DOCTORAT BRETAGNE LOIRE / MATHSTIC

Université de Bretagne Occidentale

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE BRETAGNE OCCIDENTALE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques appliquées*

Par
Dominique ABGRALL

Studies of change-point detection: on-line scheme for discrete Poisson models and off-line test for parametric mixtures. Application to insurance problems.

Thèse présentée et soutenue à Brest, le 16 Juin 2021
Unité de recherche : Laboratoire de Mathématiques de Bretagne Atlantique (UMR 6205)

Rapporteurs avant soutenance :

Patrice BERTAIL	Professeur des universités, Laboratoire MODAL'X, Université Paris Nanterre
Olivier LOPEZ	Professeur des universités, Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université

Composition du Jury :

Président :	Olivier LOPEZ	Professeur des universités, LPSM, Sorbonne Université
Examinateurs :	Patrice BERTAIL	Professeur des universités, MODAL'X, Université Paris Nanterre
	Paul DOUKHAN	Professeur des universités, Lab AGM, Université Cergy Pontoise
	Pierre AILLIOT	Maître de conférences, LMBA, Université de Bretagne Occidentale
	Franck VERMET	Maître de conférences, HDR, LMBA, Université de Bretagne Occidentale
Directrice de thèse :	Catherine RAINER	Maître de conférences, HDR, LMBA, Université de Bretagne Occidentale
Co-directeur de thèse :	Brice FRANKE	Professeur des universités, LMBA, Université de Bretagne Occidentale
Co-encadrante de thèse :	Marine HABART	Docteur et Actuaire Agrégé, AXA GIE

Invité(s) :

Matthias LOEWE	Professeur des universités, Université de Münster
Ali FARDOUN	Maître de conférences, HDR, LMBA, Université de Bretagne Occidentale

Table of Contents

Acknowledgements

La thèse de doctorat est une expérience riche et dense, surtout vers la fin. Cette aventure a donné lieu à des rencontres fructueuses de multiples manières.

En premier lieu, je voudrais remercier mes directeurs et encadrante qui m'ont accompagné et soutenu pendant ces travaux. Je remercie tout d'abord ma directrice de thèse Catherine Rainer pour avoir accepté de m'encadrer, pour ses conseils toujours renouvelés, la rigueur dans ses exigences, et surtout sa patience. Je remercie également Brice Franke, mon co-directeur, pour son temps précieux, ses conseils avisés et la vision qu'il sait insuffler dans les sujets explorés. Enfin, je remercie vivement Marine Habart, ma co-encadrante, pour son attention aux enjeux des potentielles applications avec sa casquette d'actuaire, et pour sa disponibilité pour échanger tout au long de la thèse.
Je remercie tout particulièrement Patrice Bertail et Olivier Lopez, rapporteurs de cette thèse, pour avoir accepté de prendre le temps de lire mon manuscrit. Leurs retours sur tel ou tel aspect des travaux ouvrent des possibilités très riches pour de futurs développements, merci beaucoup. Je remercie également les autres membres du jury Pierre Ailliot, Paul Doukhan et Franck Vermet. Je suis honoré qu'ils aient accepté de participer à l'évaluation de cette thèse.
Au cours des années à préparer ce manuscrit, j'ai également bénéficié des conseils et encouragements d'Ali Fardoun et Matthias Löwe, membres de mon Comité de Suivi Individuel. Je les en remercie vivement.

Avant le début de ce doctorat, j'ai été encouragé par Aliou Sow et Nicolas Bouré, mes responsables au sein d'Allianz France, qui croyaient qu'il est possible d'écrire une thèse en parallèle d'une activité professionnelle déjà intense. Je les remercie d'avoir initié cette idée, et de m'avoir donné du temps dans les moments nécessaires. Je pense également à mes collègues d'Allianz dont les encouragements ont été un soutien précieux : Guillaume Metge, Julia Simaku, Aissata Sy, Mohamed Zaimi, Nicolas Zec, et tous les autres que je n'oublie pas.
Au sein de la communauté actuarielle, lors de congrès ou de déjeuners de travail, j'ai pu échanger sur mes travaux et aussi discuter de perspectives. Je remercie tous mes collègues actuaires et en particulier Guillaume Biessy, Alexandre Boumezoued, Pierre Hazaël-Massieux, Alexis Merx et Auguste MPacko Priso.

Je souhaite remercier l'équipe de l'EURIA pour le travail effectué ensemble depuis plusieurs années : Anthony, Brice, Fabrice, Franck, Marc, Marine, Patricia et Yveline.
Pour leur soutien particulier dans mes études en Mathématiques, je remercie Guy Le Boursicaud, Mohamed Belghiti, Jean-Marc Derrien, Elena Issoglio, et toute l'équipe du LMBA et du département de Mathématiques de l'UBO qui m'a accompagné depuis 2006.

Enfin, je souhaite remercier tous ceux qui, de près ou de loin, m'ont encouragé. Je pense en particulier à la famille Bellec, Alexis et Guillaume, Hugues et Myriam, et surtout à mes parents et ma famille.

Synthèse

Cette thèse s'intéresse à deux études distinctes de techniques de détection de rupture. Ces techniques permettent de déterminer si, pour un phénomène aléatoire dont la loi sous-jacente initiale est connue, un changement se produit ou non lors de son observation.

Le manuscrit comporte une introduction et trois chapitres rédigés en anglais. Les motivations, ainsi que les résultats des deux problèmes traités, sont présentés et discutés dans l'introduction. Les préliminaires du Chapitre 1 introduisent de manière détaillée les concepts et théorèmes sur lesquels les travaux s'appuient. Le Chapitre 2 traite de l'estimation séquentielle du paramètre de la distribution après le changement dans le cadre d'une séquence de variables aléatoires suivant une loi de Poisson. Des applications détaillées sont données, en particulier pour l'étude de la mortalité française. Le Chapitre 3 introduit un test d'hypothèse alternatif qui vise à détecter un changement dans la première composante d'un mélange fini de lois paramétriques. Le résultat principal est un théorème limite fonctionnel qui donne la loi limite de la statistique du test alternatif sous forme de transformation d'un mouvement brownien multidimensionnel. Ce nouveau test est comparé numériquement à un test standard basé sur un ratio de vraisemblance (cf. e.g. Csörgő and Horváth (1997)). Une application à des données réelles en assurance IARD est donnée à but d'illustration du test alternatif.

Les techniques de détection de rupture sont aujourd'hui utilisées pour de multiples applications. Dans le cas discret, leur but est d'étudier la séquence de variables aléatoires indépendantes $(X_n)_{n\geq 1}$ associée à une distribution initiale $\mathbb{P}_{\boldsymbol{\theta}}$, où $\boldsymbol{\theta}$ est dans un ensemble de paramètres éligibles Θ. Les procédures de détection cherchent à dire si et quand cette distribution change. Chaque variable aléatoire X_n suit une distribution $\mathbb{P}_{\theta^n}$, où $\theta^n \in \Theta$. Le temps de rupture est dénoté par ν : l'échantillon est de même loi avant ν, i.e. $\boldsymbol{\theta} = \theta^1 = \cdots = \theta^\nu$; et change de loi ensuite, i.e. $\theta^n \neq \boldsymbol{\theta}$ pour $n > \nu$. Il existe différentes familles de techniques : nous nous intéressons seulement aux procédures de détection en ligne dites optimales au sens du délai de détection (Chapitre 2) et aux tests d'hypothèse qui testent l'existence d'un changement dans un échantillon fermé (Chapitre 3).

Premier problème : Estimation séquentielle du paramètre de la distribution après le changement pour une séquence de variables aléatoires suivant une loi de Poisson

Parmi les procédures de détection en ligne, les procédures du CUSUM et de Shiryaev-Roberts sont aujourd'hui les plus utilisées en raison de nombreux résultats qui montrent leur optimalité au sens du délai de détection (Page (1954), Shiryaev (1961), Roberts (1966)). Dans le cas discret et pour un temps de rupture déterministe, Moustakides (1986) a montré que la procédure du CUSUM est optimale pour minimiser le '*pire pire délai de détection*' de Lorden (1971). Pour le cas Bayésien, c'est-à-dire quand le temps de rupture est aléatoire, Pollak and Tartakovsky (2009) ont montré que la pro-

cédure de Shiryaev-Roberts est optimale pour minimiser le délai de détection moyen. D'autres résultats existent pour le cadre continu (cf. Section 1.3 et les références qui y figurent). Ces deux procédures du CUSUM et de Shiryaev-Roberts font l'hypothèse que les paramètres définissant les distributions avant et après le changement sont connus. En pratique, s'il semble raisonnable de connaître la distribution avant le changement, il est moins vraisemblable de connaître à l'avance le paramètre après changement. Quelques références dans la littérature s'intéressent à l'estimation du paramètre après changement : Wu (2005) donne une forme explicite d'un estimateur du maximum de vraisemblance de la moyenne après le changement dans le cas d'une procédure du CUSUM. D'autres approches par maximum de vraisemblance étudient les propriétés asymptotiques d'estimateurs de la loi post-changement (cf. Section 2.1 et les références qui y figurent).

Dans le Chapitre 2, nous introduisons un estimateur alternatif pour le paramètre de la distribution après changement en réutilisant les propriétés de la statistique de détection de la procédure de Shiryaev-Roberts. Nous considérons une séquence indépendante de variables aléatoires $(X_n)_{n\geq 1}$ qui suit une loi de Poisson. La séquence est identiquement distribuée de paramètre λ avant le temps de changement ν, et identiquement distribuée de paramètre $\lambda\boldsymbol{\rho}$ après le temps de changement. Le paramètre $\lambda > 0$ est connu et déterministe, alors que les paramètres $\boldsymbol{\rho} > 0$ et $\nu \in \mathbb{N} \cup \{\infty\}$ sont déterministes mais inconnus. Ici, $\boldsymbol{\rho}$ représente le ratio des intensités des lois de Poisson après et avant le changement. On dénote par f_ρ la fonction de densité d'une variable aléatoire qui suit une loi de Poisson d'intensité $\lambda\rho$, définie pour tout $\rho > 0$: $f_\rho(x) = e^{-\lambda\rho}(\lambda\rho)^x/x!$, $x \in \mathbb{N}$.

La procédure de Shiryaev-Roberts permet de détecter un changement sous l'hypothèse que les paramètres avant et après changement sont connus. Elle est définie par la statistique séquentielle $S_n(\rho)$ telle que $S_0 := 0$ et

$$S_n(\rho) := \sum_{i=1}^{n} \prod_{k=i}^{n} \frac{f_\rho(X_k)}{f_1(X_k)} = (1 + S_{n-1}(\rho)) \, \frac{f_\rho(X_n)}{f_1(X_n)}, \ n \geq 1.$$

Elle est définie de sorte à rester proche de 0 avant le changement et à augmenter fortement après le changement. Nous proposons d'utiliser cette statistique pour estimer le paramètre post-changement. Pour $n \geq 0$ fixé, nous définissons l'estimateur $\hat{\rho}_S(n)$ du paramètre de changement d'intensité $\boldsymbol{\rho}$, par

$$\hat{\rho}_S(n) := \arg\max_{\rho>0} S_n(\rho). \tag{1}$$

A notre connaissance, l'idée de réutiliser la statistique de détection pour l'inférence du paramètre post-changement ne semble pas présente dans la littérature. Nous montrons le résultat suivant.

Théorème 1. *Pour un temps de rupture $0 \leq \nu < \infty$ fixé mais inconnu, l'estimateur $\hat{\rho}_S(n)$ défini par* (1) *est convergent.*

La comparaison avec d'autres estimateurs classiques du maximum de vraisemblance ou de minimisation d'erreur quadratique montre que :

- ◇ Comparé à un panel d'estimateurs usuels, l'estimateur $\hat{\rho}_S(n)$ converge significativement plus vite vers le vrai paramètre après changement,

- ◇ Sa variance est bien plus faible que les estimateurs usuels après le changement, avec une variance asymptotique proche.

Nous pouvons utiliser l'estimateur $\hat{\rho}_S(n)$ pour appliquer la procédure de Shiryaev-Roberts qui a maintenant $S_n(\hat{\rho}_S(n))$ pour statistique séquentielle. Cela définit une procédure de détection d'un **changement de niveau** pour une séquence de variables aléatoires qui suivent une loi de Poisson d'intensité constante, où il n'est pas nécessaire de connaître le paramètre post-changement.

L'estimateur $\hat{\rho}_S(n)$ est défini pour maximiser la réponse de la statistique de détection. Cela peut être un point de départ pour traiter la question ouverte de l'optimalité de la procédure pour le cas où il n'est pas nécessaire de connaître le paramètre post-changement.

Un des avantages des procédures comme le CUSUM ou celle de Shiryaev-Roberts est d'être re-calculables sans effort pour chaque nouvelle observation qui arrive, c'est-à-dire sans re-parcourir tout l'échantillon. Par définition de $\hat{\rho}_S(n)$, notre procédure nécessite ce re-calcul. Avec le fait que $\hat{\rho}_S(n)$ converge rapidement vers le vrai paramètre, elle est bien adaptée à des applications pour lesquelles la taille n de l'échantillon ne croît pas trop vite : par exemple, pour l'étude de la mortalité annuelle, ou d'autres applications actuarielles où les nouvelles observations arrivent à un rythme hebdomadaire, mensuel, annuel.

Nous explorons également un autre type de séquence aléatoire où l'intensité de la loi de Poisson croît/décroît à une vitesse constante. Dans ce deuxième modèle, la séquence (X_n) suit une loi de Poisson de paramètre λ_n où $\lambda_n := \alpha\lambda_{n-1}$ avant le changement et $\lambda_n := \alpha'\lambda_{n-1}$ après le changement, avec $\alpha \neq \alpha'$. Les paramètres λ_0 et α sont déterministes et connus. Les paramètres ν et α' sont déterministes mais inconnus. Ici, nous ne proposons pas d'estimateur pour le paramètre après changement mais considérons plutôt l'estimateur du maximum de vraisemblance qui, numériquement, semble être un candidat raisonnable. La procédure de Shiryaev-Roberts, dans ce cadre, permet de détecter un **changement de tendance** pour une séquence de variables aléatoires qui suivent une loi de Poisson d'intensité linéairement croissante ou décroissante.

La loi de Poisson est un cadre couramment utilisé pour la modélisation de la mortalité. A l'aide de techniques de normalisation des données, nous appliquons les deux procédures de détection de niveau et de tendance à la mortalité nationale française sur plusieurs évènements marquants de l'histoire : la grippe espagnole de 1918 (Caselli et al. (1987), Spreeuwenberg et al. (2018)) ; la baisse de la mortalité dans les années 1960 liée à la révolution cardiovasculaire (Vallin and Meslé (2010)) ; et la canicule de 2003 (Robine et al. (2008)). Nous étudions également la mortalité d'un portefeuille de 15 000 rentiers entre 2003 et 2014. Les résultats principaux nous permettent d'identifier des changements dans la mortalité, de quantifier l'amplitude de ces changements ainsi qu'un délai de détection. La procédure pour détecter un changement de niveau détecte correctement des changements persistants, même minimes. La procédure qui détecte un changement de tendance apparaît sensible à une séquence (même courte) d'observations croissantes/décroissantes.

Second problème : Test de ratio de vraisemblance pondérée détectant un changement dans une composante d'un mélange de lois paramétriques

Les modèles basés sur des mélanges de lois sont devenus assez populaires dans la littérature de statistiques ces dernières décennies car ils permettent de décrire un phénomène contenant plusieurs sous-populations statistiques. L'estimation des paramètres d'un mélange de loi est un problème qui intéresse à la fois les statisticiens théoriques à travers, par exemple, des questions d'identifiabilité mais aussi les praticiens via des questions algorithmiques (cf. Section 1.1 et les références qui y figurent).

Nous considérons un échantillon de n variables aléatoires indépendantes $(X_i)_{1\leq i\leq n}$ à valeurs dans un espace vectoriel réel $\mathcal{X}$. Chaque variable X_i suit un mélange fini de lois paramétriques avec $2 < m < \infty$ composantes si, sous certaines conditions de régularité, pour $f_1, \ldots, f_m$ des fonctions de densité sur $\mathcal{X}$ fixées, la distribution $\mathbb{P}_\theta$ de X_i admet la densité

$$f(x, \theta) := \sum_{k=1}^{m} p_k f_k(x, \lambda_k),\ x \in \mathcal{X}$$

où le paramètre $\theta = (p_1, \ldots, p_{m-1}, \lambda_1, \ldots, \lambda_m)$ est dans l'ensemble des paramètres possibles $\Theta = \Theta_0 \times \prod_{k=1}^{m} \Theta_k$, sous-ensemble d'un espace Euclidien de dimension d, et $p_m := 1 - \sum_{k=1}^{m-1} p_k$. Le nombre de composantes m est connu et déterministe. Chaque paramètre λ_k est dans un ensemble Θ_k et le vecteur des poids $(p_1, \ldots, p_{m-1})$ est dans l'ensemble ouvert $\Theta_0 := \{(p_1, \ldots, p_{m-1}) \in (0,1)^{m-1}, \sum_{k=1}^{m-1} p_k < 1\}$. Nous faisons l'hypothèse que le mélange est identifiable : il est possible de distinguer les composantes du mélange les unes des autres et de les ordonner (McLachlan and Peel (2000)).

Le but du chapitre est de détecter un changement dans la première composante du mélange lorsqu'il y a au plus un changement dans l'échantillon (modèle AMOC : *at most one change*). La détection de rupture pour des mélanges de lois est courante car elle fait partie des techniques générales de détection de rupture (Csörgő and Horváth (1997), van der Vaart (1998)). Or, peu de travaux lient spécifiquement les deux sujets (Andrews and Ploberger (1994), Hansen (1996), Pons (2009), Zou et al. (2015)). Le test d'hypothèse standard basé sur un ratio de vraisemblance, exposé dans Csörgő and Horváth (1997), peut être adapté pour construire un test d'hypothèse dit **benchmark**, dédié à la détection de rupture dans la première composante du mélange. Toutefois, d'un point de vue pratique, deux problèmes émergent lors des applications. Tout d'abord, la statistique de test contient un problème d'optimisation sans solution explicite et pour lequel il n'existe pas d'algorithme dédié. Ensuite, la résolution directe du problème d'optimisation avec un algorithme standard est gloutonne en temps de calcul jusqu'à être déraisonnable pour des échantillons de grande taille, tout en convergeant vers une solution peu acceptable. C'est pourquoi nous cherchons à construire un test qui s'écrit comme une fonction simple d'estimateurs standards pour lesquels un algorithme d'estimation dédié peu se trouver aisément (par exemple l'algorithme EM de Dempster et al. (1977) ou ses dérivés).

Dans le Chapitre 3, nous introduisons un test d'hypothèse qui répond à cet objectif. Supposons que l'échantillon $(X_i)_{1\leq i\leq n}$ suive un mélange fini de lois paramétriques de sorte que, si un changement se produit, il est unique et inconnu. Contrairement au premier travail, nous nous intéressons ici à de grands échantillons, et sommes donc

amenés à étudier le comportement limite. Nous supposerons donc que l'expérience se passe dans un intervalle de temps $[0,1]$ et chaque variable X_i est observée au temps i/n, $1 \leq i \leq n$. Nous imposons que, si un changement survient, le temps de rupture est contenu dans l'intervalle $[\bar{s}, 1-\bar{s}] \subset (0,1)$, où $0 < \bar{s} < 1/2$ est connu et déterministe : le changement ne peut se produire trop près de 0 ni de 1. Nous définissons le test d'hypothèse par :

1. L'hypothèse nulle H_0 où il n'y a pas de rupture : le mélange est défini par le vrai paramètre $\boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{m-1}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m) \in \Theta$.
2. L'hypothèse alternative H_1 : un changement survient au temps s, $s \in [\bar{s}, 1-\bar{s}]$.

Dans la suite, nous nous placerons toujours sous l'hypothèse nulle. Nous imposons également que l'estimateur du maximum de vraisemblance du paramètre θ qui définit le mélange est fortement convergent.

Considérons la fonction de poids au point $x \in \mathcal{X}$ pour $\theta \in \Theta$, définie par

$$w(x,\theta) := \frac{p_1 f_1(x,\lambda_1)}{f(x,\theta)},$$

et introduisons *le log-ratio de vraisemblance pondéré* $\Lambda_{s,n}$ défini pour $s \in [\bar{s}, 1-\bar{s}]$ et $n \geq 1$ par

$$\Lambda_{s,n} := \log\left(\frac{\prod_{i=1}^{\lfloor sn \rfloor} f_1(X_i, \hat{\lambda}_{0,s,1})^{w(X_i,\hat{\theta}_{0,s})} \prod_{j=\lfloor sn \rfloor+1}^{n} f_1(X_j, \hat{\lambda}_{s,1,1})^{w(X_j,\hat{\theta}_{s,1})}}{\prod_{i=1}^{n} f_1(X_i, \hat{\lambda}_1)^{w(X_i,\hat{\theta})}}\right),$$

où $\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_{m-1}, \hat{\lambda}_1, \ldots, \hat{\lambda}_m)$ est l'estimateur du maximum de vraisemblance pour $\boldsymbol{\theta}$ sur l'échantillon tout entier, et, pour $s \in [\bar{s}, 1-\bar{s}]$, $\hat{\theta}_{0,s}$ et $\hat{\theta}_{s,1}$ sont respectivement les estimateurs du maximum de vraisemblance pour $\boldsymbol{\theta}$ sur les sous-échantillons $(X_i)_{1 \leq i \leq \lfloor sn \rfloor}$ et $(X_i)_{\lfloor sn \rfloor+1 \leq i \leq n}$. Les fonctions de poids $w(.,.)$ dans $\Lambda_{s,n}$ permettent de zoomer sur les observations qui ont le plus de chance d'appartenir à la première composante. La statistique de test est définie par

$$S_n := \sup_{s \in [\bar{s}, 1-\bar{s}]} \Lambda_{s,n}.$$

Nous dénotons ce test de ratio de vraisemblance pondérée par **WLT** (*Weighted Likelihood Test*). Le test ainsi défini est une fonction simple d'estimateurs standards du maximum de vraisemblance pour lesquels des algorithmes dédiés existent (Dempster et al. (1977), Benaglia et al. (2009)). Son application à des données ne présente donc pas de difficulté de mise en œuvre.

La distribution limite de la statistique de test S_n est obtenue par une méthode similaire à celle de Davis et al. (1995). Nous commençons par établir certaines propriétés asymptotiques des estimateurs $\hat{\theta}_{0,s}$ et $\hat{\theta}_{s,1}$ sous l'hypothèse nulle à l'aide d'arguments classiques de l'étude des estimateurs du maximum de vraisemblance (Lehmann and Casella (1998)). Ces résultats liminaires nous permettent d'établir un théorème limite fonctionnel pour le processus càd-làg $(\Lambda_{s,n})_{s \in [\bar{s}, 1-\bar{s}]}$ à l'aide d'applications multiples du *Continuous Mapping Theorem* et d'une delta-méthode fonctionnelle adaptée pour l'espace métrique de Skorokhod [1] (Billingsley (1999), van der Vaart (1998)). La distribution limite de la statistique S_n est obtenue ci-dessous comme une conséquence du résultat fonctionnel.

1. Il s'agit ici de l'espace des fonctions càd-làg sur $[\bar{s}, 1-\bar{s}]$ à valeurs dans un espace Euclidien.

Théorème 2. *Sous l'hypothèse nulle et certaines conditions de régularité détaillées dans la Section 3.2.1, si l'estimateur $\hat{\theta}$ est fortement convergent, alors*

$$S_n \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}(W_s - sW_1)}{s(1-s)}$$

où $(W_s)_{s\in[0,1]}$ est un mouvement Brownien standard de dimension $2d+d^2$ et l'application $\boldsymbol{q}$ est une forme quadratique définie dans la Section 3.3.2 par l'équation (3.33).

Ainsi le processus limite est une forme quadratique du pont Brownien. Ce type de limite est similaire à celles obtenues par exemple dans Davis et al. (1995), Csörgő and Horváth (1997) ou Dehling et al. (2014).

La loi de la variable aléatoire limite $\sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}(W_s-sW_1)}{s(1-s)}$ reste une question ouverte. Toutefois, la complexité de la fonction $\boldsymbol{q}$ ne présage pas de solution simple à ce problème et des simulations numériques sont suffisantes pour calibrer un seuil de détection pour le test WL.

Suite à de premières applications numériques simulées, nous suggérons une extension du test pour améliorer ses performances, notamment en réduisant l'erreur de type II. Pour cela, nous introduisons une version ajustée $\Lambda_n^* := (\Lambda_{s,n}^*)_{s\in[\bar{s},1-\bar{s}]}$ du processus sous-jacent où

$$\begin{aligned}\Lambda_{s,n}^* := \frac{c_{1,n}}{c_{s,n}} \left(\sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) \log f_1(X_i, \hat{\lambda}_{0,s,1}) + \sum_{i=\lfloor sn \rfloor+1}^{n} w(X_i, \hat{\theta}_{s,1}) \log f_1(X_i, \hat{\lambda}_{s,1,1}) \right) \\ - \sum_{i=1}^{n} w(X_i, \hat{\theta}) \log f_1(X_i, \hat{\lambda}_1)\end{aligned}$$

et où, pour s fixé dans $[\bar{s}, 1]$, $c_{s,n} := \sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) + \sum_{i=\lfloor sn \rfloor+1}^{n} w(X_i, \hat{\theta}_{s,1})$ représente la contribution de l'échantillon aux deux premières sommes de $\Lambda_{s,n}$, et $c_{1,n}$ la contribution à la dernière. Nous montrons que la contribution $c_{s,n}$ converge p.s. vers le paramètre de poids p_1 de la première composante, uniformément en $s \in [\bar{s}, 1]$. La statistique de ce test, dénoté par **EWLT**, devient $S_n^* := \sup_{s\in[\bar{s},1-\bar{s}]} \Lambda_{s,n}^*$. Cet ajustement n'est pas sans impact car cela ajoute un aléa supplémentaire. La distribution limite de S_n^* reste un supremum d'une forme quadratique d'un mouvement brownien mais maintenant de dimension $3d + d^2$.

Théorème 3. *Sous l'hypothèse nulle et certaines conditions de régularité détaillées dans la Section 3.2.1, si l'estimateur $\hat{\theta}$ est fortement convergent, alors*

$$S_n^* \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}^*(W_s - sW_1)}{s(1-s)}$$

où $(W_s)_{s\in[0,1]}$ est un mouvement Brownien standard de dimension $3d+d^2$ et l'application $\boldsymbol{q}^*$ est une forme quadratique définie dans la Section 3.4 par l'équation (3.39).

Si l'échantillon $(X_i)_{1\le i\le n}$ suit un **mélange Gaussien unidimensionnel**, alors le mélange est défini par un paramètre θ de la forme suivante :

$$\theta = (p_1, \ldots, p_{m-1}, \mu_1, \sigma_1, \ldots, \mu_m, \sigma_m)$$

tel que, pour chaque k-ème composante, $\mu_k \in \mathbb{R}$ est le paramètre de moyenne et $\sigma_k^2 \in \mathbb{R}_+^*$ le paramètre de variance. Nous imposons que les moyennes soient strictement croissantes et, comme dans Hathaway (1985), que les variances soient bornées : $\min\{\sigma_j/\sigma_k,\ 1 \leq j,k \leq m\} > \boldsymbol{b}$, où $0 < \boldsymbol{b} \leq 1$ est une *borne de dispersion* supposée connue. Par Hathaway (1985), l'estimateur $\hat{\theta}$ est alors fortement convergent et nous pouvons montrer que les conditions de validité des Théorèmes 2 et 3 ci-dessus sont vérifiées.

Nous présentons deux applications pour le cas Gaussien où nous nous intéressons à des échantillons de grande taille (plus de 10 000 observations).
Une première application, basée sur des simulations d'un mélange à 3 composantes, permet de montrer que les tests WL et EWL ont une erreur de type II bien plus faible que celle du test benchmark. Les mauvaises performances de ce dernier s'expliquent essentiellement par le fait qu'il n'existe pas d'algorithme dédié pour résoudre le problème d'optimisation qu'il contient, alors que les deux tests WL et EWL reposent sur des algorithmes standards. De plus, le temps nécessaire au calcul du test benchmark augmente considérablement avec la taille de l'échantillon.
Lorsque nous étudions le comportements des tests pour un changement dans une autre composante que la première, nous constatons que le test EWL reste le meilleur candidat car la fréquence de détection dans ce cas est la plus faible. Toutefois, dans certains cas, les trois tests détectent à tort un changement : corriger cet effet pourrait donner lieu à une extension de nos travaux. De plus, le test benchmark semble avoir de meilleures propriétés lorsque le problème d'optimisation qu'il contient est correctement résolu. Ainsi, il serait intéressant d'améliorer ce test en élaborant un algorithme dédié à son problème d'optimisation via, par exemple, une adaptation de l'algorithme EM.
La seconde application est consacrée à un cas concret pour la détection d'un changement dans la distribution de la variation de la charge de sinistres corporels en assurance non-vie. Cette variation de charge peut se modéliser par un mélange Gaussien à 12 composantes et l'entreprise d'où émerge le problème s'intéresse à vérifier qu'un changement se produit dans la 5ème composante. L'application des tests WL et EWL confirme la détection du changement, ce qui a permis à l'entreprise de poursuivre la recherche des causes sous-jacentes.

Les applications montrent que les deux tests WL et EWL sont performants (faible erreur de type II). De plus, les résultats obtenus sous l'hypothèse nulle dans les Théorèmes 2 et 3 permettent de calibrer aisément des seuils de détection via une simulation de mouvements browniens (temps de calcul marginal divisé par 10 000). Au-delà de ces deux constats prometteurs, les travaux réalisés montrent des possibilités d'extensions, notamment via une amélioration des algorithmes permettant l'application du test benchmark.

Introduction

This thesis concerns two distinct studies of change-point detection techniques. Such techniques consider the question of deciding whether a change happens or not, when observing over time a random phenomenon with known initial properties.

Likewise this introduction, the manuscript is divided into three chapters. The key notions and theorems that we rely on in this thesis are introduced and put into perspective in Chapter 1. Chapter 2 corresponds to the article A. et al. (2018). It deals with the sequential estimation of the post-change distribution when observing an online sequence of Poisson random variables. Detailed applications to the study of French mortality are provided. Chapter 3 introduces an alternative hypothesis test that aims to detect a change-point in the first component of a finite parametric mixture, for a closed sample where at most one change occurs. The main result consists in a functional limit theorem for the distribution of the test statistic. A short application to Property and Casualty insurance data is provided.

The purposes of this introduction are to explain the context of this work, to present the main results of the two distinct topics detailed in Chapters 2 and 3, and to discuss unaddressed problems and remaining open questions.

1 Context

The concept of change detection originated in the field of quality control (Girshick and Rubin (1952), Page (1954)) in the 1950's and is used today in many applied fields such as epidemiology, insurance, server protection or finance [2]. In the discrete case, the common goal of these applications is to study the random sequence $(X_n)_{n\geq 1}$ associated with a known initial distribution $\mathbb{P}_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta}$ in some set of eligible parameters Θ. Detection procedures aim to determine *if and when* this initial distribution changes. Each random variable X_n follows a distribution $\mathbb{P}_{\theta^n}$, where $\theta^n \in \Theta$. The time of the change ν is called *change-point*: the sample is identically distributed before ν, i.e. $\boldsymbol{\theta} = \theta^1 = \cdots = \theta^\nu$; and, after the change-point, $\theta^n \neq \boldsymbol{\theta}$ for $n > \nu$ (possibly but not necessarily i.i.d.). One can divide change-point detection techniques into three families:

- ◇ Off-line hypothesis tests that state whether or not there are one or more change-points in a closed observed sample $(x_i)_{1\leq i\leq n}$, for n fixed. The work of Chapter 3 falls into this category where: the underlying distribution $\mathbb{P}_\theta$, $\theta \in \Theta$, is assumed to be a finite parametric mixture and the test aims to detect the presence of at most one change.
- ◇ Inference methods for the parameters of a model that assumes that a closed sample contains one or more change-points.
- ◇ On-line detection schemes that aim to detect a change-point when observing the sequence of outcomes $(x_n)_{n\geq 1}$ as they arrive. In this category, Chapter 2

2. For more details, see e.g. Basseville and Nikiforov (1993), Brodsky and Darkhovsky (1993), Csörgő and Horváth (1997), Chatterjee (2012), Pons (2018), Frühwirth-Schnatter et al. (2019), Truong et al. (2020) and the references therein.

introduces a sequential estimator for the post-change distribution of a sequence of Poisson random variables.

Since they cover two distinct topics, we introduce separately the motivations and the results of Chapters 2 and 3 in the following two sections.

2 Discrete Poisson case: a sequential estimator of the post-change parameter

This section introduces Chapter 2.

2.1 Motivation

Among the on-line detection schemes for discrete sequences, a *quickest change-point detection scheme* is a sequential procedure that aims to detect a change in the distribution of the phenomenon as quickly as possible without raising too many *false alarms*, i.e. when the procedure detects a change while none occurred (Poor and Hadjiliadis (2009), Tartakovsky et al. (2015)).

We consider an experiment where we observe a sequence $(X_n)_{n\geq 1}$ of independent random variables, with values in a real vector space. The time of change or *change-point* is denoted by ν and is unknown: random (Bayesian framework) or deterministic. We assume that the sequence $(X_n)_{1\leq n\leq \nu}$ is identically distributed with parametric distribution $\mathbb{P}_{\boldsymbol{\theta}}$ and density function $f_{\boldsymbol{\theta}}$, and the sequence $(X_n)_{n\geq \nu+1}$ is identically distributed with distribution $\mathbb{P}_{\boldsymbol{\theta}'}$ and density function $f_{\boldsymbol{\theta}'}$. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are in some set Θ of possible parameters and, in general, the distribution of the sequence is denoted by $\mathbb{P}_\nu$. A *change-point detection scheme* is a procedure defined by a detection sequence S_n and a threshold s^*. An alarm is triggered as soon as $S_n > s^*$, defining a stopping time $n^* := \inf\{n \geq 1, S_n > s^*\}$. The threshold s^* is usually chosen with respect to a false alarm constraint. For example, one can consider a constraint for the *average delay of detection* when no change occurs: $\mathbb{E}_{\{\nu=\infty\}}[n^*] \geq \eta$, where $\eta > 1$ is defined by the user.

Today, there are two well-known quickest change-point detection schemes. Page (1954) introduced the *CUSUM detection scheme*. Moustakides (1986) proved that, for a deterministic change-point, this scheme is optimal in minimizing the '*worst worst*' detection delay $\sup_{\nu\geq 0} \operatorname{ess\,sup} \mathbb{E}_\nu\left[(n^*-\nu)^+ | X_1, \ldots, X_\nu\right]$. This delay criterion was introduced by Lorden (1971) [3] who first showed its asymptotic optimality, i.e. when the constraint η for the average delay of detection tends to infinity. The optimality of this scheme has been proved for a continuous-time setup for different types of processes: for example Beibel (1996) and Shiryaev (1996) in a Brownian motion context, Moustakides (2004) for statistics of Itô processes or El Karoui et al. (2017) for an inhomogeneous Poisson process.
After early attempts by Girshick and Rubin (1952), another common procedure has been proposed independently by Shiryaev (1963) (also in Shiryaev (1961)) for the continuous-time case and Roberts (1966) for the discrete case. The *Shiryaev-Roberts detection scheme* is a change-point detection scheme where the detection sequence is

3. Known as the *Lorden's criterion*. Other criteria can be found in the literature: e.g. Pollak (1985), Tartakovsky et al. (2015) and the references therein.

given by $S_0 := 0$ and

$$S_n(\boldsymbol{\theta}, \boldsymbol{\theta}') := (1 + S_{n-1}(\boldsymbol{\theta}, \boldsymbol{\theta}')) \frac{f_{\boldsymbol{\theta}'}(X_n)}{f_{\boldsymbol{\theta}}(X_n)}, \ n \geq 1. \tag{2}$$

Note that the detection statistic $S_n(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is defined such that it stays close to 0 before the change and increases strongly after the change. Pollak and Tartakovsky (2009) showed that, for a fixed $\eta > 1$, if the change-point ν is a generalized random variable with a uniform improper prior distribution, then, among all change-point detection procedures satisfying $\mathbb{E}_{\{\nu=\infty\}}[n^*] \geq \eta$, the Shiryaev-Roberts detection scheme is optimal in minimizing the average detection delay given by $\sum_{k=0}^{\infty} \mathbb{E}_{\{\nu=k\}} \left[(n^* - k)^+\right] / \mathbb{E}_{\{\nu=\infty\}}[n^*]$. Optimality is still valid for the case where ν has a geometric distribution (Shiryayev (1978)). The same result has been obtained for a continuous-time setup for different types of processes: for example Shiryaev (1963) and Feinberg and Shiryaev (2006) in the context of Brownian motions.

Since their introduction in the 1950's, the CUSUM and the Shiryaev-Roberts schemes are used in many applied fields (see Chatterjee (2012) and the references therein). A few other procedures are also used for change-point detection. For example Roberts (1959) suggests a procedure based on EWMA (exponentially weighted moving averages).

The Shiryaev-Roberts procedure assumes that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are known. In practice, it is reasonable to assume that the distribution before the change, defined by the parameter $\boldsymbol{\theta}$, is known because in most situations the practitioner can use the first observations to infer $\boldsymbol{\theta}$. However, in most cases, the post-change distribution is unknown without any practical way to infer $\boldsymbol{\theta}'$.

For exponential families, Foster and George (1993) suggest some estimators for the mean before the change. Wu (2005) provides an explicit form of the Maximum Likelihood Estimator (MLE) of the post-change mean in the case of the CUSUM scheme, and its asymptotic properties. More recently, Fotopoulos et al. (2010) studied the asymptotic distributions of the MLE under the Gaussian framework. More MLE designed for specific frameworks are provided e.g. in Wu (2016a), Wu (2016b) and Frick et al. (2014).

2.2 Our contribution

Our contribution consists in the introduction of an alternative estimator for the post-change parameter in the case of an i.i.d. discrete sample with Poisson distribution. This estimator is based on the detection statistic of the Shiryaev-Roberts procedure. We show that it is a consistent estimator for the post-change parameter. Numerical simulations indicate that, compared to the usual Maximum Likelihood Estimator, it has a significantly reduced bias and variance just after the change with identical asymptotic properties.

2.2.1 Detection of a change of level

With the previous notations, consider the sequence of independent random variables $(X_n)_{n\geq 1}$ such that X_n follows a Poisson distribution with parameter λ for $1 \leq n \leq \nu$, and with parameter $\lambda\boldsymbol{\rho}$ for $n \geq \nu + 1$. The parameter $\lambda > 0$ is deterministic and supposed to be known, while $\boldsymbol{\rho} > 0$ and $\nu \in \mathbb{N} \cup \{\infty\}$ are deterministic but unknown. The time of the change is denoted by ν and $\boldsymbol{\rho}$ is the ratio of the intensities after and before the change. Here we use the term density function with respect to the counting measure on $\mathbb{N}$. We denote by f_ρ the density function of a Poisson random variable with intensity $\lambda\rho$ defined for any $\rho > 0$ by $f_\rho(x) = e^{-\lambda\rho}(\lambda\rho)^x/x!$, $x \in \mathbb{N}$.

The Shiryaev-Roberts sequence aims to detect a change, knowing the parameters before and after the change. We suggest to use its optimality property in order to infer the intensity shift parameter $\boldsymbol{\rho}$. For given $\lambda > 0$, $n \geq 0$ known, and $\nu \geq 0$ deterministic but unknown, we define the estimator $\hat{\rho}_S(n)$ by

$$\hat{\rho}_S(n) := \arg\max_{\rho>0} S_n(\rho), \tag{3}$$

where $S_n(\rho) = S_n(\lambda, \lambda\rho)$ is the Shiryaev-Roberts sequence, as defined in (2). To our knowledge, using the detection statistic for the inference of $\boldsymbol{\rho}$ does not appear in the literature. One first property of the estimator $\hat{\rho}_S(n)$ is that it does not depend on the change-point ν. Therefore, we do not need to know when the change occurs in order to infer the post change intensity. Using again the statistic in (2), we define a new detection procedure by the detection sequence $S_n(\lambda, \lambda\hat{\rho}_S(n))$. In this procedure, only the parameter before the change needs to be known.

We show that $\hat{\rho}_S(n)$ is a consistent estimator for $\boldsymbol{\rho}$. For that purpose we consider the following sub-problem based on the likelihood ratio

$$R_{i,n}(\rho) := \prod_{k=i}^{n} \frac{f_\rho(X_k)}{f_1(X_k)},$$

where we remark that $S_n(\rho)$ is the sum of the variables $R_{i,n}(\rho)$ for $1 \leq i \leq n$. If we set

$$\hat{\rho}_{i,n} := \lambda^{-1} \sum_{k=i}^{n} \frac{X_k}{n-i+1},$$

we can show that $\hat{\rho}_{i,n} = \arg\max_{\rho>0} R_{i,n}(\rho)$. In addition, the sequence $(X_k)_{k\geq 1}$ differs from an independent identically distributed sequence only through a finite number of terms. Thus the law of large numbers holds: for all $i \geq 1$, $\hat{\rho}_{i,n}$ converges $\mathbb{P}$-almost surely to $\boldsymbol{\rho}$. Unfortunately, we can also show that the convergence cannot be uniform in $i \geq 1$. In a first Lemma below, we show that $\hat{\rho}_S(n)$ is somehow related to the $\hat{\rho}_{i,n}$, $1 \leq i \leq n$.

