour calculer la matrice de corrélations entre les W_k. Pour ce faire, nous avons utilisé les paramètres $P_x = 10, P_y = 5$ et $c = 1{,}07$. Nous avons donc créé P_x années d'observations de moyenne μ^* et P_y années d'observations de moyennes corrompues avec un facteur c. En outre,

chaque année d'observations avait 55 observations qui suivaient chacune la loi Gamma$(1, 1/\mu^*)$. Nous avons répété le processus 10^7 fois, où à chaque fois, nous avons créé les W_k.

TABLEAU 3.2. Corrélations simulées des W_k.

	W_1	W_2	W_3	W_4	W_5
W_1	1,000	0,681	0,379	0,088	-0,188
W_2	0,681	1,000	0,699	0,404	0,126
W_3	0,379	0,699	1,000	0,708	0,428
W_4	0,088	0,404	0,708	1,000	0,722
W_5	-0,188	0,126	0,428	0,722	1,000

Comme nous pouvons l'observer au tableau 3.2, la corrélation entre les W_k semble diminuer lorsque la fenêtre mobile avance. Ceci n'est pas très surprenant, car plus les W_k sont éloignés dans le temps, moins ils ont de termes en commun. En outre, la corrélation entre W_1 et W_5 peut sembler bizarre, mais elle est tout à fait normale. En effet, si W_1 est grand, ceci implique que son numérateur est grand comparé à son dénominateur. Ceci jumelé au fait que le numérateur se retrouve progressivement au dénominateur explique le fait que plus W_1 est grand, plus W_5 aura tendance à être petit. Pour avoir une meilleure idée de ce qui se passe, nous pouvons observer la figure 3.2 qui représente la corrélation entre W_1 et W_k, $k = 1,...,16$. Comme nous pouvons nous en attendre, la corrélation est nulle entre W_1 et W_{16}, car ils n'ont plus de termes communs.

Nous avons aussi décidé de calculer (estimer) la vraie corrélation. Pour y arriver, nous avons calculé la matrice de covariances et ensuite nous l'avons transformée en matrice de corrélations. Tout d'abord, rappelons que

$$\mathrm{Cov}\left(W_i, W_j\right) = \mathrm{E}\left[W_i W_j\right] - \mathrm{E}\left[W_i\right]\mathrm{E}\left[W_j\right],$$

où les W_k sont fonction des $\mathcal{S}_k$. En outre, comme les observations corrompues suivent une loi gamma de paramètre 1 et θ/c, nous avons que la somme des observations pour l'année corrompue k suit la loi Gamma$\left(N_y, \theta/c\right)$. Comme la covariance n'est rien d'autre qu'une soustraction d'espérance, nous avons utilisé la méthode de Monte-Carlo afin d'estimer la covariance théorique. Les résultats pour la vraie corrélation de l'exemple présenté ci-haut sont présentés au tableau 3.3. Comme nous pouvons le voir, les résultats sont très similaires aux résultats du tableau 3.2.

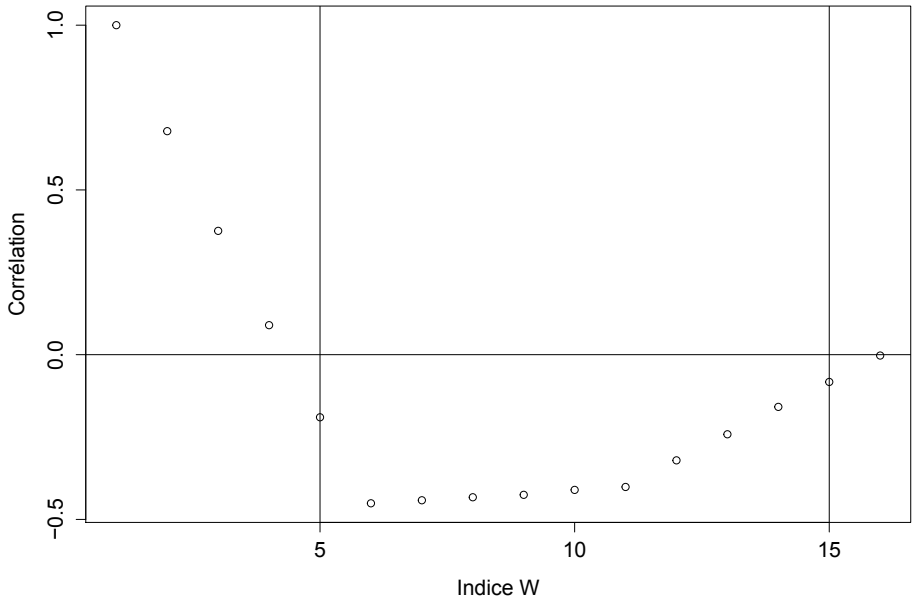

FIGURE 3.2. Corrélations entre W_1 et W_k, k = 1,...,16. Pour $P_x = 10, P_y = 5, c = 1{,}07$.

TABLEAU 3.3. Corrélations théoriques des W_k.

	W_1	W_2	W_3	W_4	W_5
W_1	1,000	0,678	0,374	0,086	-0,189
W_2	0,678	1,000	0,694	0,405	0,130
W_3	0,374	0,694	1,000	0,710	0,434
W_4	0,086	0,405	0,710	1,000	0,723
W_5	-0,189	0,130	0,434	0,723	1,000

Comme nous venons de le démontrer, les W_k sont dépendants entre eux. Ceci fait en sorte que nous devons trouver une autre méthode que celle proposée si nous voulons arriver à calculer l'ARL de manière théorique. Ceci sera l'objectif de la prochaine section.

3.7.2.3. *Deuxième tentative*

Maintenant que nous avons réalisé que les W_k sont dépendants entre eux, nous allons tenter une nouvelle approche. En fait, nous avons que

$$\mathbb{P}(G = 1) = \mathbb{P}(W_1 \geq W_0),$$
$$\mathbb{P}(G = 2) = \mathbb{P}(W_2 \geq W_0, W_1 < W_0),$$
$$\mathbb{P}(G = 3) = \mathbb{P}(W_3 \geq W_0, W_2 < W_0, W_1 < W_0),$$

$$\vdots$$

$$\mathbb{P}(G = k) = \mathbb{P}\left(W_k \geq W_0, W_{k-1} < W_0, ..., W_1 < W_0\right).$$

Toutefois, comme nous ne connaissons pas la loi de W_k lorsque nous sommes en présence d'observations corrompues ($c > 0$), nous allons utiliser la variable U_k définie précédemment. À cet effet, rappelons que $U_k \sim \texttt{Beta}'(aN_y, \alpha + aN_x)$. En généralisant l'expression de U_k avec la notation $\mathcal{S}_\ell$, nous obtenons

$$\begin{aligned}
U_k &= \frac{\sum_{i=1}^{P_y} S_{y_i}}{c\left(\beta + \sum_{i=1}^{P_x-k+1} S_{x_i} + \frac{1}{c}\sum_{i=P_x-k+2}^{P_x} S_{x_i}\right)} \\
&= \frac{\sum_{i=P_x+k}^{P_x+k+P_y-1} \mathcal{S}_i}{c\left(\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i \ell_i\right)} \\
&= W_k \times \frac{\left[\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i\right]}{c\left[\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i \ell_i\right]} \\
&= W_k \times C_k
\end{aligned} \tag{3.7.2}$$

où

$$\begin{aligned}
W_k &= \frac{\sum_{i=P_x+k}^{P_x+k+P_y-1} \mathcal{S}_i}{\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i}, \\
\ell_i &= \begin{cases} 1 & \text{si} \quad i \leq P_x, \\ \frac{1}{c} & \text{sinon}, \end{cases} \\
C_k &= \frac{\left[\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i\right]}{c\left[\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i \ell_i\right]}.
\end{aligned} \tag{3.7.3}$$

En d'autres mots, $\ell_i = 1/c$ si l'indice i représente une année d'observations corrompues.

Afin de simplifier le problème, nous supposerons, dans ce qui suit, que $P_x = 3$ et que $P_y = 1$. Notons que la même approche pourrait être utilisée avec des P_x et P_y différents. Nous pouvons voir au tableau 3.4, un résumé des transformations pour quelques valeurs de k. De plus, notons que

$$\begin{aligned}
\mathcal{S}_4 &= U_1 \times c\left(\beta + \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3\right), \\
\mathcal{S}_5 &= U_2 \times c\left(\beta + \mathcal{S}_2 + \mathcal{S}_3 + \frac{1}{c}\mathcal{S}_4\right), \\
\mathcal{S}_6 &= U_3 \times c\left(\beta + \mathcal{S}_3 + \frac{1}{c}\left[\mathcal{S}_4 + \mathcal{S}_5\right]\right),
\end{aligned}$$

TABLEAU 3.4. Résumé des transformations en fonction de k.

k	W_k	C_k	U_k
1	$\frac{\mathcal{S}_4}{\beta+\mathcal{S}_1+\mathcal{S}_2+\mathcal{S}_3}$	$\frac{1}{c}$	$\frac{\mathcal{S}_4}{c[\beta+\mathcal{S}_1+\mathcal{S}_2+\mathcal{S}_3]}$
2	$\frac{\mathcal{S}_5}{\beta+\mathcal{S}_2+\mathcal{S}_3+\mathcal{S}_4}$	$\frac{\beta+\mathcal{S}_2+\mathcal{S}_3+\mathcal{S}_4}{c\left[\beta+\mathcal{S}_2+\mathcal{S}_3+\frac{1}{c}\mathcal{S}_4\right]}$	$\frac{\mathcal{S}_5}{c\left[\beta+\mathcal{S}_2+\mathcal{S}_3+\frac{1}{c}\mathcal{S}_4\right]}$
3	$\frac{\mathcal{S}_6}{\beta+\mathcal{S}_3+\mathcal{S}_4+\mathcal{S}_5}$	$\frac{\beta+\mathcal{S}_3+\mathcal{S}_4+\mathcal{S}_5}{c\left[\beta+\mathcal{S}_3+\frac{1}{c}(\mathcal{S}_4+\mathcal{S}_5)\right]}$	$\frac{\mathcal{S}_6}{c\left[\beta+\mathcal{S}_3+\frac{1}{c}(\mathcal{S}_4+\mathcal{S}_5)\right]}$
4	$\frac{\mathcal{S}_7}{\beta+\mathcal{S}_4+\mathcal{S}_5+\mathcal{S}_6}$	$\frac{\beta+\mathcal{S}_4+\mathcal{S}_5+\mathcal{S}_6}{c\left[\beta+\frac{1}{c}(\mathcal{S}_4+\mathcal{S}_5+\mathcal{S}_6)\right]}$	$\frac{\mathcal{S}_7}{c\left[\beta++\frac{1}{c}(\mathcal{S}_4+\mathcal{S}_5+\mathcal{S}_6)\right]}$

$$\mathcal{S}_7 = U_4 \times c\left(\beta + \frac{1}{c}\left[\mathcal{S}_4 + \mathcal{S}_5 + \mathcal{S}_6\right]\right).$$

Ceci implique que U_2 peut s'écrire comme une fonction de U_1, U_3 comme une fonction de U_2 et U_1, etc. En utilisant l'équation (3.7.2), nous pourrons trouver les bornes adéquates pour les probabilités en fonction de la variable U_k. Toutefois, comme nous venons de le voir, U_k dépend des $(k-1)$ U_k précédents et donc, les bornes seront dépendantes entre elles.

En appliquant le changement de variable, nous obtenons

$$\begin{aligned}
\mathbb{P}(G=1) &= \mathbb{P}\left(U_1 \geq \frac{W_0}{c}\right) \\
&= \int_0^{\infty} f(u_1)du_1, \\
\mathbb{P}(G=2) &= \mathbb{P}\left(U_2 \geq W_0C_2, U_1 < \frac{W_0}{c}\right) \\
&= \int_0^{\frac{W_0}{c}} \int_{b_2(u_1)}^{\infty} f(u_2)f(u_1)du_2du_1, \\
\mathbb{P}(G=3) &= \mathbb{P}\left(U_3 \geq W_0C_3, U_2 < W_0C_2, U_1 < \frac{W_0}{c}\right) \\
&= \int_0^{\frac{W_0}{c}} \int_0^{b_2(u_1)} \int_{b_3(u_1,u_2)}^{\infty} f(u_3)f(u_2)f(u_1)du_3du_2du_1, \\
&\vdots \\
\mathbb{P}(G=k) &= \mathbb{P}\left(U_k \geq W_0C_k, U_{k-1} < W_0C_{k-1}, ..., U_1 < \frac{W_0}{c}\right) \\
&= \int_0^{\frac{W_0}{c}} \cdots \int_0^{b_{k-1}(u_1,...,b_{k-2})} \int_{b_k(u_1,...,u_{k-1})}^{\infty} \prod_{i=1}^{k} f(u_i) \prod_{i=1}^{k} du_i,
\end{aligned}$$

où $b_k(\cdot)$ est la k^e borne trouvée à l'aide de l'équation (3.7.2).

Comme il n'est pas possible de résoudre ces intégrales analytiquement, nous devrons les estimer à l'aide d'une méthode d'intégration numérique. Encore une fois, nous avons choisi la méthode de Monte-Carlo afin d'estimer ces intégrales.

Pour faire un petit résumé, le but premier de notre démarche était de trouver l'ARL théorique qui est définie comme $E[G]$. Afin de trouver l'ARL, il nous faudrait donc trouver théoriquement une infinité de probabilités. Ce que nous espérons, toutefois, c'est que les probabilités associées à la fonction de masse de G diminuent très rapidement pour qu'on puisse avoir une bonne estimation de l'ARL, et ce, sans devoir estimer un nombre important de probabilités.

Nous allons maintenant passer à un exemple pour voir si la méthode théorique peut être appliquée pour trouver l'ARL. Tout comme précédemment, fixons $P_x = 3, P_y = 1$ et $c = 1{,}07$. Les résultats obtenus en appliquant la méthode de Monte-Carlo avec $1{,}25 \times 10^{10}$ itérations ont été résumés à la figure 3.3, représentant les probabilités obtenues ainsi que leur intervalle de confiance respectif. Comme nous pouvons le voir, les estimations des probabilités sont

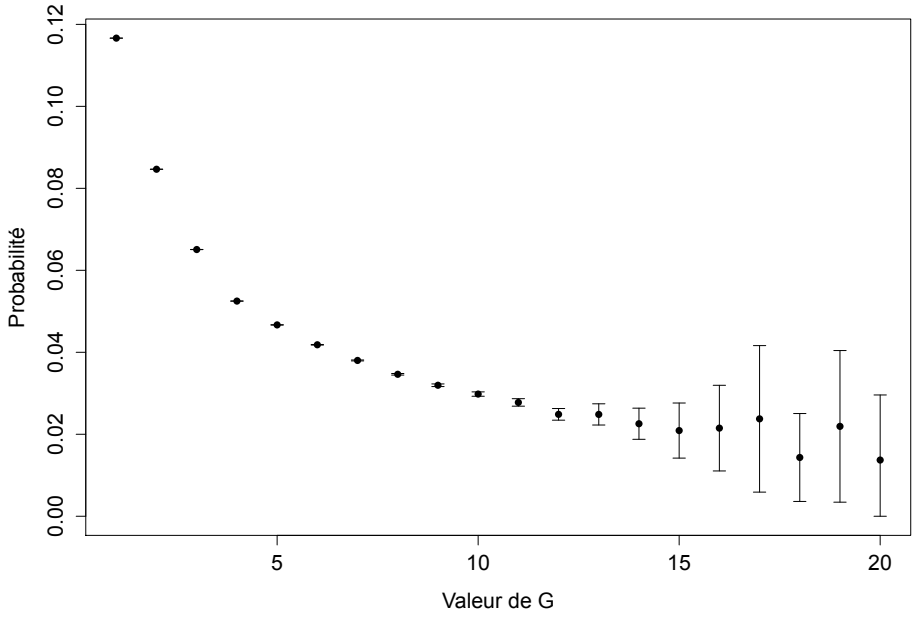

FIGURE 3.3. Probabilités ainsi que leur intervalle de confiance associé.

très précises pour les 10-12 premières estimations et commencent à se détériorer rapidement par la suite.

L'explication concernant le manque de précision pour des $G \geq 10$ est fort simple et réside dans le fait que pour déterminer la $\mathbb{P}(G = k)$, nous devons estimer une intégrale multiple de dimension k dont les bornes sont interdépendantes. Ceci implique que nous devons satisfaire de plus en plus de conditions afin de trouver une observation de dimension k afin d'estimer l'intégrale. Par conséquent, même en disposant de beaucoup de temps et de ressources, cette méthode reste limitée.

De plus, à titre d'indication, $\sum_{k=1}^{20} \mathbb{P}(G = k) \approx 0{,}76$. Ceci indique que les probabilités descendent à un rythme très lent, ce qui n'est pas favorable à ce que nous voulons accomplir. Effectivement, en ayant calculé les 20 premières probabilités, nous avons cumulé seulement 76%, ce qui n'est pas suffisant pour avoir une bonne estimation de l'ARL. La première idée serait d'évaluer les prochaines probabilités, mais ceci serait trop couteux en temps. Nous devons donc penser à une autre approche qui, sur la base des probabilités déjà trouvées, pourrait nous donner une bonne estimation pour les probabilités subséquentes, et ainsi, nous serions en mesure d'estimer l'ARL.

3.7.2.4. *Approche par la loi géométrique*

Lorsque nous sommes rendus à l'essai $P_x + 1$, ceci signifie que les deux densités *a posteriori* que nous comparons sont complètement basées sur des observations corrompues. Donc, la probabilité de rejeter le modèle devrait se rapprocher de $\alpha^* = 0{,}05$ et être relativement stable pour les essais subséquents. Évidemment, comme les W_k sont dépendants entre eux, il se peut que la probabilité de rejeter soit un peu différente de α^*, mais elle devrait s'y rapprocher. En fait, comme G est définie comme le nombre de coups avant de rejeter l'hypothèse, G suivrait une loi géométrique si les épreuves étaient indépendantes. L'idée de la prochaine section est donc de regarder si nous pouvons, malgré la dépendance, modéliser la $\mathbb{P}(G = k), k \geq P_x + 1$, comme étant une loi géométrique. Nous aimerions souligner que pour les P_x premiers essais, nous ne pouvons pas approximer les probabilités par une loi géométrique, car les probabilités varieront de manière considérable à chaque essai. Nous garderons donc les approximations trouvées par les intégrations numériques. La fonction de masse de G se résume donc en

$$\mathbb{P}(G = k) = \begin{cases} p_k & \text{si} \quad k \leq P_x, \\ \left(1 - \sum_{i=1}^{P_x} p_i\right) \times \mathbb{P}(\mathcal{G} = k - P_x) & \text{sinon,} \end{cases} \tag{3.7.4}$$

où les p_i sont les estimations trouvées avec la méthode d'intégration et où $\mathcal{G} \sim Geo(p)$. Nous devons multiplier par $\left(1 - \sum_{i=1}^{P_x} p_i\right)$ la fonction de masse de $\mathcal{G}$ afin que la fonction de masse de G somme à un.

Nous allons maintenant regarder si l'hypothèse de la loi géométrique, pour les essais subséquents à l'essai P_x, est cohérente avec les probabilités obtenues à la section précédente et présentées à la figure 3.3. En prenant le logarithme de la fonction de masse de G pour $k \geq P_x + 1$, nous obtenons

$$\begin{aligned}
&\log\left[\left(1 - \sum_{i=1}^{P_x} p_i\right) \times \mathbb{P}\left(\mathcal{G} = k - P_x\right)\right] \\
&= \log\left[\left(1 - \sum_{i=1}^{P_x} p_i\right) \times p\left(1-p\right)^{k-P_x-1}\right] \\
&= \log\left[\left(1 - \sum_{i=1}^{P_x} p_i\right) \times p\right] + \left(k - P_x - 1\right)\log\left[1-p\right] \\
&= \log\left[\frac{\left(1 - \sum_{i=1}^{P_x} p_i\right) \times p}{1-p}\right] + \left(k - P_x\right)\log\left[1-p\right].
\end{aligned}$$

De plus, en posant

$$\beta_0 = \log\left[\frac{\left(1 - \sum_{i=1}^{P_x} p_i\right) \times p}{1-p}\right]$$

et

$$\beta_1 = \log\left(1-p\right),$$

nous obtenons

$$\log\left[\mathbb{P}\left(G = k\right)\right] = \begin{cases} \log(p_k) & \text{si} \quad k \leq P_x, \\ \beta_0 + (k - P_x)\beta_1 & \text{sinon.} \end{cases}$$

Le logarithme de la fonction de masse doit donc être linéaire en k pour $k \geq P_x + 1$. En regardant la figure 3.4, nous voyons bien que la relation semble être linéaire pour $k \geq P_x + 1$, c'est-à-dire, à droite de la ligne bleue pointillée.