Lemma 1. *For any $n \geq 1$, $\hat{\rho}_S(n) \in [\min_{1\leq i\leq n} \hat{\rho}_{i,n}, \max_{1\leq i\leq n} \hat{\rho}_{i,n}]$.*

From this Lemma, we obtain the following result.

Theorem 2. *For any deterministic $0 \leq \nu < \infty$, the estimator $\hat{\rho}_S(n)$ defined in (3) is consistent.*

For the case when $\nu = 0$, the proof uses the three following results, given for some fixed $\epsilon > 0$, and, as $n \to \infty$, some large n_0 and L_{n_0}:

- ◇ For any $i \geq 1$, $\hat{\rho}_{i,n}$ converges $\mathbb{P}$-almost surely to $\boldsymbol{\rho}$. We show that, for all $i \leq n - n_0$ and all $\rho < \boldsymbol{\rho} - \epsilon$, with probability close to one, $R_{i,n}(\boldsymbol{\rho} - \epsilon/2) \geq R_{i,n}(\rho)$.

- ◇ For any $\rho > 0$ and any $i \geq 1$, $R_{i,n}(\rho)$ tends $\mathbb{P}$-almost surely to ∞. We show that, with probability close to one, $\sum_{i=n-n_0+1}^{n} R_{i,n}(\rho) \leq L_{n_0}$.

- ◇ The difference $R_{i,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) - R_{i,n}(\boldsymbol{\rho} - \epsilon)$ tends $\mathbb{P}$-almost surely to ∞. Then, with probability close to one, $R_{1,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) \geq 2L_{n_0} + R_{1,n}(\boldsymbol{\rho} - \epsilon)$.

We recall that $S_n(\rho) = \sum_{i=1}^{n} R_{i,n}(\rho)$. We can combine the three results above. With some reorganization, we show that, for n large enough and all $\rho < \boldsymbol{\rho} - \epsilon$,

$$S_n\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) = R_{1,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) + \sum_{i=2}^{n-n_0} R_{i,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) + \sum_{i=n-n_0+1}^{n} R_{i,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) \geq S_n(\rho).$$

By a symmetric argument, $S_n(\boldsymbol{\rho} + \epsilon/2) \geq S_n(\rho)$ for all $\rho > \boldsymbol{\rho} + \epsilon$, and the result follows by definition of $\hat{\rho}_S(n)$. We extend this result to any deterministic $0 \leq \nu < \infty$ by using the fact that there is only a finite number of X_i with intensity λ, and a growing number of X_i with intensity $\lambda\rho$.

In Figures 1 and 2 we numerically compare the estimator $\hat{\rho}_S(n)$ to usual estimators (maximum likelihood estimator, least square error estimator). It appears that:

- ◇ The average bias of $\hat{\rho}_S(n)$ decreases faster than in the case of the usual estimators;

- ◇ The variance of $\hat{\rho}_S(n)$ is always lower, also all variances are asymptotically close;

- ◇ As expected, $\hat{\rho}_S(n)$ is numerically close to one in average before the change.

These properties, numerical and theoretical, make this estimator a first choice compared to the usual estimators considered in the investigation for the inference of the post change parameter, especially just after the change.

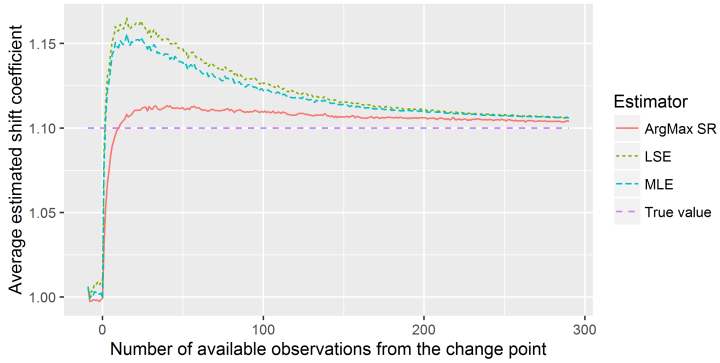


Figure 1 – Average sequential estimation of $\boldsymbol{\rho}$. *Centered on the time of change.*

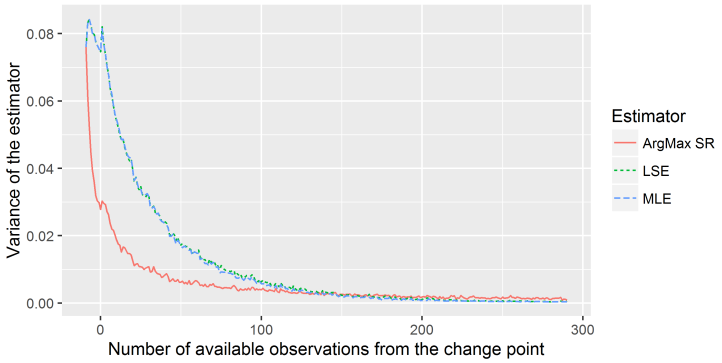


Figure 2 – Empirical sequential variance of estimators of $\boldsymbol{\rho}$. *Centered on the time of change.*

2.2.2 Detection of a change of trend

We also explored another setup for the change-point detection procedure where the intensity of the Poisson random variables sequence is increasing/decreasing at a steady pace. We look for a change in the trend and suggest a way for the inference of the post-change parameters. For that purpose, consider an independent random sequence $(X_n)_{n\geq 1}$ such that X_n follows a Poisson distribution with parameter λ_n where $\lambda_n := \alpha\lambda_{n-1}$ before the change and $\lambda_n := \alpha'\lambda_{n-1}$ after the change. We assume that α and α' cannot be equal. In this context, λ_0 and α are deterministic and known. The parameters ν and α' are deterministic but unknown. We provide an adaptive procedure that uses estimates for both ν and α'. For fixed $n \geq 1$, the associated sequence $\xi = (\xi_{i,n})_{1\leq i\leq n}$ is given for $i \in \{1, \ldots, n\}$ by

$$\xi_{i,n} := \sum_{k=1}^{i} \prod_{j=k}^{i} \frac{f^j_{\hat{\theta}_n}}{f^j_{\theta_0}}(X_j).$$

Here $\hat{\theta}_n := (\hat{\alpha}'_\xi(n), \hat{\nu}_\xi(n))$ is an estimator of the couple (α', ν). We denote by $(f^i_{\hat{\theta}_n})_{1\leq i\leq n}$ the sequence of density functions, with respect to the counting measure on $\mathbb{N}$, of independent Poisson random variables with intensity $\lambda_i := \alpha\lambda_{i-1}$ before the estimated change-point $\hat{\nu}_\xi(n)$ and with intensity $\lambda_i := \hat{\alpha}'_\xi(n)\lambda_{i-1}$ after the estimated change-point $\hat{\nu}_\xi(n)$. In practice, we used the maximum likelihood estimator (MLE) for $\hat{\theta}_n$. Numerical simulations seem to indicate that, in this case, the MLE is among the best options we have for the estimations of α' and ν.

2.2.3 Applications

A Poisson framework is commonly used for the mortality (Rhodes and Freitas (2004) and Tomas and Planchet (2015)). After data normalization as described in Zucchini and MacDonald (2009) and Mei et al. (2011), we studied the French national mortality through the following events:

- ◇ The 2003 French heatwave,
- ◇ The 1918 Spanish flu,
- ◇ The decrease of the mortality rate in the 60's as a consequence of the revolution in cardiovascular care.

During the summer 2003, the death toll of the heat wave exceeded 70 000 in Europe (Robine et al. (2008)). In France, it impacted significantly the mortality of elderly people. We studied the age group 85-90 years old. No change of level is detected when the 2003 peak occurs. However, the decrease of the mortality that follows (from 12.3% in 2000 to 9.9% in 2006) is detected in 2005 ($\hat{\rho}_S(2005) = 0.89$). As expected, a persistent change of level is detected but not a sudden variation. With the procedure for detecting a change of trend, surprisingly, the two years 2002 and 2003 are sufficient to raise an alarm for the observed increase (relative increase of +3.5% per year) since the observations are aligned enough. This is a clear example of a peak/compensation phenomenon. Once we identified the peak and the compensation period through multiple applications of the procedure, we can test whether the regularly decreasing period has the same trend as the initial period.
The death toll of the 1918 Spanish flu is evaluated at 17.4 million worldwide (Spreeuwenberg et al. (2018)). We studied the case of French women (Caselli et al. (1987)). Observations suggest that the mortality is stable before the event. A strong peak is visible in 1918. The change of level is detected in 1918, the same year it happens, with $\hat{\rho}_S(1918) = 1.5$.
As a consequence of the revolution in cardiovascular care, the French mortality rate decreased in the 60's (Vallin and Meslé (2010), Cutler and Meara (2001)). We studied the case of women between 55 and 75 years old. Didou (2011) noticed that the female mortality rates are slightly decreasing at a steady peace just before 1960 and the decrease accelerates in the 60's. The trend changes around 1976/1978 and the procedure rises an alarm in 1986. The estimated decrease of mortality in 1960 is about 1.5% per year and becomes 3.2% after the change, i.e. starting from 1978.

We also studied a portfolio of life annuitants from a real insurance portfolio. The dataset contains about 15 000 life annuities. Because of the low size of the portfolio, we assume that the expected mortality is given by the French regulatory tables. The procedure for a change of level does not raise any alarm: no persistent change of level seems to occur for the whole portfolio. A close analysis of the observations might suggest that there is a deviance between the observed and predicted deaths from 2011. With the procedure for detecting a change of trend, an alarm is raised at mid-year 2013. The initial decreasing trend of 0.6% changes to an annual increase of the mortality of 0.7% from the last quarter of 2005. A re-run from the second half of 2013 in order to detect a late change does not raise any alarm. We conclude that the studied portfolio of life annuitants does not diverge significantly from its reference mortality table.

In summary, the use of the procedure for detecting a change of level is assessed to be efficient for detecting small persistent deviations. The procedure for the detection of a change of trend is quite sensitive and detects small persistent changes of trend, even over a short period of time.

2.3 Discussion

We recall that the usual detection schemes assume that the post-change distribution is known and our contribution consisted in introducing a consistent sequential estimator for the post-change parameter in the case of an independent Poisson sequence. We numerically found out that this estimator seems to converge faster to the true parameter, with lower variance. We used this estimator in a Shiryaev-Roberts scheme.

Such schemes are built to be updated very rapidly as a new observation arrives. Our estimator is defined such that, for each new observation, its computation requires to re-run many times the scheme from the first observation. Therefore it is rather dedicated to applications where the frequency of arrival of new observations is low enough. For many actuarial applications, and especially the study of mortality as shown in the applications, this is not really a problem because the dataset remains small enough (or at least can be seen as a small sample). For other applications where the frequency of arrival of new observations overcomes the computational capacities, one can consider different aggregation techniques (e.g. using a moving average or subsamples) in order to allow our approach to be computed.

The Shiryaev-Roberts detection scheme is optimal in minimizing the average detection delay when:

- ◇ The change-point is a generalized random variable with a uniform improper prior distribution,
- ◇ The parameters before and after the change are known,
- ◇ The average delay of false alarm is below some fixed constant.

In our framework, we assume that we do not know the post-change parameter. Then, is the procedure still optimal? By definition, for each new observation, the highest possible value of the Shiryaev-Roberts statistic is attained with our estimator. Since the false alarm detection delay is also impacted, this does not ensure that our statistic gives the quickest response. Still, with an appropriate threshold that depends on n, one can control accurately the false alarms and still expect a quick answer from the procedure and benefit from the estimation properties of our estimator.

3 Weighted likelihood test for a change in one component of a parametric mixture

This section introduces Chapter 3.

3.1 Motivation

We consider an experiment where we observe a sample of n independent continuous random variables $(X_i)_{1\leq i\leq n}$ with values in a real vector space $\mathcal{X}$. For a parameter $\theta = (p_1, \ldots, p_{m-1}, \lambda_1, \ldots, \lambda_m)$ in a set of eligible parameters Θ, the variable X_1 follows a **finite parametric mixture** with m components if it has a density of the form

$$f(x, \theta) := \sum_{k=1}^{m} p_k f_k(x, \lambda_k),\ x \in \mathcal{X},$$

where $f_1, \ldots, f_m$ are some fixed density functions on $\mathcal{X}$ and $p_m := 1 - \sum_{k=1}^{m-1} p_k$.

The concept of mixture distributions arises in many fields and has been popular in the literature for the last decades, as it allows to describe experiments with different sub-populations (Pearson (1894)). The earliest books on finite mixtures (Everitt and Hand (1981), Titterington et al. (1985)) already include a few recurring topics that emerge when studying mixtures: the identifiability of the mixture model; the estimation of the mixture parameters (e.g. with the EM algorithm by Dempster et al. (1977),

the Bayesian framework, the likelihood approach or the method of moments); and the question of determining the number of components in the mixture. Later books, as in McLachlan and Peel (2000), Frühwirth-Schnatter (2006), Pons (2009) or Lachos Dávila et al. (2018), provide an extensive overview of the literature with the solutions of the above mentioned questions and some innovations: Hidden Markov Models, further investigations in the early topics, speed of convergence of algorithms, multivariate mixtures, and other specific topics. Recently, Frühwirth-Schnatter et al. (2019) gathers most of the recent work on the topic in both theoretical and applied fields: newest versions of the EM algorithm, Bayesian inference and the role of Monte-Carlo Markov Chains, non-parametric mixtures, the use of clustering for modeling mixtures, the case of heavy tails and skewness, and various applications in image analysis, genomics, astronomy or finance.

In this work, we assume that:

- ◇ The set of eligible parameters $\Theta = \Theta_0 \times \prod_{k=1}^{m} \Theta_k$ is a subset of a d-dimensional Euclidean space and each parameter λ_k belongs to a set Θ_k.
- ◇ The number of components $2 < m < \infty$ is deterministic and known.
- ◇ The vector of weight parameters $(p_1, \ldots, p_{m-1})$ belongs to the open set $\Theta_0 := \{(p_1, \ldots, p_{m-1}) \in (0,1)^{m-1}, \sum_{k=1}^{m-1} p_k < 1\}$.
- ◇ The mixture is identifiable in the sense that different values of θ lead to different laws $\mathbb{P}_\theta$.

The last assumption means that, in particular, we can order and distinguish the components of the mixture (Redner (1981), Feng and McCulloch (1996), McLachlan and Peel (2000)).

Considering the closed sample $(X_i)_{1 \leq i \leq n}$, we are interested in **detecting a change in the parameters of the distribution that describes the first component of the mixture** for the case when there is *at most one change* (AMOC model). This is motivated by examples from applications. Here is one possible example: consider a study where the initial purpose is to model a population but, during initial phases (gathering the data, first data visualization, etc.), the modeler understands that the population of interest is actually one part of a wider group and they cannot be distinguished. The initial interest does not fade and the model is now constrained to consider the larger group. More generally, detecting a change in only one component makes sense as soon as, when studying a phenomenon, each component plays a specific role with specific implications in practice.

The detection of a change-point in a closed sample is a standard problem for which techniques already exist, such as the ones in the book of Csörgő and Horváth (1997) [4]. Dedicated techniques for mixtures are not so common because such models raise already so many difficulties when it comes to the inference of the parameters (including the number of components). To our knowledge, very few references seem to specifically detect change-points in mixtures with likelihood ratio-based techniques [5] (Andrews and

4. For a review of techniques, see also Barber et al. (2011), Chen and Gupta (2012), Killick and Eckley (2014) and the references therein.

5. Some of the existing work is dedicated to a Bayesian framework and therefore not in the scope of this Chapter. See e.g. Giordani and Kohn (2008), Pandya and Jadav (2009), Pandya and Jadav

Ploberger (1994), Hansen (1996), Pons (2009), Zou et al. (2015)). It turns out that standard change-point detection techniques can be adapted for finite parametric mixtures (Csörgő and Horváth (1997), van der Vaart (1998), Pons (2018)).

For AMOC models, the standard general approach based on a likelihood ratio is, for example, exposed in Csörgő and Horváth (1997). For that purpose, consider that each parameter $\theta = (a, b) \in \Theta$ is defined by two sub-parameters a and b. The null hypothesis H_0 of the test assumes that no change happens, i.e. $\theta^1 = \cdots = \theta^n$. The alternative hypothesis H_1 assumes that, for some k, a change occurs between the k-th and the $k+1$-th observation for the sub-parameter a, i.e. $a^1 = \cdots = a^k \neq a^{k+1} = \cdots = a^n$ and $b^1 = \cdots = b^n$. Here b is called a *nuisance* parameter. The log-likelihood ratio associated with the test is defined by

$$\Lambda_{k,n}^{BM} := \log \left(\frac{\sup_{(a,b),(a',b)\in\Theta} \prod_{i=1}^{k} f(X_i, (a,b)) \prod_{i=k+1}^{n} f(X_i, (a',b))}{\sup_{(a,b)\in\Theta} \prod_{i=1}^{n} f(X_i, (a,b))} \right). \tag{4}$$

The test statistic is defined as $\max_{1\leq k\leq n} 2\Lambda_{k,n}^{BM}$. By considering that the sub-parameter a represents the parameters (p_1, λ_1) of the first component of a finite mixture, this general test, referred in the following as the *benchmark* test (BM), can be used to answer the problem (see Section 3.6.1 for a detailed construction). Note that the separation between the parameters of interest a and the nuisance parameters b allows the test to focus on a change in the first component.
The optimization problem in the denominator is a Maximum Likelihood Estimation problem. It can be treated with known algorithms such as the EM algorithm introduced by Dempster et al. (1977) and pre-implemented in calculation softwares such as R (algorithm based on Benaglia et al. (2009)). However, for the supremum in the numerator of $\Lambda_{k,n}^{BM}$, when it comes to numerical applications, many computational difficulties arise:

- ◇ This optimization problem does not have an explicit solution,
- ◇ No dedicated algorithm exists to numerically solve it,
- ◇ With standard optimization algorithms, the numerical solutions are not satisfactory when working on real data or large samples,
- ◇ The run time increases so much for large sample that the statistic cannot be computed with a reasonable run time.

Note that these difficulties mainly come from the role played by the parameters a, a' and b. In order to circumvent them, we suggest in this thesis a different approach to focus on the first component. In particular, we propose using a weighted likelihood ratio that will increase the contribution of the X_i that are more likely to belong to the first component. Moreover we will build the weighted log-likelihood ratio so that is can be computed with standard estimation algorithms.

3.2 Our contribution

Our contribution consists in the introduction of two alternative hypothesis tests that are based on weighted likelihood ratios which require only known inference algorithms to be computed. The first version uses weights to help the likelihood ratio to zoom on

(2010), Wilson et al. (2013), Li et al. (2018) or Ganji and Mostafayi (2019).

the first component. In the second extended version, we added an adjustment that helps improve the type II error. With a technique from Davis et al. (1995), we derive the limit distribution of their statistics under the null hypothesis in the form of quadratic forms of multidimensional Brownian motions, with the help of a dedicated functional limit theorem. We show that the validity conditions of the main result hold for univariate finite Gaussian mixtures within the framework of Hathaway (1985). Numerical applications on simulated data illustrate the advantage of the alternative tests compared to the benchmark test defined by (4)*. An application with Property and Casualty insurance real data is provided for the alternative tests.*

3.2.1 Definition of the Weighted Likelihood Test (WLT)

With the notations above, we still consider that the elements of the sample $(X_i)_{1\leq i\leq n}$ are defined by a finite parametric mixture distribution where there is *at most one change* (AMOC), deterministic but unknown, or none. Since we are interested in the limit behavior of the sample of variables when their number tends to infinity, we suppose that the experience takes place in the time interval $[0,1]$ and each variable X_i occurs at time i/n, $1 \leq i \leq n$. We impose that, if there is a change-point, it happens at some time in the interval $[\bar{s}, 1-\bar{s}] \subset (0,1)$, where $0 < \bar{s} < 1/2$ is deterministic and known. This means that the change does not occur too close to 0 nor 1. Changes close to those values are difficult to be detected and would be of less significance, since their impact on the full sequence is small. The hypothesis test becomes:

1. The *null hypothesis* H_0 defines the case when there is no change-point: the mixture is defined by the true parameter $\boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{m-1}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m) \in \Theta$.
2. The *alternative hypothesis* H_1 : a change-point occurs at some time s, $s \in [\bar{s}, 1-\bar{s}]$, i.e. the parameters which describe the distribution of the first component are different before and after s while the other parameters of the mixture remain the same.

We will establish a central limit theorem under the assumption that H_0 holds. This is necessary to be able to determine the rejection domain while controlling the type I error.

Under some regularity conditions detailed in Section 3.2.1, the usual limit theorems for Maximum Likelihood Estimators (MLE) ensure that there exist sequences of solutions of the likelihood equations that are consistent (Mäkeläinen et al. (1981) for Gaussian mixtures, McLachlan and Peel (2000) for mixtures in general, Section 6.5 in Lehmann and Casella (1998) for the general case). Let us consider one of these consistent sequences of solutions as an *estimator for the unknown* $\boldsymbol{\theta}$ and denote it $\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_{m-1}, \hat{\lambda}_1, \ldots, \hat{\lambda}_m)$. For $s \in [\bar{s}, 1-\bar{s}]$, by the same logic, we consider the estimators of $\boldsymbol{\theta}$ over the subsamples $(X_i)_{1\leq i\leq \lfloor sn \rfloor}$ and $(X_i)_{\lfloor sn \rfloor + 1 \leq i \leq n}$, respectively denoted by $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$. We assume that $\hat{\theta}$ is strongly consistent, which, for example, holds for the Gaussian case under some conditions (Hathaway (1985)). For other cases, classical results in the literature cover a wide range of situations for providing reasonable sufficient conditions (see e.g. Redner and Walker (1984), Feng and McCulloch (1996), McLachlan and Peel (2000) and the references therein). Strong consistency remains an important restriction compared to the general case because usual regularity conditions only ensure consistency.

With the *weight function* at point $x \in \mathcal{X}$ for $\theta \in \Theta$, defined by

$$w(x,\theta) := \frac{p_1 f_1(x,\lambda_1)}{f(x,\theta)},$$

we introduce the *weighted log-likelihood ratio* denoted by $\Lambda_{s,n}$ and defined for $s \in [\bar{s}, 1-\bar{s}]$ and $n \geq 1$ by

$$\Lambda_{s,n} := \log\left(\frac{\prod_{i=1}^{\lfloor sn \rfloor} f_1(X_i, \hat{\lambda}_{0,s,1})^{w(X_i,\hat{\theta}_{0,s})} \prod_{j=\lfloor sn \rfloor+1}^{n} f_1(X_j, \hat{\lambda}_{s,1,1})^{w(X_j,\hat{\theta}_{s,1})}}{\prod_{i=1}^{n} f_1(X_i, \hat{\lambda}_1)^{w(X_i,\hat{\theta})}}\right). \tag{5}$$

Note that, for an observation X_i with distribution parameter θ, the weight $w(X_i,\theta) = p_1 f_1(X_i,\lambda_1)/f(X_i,\theta)$ is the probability that X_i comes from the first component. Conditionally to this fact, the log-likelihood of X_i is given by $\log f_1(X_i,\lambda_1)$. We can extend this logic to all the components of the mixture such that the vector

$$\left(\frac{p_1 f_1(X_i,\lambda_1)}{f(X_i,\theta)} \log f_1(X_i,\lambda_1), \ldots, \frac{p_m f_m(X_i,\lambda_m)}{f(X_i,\theta)} \log f_m(X_i,\lambda_m)\right)$$

can somehow be interpreted as the contributions of the m components to the log-likelihood of X_i. Thus the expression $w(X_i,\theta)\log f_1(X_i,\lambda_1)$ in $\Lambda_{s,n}$ reflects the contribution of the first component. As a consequence, the response of the statistic is magnified when a change occurs in the first component.
The *test statistic* is then defined by

$$S_n := \sup_{s\in[\bar{s},1-\bar{s}]} \Lambda_{s,n}.$$

We refer to this test as the **WLT** (Weighted Likelihood Test). As expected, it can be computed with the help of the usual MLE algorithm such as the EM algorithm (McLachlan and Peel (2000), Benaglia et al. (2009)). Our main result consists in deriving the limit distribution of the test statistic under the null hypothesis (Theorem 4) with the help of a dedicated functional limit theorem (Theorem 3).

3.2.2 Main result

With a similar approach as in the work of Davis et al. (1995), the càd-làg real-valued process $\Lambda_n := (\Lambda_{s,n})_{[\bar{s},1-\bar{s}]}$ can be decomposed as $\Lambda_{s,n} = Q^1_{s,n} + Q^2_{s,n} - Q^1_{1,n}$, for $s \in [\bar{s}, 1-\bar{s}]$, where

$$Q^1_{s,n} := \sum_{i=1}^{\lfloor sn \rfloor} \Big(w(X_i,\hat{\theta}_{0,s})\log f_1(X_i,\hat{\lambda}_{0,s,1}) - w(X_i,\boldsymbol{\theta})\log f_1(X_i,\boldsymbol{\lambda}_1)\Big) - \boldsymbol{u}^T \boldsymbol{I}^{-1} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i,\boldsymbol{\theta}),\ s \in [\bar{s},1],$$

$$Q^2_{s,n} := \sum_{i=\lfloor sn \rfloor+1}^{n} \Big(w(X_i,\hat{\theta}_{s,1})\log f_1(X_i,\hat{\lambda}_{s,1,1}) - w(X_i,\boldsymbol{\theta})\log f_1(X_i,\boldsymbol{\lambda}_1)\Big) - \boldsymbol{u}^T \boldsymbol{I}^{-1} \sum_{i=\lfloor sn \rfloor+1}^{n} D_\theta(\log f)(X_i,\boldsymbol{\theta}),\ s \in [\bar{s},1-\bar{s}].$$

Here D_θ denotes the gradient operator in $\theta \in \mathbb{R}^d$ and $\boldsymbol{I}$ the Fisher information matrix under the null hypothesis. Further note that the finite constant

$$\boldsymbol{u} := \mathbb{E}_{H_0}\left[D_\theta(w \log f_1)(X_1, \boldsymbol{\theta})\right] \in \mathbb{R}^d$$

is non null in general. The random processes $(Q^1_{s,n})_{s\in[\bar{s},1-\bar{s}]}$ and $(Q^2_{s,n})_{s\in[\bar{s},1-\bar{s}]}$ have a similar structure that differs only by the sub-sample considered. Therefore, in order to obtain a limit distribution for the process Λ_n, we study $Q^1_n := (Q^1_{s,n})_{s\in[\bar{s},1]}$ and extend the result to Λ_n.

We first show the following functional limit result.

Theorem 3. *Under the null hypothesis and some regularity conditions detailed in Section 3.2.1, in the Skorokhod metric space of càd-làg real-valued functions on* $[\bar{s}, 1]$*, the process*

$$Q^1_n \xrightarrow[n\to\infty]{\mathcal{D}} \left(\frac{1}{s}\boldsymbol{q}(W_s)\right)_{s\in[\bar{s},1]},$$

where $W := (W_s)_{s\in[0,1]}$ *is a standard* $2d + d^2$*-dimensional Brownian motion and* $\boldsymbol{q}$ *is a quadratic form that we shall define in Section 3.3.2 by Equation* (3.33)*.*

In order to prove this result, we start by showing that, by a Taylor-Lagrange development of $\hat{\theta}_{0,s}$ around $\boldsymbol{\theta}$, $Q^1_{s,n}$ can be expressed as a function of:

- ◇ the difference $\hat{\theta}_{0,s} - \boldsymbol{\theta}$,
- ◇ an average $\hat{u}_{0,s}$ of i.i.d. $\mathbb{R}^d$ valued random vectors,
- ◇ and a random variable depending on s and n that converges a.s. and uniformly in $s \in [\bar{s}, 1]$ to some constant.

The difference $\hat{\theta}_{0,s} - \boldsymbol{\theta}$ can itself be expressed as a function of:

- ◇ an average $\hat{\iota}_{0,s}$ of i.i.d. $\mathbb{R}^d$ valued random vectors,
- ◇ an average $\hat{I}_{0,s}$ of i.i.d. $d \times d$ random matrices,
- ◇ and a random variable depending on s and n that converges a.s. and uniformly in $s \in [\bar{s}, 1]$ to some constant.

An important ingredient of this decomposition is, as in Lehmann and Casella (1998), a $d \times d$-matrix $\hat{A}^{-1}_{0,s}$ which converges to the inverse of the Fisher information matrix $\boldsymbol{I}^{-1}$ (uniformly in s). It follows that $Q^1_{s,n}$ can be expressed as a function of:

- ◇ the triple $\hat{\xi}_{0,s} = (\hat{\iota}_{0,s}, \hat{u}_{0,s} - \boldsymbol{u}, \hat{I}_{0,s})$,
- ◇ and some random variable depending on s and n that converges a.s. and uniformly in $s \in [\bar{s}, 1]$ to some constant.

As an application of Donsker's Theorem in the Skorokhod metric space (Billingsley (1999)), we can show that the random process $(\hat{\xi}_{0,s})_{s\in[\bar{s},1]}$ converges weakly to the process $(\frac{1}{s}\Sigma W_s)_{s\in[\bar{s},1]}$, where Σ^2 is the covariance matrix under H_0 of the i.i.d. terms in the average $\hat{\xi}_{0,s}$.
This weak convergence is extended to the process Q^1_n with arguments based on a extended functional version of Slutsky's Theorem 1.13, the Continuous Mapping Theorem 1.11 and a Functional Delta Method (Corollary 1.16) inspired from van der Vaart (1998) and adapted for the metric Skorokhod space. In particular, the process Q^1_n is a function of $(\hat{A}^{-1}_{0,s})_{s\in[\bar{s},1]}$. In order to handle the inversion of the matrix $\hat{A}_{0,s}$, the functional delta

method is applied to a map defined for a càd-làg process $(x_s, y_s, M_s)_{s\in[\bar{s},1]}$, with x_s and y_s d-dimensional vectors and M_s a $d \times d$ matrix, by

$$(x_s, y_s, M_s)_{s\in[\bar{s},1]} \mapsto (x_s, y_s, M_s^{-1})_{s\in[\bar{s},1]}.$$

To this end, we extend the derivative of the application that inverts a matrix (Abraham et al. (1988), Dudley and Norvaiša (2011)).

With similar arguments, we can extend to Λ_n the result obtained for Q_n^1. For that purpose, we remark that the process $(Q_{s,n}^1, Q_{s,n}^2, -Q_{1,n}^1)_{s\in[\bar{s},1-\bar{s}]}$ can be seen as a function of the process

$$\left(\hat{\xi}_{0,s}, \frac{n}{n-\lfloor sn\rfloor}\hat{\xi}_{0,1} - \frac{\lfloor sn\rfloor}{n-\lfloor sn\rfloor}\hat{\xi}_{0,s}, \hat{\xi}_{0,1}\right)_{s\in[\bar{s},1-\bar{s}]}.$$

After deriving the limit distribution of $(Q_{s,n}^1, Q_{s,n}^2, -Q_1^1)_{s\in[\bar{s},1-\bar{s}]}$ with similar arguments as for Q_n^1, the following main result is obtained by a last application of the Continuous Mapping Theorem.

Theorem 4. *Under the null hypothesis H_0 and some regularity conditions detailed in Section 3.2.1, if $\hat{\theta}$ is strongly consistent, then*

$$S_n \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}(W_s - sW_1)}{s(1-s)},$$

where $W := (W_s)_{s\in[0,1]}$ is a standard $2d + d^2$-dimensional Brownian motion and $\boldsymbol{q}$ is a quadratic form that we shall define in Section 3.3.2 by Equation (3.33).

It follows that the limit process is a quadratic form of a Brownian bridge, which is quite standard in other frameworks. For the *benchmark* test (BM) defined by (4), Csörgő and Horváth (1997) show that $\max_{1\le k\le n} 2\Lambda_{k,n}^{BM}$ converges weakly to the supremum of the sum of d_a squared Brownian bridges, where d_a is the dimension of the a parameter. See Shorack and Wellner (1986) for an example with $d_a = 1$. A similar result is also given in the alternative work of Davis et al. (1995), or later applications as in Dehling et al. (2014).

We already pointed out that the WL test can be computed with the help of the usual MLE algorithm such as the EM algorithm (McLachlan and Peel (2000), Benaglia et al. (2009)). From a practical point of view, this is an improvement compared to the benchmark test. However, the introduction of the weights increases significantly the dimension of the Brownian bridge. For the WLT, it is $2d + d^2$, where d is the dimension of the parameter θ, while, in the benchmark test, it is d_a, the dimension of the subparameter a in (4) (e.g. $d = 3m$ and $d_a = 4$ for an univariate Gaussian mixture with m components).

3.2.3 Extension: scaling the contributions in the likelihood ratio (EWLT)

Early numerical applications showed that the test defined by (5) can be improved by multiplying $\Lambda_{s,n}$ by an adjustment factor. To this end, we start by defining for a fixed $s \in [\bar{s}, 1]$ the contribution $c_{s,n}$ as follows:

$$c_{s,n} := \sum_{i=1}^{\lfloor sn\rfloor} w(X_i, \hat{\theta}_{0,s}) + \sum_{i=\lfloor sn\rfloor+1}^{n} w(X_i, \hat{\theta}_{s,1}).$$

Remark that the log-ratio $\Lambda_{s,n}$ is the difference of $(Q^1_{s,n} + Q^2_{s,n})$ and $Q^1_{1,n}$. Then, $c_{s,n}$ is the contribution of the sample to the term $(Q^1_{s,n} + Q^2_{s,n})$, and $c_{1,n} = \sum_{i=1}^n w(X_i, \hat{\theta})$ is the contribution of the sample to the term $Q^1_{1,n}$. Under the null hypothesis, we can show that the random variable $c_{s,n}/n$ converges a.s. to the parameter $\boldsymbol{p}_1$, uniformly in $s \in [\bar{s}, 1]$. It follows that the contributions $c_{s,n}/n$ and $c_{1,n}/n$ tend to be the same as $n \to \infty$. However, when a change occurs, they differ in some cases, which leads to an increase of the type II error (false negative) as a side effect. Therefore, we suggest to scale the total contributions. We define a new log-ratio process $\Lambda^*_n := (\Lambda^*_{s,n})_{s\in[\bar{s},1-\bar{s}]}$ by

$$\Lambda^*_{s,n} := \frac{c_{1,n}}{c_{s,n}} \left(\sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) \log f_1(X_i, \hat{\lambda}_{0,s,1}) + \sum_{i=\lfloor sn \rfloor+1}^{n} w(X_i, \hat{\theta}_{s,1}) \log f_1(X_i, \hat{\lambda}_{s,1,1}) \right) - \sum_{i=1}^{n} w(X_i, \hat{\theta}) \log f_1(X_i, \hat{\lambda}_1) \tag{6}$$

that can be reorganized as follows

$$\Lambda^*_{s,n} = \frac{c_{1,n}}{c_{s,n}} \Lambda_{s,n} - \frac{\frac{1}{n}\sum_{i=1}^n w(X_i, \hat{\theta}) \log f_1(X_i, \hat{\lambda}_1)}{\frac{1}{n} c_{s,n}} \left(\sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) + \sum_{i=\lfloor sn \rfloor+1}^{n} w(X_i, \hat{\theta}_{s,1}) - \sum_{i=1}^{n} w(X_i, \hat{\theta}) \right).$$

The ratio $c_{1,n}/c_{s,n}$ converges a.s. to 1 uniformly in $s \in [\bar{s}, 1]$. Moreover the factor $(1/n \ \sum_{i=1}^n w(X_i, \hat{\theta}) \log f_1(X_i, \hat{\lambda}_1))/(c_{s,n}/n)$ converges a.s. to some finite constant, uniformly in $s \in [\bar{s}, 1]$. Note also that the sum

$$\sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) + \sum_{i=\lfloor sn \rfloor+1}^{n} w(X_i, \hat{\theta}_{s,1}) - \sum_{i=1}^{n} w(X_i, \hat{\theta})$$

has the same form as $\Lambda_{s,n}$, but without the factor $\log f_1(X_i, \hat{\lambda}_{\cdot,\cdot,1})$. It follows that, defining the test statistic by $S^*_n := \sup_{s\in[\bar{s},1-\bar{s}]} \Lambda^*_{s,n}$, its limit distribution is obtained with similar arguments as for Theorem 4. We refer to this test as the **EWLT** (Extended Weighted Likelihood Test).

Theorem 5. *Under the null hypothesis H_0 and some regularity conditions detailed in Section 3.2.1, if $\hat{\theta}$ is strongly consistent, then*

$$S^*_n \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}^*(W_s - sW_1)}{s(1-s)},$$

where $W := (W_s)_{s\in[0,1]}$ is a standard $3d + d^2$-dimensional Brownian motion and $\boldsymbol{q}^*$ is a quadratic form that we shall define in Section 3.4 by Equation (3.39).

This extension improves significantly the detection quality. In particular the type II error is small, compared to the first version of the test.

3.2.4 Example: the case of univariate finite Gaussian mixtures

If the sample $(X_i)_{1\le i\le n}$ follows an *univariate finite Gaussian mixture*, then each X_i is defined by a parameter θ^i in the set of eligible parameters

$$\Theta \ni \theta = (p_1, \ldots, p_{m-1}, \mu_1, \sigma_1, \ldots, \mu_m, \sigma_m)$$

such that, for each k-th component, $\mu_k \in \mathbb{R}$ is the mean parameter and $\sigma_k \in \mathbb{R}_+^*$ is the standard deviation parameter. We impose that the means are strictly increasing:

$$\mu_1 < \mu_2 < \cdots < \mu_m,$$

and, as in Hathaway (1985), that the variances are bounded:

$$\min\left\{\frac{\sigma_j}{\sigma_k},\ 1 \leq j,k \leq m\right\} > \boldsymbol{b},$$

where $0 < \boldsymbol{b} \leq 1$ is a *dispersion boundary*. We suppose that the true parameter $\boldsymbol{\theta}$ satisfies these two conditions. By Hathaway (1985) the MLE $\hat{\theta}$ is strongly consistent, and we can show that the conditions of Theorems 4 and 5 hold in this situation.

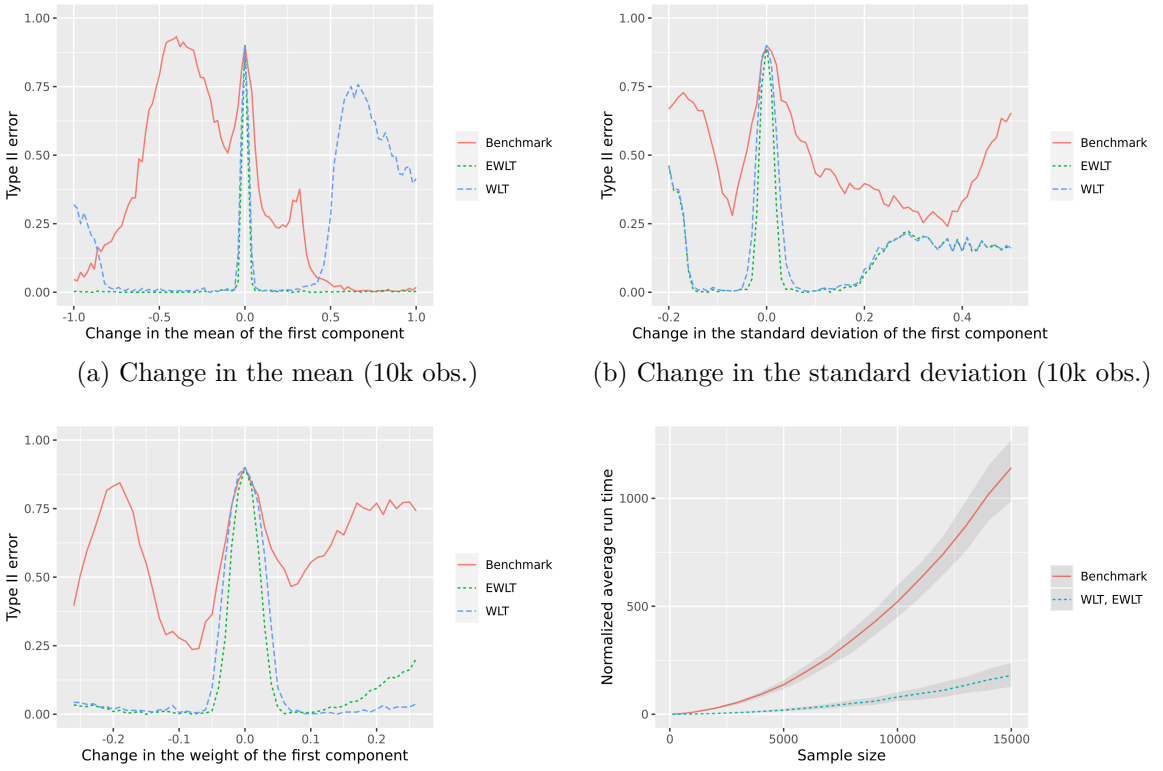
3.2.5 Applications

We provide two distinct applications for the case of an univariate finite Gaussian mixture. We are interested in detecting a change that is not visible to the naked eye (small) but also not too close to 0 (no impact in practice), for large samples (over 10 000 observations).