En utilisant la méthode des moindres carrés pour trouver $\widehat{\beta}_0$ et $\widehat{\beta}_1$, il nous serait possible d'obtenir un estimateur pour p. Toutefois, nous devrons considérer une contrainte supplémentaire. En effet, comme $p = 1 - \exp\{\beta_1\}$, nous avons que

$$\beta_0 = \log\left[\left(1 - \sum_{i=1}^{P_x} p_i\right) \times \frac{\left(1 - \exp\{\beta_1\}\right)}{1 - \left(1 - \exp\{\beta_1\}\right)}\right]$$

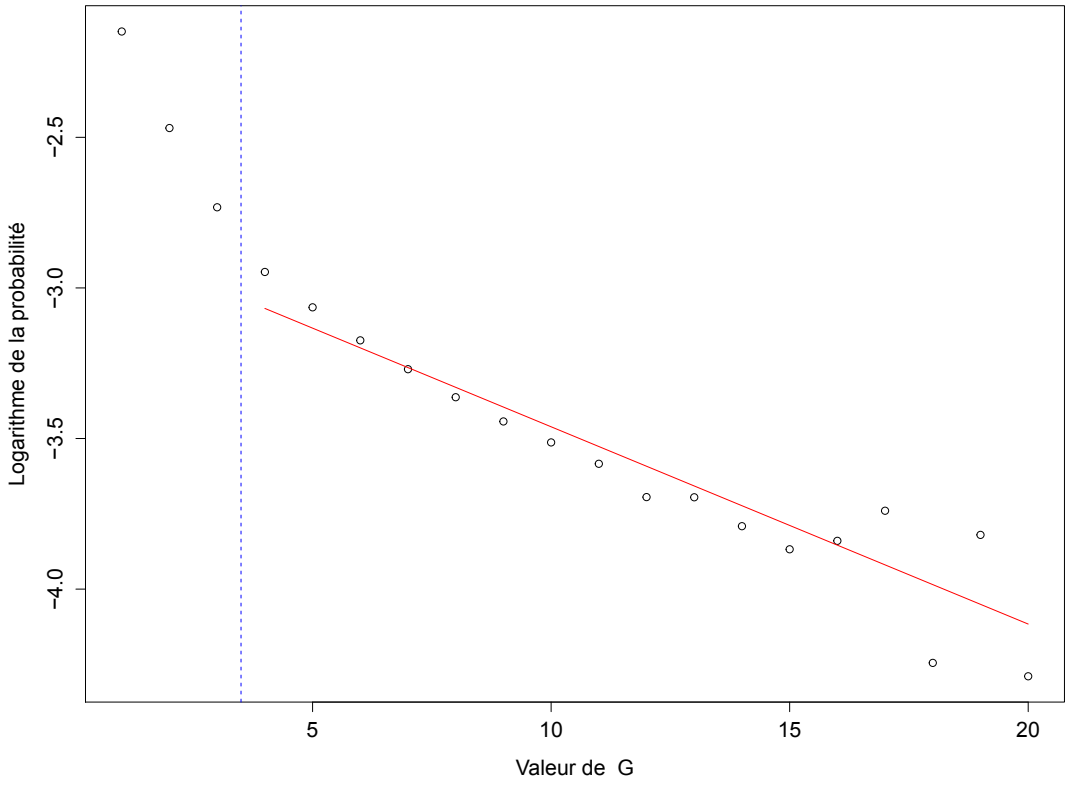

FIGURE 3.4. Graphique du logarithme des probabilités pour les probabilités 1 à 20 et pour $P_x = 3, P_y = 1, c = 1{,}07$.

$$= \log\left[1 - \sum_{i=1}^{P_x} p_i\right] + \log\left(\frac{1}{\exp\{\beta_1\}} - 1\right), \tag{3.7.5}$$

ce qui constitue une contrainte supplémentaire pour le modèle d'optimisation. Nous devons donc minimiser

$$\begin{aligned}
&\sum_{i=P_x+1}^{20} [p_i - \log(\mathbb{P}(G = i))]^2 \\
&= \sum_{i=P_x+1}^{20} [p_i - \beta_0 - (k - P_x)\beta_1]^2 \\
&= \sum_{i=P_x+1}^{20} \left[p_i - \log\left(1 - \sum_{i=1}^{P_x} p_i\right) - \log\left(\frac{1}{\exp\{\beta_1\}} - 1\right) - (k - P_x)\beta_1\right]^2. \qquad (3.7.6)
\end{aligned}$$

En minimisant numériquement l'expression (3.7.6), nous obtenons

$$\widehat{\beta}_1 = -0{,}0655.$$

De plus, en utilisant l'équation (3.7.5), nous obtenons

$$\widehat{\beta}_0 = -3{,}0025,$$

ce qui nous mène à

$$\widehat{p}_{MC} = 0{,}0634.$$

La droite représentée par $\widehat{\beta}_0$ et $\widehat{\beta}_1$ est la droite présentée à la figure 3.4. Comme nous pouvons le voir, l'ajustement est très bon jusqu'à la 16^e observation. Évidemment, le fait que la relation soit moins linéaire pour les quatre dernières estimations reflète la moins bonne qualité de ces estimations.

En regardant la figure 3.5, nous pouvons apprécier la qualité de l'ajustement (représentée par la courbe rouge) pour la fonction de masse de G pour $k \geq P_x + 1$. La méthode semble donc être efficace.

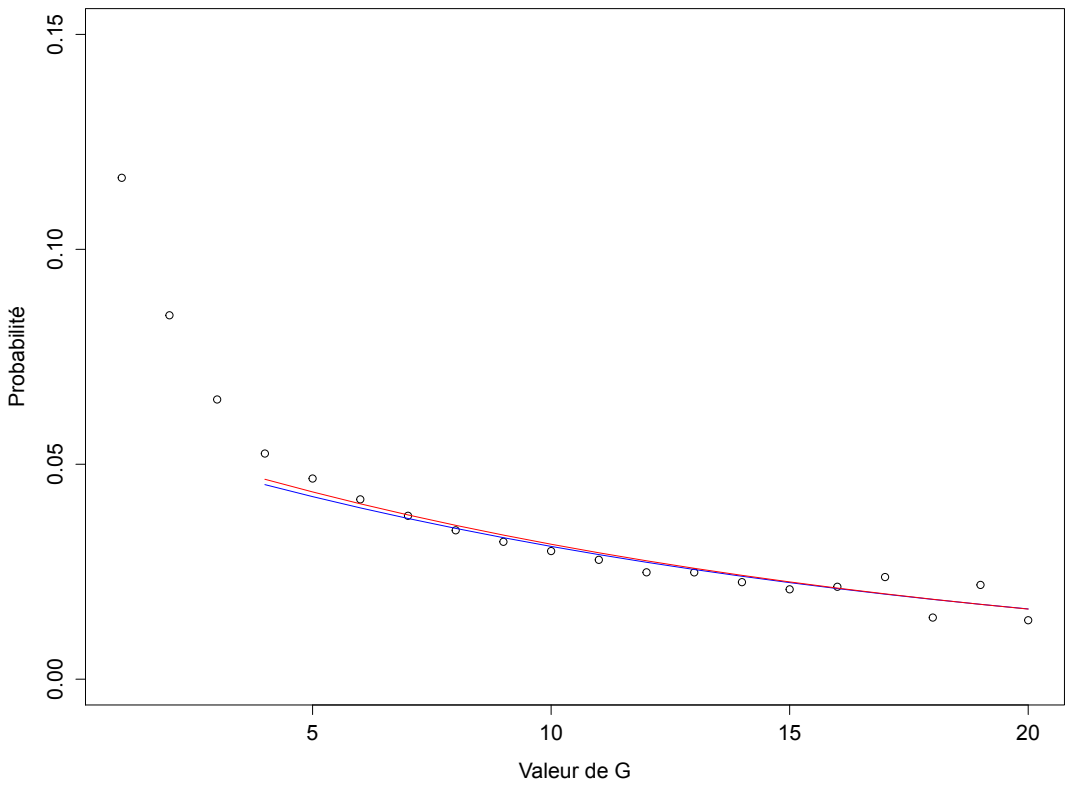

FIGURE 3.5. Graphique des estimations des probabilités avec les ajustements géométriques. En rouge approche par la méthode des moindres carrés ($\widehat{p}_{MC} = 0{,}0634$) et en bleu l'approche des ratios ($\widehat{p}_{Ratio} = 0{,}0617$) .

Toujours en se basant sur les propriétés de la loi géométrique, nous pouvons facilement trouver un autre estimateur de p. En fait, en regardant l'expression (3.7.4), nous pouvons noter que pour $k \geq P_x + 1$, nous avons

$$\begin{aligned}\frac{\mathbb{P}(G = k+1)}{\mathbb{P}(G = k)} &= \frac{\left(1 - \sum_{i=1}^{P_x} p_i\right)}{\left(1 - \sum_{i=1}^{P_x} p_i\right)} \times \frac{p(1-p)^{P_x - k}}{p(1-p)^{P_x - k - 1}} \\ &= 1 - p.\end{aligned}$$

Ceci implique que le ratio de deux probabilités consécutives devrait être stable et égale à $1 - p$. En regardant la figure 3.6, nous pouvons voir que les ratios sont relativement constants jusqu'aux estimations qui sont moins précises (16-20). À cette figure, la ligne rouge représente la moyenne des ratios obtenus. Nous allons donc utiliser la moyenne des ratios de probabilités obtenues pour acquérir notre deuxième estimateur $\widehat{p}_{\text{Ratio}}$. En appliquant cette méthode, nous obtenons $\widehat{p}_{\text{Ratio}} = 0{,}0617$. Encore une fois, nous pouvons apprécier la qualité

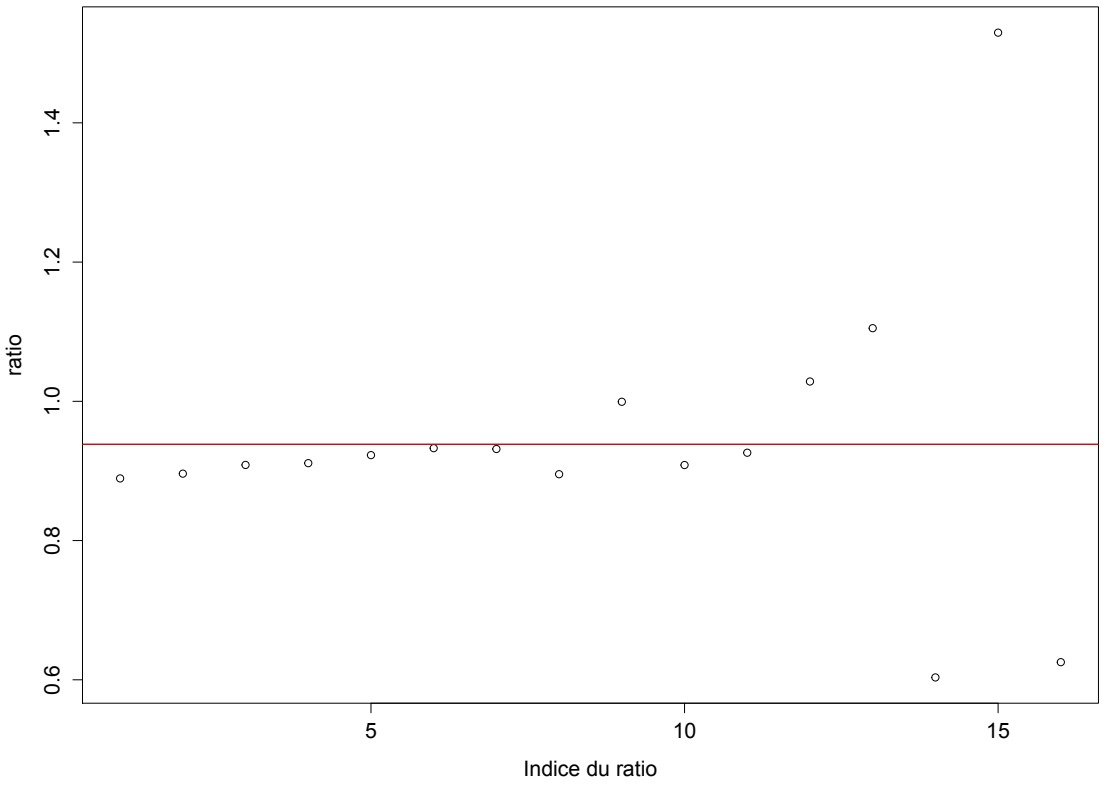

FIGURE 3.6. Graphique des ratios pour les probabilités 4 à 20.

de l'ajustement pour l'estimation de la fonction de masse en prenant $\widehat{p}_{\text{Ratio}}$ représentée par la courbe bleue sur la figure 3.5.

Comme nous pouvons le voir, les deux estimations proposées pour p sont de bons candidats. De plus, comme nous pouvions nous en attendre, ils sont proches de $\alpha^* = 0{,}05$. Nous avons donc que l'hypothèse selon laquelle la fonction de masse de G peut s'écrire comme à l'équation (3.7.4) est plausible. Ceci signifie donc que la dépendance influence moins le résultat que nous le pensions et la raison de ce phénomène sera examinée dans la prochaine section.

Il reste maintenant à trouver l'ARL. En regardant la fonction de masse de G présentée à l'équation (3.7.4), nous avons que

$$\begin{aligned}
\mathrm{ARL} &= \mathrm{E}[G] \\
&= \sum_{k=1}^{\infty} k \times \mathbb{P}(G = k) \\
&= \sum_{k=1}^{P_x} k \times p_k + \left(1 - \sum_{k=1}^{P_x} p_k\right) \times \left(\sum_{k=P_x+1}^{\infty} k \times \mathbb{P}(\mathcal{G} = k - P_x)\right). \qquad (3.7.7)
\end{aligned}$$

En utilisant l'équation (3.7.7), nous trouvons les estimations suivantes pour l'ARL($c = 1{,}07, P_x = 3, P_y = 1$)

$$\widehat{\mathrm{ARL}}_{MC}(c = 1{,}07, P_x = 3, P_y = 1) = 14{,}25,$$
$$\widehat{\mathrm{ARL}}_{Ratio}(c = 1{,}07, P_x = 3, P_y = 1) = 14{,}57.$$

À titre de comparaison, nous avons procédé à une simulation et nous avons obtenu une $\widehat{\mathrm{ARL}}_{Simu}(c = 1{,}07, P_x = 3, P_y = 1) = 15{,}60$, ce qui est assez proche des résultats théoriques que nous avons obtenus suite à plusieurs approximations. Notons que les détails concernant les simulations seront expliqués dans la section 3.7.3.

3.7.2.5. *Pourquoi l'approche géométrique fonctionne-t-elle ?*

Suite aux résultats obtenus à la section précédente, il semblerait que la dépendance n'affecte pas la loi de G pour $k \geq P_x + 1$. L'objectif de cette section sera d'expliquer ce phénomène.

À la section 3.7.2.1, nous avons montré que

$$U_k \sim \mathrm{Beta}'(aN_y, \alpha + aN_x), \quad \forall k$$

et comme

$$U_k = W_k \times C_k,$$

nous avons que

$$f_{W_k}(w_k) = f_{U_k}(w_k \times C_k) \times C_k$$

$$= \frac{\Gamma(\alpha + aN_x + aN_y)}{\Gamma(aN_y)\,\Gamma(\alpha + aN_x)} \times \frac{(w_k C_k)^{\alpha + aN_x - 1}}{(1 + w_k C_k)^{\alpha + aN_x + aN_y}} \times C_k,$$

où C_k est définie comme à l'équation (3.7.3). Ceci implique que $W_k \sim Beta'(aN_y, \alpha + aN_x, 1, 1/C_k)$. En somme, la loi de W_k dépend des C_k, qui dépendent des P_x années d'observations passées. La variable aléatoire W_k change donc théoriquement de loi pour chaque k. Or, à l'équation (3.7.3), nous avons trouvé que

$$C_k = \frac{\left[\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i\right]}{c\left[\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i \ell_i\right]}.$$

Cette expression se simplifie en

$$C_k = \frac{\left[\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i\right]}{\left[c\beta + \sum_{i=k}^{P_x+k-1} \mathcal{S}_i\right]},$$

lorsque $k \geq P_x + 1$, car toutes les années de la densité *a posteriori* sont corrompues. Ceci implique que toutes les C_k suivent la même loi pour $k \geq P_x + 1$. Or, β étant fixe, ceci indique que la loi des C_k dépend seulement de c pour $k \geq P_x + 1$. Au tableau 3.5, nous pouvons voir l'espérance ainsi que la variance théorique de la variable aléatoire C_k ($k \geq P_x + 1$) en fonction de c. L'espérance et la variance ont été estimées par la méthode de Monte-Carlo en utilisant un million d'itérations. Comme nous pouvons voir, la variance de C_k est très petite

TABLEAU 3.5. Espérances et variances théoriques de C_k pour différentes valeurs de c et pour $k \geq P_x + 1$.

c	Espérance	Variance
1	1,000	0,000
1,01	0,999	$1{,}960 \times 10^{-9}$
1,03	0,998	$1{,}697 \times 10^{-8}$
1,07	0,996	$8{,}578 \times 10^{-8}$
1,1	0,994	$1{,}658 \times 10^{-7}$
1,13	0,993	$2{,}648 \times 10^{-7}$
1,27	0,987	$9{,}073 \times 10^{-7}$
1,34	0,985	$1{,}290 \times 10^{-6}$
1,4	0,983	$1{,}636 \times 10^{-6}$

et ceci implique que la valeur de C_k variera très peu. Cela nous donne que les $W_k \sim Beta'(aN_y, \alpha + aN_x, 1, 1/C^*)$, où C^* est considérée comme une constante

fixe pour $k \geq P_x + 1$. Nous pouvons donc faire l'hypothèse que

$$W_k \sim \begin{cases} \text{Beta}'(aN_y, \alpha + aN_x, 1, 1/C_k) & \text{si } k \leq P_x, \\ \text{Beta}'(aN_y, \alpha + aN_x, 1, 1/C^*) & \text{sinon.} \end{cases}$$

Nous avons donc que les W_k suivent toutes la même loi pour $k \geq P_x + 1$. Ceci implique que les W_k sont approximativement identiquement distribuées pour $k \geq P_x + 1$. En somme, si $k > P_x$ et $\ell > P_x$, nous avons que

$$\begin{aligned} \mathbb{P}(W_k > W_0) &= \mathbb{P}(W_\ell > W_0) \\ &= p, \end{aligned}$$

qui est le paramètre de la loi géométrique. La probabilité de rejeter est donc approximativement égale à chaque essai (pour $k \geq P_x + 1$).

3.7.2.6. *Résumé et limitation*

En résumé, il est possible d'estimer l'ARL de manière théorique. Il suffit de disposer d'assez de puissance de calcul afin de pouvoir estimer de manière précise les P_x premières probabilités ainsi que quelques autres probabilités subséquentes, afin de pouvoir estimer le paramètre p de la loi géométrique. Par la suite, il suffit de calculer l'ARL à l'aide de l'expression (3.7.7).

Bien que cette méthode fonctionne, il pourrait être difficile de l'appliquer en pratique. En effet, plus P_x sera grand, plus le temps de calcul sera important pour avoir une bonne estimation du paramètre p de la loi géométrique.

3.7.3. ARL simulée

Dans cette section, nous nous consacrerons au calcul de l'ARL pour les deux scénarios en utilisant la simulation. Comme l'ARL dépendra de P_x, P_y et c, nous la noterons $\text{ARL}^{\ell}_{\text{Simu}}(P_x, P_y, c)$, où $\ell = 1$ pour le premier scénario et $\ell = 2$ pour le deuxième scénario.