First, with numerical simulations of a Gaussian mixture with 3 components where we know the parameters before and after the change in the first component, we compare our two tests, the WL test defined as Λ_n in (5) and the EWL test defined as Λ_n^* in (6), to the *benchmark* test (BM) defined as Λ_n^{BM} in (4). The results are obtained by multiple re-simulations of samples (see Section 3.6.2 for a detailed setup):

- ◇ When a change occurs in the first component, we look at the type II error (proportion of false negative) to evaluate the test quality (Figure 3). For large samples, the benchmark test performs poorly with a significantly longer run time essentially due to the fact that it is computed with standard algorithms that are not solving properly the supremum in the numerator of Λ_n^{BM}. Since their estimation algorithms are more robust, our two tests have both very low type II errors for small changes in the parameters on a sample of 10k observation and, in most cases, the EWLT performs notably better than the WLT.
- ◇ For large samples of more than 10k observations, regardless of algorithmic issues, the three tests have similar type II errors. For samples with 1k observations, the EWLT and the benchmark test are comparable but the WLT seems less effective.
- ◇ As a nice-to-have, we also studied the detection frequency of each test when a change occurs in the second or third component where the quality of the test is characterized by a low detection frequency. The main observation is that the EWLT shows the best results and, for large samples of 10k observations, the WLT is better than the benchmark test (the benchmark test still behaves poorly for large samples). However, we remark that the three tests show systematic patterns of high detection frequency for a change in the mean and the standard deviation: there is still some room for improvement regarding this criterion.

We conclude that the EWLT has the best numerical properties out of the three tests.



(a) Change in the mean (10k obs.)

(b) Change in the standard deviation (10k obs.)

(c) Change in the weight parameter (10k obs.)

(d) Run time of the WL, EWL and benchmark tests for an increasing sample size.

Figure 3 – Type II error and run time for a change in the first component.

The second application is an illustration of the WL and the EWL tests to real data from the insurance industry, in particular the bodily injuries from the motor claims. We denote by C_t the real-valued random variable that represents the amount that, at time t, the insurer expects to pay eventually ($t = 0$ being the declaration date of the claim). This amount varies over time when the claim is reviewed, until it is settled. In this application, we know that a change of the revision process happened at some point in the past. The question is then to determine whether or not this change impacted significantly the observed variations of claim amount over time. With $sgn(.)$ the function that gives the sign of a real number taking respectively the values -1, 0 and 1 when this number is negative, null or positive, we consider the random variable $Z = \text{sgn}(C_1 - C_{0.5}) \log(1 + |C_1 - C_{0.5}|)$ that gives the variation of the claim amount between the 6th and the 12th month in log-scale. Past data shows that Z can be modeled by a finite parametric mixture with 12 components. For internal reasons, the insurance company is interested in the 5th component of the mixture, i.e. slight decreases of claim amounts. The change detection is performed on a sample of over 15k observations and a significant change is detected with both tests. The analysis of the results indicates that it seems to occur from the time 1.07. This conclusion allowed the insurance company to investigate further the quantification of the change.

This application shows that the WL and EWL tests can be used in the industry for the monitoring of changes, when they are unexpected but also to assess their significativity when they are known or suspected.

3.3 Discussion

In this last section, we briefly discuss a few issues that we encountered while writing this second part of the thesis. We also suggest possible extensions to improve or extend the results obtained.

We start with a discussion on the choice of the log-likelihood ratio used to detect a change in the first component of the mixture. The log-likelihood ratios of the benchmark and the WL tests have a similar general structure: a ratio of two cases where the numerator splits the sample in two and the denominator considers the whole sample. However, they differ in their way to isolate the effects of the first component, and in the choice of the likelihood function.
The benchmark test defined in (4) distinguishes the parameter of interest a (first component) from the nuisance parameter b (other components). The log-likelihood retained is the one from the whole model: $\log f(X_i,(a,b))$, with $\theta=(a,b)$.
In the WL test and its log-likelihood ratio $\Lambda_{s,n}$ defined in (5), we chose to isolate the first component with the weight functions $w(X_i,\theta)$. This weight is the probability of the observation X_i to belong to the first component, knowing that it has distribution parameter θ. In the log-ratio, we remark that $\log f_1(X_i,\lambda_1)$ is the log-likelihood of the observation X_i, knowing that X_i belongs to the first component. Intuitively, it follows that $w(X_i,\theta)\log f_1(X_i,\lambda_1)$ can be seen as the contribution of the first component in the likelihood of X_i.
During the first stages of my thesis, we studied the properties of the following log-likelihood ratio, an **early version of the test**,

$$\Lambda_{s,n}^{\mathrm{v0}} := \log\left(\frac{\sup_{\lambda_1}\prod_{i=1}^{\lfloor sn\rfloor} f_1(X_i,\lambda_1)^{w(X_i,\hat\theta_{0,s})}\ \sup_{\lambda_1}\prod_{j=\lfloor sn\rfloor+1}^{n} f_1(X_j,\lambda_1)^{w(X_j,\hat\theta_{s,1})}}{\sup_{\lambda_1}\prod_{i=1}^{n} f_1(X_i,\lambda_1)^{w(X_i,\hat\theta)}}\right).$$

In this early version, the logic in the same as in the WL test except that we consider problems of the form $\sup_{\lambda_1}\prod_i f_1(X_i,\lambda_1)^{w(X_i,\hat\theta_\cdot)}$ instead of $\prod_i f_1(X_i,\hat\lambda_1)^{w(X_i,\hat\theta_\cdot)}$ in the WL test, where $\hat\theta_\cdot$ denotes the estimators $\hat\theta_{0,s}$, $\hat\theta_{s,1}$ and $\hat\theta$ of each problem. For the Gaussian case, we showed that the optimization problem of the form $\arg\max_{\lambda_1}\prod_i f_1(X_i,\lambda_1)^{w(X_i,\hat\theta_\cdot)}$ can be solved with an explicit solution denoted by $\hat\lambda_\cdot^*$. We also showed that the solutions $\hat\lambda_\cdot^*$ are consistent estimators of the true parameter $\boldsymbol{\lambda}$. In other words, the early version considers elements of the form $\prod_i f_1(X_i,\hat\lambda_\cdot^*)^{w(X_i,\hat\theta_\cdot)}$. The difference from the WL test lies only in the difference between $\hat\lambda_\cdot^*$ and $\hat\lambda_\cdot$.
While working on the weak limit theorem, we noticed that the estimator $\hat\lambda_\cdot^*$ brings additional complexity that did not seem useful. Thus we retained the version given in the WL test which is easier to implement from the practitioner's point of view. In the end, numerical applications show that the WL and EWL tests perform better than the benchmark test for large samples.

The WLT allows us to detect a change in the first component of a finite parametric mixture. We remark that the detection of a change in other components can be obtained by simply switching the labels in the definition of the mixture. Among its interesting properties, we highlight the run time, especially when one wants to **run a test over more than one components of the mixture**. It is reasonable to assume that we could be interested in detecting a change in a limited number of components.

It is however important not to detect on too many components at the same time. Under the assumption that the components are independent, the probability to detect a change for at least one of them tends to one, under the null hypothesis and as the number of components tends to infinity. For a fixed number of components m, it makes sense to adapt the detection threshold in order to control the probability of false alarm, which means that *a change is detected for at least one component under* H_0.

Given the complexity of the function $\boldsymbol{q}$ defined in (3.33), we did not address directly the question of the distribution of the limit variable $\sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}(W_s - sW_1)}{s(1-s)}$. This is a reasonable choice since numerical simulations are sufficient to compute a detection threshold.

In this work, we focused on the detection of a change in the first component of a finite parametric mixture. The benchmark test would also be a valid candidate if the optimization problem given by the numerator of the log-ratio $\Lambda_{k,n}^{BM}$ defined in (4) could be solved numerically by an adequate algorithm. With this idea, one could consider a customized EM algorithm. It could be designed as follows:

1. Initialization;
2. Perform one iteration of an EM algorithm that optimizes all the components except the first one over the whole sample, while considering the rest of the model fixed;
3. Perform one iteration of an EM algorithm that optimizes the first component over the left part of the sample, while considering the rest of the model fixed;
4. Perform one iteration of an EM algorithm that optimizes the first component over the right part of the sample, while considering the rest of the model fixed;
5. If convergence is not reached, go back to step 2.

In this rough sketch, each step ensures to increase the likelihood. The remaining question is to make sure that steps 2 to 4 can be formally expressed as in an EM algorithm. In the early developments of my thesis, we considered similar variants of the EM algorithm in order to estimate the parameters and the change-point for a finite parametric mixture that contains at most one change. The key is to ensure that each steps guarantees an increase of the likelihood.

From the numerical applications, the WL and EWL tests are valid candidates when looking for a change in one component of a finite parametric mixture. In addition, the results obtained under the null hypothesis in Theorems 4 and 5 allow us to reduce significantly the calibration run time of the detection thresholds: the marginal run time of one simulation is divided by 10 000. Beyond these promising results, the topics discussed in this last section open possibilities for other techniques which are worth being explored.

Chapter 1

Preliminary

In this Chapter, we introduce the key notions and theorems that we rely on in Chapters 2 and 3. Most of the proofs are omitted but exact references are provided.

1.1 Maximum Likelihood Estimation

We start with classical concepts and results on Maximum Likelihood Estimation, mainly based on Lehmann and Casella (1998).

1.1.1 Asymptotic existence and consistency of the MLE

We consider an experiment where we observe a sample of n independent continuous random variables $X = (X_i)_{1\leq i\leq n}$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with values in some set $\mathcal{X}$, subset of an Euclidean space, endowed with Lebesgue's measure. Each X_i, $1 \leq i \leq n$, follows a finite parametric distribution $\mathbb{P}_{\theta^i}$, where θ^i belongs to a set of eligible parameters Θ that we can identify with a subset of $\mathbb{R}^d$, $d \geq 1$. The parameters $\theta^i = (\theta^i_1, \dots, \theta^i_d)$ fully define the distribution. We start by giving some regularity conditions on the distribution and the set Θ.

Conditions 1.1 (Cramér (1946), Lehmann and Casella (1998))**.** *We assume that:*

(a) For $1 \leq i \leq n$, the parametric distribution $\mathbb{P}_{\theta^i}$ admits a parametric density function $f(., \theta^i)$ with respect to the Lebesgue measure.

(b) The distributions $(\mathbb{P}_{\theta^i})_{1\leq i\leq n}$ have common support. In other words, the set $\{x \in \mathcal{X}, f(x, \theta) > 0\}$ does not depend on θ.

(c) Identifiability: the distributions $\mathbb{P}_\theta$ are distinct[1]*.*

The set of possible parameters Θ is usually defined so that regularity conditions are valid. Under conditions 1.1, we can define the likelihood of the sample X.

Definition 1.1. *For a given $\theta \in \Theta$, the **likelihood** of a sample X is the random variable defined by $f(X, \theta) = \prod_{i=1}^n f(X_i, \theta)$ and the **log-likelihood** $L(X, \theta)$ is defined by*

$$L(X, \theta) := \log f(X, \theta) = \sum_{i=1}^{n} \log f(X_i, \theta).$$

With conditions 1.1, the log-likelihood $L(X, \theta)$ is well defined on the common support of the densities $f(., \theta)$: the $X_i(\omega)$, $\omega \in \Omega$ and $1 \leq i \leq n$, can only take values in the set $\{x \in \mathcal{X}, f(x, \theta) > 0\}$ that does not depend on θ.

1. This ensures that the parameter θ can be estimated consistently. See e.g. Csörgő and Horváth (1997) or Lehmann and Casella (1998).

Let us assume that the sample X is independent identically distributed, i.e. the X_i have the same distribution with $\boldsymbol{\theta} := \theta^1 = \cdots = \theta^n$. Then we define the Maximum Likelihood Estimator as follows (Norden (1973), Kotz et al. (1985)).

Definition 1.2. *For the sample X, the **Maximum Likelihood Estimator** (MLE) $\hat{\theta}$ of the true parameter $\boldsymbol{\theta}$ is defined by*

$$\hat{\theta} := \arg\max_{\theta \in \Theta} L(X, \theta)$$

if it exists and is unique.

The existence and uniqueness of the MLE are not obvious (Mäkeläinen et al. (1981), Lehmann and Casella (1998)). Since this is an optimization problem, in practice, one might look at a candidate among the roots of the equations given by the partial derivatives of the log-likelihood. These equations are called the **likelihood equations** and exist only if $L(X, \theta)$ is continuously differentiable in θ. More generally, with the conditions 1.1 and conditions 1.2 that we give below, we can ensure that, even if the MLE does not exist and/or is not unique, we can find some consistent sequence of roots.

Definition 1.3. *A sequence of estimators of $\boldsymbol{\theta}$ is said to be **consistent** if, as $n \to \infty$, it converges in probability to $\boldsymbol{\theta}$. It is said to be **strongly consistent** if it converges almost surely to $\boldsymbol{\theta}$.*

Consistency can be understood as the asymptotic unbiasedness. Some additional conditions involve the Fisher information matrix defined as follows.

Definition 1.4. *The **Fisher information matrix** $I(\theta)$ is defined for $1 \leq j, k \leq d$ and $\theta = (\theta_1, \ldots, \theta_d) \in \Theta$ by*

$$I_{jk}(\theta) := Cov\left(\frac{\partial}{\partial \theta_j} \log f(X_1, \theta), \frac{\partial}{\partial \theta_k} \log f(X_1, \theta)\right).$$

We impose some regularity conditions on the first and second derivatives of the log-likelihood.

Conditions 1.2 (Lehmann and Casella (1998))**.** *We assume that:*

(a) There exists an open subset Θ' of Θ such that $\boldsymbol{\theta} \in \Theta'$ and, for almost all $x \in \mathcal{X}$, the function $\theta \mapsto f(x, \theta)$ is three times continuously differentiable for all $\theta \in \Theta'$.

(b) For all $\theta \in \Theta'$, the Fisher information matrix $I(\theta)$ is positive definite and, for $1 \leq j, k \leq d$, $|I_{jk}(\theta)| < \infty$.

(c) The density function f verifies, for $1 \leq j, k \leq d$ and for all $\theta \in \Theta'$,

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial \theta_j} \log f(X_1, \theta)\right] = 0$$

and the Fisher information matrix $I(\theta)$ verifies

$$I_{jk}(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial \theta_j} \log f(X_1, \theta)\ \frac{\partial}{\partial \theta_k} \log f(X_1, \theta)\right] = \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X_1, \theta)\right].$$

(d) For all $1 \leq j, k, l \leq d$, we can find some function M that does not depend on θ such that, for all $\theta \in \Theta'$ and all $x \in \mathcal{X}$,

$$\left| \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \log f(x, \theta) \right| \leq M(x)$$

with $\mathbb{E}_{\boldsymbol{\theta}}[M(X_1)] < \infty$.

Under these conditions, the following theorem provides an asymptotic result for likelihood based estimators.

Theorem 1.5 (Theorem 5.1 in Lehmann and Casella (1998), Section 6.5)**.** *Under conditions 1.1 and 1.2, consider an independent identically distributed sample X with true parameter $\boldsymbol{\theta} \in \Theta'$. Then, with probability tending to one as $n \to \infty$, there exists at least one consistent sequence $\hat{\theta}_n$ of solutions of the likelihood equations such that $\sqrt{n}(\hat{\theta}_n - \boldsymbol{\theta})$ is asymptotically normal with mean 0 and covariance matrix $I(\boldsymbol{\theta})^{-1}$.*

In Chapter 3, the Fisher information matrix $I(\boldsymbol{\theta})$ under the null hypothesis is denoted by $\boldsymbol{I}$.

In the next section, we provide additional results on consistency and strong consistency of MLE in the case of finite parametric mixtures.

1.1.2 Finite parametric mixtures

The concept of mixture distributions arises in many fields and has been popular in the literature for the last decades, as it allows to describe experiments with different sub-populations (Frühwirth-Schnatter et al. (2019) and the references therein).

Let us consider an experiment where we observe a sample of n independent continuous random variables $X = (X_i)_{1 \leq i \leq n}$. Fix $2 < m < \infty$, deterministic and known.

Definition 1.6. *We say that X_1 follows a **finite parametric mixture** with distribution $\mathbb{P}_\theta$ with m components if conditions 1.1, 1.2(a)-1.2(c) are valid and, for $f_1, \ldots, f_m$ some fixed density functions on $\mathcal{X}$, the distribution $\mathbb{P}_\theta$, admits the density*

$$f(x, \theta) := \sum_{k=1}^{m} p_k f_k(x, \lambda_k), \ x \in \mathcal{X},$$

where $(p_1, \ldots, p_{m-1})$ belongs to the open set

$$\Theta_0 := \left\{ (p_1, \ldots, p_{m-1}) \in (0,1)^{m-1}, \sum_{k=1}^{m-1} p_k < 1 \right\},$$

with $p_m := 1 - \sum_{k=1}^{m-1} p_k$ and $\theta = (p_1, \ldots, p_{m-1}, \lambda_1, \ldots, \lambda_m)$ belongs to the set of eligible parameters $\Theta = \Theta_0 \times \prod_{k=1}^{m} \Theta_k$.

Remark that the identifiability of the mixture (see conditions 1.1(c)) is not obvious in general. For example, the mixtures with two components and respective parameters $(p_1, p_2, \lambda_1, \lambda_2)$ and $(p_2, p_1, \lambda_2, \lambda_1)$ have same distribution (McLachlan and Peel (2000)).

For finite samples of Gaussian mixtures, the Maximum Likelihood Estimator does not exist since its likelihood is unbounded (Day (1969)). However the conditions of Theorem 1.5 still hold [2], allowing to find a consistent sequence of roots of the likelihood equations. We finish this section by a result from Hathaway (1985) that provides the strong consistency of the MLE for an **univariate Gaussian mixture** under some conditions on the parameter set. We consider X an independent identically distributed closed sample of n real-valued random variables with a Gaussian mixture distribution with m components of true parameter $\boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{m-1}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)$ in the set of possible parameters $\Theta \subset \Theta_0 \times (\mathbb{R} \times \mathbb{R}_+^*)^m$. We assume that Θ is defined such that the mixture is identifiable. For $b \in (0, 1]$, define the subset $\Theta_b \subset \Theta$ by

$$\Theta_b := \left\{ \theta \in \Theta, \min_{1 \leq k, k' \leq m} \frac{\sigma_k}{\sigma_{k'}} \geq b \right\}.$$

Theorem 1.7 (Theorem 3.3 in Hathaway (1985))**.** *Let $b \in (0, 1]$ such that the true parameter $\boldsymbol{\theta}$ belongs to Θ_b. Then the MLE $\hat{\theta}_n$ that maximizes the log-likelihood of the sample X over Θ_b exists and is strongly consistent.*

1.2 Weak convergence of càd-làg random functions on $[0, 1]$

This section is mainly based on Billingsley (1999) and van der Vaart (1998). We give some results on the weak convergence of sequences of càd-làg processes. For that purpose we introduce briefly the topologies of the metric spaces $\mathbb{C}$ and $\mathbb{D}$ before stating standard limit theorems.

1.2.1 The spaces $\mathbb{C}$ and $\mathbb{D}$ of random functions

We start by a general definition of a random element in a metric space $\mathbb{S}$. We denote by $\mathcal{S}$ the Borel σ-algebra of $\mathbb{S}$.

Definition 1.8. *A **random element** on $(\mathbb{S}, \mathcal{S})$ is a map from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathbb{S}$, which is $(\mathcal{S}, \mathcal{F})$-measurable. If $\mathbb{S}$ is a metric space of functions, then X is called a **random function**, random process or stochastic process.*

We consider two particular metric spaces of functions. First we denote by $\mathbb{C}(E, F)$ the **space of real-valued continuous functions** on $E \subseteq [0, 1]$ with values in a normed Euclidean space F, with its associated Borel σ-field and the **uniform norm** $\|.\|_{\mathbb{C}(E,F)}$ defined for a function f in $\mathbb{C}(E, F)$ by

$$\|f\|_{\mathbb{C}(E,F)} := \sup_{t \in E} \|f(t)\|_F, \tag{1.1}$$

where $\|.\|_F$ is the norm on F. In $\mathbb{C}(E, F)$, we say that a sequence of functions f_n converges to a function f if $\|f_n - f\|_{\mathbb{C}(E,F)} \to 0$ as $n \to \infty$.

Among well-known random functions with values in $\mathbb{C}([0, 1], \mathbb{R})$ we can already introduce the **standard Brownian motion** or Wiener process [3], denoted by $(W_t)_{t \in [0,1]}$ and satisfying that $W_0 = 0$ and the following three properties:

2. See e.g. Example 6.10 in Lehmann and Casella (1998).
3. See e.g. Section 8.1 in Durrett (2010).

- ◇ For any finite $k \geq 2$ and any $0 \leq t_0 < \cdots < t_k \leq 1$, the random variables $W_{t_0}, W_{t_1} - W_{t_0}, \ldots, W_{t_k} - W_{t_{k-1}}$ are independent;
- ◇ For any $0 \leq t < s \leq 1$ the distribution of the random variable $W_s - W_t$ is Gaussian with mean zero and variance $s - t$;
- ◇ The application $t \mapsto W_t$ is continuous with probability 1.

The space of càd-làg functions, defined on E with values in F, is denoted by $\mathbb{D}(E,F)$ and referred as the **Skorokhod metric space** with the Skorokhod metric $d_{\mathbb{D}(E,F)}(.,.)$ defined for f_1 and f_2 in $\mathbb{D}(E,F)$ by

$$d_{\mathbb{D}(E,F)}(f_1,f_2) := \inf_{\tau\in\Gamma_E} \max\left\{\sup_{t\in E}|\tau(t)-t|, \sup_{t\in E}\|f_1(t)-f_2\circ\tau(t)\|_F\right\}.$$

Here Γ_E is the set of continuous and strictly increasing bijections from E to itself. In $\mathbb{D}(E,F)$, a sequence of functions f_n converges to a function f if $d_{\mathbb{D}(E,F)}(f_n,f) \to 0$ or, equivalently, if we can find some sequence $\tau_n \in \Gamma_E$ such that $\sup_{t\in E}|\tau_n(t)-t| \to 0$ and $\sup_{t\in E}\|f(t)-f_n\circ\tau_n(t)\|_F \to 0$.

$\mathbb{C}(E,F)$ is a subspace of $\mathbb{D}(E,F)$ and the Skorokhod metric coincides with the uniform norm on $\mathbb{C}(E,F)$. We refer to Billingsley (1999) for a detailed construction of the topologies of the spaces $\mathbb{C}(E,F)$ and $\mathbb{D}(E,F)$.

Definition 1.9. *A sequence X_n of random functions with values in $\mathbb{C}(E,F)$, resp. $\mathbb{D}(E,F)$, is said to converge **in distribution** or **weakly** to X if $\mathbb{E}[\varphi(X_n)]$ converges to $\mathbb{E}[\varphi(X)]$ for every bounded continuous function φ from $\mathbb{C}(E,F)$, resp. $\mathbb{D}(E,F)$, to $\mathbb{R}$. We write $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$.*
*X_n is said to converge **in probability** to X in $\mathbb{C}(E,F)$, resp. $\mathbb{D}(E,F)$, if, for any $\epsilon > 0$, $\mathbb{P}[\|X_n - X\|_{\mathbb{C}(E,F)} > \epsilon] \to 0$, resp. $\mathbb{P}[d_{\mathbb{D}(E,F)}(X_n,X) > \epsilon] \to 0$. We write $X_n \xrightarrow[n\to\infty]{P} X$.*
*X_n is said to converge **almost surely** to X if, for almost every $\omega \in \Omega$, $X_n(\omega)$ converges to $X(\omega)$ in the sense of the convergence in $\mathbb{C}(E,F)$, resp. $\mathbb{D}(E,F)$.*

The convergence in distribution is often associated to the Portmanteau Theorem [4] which gives equivalent definitions.

1.2.2 Limit theorems

In this section, we provide some limit theorems that we encounter in Chapter 3, mainly for proving the weak convergence of càd-làg processes. We start with **Donsker's theorem** in the Skorokhod metric space $\mathbb{D}([0,1],\mathbb{R})$.

Theorem 1.10 (Donsker's Theorem in $\mathbb{D}([0,1],\mathbb{R})$, Theorem 14.1 in Billingsley (1999))**.** *If $X_1, X_2, \ldots, X_n$ are independent identically distributed random variables with mean zero and variance $\sigma^2 > 0$, then, the random functions $t \in [0,1] \mapsto Y_t^n := \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor} X_i$ defined in the Skorokhod space $\mathbb{D}([0,1],\mathbb{R})$ are such that the process $(Y_t^n)_{t\in[0,1]}$ converges weakly to a standard Brownian motion as $n \to \infty$.*

Further we introduce below the **Continuous Mapping Theorem** for random elements in metric spaces.

4. See e.g. Lemma 18.9 in van der Vaart (1998).

We assume in the following that the metric space $\mathbb{S}$ is separable, which is the case for the spaces $\mathbb{C}(E,F)$ and $\mathbb{D}(E,F)$ with the topologies defined above (Billingsley (1999)). Moreover we denote by $\mathbb{S}\times\mathbb{S}$ the separable product space with metric

$$d_{\mathbb{S}\times\mathbb{S}} : (X,Y) \mapsto \max\{d_{\mathbb{S}}(X), d_{\mathbb{S}}(Y)\}.$$

For more details on the construction of the product space, we refer to Section 2 in Billingsley (1999).

Theorem 1.11 (Continuous Mapping Theorem, Theorem 18.11 in van der Vaart (1998))**.** *Consider a sequence X_n of random elements with values in subsets $\mathbb{S}_n$ of a separable metric space $\mathbb{S}$ and a random element X with values in $\mathbb{S}_0$. Consider also a sequence of maps f_n from $\mathbb{S}_n$ to another metric space F such that, for every sequence $x_n \in \mathbb{S}_n$ for which we can find a convergent subsequence $x_{n'}$ with limit $x \in \mathbb{S}_0$, $f_{n'}(x_{n'})$ converges to $f(x)$. It holds that:*

- ◇ *If $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$, then $f_n(X_n) \xrightarrow[n\to\infty]{\mathcal{D}} f(X)$,*
- ◇ *If $X_n \xrightarrow[n\to\infty]{P} X$, then $f_n(X_n) \xrightarrow[n\to\infty]{P} f(X)$,*
- ◇ *If $X_n \xrightarrow[n\to\infty]{a.s.} X$, then $f_n(X_n) \xrightarrow[n\to\infty]{a.s.} f(X)$.*

Based on the Continuous Mapping Theorem, **Slutsky's Theorem** is a standard result for the convergence of random variables since it allows to conclude on the convergence of sums and products of random variables when one variable converges in distribution and the other one in probability (Slutsky (1925)). We give here its version for random elements.

Theorem 1.12 (Theorem 3.1 in Billingsley (1999))**.** *Assume that $\mathbb{S}$ is a separable metric space with metric $d_{\mathbb{S}}$ and $(X_n, Y_n)_{n\geq 1}$ is a sequence of random elements of $\mathbb{S}\times\mathbb{S}$. If X_n converges weakly to X and $d_{\mathbb{S}}(X_n, Y_n)$ converges in probability to zero, then Y_n converges weakly to X.*

In addition to this result, van der Vaart (1998) provides a panel of convergence properties. We give one of them in the result below as an extended functional version of Slutsky's Theorem.

Theorem 1.13 (Extended Slutsky's Theorem, Theorem 18.10 in van der Vaart (1998))**.** *Let X_n and Y_n be two sequences of random elements, and X and Y two random elements, all with values in the separable metric space $\mathbb{S}$ with metric $d_{\mathbb{S}}$. Then, if X_n converges weakly to X and Y_n converges in probability to a constant c in $\mathbb{S}$, then (X_n, Y_n) converges weakly to (X, c).*

Together with the Continuous Mapping Theorem, this result provides a functional Slutsky's Theorem. For example, let us consider two sequences (X_t^n) and (Y_t^n) of continuous real-valued random functions of t such that, as $n \to \infty$, the sequence (X_t^n) converges weakly to the random function (X_t) and (Y_t^n) converges in probability to a constant function (c), $c \in \mathbb{R}$. Since the application $((x_t),(y_t)) \mapsto (x_t\, y_t)$ is continuous in the sense of the conditions of the Continuous Mapping Theorem, it follows that $(X_t^n Y_t^n)$ converges weakly to (cX_t).

The applications based on the Continuous Mapping Theorem are numerous. Among them, we introduce the functional **Delta Method**, a technique that approximates

asymptotically the distribution of $\Phi(X)$ where X is a random function in a normed space (Doob (1935), van der Vaart (1998)). For that purpose we introduce the notion of Hadamard differentiability for maps between two normed spaces.

Definition 1.14. *For D and E two normed spaces, a map Φ from $D_\Phi \subseteq D$ to E is **Hadamard differentiable** in $\theta \in D_\Phi$ if we can find a continuous linear map Φ'_θ from D to E such that, as $t \to 0$,*

$$\left\| \frac{\Phi(\theta + t\zeta_t) - \Phi(\theta)}{t} - \Phi'_\theta(\zeta) \right\|_E \to 0$$

*for any application $t \mapsto \zeta_t$ such that $\zeta_t \to \zeta$ and that, for all small t, $\theta + t\zeta_t$ belongs to D_Φ. The map $\Phi'_\theta(.)$ is the **differential** of Φ at θ.*

Theorem 1.15 (Functional Delta Method in normed spaces, Theorem 20.8 in van der Vaart (1998))**.** *For D and E two normed linear spaces, consider a map Φ from a subset D_Φ of D to E, that is Hadamard differentiable at $\theta \in D_\Phi$ with differential denoted by $\Phi'_\theta(.)$. Consider also a sequence of random maps X_n with values in D_Φ and a sequence of numbers a_n which tends to infinity as $n \to \infty$. If, as $n \to \infty$, the sequence $a_n(X_n - \theta)$ converges weakly to some random map X, then the sequence $a_n(\Phi(X_n) - \Phi(\theta))$ converges weakly to the random map $\Phi'_\theta(X)$.*

Proof. The proof is a direct application of the Continuous Mapping Theorem with $f_n(\zeta) := a_n(\Phi(\theta + a_n^{-1}\zeta) - \Phi(\theta))$ where the maps f_n are defined on

$$\mathbb{S}_n := \{\zeta : \theta + a_n^{-1}\zeta \in D_\Phi\}.$$

The Hadamard differentiability ensures that the conditions of the Continuous Mapping Theorem hold. □

The space $\mathbb{C}([0,1], \mathbb{R}^d)$ is a linear normed space with the norm $\|.\|_{\mathbb{C}([0,1],\mathbb{R}^d)}$ defined in (1.1) and therefore falls in the scope of Theorem 1.15. However the Skorokhod metric space $\mathbb{D}([0,1], \mathbb{R}^d)$ is not a normed space. We give here a Corollary of the result from van der Vaart (1998) for càd-làg processes.

Corollary 1.16 (Functional Delta Method in the Skorokhod metric space)**.** *For $0 < d_1, d_2 < \infty$, consider a map $\Phi : D_\Phi \subseteq \mathbb{D}([0,1], \mathbb{R}^{d_1}) \to \mathbb{D}([0,1], \mathbb{R}^{d_2})$. Consider also a sequence of random maps X_n with values in D_Φ and a sequence of numbers a_n which tends to infinity as $n \to \infty$. If, as $n \to \infty$,*

- ◇ *the sequence $a_n(X_n - \theta)$ converges weakly to some random map X,*
- ◇ *we can find some linear map $\Phi'_\theta(.)$ from $\mathbb{D}([0,1], \mathbb{R}^{d_1})$ to $\mathbb{D}([0,1], \mathbb{R}^{d_2})$ such that for every sequence $\zeta_n \in \{z : \theta + a_n^{-1}z \in D_\Phi\}$ for which we can find a subsequence $\zeta_{n'}$ that converges in $\mathbb{D}([0,1], \mathbb{R}^{d_1})$ to ζ, the sequence $a_{n'}(\Phi(\theta + a_{n'}^{-1}\zeta_{n'}) - \Phi(\theta))$ converges in $\mathbb{D}([0,1], \mathbb{R}^{d_2})$ to $\Phi'_\theta(\zeta)$,*

then the sequence $a_n(\Phi(X_n) - \Phi(\theta))$ converges weakly to the random map $\Phi'_\theta(X)$.

Proof. As for Theorem 1.15, the proof is an application of the Continuous Mapping Theorem where the Hadamard differentiability is replaced by the second condition. □

In Chapter 3, this result is applied to a sequence of random d-by-d square matrices and a map that inverts the matrices (t by t). For that purpose, we introduce the following result that provides the differential of the application that inverts a square matrix.

For $0 < d < \infty$ fixed, let us consider the space $gl_d(\mathbb{R})$ of real-valued d-by-d square matrices. We consider the entrywise 2-norm $\|.\|_2$, also called the **Frobenius norm**, that is defined for $A = (a_{ij})_{1\leq i,j\leq d} \in gl_d(\mathbb{R})$ by

$$\|A\|_2^2 = \sum_{1\leq i,j\leq d} a_{ij}^2.$$

With the identity matrix and the entrywise 2-norm, the normed space $(gl_d(\mathbb{R}), \|.\|_2)$ is an unital Banach algebra. The open sub-space of **invertible matrices** in $gl_d(\mathbb{R})$ is denoted by $GL_d(\mathbb{R})$.

Theorem 1.17 (Lemma 2.5.5 in Abraham et al. (1988))**.** *The application from $GL_d(\mathbb{R})$ to $GL_d(\mathbb{R})$ is infinitely differentiable and its first differential at point $A \in GL_d(\mathbb{R})$ is the application defined for $H \in gl_d(\mathbb{R})$ by $H \mapsto -A^{-1}HA^{-1}$.*

This result is obtained from an expansion of the inverse of a matrix as a Neumann series, as given in the following theorem.

Theorem 1.18 (Theorem 4.16 in Dudley and Norvaiša (2011))**.** *Consider some invertible matrix $A \in GL_d(\mathbb{R})$. If $H \in gl_d(\mathbb{R})$ is such that $\|H\|_2 < 1/\|A^{-1}\|_2$, then $A+H$ is invertible and*

$$(A+H)^{-1} = \sum_{k=0}^{\infty} (-1)^k (A^{-1}H)^k A^{-1}.$$

This closes the topic on weak convergence of càd-làg processes in the Skorokhod metric space.

1.3 Change-point detection techniques

We consider a sequence $(X_n)_{n\geq 1}$ of independent real-valued random variables. Each random variable X_n follows a distribution $\mathbb{P}_{\theta^n}$, where θ^n belongs to the set Θ of eligible parameters. The sequence is associated with a known initial distribution $\mathbb{P}_{\boldsymbol{\theta}}$ and a change occurs at time ν, called *change-point*. It means that the sample is identically distributed before ν, i.e. $\boldsymbol{\theta} = \theta^1 = \cdots = \theta^\nu$; and, after the change-point, $\theta^n \neq \boldsymbol{\theta}$ for $n > \nu$ (possibly but not necessarily i.i.d.). In this section, for $\theta \in \Theta$, the distribution $\mathbb{P}_\theta$ is assumed to admit a density function f_θ with respect to some measure (often the Lebesgue measure for the continuous case, and the counting measure on $\mathbb{N}$ for the discrete case).

Detection procedures aim to determine *if and when* the initial distribution changes. In this section, we start by introducing a standard on-line detection scheme called the *Shiryaev-Roberts procedure*. Then, we provide a standard off-line test for the detection of at most one change in a closed sample.

1.3.1 The Shiryaev-Roberts detection scheme

With the notations above, consider an experiment where we observe, as they arrive, outcomes of the random sequence $(X_n)_{n\geq 1}$. We assume that only one change occurs and the distribution after the change-point ν is denoted by $\mathbb{P}_{\boldsymbol{\theta}'}$.

Definition 1.19. *A **change-point detection scheme** is a procedure defined by a random sequence S_n called the **detection sequence**, and a **threshold** s^*. The procedure states that an alarm is raised as soon as $S_n > s^*$, defining a stopping time $n^* := \inf\{n \geq 1, S_n > s^*\}$.*

Given a delay criterion, *quickest change-point detection schemes* aim to detect a change in the distribution *as quickly as possible*, without raising too many false alarms (detection when no change occurred).

Definition 1.20 (Shiryaev (1961), Roberts (1966))**.** *The **Shiryaev-Roberts procedure** is a change-point detection scheme with a detection sequence given by*

$$\begin{aligned} &S_0 := 0, \\ &S_n(\boldsymbol{\theta}, \boldsymbol{\theta}') := \sum_{i=1}^{n} \prod_{k=i}^{n} \frac{f_{\boldsymbol{\theta}'}(X_k)}{f_{\boldsymbol{\theta}}(X_k)} = (1 + S_{n-1}(\boldsymbol{\theta}, \boldsymbol{\theta}')) \, \frac{f_{\boldsymbol{\theta}'}(X_n)}{f_{\boldsymbol{\theta}}(X_n)}, \ n \geq 1. \end{aligned}$$

Remark that, as long as $n \leq \nu$, the statistic $S_n(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is close to 0. After the change, it quickly takes very high values. One advantage of this procedure is the recursive form of the statistic $S_n(\boldsymbol{\theta}, \boldsymbol{\theta}')$. It can be easily computed with its previous value $S_{n-1}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and the new outcome x_n. Another advantage is its optimality in minimizing the average detection delay, as given in the following result.

Theorem 1.21 (Pollak and Tartakovsky (2009))**.** *Fix $\eta > 1$. If the change-point ν is a random variable with a uniform improper prior distribution, then, among all change-point detection procedures such that $\mathbb{E}_{\{\nu=\infty\}}[n^*] \geq \eta$, the Shiryaev-Roberts detection scheme is optimal in minimizing the average detection delay $\frac{\sum_{k=0}^{\infty} \mathbb{E}_{\{\nu=k\}}\left[(n^*-k)^+\right]}{\mathbb{E}_{\{\nu=\infty\}}[n^*]}$.*

Optimality is also valid for other frameworks (see e.g. Shiryaev (1963), Feinberg and Shiryaev (2006), Polunchenko and Tartakovsky (2010)).

1.3.2 A standard hypothesis test for the detection of at most one change in a closed sample

Let us consider an experiment where we observe the outcomes of a closed sample $(X_i)_{1\leq i\leq n}$ of independent real-valued random variables that contains *at most one change* (AMOC). Let us assume that each parameter $\theta = (a, b) \in \Theta$ is defined by two sub-parameters a and b. We introduce a standard likelihood-based hypothesis test that aims to detect if a change occurs in the first sub-parameter a (see e.g. Section 1.1 in Csörgő and Horváth (1997)). Here b is called a *nuisance* parameter. We test

- ◇ the null hypothesis where no change happens, i.e. $\theta^1 = \cdots = \theta^n$,
- ◇ against the alternative hypothesis where at most one change occurs, i.e. there exists some $1 < k \leq n$ such that $a^1 = \cdots = a^k \neq a^{k+1} = \cdots = a^n$ and $b^1 = \cdots = b^n$.

For that purpose, we consider the log-likelihood ratio associated with the test and defined by

$$\Lambda_{k,n} := \log\left(\frac{\sup_{(a,b),(a',b)\in\Theta}\prod_{i=1}^{k} f(X_i,(a,b))\prod_{i=k+1}^{n} f(X_i,(a',b))}{\sup_{(a,b)\in\Theta}\prod_{i=1}^{n} f(X_i,(a,b))}\right).$$

The test statistic is defined by $\max_{1\leq k\leq n} 2\Lambda_{k,n}$. Under the null hypothesis, its limit distribution takes the shape of a supremum over a function of a Brownian motion: see e.g. Corollary 1.1.1, Theorems 1.3.1. and 1.3.2 in Csörgő and Horváth (1997). Such a result can be found in other similar work for different frameworks such as in Davis et al. (1995) or Dehling et al. (2014).

This ends the introduction of the tools we encounter in this thesis.

Chapter 2

Discrete Poisson case: a sequential estimator of the post-change parameter

Exploring the longevity risk using statistical tools derived from the Shiryaev-Roberts procedure [1]

European Actuarial Journal (2018),
https://doi.org/10.1007/s13385-018-0168-4
Joint work with Marine Habart [2]*, Catherine Rainer* [3] *and Aliou Sow* [4]

2.1 Introduction and general framework

2.1.1 Introduction

Within the usual techniques to monitor mortality and longevity risks [5], statistical sequential tools have been used only recently in the actuarial field [6] even if applications of the change-point theory appeared a little bit earlier for other topics [7]. In this chapter, we focus on the sequential procedure developed by Shiryaev (1963) and Roberts (1966) within the discrete time framework of Polunchenko and Tartakovsky (2010) and provide two adaptive procedures.