3.7.3.1. *Scénario S1*

Comme expliqué plus tôt, pour ce scénario, la moyenne des observations sera multipliée par un facteur c à partir du point de rupture. Les observations qui seront corrompues, suivront donc la loi gamma de paramètre 1 et θ/c.

Afin d'utiliser des paramètres similaires à l'approche classique, nous avons exprimer c en fonction de δ, où δ est celui défini à l'équation (2.2.3) et utilisé afin de quantifier l'augmentation de la moyenne en écart-type. Nous avons donc que $c(\delta) = 1 + \delta/\sqrt{n}$, où le n sera considéré comme étant fixe et égal à

55 pour les simulations. En d'autres mots, chaque année (été) aura 55 jours de précipitations.

Voici comment nous avons procédé pour les simulations. Pour chaque itération, nous avons fait les étapes suivantes :

(1) Nous avons créé 10 années d'observations non corrompues de moyenne μ^*, où μ^* a été fixé à la moyenne des 10 premières années d'observations du jeu de données Ouranos.

(2) Nous avons estimé (α, β), les hyper-paramètres, en utilisant la méthode des moments bayésiens présentée à la section 3.6, en utilisant les années non corrompues.

(3) Nous avons créé 100 années d'observations de moyenne corrompue par un facteur c. Ce nombre pouvant être ajusté dans le cas où nous avions besoin de plus d'observations.

(4) Nous avons fait avancer la fenêtre mobile jusqu'à ce qu'on détecte la non-stationnarité et nous avons noté le nombre d'essais. En résumé, nous avons donc calculé $W_1, ..., W_k$, jusqu'à ce que $W_k \geq W_0$.

L'$ARL^1_{Simu}(P_x, P_y, c)$ constitue la moyenne du nombre d'essais avant qu'on détecte la non-stationnarité. Pour chaque combinaison possible de paramètres, nous avons procédé à 10^7 itérations. Les résultats obtenus sont présentés au tableau 3.6.

Lorsque $c = 1$, nous sommes théoriquement sous le modèle M_0. De ce fait, si les essais étaient indépendants entre eux, l'ARL devrait être très près de 20 $(1/\alpha^*)$ pour n'importe quelle combinaison de P_x et P_y. Comme nous l'avons souligné à la section précédente, ceci n'est pas le cas. En effet, plus P_y augmente, plus la dépendance est importante et moins le niveau est respecté. Afin de choisir des paramètres qui respectent le plus possible le niveau, nous allons nous limiter au cas où $P_y = 1$ ou $P_y = 2$. De plus, nous pouvons observer que la variation en P_x est surtout importante lorsque nous passons de $P_x = 10$ à $P_x = 20$, mais que sinon, elle n'est pas très importante. C'est pourquoi, nous allons restreindre notre étude aux cas $P_x = 10$ et $P_x = 20$.

Afin de mieux comprendre les résultats obtenus dans le tableau 3.6, nous avons décidé de procéder à un exemple. En fait, pour c=1,07, $P_x = 20$ et $P_y = 1$, ceci signifie que nous avons besoin de $P_x + P_y + ARL = 31{,}33$ années d'observations avant d'observer le changement. De plus, dans le cas où $c = 1{,}07$, $P_x = 20$ et $P_y = 2$, nous allons avoir besoin de 32,83 années d'observations avant de détecter un changement de 7% dans la moyenne.

TABLEAU 3.6. ARL simulée pour le scénario S1 et pour différents c, P_x, P_y.

(A) $\delta = 0, c = 1{,}00$.

P_x, P_y	1	2	3	4	5
10	18,06	21,48	25,23	28,90	32,43
20	18,56	22,19	26,18	30,08	33,94
30	18,93	22,79	26,90	30,96	34,93
40	19,19	23,22	27,51	31,68	35,80
50	19,37	23,53	27,95	32,27	36,44

(B) $\delta = 0{,}10, c = 1{,}01$.

P_x, P_y	1	2	3	4	5
10	17,17	20,26	23,69	27,00	30,09
20	17,29	20,43	23,96	27,45	30,83
30	17,33	20,54	24,11	27,62	31,04
40	17,36	20,60	24,18	27,70	31,16
50	17,37	20,65	24,23	27,77	31,19

(C) $\delta = 0{,}25, c = 1{,}03$.

P_x, P_y	1	2	3	4	5
10	15,43	17,83	20,61	23,27	25,74
20	14,79	16,96	19,58	22,21	24,76
30	14,35	16,34	18,76	21,17	23,52
40	14,03	15,89	18,12	20,35	22,53
50	13,81	15,54	17,65	19,74	21,77

(D) $\delta = 0{,}50, c = 1{,}07$.

P_x, P_y	1	2	3	4	5
10	12,02	13,10	14,64	16,13	17,48
20	10,33	10,83	11,81	12,83	13,83
30	9,43	9,62	10,21	10,86	11,48
40	8,91	8,92	9,29	9,70	10,10
50	8,60	8,51	8,75	9,01	9,25

(E) $\delta = 0{,}75, c = 1{,}10$.

P_x, P_y	1	2	3	4	5
10	9,66	9,90	10,64	11,40	12,05
20	7,67	7,37	7,54	7,76	8,00
30	6,79	6,30	6,21	6,17	6,15
40	6,37	5,80	5,59	5,44	5,29
50	6,13	5,54	5,27	5,05	4,85

(F) $\delta = 1{,}00, c = 1{,}13$.

P_x, P_y	1	2	3	4	5
10	7,57	7,21	7,37	7,58	7,76
20	5,65	4,96	4,71	4,55	4,43
30	4,96	4,24	3,87	3,60	3,37
40	4,67	3,95	3,55	3,23	2,97
50	4,52	3,80	3,39	3,06	2,79

(G) $\delta = 2{,}00, c = 1{,}27$.

P_x, P_y	1	2	3	4	5
10	2,35	1,71	1,45	1,31	1,22
20	1,92	1,43	1,22	1,12	1,06
30	1,85	1,38	1,18	1,09	1,04
40	1,82	1,36	1,17	1,08	1,04
50	1,80	1,35	1,16	1,07	1,03

(H) $\delta = 2{,}50, c = 1{,}34$.

P_x, P_y	1	2	3	4	5
10	1,60	1,22	1,10	1,05	1,02
20	1,47	1,15	1,05	1,02	1,01
30	1,44	1,13	1,04	1,01	1,00
40	1,43	1,12	1,04	1,01	1,00
50	1,42	1,12	1,03	1,01	1,00

(I) $\delta = 3{,}00, c = 1{,}40$.

P_x, P_y	1	2	3	4	5
10	1,32	1,08	1,03	1,01	1,00
20	1,27	1,06	1,01	1,00	1,00
30	1,26	1,05	1,01	1,00	1,00
40	1,25	1,05	1,01	1,00	1,00
50	1,25	1,05	1,01	1,00	1,00

À la figure 3.7, nous pouvons apercevoir le comportement de l'ARL pour différentes combinaisons de P_x et P_y en fonction de c. Il est à noter que les lignes verticales représentent les différentes valeurs de c que nous avons utilisées pour les simulations. Règle générale, nous pouvons observer que les lignes

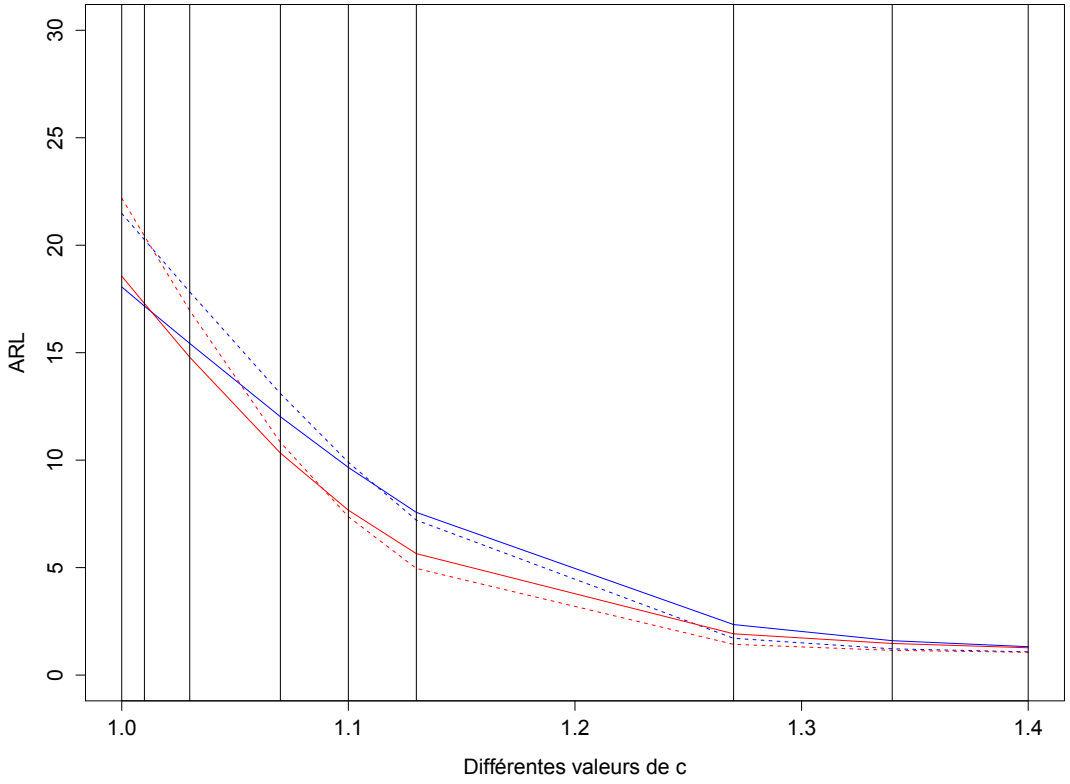

FIGURE 3.7. Graphique de l'ARL pour différentes valeurs de P_x et P_y en fonction de c. Les courbes bleues représentent le cas $P_x = 10$ et les courbes rouges représentent le cas $P_x = 20$. Les courbes pleines représentent le cas $P_y = 1$ et les courbes pointillées représentent le cas $P_y = 2$.

rouges sont inférieures aux lignes bleues. Nous utiliserons donc $P_x = 20$. Pour $P_x = 20$, nous avons que la différence entre les courbes $P_y = 1$ et $P_y = 2$ est très petite. Nous allons donc sélectionner $P_y = 1$. Finalement, les paramètres $P_x = 20$ et $P_y = 1$ seront ceux utilisés dans le cadre de notre analyse.