We start by the introduction of the mathematical framework with a specific Poisson model, designed here for the study of the mortality with a similar approach as in Rhodes and Freitas (2004) and Tomas and Planchet (2015). In Sections 2.2 and 2.3, we provide two adaptive procedures. For both procedures, the time of change is assumed to be deterministic but unknown; and the mortality is given by independent Poisson random variables. Distributions before the change are assumed to be known. In the first procedure, we assume that the mortality is given by a sequence of observations from independent identically distributed Poisson random variables with constant intensity λ before the change. After the change-point ν, the intensity is the only parameter that

1. A. et al. (2018)

2. Actuary from Institut des Actuaires (Actuaire Agrégé), Ph.D, Telecom Bretagne & EURIA, marine.habart@gmail.com.

3. Maître de conférences, Ph.D, Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Occidentale, catherine.rainer@univ-brest.fr.

4. Actuary from Institut des Actuaires (Actuaire Certifié), Telecom Bretagne & ISFA, aliou_s@yahoo.fr.

5. See e.g. De Jong and Boyle (1983), Olivieri et al. (2002) or and Planchet and Tomas (2014), Section 3.2..

6. See e.g. Gandy et al. (2005), El Karoui et al. (2017), Croix et al. (2015), Mouyopa Djitta (2015) and suggested in Tomas and Planchet (2014) as a possible extension.

7. See e.g. Matthews et al. (1985), Servier (2010), Oueslati and Lopez (2013).

changes and this change is assumed to be unknown but deterministic. This means that we look for a sudden but persistent change of level. Because the original tool requires to know the post-change distribution, we suggest a consistent estimator designed for the specific case of the Shiryaev-Roberts detection procedure. A detailed proof of its consistency is provided and simulations assess that it has lower bias and lower variance just after the change-point. In the second procedure, the mortality is assumed to decrease at a steady pace: the intensity of the Poisson random variables is decreasing with a constant rate. After the change, the rate is different, deterministic but unknown. This means that, for an application to mortality, we look for a change in the trend of the log mortality rate. Here, the change-point and the trend coefficient are estimated by usual MLE. Eventually we show that these innovative approaches are practical tools in order to explore mortality data, especially when nothing is known about the post-change distributions.

An important part of this work is devoted in Section 2.4 to the application of our methodology on real data, in a context where the change is obvious, using specific methodologies to adjust the data as in Mei et al. (2011). For the mortality risk, we focus on the 2003 French heatwave and the 1918 Spanish Flu. For the longevity risk, i.e. with the second procedure, we look at the 2003 heatwave from a different perspective; and we study the French female mortality in the 1960's where a clear change of trend is identifiable [8]. In addition, for both risks, we analyze a real insurance portfolio where no specific information might help us to understand the change, and where the change itself does not seem perceptible (level and trend analysis). The main results allow us to identify the change-points of the mortality when they happen and to quantify the minimum lag before clear identification of the phenomena. These examples illustrate the main properties of the model in the case of actuarial applications.

Variants of the suggested approaches are also widely expressed in Section 2.5 and the suggested estimators are compared to benchmark methodologies in Appendix 2.7.1.

2.1.2 Mathematical framework

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $x = (x_n)_{n\geq 1}$ a sequence of outcomes of a sequence of independent random variables $X = (X_n)_{n\geq 1}$ such that

$$\begin{aligned} X_n &\sim \text{Poisson}(\lambda), \quad 1 \leq n \leq \nu, \\ &\sim \text{Poisson}(\lambda\boldsymbol{\rho}), \quad \nu + 1 \leq n, \end{aligned}$$

where $\lambda > 0$ deterministic and known, and $\boldsymbol{\rho} > 0$ and $\nu \in \mathbb{N} \cup \{\infty\}$ are deterministic but unknown. We denote by ν the time step of the change and by $\boldsymbol{\rho}$ the ratio of the intensities after and before the change. The **Shiryaev-Roberts random sequence** $S(\boldsymbol{\rho}) = (S_n(\boldsymbol{\rho}))_{n\geq 1}$ is introduced by both Shiryaev (1963) and Roberts (1966). Using the sequential framework of Polunchenko and Tartakovsky (2010) it is defined for all $n \geq 1$ by

$$S_n(\boldsymbol{\rho}) = \sum_{i=1}^{n} \prod_{k=i}^{n} \frac{f_{\boldsymbol{\rho}}(X_k)}{f_1(X_k)} = (1 + S_{n-1}(\boldsymbol{\rho})) \frac{f_{\boldsymbol{\rho}}(X_n)}{f_1(X_n)}$$

8. This change is a consequence of the revolution in cardiovascular care, see Vallin and Meslé (2010), Didou (2011) and Cutler and Meara (2001).

with $S_0(\boldsymbol{\rho}) := 0$ and f_ρ the density function of a Poisson random variable with intensity $\lambda\rho$, for any $\rho > 0$:

$$f_\rho(x) = e^{-\lambda\rho}\frac{(\lambda\rho)^x}{x!},\ x \in \mathbb{N}. \tag{2.1}$$

Here we use the term density function with respect to the counting measure on $\mathbb{N}$.

In the original work of Shiryaev (1963) and Roberts (1966), both parameters λ and $\boldsymbol{\rho}$ are known. The Shiryaev-Roberts procedure states that $S_n(\boldsymbol{\rho})$ is computed as long as it does not exceed some threshold s_n^*. The **detection time step** $T^*(s_n^*)$ is the time step of the first observation that triggers the detection, and is defined by

$$T^*(s_n^*) = \inf\{n \geq 1,\ S_n(\boldsymbol{\rho}) > s_n^*\}$$

with $\inf \emptyset = \infty$. The **threshold sequence** s_n^* is usually chosen with respect to a **false alarm constraint**. This constraint can be determined by the **probability of false alarm** α and its corresponding threshold s_n^* by

$$\mathbb{P}\left[\max_{1\leq i\leq n} \boldsymbol{S}_i(\boldsymbol{\rho}) \geq s_n^*\right] = \alpha. \tag{2.2}$$

where $(\boldsymbol{S}_i(\boldsymbol{\rho}))_{1\leq i\leq n}$ is the Shiryaev-Roberts sequence when no change occurs (i.e. the procedure applied to some sequence $\boldsymbol{X}$ for which no change occurs). Here α is a parameter that needs to be defined by the user of the procedure.

This procedure is an accurate tool for measuring the deviation from the initial intensity, as shown by its optimality properties already known in the field of change-point detection. For example, Pollak and Tartakovsky (2009) proves its optimality for minimizing the cumulative average delay to detection [9] $\sum_{k=0}^{\infty} E_k[(T^*(s^*) - k)^+]/E_\infty[T^*(s^*)]$ for sequential observations, where the change-point ν is unknown and where the threshold sequence is reduced to a fixed constant s^*. However, it requires to know the post-change distribution. In practice, the shift parameter ρ is unknown and we would like to know simultaneously if there is a change and what would be the shift associated with this change. In this case the application of detection procedures requires the use of adaptive procedures. Wu (2015) gives an overview of existing methods for different detection procedures and Pollak (2009) provides a common adaptive procedure for the Shiryaev-Roberts approach.

More generally, for exponential families, Foster and George (1993) suggest estimators for the mean before the change, and Wu (2005) provides estimators for the post-change mean in the case of the CUSUM chart. More recently, Fotopoulos et al. (2010) studied the asymptotic distributions of the MLE under the Gaussian framework, and estimators designed for specific frameworks are provided e.g. in Wu (2016a), Wu (2016b) and Frick et al. (2014). Extensive asymptotic results for the change-point detection can be found in Csörgő and Horváth (1997).

In this work, we provide two adaptive procedures. In the first one, we assume that the current intensity is stable and we look for a sudden but persistent change of level. In the second model, the intensity evolves at a steady pace, and we look for a change of the intensity constant increase/decrease rate.

9. Where E_k is the expectancy when $\nu = k$.

2.2 Adaptive procedure for detecting a change of level

Under the general framework given in Section 2.1, for given $\lambda > 0$, $n \geq 0$ known, and $\nu \geq 0$ deterministic but unknown, we suggest the estimator $\hat{\rho}_S(n)$ for the intensity shift parameter $\boldsymbol{\rho}$, defined as

$$\hat{\rho}_S(n) := \arg\max_{\rho>0} S_n(\rho),$$

where $(S_n(\rho))$ is the Shiryaev-Roberts sequence:

$$S_n(\rho) := \sum_{i=1}^{n} \prod_{k=i}^{n} \frac{f_\rho(X_k)}{f_1(X_k)} \tag{2.3}$$

and f_ρ the density function of a Poisson random variable with intensity $\lambda\rho$ from (2.1).

Remark that, as for the Shiryaev-Roberts sequence, the estimator $\hat{\rho}_S(n)$ does not depend on the change-point ν. Therefore, we do not need to know when the change occurs in order to infer the post change intensity. The corresponding adaptive procedure uses the detection sequence $\tilde{S}_n = (\tilde{S}_{i,n})_{1\leq i\leq n}$ defined by

$$\tilde{S}_{i,n} = \sum_{k=1}^{i} \prod_{j=k}^{i} \frac{f_{\hat{\rho}_S(n)}(X_j)}{f_1(X_j)}. \tag{2.4}$$

$\tilde{S}_{i,n}$ is computed as long as $\tilde{S}_{i,n} \leq \tilde{s}_n^*$, where the threshold $\tilde{s}_n^*$ is computed as in Equation (2.2). This implies that for each new observation (i.e. n increasing), the sequence $\tilde{S}_n$ is re-evaluated. Thus, as soon as the alarm is raised, the procedure provides an estimate of the post change intensity.

In the two following paragraphs, we show that $\hat{\rho}_S(n)$ is consistent. We also assess in Appendix 2.7.1 that for the studied cases the average bias of $\hat{\rho}_S(n)$ decreases faster than the one of the usual MLE, the variance of $\hat{\rho}_S(n)$ is always lower than the one of the MLE and both variances are asymptotically similar.

2.2.1 Steps of the adaptive procedure for detecting a change of level

In summary, the adaptive procedure is given in the following 8 steps:

1. Set the time step to $n = 1$;
2. Estimate the coefficient of change $\hat{\rho}_S(1)$;
3. Compute the Shiryaev-Roberts sequence $\left(\tilde{S}_{i,1}(\hat{\rho}_S(1))\right)_{i=1}$;
4. Compute the threshold $\tilde{s}_1^*$;
5. If the Shiryaev-Roberts sequence overcomes the threshold, raise an alarm. Otherwise, increment the time step n;
6. Estimate the coefficient of change $\hat{\rho}_S(n)$;
7. Compute the Shiryaev-Roberts sequence $\left(\tilde{S}_{i,n}(\hat{\rho}_S(n))\right)_{1\leq i\leq n}$;
8. Go back to step 4;

$\boldsymbol{\rho}$ estimation is provided by $\hat{\rho}_S(n)$ when stopping. For practical considerations on the threshold computation, see Sections 2.1.2 and 2.4.

2.2.2 Consistency of the sequence $\hat{\rho}_S(n)$

Definition 2.1. *An estimator $\hat{\rho}(n)$ of $\boldsymbol{\rho}$ is said to be **consistent** iif $\hat{\rho}(n) \xrightarrow[n\to\infty]{P} \boldsymbol{\rho}$.*

For all $1 \leq i \leq n$, we set

$$\hat{\rho}_{i,n} := \frac{1}{\lambda(n-i+1)} \sum_{k=i}^{n} X_k$$

and introduce the likelihood ratio for the sequence $X_i, X_{i+1}, \ldots, X_n$

$$\begin{aligned} R_{i,n}(\rho) := \prod_{k=i}^{n} \frac{f_\rho(X_k)}{f_1(X_k)} &= e^{-\lambda(n-i+1)(\rho-1)+\log(\rho)\sum_{k=i}^{n} X_k} \\ &= e^{\lambda(n-i+1)(\hat{\rho}_{i,n}\log(\rho)-\rho+1)}. \end{aligned}$$

Remark that, for any $i \geq 1$, $\arg\max_{\rho>0} R_{i,n}(\rho) = \arg\max_{\rho>0} \log\left(R_{i,n}(\rho)\right)$.
Since $\frac{d}{d\rho}\log\left(R_{i,n}(\rho)\right) = 0$ for $\rho = \frac{1}{\lambda(n-i+1)}\sum_{k=i}^{n} X_k$, it follows that the maximum of $\rho \mapsto R_{i,n}(\rho)$ is attained at $\hat{\rho}_{i,n}$.

Remark also that the sequence $(X_k)_{k\geq 1}$ differs from an i.i.d sequence only through a finite number of items. Thus the law of large numbers holds: for all $i \geq 1$, $\mathbb{P}$-a.s.,

$$\lim_{n\to\infty} \hat{\rho}_{i,n} = \boldsymbol{\rho}. \tag{2.5}$$

Remark 2.2. *The sequence $(\hat{\rho}_{i,n})_{n\geq 1}$ does not converges uniformly in $i \geq 1$.*

Proof. The sequence $(\hat{\rho}_{i,n})_{n\geq 1}$ converges uniformly in $i \geq 1$ iif there exists $p > 0$ such that: For any $\epsilon, \eta > 0$ there exists $N \geq 1$ such that, for all $n \geq N$,

$$\mathbb{P}\left[\left|\sup_{1\leq i\leq n} \hat{\rho}_{i,n} - p\right| > \epsilon\right] < \eta.$$

Notice first that for all fixed $n \geq 1$, $\sup_{1\leq i\leq n} \hat{\rho}_{i,n} \geq \hat{\rho}_{n,n} := X_n$. Since $(X_n)_{n\geq 1}$ are independent of same law with unbounded support, X_n can take any positive integer value with strictly positive probability. Fix $p > 0$ and consider any $\epsilon > 0$. Then define $A_n := \{X_n > p + \epsilon\}$ for any $n \geq 1$. Since $\mathbb{P}[A_n]$ is strictly positive and independent from n, we can find $\eta > 0$ independent from n such that $\mathbb{P}[A_n] > \eta$. On A_n, we have $|X_n - p| > \epsilon$ and also $\left|\sup_{1\leq i\leq n} \hat{\rho}_{i,n} - p\right| > \epsilon$.

This is valid for any $n \geq 1$: For any $p > 0$, we can find $\epsilon, \eta > 0$ such that for all $N \geq 1$, there exists at least one $n \geq N$ such that $\mathbb{P}\left[\left|\sup_{1\leq i\leq n} \hat{\rho}_{i,n} - p\right| > \epsilon\right] > \eta$. The result follows. □

Let us now consider the sum of the ratios $R_{i,n}(\rho) > 0$ such that, from (2.3),

$$S_n(\rho) = \sum_{i=1}^{n} R_{i,n}(\rho).$$

We recall that the suggested estimator of the parameter $\boldsymbol{\rho}$ is

$$\hat{\rho}_S(n) = \arg\max_{\rho>0} S_n(\rho).$$

In a first Lemma 2.3, we establish some basic bounds for $\hat{\rho}_S(n)$.

Lemma 2.3. *For any $n \geq 1$, $\hat{\rho}_S(n) \in \left[\min_{1\leq i\leq n} \hat{\rho}_{i,n}, \max_{1\leq i\leq n} \hat{\rho}_{i,n}\right]$.*

Proof. Since, for all $1 \leq i \leq n$, $\arg\max_{\rho>0} R_{i,n}(\rho) = \hat{\rho}_{i,n}$, the map $\rho \mapsto R_{i,n}(\rho)$ is strictly increasing when $\rho < \hat{\rho}_{i,n}$ and strictly decreasing when $\rho > \hat{\rho}_{i,n}$. It follows that the function $\rho \mapsto S_n(\rho)$ is also strictly increasing when $\rho < \min\{\hat{\rho}_{i,n},\ 1 \leq i \leq n\}$ and strictly decreasing when $\rho > \max\{\hat{\rho}_{i,n},\ 1 \leq i \leq n\}$. The lemma follows. □

We need the next Lemma to get a more precise localization of $\hat{\rho}_S(n)$.

Lemma 2.4. *For any $\epsilon > 0$ and $i \geq 1$,*

$$\lim_{n\to\infty} R_{i,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) - R_{i,n}(\boldsymbol{\rho} - \epsilon) = +\infty \ a.s. \tag{2.6}$$

and

$$\lim_{n\to\infty} R_{i,n}\left(\boldsymbol{\rho} + \frac{\epsilon}{2}\right) - R_{i,n}(\boldsymbol{\rho} + \epsilon) = +\infty \ a.s.. \tag{2.7}$$

Proof. We show only (2.6), since (2.7) is obtained by symmetric arguments. Moreover, we can chose $i = 1$, because the other cases are similar.

Let us consider $\omega \in \Omega$ for which (2.5) is satisfied. In the proof, we omit the notation ω. We fix $\epsilon > 0$ and define

$$\begin{aligned} x_n &:= \lambda n\left(\hat{\rho}_{1,n}\log(\rho_1) - \rho_1 + 1\right), \\ y_n &:= \lambda n\left(\hat{\rho}_{1,n}\log(\rho_2) - \rho_2 + 1\right), \end{aligned}$$

where $\rho_1 := \boldsymbol{\rho} - \frac{\epsilon}{2}$ and $\rho_2 := \boldsymbol{\rho} - \epsilon$. Then $R_{1,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) - R_{1,n}(\boldsymbol{\rho} - \epsilon) = e^{x_n} - e^{y_n}$. Since $e^{x_n} - e^{y_n} \geq x_n - y_n$ when $x_n - y_n \geq 0$, the Lemma is proven, as soon we have shown that

$$x_n - y_n \xrightarrow[n\to\infty]{} +\infty. \tag{2.8}$$

We use the convergence $\hat{\rho}_{1,n} \xrightarrow[n\to\infty]{} \boldsymbol{\rho}$ to find $N \geq 1$ large enough such that, for all $n > N$, $\hat{\rho}_{1,n} > \rho_1 + \frac{\epsilon}{4}$. For these large n, we get

$$\begin{aligned} x_n - y_n &= \lambda n\left(\rho_2 - \rho_1 + \hat{\rho}_{1,n}(\log\rho_1 - \log\rho_2)\right) \\ &> \lambda n\left(\rho_2 - \rho_1 + \rho_1(\log\rho_1 - \log\rho_2) + \frac{\epsilon}{4}(\log\rho_1 - \log\rho_2)\right) \\ &= \lambda n\left((\rho_2 - \rho_1\log\rho_2) - (\rho_1 - \rho_1\log\rho_1) + \frac{\epsilon}{4}(\log\rho_1 - \log\rho_2)\right) \end{aligned}$$

where $(\rho_2 - \rho_1\log\rho_2) - (\rho_1 - \rho_1\log\rho_1) > 0$, because $\rho \mapsto \rho - \rho'\log\rho$ is decreasing for $0 < \rho \leq \rho'$, and $\log\rho_1 - \log\rho_2 > 0$. Hence

$$(\rho_2 - \rho_1\log\rho_2) - (\rho_1 - \rho_1\log\rho_1) + \frac{\epsilon}{4}(\log\rho_1 - \log\rho_2) > 0$$

and (2.8) follows. □

Remark 2.5. *Making vary $\epsilon > 0$ in* (2.6) *and* (2.7)*, we get as a by-product of this lemma that, for all $\rho > 0$ and all $i \geq 1$,*

$$\lim_{n\to\infty} R_{i,n}(\rho) = +\infty \ \ \mathbb{P}\text{-}a.s..$$

We first show the convergence of $(\hat{\rho}_S(n))_{n\geq 1}$ to $\boldsymbol{\rho}$ for $\nu = 0$. In this case $(X_n)_{n\geq 1}$ is an i.i.d. sequence with common Poisson law of parameter $\lambda\boldsymbol{\rho}$.

Proposition 2.6. *For $\nu = 0$, $\hat{\rho}_S(n) \xrightarrow[n\to\infty]{P} \boldsymbol{\rho}$.*

Proof. Fix $\epsilon, \eta > 0$. The result is established as soon as we have shown that there exists $N \geq 1$ large enough such that, for all $n \geq N$,

$$\mathbb{P}[|\hat{\rho}_S(n) - \boldsymbol{\rho}| \leq \epsilon] \geq 1 - 2\eta.$$

1) We start again from the law of large numbers: Set $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, n \geq 1$. By the $\mathbb{P}$-a.s. convergence of $(\bar{X}_n)_{n\geq 1}$ to $\lambda\boldsymbol{\rho}$, we can find some $n_0 > 1$ such that

$$\mathbb{P}\left[\forall n > n_0, \left|\frac{1}{\lambda}\bar{X}_n - \boldsymbol{\rho}\right| \leq \frac{\epsilon}{2}\right] \geq 1 - \frac{\eta}{3}$$

(see, for instance, Lemma 5.6 in Karr (1993)). In particular, for fixed $n \geq n_0$, it holds that

$$\mathbb{P}\left[\forall i \text{ s.t. } n - i \geq n_0, \left|\frac{1}{\lambda}\bar{X}_{n-i+1} - \boldsymbol{\rho}\right| \leq \frac{\epsilon}{2}\right] \geq 1 - \frac{\eta}{3}. \tag{2.9}$$

Now remark that the vectors $\left(\frac{1}{\lambda}\bar{X}_1, \ldots, \frac{1}{\lambda}\bar{X}_n\right)$ and $(\hat{\rho}_{n,n}, \hat{\rho}_{n-1,n}, \ldots, \hat{\rho}_{1,n})$ have same law. This implies that, if we set

$$A_n^1 = \left\{\forall i \leq n - n_0,\ |\hat{\rho}_{i,n} - \boldsymbol{\rho}| \leq \frac{\epsilon}{2}\right\}, \tag{2.10}$$

relation (2.9) is equivalent to

$$\mathbb{P}[A_n^1] \geq 1 - \frac{\eta}{3}.$$

2) For n_0 defined in the previous step, let $L > 0$ such that

$$\mathbb{P}\left[\forall i \leq n_0,\ R_{i,n_0}(\hat{\rho}_{i,n_0}) \leq \frac{L}{n_0}\right] \geq 1 - \frac{\eta}{3}. \tag{2.11}$$

Since, for all $i \geq 1$, the map $\rho \mapsto R_{i,n_0}(\rho)$ attains its maximum at $\hat{\rho}_{i,n_0}$, relation (2.11) implies that

$$\mathbb{P}\left[\forall \rho > 0, \sum_{i=1}^{n_0} R_{i,n_0}(\rho) \leq L\right] \geq 1 - \frac{\eta}{3}.$$

Finally, for all $n \geq n_0$, the vectors $(R_{1,n_0}, \ldots, R_{n_0,n_0})$ and $(R_{n-n_0+1,n}, \ldots, R_{n,n})$ have same law, and we deduce that, for all $n \geq n_0$,

$$\mathbb{P}\left[A_n^2\right] \geq 1 - \frac{\eta}{3}, \text{ for } A_n^2 := \left\{\forall \rho > 0, \sum_{i=n-n_0+1}^{n} R_{i,n}(\rho) \leq L\right\}. \tag{2.12}$$

3) By Lemma 2.4, for L defined in step 2), we can find $n_1 \geq 1$ such that, for all $n \geq n_1$,

$$\mathbb{P}\left[A_n^3\right] \geq 1 - \frac{\eta}{3}, \text{ with } A_n^3 := \left\{R_{1,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) - R_{1,n}(\boldsymbol{\rho} - \epsilon) \geq 2L,\right\}. \tag{2.13}$$

4) Combining now 1) - 3), we get, for all $n \geq \max\{n_0, n_1\}$, $\mathbb{P}[A_n] \geq 1 - \eta$ for $A_n := A_n^1 \cap A_n^2 \cap A_n^3 \subset \Omega$.

Since $\rho \mapsto R_{i,n}(\rho)$ is increasing for all $\rho < \hat{\rho}_{i,n}$, on A_n, for all $\rho < \boldsymbol{\rho} - \epsilon$ and for all $i \leq n - n_0$, relation (2.10) implies that

$$R_{i,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) \geq R_{i,n}(\rho)$$

and (2.13) implies that

$$R_{1,n}\left(\boldsymbol{\rho}-\frac{\epsilon}{2}\right) \geq R_{1,n}(\rho)+2L.$$

Thus

$$\sum_{i=1}^{n-n_0} R_{i,n}\left(\boldsymbol{\rho}-\frac{\epsilon}{2}\right) \geq \sum_{i=1}^{n-n_0} R_{i,n}(\rho)+2L.$$

Combined with (2.12), it follows that

$$\begin{aligned}\sum_{i=1}^{n} R_{i,n}\left(\boldsymbol{\rho}-\frac{\epsilon}{2}\right) &\geq \sum_{i=1}^{n} R_{i,n}(\rho)+2L+\left(\sum_{i=n-n_0+1}^{n} R_{i,n}\left(\boldsymbol{\rho}-\frac{\epsilon}{2}\right)-\sum_{i=n-n_0+1}^{n} R_{i,n}(\rho)\right)\\ &\geq \sum_{i=1}^{n} R_{i,n}(\rho)+L. \end{aligned} \tag{2.14}$$

This means that on A_n, for all $\rho < \boldsymbol{\rho}-\epsilon$,

$$S_n\left(\boldsymbol{\rho}-\frac{\epsilon}{2}\right) > S_n(\rho).$$

We conclude that, for all $n \geq \max\{n_0, n_1\}$, $\mathbb{P}[\hat{\rho}_S(n) \geq \boldsymbol{\rho}-\epsilon] \geq 1-\eta$. By symmetric arguments, we prove that, for n big enough, $\mathbb{P}[\hat{\rho}_S(n) \leq \boldsymbol{\rho}+\epsilon] \geq 1-\eta$. And combining both, the result follows. □

The convergence of the estimator for any deterministic finite $\nu \geq 1$ is a slightly wider case of Proposition 2.6.

Proposition 2.7. *For any deterministic finite $\nu \geq 1$, $\hat{\rho}_S(n) \xrightarrow[n\to\infty]{P} \boldsymbol{\rho}$.*

Proof. With the same notations as above, for $n > \nu$, the Shiryaev-Roberts sequence can be written as

$$S_n(\rho) = \sum_{i=1}^{n} R_{i,n}(\rho) = R_{\nu+1,n}(\rho)\sum_{i=1}^{\nu} R_{i,\nu}(\rho) + \sum_{i=\nu+1}^{n} R_{i,n}(\rho). \tag{2.15}$$

Set again $\hat{\rho}_{i,n} = \arg\max_{\rho>0} R_{i,n}(\rho) = \frac{1}{\lambda(n-i+1)}\sum_{k=i}^{n} X_k$.

1) Since the law of large numbers is still valid, for any $\epsilon, \eta > 0$, we can find $n_0 \geq 1$ such that, for all $n \geq n_0$,

$$\mathbb{P}\left[\tilde{A}_n^1\right] \geq 1-\frac{\eta}{4}, \text{ for } \tilde{A}_n^1 := \left\{\forall i \leq n-n_0,\ |\hat{\rho}_{i,n}-\boldsymbol{\rho}| \leq \frac{\epsilon}{2}\right\}. \tag{2.16}$$

2) Further, since $(X_n)_{n\geq\nu+1}$ is i.i.d. with Poisson law of parameter $\lambda\boldsymbol{\rho}$, (2.14) from proof of Proposition 2.6 ensures that we can find some $n_1 > \nu$ and $L > 0$ such that, for all $n \geq n_1$,

$$\mathbb{P}\left[\tilde{A}_n^2\right] \geq 1-\frac{\eta}{4}, \text{ for } \tilde{A}_n^2 := \left\{\forall \rho < \boldsymbol{\rho}-\epsilon,\ \sum_{i=\nu+1}^{n} R_{i,n}\left(\boldsymbol{\rho}-\frac{\epsilon}{2}\right) \geq \sum_{i=\nu+1}^{n} R_{i,n}(\rho)+L\right\}. \tag{2.17}$$

3) Now, for $n \geq n_0$, let ρ_n^- such that $\mathbb{P}\left[\rho_n^- \leq \min_{1\leq i\leq n} \hat{\rho}_{i,n}\right] \geq 1 - \frac{\eta}{8}$ and $0 < k < K$ such that $\mathbb{P}\left[\forall \rho \in [\rho_n^-, \boldsymbol{\rho} - \epsilon),\ k < \sum_{i=1}^{\nu} R_{i,\nu}(\rho) < K\right] \geq 1 - \frac{\eta}{8}$. Then

$$\mathbb{P}\left[\tilde{A}_n^3\right] \geq 1 - \frac{\eta}{4},\ \text{for } \tilde{A}_n^3 := \left\{\rho_n^- \leq \min_{1\leq i\leq n} \hat{\rho}_{i,n} \text{ and, } \forall \rho \in [\rho_n^-, \boldsymbol{\rho} - \epsilon),\ k < \sum_{i=1}^{\nu} R_{i,\nu}(\rho) < K\right\}. \tag{2.18}$$

4) From Lemma 2.4 and Remark 2.5, we can find $n_2 \geq 1$ such that, for all $n \geq n_2$,

$$\mathbb{P}\left[\tilde{A}_n^4\right] \geq 1 - \frac{\eta}{4},\ \text{with } \tilde{A}_n^4 := \left\{\frac{R_{\nu+1,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right)}{R_{\nu+1,n}(\boldsymbol{\rho} - \epsilon)} \geq \frac{K}{k}\right\}. \tag{2.19}$$

5) Combining 1) - 4), for $n \geq \max\{n_0, n_1, n_2\}$, we define $\tilde{A}_n := \tilde{A}_n^1 \cap \tilde{A}_n^2 \cap \tilde{A}_n^3 \cap \tilde{A}_n^4 \subset \Omega$. Then $\mathbb{P}[\tilde{A}_n] \geq 1 - \eta$. From (2.16), (2.19) and because $\rho \mapsto R_{i,n}(\rho)$ is increasing for all $\rho < \hat{\rho}_{i,n}$, it follows that, on $\tilde{A}_n$, for all $\rho < \boldsymbol{\rho} - \epsilon$,

$$R_{\nu+1,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) \geq \frac{K}{k} R_{\nu+1,n}(\rho).$$

Combined with (2.18), this gives, for all $\rho \in [\rho_n^-, \rho - \epsilon)$,

$$\begin{aligned} R_{\nu+1,n}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) \sum_{i=1}^{\nu} R_{i,\nu}\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) &\geq R_{\nu+1,n}(\rho) \frac{K}{k} \frac{\sum_{i=1}^{\nu} R_{i,\nu}(\boldsymbol{\rho} - \frac{\epsilon}{2})}{\sum_{i=1}^{\nu} R_{i,\nu}(\rho)} \sum_{i=1}^{\nu} R_{i,\nu}(\rho) \\ &\geq R_{\nu+1,n}(\rho) \sum_{i=1}^{\nu} R_{i,\nu}(\rho). \end{aligned}$$

This relation together with (2.17) and (2.15) gives, for all $\rho \in [\rho_n^-, \boldsymbol{\rho} - \epsilon)$,

$$S_n\left(\boldsymbol{\rho} - \frac{\epsilon}{2}\right) > S_n(\rho).$$

Thus, on $\tilde{A}_n$, $\hat{\rho}_S(n) \notin [\rho_n^-, \boldsymbol{\rho} - \epsilon)$. From the choice of ρ_n^-, it follows that, still on $\tilde{A}_n$, $\hat{\rho}_S(n) \geq \boldsymbol{\rho} - \epsilon$.

We conclude in the same way as for Proposition 2.6. □

2.3 Adaptive procedure for detecting a change of trend

In this section, we explore a new set up for the change-point detection procedure where the intensity of the Poisson random variables sequence is increasing/decreasing at a steady pace. We look for a change in the trend and suggest a way for the inference of the post-change parameters.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X = (X_n)_{n\geq 1}$ an independent random sequence such that $X_n \sim \text{Poisson}(\lambda_n)$, $n \geq 1$ where $\lambda_n := \alpha\lambda_{n-1}$ for $1 \leq n \leq \nu$ and $\lambda_n := \alpha'\lambda_{n-1}$ for $\nu + 1 \leq n$, with $1 \leq \nu < n$, $\lambda_0 > 0$, $\alpha > 0$, $0 < \alpha' := \alpha\boldsymbol{\rho}$, $\boldsymbol{\rho} > 0$ and $\alpha \neq \alpha'$. In this context, λ_0 and α are deterministic and known. ν and α' are deterministic but unknown.

The useful property of the Shiryaev-Roberts sequence to get rid of the change-point does not apply [10]. Therefore, we provide an adaptive procedure that uses estimates for both ν and $\boldsymbol{\rho}$. For fixed $n \geq 1$, the associated sequence $\xi = (\xi_{i,n})_{1 \leq i \leq n}$ is given for $i \in \{1, \ldots, n\}$ by

$$\xi_{i,n} := \sum_{k=1}^{i} \prod_{j=k}^{i} \frac{f^j_{\hat{\theta}_n}}{f^j_{\theta_0}}(X_j). \tag{2.20}$$

Here $(f^i_{\hat{\theta}_n})_{1 \leq i \leq n}$ denotes a sequence of density functions of independent Poisson random variables with intensity $\lambda_i := \alpha\lambda_{i-1}$, $1 \leq i \leq \hat{\nu}_\xi(n)$ and $\lambda_i := \alpha\hat{\rho}_\xi(n)\lambda_{i-1}$, $\hat{\nu}_\xi(n) + 1 \leq i \leq n$, and $\hat{\theta}_n := (\hat{\rho}_\xi(n), \hat{\nu}_\xi(n))$ an estimator of the couple (ρ, ν). In practice, $\hat{\theta}_n$ can be the maximum likelihood estimator. In some cases however, it can be useful to minimize instead the quadratic error, especially when the change of trend occurs for strongly decreasing intensities [11].

The procedure states that $\xi_{i,n}$ is computed as long as $\xi_{i,n} \leq x^*_n$, where x^*_n is a threshold sequence defined as in Equation (2.2).

In practice, $\hat{\rho}_\xi(n)$ and $\hat{\nu}_\xi(n)$ are estimated for each new observation using parallel computing as in Knaus et al. (2009). In fact, Pollak (2009) pointed out that when the Shiryaev-Roberts sequence grows, the computational time explodes. For example, when computing the original Shiryaev-Roberts procedure, a sample with 100 observations requires 11 times more than a benchmark sample with 10 observations. Because of the estimation of $\boldsymbol{\rho}$, the adaptive procedure for detecting a change of level with 10 observations requires 27 times more computing time than the benchmark sample and 260 times more with 100 observations. The adaptive procedure for detecting a change of trend is consuming even more computations due to the fact that, for each new observation, we need to re-compute the whole sequence. Compared to the benchmark sample, its computing time is 40 times higher with 10 observations and 1020 times higher with 100 observations. These figures were estimated using the same computing method, i.e. without parallelization. Therefore, when the flow of analyzed data increases significantly, reducing the computing time becomes a major point of attention.

Steps of the adaptive procedure for detecting a change of the trend

In summary, the adaptive procedure is given in the following 7 steps:

1. Set the time step to $n = 1$;
2. Estimate $\hat{\rho}_\xi(1)$ and $\hat{\nu}_\xi(1)$;
3. Compute the Shiryaev-Roberts sequence $(\xi_{i,1})_{i=1}$;
4. Compute a threshold $\tilde{x}^*_1$;
5. If the sequence $(\xi_{i,n})_{1 \leq i \leq n}$ overcomes the threshold, raise an alarm and stop the procedure. Otherwise, increment the time step n;

10. In the detection procedure for the level, our approach does not require to know the time of the change for the estimation of the change coefficient ρ, see Section 2.2, page 48.

11. As a matter of fact, MLE algorithms might converge towards a local optimum and looking for improvements on the first observations of the sample. In general, we recommend a careful implementation of the estimation algorithms and a specific study of their convergence. In addition, when the computing time is reasonable, we recommend scanning all possible time changes and estimating $\boldsymbol{\rho}$ alone; this might improve considerably the likelihood.

6. Estimate $\hat{\rho}_\xi(n)$ and $\hat{\nu}_\xi(n)$, and compute the sequence $(\xi_{i,n})_{1\leq i\leq n}$;
7. Go back to step 4;

Estimations of $\boldsymbol{\rho}$ and ν are provided by $\hat{\rho}_\xi(n)$ and $\hat{\nu}_\xi(n)$ when stopping. For practical considerations on the threshold computation, see Sections 2.1.2 and 2.4.

2.4 Study cases

In this section, we provide a few applications for the adaptive procedures. A Poisson framework is used for the mortality since it is a common assumption in the actuarial literature [12]. We also adjust the data with a proportional rescaling in order to align it to the chosen Poisson distribution: (i) As suggested in Zucchini and MacDonald (2009), Section 1.2.1, or Mei et al. (2011), Section 4, the data is first normalized by the population size in order to eliminate exposure effects. (ii) Secondly, we multiply the observed number of deaths by the coefficient that allows the empirical mean to be equal to the empirical standard deviation [13].

Two kind of study cases are provided. First, the adaptive procedure for detecting a change of level is applied to national and annual data from HMD [14]. It shows that, as expected, persistent changes of level occurred in the past and are correctly identified by the procedure. Therefore, we suggest that the procedure may be used as a risk management tool: in practice, raising an alarm is only a start and a detailed analysis of the causes of the alarm should follow.

We also studied some peculiar events that are obviously not persistent changes of level: when the population is homogeneous over time, at national scale, catastrophic events are easily associated to a clear cause (e.g. the Spanish flu). When they occur, an alarm is raised if they are far enough on the distribution tail. We illustrate that most of strong but not extreme variations that happen only one time are not detected, unlike slight but persistent changes that trigger the alarm. Therefore, the use of this tool makes sense when it is used along other controls such as confidence intervals (for the detection of strong variations).

In a second part, we focus on the change of the trend: here, we are interested in situations when the mortality decreases faster than expected. We notice first that the adaptive procedure given in Section 2.3 is sensitive to changes: a few regularly aligned divergent observations can raise an alarm when the data is historically stable. Then we recommend a cautious calibration of the procedure (here through the probability of false alarm). We also illustrate the case of specific events called **peak/compensation phenomena** and defined as (i) a sequence of regular decrease of the mortality followed by (ii) a one-time increase called the *peak*, (iii) a very short period of lower mortality that is the *compensation period* and (iv) a regular decrease of the mortality. For this kind of events, the aim of the procedure is to check whether the trend of the period (i)

12. See e.g. Rhodes and Freitas (2004) and Tomas and Planchet (2015).
13. This rescaling restores the right quantile for the application of a Poisson model. Notice that in the case of actuarial applications for insurer portfolios, data quality and full understanding are required to assess the use of such a methodology.
14. Human Mortality Database, http://www.mortality.org/.

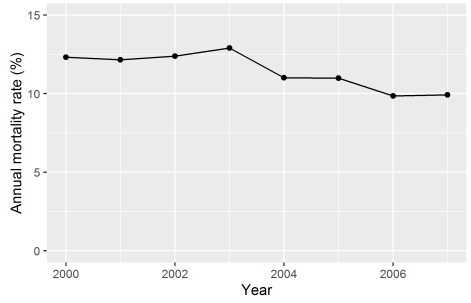
is identical to the one of the period (iv). An example is provided for the 2003 French heatwave.

Threshold calibration

In the following, we apply the threshold calibration from (2.2) where we set $\alpha = 0.01\%$. The threshold is estimated through simulations of the procedure in the case when the no change occurs.

2.4.1 Detecting a change of the level

Regarding the theoretical requirements, the adaptive procedure for detecting a change of level should be used in the context of a stable mortality over time. In addition, the intensity after the change should also be stable and persistent. These two constraints are strong but common requirements for the study of mortality at national scale. The first example illustrates the case when the intensity decreases during the years that followed a noticeable event. We study in a second phase the 1918 Spanish flu: in case of such extreme events, an alarm is raised anyway. In the third example, we look at the mortality of the portfolio from an insurance company, when nothing is known about the reasons of the observed variations and no change seems to occur. These chosen applications highlight the main properties and limits of the procedure.



(a) Actual mortality rate (Human Mortality Database).

n	$\hat{\rho}_S(n)$	$\tilde{S}_{i,n}$	$\tilde{s}^*_n$	Detection
2000	1.00	1.00	1.001	No
2001	0.99	2.1	5.7	No
2002	1.00	3.0	6.7	No
2003	1.02	5.9	84.7	No
2004	0.89	179	9 481	No
2005	0.89	42 904	10 058	Yes

(b) Detection procedure results

Figure 2.1 – Mortality rate of French civilians (85-90 years old) between 2000 and 2005.

The **2003 French heatwave** impacted significantly the mortality of elderly people: the event is the most noticeable for the age group 85-90 years old (French national mortality data, men and women), see Figure 2.1a. No change of level is detected when the 2003 peak occurs (annual mortality rate: 12.9%, $\hat{\rho}_S(2003) = 1.02$). However, the decrease of the mortality that follows (from 12.3% in 2000 to 9.9% in 2006) is detected in 2005 (see Table 2.1b, $\hat{\rho}_S(2005) = 0.89$). As expected, a persistent change of level is detected but not a sudden variation.

If we look at the probability of the events under the Poisson framework, the mortality rate observed in 2003 is a quantile with probability 92% while the low rate observed in 2005 is a quantile with probability 0.05%. In particular, we can calculate that the false alarm probability should be higher than 22% in order to raise an alarm. Therefore,

persistent changes are detected rather than one-time reasonable events.

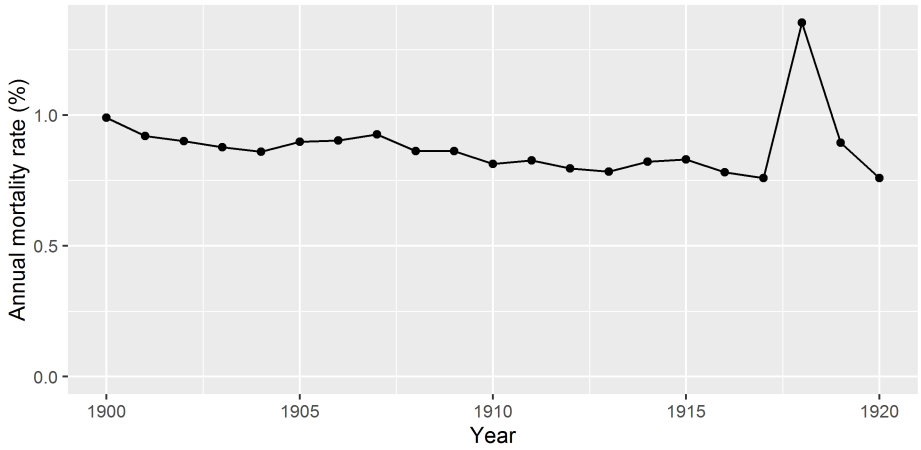


Figure 2.2 – Mortality rate of civilian French women (30-50 years old) between 1910 and 1919 (Human Mortality Database).

The **1918 Spanish flu** is an extreme event observed here for French women between 1900 and 1920 (non military, between 30 and 50 years old [15]). During this time period, the slight mortality improvements mainly come from the development of public health national programs and of specific procedures for the treatment of sick people. In Figure 2.2, observations suggest that the mortality is stable enough for this age group, except for the peak in 1918 that we want to challenge.

The level before the change is set using a 10 years window of observations, i.e. between 1900 and 1909, where the mortality rate is about 0.90%, and the probability of false alarm is set to 0.1%. The change is detected in 1918 [16], the same year it happens, with $\hat{\rho}_S(1918) = 1.5$. Usual statistical tools also provide strong results in this situation, as expected: under the Poisson framework, the upper bound of the confidence interval with probability 0.1% is a mortality rate of 1.02%. For the year 1918, the Standard Mortality Ratio, defined as the ratio between the observed number of deaths and the expected number of deaths, is very close to the estimation of $\boldsymbol{\rho}$ (here $SMR(1918) = 150\%$) while it is never over 100% before the peak since the mortality is slightly decreasing.

The low variance of the data before the detection increases the chance to raise an alarm when a strong variation is observed. This example shows that extreme events such as this one are detected as soon as they are observed [17].

15. See Caselli et al. (1987), page 44.

16. In order to focus on this event, the procedure is set to raise an alarm only in the case of an increased mortality level.

17. The same study on civilian men show that the mortality tripled from 1914 to 1918 before coming back to the same level as during the beginning of the century. The detection results are similar as for women: the alarm is raised in 1914 due to the war. Therefore, even if the change is persistent, it is detected the first year it occurs.

Figure 2.3 – Expected (filled) and observed (dotted) number of deaths between 2003 and 2014, data from a French insurance portfolio of life annuitants.

In the third example we study a **portfolio of real life annuitants** that contains about 15 000 life annuities with 50% of men and an average age of 77 years. Data is available from 2003 to 2014 by quarter. Most of the annuitants are between 60 and 90 years old. Because of the low size of the portfolio, we assume that the expected mortality is given by the French regulatory tables TGH-TGF-05 that are extended to recent generations. The purpose of our analysis is to assess whether the regulatory table is suited for the portfolio. A first analysis shows that, for the whole sample (men and women together), the quarterly number of predicted deaths is a reasonable average assumption since the observed number of deaths is close enough, with expected variations due to the size of the portfolio (see Figure 2.3). The procedure does not raise any alarm: no persistent change of level seems to occur for the whole portfolio.

A close analysis of the observations might suggest that there is a deviance between the observed and predicted deaths from 2011. In the following paragraph, we focus on detecting a change of trend and, for the life annuitant portfolio, we bring more information about the change of mortality over the studied period.

2.4.2 Detecting a change of the trend

We focus on three events: the decrease of the mortality in the 60's for women between 60 and 65 years old, the decrease of the mortality trend that followed the French heatwave of 2003, and the analysis of the insurance portfolio of life annuitants. In these examples, we show that the suggested procedure is sensitive to any sequence of observations that are diverging from the expected trend in the same direction: the calibration of the threshold becomes crucial in the detection process. Thus, persistent trends are detected quickly. We also illustrate the peak/compensation phenomenon defined in Section 2.4, page 55.

Initial trends used to run the procedure are calibrated on the 10 years window before the starting point. Post change trends and the time of change are provided by the detection procedure itself.

Decrease of the mortality rate in the 60's

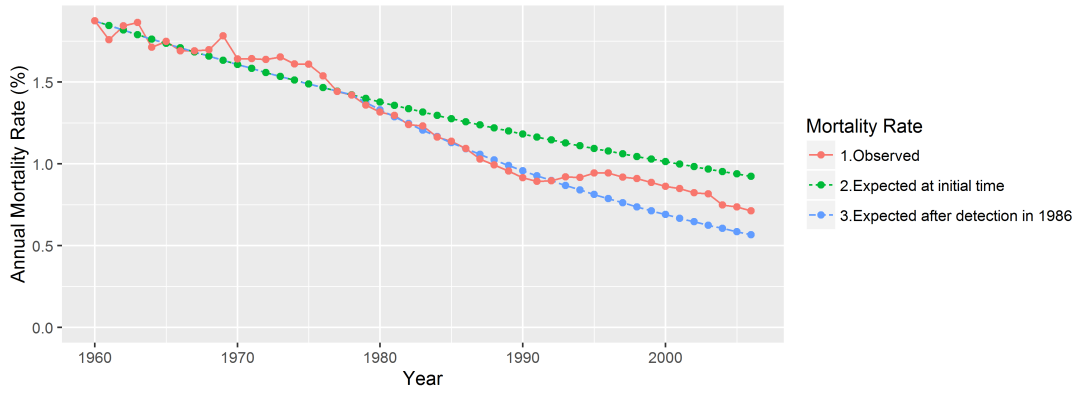


Figure 2.4 – Detection of the change of mortality trend for French women between 55 and 75 years old, from 1960

Using the adaptive procedure of Section 2.3, we look at national French data from HMD between 1960 and nowadays. Didou (2011) noticed that the female mortality rates are slightly decreasing at a steady peace just before 1960 [18] and the decrease accelerates in the 60's. Therefore, we chose to illustrate the procedure with the observed deaths of women between 55 and 75 years old. The Figure 2.4 shows that the trend changes around 1976/1978 and the procedure rises the alarm in 1986 (see Figure 2.5). The estimated decrease of mortality in 1960 is about 1.5% per year and becomes 3.2% after the change, i.e. starting from 1978.

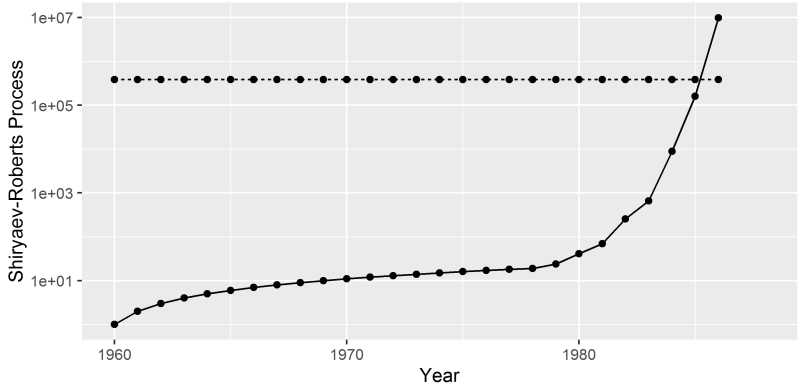


Figure 2.5 – Shiryaev-Roberts Process (filled) and threshold (dotted) for French women between 55 and 75 years old between 1960 and 1986.

In practice, the procedure is very sensitive to any change of trend: the choice for the probability of false alarm becomes a lever to distinguish long or short term changes. In this case, we look for long term changes of the trend. That is why we deliberately set the probability of false alarm to a low value (0.01%).

18. This is a consequence of the revolution in cardiovascular care, see Figure 1 from Vallin and Meslé (2010). It is also noticeable in other countries, as developed in Cutler and Meara (2001).

2003 French Heatwave

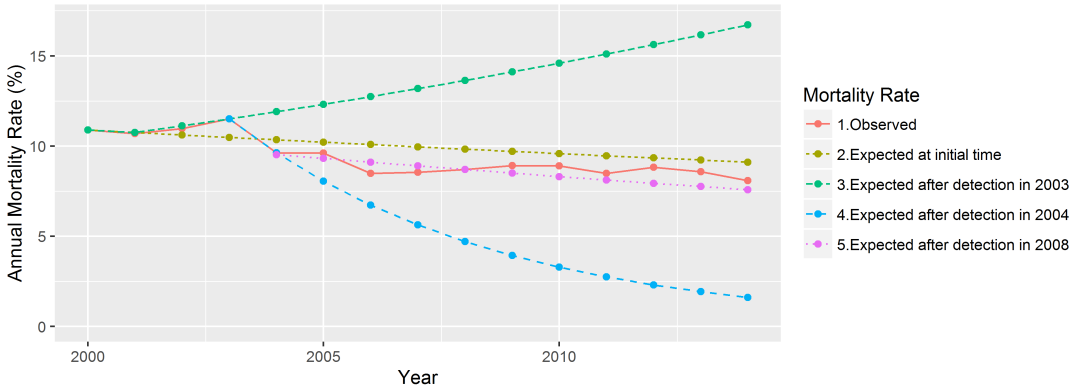


Figure 2.6 – Detection of the change of mortality trend for French population between 85 and 90 years old

In Section 2.4.1, we noticed that the peak of mortality in 2003 was not detected with the procedure for detecting a change of level. Surprisingly, the two years 2002 and 2003 are sufficient to raise an alarm for the observed increase (relative increase of +3.5% per year) since the observations are aligned enough. This is a clear example of a peak/compensation phenomenon, as defined in Section 2.4, page 55. Once we identified the peak (ii) and the compensation period (iii) through multiple applications of the procedure, we can test whether the regularly decreasing period (iv) has the same trend as the initial period (i). The procedure stops actually in 2008 and sates that the change of trend occurred in 2004 with an annual decrease of 2.3% (see Figure 2.6).

Life annuitants insurance portfolio

In Section 2.4.1, page 58, the procedure for detecting a change of level did not raise any alarm for the life annuitants portfolio. With the adaptive procedure for the trend, an alarm is raised in 2013 Q2 [19]. The initial decreasing trend of 0.6% changes to an annual increase of the mortality of 0.7% from 2005 Q3. A re-run from 2013 Q2 in order to detect a late change does not raise any alarm.

We want to highlight here the fact that the calibration of the initial trend impacts the result and its interpretation. Due to the very limited dataset at our disposal, we chose to keep 10 observations to estimate the initial trend. The observed data is also adjusted from the expected mortality in order to take into account the exposure of the underlying risk by age. With this limit, we conclude that a slight change of trend (from decreasing to increasing) happens in 2005 Q3. Further analysis have to be conducted by the insurer in order to monitor carefully this change.

In summary, the studied portfolio of life annuitants does not diverge significantly from its reference mortality table. More broadly, the use of the trend procedure is assessed to be efficient for detecting small changes of deviation from the initial trend, especially when the data is unstable or when one-time events happen. Indeed, alarms can bee raised very fast (e.g. for the 2003 heatwave). Or, as seen in the case of the 60's mortality,

19. Q1 is the first quarter of the year, Q2 the second quarter of the year, etc..

when the trend is stable, alarms are not raised too often and we can observe a long delay before detection.

2.5 Extension: weighted likelihood ratio

We put forward two possible extensions for the detection of change (level and trend). First, we suggest using Hidden Markov Models to take into account over-dispersion of the data. No application is provided in this chapter since a large enough amount of data is required to calibrate the distributions before and after the change. We also suggest the possibility of weighting the likelihood ratios by the exposure to the risk in the same way than Mei et al. (2011), Section 4.

Intuitively, when we observe data from an insurance portfolio, it appears that there are some fluctuations in the number of insured over time. In this case, actuarial methods suggest the use of weights that will take into account the exposure (i.e. the size of the portfolio) in the likelihood ratio. Let $n \geq 1$. With the same notations as in (2.4), we define an alternative sequence $(WL_{i,n})_{1\leq i\leq n}$ for the level procedure by

$$WL_{i,n} := \sum_{k=1}^{i} \prod_{j=k}^{i} \left(\frac{f_{\hat{\rho}_S(n)}(X_j)}{f_1(X_j)} \right)^{l_j}$$

where l_j, $j \geq 1$, is the size of the portfolio for each time step. Identically, with the same notations as in (2.20), we define an alternative sequence $(WT_{i,n})_{1\leq i\leq n}$ for the trend procedure by

$$WT_{i,n} := \sum_{k=1}^{i} \prod_{j=k}^{i} \left(\frac{f^j_{\hat{\theta}_n}}{f^j_{\theta_0}}(X_j) \right)^{l_j}.$$

The results of the application of these procedures are provided in Appendix 2.7.2, page 67.

2.6 Conclusion

In this chapter, we provided two adaptive procedure for the detection of a change of level and a change of trend within a discrete time Poisson framework. The sequential procedures are derived from the one developed by Shiryaev (1963) and Roberts (1966).

For both procedures, the time of change is assumed to be deterministic but unknown. The mortality is given by independent Poisson random variables.and the distribution before the change is assumed to be known. In the first approach, we suggest an estimator for the intensity after the change designed for the specific case of the Shiryaev-Roberts detection procedure. We establish that it is a consistent estimator. In the second approach, the change-point and the trend coefficient are estimated by usual MLE. In fine, we show that these approaches are practical tools for the mortality data exploration, especially when nothing is known about the post-change distribution.

Actuarial applications are provided in both cases. First, the adaptive procedure for detecting a change of level is applied to national and annual data in order to illustrate its

main characteristics. We conclude that a change of level might occur for insurer portfolios when the underlying population is modified over time (persistent data disruptions, even small ones). In this case, the procedure becomes a tool of risk management for actuaries and any raised alarm should initiate a complete analysis of its source. We show also that, as expected, most of strong variations that happen only one time are not detected except when they are extreme enough. Therefore, we recommend the use of this tool, combined with usual controls such as confidence intervals.

Then we focus on the change of trend for the longevity risk, i.e. we are interested in situations when the mortality decreases faster than expected. The adaptive procedure given in Section 2.3 is quite sensitive to changes. The probability of false alarm, set for the control of false detection, can be used as a lever to focus on short or long term changes. In addition, restarting the procedure allows a systematic analysis of specific events such as the peak/compensation phenomena (e.g. the 2003 French heatwave).

In Section 2.5, we suggest some extensions to our work: the question of the sequential estimation of the post change distribution is still wide and is worth to be explored.

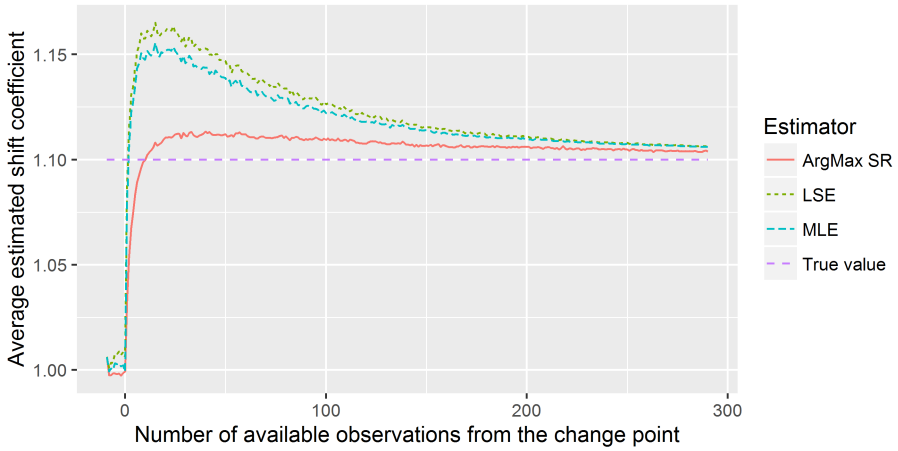
2.7 Appendices

2.7.1 Benchmarking

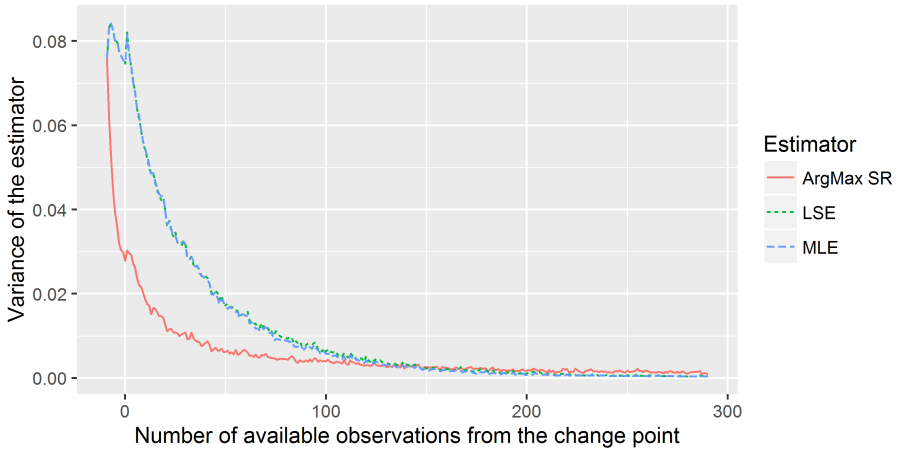
2.7.1.1 Change of level

The first intuition about the usefulness of an estimator that maximizes the Shiryaev-Roberts sequence is well illustrated in this paragraph. From simulations, the estimator converges faster to the true parameter (Figure 2.7a) with lower variance (Figure 2.7b). So far, for all the studied cases, we always observed this property despite the fact that it is not mathematically proven yet. In both Figures, the suggested estimator ($ArgMaxSR$) is compared to the Maximum Likelihood Estimator of the change-point model (MLE) and the ordinary least square estimator (LSE). This property is illustrated for parameters value from the 2003 French heatwave phenomenon. Figure 2.8a illustrates one observation of the random sequence, where no change is perceptible: the detection procedure is the only way to identify a persistent change. Here the estimator converges slowly toward a stable value since the change is very small.

Notice that Figure 2.7a would not be sufficient to identify a change in the context of stable data because we illustrated here average sequential estimates for the change coefficient: a single simulation of the sequential estimation is still quite volatile in practice.

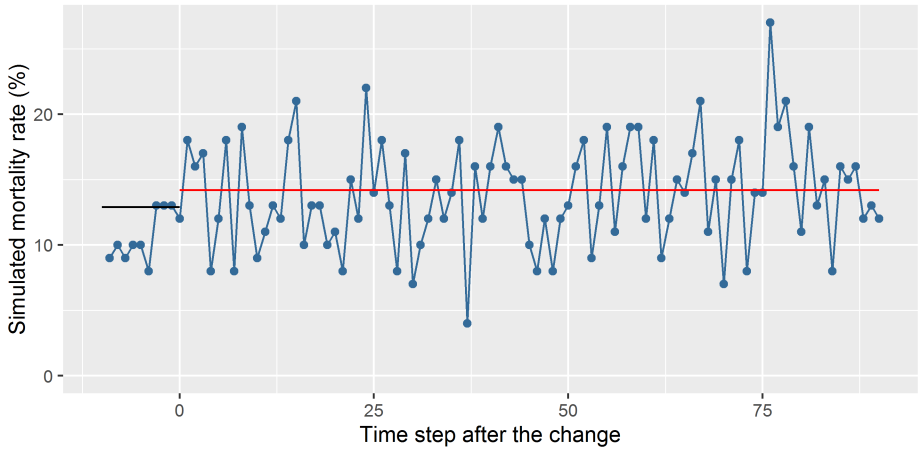


(a) Average sequential estimation of $\boldsymbol{\rho}$ by number of available observations, with $\nu = 10$, $\boldsymbol{\rho} = 1.1$, and 10 000 simulations. *Centered on the time of change.*

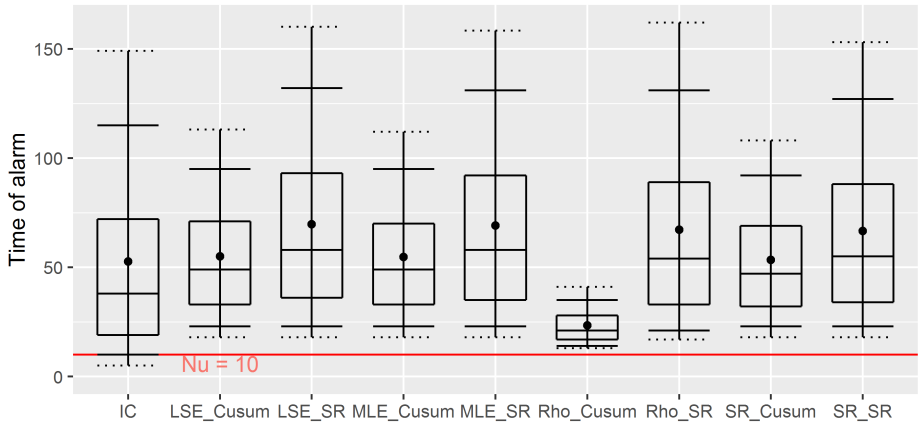


(b) Empirical sequential variance of estimators of $\boldsymbol{\rho}$ by number of available observations, with $\nu = 10$, $\boldsymbol{\rho} = 1.1$, and 10 000 simulations. *Centered on the time of change.*

Figure 2.7 – Benchmarking of the level procedure (1/2)



(a) Illustration of a realisation of the random experiment: $\nu = 10$, $\boldsymbol{\rho} = 1.1$. *Centered on the time of change.*



(b) Distribution of the alarm time, with $\nu = 10$, $\boldsymbol{\rho} = 1.1$, and 10 000 simulations. **Legend**: Point = mean. Box plots with 0.05, 0.1, 0.25, 0.5, 0.75, 0.9 and 0.95 probabilities.
Each scenario indicates first the methodology for the calibration of the sequential estimator (IC = Confidence Interval i.e. without any estimation of the change coefficient, MLE = Maximum of Likelihood Estimator, LSE = Least Square Estimator, SR = our estimator, Rho = case where the change coefficient is known) and then the procedure applied for the detection (CUSUM or Shiryaev-Roberts)..

Figure 2.8 – Benchmarking of the level procedure (2/2)

With 10 000 simulations of a sequence of Poisson random variables with a change at the time step $\nu = 10$ and $\boldsymbol{\rho} = 1.1$, we also observed that, in addition to a faster and more stable convergence to the true parameter, for a given sample size, the last value of the Shiryaev-Roberts sequence of the suggested estimator is always greater than the one from the MLE. This seems natural since the suggested estimator maximizes this value.

In addition, because we control the probability of false detection for each time step through the calibration of the threshold, with consideration of the methodology, the choice of estimator does not affect the probability of false detection. Figure 2.8b shows that not knowing the post change parameter affects strongly the delay before detection, especially the variance of the alarm time. It also shows that our estimator is still the

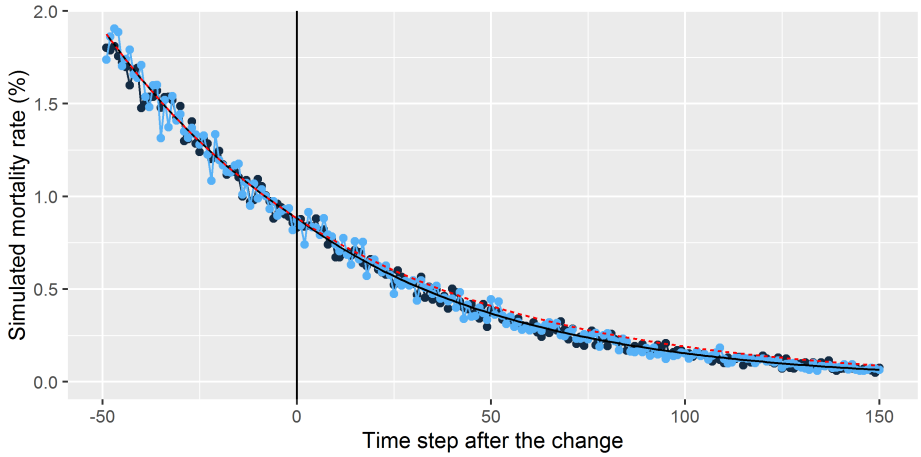
best among the pool of three candidates. Since we measure here the average delay of detection, the CUSUM procedure is optimal as expected (or close to it since El Karoui et al. (2017) proved it for the continuous time case). Therefore, the best strategy (in terms of what we studied, not in general) would be to use our estimator and then to detect the change with the CUSUM procedure, for minimizing the average delay before detection.

2.7.1.2 Change of trend

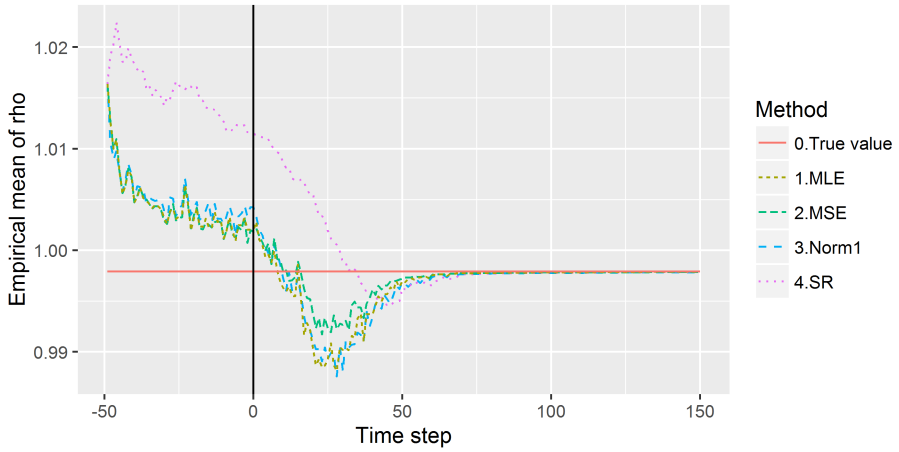
In this paragraph, we show that estimating the change of trend with the MLE for both the time of change and the change coefficient is more adequate than using the methodology applied for the level. In order to illustrate this point, we chose to replicate the case of the application of the methodology for the clear change of trend in the 60's given in Section 2.4.2, page 58. Then we set $\nu = 50$, $\alpha = 98.47\%$, $\alpha' = 98.27\%$ and thus $\boldsymbol{\rho} = 99.79\%$.

Figure 2.9a illustrates two simulations of the setup. Here again, the change is not noticeable and the detection procedure is required to identify it. Figure 2.9b and Figure 2.9c) show that the best estimator is the MLE (and no longer the one that maximizes the Shiryaev-Roberts sequence). In fact, in the level context, our estimator did not involve any estimation of the time of change ν while the MLE did it. Consequently, the variance was much lowered. Here, both methods have to estimate ν. Therefore maximizing the Shiryaev-Roberts sequence does not bring any improvement and the MLE is the best estimator to use. We also compared it to the Least Square Estimator (LSE) that minimizes the quadratic error and the Norm1 estimator that minimizes the absolute error.

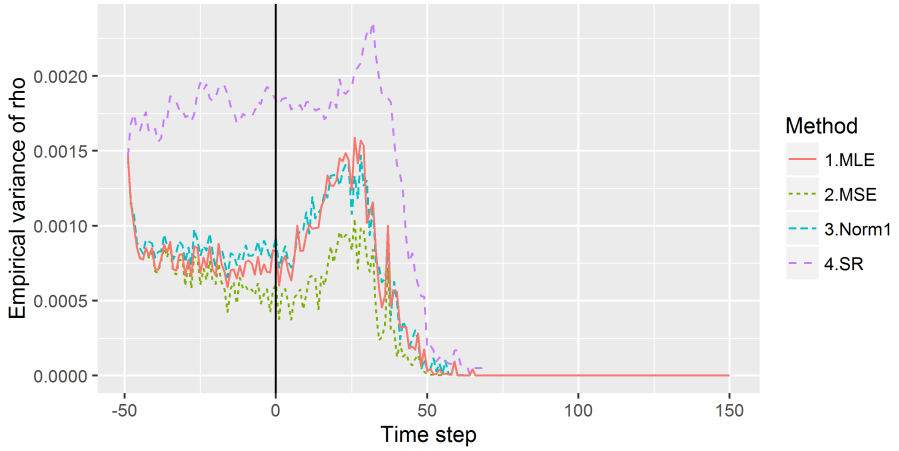
Hence we recommend the use of the MLE for this framework.



(a) Illustration of a realisation of the random experiment in the case of a change of trend: $\nu = 50$, $\boldsymbol{\rho} = 99.79\%$. Actual (filled) and Expected initially (dotted) intensities. *Centered on the time of change.*



(b) Average sequential estimation of $\boldsymbol{\rho}$ by time step: $\nu = 50$, $\boldsymbol{\rho} = 99.79\%$ and 1 000 simulations. *Centered on the time of change.*



(c) Empirical sequential variance of estimators of $\boldsymbol{\rho}$ by time step: $\nu = 50$, $\boldsymbol{\rho} = 99.79\%$ and 1 000 simulations. *Centered on the time of change.*

Figure 2.9 – Benchmarking of the trend procedure

2.7.2 Application of the weighted likelihood ratio procedures

The extensions suggested in Section 2.5, page 61, are illustrated in Table 2.1 for the level procedure and in Table 2.2 for the trend procedure.

It appears that, in our examples, taking into account the population weights did not impact significantly the results. However, results provided in Mei et al. (2011) show that when the population size varies (at a steady pace in the examples provided by the authors), the detection lag can be significantly different. Therefore, we expect those weighting to be crucial in the context of low and/or unstable data.

		with exposure weights		without exposure weights	
Study case	λ [20]	Alarm	$\lambda\hat{\rho}$	Alarm	$\lambda\hat{\rho}$
2003 heatwave	12.3%	2005	11.0%	2005	11.0%
Spanish flu	0.86%	1918	1.30%	1918	1.30%
Insurance portfolio	0.3%	No detection	N/A	No detection	N/A

Table 2.1 – Results of the level procedure with exposure weights.

		with exposure weights			without exposure weights		
Study case	α	Alarm	$\alpha\hat{\rho}$	$\hat{\nu}$	Alarm	$\alpha\hat{\rho}$	$\hat{\nu}$
60's mortality	-3.23%	1986	-1.53%	1978	1986	-1.53%	1978
2003 heatwave	-1.27%	2003	+3.46%	2001	2003	+3.46%	2001
Insurance portfolio	-0.62%	2013 Q3	+0.74%	2005 Q4	2013 Q2	+0.66%	2005 Q3

Table 2.2 – Results of the trend procedure with exposure weights.

Chapter 3

Weighted likelihood test for a change in one component of a parametric mixture

3.1 Introduction

Finite parametric mixtures of distributions play a central role in applied statistics, as they allow to describe experiments with different sub-populations [1]. The detection of at most one change-point in a closed sample is a standard problem in statistics but, to our knowledge, very few references specifically address this topic for mixtures with likelihood ratio-based techniques [2] (Andrews and Ploberger (1994), Hansen (1996), Pons (2009), Zou et al. (2015)). General and standard techniques, as exposed in Csörgő and Horváth (1997), can be adapted for finite parametric mixtures. However, when it comes to numerical applications, we observed that the standard approach raises many computational difficulties.

In this chapter, we consider a sample of n independent random variables that follow a finite mixture distribution with parametric components. The sample might contain *at most one change* (AMOC) in the parameters of the first component. If there is a change, the r.v. are identically distributed before and after the change-point: the parameters which describe the distribution of the first component are different before and after the break-point while the other parameters of the mixture remain the same. For example, a shift occurs in the mean or in the standard deviation of the first component in the case of a Gaussian mixture. We want to test whether there is a change or not. In order to circumvent the problems raised by the standard technique, we suggest a different approach that takes the form of a weighted likelihood test (WLT) [3]. In particular, the WL test can be computed using standard estimation algorithms. With a technique from Davis et al. (1995), we derive the limit distribution of its statistic under the null hypothesis in the form of a quadratic form of a multidimensional Brownian motion.

We start in Section 3.2 by the introduction of the model and the validity conditions

1. See e.g. Pearson (1894), Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Peel (2000), Frühwirth-Schnatter (2006), Pons (2009) or Lachos Dávila et al. (2018), Frühwirth-Schnatter et al. (2019).

2. Some of the existing work is dedicated to a Bayesian framework and therefore not in the scope of this chapter. See e.g. Giordani and Kohn (2008), Pandya and Jadav (2009), Pandya and Jadav (2010), Wilson et al. (2013), Li et al. (2018) or Ganji and Mostafayi (2019).

3. Weighted likelihood approaches are used in many contexts: see e.g. Dickey (1971), Hu and Zidek (2002), Amisano and Giacomini (2007), Basu et al. (2011), Song et al. (2018) and the references therein.

required for our main result. In particular, we impose that the change-point cannot occur too close to the first nor the last observation of the closed sample. In addition, the Maximum Likelihood Estimator (MLE) for the parameters of the mixture has to be strongly consistent. As in Davis et al. (1995) and Csörgő and Horváth (1997), the test is based on a likelihood ratio. The main difference from the standard approach lies in the presence of weight functions that allow to focus on the first component of the mixture. In Section 3.3, under the null hypothesis, we first obtain asymptotic properties of the MLE (Lehmann and Casella (1998)) before deriving a functional limit result for one term of the log-likelihood ratio in Theorem 3.16. This result is based on multiple applications of the Continuous Mapping Theorem and a Functional Delta Method in the Skorokhod metric space of càd-làg functions (Billingsley (1999), van der Vaart (1998)). In Theorem 3.17, the limit distribution of the test statistic is obtained as a consequence. In Section 3.4, we suggest an extension of the test (EWLT) where we scale the contribution of the sample to the weighted likelihood ratio. This improves significantly the detection frequency of the test in the case of a change (lower type II error). In Section 3.5, we show that validity conditions hold for univariate finite Gaussian mixtures within the framework of Hathaway (1985).

Applications in Section 3.6 consist in two parts. First, with numerical simulations, we illustrate the properties of the WL and EWL tests and compare them to a *benchmark* test (BM) obtained from an application of the standard test (e.g. exposed in Csörgő and Horváth (1997)). Both WL and EWL tests have notably lower type II errors, especially for large samples of over 10 000 observations. Overall, the EWLT performs significantly better than the other candidates. The second application is an illustration of the WL and EWL tests on Property and Casualty insurance real data. The tests are applied for the detection of a change in the variation over six months of the claim amount. In insurance problems, this application indicates that the proposed tests can be used for the monitoring of changes, when they are unexpected, and also to assess their significativity when they are known or suspected.

In Section 3.7, we give an overview of the conclusions and perspectives of this work.

3.2 Description of the model, assumptions and notations

3.2.1 Model and assumptions

We consider an experiment where we observe a sample of n independent continuous random variables $X = (X_i)_{1 \leq i \leq n}$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with values in some set $\mathcal{X}$, subset of an Euclidean space, endowed with Lebesgue's measure. Each X_i, $1 \leq i \leq n$, follows a finite **mixture distribution** with $2 < m < \infty$ parametric components $\mathbb{P}_{\theta^i}$, where θ^i belongs to a convex set of eligible parameters Θ. More precisely, for m a fixed, deterministic and known integer, the elements of Θ are of following type: $\Theta \ni \theta = (p_1, \ldots, p_{m-1}, \lambda_1, \ldots, \lambda_m)$, with $(p_1, \ldots, p_{m-1})$ in the open set

$$\Theta_0 := \left\{ (p_1, \ldots, p_{m-1}) \in (0,1)^{m-1}, \sum_{k=1}^{m-1} p_k < 1 \right\}$$

and, for each $k \in \{1, \dots, m\}$, $\lambda_k \in \Theta_k$, with Θ_k an open convex subset of some $\mathbb{R}^{d_k}$, with $d_k \geq 1$. Set $d := m-1+\sum_{k=1}^{m} d_k$: then $\Theta = \Theta_0 \times \prod_{k=1}^{m} \Theta_k$ is an open convex subset of $\mathbb{R}^d$.

Finally, given $f_1, \dots, f_m$ some fixed density functions on $\mathcal{X}$, the distribution $\mathbb{P}_\theta$, $\theta \in \Theta$, admits the density

$$f(x, \theta) := \sum_{k=1}^{m} p_k f_k(x, \lambda_k), \ x \in \mathcal{X},$$

with $p_m := 1 - \sum_{k=1}^{m-1} p_k$. We first assume that the distributions $\{\mathbb{P}_\theta, \theta \in \Theta\}$ are all distinct. This means in particular that the mixture should be identifiable [4]. We also add some usual assumptions on the regularity of the components of the mixture:

- ◇ The distributions defined by $\{f(., \theta), \theta \in \Theta\}$ have common support, i.e. the set $\{x \in \mathcal{X}, f(x, \theta) > 0\}$ does not depend on θ;
- ◇ For almost all $x \in \mathcal{X}$, the function $\theta \mapsto f(x, \theta)$ is twice continuously differentiable in $\theta \in \Theta$ with partial derivatives bounded by a non-negative integrable function of x that does not depend on θ.

In this experiment, the sample is identically distributed before and after the change-point. There is *at most one change* (AMOC), deterministic but unknown, or none. Since we are interested in the limit behavior of the sample of variables when their number tends to infinity, we suppose that the experience takes place in the time interval $[0, 1]$ and each variable X_i occurs at time i/n, $1 \leq i \leq n$.
With this approach, if there is a change-point, it occurs at a time denoted by $\boldsymbol{s} \in (0, 1)$ such that the parameters which describe the distribution of the first component are different before and after the break-point while the other parameters of the mixture remain the same. We write $\boldsymbol{\theta} := \theta^1 = \cdots = \theta^{\lfloor \boldsymbol{s}n \rfloor} \neq \theta^{\lfloor \boldsymbol{s}n \rfloor + 1} = \cdots = \theta^n$, where, for any $x \in [0, 1]$, $\lfloor x \rfloor$ denotes the integer part of x.
If there is no change-point, then X is an i.i.d. sample with $\boldsymbol{\theta} = \theta^1 = \cdots = \theta^n$.

For the sake of simplicity, we impose that there is at most one change. A setting with more than one change-point can be extended with the same logic.

In the sequel we suppose that the following holds for the change-point $\boldsymbol{s}$:

Assumption 3.1. *If there is a change, the change-point $\boldsymbol{s}$ is contained in $[\bar{s}, 1 - \bar{s}]$ where $0 < \bar{s} < 1/2$ is deterministic and known, i.e. the change does not occur too close to 0 nor 1.*

Let us construct the following hypothesis test:

1. The **null hypothesis** H_0 defines the case when there is no change-point;
2. The composite alternative hypothesis H_1: a change-point occurs at time s, $s \in [\bar{s}, 1-\bar{s}]$, i.e. the parameters which describe the distribution of the first component are different before and after s while the other parameters of the mixture remain the same.

This setting implies that the number of components does not change. Our work provides results when H_0 holds; so they do not depend on what happens after the change. However, the test statistic we shall define is designed to amplify the change in one of the

4. See for example Redner (1981), Feng and McCulloch (1996), and Section 1.14 in McLachlan and Peel (2000) for discussion on the identifiability of mixtures.

components of the mixture. Hence it makes sense to assume that the distribution after the change is a well defined mixture (identifiable) with the same number of components.

After defining a statistic, we will establish a central limit theorem under the assumption that H_0 holds. This is necessary to be able to determine the rejection domain while controlling the type I error (proportion of false positives). Therefore, **in the following, we assume that H_0 holds**, i.e. X is independent identically distributed with distribution $\mathbb{P}_{\boldsymbol{\theta}}$. We also consider that the following assumptions hold:

Assumption 3.2. *For almost all $x \in \mathcal{X}$,*

- ◇ ***Regularity**: the density $f(x, \theta)$ is three times differentiable in $\theta = (\theta_1, \ldots, \theta_d) \in \Theta$.*
- ◇ ***Integrability**: for $1 \leq j \leq d$,*

$$\mathbb{E}_{H_0}\left[\frac{\partial}{\partial \theta_j} \log f(X_1, \boldsymbol{\theta})\right] = 0.$$

- ◇ ***Continuity**: for any $1 \leq j, k, l \leq d$, the applications*

$$\theta \mapsto \frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l} \log f(x, \theta)$$

are continuous on Θ.

- ◇ ***Dominance**: we can find some function $\kappa_1(x)$, $x \in \mathcal{X}$, that does not depend on θ and is such that, for all $1 \leq j, k, l \leq d$ and all θ in Θ,*

$$\left|\frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l} \log f(x, \theta)\right| \leq \kappa_1(x)$$

with $\mathbb{E}_{H_0}[\kappa_1(X_1)] < \infty$.