3.7.3.2. *Scénario S2*

Pour ce scénario, nous avons qu'à partir du point de rupture, la moyenne des observations augmente linéairement en fonction des années selon le modèle présenté à l'équation (2.1.1).

Voici comment nous avons procédé pour les simulations. Pour chaque itération, nous avons procédé aux étapes suivantes :

(1) Nous avons créé 140 années d'observations de moyenne μ_i, où μ_i est définie selon l'équation (2.1.1) avec $t_0 = 10$. Le nombre d'années générées pouvant être augmenté au besoin.

(2) Nous avons estimé (α, β), les hyper-paramètres, en utilisant la méthode des moments bayésiens présentée à la section 3.6, en utilisant les 10 premières années.

(3) Nous avons fait avancer la fenêtre mobile jusqu'à ce qu'on détecte la non-stationnarité et nous avons noté le nombre d'essais. En résumé, nous avons donc calculé $W_1, ..., W_k$, jusqu'à ce que $W_k \geq W_0$.

L'$ARL^2_{Simu}(P_x, P_y)$ constitue la moyenne du nombre d'essais avant qu'on détecte la non-stationnarité. Pour chaque combinaison possible de paramètres, nous avons procédé à 10^7 itérations. Les résultats obtenus sont présentés au tableau 3.7. Comme nous pouvons le remarquer, nous avons encore une fois

TABLEAU 3.7. ARL simulée pour le scénario S2 et pour différents P_x, P_y.

P_x, P_y	1	2	3	4	5
10	16,23	18,36	20,55	22,41	23,91
20	15,98	17,94	19,94	21,65	23,07
30	15,83	17,70	19,55	21,09	22,34
40	15,73	17,54	19,29	20,73	21,88
50	15,65	17,43	19,12	20,51	21,59

que la variation de l'ARL en P_x est surtout importante lorsque nous passons de $P_x = 10$ à $P_x = 20$, mais que sinon, elle n'est pas très importante. Nous allons donc restreindre le paramètre P_x aux valeurs 10 et 20. De plus, nous pouvons observer que l'ARL, pour $P_x = 20$, est inférieur à l'ARL pour $P_x = 10$. Nous allons donc sélectionner $P_x = 20$ tout comme dans le premier scénario. Il nous reste donc à sélectionner la valeur de P_y. En regardant plus attentivement la ligne $P_x = 20$, nous pouvons voir que l'ARL augmente en fonction de P_y. Nous choisirons donc, encore une fois, le couple $P_x = 20$ et $P_y = 1$.

Finalement, nous sommes arrivés à la conclusion que les paramètres $P_x = 20$ et $P_y = 1$ sont les paramètres optimaux pour détecter la non-stationnarité sur la base du critère de l'ARL. Cette conclusion étant valable pour le premier et deuxième scénario.

3.8. EXEMPLE PRATIQUE POUR L'APPROCHE BAYÉSIENNE

Cette section aura pour objectif de montrer une application de la méthode que nous avons développée sur un petit jeu de données. Pour ce faire, nous avons utilisé le même jeu de données que pour les approches par carte de contrôle et qui a été résumé au tableau 2.3. Pour cet exemple, nous avons utilisé les paramètres $P_x = 10$ et $P_y = 1$. De plus, notons qu'il y a 55 observations par année.

Premièrement, afin de pouvoir utiliser l'approche bayésienne, nous devons estimer les hyper-paramètres. Pour ce faire, nous utiliserons la méthode des moments bayésiens, présentée à la section 3.6, sur les dix premières années d'observations (non corrompues). En appliquant cette méthode, nous obtenons $\widehat{\alpha} = 71{,}38$ et $\widehat{\beta} = 670{,}12$.

Deuxièmement, nous devons calculer les différents W_k et vérifier si $W_k \geq W_0$, où W_0 est le quantile de la loi $\texttt{Beta}'(N_y, \widehat{\alpha} + N_x)$ associé à un niveau $\alpha^* = 0{,}05$. De plus, rappelons que $N_x = P_x * 55 = 550$ et que $N_y = P_y * 55 = 55$. Avec ces paramètres, nous obtenons $W_0 = 0{,}11$. Au tableau 3.8, nous pouvons voir les différentes valeurs de W obtenues. D'ailleurs, à titre de démonstration, voici comment nous avons trouvé W_1 :

$$\begin{aligned} W_1 &= \frac{N_y \times \overline{x}_{11\cdot}}{\widehat{\beta} + N_x \times \sum_{i=1}^{10} \overline{x}_{i\cdot}} \\ &= \frac{55 \times 7{,}90}{670{,}12 + 550 \times (9{,}96 + 7{,}72 + \ldots + 10{,}02)} \\ &= 0{,}074. \end{aligned}$$

Avec la méthode bayésienne, nous pouvons voir que nous rejetons dès le deuxième coup. À la figure 3.8, nous pouvons apprécier visuellement la valeur des W_k ainsi que la variable W_0 qui est représentée à l'aide de la ligne rouge horizontale.

TABLEAU 3.8. Différentes valeurs de W_k.

k	1	2	3	4	5	6	7	8	9	10	11
W_k	0,074	0,110	0,101	0,105	0,095	0,087	0,087	0,130	0,080	0,081	0,080

Comme nous l'avons mentionné précédemment, l'approche bayésienne est de moins en moins performante au fur et à mesure que la fenêtre mobile avance dans le temps. En effet, plus la fenêtre mobile avance dans le temps, plus le vecteur d'observations $\underset{\sim}{x}$ est corrompu. Ceci fait en sorte que nous comparons deux densités *a posteriori* qui se ressemblent de plus en plus. À la figure 3.8, nous pouvons observer la performance décroissante de la méthode lorsque

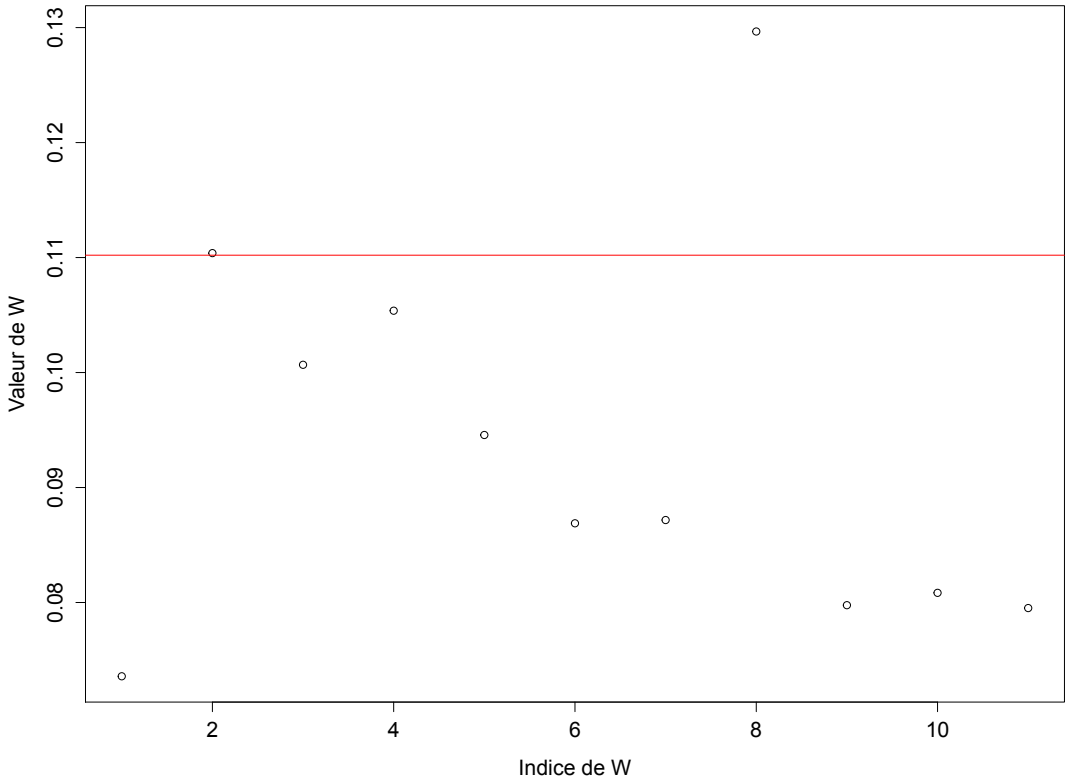

FIGURE 3.8. Graphique des différents W.

la fenêtre mobile progresse dans le temps. Nous aimerions souligner que la brusque augmentation de la valeur pour W_8 est due à une moyenne de 14,51 pour la 18^{e} années d'observations, ce qui constitue de loin le maximum des autres moyennes rencontrées (la deuxième plus grande étant 11,63).

Chapitre 4

SIMULATION ET EXEMPLE AVEC DONNÉES OURANOS

Dans cette section, nous allons, en premier lieu, comparer l'approche classique et bayésienne à l'aide des simulations et en utilisant les paramètres optimaux trouvés dans les sections précédentes. Par la suite, nous procéderons à l'analyse des véritables données Ouranos.

4.1. COMPARAISON DES APPROCHES CLASSIQUE ET BAYÉSIENNE

Au cours des sections précédentes, nous avons tenté de calculer les ARL de façon théorique et nous avons comparé les résultats obtenus avec ceux obtenus lors des simulations. Comme nous l'avons remarqué, les résultats théoriques se rapprochent, dans une certaine mesure, des résultats de simulations, mais comme plusieurs hypothèses étaient plus ou moins satisfaites, il était normal que nous n'obtenions pas une parfaite cohésion entre la théorie et les simulations. Par conséquent, nous allons nous baser sur les résultats des simulations.

De plus, par le passé, nous avons calculé de manière indépendante les ARL pour les trois méthodes, c'est-à-dire, les jeux de données utilisées étaient différents pour chaque méthode. Afin de comparer l'ARL d'une manière un peu plus concise, nous avons décidé de procéder à une dernière simulation qui fera en sorte que les méthodes seront comparées sur les mêmes jeux de données. Voici comment nous avons procédé pour cette simulation. Pour chaque itération, nous avons procédé aux étapes suivantes :

(1) Nous avons créé 140 années d'observations de moyenne conforme au scénario choisi (voir la section 2.1.5 pour de l'information sur les scénarios) avec $t_0 = 10$.