Let $\boldsymbol{I}$ be the **Fisher information matrix** defined by

$$\boldsymbol{I}_{j,k} := \mathrm{Cov}\left(\frac{\partial}{\partial \theta_j} \log f(X_1, \boldsymbol{\theta}), \frac{\partial}{\partial \theta_k} \log f(X_1, \boldsymbol{\theta})\right), \ 1 \leq j, k \leq d. \tag{3.1}$$

The assumptions we state on $\boldsymbol{I}$ are the following:

Assumption 3.3. *The matrix $\boldsymbol{I}$ is positive definite with finite elements and*

$$\boldsymbol{I}_{j,k} = \mathbb{E}_{H_0}\left[\frac{\partial}{\partial \theta_j} \log f(X_1, \boldsymbol{\theta}) \frac{\partial}{\partial \theta_k} \log f(X_1, \boldsymbol{\theta})\right] = -\mathbb{E}_{H_0}\left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(X_1, \boldsymbol{\theta})\right].$$

We define the **log-likelihood** of the sample X as a function of the d-dimensional vector $\theta \in \Theta$ by

$$\theta \mapsto L(X, \theta) := \sum_{i=1}^{n} \log f(X_i, \theta).$$

It is well defined since, by assumption, the $X_i(\omega)$, $\omega \in \Omega$ and $1 \leq i \leq n$, can only take values in the set $\{x \in \mathcal{X}, f(x, \theta) > 0\}$.

With Assumptions 3.2 and 3.3, from the usual Limit Theorems [5] for Maximum Likelihood Estimators (MLE), there exist sequences of solutions of the likelihood equations $\frac{\partial}{\partial\theta_j}L(X,\theta) = 0$, $1 \leq j \leq d$, that exist with probability tending to one as $n \to \infty$ and are consistent. Let us select one of these consistent sequences of solutions as an **estimator for the unknown $\boldsymbol{\theta}$** and denote it $\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_{m-1}, \hat{\lambda}_1, \ldots, \hat{\lambda}_m)$.

Remark 3.1. *If the solution exists and is unique, then it is a Maximum Likelihood Estimator (MLE). However the existence and unicity of $\hat{\theta}$ for finite samples of mixtures are not obvious: for example, the MLE of a Gaussian mixture might not even exist for a finite sample since the likelihood can be unbounded*[6]*. For numerical applications, it is clear that the search of all roots of the likelihood equations would require unreasonable time. Thus the question of uniqueness is not really the problem. For the computation, one might use the EM algorithm*[7] *or some other approach*[8]*. Algorithms, when they converge, ensure to provide some local maximum but there is no guarantee to find the global maximum. We will see in Section 3.5 that, for univariate Gaussian mixtures, one can find some Θ that ensures the existence of the estimator $\hat{\theta}$ for finite samples.*

In addition to the usual regularity conditions given in Assumptions 3.2 and 3.3, we assume that:

Assumption 3.4. *When H_0 holds, the estimator $\hat{\theta}$ converges almost surely to $\boldsymbol{\theta}$ as $n \to \infty$, i.e. it is strongly consistent.*

This is an important restriction compared to the general case because Assumptions 3.2 and 3.3 only ensure the convergence in probability (Lehmann and Casella (1998)). For the applications in this chapter, we will consider the Gaussian case and discuss it in Section 3.5, using the result from Hathaway (1985). For other cases, one might use the classical results in the literature that cover a wide range of reasonable sufficient conditions. One of the first results has been given by Theorem 2 in Wald (1949) when Θ is compact. Theorem 4 in Redner (1981) weakens Wald's conditions for Θ a compact subset of the quotient topological space of all possible parameters, with a dedicated result for finite mixtures in Theorem 5. Feng and McCulloch (1996) proves it for Θ a compact subset of the Euclidean space of possible parameters, i.e. showing that the question of identifiability does not impact the convergence properties of maximum likelihood estimators. Other approaches have been proposed by Kiefer and Wolfowitz (1956) or Redner and Walker (1984). For a wider discussion on mixtures, refer to Section 2.5 in McLachlan and Peel (2000).

Remark 3.2. *Since we study asymptotic properties, the main point in Assumption 3.2 is the integrability of the function κ_1. With the continuity condition and the convergence of $\hat{\theta}$ from Assumption 3.4, almost surely, the boundedness itself follows for n large enough.*

5. See e.g. Theorem 1.5.

6. For the existence of MLE of Gaussian mixtures, see e.g. Example 6.10 in Lehmann and Casella (1998), Section 6.6. For a discussion on uniqueness of MLE in general, see Mäkeläinen et al. (1981). Chapter 2 in McLachlan and Peel (2000) gives an overview of these questions for mixtures.

7. See e.g. Dempster et al. (1977), Wu (1983), Hathaway (1983), Redner and Walker (1984), Benaglia et al. (2009) and the references therein.

8. For an overview, see e.g. Section 1.13 in McLachlan and Peel (2000), Tanaka (2009), Chen (2017) and the references therein.

For $s \in [\bar{s}, 1-\bar{s}]$, by the same logic as for the estimator $\hat{\theta}$, we consider the estimators of $\boldsymbol{\theta}$ over the subsamples $(X_i)_{1\leq i\leq \lfloor sn \rfloor}$ and $(X_i)_{\lfloor sn \rfloor+1\leq i\leq n}$, respectively denoted by $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$. For a fixed $s \in (0,1)$, they have the same properties as $\hat{\theta}$ when $n \to \infty$.

To focus our study on only one component of the mixture (the first component), we design a specific weight function that, in the detection statistic, allows to overweight the density of an observation x when the density of the first component is dominant compared to others. For that purpose we define the weight function at point $x \in \mathcal{X}$ for $\theta = (p_1, \ldots, p_{m-1}, \lambda_1, \ldots, \lambda_m) \in \Theta$ by

$$w(x,\theta) := \frac{p_1 f_1(x,\lambda_1)}{f(x,\theta)}. \tag{3.2}$$

The function w is well defined since, by assumption, the random variables X_i, $1 \leq i \leq n$, can only take values in the set $\{x \in \mathcal{X}, f(x,\theta) > 0\}$. In addition, by definition, for any $x \in \mathcal{X}$ and any $\theta \in \Theta$,

$$0 \leq w(x,\theta) \leq 1. \tag{3.3}$$

As a consequence of Assumption 3.2, the application

$$w \log f_1 : (x,\theta) \mapsto w(x,\theta) \log f_1(x,\lambda_1) \tag{3.4}$$

is twice differentiable in $\theta \in \Theta$, and, for all $1 \leq j,k \leq d$ and almost all $x \in \mathcal{X}$, the application

$$\theta \in \Theta \mapsto \frac{\partial^2}{\partial\theta_j \partial\theta_k}(w \log f_1)(x,\theta)$$

is continuous in $\boldsymbol{\theta}$.

Here is an assumption concerning the application w:

Assumption 3.5. ***(Dominance)*** *There exist some convex subset $\Theta' \subset \Theta$ such that $\boldsymbol{\theta}$ is in the interior of Θ', and an application κ_2 from $\mathcal{X}$ to $\mathbb{R}$ that does not depend on θ, such that, for all $1 \leq j,k \leq d$, θ in Θ' and for almost all $x \in \mathcal{X}$,*

$$\left| \frac{\partial^2}{\partial\theta_j \partial\theta_k}(w \log f_1)(x,\theta) \right| \leq \kappa_2(x)$$

with $\mathbb{E}_{H_0}[\kappa_2(X_1)] < \infty$.

For the same reasons as for Assumption 3.2, the essential point here is the integrability of the dominating function κ_2.

3.2.2 Definition of the Weighted Likelihood Test (WLT)

We now define the test statistic. First we introduce $\Lambda_n := (\Lambda_{s,n})_{s\in[\bar{s},1-\bar{s}]}$, the underlying càd-làg stochastic process of the detection statistic:

$$\begin{aligned}\Lambda_{s,n} &:= \log\left(\frac{\prod_{i=1}^{\lfloor sn\rfloor} f_1(X_i,\hat{\lambda}_{0,s,1})^{w(X_i,\hat{\theta}_{0,s})}\prod_{j=\lfloor sn\rfloor+1}^{n} f_1(X_j,\hat{\lambda}_{s,1,1})^{w(X_j,\hat{\theta}_{s,1})}}{\prod_{i=1}^{n} f_1(X_i,\hat{\lambda}_1)^{w(X_i,\hat{\theta})}}\right) \\ &= \sum_{i=1}^{\lfloor sn\rfloor} w(X_i,\hat{\theta}_{0,s})\log f_1(X_i,\hat{\lambda}_{0,s,1}) + \sum_{i=\lfloor sn\rfloor+1}^{n} w(X_i,\hat{\theta}_{s,1})\log f_1(X_i,\hat{\lambda}_{s,1,1}) \\ &\quad - \sum_{i=1}^{n} w(X_i,\hat{\theta})\log f_1(X_i,\hat{\lambda}_1).\end{aligned} \tag{3.5}$$

Note that, for an observation X_i with distribution parameter θ, the weight $w(X_i,\theta)$ is the probability that X_i comes from the first component. Conditionally to this fact, the log-likelihood of X_i is given by $\log f_1(X_i,\lambda_1)$. Thus the expression $w(X_i,\theta)\log f_1(X_i,\lambda_1)$ in $\Lambda_{s,n}$ somehow reflects the contribution of the first component in the likelihood of X_i. As a consequence, the response of the statistic is magnified when a change occurs in the first component.
In addition, the process Λ_n is defined on $[\bar{s},1-\bar{s}]$ in order to ensure that, for n large enough, an asymptotic behavior can be obtained for each sum. The **test statistic** is then defined by

$$S_n := \sup_{s\in[\bar{s},1-\bar{s}]} \Lambda_{s,n}. \tag{3.6}$$

We refer to this test as the **WLT** (Weighted Likelihood Test). The test procedure states that there is no change-point for the first component (i.e. we accept H_0) when S_n is smaller than some threshold L_α chosen with respect to a **false alarm constraint**. This false alarm can be obtained from the probability of false alarm $\alpha\in(0,1)$ such that L_α is the α-percentile of the limit distribution of S_n when H_0 holds.

The main purpose of this chapter is to derive the limit distribution of S_n when H_0 holds. We follow the work of Davis et al. (1995) and look at $\Lambda_{s,n}$ as a stochastic process. In the next section, we start by a focus on the properties of the estimators $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$. Then we derive the limit distribution of the detection statistic S_n in Theorem 3.17.

3.2.3 Notations

We denote by $D_\theta(.)$, $D^2_\theta(.)$ and $D^3_\theta(.)$ respectively the vector, matrix and hypermatrix differential operators in $\theta\in\mathbb{R}^d$.
For $\theta,\tilde{\theta}\in\mathbb{R}^d$, we denote by $[\theta,\tilde{\theta}]$ the segment $[\theta,\tilde{\theta}] := \{\lambda\theta+(1-\lambda)\tilde{\theta},\ \lambda\in[0,1]\}$.

$gl_d(\mathbb{R})$ denotes the set of matrices of size $d\times d$ with real coefficients and $GL_d(\mathbb{R})$ the set of invertible $d\times d$-matrices with real coefficients.
For a given matrix M, its i-th line is denoted by $M_{i,.}$ and its j-th column is denoted by $M_{.,j}$. The same logic is used for hypermatrices: for a given $J\in\mathbb{R}^{d\times d\times d}$ and $1\le i\le d$, we denote by $J_{i,.,.} := (J_{i,j,k,})_{1\le j,k\le d}$ the $d\times d$-matrix obtained from J.
For a given matrix M, we denote $(M^{-1})^T$ by $M^{-1\,T}$.

For $d_1, d_2 > 0$, we endow the space $F = \mathbb{R}^{d_1} \times gl_{d_2}(\mathbb{R})$ with the norm $\|.\|_2$ defined for the pair $x = (y, Z) \in \mathbb{R}^{d_1} \times gl_{d_2}(\mathbb{R})$ by $\|x\|_2^2 := \sum_{i=1}^{d_1} y_i^2 + \sum_{1 \leq i,j \leq d_2} Z_{i,j}^2$. The norm used for y in $\mathbb{R}^{d_1}$ is the Euclidean norm. The norm used for $Z \in gl_{d_2}(\mathbb{R})$ is the *entrywise* 2-norm, also known as the Frobenius norm.

The space of càd-làg functions, defined on some interval $E \subseteq [0,1]$ with values in F, is denoted by $\mathbb{D}(E,F)$ and referred as the Skorokhod metric space with the Skorokhod metric $d_{\mathbb{D}(E,F)}(.,.)$ defined for ζ_1 and ζ_2 in $\mathbb{D}(E,F)$ by

$$d_{\mathbb{D}(E,F)}(\zeta_1, \zeta_2) := \inf_{\tau \in \Gamma_E} \max \left\{ \sup_{s \in E} |\tau(s) - s|, \sup_{s \in E} \|\zeta_1(s) - \zeta_2 \circ \tau(s)\|_2 \right\} \tag{3.7}$$

with Γ_E the set of continuous and strictly increasing bijections from E to itself. For some arguments, we also consider the norm $\|.\|_2$ on $\mathbb{D}(E,F)$ defined for $\zeta \in \mathbb{D}(E,F)$ by $\|\zeta\|_2 := \sup_{s \in E} \|\zeta(s)\|_2$.
Refer to Section 12 in Billingsley (1999) for a detailed construction of the Skorokhod topology and the space $\mathbb{D}(E,F)$.

If Σ^2 is a covariance matrix, then it is positive semi-definite, and Σ will denote the unique positive semi-definite square root of Σ^2.

A glossary of notations is given in Appendix 3.9.

3.3 Limit distribution of the test statistic

In this section we shall determine the limit distribution of the process Λ_n as n tends to infinity.
Let us consider the constant

$$\boldsymbol{u} := \mathbb{E}_{H_0} \left[D_\theta (w \log f_1)(X_1, \boldsymbol{\theta}) \right] \in \mathbb{R}^d, \tag{3.8}$$

where the application $w \log f_1$ is defined in (3.4). By Assumption 3.5, $\boldsymbol{u}$ is finite. We indicate that $\boldsymbol{u}$ is not null in general: we numerically established that $\boldsymbol{u}$ is strictly positive for some examples of Gaussian mixtures [9]. The càd-làg real-valued process Λ_n defined on $[\bar{s}, 1-\bar{s}]$ can be decomposed as follows:

$$\Lambda_{s,n} = Q_{s,n}^1 + Q_{s,n}^2 - Q_{1,n}^1, \ s \in [\bar{s}, 1-\bar{s}], \tag{3.9}$$

9. Using the Strong Law of Large Numbers, a simple numerical simulation for a Gaussian mixture with 3 components shows that, in general, the constant $\boldsymbol{u}$ is not null. See Appendix 3.8.1 for an illustration.

where $Q_n^1 = (Q_{s,n}^1)_{s\in[\bar{s},1]}$ and $Q_n^2 = (Q_{s,n}^2)_{s\in[\bar{s},1-\bar{s}]}$ are càd-làg real-valued processes defined by

$$\begin{aligned} Q_{s,n}^1 &:= \sum_{i=1}^{\lfloor sn\rfloor} \Big(w(X_i, \hat{\theta}_{0,s}) \log f_1(X_i, \hat{\lambda}_{0,s,1}) - w(X_i, \boldsymbol{\theta}) \log f_1(X_i, \boldsymbol{\lambda}_1) \Big) \\ &\quad - \boldsymbol{u}^T \boldsymbol{I}^{-1} \sum_{i=1}^{\lfloor sn\rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta}),\ s \in [\bar{s}, 1], \\ Q_{s,n}^2 &:= \sum_{i=\lfloor sn\rfloor+1}^{n} \Big(w(X_i, \hat{\theta}_{s,1}) \log f_1(X_i, \hat{\lambda}_{s,1,1}) - w(X_i, \boldsymbol{\theta}) \log f_1(X_i, \boldsymbol{\lambda}_1) \Big) \\ &\quad - \boldsymbol{u}^T \boldsymbol{I}^{-1} \sum_{i=\lfloor sn\rfloor+1}^{n} D_\theta(\log f)(X_i, \boldsymbol{\theta}),\ s \in [\bar{s}, 1-\bar{s}]. \end{aligned} \tag{3.10}$$

The process Q_n^1 is defined on $[\bar{s}, 1]$ in order to include the last term of $\Lambda_{s,n}$ in (3.9) while Q_n^2 needs only to be defined on $[\bar{s}, 1-\bar{s}]$. In addition, we remark that the random processes $(Q_{s,n}^1)_{s\in[\bar{s},1-\bar{s}]}$ and Q_n^2 have a similar structure that differs only by the sub-sample considered. Therefore, in the following, we study the limit of Q_n^1 and simply extends the arguments to obtain the limit of Λ_n. However, before that, we need to establish some basic properties for the estimators $\hat{\theta}$, $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$ defined in Section 3.2.

3.3.1 The estimators $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$

Before proving the main result, we start with some preliminaries concerning the estimators $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$.

From Assumption 3.4, we already know that the estimator $\hat{\theta}$ converges almost surely to $\boldsymbol{\theta}$ when $n \to \infty$. With the following result inspired from Proposition 3.3 in Dehling et al. (2014), we can extend this convergence property to $\hat{\theta}_{0,s}$, $s \in [\bar{s}, 1]$, and to $\hat{\theta}_{s,1}$, $s \in [\bar{s}, 1-\bar{s}]$.

Lemma 3.3. *If a sequence $(u_n)_{n\geq 1} \subset \mathbb{R}^d$ converges to some finite limit u, then the sequence $u_{\lfloor sn\rfloor}$ converges to u, uniformly in $s \in [\bar{s}, 1]$.*

Proof. Fix $\epsilon > 0$. Let N such that for all $n \geq N$, $|u_n - u| \leq \epsilon$ and set $N' := \left\lfloor \frac{N}{\bar{s}} \right\rfloor + 1$. Then, for any $n \geq N'$, $[\bar{s}, 1] \subset [\frac{N}{n}, 1]$, thus, for any $n \geq N'$, $\lfloor ns\rfloor \geq N$ and, by the choice of N, $\left|u_{\lfloor sn\rfloor} - u\right| \leq \epsilon$. The result follows. □

Corollary 3.4. *If a sequence $(u_n)_{n\geq 1} \subset \mathbb{R}^d$ converges to some finite limit u, then the sequence $u_{\lfloor sn\rfloor}$ converges to u, uniformly in $s \in [\bar{s}, 1-\bar{s}]$.*

Reasoning ω by ω, this result implies directly that the almost sure convergence of $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$ is ω-wise uniform in s. This will represent a key property for the main result.

Proposition 3.5. *Under H_0 and Assumptions 3.1-3.4, the estimator $\hat{\theta}_{0,s}$ (resp. $\hat{\theta}_{s,1}$) converges almost surely to $\boldsymbol{\theta}$, uniformly in $s \in [\bar{s}, 1]$ (resp. in $s \in [\bar{s}, 1-\bar{s}]$).*

For n large enough, it is possible to obtain an explicit form for $\hat{\theta}_{0,s}$. Indeed, the estimator $\hat{\theta}$ is a sequence of solutions of the likelihood equations $D_\theta L(X, \theta) = 0$. Therefore, we can follow the ideas from the proof of the usual limit theorems for maximum likelihood estimators (see e.g. Theorem 1.5).

Corollary 3.6. *Under H_0 and Assumptions 3.1-3.4, almost surely, the estimator $\hat{\theta}$ exists for n large enough. Moreover, for almost all $\omega \in \Omega$, we can find some $N(\omega) \geq 1$ that does not depend on $s \in [\bar{s}, 1-\bar{s}]$ such that, for all $n \geq N(\omega)$, the three estimators $\hat{\theta}$, $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$ are respectively the unique solutions of the likelihood equations*

$$D_\theta L(X,\theta) = 0,\ D_\theta L((X_1,\dots,X_{\lfloor sn \rfloor}),\theta) = 0,\ \text{and } D_\theta L((X_{\lfloor sn \rfloor+1},\dots,X_n),\theta) = 0. \tag{3.11}$$

Proof. The existence of $\hat{\theta}$ follows from the almost sure convergence. Indeed, by assumption, Θ is an open convex subset of $\mathbb{R}^d$ in which θ_0 belongs. Then, we obtain that, for n large enough, $\hat{\theta}$ also belongs to this open convex set. The proof for the three estimators is a direct application of Proposition 3.5. □

Remark 3.7. *It is clear that the number N in Corollary 3.6 depends on ω. However, since, in this subsection, we always work ω by ω on the set of full probability where the three estimators converge, this will not pose any problem.*

In the sequel, the expression "for n large enough" will always implicitly imply that $\hat{\theta}$ belongs to Θ and solves (3.11). In particular, due to the regularity Assumption 3.2, the following Taylor expansion is well defined as soon as $\hat{\theta}$ belongs to Θ: for $1 \leq j \leq d$,

$$\begin{aligned} D_\theta L(X,\hat{\theta})_j =& D_\theta L(X,\boldsymbol{\theta})_j + \sum_{k=1}^{d} D^2_\theta L(X,\boldsymbol{\theta})_{j,k} \left(\hat{\theta}_k - \boldsymbol{\theta}_k\right) \\ &+ \frac{1}{2}\sum_{k=1}^{d}\sum_{l=1}^{d} \left(\hat{\theta}_l - \boldsymbol{\theta}_l\right) D^3_\theta L(X,\theta')_{j,k,l} \left(\hat{\theta}_k - \boldsymbol{\theta}_k\right) \end{aligned} \tag{3.12}$$

for some θ' on the segment $[\hat{\theta}, \boldsymbol{\theta}] \subset \mathbb{R}^d$.

Set

$$\hat{A} := -\frac{1}{n}\sum_{i=1}^{n}\left(D^2_\theta(\log f)(X_i,\boldsymbol{\theta}) \ +\ \frac{1}{2}\sum_{l=1}^{d}\left(\hat{\theta}_l - \boldsymbol{\theta}_l\right) D^3_\theta(\log f)(X_i,\theta')_{l,.,.}\right). \tag{3.13}$$

Because of (3.11), the left hand side of (3.12) vanishes. Thus, replacing $L(.,\boldsymbol{\theta})$ by its explicit expression, we get the equality between the two vectors

$$\hat{A}\left(\hat{\theta} - \boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} D_\theta(\log f)(X_i,\boldsymbol{\theta}).$$

The limit of $\hat{A}$, when n tends to infinity, is given by the next Proposition. We follow the standard proof of the usual limit theorems for maximum likelihood estimators (e.g. Theorem 1.5) and extend it to the almost sure convergence.

Proposition 3.8. *Under H_0 and Assumptions 3.2-3.4, the matrix $\hat{A}$ converges almost surely to the Fisher Information Matrix $\boldsymbol{I}$.*

Proof. Assumption 3.2 ensures that, for all $1 \leq j,k,l \leq d$ and almost all $x \in \mathcal{X}$, $|D^3_\theta L(x,\theta)_{j,k,l}|$ is bounded by $\kappa_1(x)$, uniformly in $\theta \in \Theta$. Thus, from the Strong Law of Large Numbers, for any $\theta \in \Theta$ and all $1 \leq j,k,l \leq d$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} D^3_\theta(\log f)(X_i,\theta)_{j,k,l}\right| \leq \frac{1}{n}\sum_{i=1}^{n}\kappa_1(X_i) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_{H_0}[\kappa_1(X_1)] < \infty.$$

Since, by Assumption 3.4, $\hat{\theta} \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\theta}$, it follows that

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{l=1}^{d}\left(\hat{\theta}_l - \boldsymbol{\theta}_l\right) D^3_\theta(\log f)(X_i,\theta')_{l,.,.} \xrightarrow[n\to\infty]{a.s.} 0.$$

Applying once more the Strong Law of Large Numbers and using Assumption 3.3, we get, for all $1 \le j,k \le d$,

$$\frac{1}{n}\sum_{i=1}^{n} D^2_\theta(\log f)(X_i,\boldsymbol{\theta})_{j,k} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_{H_0}\left[D^2_\theta(\log f)(X_1,\boldsymbol{\theta})\right] = -\boldsymbol{I}_{j,k}.$$

In conclusion, with (3.13), $\hat{A} \xrightarrow[n\to\infty]{a.s.} \boldsymbol{I}$. □

Corollary 3.9. *Almost surely, the inverse matrix $\hat{A}^{-1}$ exists for large n, and converges to the inverse Fisher Information Matrix $\boldsymbol{I}^{-1}$ as $n \to \infty$.*

Proof. It follows from Proposition 3.8 that $\det(\hat{A}) \xrightarrow[n\to\infty]{a.s.} \det(\boldsymbol{I})$. Now recall that, by Assumption 3.3, $\boldsymbol{I}$ is definite positive and, in particular $\det(\boldsymbol{I}) > 0$. It follows that, for n large enough, $\det(\hat{A}) \neq 0$ and $\hat{A}^{-1}$ exists. The result follows. □

In the same way as above, for any $s \in [\bar{s}, 1]$, there exists some point $\theta'_{0,s}$ on the segment $\left[\hat{\theta}_{0,s}, \boldsymbol{\theta}\right]$, such that the matrix

$$\hat{A}_{0,s} := -\frac{1}{\lfloor sn \rfloor}\sum_{i=1}^{\lfloor sn \rfloor}\left(D^2_\theta(\log f)(X_i,\boldsymbol{\theta}) + \frac{1}{2}\sum_{l=1}^{d}\left(\hat{\theta}_{0,s;l} - \boldsymbol{\theta}_l\right) D^3_\theta(\log f)(X_i,\theta'_{0,s})_{l,.,.}\right) \tag{3.14}$$

satisfies

$$\hat{A}_{0,s}(\hat{\theta}_{0,s} - \boldsymbol{\theta}) = \frac{1}{\lfloor sn \rfloor}\sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i,\boldsymbol{\theta}). \tag{3.15}$$

And, for any $s \in [\bar{s}, 1-\bar{s}]$, there exists some point $\theta'_{s,1}$ on the segment $\left[\hat{\theta}_{s,1}, \boldsymbol{\theta}\right]$, such that the matrix

$$\hat{A}_{s,1} := -\frac{1}{n - \lfloor sn \rfloor}\sum_{i=\lfloor sn \rfloor+1}^{n}\left(D^2_\theta(\log f)(X_i,\boldsymbol{\theta}) + \frac{1}{2}\sum_{l=1}^{d}\left(\hat{\theta}_{s,1;l} - \boldsymbol{\theta}_l\right) D^3_\theta(\log f)(X_i,\theta'_{s,1})_{l,.,.}\right)$$

satisfies

$$\hat{A}_{s,1}(\hat{\theta}_{s,1} - \boldsymbol{\theta}) = \frac{1}{n - \lfloor sn \rfloor}\sum_{i=\lfloor sn \rfloor+1}^{n} D_\theta(\log f)(X_i,\boldsymbol{\theta}).$$

To sum up, the following proposition provides an explicit expression for $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$ and the convergence of $\hat{A}_{0,s}$ and $\hat{A}_{s,1}$.

Proposition 3.10. *Under H_0 and Assumptions 3.1-3.4, almost surely, for n large enough,*

- ◇ *for all $s \in [\bar{s}, 1]$, the matrix $\hat{A}_{0,s}$ is invertible and*

$$\hat{\theta}_{0,s} - \boldsymbol{\theta} = \hat{A}^{-1}_{0,s}\frac{1}{\lfloor sn \rfloor}\sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i,\boldsymbol{\theta}), \tag{3.16}$$

with $\mathbb{E}_{H_0}\left[D_\theta(\log f)(X_1,\boldsymbol{\theta})\right] = 0$, and where $\hat{A}^{-1}_{0,s}$ converges almost surely to $\boldsymbol{I}^{-1}$, uniformly in $s \in [\bar{s}, 1]$,

◇ *for all* $s \in [\bar{s}, 1-\bar{s}]$ *the matrix* $\hat{A}_{s,1}$ *is invertible and*

$$\hat{\theta}_{s,1} - \boldsymbol{\theta} = \hat{A}_{s,1}^{-1} \frac{1}{n - \lfloor sn \rfloor} \sum_{i=\lfloor sn \rfloor + 1}^{n} D_\theta(\log f)(X_i, \boldsymbol{\theta})$$

with $\mathbb{E}_{H_0}\left[D_\theta(\log f)(X_1, \boldsymbol{\theta})\right] = 0$*, and where* $\hat{A}_{s,1}^{-1}$ *converges almost surely to* $\boldsymbol{I}^{-1}$*, uniformly in* $s \in [\bar{s}, 1-\bar{s}]$*.*

Proof. For n large enough, $\hat{A}_{0,s}^{-1}$ is well defined and Equation (3.16) follows directly from (3.15). Assumption 3.2 guarantees that the expectation $\mathbb{E}_{H_0}\left[D_\theta(\log f)(X_1, \boldsymbol{\theta})\right]$ vanishes. Finally we use Corollary 3.9 and Lemma 3.3 to obtain the almost sure convergence of $\hat{A}_{0,s}^{-1}$ to $\boldsymbol{I}^{-1}$, uniformly in $s \in [\bar{s}, 1]$.
The proof of the analogue result for $\hat{\theta}_{s,1}$ and $\hat{A}_{s,1}$ is the same. □

The explicit expression obtained for $\hat{\theta}_{0,s} - \boldsymbol{\theta}$ and $\hat{\theta}_{s,1} - \boldsymbol{\theta}$ already points out the direction of the next steps: since $\hat{A}_{0,s}^{-1}$ and $\hat{A}_{s,1}^{-1}$ converge almost surely to $\boldsymbol{I}^{-1}$, uniformly in s, and with $\mathbb{E}_{H_0}\left[D_\theta(\log f)(X_1, \boldsymbol{\theta})\right] = 0$, we will be able to establish Donsker-type result for $\hat{\theta}_{0,s} - \boldsymbol{\theta}$ and $\hat{\theta}_{s,1} - \boldsymbol{\theta}$. This can be used to derive a Donsker-type result for $Q_{s,n}^1$ and $Q_{s,n}^2$.

We will need the following variant of Glivenko-Cantelli's Theorem that exploits the almost sure convergence of $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$ to $\boldsymbol{\theta}$.

Lemma 3.11. *Consider an application* $h : (x, \theta) \in \mathcal{X} \times \Theta \mapsto h(x, \theta) \in \mathbb{R}$ *and a convex subset* $\mathcal{O}$ *of* Θ*, such that* $\boldsymbol{\theta}$ *is in the interior of* $\mathcal{O}$ *and*

1. *for almost all* $x \in \mathcal{X}$*, the application* $\theta \mapsto h(x, \theta)$ *is continuous on* $\mathcal{O}$*,*
2. *we can find some application* $\mathcal{X} \ni x \mapsto \kappa_3(x)$*, such that, for all* θ *in* $\mathcal{O}$*,* $|h(x, \theta)| \leq \kappa_3(x)$ *and* $\mathbb{E}_{H_0}[|\kappa_3(X_1)|] < \infty$*.*

Then, under H_0 *and Assumptions 3.1-3.4, one has* $\mathbb{E}_{H_0}\left[|h(X_1, \boldsymbol{\theta})|\right] < \infty$ *and*

◇ *for* $\theta'_{0,s} \in [\hat{\theta}_{0,s}, \boldsymbol{\theta}]$*,* $\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} h(X_i, \theta'_{0,s}) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_{H_0}\left[h(X_1, \boldsymbol{\theta})\right]$*, uniformly in* $s \in [\bar{s}, 1]$*,*

◇ *for* $\theta'_{s,1} \in [\hat{\theta}_{s,1}, \boldsymbol{\theta}]$*,* $\frac{1}{n - \lfloor sn \rfloor} \sum_{i=\lfloor sn \rfloor + 1}^{n} h(X_i, \theta'_{s,1}) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_{H_0}\left[h(X_1, \boldsymbol{\theta})\right]$*, uniformly in* $s \in [\bar{s}, 1-\bar{s}]$*.*

Proof. We only show the case $\theta'_{0,s} \in [\hat{\theta}_{0,s}, \boldsymbol{\theta}]$, $s \in [\bar{s}, 1]$. By the second condition of the lemma, $\boldsymbol{\theta} \in \mathcal{O}$ implies that $|h(x, \boldsymbol{\theta})| \leq \kappa_3(x)$ for all $x \in \mathcal{X}$, thus $\mathbb{E}_{H_0}\left[|h(X_1, \boldsymbol{\theta})|\right] \leq \mathbb{E}_{H_0}[|\kappa_3(X_1)|] < \infty$. Let us fix some $\epsilon > 0$ small enough so that, with $B(\boldsymbol{\theta}, \epsilon)$ the closed ball centered in $\boldsymbol{\theta}$ with radius ϵ, $B(\boldsymbol{\theta}, \epsilon) \cap \Theta$ is strictly contained in the subset $\mathcal{O}$. This is possible since, from the first condition of the lemma, $\mathcal{O}$ is a convex subset of Θ such that $\boldsymbol{\theta}$ is in the interior of $\mathcal{O}$. From Proposition 3.5, $\hat{\theta}_{0,s} \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\theta}$, uniformly in $s \in [\bar{s}, 1]$. Therefore, almost surely, we can find some $N \geq 1$ such that for all $n \geq N$ and for all $s \in [\bar{s}, 1]$, $\hat{\theta}_{0,s} \in B(\boldsymbol{\theta}, \epsilon) \cap \Theta$. Since $\theta'_{0,s}$ is a point on the segment $[\hat{\theta}_{0,s}, \boldsymbol{\theta}]$, it also belongs to $B(\boldsymbol{\theta}, \epsilon) \cap \Theta$. It follows that

$$\left| \frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} h(X_i, \theta'_{0,s}) - \mathbb{E}_{H_0}\left[h(X_1, \boldsymbol{\theta})\right] \right|$$

$$\leq \sup_{\theta \in B(\boldsymbol{\theta}, \epsilon) \cap \Theta} \left| \frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} h(X_i, \theta) - \mathbb{E}_{H_0}\left[h(X_1, \theta)\right] \right| + \left| \mathbb{E}_{H_0}\left[h(X_1, \theta'_{0,s})\right] - \mathbb{E}_{H_0}\left[h(X_1, \boldsymbol{\theta})\right] \right|. \tag{3.17}$$

With conditions 1. and 2., thanks to the dominated convergence theorem, the application $\theta \mapsto \mathbb{E}_{H_0}[h(X_1, \theta)]$ is continuous on $\mathcal{O}$. Since $\boldsymbol{\theta} \in \mathcal{O}$, $\hat{\theta}_{0,s} \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\theta}$, uniformly in $s \in [\bar{s}, 1]$ and, for all $n \geq 1$, $\theta'_{0,s} \in [\hat{\theta}_{0,s}, \boldsymbol{\theta}]$, we have also $\theta'_{0,s} \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\theta}$. Therefore the second term of the right hand side of (3.17) converges almost surely to 0, uniformly in $s \in [\bar{s}, 1]$, when $n \to \infty$.

To conclude the proof, we show that the first term also converges almost surely to 0, uniformly in $s \in [\bar{s}, 1]$, when $n \to \infty$. For all fixed $\theta \in B(\boldsymbol{\theta}, \epsilon) \cap \Theta$, the following convergence is an application of the Strong Law of Large Numbers:

$$Y_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} h(X_i, \theta) - \mathbb{E}_{H_0}[h(X_1, \theta)] \xrightarrow[n\to\infty]{a.s.} 0.$$

We deduce from assumptions 1. and 2. that $\theta \to Y_n(\theta)$ is continuous. Since $B(\boldsymbol{\theta}, \epsilon) \cap \Theta$ is compact, we get the convergence of the supremum:

$$\sup_{\theta \in B(\boldsymbol{\theta}, \epsilon) \cap \Theta} |Y_n(\theta)| \xrightarrow[n\to\infty]{a.s.} 0.$$

And finally we can conclude by Lemma 3.3. □

This Lemma concludes the collection of properties that are required for the estimators $\hat{\theta}_{0,s}$ and $\hat{\theta}_{s,1}$. In the following section, we focus on the process Q_n^1 and its limit distribution.

3.3.2 Limit distribution of Q_n^1

The process Q_n^1 is the first of the three terms defining the process Λ_n in (3.9). Let us recall here its expression: for $s \in [\bar{s}, 1]$,

$$\begin{aligned} Q_{s,n}^1 := &\sum_{i=1}^{\lfloor sn \rfloor} \Big(w(X_i, \hat{\theta}_{0,s}) \log f_1(X_i, \hat{\lambda}_{0,s,1}) - w(X_i, \boldsymbol{\theta}) \log f_1(X_i, \boldsymbol{\lambda}_1) \Big) \\ &- \boldsymbol{u}^T \boldsymbol{I}^{-1} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta}), \end{aligned}$$

with $\boldsymbol{u} := \mathbb{E}_{H_0}[D_\theta(w \log f_1)(X_1, \boldsymbol{\theta})]$.

We want to derive its limit distribution with a Donsker-type result. In the following, our purpose is to reorganize the terms so that the process $Q_{s,n}^1$ is somehow the product of a matrix that converges almost surely uniformly in $s \in [\bar{s}, 1]$ to some constant and of a vector that converges weakly to some random process. With the help of the Extended Slutsky's Theorem 1.13, one can target to derive the limit process of $Q_{s,n}^1$. The difficulty here is that the random variable $Q_{s,n}^1$ is not a sum of independent terms, because of the presence of the estimator $\hat{\theta}_{0,s}$, which depends itself on the whole sub-sample $(X_1, X_2, \ldots, X_{\lfloor sn \rfloor})$. Therefore, with Proposition 3.10, it seems logical to develop

$$\theta = (p_1, \ldots, p_{m-1}, \lambda_1, \ldots, \lambda_m) \mapsto w(x, \theta) \log f_1(x, \lambda_1)$$

around $\boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{m-1}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m)$. This is possible by Assumption 3.5 and because, by Proposition 3.5, $\hat{\theta}_{0,s}$ converges almost surely to $\boldsymbol{\theta}$, uniformly in $s \in [\bar{s}, 1]$, as $n \to \infty$.

It follows that, with a Taylor-Lagrange decomposition [10], almost surely, we can find some $N \geq 1$ (depending on ω) and some $\theta'_{0,s} \in [\hat{\theta}_{0,s}, \boldsymbol{\theta}]$ (depending on ω, n and s), such that, for $n \geq N$ and for $s \in [\bar{s}, 1]$,

$$\begin{aligned}
&\sum_{i=1}^{\lfloor sn \rfloor} \Big(w(X_i, \hat{\theta}_{0,s}) \log f_1(X_i, \hat{\lambda}_{0,s,1}) - w(X_i, \boldsymbol{\theta}) \log f_1(X_i, \boldsymbol{\lambda}_1) \Big) \\
&= \sum_{i=1}^{\lfloor sn \rfloor} \Big(D_\theta(w \log f_1)(X_i, \boldsymbol{\theta})^T (\hat{\theta}_{0,s} - \boldsymbol{\theta}) + (\hat{\theta}_{0,s} - \boldsymbol{\theta})^T D^2_\theta(w \log f_1)(X_i, \theta'_{0,s})(\hat{\theta}_{0,s} - \boldsymbol{\theta}) \Big).
\end{aligned}$$

So far, we can rewrite $Q^1_{s,n}$ as follows

$$\begin{aligned}
Q^1_{s,n} = &\left(\sum_{i=1}^{\lfloor sn \rfloor} D_\theta(w \log f_1)(X_i, \boldsymbol{\theta})^T \right) (\hat{\theta}_{0,s} - \boldsymbol{\theta}) \\
&+ (\hat{\theta}_{0,s} - \boldsymbol{\theta})^T \left(\sum_{i=1}^{\lfloor sn \rfloor} D^2_\theta(w \log f_1)(X_i, \theta'_{0,s}) \right) (\hat{\theta}_{0,s} - \boldsymbol{\theta}) \\
&- \boldsymbol{u}^T \boldsymbol{I}^{-1} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta}).
\end{aligned}$$

From Assumption 3.5, we know that $\mathbb{E}_{H_0}\left[|D_\theta(w \log f_1)(X_1, \boldsymbol{\theta})| \right] < \infty$. Thus we can center the right side sum of the first term by $\boldsymbol{u}$:

$$\begin{aligned}
Q^1_{s,n} = &\left(\sum_{i=1}^{\lfloor sn \rfloor} \Big(D_\theta(w \log f_1)(X_i, \boldsymbol{\theta})^T - \boldsymbol{u}^T \Big) \right) (\hat{\theta}_{0,s} - \boldsymbol{\theta}) \\
&+ (\hat{\theta}_{0,s} - \boldsymbol{\theta})^T \left(\sum_{i=1}^{\lfloor sn \rfloor} D^2_\theta(w \log f_1)(X_i, \theta'_{0,s}) \right) (\hat{\theta}_{0,s} - \boldsymbol{\theta}) \\
&+ \lfloor sn \rfloor \boldsymbol{u}^T (\hat{\theta}_{0,s} - \boldsymbol{\theta}) - \boldsymbol{u}^T \boldsymbol{I}^{-1} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta}).
\end{aligned}$$

Choosing n large enough, Proposition 3.10 provides us an explicit expression of $\hat{\theta}_{0,s} - \boldsymbol{\theta}$, which permits us to rewrite $Q^1_{s,n}$ as follows:

$$\begin{aligned}
Q^1_{s,n} = &\left(\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} \Big(D_\theta(w \log f_1)(X_i, \boldsymbol{\theta})^T - \boldsymbol{u}^T \Big) \right) \lfloor sn \rfloor \hat{A}^{-1}_{0,s} \left(\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta}) \right) \\
&+ \left(\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta})^T \right) \lfloor sn \rfloor \ {\hat{A}^{-1}_{0,s}}^T \left(\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D^2_\theta(w \log f_1)(X_i, \theta'_{0,s}) \right) \\
&\quad \hat{A}^{-1}_{0,s} \left(\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta}) \right) \\
&+ \boldsymbol{u}^T \left(\hat{A}^{-1}_{0,s} - \boldsymbol{I}^{-1} \right) \lfloor sn \rfloor \left(\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta}) \right). \qquad (3.18)
\end{aligned}$$

Remark that, in this reformulation of $Q^1_{s,n}$, we can recognize several sums of centered i.i.d random variables which, multiplied by $\sqrt{n}$, can be treated by Donsker's Theorem

10. See e.g. Theorem 5.3 in Coleman (2012).

and produce at the limit a multi-dimensional Brownian motion. Further, there are the variables $\hat{A}_{0,s}^{-1}$, which, by Proposition 3.10, converge a.s. to the inverse Fischer information $\boldsymbol{I}^{-1}$. The last line of (3.18), decomposing $\lfloor sn \rfloor$ into $\sqrt{n}\frac{\lfloor sn \rfloor}{n}\sqrt{n}$, makes appear the term $\sqrt{n}\left(\hat{A}_{0,s}^{-1} - \boldsymbol{I}^{-1}\right)$. Its limit, when n tends to ∞ and the link with the other components of $Q_{s,n}^1$ have to be analyzed separately, before we combine all these terms to compute the limit of the process $(Q_{s,n}^1)_{s\in[\bar{s},1]}$ in terms of a transformation of a Brownian motion.