(2) Nous avons estimé (α, β), les hyper-paramètres, en utilisant la méthode des moments bayésiens présentée à la section 3.6, en utilisant les 10 premières années.

(3) Nous avons estimé μ_0 et par le fait même σ^*, en utilisant l'équation (2.2.2) sur les 10 premières années d'observations.

(4) Pour les approches classiques : nous avons appliqué la méthode du contrôle de qualité standard ainsi que du CUSUM et nous avons noté à partir de quel essai leur seuil respectif a été dépassé.

(5) Pour l'approche bayésienne : nous avons fait avancer la fenêtre mobile jusqu'à ce qu'on détecte la non-stationnarité et nous avons noté le nombre d'essais. En résumé, nous avons donc calculé $W_1, ..., W_k$, jusqu'à ce que $W_k \geq W_0$.

De plus, les paramètres utilisés pour cette simulation sont les paramètres optimaux trouvés dans les sections précédentes. En somme, pour l'approche classique du CUSUM, nous avons fixé $k = 0{,}7$ et $h = 1{,}1$, alors que pour l'approche bayésienne, nous avons fixé $P_x = 20$ et $P_y = 1$. Notons que le contrôle de qualité standard n'avait aucun paramètre à fixer. Finalement, pour chacun des scénarios, nous avons procédé à un total de 10^8 itérations.

De plus, comme dans les sections précédentes, nous avons fait les simulations en supposant que $a = 1$ et donc, que les données étaient de loi exponentielle.

4.1.1. Scénario S1

Les ARL obtenues pour la simulation du scénario S1 sont présentées au tableau 4.1.

TABLEAU 4.1. ARL simulée pour le scénario S1 avec $P_x = 20, P_y = 1, k = 0{,}7$ et $h = 1{,}1$, pour différents c.

c	1	1,01	1,03	1,07	1,1	1,13	1,27	1,34	1,4
δ	0	0,10	0,25	0,50	0,75	1,00	2,00	2,50	3,00
Bayes	18,56	16,83	14,35	10,60	7,59	5,38	1,92	1,48	1,26
Standard	27,38	21,56	15,48	9,48	6,32	4,47	1,84	1,46	1,25
CUSUM	26,30	20,18	14,03	8,34	5,47	3,90	1,76	1,43	1,25

Pour la colonne $c = 1$, nous nous attendions à avoir des ARL autour de 20. Comme nous pouvons le voir, la méthode bayésienne est légèrement au-dessous de l'ARL désirée alors que les deux méthodes classiques sont au-dessus. Lorsque c augmente, nous avons que l'approche bayésienne performe

légèrement moins bien que les deux approches classiques qui sont très proches. Évidemment, lorsque c est grand, les trois approches détectent le changement très vite.

4.1.2. Scénario S2

Pour le scénario S2, les résultats de la simulation sont présentés au tableau 4.2. Tel que nous pouvons le voir, les ARL sont très près l'un de l'autre. Ce qui

TABLEAU 4.2. ARL pour le scénario deux avec $P_x = 20, P_y = 1, k = 0{,}7$ et $h = 1{,}1$.

	Bayes	Standard	CUSUM
ARL	15,96	17,16	15,95

laisse sous-entendre que les trois méthodes donnent des résultats similaires pour ce scénario.

4.2. EXEMPLE AVEC DONNÉES OURANOS

Dans cette section, nous allons finalement utiliser les outils que nous avons développés dans ce mémoire sur le véritable jeu de données Ouranos. Pour ce faire, nous commencerons par traiter de l'approche classique et nous suivrons avec l'approche bayésienne.

4.2.1. Quelques précisions sur les outils

Les données Ouranos comportent 140 années d'observations allant de 1961 à 2100. Comme nous l'avons expliqué, les dix premières années seront utilisées pour estimer les hyper-paramètres de l'approche bayésienne ainsi que la moyenne et l'écart-type des méthodes classiques. De plus, pour ce qui est de la méthode bayésienne, nous avons besoin de $P_x = 20$ années d'observations pour pouvoir modéliser la première densité *a posteriori*. Ceci fait en sorte que l'approche bayésienne peut commencer à détecter la non-stationnarité à partir de 1991 alors que l'approche classique peut le faire à partir de l'année 1971.

Aussi, dans les sections précédentes, le nombre d'observations par années était fixe. Évidemment, ceci ne sera pas le cas avec les données Ouranos. En effet, rappelons que les observations dont nous disposons sont seulement les journées où il y a eu des précipitations. Ceci fera en sorte que les seuils respectifs de chaque méthode ne seront plus constants. Plus précisément, W_0 qui est le quantile de la loi $\mathrm{Beta}'(aN_y, \alpha + aN_x)$ dépend du nombre d'observations à travers N_x et N_y. Pour ce qui est de l'approche classique, nous avons que leur

seuil respectif, à chaque essai, dépend de $\sigma^* = \sigma/\sqrt{n}$, où le n est le nombre de journées de précipitations pour l'année que nous testons.

Dans les chapitres précédents, nous avons supposé que $\alpha = 1$ pour la totalité des simulations. Toutefois, dans cette section, nous utiliserons $\alpha = 0{,}8$, cette valeur ayant été obtenue par la méthode du maximum de vraisemblance.

4.2.2. Contrôle de qualité standard

En utilisant le contrôle de qualité standard, nous avons obtenu les résultats présentés à la figure 4.1. Comme nous pouvons le voir, il y a des observations dépassant le seuil dès l'année 1981. Néanmoins, ce n'est qu'à partir de l'année 1994 que nous pouvons observer une forte proportion d'années au-dessus du seuil. À cet effet, le seuil est défini comme à l'équation (2.2.1). De plus, nous aimerions souligner que la ligne verticale bleue est la ligne qui délimite à partir de où l'approche bayésienne peut commencer à détecter. En résumé, nous pouvons observer qu'il y a, sans aucun doute, une augmentation dans la moyenne et nous considérerons que 1994 est l'année où nous avons quitté la stationnarité.

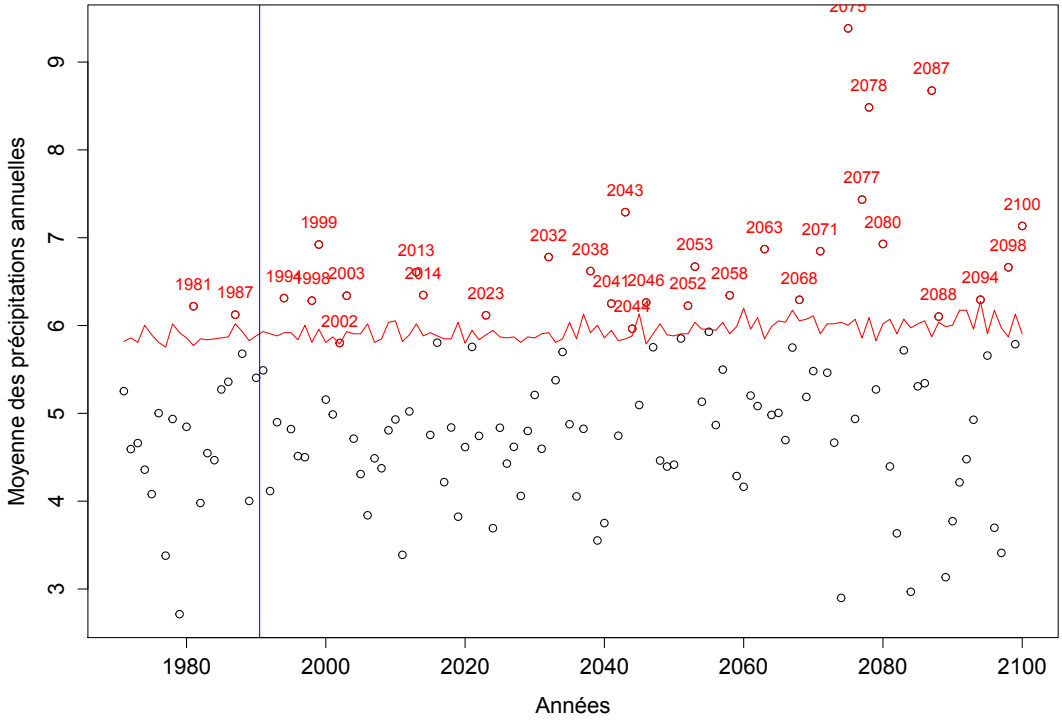

FIGURE 4.1. Graphique pour la carte de contrôle standard.

4.2.3. Contrôle de qualité CUSUM

Nous allons maintenant passer à l'application du CUSUM sur le jeu de données Ouranos. Comme nous pouvons le voir à la figure 4.2, il y a une très forte proportion d'années qui se retrouvent au-dessus du seuil. À cet effet, rappelons que le seuil est défini comme étant $h\sigma^*$. De plus, nous pouvons apprécier une tendance croissante, ce qui indique que la moyenne du modèle a clairement augmenté. Encore une fois, le point de rupture se situe proche de 1994.