Firstly, for all $s \in [\bar{s}, 1]$, we define the triple $\hat{\xi}_{0,s} := (\hat{\iota}_{0,s}, \hat{u}_{0,s} - \boldsymbol{u}, \hat{I}_{0,s}) \in (\mathbb{R}^d)^2 \times gl_d(\mathbb{R})$ by

$$\begin{aligned}
\hat{\iota}_{0,s} :=& \frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(\log f)(X_i, \boldsymbol{\theta}) \\
\hat{u}_{0,s} :=& \frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta(w \log f_1)(X_i, \boldsymbol{\theta}) \\
\hat{I}_{0,s} :=& -\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta^2(\log f)(X_i, \boldsymbol{\theta})
\end{aligned} \tag{3.19}$$

which allows us to rewrite $Q_{s,n}^1$ as

$$\begin{aligned}
Q_{s,n}^1 =& \frac{\lfloor sn \rfloor}{n} \sqrt{n}(\hat{u}_{0,s} - \boldsymbol{u})^T \, \hat{A}_{0,s}^{-1} \sqrt{n}\hat{\iota}_{0,s} \\
&+ \frac{\lfloor sn \rfloor}{n} \sqrt{n}\hat{\iota}_{0,s}^T \; \hat{A}_{0,s}^{-1\,T} \left(\frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta^2(w \log f_1)(X_i, \theta'_{0,s}) \right) \hat{A}_{0,s}^{-1} \sqrt{n}\hat{\iota}_{0,s} \\
&+ \frac{\lfloor sn \rfloor}{n} \boldsymbol{u}^T \sqrt{n} \left(\hat{A}_{0,s}^{-1} - \boldsymbol{I}^{-1} \right) \sqrt{n}\hat{\iota}_{0,s}.
\end{aligned} \tag{3.20}$$

Another crucial ingredient of our following discussion is the covariance matrix $\Sigma^2 \in gl_{2d+d^2}(\mathbb{R})$ under H_0 of the triple

$$\left(D_\theta(\log f)(X_1, \boldsymbol{\theta}), D_\theta(w \log f_1)(X_1, \boldsymbol{\theta}) - \boldsymbol{u}, -D_\theta^2(\log f)(X_1, \boldsymbol{\theta}) - \boldsymbol{I} \right)$$

reorganized in a $2d + d^2$-dimensional real vector. As a covariance matrix, Σ^2 is positive semi-definite. Then Σ will denote the unique positive semi-definite square root of Σ^2.

Lemma 3.12. *Set*

$$\boldsymbol{\xi} := (0, 0, \boldsymbol{I}) \in (\mathbb{R}^d)^2 \times GL_d(\mathbb{R}). \tag{3.21}$$

Under H_0, for all $s \in [\bar{s}, 1]$, $\mathbb{E}_{H_0}[\hat{\xi}_{0,s}] = \boldsymbol{\xi}$, the sequence of random variables $\hat{\xi}_{0,s}$ converges a.s. to $\boldsymbol{\xi}$, uniformly in $s \in [\bar{s}, 1]$, and the process $\sqrt{n}\left(\hat{\xi}_{0,s} - \boldsymbol{\xi}\right)_{s\in[\bar{s},1]}$ converges weakly in the Skorokhod metric space of càd-làg paths $\mathbb{D}_{[\bar{s},1]} := \mathbb{D}([\bar{s}, 1], (\mathbb{R}^d)^2 \times gl_d(\mathbb{R}))$, as follows

$$\sqrt{n}\left(\hat{\xi}_{0,s} - \boldsymbol{\xi}\right)_{s\in[\bar{s},1]} \xrightarrow[n\to\infty]{\mathcal{D}} \left(\frac{1}{s}\Sigma W_s\right)_{s\in[\bar{s},1]},$$

where $W := (W_s)_{s\in[0,1]}$ is a standard $2d + d^2$-dimensional Brownian motion and ΣW_s is reorganized as a triple in $(\mathbb{R}^d)^2 \times gl_d(\mathbb{R})$.

Proof. Under H_0, the random vector $\hat{\xi}_{0,s}$ is the sum of independent identically distributed random variables. In addition, we already know that

- ◇ from Assumptions 3.2 and 3.3, $\mathbb{E}_{H_0}\left[D_\theta(\log f)(X_1,\boldsymbol{\theta})\right]=0$ and $\mathbb{E}_{H_0}\left[-D^2_\theta(\log f)(X_1,\boldsymbol{\theta})\right]=\boldsymbol{I}$,
- ◇ by the expression for $\boldsymbol{u}$ in (3.8), $\mathbb{E}_{H_0}\left[D_\theta(w\log f_1)(X_1,\boldsymbol{\theta})\right]=\boldsymbol{u}$.

Therefore $\mathbb{E}_{H_0}[\hat{\xi}_{0,s}]=\boldsymbol{\xi}$, and the uniform a.s. convergence of the random variables $\hat{\xi}_{0,s}$ to $\boldsymbol{\xi}$ is a direct consequence of Lemma 3.11. The second part of the lemma follows then from Donsker's Theorem 1.10 and the Extended Slutsky's Theorem 1.13. □

The next Lemma will help us to handle the term $(\hat{A}^{-1}_{0,s}-\boldsymbol{I}^{-1})$ which appears in the last line of the expression (3.20) of $Q^1_{0,s}$.

To this aim we introduce the hypermatrix

$$\boldsymbol{J}:=\mathbb{E}_{H_0}\left[D^3_\theta(\log f)(X_1,\boldsymbol{\theta})\right]\in\mathbb{R}^{d\times d\times d}. \tag{3.22}$$

Lemma 3.13. *Almost surely, for large n (depending on ω), the variable $\hat{A}_{0,s}$ can be written as*

$$\hat{A}_{0,s}=\hat{I}_{0,s}-\frac{1}{2}\sum_{l=1}^{d}\left(\hat{\imath}^T_{0,s}\left(\hat{A}^{-1\,T}_{0,s}\right)_{.,l}\right)(\hat{J}_{0,s})_{.,.,l}, \tag{3.23}$$

for all $s\in[\bar{s},1]$, where $\hat{J}_{0,s}$ is the hypermatrix defined by

$$\hat{J}_{0,s}:=\frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor}D^3_\theta(\log f)(X_i,\theta'_{0,s}).$$

In addition, under H_0, almost surely, $\hat{J}_{0,s}$ converges to the hypermatrix $\boldsymbol{J}$, uniformly in $s\in[\bar{s},1]$.

Proof. Recall that the explicit expression of $\hat{A}_{0,s}$ in (3.14) depends itself on $\hat{\theta}_{0,s}-\boldsymbol{\theta}$, which, by Proposition 3.10, almost surely, can once more be replaced by

$$\hat{A}^{-1}_{0,s}\frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor}D_\theta(\log f)(X_i,\boldsymbol{\theta})$$

for n large enough. This gives:

$$\begin{aligned}\hat{A}_{0,s}=&\left(-\frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor}D^2_\theta(\log f)(X_i,\boldsymbol{\theta})\right)\\&-\frac{1}{2}\sum_{l=1}^{d}\left(\left(\frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor}D_\theta(\log f)(X_i,\boldsymbol{\theta})^T\right)\left(\hat{A}^{-1\,T}_{0,s}\right)_{.,l}\right)\left(\frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor}D^3_\theta(\log f)(X_i,\theta'_{0,s})_{l,.,.}\right).\end{aligned}$$

The result given in (3.23) follows. With Assumption 3.2, the convergence of $\hat{J}_{0,s}$ is a direct application of Lemma 3.11, taking the parameter set Θ as $\mathcal{O}$. □

Set now, for all $s\in[\bar{s},1]$ and all $n\geq 1$,

$$\hat{\xi}'_{0,s}:=(\hat{\imath}_{0,s},\hat{u}_{0,s}-\boldsymbol{u},\hat{A}^{-1}_{0,s}). \tag{3.24}$$

It is easy to see with Proposition 3.10 that, when n tends to infinity, $\hat{\xi}'_{0,s}$ converges a.s. to

$$\boldsymbol{\xi}' := (0, 0, \boldsymbol{I}^{-1}).$$

The aim of the next theorem is to establish a Donsker-type result for the process $(\hat{\xi}'_{0,s})_{s\in[\bar{s},1]}$.

Theorem 3.14. *Under H_0, the process $\sqrt{n}(\hat{\xi}'_{0,s} - \boldsymbol{\xi}')_{s\in[\bar{s},1]}$ converges weakly to the process $(\frac{1}{s}\boldsymbol{g}(\Sigma W_s))_{s\in[\bar{s},1]}$ in $\mathbb{D}_{[\bar{s},1]}$, where $\boldsymbol{g}$ is the linear map defined for $(\iota, u, I) \in (\mathbb{R}^d)^2 \times gl_d(\mathbb{R})$ by*

$$\boldsymbol{g}(\iota, u, I) := \left(\iota, u, -\boldsymbol{I}^{-1}\left(I - \frac{1}{2}\sum_{l=1}^{d}\left(\iota^T\left(\boldsymbol{I}^{-1\,T}\right)_{.,l}\right)\boldsymbol{J}_{.,.,l}\right)\boldsymbol{I}^{-1}\right), \tag{3.25}$$

and ΣW_s is reorganized as a triple in $(\mathbb{R}^d)^2 \times gl_d(\mathbb{R})$.

Proof. (i) Recall that, by Lemma 3.12, $\sqrt{n}\left(\hat{\xi}_{0,s} - \boldsymbol{\xi}\right)_{s\in[\bar{s},1]}$ converges weakly to $(\frac{1}{s}\Sigma W_s)_{s\in[\bar{s},1]}$ in $\mathbb{D}_{[\bar{s},1]}$, while, by Proposition 3.10 and Lemma 3.13, the couple $(\hat{A}^{-1}_{0,s}, \hat{J}_{0,s})$ converges a.s. to $(\boldsymbol{I}^{-1}, \boldsymbol{J})$ uniformly in $s \in [\bar{s}, 1]$. It follows by the Extended Slutsky's Theorem 1.13 that the random process $(\sqrt{n}(\hat{\xi}_{0,s} - \boldsymbol{\xi}), \hat{A}^{-1}_{0,s}, \hat{J}_{0,s})_{s\in[\bar{s},1]}$ converges weakly to $(\frac{1}{s}\Sigma W_s, \boldsymbol{I}^{-1}, \boldsymbol{J})_{s\in[\bar{s},1]}$ in $\mathbb{D}([\bar{s}, 1], (\mathbb{R}^d)^2 \times gl_d(\mathbb{R}) \times GL_d(\mathbb{R}) \times \mathbb{R}^{d\times d\times d})$.

(ii) Using Lemma 3.13, almost surely for n large enough (i.e. n depends on ω but not on s), we can write

$$\hat{\xi}'_{0,s} = \varphi \circ g(\hat{\xi}_{0,s}; \hat{A}^{-1}_{0,s}, \hat{J}_{0,s}),\ s \in [\bar{s}, 1],$$

where, for all $(\iota, u, I; A, J) \in (\mathbb{R}^d)^2 \times (gl_d(\mathbb{R}))^2 \times \mathbb{R}^{d\times d\times d}$, we define

$$g(\iota, u, I; A, J) := \left(\iota, u, I - \frac{1}{2}\sum_{l=1}^{d}\left(\iota^T(A^T)_{.,l}\right)J_{.,.,l}\right), \tag{3.26}$$

and, for $(\iota, u, I) \in (\mathbb{R}^d)^2 \times GL_d(\mathbb{R})$,

$$\varphi(\iota, u, I) := (\iota, u, I^{-1}).$$

Remark that $g(0, u, I; A, J) = (0, u, I)$ for all $(u, I; A, J) \in \mathbb{R}^d \times (gl_d(\mathbb{R}))^2 \times \mathbb{R}^{d\times d\times d}$. In particular,

$$\boldsymbol{\xi} = g(\boldsymbol{\xi}; \hat{A}^{-1}_{0,s}, \hat{J}_{0,s})$$

because $\boldsymbol{\xi} = (0, 0, \boldsymbol{I})$ by definition in (3.21). Since $g(\iota, u, I; A, J)$ is linear in (ι, u, I), then the following equality holds for each $s \in [\bar{s}, 1]$:

$$\sqrt{n}(g(\hat{\xi}_{0,s}; \hat{A}^{-1}_{0,s}, \hat{J}_{0,s}) - \boldsymbol{\xi}) = g(\sqrt{n}(\hat{\xi}_{0,s} - \boldsymbol{\xi}), \hat{A}^{-1}_{0,s}, \hat{J}_{0,s}).$$

By (i) and the Continuous Mapping Theorem 1.11, the process

$$\sqrt{n}(g(\hat{\xi}_{0,s}; \hat{A}^{-1}_{0,s}, \hat{J}_{0,s}) - \boldsymbol{\xi})_{s\in[\bar{s},1]}$$

converges weakly to $g(\frac{1}{s}\Sigma W_s; \boldsymbol{I}^{-1}, \boldsymbol{J})_{s\in[\bar{s},1]}$.

(iii) By Lemma 3.12, the sequence of random variables $\hat{\xi}_{0,s}$ converges a.s. to $\boldsymbol{\xi}$, uniformly in $s \in [\bar{s}, 1]$. With (i), the triple $(\hat{\xi}_{0,s}, \hat{A}^{-1}_{0,s}, \hat{J}_{0,s})$ also converges a.s. to $(\boldsymbol{\xi}, \boldsymbol{I}^{-1}, \boldsymbol{J})$,

uniformly in $s \in [\bar{s}, 1]$. Then, again by Theorem 1.11, $g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s})$ converges a.s. to $\boldsymbol{\xi}$, uniformly in $s \in [\bar{s}, 1]$.

(iv) Remark that, from Proposition 3.10 and Lemma 3.13, $g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s})$ is well defined when $\hat{A}_{0,s}$ is invertible. That is the case almost surely for n large enough, uniformly in $s \in [\bar{s}, 1]$: i.e. n depends on ω but not on s.

From Assumption 3.3, $\boldsymbol{I}$ is positive definite with finite components. So $\boldsymbol{I}^{-1}$ is also positive definite and $0 < \|\boldsymbol{I}^{-1}\|_2^{-1} < \infty$. Fix some $0 < r < \|\boldsymbol{I}^{-1}\|_2^{-1}$ such that the closed ball $B(\boldsymbol{\xi}, r)$ centered in $\boldsymbol{\xi}$ with radius r is included in $(\mathbb{R}^d)^2 \times GL_d(\mathbb{R})$. With (iii), we see that, almost surely, the following holds for n large enough, uniformly in $s \in [\bar{s}, 1]$:

$$\hat{A}_{0,s} \text{ is invertible and } \left\| g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s}) - \boldsymbol{\xi} \right\|_2 < r. \tag{3.27}$$

Let $\hat{\xi}''_{0,.}$ denote the process on $[\bar{s}, 1]$ defined for all $\omega \in \Omega$, all $n \geq 1$ and all $s \in [\bar{s}, 1]$ as follows:

$$\hat{\xi}''_{0,s}(\omega) := \begin{cases} g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s})(\omega) & \text{if (3.27) holds,} \\ \boldsymbol{\xi} & \text{otherwise.} \end{cases}$$

Then, almost surely, $\sqrt{n}(\hat{\xi}''_{0,s} - g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s}))$ is equal to 0 for n large enough, uniformly in $s \in [\bar{s}, 1]$. We denote by $\boldsymbol{\xi}_.$ the constant process such that $\boldsymbol{\xi}_s = \boldsymbol{\xi}$ for all $s \in [\bar{s}, 1]$. Therefore, by the conclusion from (ii) and Theorems 1.13 and 1.11,

$$\sqrt{n}(\hat{\xi}''_{0,.} - \boldsymbol{\xi}_.) = \sqrt{n}(g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s})_{s\in[\bar{s},1]} - \boldsymbol{\xi}_.) + \sqrt{n}(\hat{\xi}''_{0,.} - g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s})_{s\in[\bar{s},1]})$$

converges weakly to $g(\frac{1}{s}\Sigma W_s; \boldsymbol{I}^{-1}, \boldsymbol{J})_{s\in[\bar{s},1]}$.

(v) Let us denote by Φ the function from $\mathbb{D}([\bar{s}, 1], (\mathbb{R}^d)^2 \times GL_d(\mathbb{R})) \subset \mathbb{D}_{[\bar{s},1]}$ onto itself defined by:

$$\Phi(\zeta)_s := \varphi(\zeta_s), \ s \in [\bar{s}, 1]. \tag{3.28}$$

It follows from the definition of $\hat{\xi}'_{0,s}$ in (3.24) that

$$\begin{aligned} &\sqrt{n}(\hat{\xi}'_{0,s} - \boldsymbol{\xi}')_{s\in[\bar{s},1]} \\ &= \sqrt{n}(\Phi(\hat{\xi}''_{0,.}) - \Phi(\boldsymbol{\xi}_.)) + \sqrt{n}(\Phi(g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s})_{s\in[\bar{s},1]}) - \Phi(\hat{\xi}''_{0,.})). \end{aligned} \tag{3.29}$$

With (iii) and (iv), almost surely, $\sqrt{n}(\varphi(g(\hat{\xi}_{0,s}; \hat{A}_{0,s}^{-1}, \hat{J}_{0,s})) - \varphi(\hat{\xi}''_{0,s}))$ is equal to 0 for n large enough, uniformly in $s \in [\bar{s}, 1]$. Once more, by Theorems 1.13 and 1.11, both processes $\sqrt{n}(\hat{\xi}'_{0,s} - \boldsymbol{\xi}')_{s\in[\bar{s},1]}$ and $\sqrt{n}(\Phi(\hat{\xi}''_{0,.}) - \Phi(\boldsymbol{\xi}_.))$ have the same limit distribution.

The remainder of the proof is based on Corollary 1.16, a functional delta method in the Skorokhod metric space. This result, applied to the map Φ, would conclude the proof and provide the limit distribution of $\sqrt{n}(\Phi(\hat{\xi}''_{0,.}) - \Phi(\boldsymbol{\xi}_.))$.

The first condition of Corollary 1.16 holds by (iv) since $\sqrt{n}(\hat{\xi}''_{0,.} - \boldsymbol{\xi}_.)$ converges weakly to $g(\frac{1}{s}\Sigma W_s; \boldsymbol{I}^{-1}, \boldsymbol{J})_{s\in[\bar{s},1]}$.

To conclude the proof, it now sufficient to show that the second condition of Corollary 1.16 also holds. For that purpose, we start by noticing that, by (iv), for all $n \geq 1$, the

process $\sqrt{n}(\hat{\xi}''_{0,.} - \boldsymbol{\xi}_.)$ is in the closed ball

$$B(0, r\sqrt{n}) := \{\zeta \in \mathbb{D}_{[\bar{s},1]}, \|\zeta\|_2 \leq r\sqrt{n}\},$$

where 0 is the null function on $[\bar{s}, 1]$. Let us consider the sequence of applications γ_n defined for ζ_n in $B(0, r\sqrt{n}) \subset \mathbb{D}_{[\bar{s},1]}$ by

$$\gamma_n(\zeta_n) := \sqrt{n}\left(\Phi\left(\boldsymbol{\xi}_. + \frac{1}{\sqrt{n}}\zeta_n\right) - \Phi(\boldsymbol{\xi}_.)\right).$$

Further denote the differential of Φ at $\boldsymbol{\xi}_.$ by $D\Phi$. The differential is a function from $\mathbb{D}_{[\bar{s},1]}$ onto itself, defined for $\zeta = (\zeta_1, \zeta_2, \zeta_3)$ in $\mathbb{D}_{[\bar{s},1]}$ by [11]

$$D\Phi(\zeta)_s := (\zeta_1(s), \zeta_2(s), -\boldsymbol{I}^{-1}\zeta_3(s)\boldsymbol{I}^{-1}),\ s \in [\bar{s}, 1].$$

It is then sufficient to show that the convergence of every sequence $\zeta_n \in B(0, r\sqrt{n})$ to $\zeta \in \mathbb{D}_{[\bar{s},1]}$ implies the convergence of $\gamma_n(\zeta_n)$ to $D\Phi(\zeta)$.

Let us consider some sequence $\zeta_n = (\zeta_{1,n}, \zeta_{2,n}, \zeta_{3,n}) \in B(0, r\sqrt{n})$ and some path $\zeta = (\zeta_1, \zeta_2, \zeta_3) \in \mathbb{D}_{[\bar{s},1]}$ such that $d_{\mathbb{D}_{[\bar{s},1]}}(\zeta_n, \zeta) \to 0$ as $n \to \infty$. From the definition of the Skorokhod metric in (3.7), this means [12] that there exists some sequence of strictly increasing bijections τ_n^* from $[\bar{s}, 1]$ onto itself such that, as $n \to \infty$,

$$\sup_{s\in[\bar{s},1]} |\tau_n^*(s) - s| \to 0 \quad \text{and} \quad \sup_{s\in[\bar{s},1]} \|\zeta_n(\tau_n^*(s)) - \zeta(s)\|_2 \to 0. \tag{3.30}$$

To conclude the proof, we need only to show that $\sup_s \|\gamma_n(\zeta_n(\tau_n^*(.))_s) - D\Phi(\zeta)_s\|_2 \to 0$. First, we remark that

$$\gamma_n(\zeta_n(\tau_n^*(s))_{s\in[\bar{s},1]}) = \left(\zeta_{1,n}(\tau_n^*(s)), \zeta_{2,n}(\tau_n^*(s)), \sqrt{n}\left(\left(\boldsymbol{I} + \frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^{-1} - \boldsymbol{I}^{-1}\right)\right)_{s\in[\bar{s},1]}$$

and $D\Phi(\zeta) = (\zeta_1(s), \zeta_2(s), -\boldsymbol{I}^{-1}\zeta_3(s)\boldsymbol{I}^{-1})_{s\in[\bar{s},1]}$. By (3.30) and the definition of $\|.\|_2$ in Section 3.2.3, it is sufficient to show that $\sup_s \|\zeta_{3,n}(\tau_n^*(s)) - \zeta_3(s)\|_2 \to 0$ implies

$$\sup_s \left\|\sqrt{n}\left(\left(\boldsymbol{I} + \frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^{-1} - \boldsymbol{I}^{-1}\right) + \boldsymbol{I}^{-1}\zeta_3(s)\boldsymbol{I}^{-1}\right\|_2 \to 0.$$

Because τ_n^* is a bijection from $[\bar{s}, 1]$ onto itself and $\zeta_n = (\zeta_{1,n}, \zeta_{2,n}, \zeta_{3,n})$ is in $B(0, r\sqrt{n})$, we obtain that, for all $s \in [\bar{s}, 1]$ and $n \geq 1$, the random variable $\frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}$ is in the closed ball $B(0, r)$. In addition, we chose r such that $\|\boldsymbol{I}^{-1}\|_2 < 1/r$. Because the Frobenius norm is submultiplicative [13], it follows that

$$\left\|-\boldsymbol{I}^{-1}\frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right\|_2 < 1.$$

11. We extend here the well known differential of the inversion of matrices given in Theorem 1.17.
12. See Section 12 in Billingsley (1999) for more details on the Skorokhod topology.
13. This property is a consequence of the Cauchy-Schwarz inequality. See e.g. Trefethen and Bau (1997), p.23.

Therefore, using Theorem 1.18, we can expand the term $\left(\boldsymbol{I} + \frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^{-1}$ as a Neumann series. We obtain

$$\left(\boldsymbol{I} + \frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^{-1} = \left(Id_d + \boldsymbol{I}^{-1}\frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^{-1}\boldsymbol{I}^{-1} = \left(\sum_{k\geq 0}(-1)^k \boldsymbol{I}^{-k}\left(\frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^k\right)\boldsymbol{I}^{-1}.$$

For all $s \in [\bar{s}, 1]$ and $n \geq 1$,

$$\begin{aligned}
&\sqrt{n}\left(\left(\boldsymbol{I} + \frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^{-1} - \boldsymbol{I}^{-1}\right) + \boldsymbol{I}^{-1}\zeta_3(s)\boldsymbol{I}^{-1} \\
=&\sqrt{n}\left(\sum_{k\geq 2}(-1)^k \boldsymbol{I}^{-k}\left(\frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^k\right)\boldsymbol{I}^{-1} - \boldsymbol{I}^{-1}\Big(\zeta_{3,n}(\tau_n^*(s)) - \zeta_3(s)\Big)\boldsymbol{I}^{-1}.
\end{aligned}$$

The result follows from the fact that $\sup_s \left\|\zeta_{3,n}(\tau_n^*(s)) - \zeta_3(s)\right\|_2 \to 0$ and

$$\begin{aligned}
&\sup_s \left\|\sqrt{n}\left(\sum_{k\geq 2}(-1)^k \boldsymbol{I}^{-k}\left(\frac{\zeta_{3,n}(\tau_n^*(s))}{\sqrt{n}}\right)^k\right)\boldsymbol{I}^{-1}\right\|_2 \\
\leq&\frac{1}{\sqrt{n}}\left\|\boldsymbol{I}^{-1}\right\|_2^3 \sup_s \left\|\zeta_{3,n}(\tau_n^*(s))\right\|_2^2 \sum_{k\geq 0}\left(\frac{\left\|\boldsymbol{I}^{-1}\right\|_2}{\sqrt{n}}\sup_s \left\|\zeta_{3,n}(\tau_n^*(s))\right\|_2\right)^k. \qquad (3.31)
\end{aligned}$$

We already know that the constant $\left\|\boldsymbol{I}^{-1}\right\|_2$ is finite. From Lemma 12.1 in Billingsley (1999), $\|\zeta_3\|_2$ is a finite constant since ζ is a càd-làg process with finite values in $\mathbb{D}_{[\bar{s},1]}$. It follows that $\sup_s \left\|\zeta_{3,n}(\tau_n^*(s))\right\|_2 \leq \|\zeta_3\|_2 + \sup_s \left\|\zeta_{3,n}(\tau_n^*(s)) - \zeta_3(s)\right\|_2$ converges to $\|\zeta_3\|_2$ and, for n large enough, the series above can be dominated by a convergent geometric series. Then (3.31) converges to 0 and the result follows. □

The main theorem states that the process Q_n^1 converges to a quadratic form of the Brownian motion W introduced in Theorem 3.14. For $(\iota, u, I; A, J) \in (\mathbb{R}^d)^2 \times (gl_d(\mathbb{R}))^3$, we set first:

$$q(\iota, u, I; A, J) := u^T A\iota + \iota^T A^T J A\iota + \boldsymbol{u}^T I\iota \in \mathbb{R}, \qquad (3.32)$$

where we recall the vector $\boldsymbol{u} = \mathbb{E}_{H_0}\left[D_\theta(w\log f_1)(X_1, \boldsymbol{\theta})\right]$ defined in (3.8). We also consider the application $\boldsymbol{q}$ defined as follows:

$$\boldsymbol{q}(z) := q(\boldsymbol{g}(\Sigma z); \boldsymbol{I}^{-1}, \boldsymbol{U}),\ z \in \mathbb{R}^{2d+d^2}, \qquad (3.33)$$

where $\boldsymbol{g}$ is defined by (3.25), Σz is reorganized as a triple in $(\mathbb{R}^d)^2 \times gl_d(\mathbb{R})$. Further $\boldsymbol{U}$ is the matrix defined by

$$\boldsymbol{U} := \mathbb{E}_{H_0}\left[D_\theta^2(w\log f_1)(X_1, \boldsymbol{\theta})\right]. \qquad (3.34)$$

Remark 3.15. *The map $z \in \mathbb{R}^{2d+d^2} \mapsto \boldsymbol{q}(z)$ is a quadratic form.*

Proof. Remark that, for any A and J, the map $(\iota, u, I) \mapsto q(\iota, u, I; A, J)$ is a quadratic form but not a norm. Its unique associated symmetric bilinear form[14] is given by $\Big((\iota, u, I), (\iota', u', I')\Big) \mapsto \frac{1}{2}\Big(q(\iota + \iota', u + u', I + I'; A, J) - q(\iota, u, I; A, J) - q(\iota, u, I; A, J)\Big)$. From Theorem 3.14, we already know that the map $\boldsymbol{g}$ is linear. It follows that the map

$$\begin{aligned}
(z, z') \mapsto&\frac{1}{2}\left(\boldsymbol{q}(z + z') - \boldsymbol{q}(z) - \boldsymbol{q}(z')\right) \\
=&\frac{1}{2}\Big(q(\boldsymbol{g}(\Sigma(z + z')); \boldsymbol{I}^{-1}, \boldsymbol{U}) - q(\boldsymbol{g}(\Sigma z); \boldsymbol{I}^{-1}, \boldsymbol{U}) - q(\boldsymbol{g}(\Sigma z'); \boldsymbol{I}^{-1}, \boldsymbol{U})\Big)
\end{aligned}$$

is symmetric bilinear. The result follows. □

14. See for instance Section 41 in O'Meara (2000).

We can now state our main result for Q_n^1.

Theorem 3.16. *Under H_0, the process Q_n^1 converges weakly as $n \to \infty$ to the process $(\frac{1}{s}\boldsymbol{q}(W_s))_{s\in[\bar{s},1]}$ in $\mathbb{D}([\bar{s},1],\mathbb{R})$.*

Proof. For all $s \in [\bar{s},1]$ and n large enough, from (3.20), we can reorganize the variable $Q_{s,n}^1$ as follows

$$\begin{aligned} Q_{s,n}^1 =& \frac{\lfloor sn \rfloor}{n}\Big(\sqrt{n}(\hat{u}_{0,s} - \boldsymbol{u})^T \, \hat{A}_{0,s}^{-1} \, \sqrt{n}\hat{\imath}_{0,s} + \sqrt{n}\hat{\imath}_{0,s}^T \;\; {A_{0,s}^{-1}}^T \hat{U}_{0,s} \hat{A}_{0,s}^{-1} \, \sqrt{n}\hat{\imath}_{0,s} \\ & + \boldsymbol{u}^T \sqrt{n} \left(\hat{A}^{-1} - \boldsymbol{I}^{-1} \right) \, \sqrt{n}\hat{\imath}_{0,s} \Big) \\ =& \frac{\lfloor sn \rfloor}{n} q(\sqrt{n}(\hat{\xi}'_{0,s} - \boldsymbol{\xi}'); \hat{A}_{0,s}^{-1}, \hat{U}_{0,s}), \end{aligned} \tag{3.35}$$

with

$$\hat{U}_{0,s} := \frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta^2 (w \log f_1)(X_i, \theta'_{0,s})$$

and $\hat{\xi}'_{0,s} = (\hat{\imath}_{0,s}, \hat{u}_{0,s} - \boldsymbol{u}, \hat{A}_{0,s}^{-1})$ from its definition in (3.24).
We know from Theorem 3.14 that the process $(\sqrt{n}(\hat{\xi}'_{0,s} - \boldsymbol{\xi}'))_{s\in[\bar{s},1]}$ converges weakly to the process $(\frac{1}{s}\boldsymbol{g}(\Sigma W_s))_{s\in[\bar{s},1]}$. Recall that, by Proposition 3.10, $\hat{A}_{0,s}^{-1}$ converges a.s. to $\boldsymbol{I}^{-1}$, uniformly in $s \in [\bar{s},1]$. Further, by Assumption 3.5, Lemma 3.11 can be applied to $h(x,\theta) = D_\theta^2(w \log f_1)(x,\theta)$, taking the set Θ' as $\mathcal{O}$. Therefore $\hat{U}_{0,s}$ converges a.s. to $\boldsymbol{U}$, uniformly in $s \in [\bar{s},1]$. It follows by the Extended Slutsky's Theorem 1.13, that the process

$$\left(\sqrt{n}(\hat{\xi}'_{0,s} - \boldsymbol{\xi}'), \hat{A}_{0,s}^{-1}, \hat{U}_{0,s} \right)_{s\in[\bar{s},1]}$$

converges weakly to the process $(\frac{1}{s}\boldsymbol{g}(\Sigma W_s), \boldsymbol{I}^{-1}, \boldsymbol{U})_{s\in[\bar{s},1]}$. The map q being continuous, the result follows by the Continuous Mapping Theorem 1.11. □

3.3.3 Limit distribution of the test statistic

The limit distribution of Λ_n is obtain as an extension of the result in Theorem 3.16 for Q_n^1. In the following Theorem we derive the limit distribution of the test statistic S_n.

Theorem 3.17. *Under H_0 and Assumptions 3.1-3.5, the test statistic*

$$S_n \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}(W_s - sW_1)}{s(1-s)}$$

where $(W_s)_{s\in[0,1]}$ is a standard $2d+d^2$-dimensional Brownian motion and the application $\boldsymbol{q}$ is defined in (3.33).

Proof. Recall that from (3.29) and (3.35), for $s \in [\bar{s},1]$, $Q_{s,n}^1$ can be written as follows for $s \in [\bar{s},1]$:

$$Q_{s,n}^1 = \frac{\lfloor sn \rfloor}{n} q\Big(\sqrt{n}\Big(\Phi(g(\hat{\xi}_{0,t}; \hat{A}_{0,t}^{-1}, \hat{J}_{0,t})_{t\in[\bar{s},1]}) - \Phi(\boldsymbol{\xi}_\cdot) \Big)_s ; \hat{A}_{0,s}^{-1}, \hat{U}_{0,s} \Big).$$

We remark from (3.9) that $Q_{s,n}^1$ and $Q_{s,n}^2$ have a similar structure and differ only from the fact that $Q_{s,n}^1$ depends from the sample $(X_1, \ldots, X_{\lfloor sn \rfloor})$ and the estimator $\hat{\theta}_{0,s}$,

while $Q^2_{s,n}$ depends from the sample $(X_{\lfloor sn\rfloor+1},\dots,X_n)$ and the estimator $\hat\theta_{1,s}$. With the definition of $\hat\xi_{0,s}$ in (3.19), we can write $Q^2_{s,n}$ for $s\in[\bar s,1-\bar s]$ as follows:

$$Q^2_{s,n}=\frac{n-\lfloor sn\rfloor}{n}q\left(\sqrt{n}\left(\tilde\Phi\left(g\left(\frac{n}{n-\lfloor tn\rfloor}\hat\xi_{0,1}-\frac{\lfloor tn\rfloor}{n-\lfloor tn\rfloor}\hat\xi_{0,t};\hat A^{-1}_{t,1},\hat J_{t,1}\right)_{t\in[\bar s,1-\bar s]}\right)-\tilde\Phi((\boldsymbol\xi)_{s\in[\bar s,1-\bar s]})\right)_s;\hat A^{-1}_{s,1},\hat U_{s,1}\right)$$

where $\tilde\Phi$ is the map from the set of càd-làg paths $\mathbb{D}([\bar s,1-\bar s],(\mathbb{R}^d)^2\times GL_d(\mathbb{R}))$ onto itself that coincide with Φ on $[\bar s,1-\bar s]$, i.e. for $s\in[\bar s,1-\bar s]$ and $x\in\mathbb{D}([\bar s,1-\bar s],(\mathbb{R}^d)^2\times GL_d(\mathbb{R}))$, $\tilde\Phi(x)_s:=\varphi(x_s)$. In addition, the random variable $Q^1_{1,n}$ can be written as:

$$Q^1_{1,n}=q\Big(\sqrt{n}\Big(\varphi(g(\hat\xi_{0,1};\hat A^{-1},\hat J_{0,1}))-\varphi(\boldsymbol\xi)\Big);\hat A^{-1},\hat U_{0,1}\Big).$$

From the three equations above, the process $(Q^1_{s,n},Q^2_{s,n},-Q^1_{1,n})_{s\in[\bar s,1-\bar s]}$ can be seen as a function of the triple process

$$\left(\hat\xi_{0,s},\frac{n}{n-\lfloor sn\rfloor}\hat\xi_{0,1}-\frac{\lfloor sn\rfloor}{n-\lfloor sn\rfloor}\hat\xi_{0,s},\hat\xi_{0,1}\right)_{s\in[\bar s,1-\bar s]}. \tag{3.36}$$

Recall that, by Lemma 3.12, the process $\sqrt{n}\left(\hat\xi_{0,s}-\boldsymbol\xi\right)_{s\in[\bar s,1]}$ converges weakly to $(\frac{1}{s}\Sigma W_s)_{s\in[\bar s,1]}$ in $\mathbb{D}_{[\bar s,1]}$. Then a similar central limit result holds for the triple defined in (3.36) and, by a succession of composition of the applications g, Φ and q along with arguments based on the Extended Slutsky's Theorem 1.13 and the Continuous Mapping Theorem 1.11, the result obtained for Q^1_n is extended to the process $(Q^1_{s,n},Q^2_{s,n},-Q^1_{1,n})_{s\in[\bar s,1-\bar s]}$. To show this, we reuse the arguments of Theorem 3.14. The functional delta method on the Skorokhod metric space is still applicable and the process

$$\left(\begin{array}{l}\sqrt{n}(\Phi(g(\hat\xi_{0,t};\hat A^{-1}_{0,t},\hat J_{0,t})_{t\in[\bar s,1]})-\Phi(\boldsymbol\xi_.))_s,\\ \sqrt{n}\left(\tilde\Phi\left(g\left(\frac{n}{n-\lfloor tn\rfloor}\hat\xi_{0,1}-\frac{\lfloor tn\rfloor}{n-\lfloor tn\rfloor}\hat\xi_{0,t};\hat A^{-1}_{t,1},\hat J_{t,1}\right)_{t\in[\bar s,1-\bar s]}\right)-\tilde\Phi((\boldsymbol\xi)_{t\in[\bar s,1-\bar s]})\right)_s\\ \sqrt{n}(\varphi(g(\hat\xi_{0,1};\hat A^{-1},\hat J_{0,1}))-\varphi(\boldsymbol\xi))\end{array}\right)_{s\in[\bar s,1-\bar s]}$$

converges weakly to the process

$$\left(\boldsymbol g\left(\frac{1}{s}\Sigma W_s\right),\boldsymbol g\left(\frac{1}{1-s}\Sigma\left(W_1-W_s\right)\right),\boldsymbol g(\Sigma W_1)\right)_{s\in[\bar s,1-\bar s]}$$

in the Skorokhod metric space of càd-làg functions on $[\bar s,1-\bar s]$ with values in $(\mathbb{R}^d)^2\times GL_d(\mathbb{R})$. Then, with a simple extension of the arguments of Theorem 3.16, the triple $\left(Q^1_{s,n},Q^2_{s,n},-Q^1_{1,n}\right)_{s\in[\bar s,1-\bar s]}$ converges as follows:

$$\left(Q^1_{s,n},Q^2_{s,n},-Q^1_{1,n}\right)_{s\in[\bar s,1-\bar s]}\xrightarrow[n\to\infty]{\mathcal{D}}\left(\frac{1}{s}\boldsymbol q\left(W_s\right),\frac{1}{1-s}\boldsymbol q\left(W_1-W_s\right),-\boldsymbol q(W_1)\right)_{s\in[\bar s,1-\bar s]}.$$

From Remark 3.15, $\boldsymbol q$ is a quadratic form and, from Lemma 3.22,

$$\frac{1}{s}\boldsymbol q\left(W_s\right)+\frac{1}{1-s}\boldsymbol q\left(W_1-W_s\right)-\boldsymbol q(W_1)=\frac{\boldsymbol q(W_s-sW_1)}{s(1-s)},\ s\in[\bar s,1-\bar s].$$

The result follows from a last application of Theorem 1.11 to the application that sums the elements of the triple above and takes the supremum over $[\bar s,1-\bar s]$. □

The limit distribution obtained is somehow similar to the one given by Csörgő and Horváth (1997), Corollary 1.1.1, for the i.i.d. case of an unconstrained log-likelihood ratio test. Since $\boldsymbol{q}$ is a quadratic form, $\boldsymbol{q}(W_s - sW_1)_{s\in[\bar{s},1-\bar{s}]}$ is also a quadratic form of a Brownian bridge. The introduction of the weights impacts here the dimension of the Brownian bridge that is here $2d + d^2$, while, in the standard case, the Brownian bridge is of dimension d.