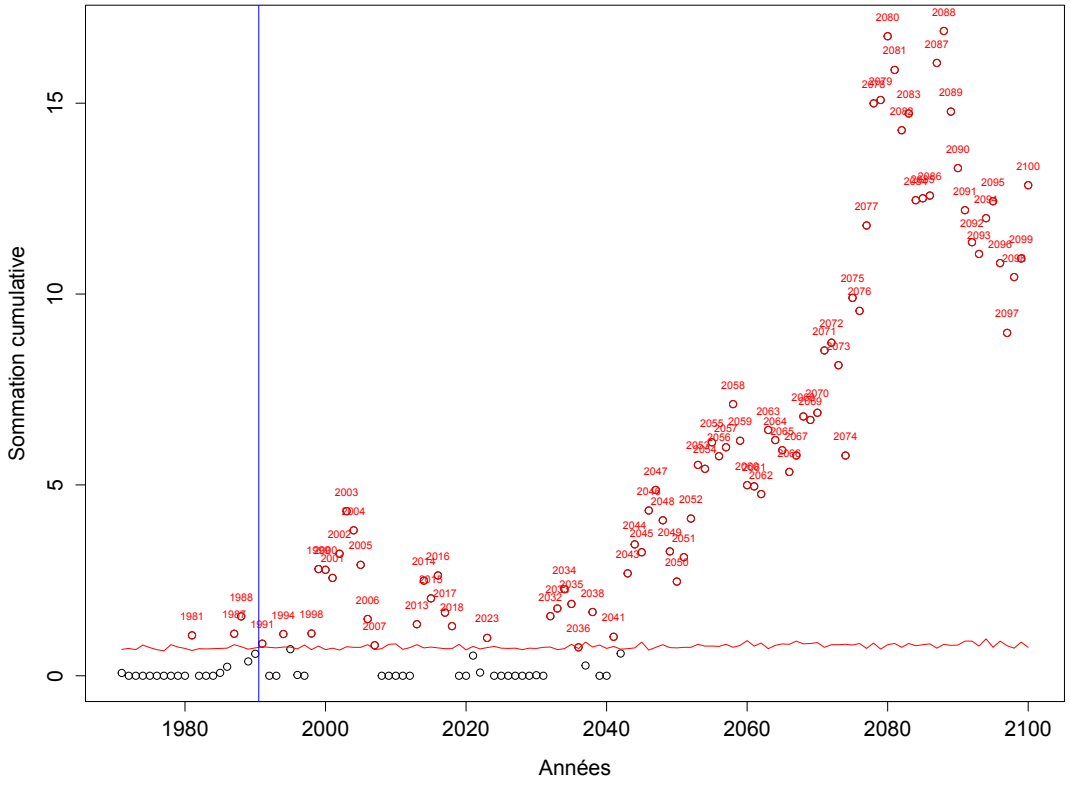

FIGURE 4.2. Graphique pour la carte de contrôle CUSUM.

4.2.4. Approche bayésienne

Nous allons maintenant regarder les résultats obtenus lorsqu'on applique la méthode bayésienne que nous avons conçue, sur le jeu de données Ouranos. En regardant la figure 4.3, nous pouvons voir qu'il y a plusieurs points au-dessus du seuil. À cet effet, rappelons que le seuil est défini comme le quantile 5% de la loi $\text{Beta}'(aN_y, \alpha + aN_x)$. Nous pouvons encore une fois conclure, sur la base de la figure, qu'il y a eu une augmentation dans la moyenne à partir de l'année 1994.

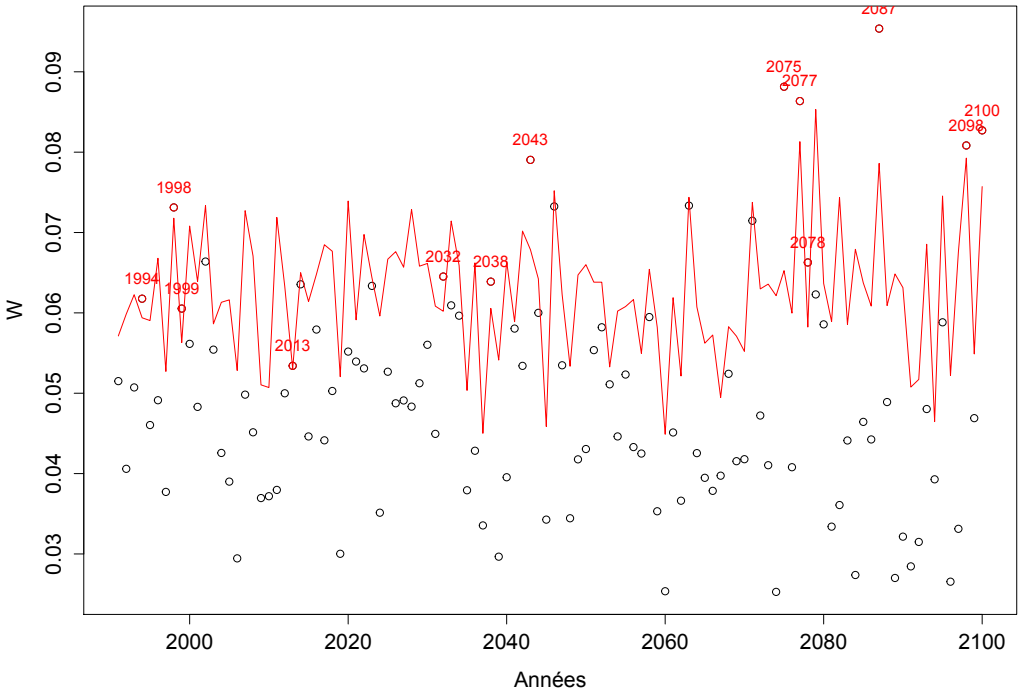

FIGURE 4.3. Graphique de la méthode bayésienne.

4.2.5. Résumé

En résumé, les trois méthodes indiquent une augmentation dans la moyenne au cours du temps. Aussi, pour ce qui est de l'année à partir de laquelle le système a changé, les trois méthodes supportent le fait que cela s'est produit proche de l'année 1994 comme nous pouvons le voir au tableau 4.3.

TABLEAU 4.3. Points de rupture potentiels avant les années 2000.

Standard	CUSUM	Bayésien
1981	1981	-
1987	1987	-
-	1988	-
-	1991	-
1994	1994	1994
1998	1998	1998
1999	1999	1999

Chapitre 5

CONCLUSION

Dans ce mémoire, nous avons utilisé le contrôle de qualité ainsi qu'une approche bayésienne afin de déterminer si la distribution des précipitations est stationnaire ou non dans le temps.

Après avoir calibré nos outils au moyen de l'ARL, nous avons comparé les trois méthodes sur la base des simulations, et ce, pour les deux scénarios. Rappelons qu'après le point de rupture, le premier scénario indique un changement fixe dans la moyenne alors que le deuxième scénario indique un changement linéaire croissant de la moyenne au cours du temps. Comme nous avons pu nous en apercevoir, l'approche bayésienne performe un peu moins bien pour le premier scénario alors que les trois méthodes performent de manière similaire pour le deuxième scénario.

Sur la base des résultats obtenus au dernier chapitre, nous avons conclu que la moyenne des précipitations du MRCC avait augmenté. De plus, selon les résultats obtenus, nous pouvons affirmer que le changement dans les précipitations s'est produit autour de l'année 1994.

Une raison qui peut expliquer le fait que l'approche bayésienne performe légèrement moins bien que l'approche classique pour le premier scénario est que pour l'approche bayésienne, la fenêtre mobile avance dans le temps. Ceci fait en sorte qu'à un moment donné, nous comparons deux densités *a posteriori* complètement corrompues. Les deux approches classiques n'ont pas le même problème puisqu'elles comparent toujours les statistiques sur la base des estimations obtenues pour les années initiales. En effet, $\widehat{\mu}_0$ ainsi que $\widehat{\sigma^*}$ sont toujours calculés en fonction des 10 premières années.

D'ailleurs, nous pourrions probablement améliorer l'efficacité de l'approche bayésienne en l'utilisant pour comparer les densités *a posteriori* de la forme $\pi(\theta|\underset{\sim}{x}_0)$ et $\pi(\theta|\underset{\sim}{x}_0, \underset{\sim}{y})$, où x_0 représente les premières années d'observations qui

sont considérées stationnaires. De cette manière, les deux approches seraient plus facilement comparables.

Dans de futurs travaux, il pourrait être intéressant de modéliser différemment les observations afin qu'on puisse prendre en compte les journées où il n'y a eu aucune précipitation. En effet, dans ce mémoire, nous avons tout simplement enlevé ces journées pour pouvoir utiliser la loi exponentielle. Dans cette optique, il serait possible d'utiliser un modèle à sur-représentation de zéros pour modéliser les précipitations. La densité des observations aurait donc la forme suivante :

$$f(x|\theta,\rho) = \rho f_0(x)) + (1-\rho) f_1(x|\theta),$$

où f_1 représente la densité exponentielle de paramètre θ, où f_0 est une mesure de Dirac sur le singleton $\{x = 0\}$ et où ρ représente la proportion de journées sans pluie. En utilisant cette modélisation, il serait possible de garder toutes les observations et ainsi, concevoir un modèle plus complet.

Bibliographie

Cam, L. M. L. et J. Neyman (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 5 : Weather Modification*. University of California Press.

Chaumont, D., G. Desrochers, J.-F. Angers, A. Frigon, G. Pacher, et R. Roy (2007). *Évolution des conditions climatiques au Québec : Développement d'un scénario climatique utilisée à des fins de prévision de la demande d'électricité au Québec sur l'horizon 2030*.

Derman, C. et S. Ross (1997). *Statistical aspects of quality control*. Academic Press.

Ghosh, J., M. Delampady, et T. Samanta (2007). *An Introduction to Bayesian Analysis : Theory and Methods*. Springer texts in statistics. Springer.

GIEC (2007). *Bilan 2007 des changements climatiques :Rapport de synthèse*. `http://www.ipcc.ch/ipccreports/ar4-wg1.htm`, consulté le 15 décembre 2013.

Gradshteyn, I. S. et I. M. Ryzhik (2007). *Table of integrals, series, and products.* (7e ed.). Academic Press.

Grigg, O., V. Farewell, et D. Spiegelhalter (2003). Use of risk-adjusted cusum and rsprtcharts for monitoring in medical contexts. *Statistical Methods in Medical Research 12*(2), 147–170.

Hanif, M., A. Hussain, N. Jamal, et M. Amir (2012). New approximation of arl in cusum control chart. *Far East Journal of Marketing and Management*.

Montgomery, D. (2007). *Introduction to statistical quality control* (6e ed.). Wiley.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika 41*(1), 100–115.

Rogerson, P. A. (2006). Formulas for the design of cusum quality control charts. *Communications in Statistics - Theory and Methods 35*(2), 373–383.

Ryu, J.-H., H. Wan, et S. Kim (2010). Optimal design of a cusum chart for a mean shift of unknown size. *Journal of Quality Technology 42*(3), 311–326.

Saniga, E., T. McWilliams, D. Davis, et J. Lucas (2006). *Economic Advantages of CUSUM Control Charts for Variables*.

Shewart, W. A. (1931). *Economic control of Quality of Manufactured Product*. New York : Van Nostrand Reinhold Co.

Siegmund, D. (1985). *Sequential Analysis*. Springer.