3.3.4 Test procedure

In practice, we observe a realization $X(\omega)$ for some $\omega \in \Omega$ of the random sample X with n observations. We propose the following test procedure:

1. Compute the estimators $\hat{\theta}$, $(\hat{\theta}_{0,s})_{s\in[\bar{s},1-\bar{s}]}$ and $(\hat{\theta}_{s,1})_{s\in[\bar{s},1-\bar{s}]}$ where $\bar{s}$ is known from Assumption 3.1. The three estimators are defined in Section 3.2.
2. Compute the process $(\Lambda_{s,n})_{s\in[\bar{s},1-\bar{s}]}$ and the test statistic S_n using their definitions in Equations (3.5) and (3.6).
3. Compute the constants $\boldsymbol{I}^{-1}$ defined in (3.1), $\boldsymbol{U}$ defined in (3.34), $\boldsymbol{J}$ defined in (3.22) and the covariance matrix Σ defined for Lemma 3.12. This requires additional developments: theoretical computation or numerical approximation.
4. Compute the distribution of $\sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}(W_s - sW_1)}{s(1-s)}$ where W is a standard Brownian motion and $\boldsymbol{q}$ is defined in (3.33). This requires additional development: theoretical computation or numerical approximation.
5. Compute the threshold L_α chosen with respect to a false alarm constraint. It can be obtained from the probability of false alarm $\alpha \in (0,1)$ such that L_α is the α-percentile of the distribution of $\sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}(W_s - sW_1)}{s(1-s)}$. We conclude that no change-point occurs if the statistic S_n is smaller than the threshold: we should reject H_0.

Since, by definition, the estimators $\hat{\theta}_{0,s}$, $\hat{\theta}_{s,1}$ and the process $\Lambda_{s,n}$ are constant piecewise, it is sufficient to estimate them on a finite set of $s \in [\bar{s}, 1-\bar{s}]$.

This test focuses on the first component of the mixture. Obviously, by definition of the mixture and of the test itself, this can be applied to any other component. Therefore, in practice, it might be relevant to run a test on each component. Alongside a standard Likelihood Ratio Test that does not focus on a specific component, the set of tests constitutes a useful detection tool for the industry.

In the next section, we suggest an extended version of this test. Numerical applications showed that such an extension increases the detection frequency when a change occurs.

3.4 Extension: scaling the contributions in the likelihood ratio (EWLT)

In this section, we introduce an extended version of the test defined in Section 3.2.2. For a fixed $s \in [\bar{s}, 1]$, we define the **contribution** $c_{s,n}$ by

$$c_{s,n} := \sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) + \sum_{i=\lfloor sn \rfloor + 1}^{n} w(X_i, \hat{\theta}_{s,1}). \tag{3.37}$$

Remark that, from its definition in (3.5), the log-ratio $\Lambda_{s,n}$ is the difference of $(Q^1_{s,n} + Q^2_{s,n})$ and $Q^1_{1,n}$. Then, $c_{s,n}$ is the contribution of the sample to the term $(Q^1_{s,n} + Q^2_{s,n})$, and $c_{1,n} = \sum_{i=1}^n w(X_i, \hat{\theta})$ is the contribution of the sample to the term $Q^1_{1,n}$. Under the null hypothesis, by Lemma 3.11 and the definition of $w(.,.)$ in (3.2), we see that

$$\frac{c_{s,n}}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_{H_0}\left[w(X_1, \boldsymbol{\theta})\right] = \int_{\mathcal{X}} \frac{\boldsymbol{p}_1 f_1(x, \boldsymbol{\lambda}_1)}{f(x, \boldsymbol{\theta})} f(x, \boldsymbol{\theta}) dx = \boldsymbol{p}_1,$$

uniformly in $s \in [\bar{s}, 1]$. It follows that, for a given $s \in [\bar{s}, 1-\bar{s}]$, the average contributions $c_{s,n}/n$ and $c_{1,n}/n$ have the same limit under the null hypothesis. Under the alternative hypothesis, we observed in numerical applications that these average contributions can play a significant role in the detection performance. Therefore, we suggest to scale our statistic with the total contributions. We define a new log-ratio process $\Lambda^*_n := (\Lambda^*_{s,n})_{s\in[\bar{s},1-\bar{s}]}$ by

$$\begin{aligned}\Lambda^*_{s,n} := \frac{c_{1,n}}{c_{s,n}} &\left(\sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) \log f_1(X_i, \hat{\lambda}_{0,s,1}) + \sum_{i=\lfloor sn \rfloor + 1}^{n} w(X_i, \hat{\theta}_{s,1}) \log f_1(X_i, \hat{\lambda}_{s,1,1}) \right) \\ &- \sum_{i=1}^{n} w(X_i, \hat{\theta}) \log f_1(X_i, \hat{\lambda}_1). \end{aligned} \tag{3.38}$$

The test statistic is then defined by $S^*_n := \sup_{s\in[\bar{s},1-\bar{s}]} \Lambda^*_{s,n}$. We refer to this test as the **EWLT** (Extended Weighted Likelihood Test).

In order to obtain a limit distribution for S^*_n, we start by noticing that

$$\begin{aligned}\Lambda^*_{s,n} &= \frac{c_{1,n}}{c_{s,n}} \Lambda_{s,n} + \left(\frac{c_{1,n}}{c_{s,n}} - 1 \right) \sum_{i=1}^{n} w(X_i, \hat{\theta}) \log f_1(X_i, \hat{\lambda}_1) \\ &= \frac{c_{1,n}}{c_{s,n}} \Lambda_{s,n} \\ &\quad - \frac{\frac{1}{n}\sum_{i=1}^n w(X_i, \hat{\theta}) \log f_1(X_i, \hat{\lambda}_1)}{\frac{1}{n} c_{s,n}} \left(\sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) + \sum_{i=\lfloor sn \rfloor + 1}^{n} w(X_i, \hat{\theta}_{s,1}) - \sum_{i=1}^{n} w(X_i, \hat{\theta}) \right).\end{aligned}$$

By Lemma 3.11, we have that the ratio $c_{1,n}/c_{s,n}$ converges a.s. to 1 uniformly in $s \in [\bar{s}, 1]$, and also that

$$\frac{\frac{1}{n}\sum_{i=1}^n w(X_i, \hat{\theta}) \log f_1(X_i, \hat{\lambda}_1)}{\frac{1}{n} c_{s,n}} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_{H_0}\left[\log f_1(Y, \boldsymbol{\lambda}_1)\right] =: \boldsymbol{\beta},$$

uniformly in $s \in [\bar{s}, 1]$, with Y a random variable with density $f_1(., \boldsymbol{\lambda}_1)$. We can show numerically that, in general, $\boldsymbol{\beta}$ is not null.
Remark that the sum

$$\sum_{i=1}^{\lfloor sn \rfloor} w(X_i, \hat{\theta}_{0,s}) + \sum_{i=\lfloor sn \rfloor + 1}^{n} w(X_i, \hat{\theta}_{s,1}) - \sum_{i=1}^{n} w(X_i, \hat{\theta})$$

has the same form as $\Lambda_{s,n}$ in (3.5), but without the factor $\log f_1(X_i, \hat{\lambda}_{.,.,1})$. Then, we already see that the limit distribution of S^*_n is obtained with similar arguments that

gave us the limit distribution of S_n in Theorem 3.17. As in Section 3.3, we start by rewriting $\Lambda^*_{s,n}$ as follows

$$\Lambda^*_{s,n} = \frac{c_{1,n}}{c_{s,n}}\Lambda_{s,n} - \frac{\frac{1}{n}\sum_{i=1}^n w(X_i,\hat{\theta})\log f_1(X_i,\hat{\lambda}_1)}{\frac{1}{n}c_{s,n}}(Q^{1*}_{s,n} + Q^{2*}_{s,n} - Q^{1*}_{1,n})$$

where, $Q^{1*}_n = (Q^{1*}_{s,n})_{s\in[\bar{s},1]}$ and $Q^{2*}_n = (Q^{2*}_{s,n})_{s\in[\bar{s},1-\bar{s}]}$ are càd-làg real-valued processes defined by

$$Q^{1*}_{s,n} := \sum_{i=1}^{\lfloor sn\rfloor}\Big(w(X_i,\hat{\theta}_{0,s}) - w(X_i,\boldsymbol{\theta})\Big) - \boldsymbol{v}^T\boldsymbol{I}^{-1}\sum_{i=1}^{\lfloor sn\rfloor} D_\theta(\log f)(X_i,\boldsymbol{\theta}),\ s\in[\bar{s},1],$$

$$Q^{2*}_{s,n} := \sum_{i=\lfloor sn\rfloor+1}^{n}\Big(w(X_i,\hat{\theta}_{s,1}) - w(X_i,\boldsymbol{\theta})\Big) - \boldsymbol{v}^T\boldsymbol{I}^{-1}\sum_{i=\lfloor sn\rfloor+1}^{n} D_\theta(\log f)(X_i,\boldsymbol{\theta}),\ s\in[\bar{s},1-\bar{s}],$$

and $\boldsymbol{v} := \mathbb{E}_{H_0}\left[D_\theta w(X_1,\boldsymbol{\theta})\right] \in \mathbb{R}^d$. With similar arguments as in Section 3.3.2, $Q^{1*}_{s,n}$ and $Q^{2*}_{s,n}$ can be expressed as functions of the triple $(\hat{\iota}_{0,s}, \hat{v}_{0,s} - \boldsymbol{v}, \hat{I}_{0,s})$ with $\hat{\iota}_{0,s}$ and $\hat{I}_{0,s}$ defined in (3.19), and

$$\hat{v}_{0,s} := \frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor} D_\theta w(X_i,\boldsymbol{\theta}).$$

It follows that $\Lambda^*_{s,n}$ can be expressed as a function of the quadruple

$$\hat{\xi}^*_{0,s} := (\hat{\iota}_{0,s}, \hat{u}_{0,s} - \boldsymbol{u}, \hat{v}_{0,s} - \boldsymbol{v}, \hat{I}_{0,s})$$

and a random variable that depends on s and n and converges a.s. uniformly in s to some finite constant.

We denote by Σ^* the unique positive semi-definite square root of the covariance matrix under the null hypothesis of the quadruple

$$\left(D_\theta(\log f)(X_1,\boldsymbol{\theta}),\ D_\theta(w\log f_1)(X_1,\boldsymbol{\theta}) - \boldsymbol{u},\ D_\theta w(X_1,\boldsymbol{\theta}) - \boldsymbol{v},\ -D^2_\theta(\log f)(X_1,\boldsymbol{\theta}) - \boldsymbol{I}\right)$$

reorganized in a $3d + d^2$-dimensional real vector. Still under the null hypothesis, the result of Lemma 3.12 can be extended to the process $(\hat{\xi}^*_{0,s})_{s\in[\bar{s},1]}$ with an application of Donsker's Theorem 1.10. With $\boldsymbol{\xi}^* := (0,0,0,\boldsymbol{I})$, the process $\sqrt{n}(\hat{\xi}^*_{0,s} - \boldsymbol{\xi}^*)_{s\in[\bar{s},1]}$ converges weakly in the Skorokhod metric space as follows

$$\sqrt{n}\left(\hat{\xi}^*_{0,s} - \boldsymbol{\xi}^*\right)_{s\in[\bar{s},1]} \xrightarrow[n\to\infty]{\mathcal{D}} \left(\frac{1}{s}\Sigma^* W_s\right)_{s\in[\bar{s},1]},$$

where $W := (W_s)_{s\in[0,1]}$ is a standard $3d + d^2$-dimensional Brownian motion and $\Sigma^* W_s$ is reorganized as a quadruple in $(\mathbb{R}^d)^3 \times gl_d(\mathbb{R})$.

As for the limit distribution of S_n in Theorem 3.17, the convergence of S^*_n is obtained by a functional delta method (Corollary 1.16) and multiple applications of the Extended Slutsky's Theorem 1.13 and the Continuous Mapping Theorem 1.11.
To this end, we start by adapting the function $\boldsymbol{g}$ defined in (3.25), for Theorem 3.14. We define the map $\boldsymbol{g}^*$ for a quadruple (ι, u, v, I) in $(\mathbb{R}^d)^3 \times gl_d(\mathbb{R})$ by

$$\boldsymbol{g}^*(\iota,u,v,I) := \left(\iota, u, v, -\boldsymbol{I}^{-1}\left(I - \frac{1}{2}\sum_{l=1}^d\left(\iota^T\left(\boldsymbol{I}^{-1\,T}\right)_{.,l}\right)\boldsymbol{J}_{.,.,l}\right)\boldsymbol{I}^{-1}\right).$$

Remark that, for a fixed s, $\boldsymbol{g}^*(\Sigma^* W_s)$ is a quadruple in $(\mathbb{R}^d)^3 \times gl_d(\mathbb{R})$. Remark also that, when we applied the delta method in the proof of Theorem 3.14, the vector u did not play a significant role. Here, this is also the case for the vector v. The result is then extended to the quadruple process $\sqrt{n}(\hat{\xi}^*_{0,s} - \boldsymbol{\xi}^*)_{s\in[\bar{s},1]}$ without additional arguments. Then, similarly to the function q defined in (3.32), we introduce the map q^*, defined for $(\iota, u, v, I; A, J, J^*) \in (\mathbb{R}^d)^3 \times (gl_d(\mathbb{R}))^4$ by

$$\begin{aligned} &q^*(\iota, u, v, I; A, J, J^*) \\ &= u^T A\iota - \boldsymbol{\beta} v^T A\iota + \iota^T A^T (J - \boldsymbol{\beta} J^*) A\iota + (\boldsymbol{u}^T - \boldsymbol{\beta}\boldsymbol{v}^T) I\iota \in \mathbb{R}. \end{aligned}$$

Last, we set

$$\boldsymbol{q}^*(z) := q^*(\boldsymbol{g}^*(\Sigma^* z); \boldsymbol{I}^{-1}, \boldsymbol{U}, \boldsymbol{V}),\ z \in \mathbb{R}^{3d+d^2}, \tag{3.39}$$

where $\Sigma^* z$ is reorganized as a quadruple in $(\mathbb{R}^d)^3 \times gl_d(\mathbb{R})$ and with $\boldsymbol{V}$ the matrix defined by

$$\boldsymbol{V} := \mathbb{E}_{H_0}\left[D^2_\theta w(X_1, \boldsymbol{\theta})\right]$$

and $\boldsymbol{U}$ defined in (3.34). With the same arguments as in Remark 3.15, $\boldsymbol{q}^*$ is also a quadratic form. The following theorem states the limit distribution obtained for S^*_n.

Theorem 3.18. *Under H_0 and Assumptions 3.1-3.5, the test statistic*

$$S^*_n \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{s\in[\bar{s},1-\bar{s}]} \frac{\boldsymbol{q}^*(W_s - sW_1)}{s(1-s)}$$

where $(W_s)_{s\in[0,1]}$ is a standard $3d+d^2$-dimensional Brownian motion and the application $\boldsymbol{q}^*$ is defined in (3.39)*.*

Proof. The proof follows the same logic as the proof of Theorem 3.17. The arguments are based on a functional delta method (Corollary 1.16) and multiple applications of the Extended Slutsky's Theorem 1.13 and the Continuous Mapping Theorem 1.11. □

The test procedure follows the same steps as described in Section 3.3.4. We will see in Section 3.6.2 that this extension improves significantly the detection quality. In particular the type II error is smaller compared to the one of the WL test.

Remark 3.19. *In Theorems 3.17 and 3.18, and in the results on which they rely, the dimension of the limit Brownian motion W can be reduced by recognizing that the matrix $\hat{I}_{0,s}$ defined in* (3.19) *is symmetric: it contains $d(d+1)/2$ distinct elements. This does not seem to have an impact on the numerical properties of the tests.*

In the following section we explicit the test in the case of an univariate Gaussian finite mixture.

3.5 Example: the univariate finite Gaussian mixture

In this section, we assume that the sample $X = (X_i)_{1\le i\le n}$ follows an univariate Gaussian mixture with m components. In addition to the weight parameters $p_1, \ldots, p_{m-1}$,

the mixture is defined by the means $\mu_1, \dots, \mu_m \in \mathbb{R}$ and the standard deviations $\sigma_1, \dots, \sigma_m \in \mathbb{R}_*^+$ of the m components. We assume that the set of eligible parameters Θ that contains $\boldsymbol{\theta}$ is the subset of $\Theta_0 \times (\mathbb{R} \times \mathbb{R}_*^+)^m$ defined by the parameters

$$\theta = (p_1, \dots, p_{m-1}, \mu_1, \sigma_1, \dots, \mu_m, \sigma_m)$$

such that:

1. The means are strictly increasing: $\mu_1 < \mu_2 < \cdots < \mu_m$.
2. There exists some **dispersion boundary** $0 < \boldsymbol{b} \leq 1$, deterministic and known, such that the variances are positive and bounded: for all $i \in \{1, \dots, m\}$, $\sigma_i > 0$ and
$$\min \left\{ \frac{\sigma_j}{\sigma_k},\ 1 \leq j, k \leq m \right\} > \boldsymbol{b}.$$

Note that Θ is an open convex [15] subset of $\mathbb{R}^{3m-1}$.

Remark 3.20. *We impose that the means of the components cannot be equal in order to ensure that the mixture is identifiable and that Θ is an open subset of $\mathbb{R}^{3m-1}$. If, under H_0, two means are equal, it is sufficient to define a model that uses one less parameter, i.e. using the same parameter for both means but with different variances.*

The second assumption is a constraint from Hathaway (1985) that ensures the strong consistency of the estimator $\hat{\theta}$ (Theorem 1.7). We obtain the following result.

Proposition 3.21. *Under Assumption 3.1 and with the parameter set Θ defined above, the validity conditions of Theorems 3.17 and 3.18 hold for a finite Gaussian mixture.*

Proof. The result is obtained as soon as we show that Assumptions 3.2-3.5 are valid. First of all, the conditions of the model introduced in Section 3.2.1 and Assumptions 3.2 and 3.3 are standard prerequisites for limit results of likelihood based estimators (McLachlan and Peel (2000)). In particular, from Example 6.10 in Lehmann and Casella (1998), one sees that the assumptions of Theorem 1.5 hold for identifiable Gaussian mixtures. With condition 1. above, it follows that Assumptions 3.2 and 3.3 hold.
With condition 2. above, Theorem 1.7 from Hathaway (1985) ensures that Assumption 3.4 is valid, i.e. the estimator $\hat{\theta}$ is strongly consistent.
Since the parameter set Θ is a convex open subset of $\mathbb{R}^d$ that contains the true parameter $\boldsymbol{\theta}$, it is possible to find a bounded convex set $\Theta' \subset \Theta$ such that $\boldsymbol{\theta}$ is in the interior of Θ'. We show that Assumption 3.5 is valid for this set Θ'. First we recall that, $\mathbb{E}_{H_0}[|X_1|^k]$ is finite for all $k \geq 0$. Thus it is sufficient to show that we can find some function $\theta \mapsto \kappa(\theta)$ with positive values, not depending on x and continuous on Θ', such that, for all $1 \leq i, j \leq d$, θ in Θ' and $x \in \mathbb{R}$,

$$\left| D_\theta^2 (w \log f_1)(x, \theta)_{i,j} \right| \leq \sum_{k=0}^{6} \kappa(\theta) |x|^k. \tag{3.40}$$

On the one hand, since $x \in \mathbb{R} \mapsto f_1(x, (\mu_1, \sigma_1))$ is the density function of a Gaussian random variable, $\log f_1(x, (\mu_1, \sigma_1))$ can be written as a second-order polynomial of x. Its coefficients are infinitely differentiable functions of (μ_1, σ_1) on $\mathbb{R} \times \mathbb{R}_*^+$. It follows that the absolute values of the first and second order partial derivatives of $f_1(x, (\mu_1, \sigma_1))$ can

15. In particular, for any $t \in [0, 1]$, for any θ and θ' in Θ, and for any $1 \leq j, k \leq m$, it holds that $\frac{t\sigma_j + (1-t)\sigma'_j}{t\sigma_k + (1-t)\sigma'_k} > \frac{t\boldsymbol{b}\sigma_k + (1-t)\boldsymbol{b}\sigma'_k}{t\sigma_k + (1-t)\sigma'_k} = \boldsymbol{b}$.

be bounded by a second-order polynomial as on the left side of (3.40). On the other hand, by (3.3), the weight function $w(x, \theta)$ takes its values in $[0, 1]$ for all x and all θ. In order to conclude the proof, we need only to bound the absolute value of the first and second partial derivatives of $w(x, \theta)$.
From the definition of w in (3.2), its first order partial derivatives can be written as

$$\frac{\partial}{\partial\theta_i} w(x,\theta) = \frac{\frac{\partial}{\partial\theta_i}\tilde{f}_1(x,\theta)}{f(x,\theta)} - w(x,\theta)\sum_{k=1}^{m}\frac{\frac{\partial}{\partial\theta_i}\tilde{f}_k(x,\theta)}{f(x,\theta)},\ 1 \leq i \leq d$$

where $\tilde{f}_k : (x, \theta) \mapsto p_k f_k(x, \lambda_k)$. We can show that

$$\frac{\frac{\partial}{\partial\theta_i}\tilde{f}_k(x,\theta)}{f(x,\theta)} = \frac{p_k f_k(x,\lambda_k)}{f(x,\theta)}\kappa_{i,k}(x,\theta),$$

where $\kappa_{i,k}(x, \theta)$ is a second-order polynomial of x with coefficients that are infinitely differentiable functions of θ on Θ. Moreover we recognize that $p_k f_k(x, \lambda_k)/f(x, \theta)$ is another weight function that takes its values in $[0, 1]$ for all x and all θ. Thus $|\frac{\partial}{\partial\theta_i} w(x, \theta)|$ can be bounded by a second-order polynomial as on the left side of (3.40). With similar arguments, the absolute value of the second partial derivatives of $w(x, \theta)$ can be bounded by a polynomial of degree four as in the left side of (3.40). The result follows. □

Remark that Assumption 3.1 does not concern the choice of distribution and therefore remains a preliminary condition to be discussed when applying the WL and EWL tests. In the following section, we give a few applications for the Gaussian case: we start with numerical illustrations of the test, compared to some standard test available in the literature. Then we show how this test can help detect if a change occurs or not in a dataset from the non-life insurance industry.

3.6 Applications

We provide two distinct applications for the case of an univariate finite Gaussian mixture. First, with numerical simulations, we illustrate the properties of the three following tests:

- ◇ the WL test, defined by the statistic S_n in (3.6) and for which Theorem 3.17 provides a limit distribution;
- ◇ the EWL test, defined by the statistic S_n^* in Section 3.4 and for which Theorem 3.18 provides a limit distribution;
- ◇ a standard *benchmark* test that we shall define in the following section.

In this numerical application, our main interest lies in the detection of changes in the first component that are not visible to the naked eye (small) but also not too close to 0 (no impact in practice), for large samples (over 10k observations). The second application is an illustration of the WL and EWL tests on a Property and Casualty insurance large dataset (15k observations).

We start by introducing the benchmark test (BM).

3.6.1 The *benchmark* test

The book of Csörgő and Horváth (1997) gathers standard likelihood-based approaches for the detection of change-points in many different frameworks. For the simple *at most one change* (AMOC) case, one can consider that each parameter $\theta = (a, b) \in \Theta$ is defined by two sub-parameters a and b. We give below a standard likelihood-based hypothesis test that aims to detect if a change occurs in the first sub-parameter a (see e.g. Section 1.1 in Csörgő and Horváth (1997)). Here b is called a *nuisance* parameter. We test:

- ◇ the null hypothesis where no change happens, i.e. $\theta^1 = \cdots = \theta^n$,
- ◇ against the alternative hypothesis where at most one change occurs, i.e. there exists some $1 \leq k \leq n$ such that $a^1 = \cdots = a^k \neq a^{k+1} = \cdots = a^n$ and $b^1 = \cdots = b^n$.

The test is defined with the help of the log-likelihood ratio

$$\Lambda_{k,n}^{cs} := \log \left(\frac{\sup_{(a,b),(a',b)\in\Theta} \prod_{i=1}^{k} f(X_i, (a,b)) \prod_{i=k+1}^{n} f(X_i, (a',b))}{\sup_{\theta\in\Theta} \prod_{i=1}^{n} f(X_i, \theta)} \right)$$

and the test statistic $\max_{1\leq k\leq n} 2\Lambda_{k,n}^{cs}$. Csörgő and Horváth (1997) provide its limit distribution [16].

For the detection of a change in the first component for a finite parametric mixture with m components, we set $a := (p_1, \lambda_1)$ and $b := (p_2, \ldots, p_{m-1}, \lambda_2, \ldots, \lambda_m)$. In other words, the separation between the parameters of interest a and the nuisance parameters b allows the test to focus on a change in the first component. In the WL and EWL tests, this role is played by the weight functions.
This setting for a and b also means that we allow the weight parameter of the first component to change ($p_1 \neq p_1'$). Looking at the numerator of $\Lambda_{k,n}^{BM}$, for both couples (a, b) and (a', b), the sum of the $m-1$ weight parameters has to be strictly below one. We impose first that $\sum_{k=1}^{m-1} p_k < 1$. Then, we assume that the relative weight of two components for $2 \leq i, j \leq m$ is the same before and after k. The log-likelihood ratio becomes:

$$\Lambda_{k,n}^{BM} := \log \left(\frac{\sup_{a,a',b} \prod_{i=1}^{k} f(X_i, (a,b)) \prod_{i=k+1}^{n} f(X_i, (a',b'))}{\sup_{\theta\in\Theta} \prod_{i=1}^{n} f(X_i, \theta)} \right) \tag{3.41}$$

with

$$b' := \left(\frac{1-p_1'}{1-p_1} p_2, \ldots, \frac{1-p_1'}{1-p_1} p_{m-1}, \lambda_2, \ldots, \lambda_m \right).$$

Thus the sum of the $m-1$ weight parameters of the couple (a', b') is also below one. We refer to the test defined by the test statistic $\max_{1\leq k\leq n} 2\Lambda_{k,n}^{BM}$ as the *benchmark* test.

3.6.2 Numerical properties

Setup

While establishing Theorem 3.17, we assumed that the null hypothesis holds in order to control the type I error (proportion of false positives). In this numerical application,

16. See Corollary 1.1.1, Theorems 1.3.1. and 1.3.2 in Csörgő and Horváth (1997).

we aim to visualize some properties of the WL and EWL tests:

- ◇ Compare the type II error (proportion of false negative) of the three tests (WL, EWL and *benchmark*),
- ◇ Understand the computational difficulties raised by each test (available algorithms, run time, inadequate convergence, etc.),
- ◇ As a nice-to-have, visualize the detection frequency of the tests when a change occurs in any other component.

We illustrate our analysis with simulations of a Gaussian mixture with 3 components such that:

- ◇ The sample size is large: $n = 1$k or 10k observations (obs.) as we are interested in applications to large datasets (see the application on real data in Section 3.6.3).
- ◇ Under the null hypothesis, the mixture used for the illustrations has equal weight parameters (1/3), equal standard deviations (0.25) and respective means -1, 0 and 1. It follows that, with a relatively small sample, the empirical density of the mixture shows clearly the three components.
- ◇ The change occurs in middle of the sample ($\boldsymbol{s} = 0.5$) [17].
- ◇ The detection threshold is the 90% percentile of test statistic under the null hypothesis (type I error).

The results are obtained with standard algorithms from the *R* software. For the benchmark test, the optimization problem of the numerator of $\Lambda_{k,n}^{BM}$ is not standard and a dedicated algorithm does not exist. We solve this problem with the generic *optim* function (Byrd et al. (1995)). For the denominator of $\Lambda_{k,n}^{BM}$ and the estimators in the WL and EWL tests, we apply the standard EM algorithm from the *mixtools* package (Benaglia et al. (2009)). We distinguish two settings for the initialization of the algorithms:

- ◇ A *standard* initialization assuming that we do not know the parameters,
- ◇ A *theoretical* initialization with the true parameters. This is possible because we are simulating the data.

The comparison of both types of initialization illustrate the capacity of the algorithms to converge to the optimal parameters. Results are obtained by multiple simulations of the random sample.

In this setup, the results obtained under the null hypothesis in Theorems 3.17 and 3.18 allow us to reduce significantly the calibration run time of the detection thresholds: the marginal run time of one simulation is divided by 10 000. For 10^6 simulations of the limit distribution, due to the approximation of the constants of the maps $\boldsymbol{q}$ and $\boldsymbol{q}^*$ defined in (3.33) and (3.39), the run time of the WLT and EWLT is divided by 8 000.

Results: detection quality under the alternate hypothesis

The alternate hypothesis is defined in Section 3.2.1 as the case when the sample contains one change in the first component: the parameters which describe the distribution

17. The simulations for the numerical comparison given in this section require an important run time. In order to reduce significantly this run time, we computed the illustrations for $s = 0.5$ (statistic and threshold) and not over the whole process. We made sure that this does not effect the conclusions. The application on the real data in Section 3.6.3 is based on a computation of the whole process.

of the first component are different before and after the change-point while the other parameters of the mixture remain the same. In this application, we illustrate the potential properties of the WL and EWL tests through three kind of changes in the parameters of the first component: a shift between -1 and +1 of the mean, a shift between -0.2 and +0.5 of the standard deviation and a shift between -0.25 and +0.25 of the weight parameter.
The type II error (proportion of false negative) of each test is obtained from multiple re-simulations of samples that contain a change in the first component (Table 3.1). A high-performance test is characterized by a low type II error. Detailed graphs are also given in Figures 3.4 and 3.5 in Appendix 3.8.3.

in %	*Theoretical* init.						*Standard* init.					
	n=1k obs.			n=10k obs.			n=1k obs.			n=10k obs.		
	WL	EWL	BM	WL	EWL	BM	WL	EWL	BM	WL	EWL	BM
Mean +0.1	34.8	15.4	11.6	0.0	0.0	0.0	44.4	21.8	42.4	**0.4**	**0.8**	30.8
Mean -0.1	36.8	14.4	9.2	0.0	0.0	0.0	42.4	19.0	29.8	**0.8**	**1.2**	50.8
Std dev +0.1	22.0	8.6	6.8	0.0	0.0	0.0	33.0	9.2	73.0	**0.4**	**0.6**	43.4
Std dev -0.1	5.6	0.4	0.0	0.0	0.0	0.0	9.8	1.0	22.6	**1.0**	**0.2**	45.4
Weight +0.1	61.2	38.8	13.8	0.0	0.0	0.0	67.2	42.0	40.8	**0.2**	**0.6**	55.4
Weight -0.1	54.2	26.6	10.4	0.0	0.0	0.0	51.8	27.4	34.2	**0.6**	**0.2**	27.8

Table 3.1 – Type II error (in %) for a change in the first component (500 re-simulations). Results are given respectively for the WLT, the EWLT and the benchmark test (BM).

First, we recall that our main interest lies in the detection of changes in the first component that are not visible to the naked eye (small) but also not too close to 0 (no impact in practice), for large samples (over 10k observations). In that sense, Table 3.1 and Figure 3.5 show that the WLT performs significantly better than the benchmark test for large samples. In addition, the EWLT improves the performance of the WLT. Regardless of algorithmic issues, i.e. with a *theoretical* initialization, the EWLT corrects side effects due to unnormalized contributions (see Section 3.4 for details on the contribution). In particular, Figures 3.4a and 3.4b illustrate the correction for a significant increase of the mean. Remark that this has a relative interest since such increase of the mean is visible to the naked eye. More generally, the three tests have a similar detection quality for large samples of 10k observations, even if the benchmark test seems better for a sample of 1k observations (Figures 3.4).
Results obtained with the *standard* initialization correspond to what one could expect when applying these tests to real data. In that case, the benchmark test fails to detect properly the change, especially with large samples. To our understanding, this is mainly due to the optimization problem in the numerator of $\Lambda_{k,n}^{BM}$ that the algorithm often fails to solve. Since their estimation algorithms are more robust, the WL and EWL tests have both very low type II errors for small changes in the parameters on a sample of 10k observation (Table 3.1 and Figure 3.5).
We also see that there is still room for improvement for large changes in the weight and the standard deviation, even if large changes do not have a strong importance since they are visible to the naked eye.

In addition to these performance results under the alternative hypothesis, Figure 3.1 shows that, as the sample size increases, the run time needed to compute the benchmark test increases considerably faster compared to the WLT. It follows that one major advantage of the WLT is that it can be computed quickly with standard algorithms, making it especially convenient for large datasets.

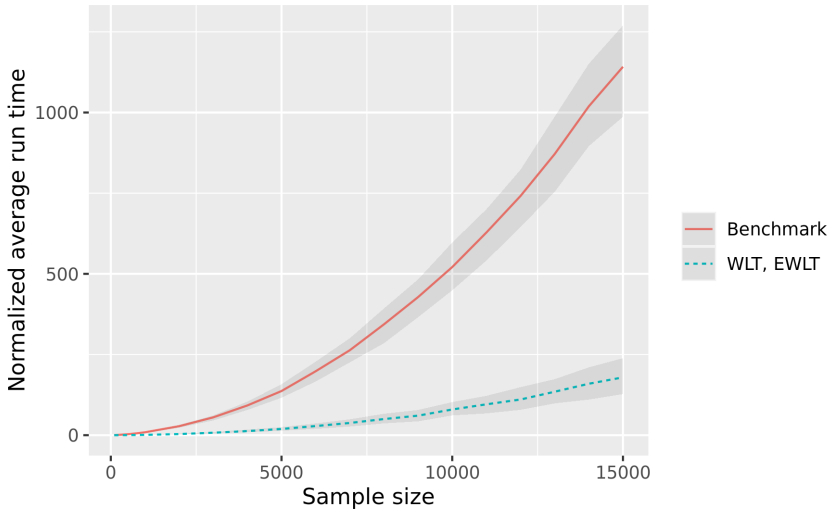


Figure 3.1 – Run time of the benchmark and the WL and EWL tests for an increasing sample size.

We conclude that the WLT is a valid candidate for the detection of a change in the first component of a Gaussian mixture. The EWLT is an improved version that significantly reduces the type II error.

Results: when a change occurs in the second or the third component

As a nice-to-have, we also studied the detection frequency of each test when a change occurs in the second or third component. We used the same shift ranges as the ones used for a change in the first component. In this context, a high-performance test is characterized by a low detection frequency. The results are given in Table 3.2 with detailed graphs in Figures 3.6, 3.7, 3.8 and 3.9 in Appendix 3.8.3.

The main observation is that the EWLT shows the best results and, for large samples of 10k observations, the WLT is better than the benchmark test. The WL and EWL tests tend to detect a change when there is a change in the standard deviation of the second or third component. This explained by the role of the weight functions in their detection statistic that zoom on some range of values around the mean of the first component: there are significantly more (or less) points that enter this range, increasing the detection frequency. We still observe that the benchmark test behaves poorly for large samples.
One could expect that a high-performance test would have a detection frequency that stays close to the type I error (10%). This is the case for the WL and EWL tests when a change occurs in the weight parameter (Figures 3.7f and 3.9f). However, from Figures 3.7 and 3.9, we remark that the three tests show systematic patterns for a wide range

in %	*Theoretical* init.						*Standard* init.					
	n=1k obs.			n=10k obs.			n=1k obs.			n=10k obs.		
	WL	EWL	BM	WL	EWL	BM	WL	EWL	BM	WL	EWL	BM
Change in the second component												
Mean +0.1	12	13	20	18	7	73	13	11	25	**16**	**8**	59
Mean -0.1	14	8	25	21	7	91	15	10	29	**23**	**6**	41
Std dev +0.1	49	22	27	99	63	91	44	20	28	**99**	**61**	43
Std dev -0.1	76	67	40	100	100	99	62	44	40	**100**	**100**	80
Weight +0.1	12	13	14	11	11	30	7	9	27	**14**	**13**	61
Weight -0.1	14	13	13	12	12	31	15	16	13	**16**	**14**	24
Change in the third component												
Mean +0.1	15	12	10	17	15	13	13	14	10	**15**	**13**	28
Mean -0.1	14	11	10	16	13	14	15	16	13	**15**	**14**	65
Std dev +0.1	14	14	11	86	64	9	11	11	10	**90**	**68**	18
Std dev -0.1	9	9	10	3	4	13	6	6	12	**7**	**6**	33
Weight +0.1	9	12	16	13	14	28	13	12	15	**11**	**10**	29
Weight -0.1	10	10	15	8	11	21	8	9	28	**12**	**12**	66

Table 3.2 – Detection frequency (in %) for a change in the second and third components (500 re-simulations). Results are given respectively for the WLT, the EWLT and the benchmark test (BM).

of possible deviations for a change in the mean and the standard deviation. It follows that there is still some room for improvement regarding this criterion.

In the next section, we study briefly an illustration to a Property and Casualty insurance large dataset (15k observations).

3.6.3 Illustration of the WL and EWL tests on P&C insurance data

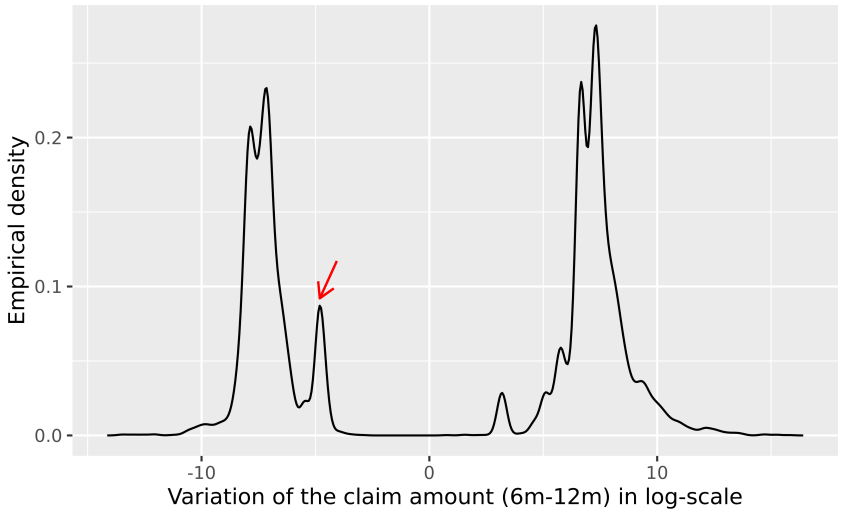
We recall that:

- ◇ the WLT is defined by the statistic S_n in (3.6) and Theorem 3.17 provides its limit distribution;
- ◇ the EWLT is defined by the statistic S_n^* in Section 3.4 and Theorem 3.18 provides its limit distribution.

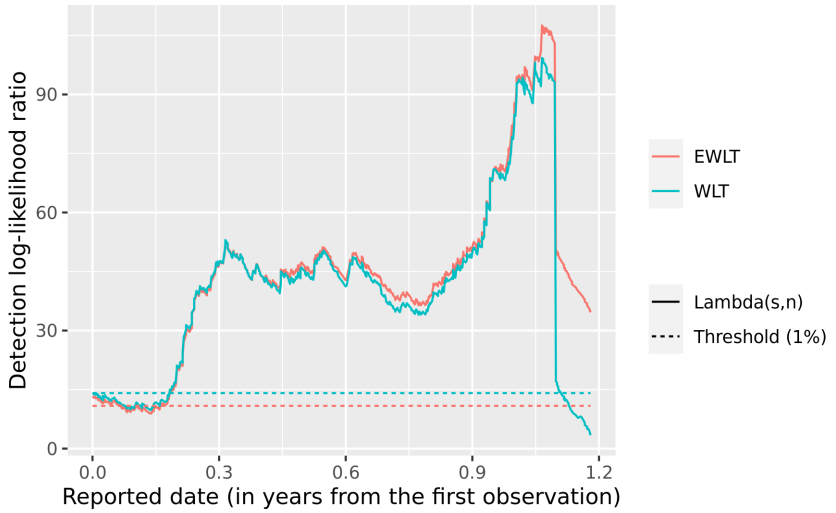
In this section, we give an example of application of these two tests to a problem from the insurance industry, in particular the bodily injuries from the motor claims. Each claim is known by the insurer from its declaration date that is the starting point of observation: we set $t = 0$, t being expressed in years. We denote by C_t the real-valued random variable that represents the amount that, at time t, the insurer expects to pay eventually. This amount varies over time when the claim is reviewed, until it is settled. The evolution of the amount C_t depends on structural factors (e.g. the type of injury), internal factors (e.g. a change in the revision policy) and external factors (e.g. new elements are known by the insurer, a court decision sets the final amount, etc.). For bodily injuries, the claims takes time to resolve (over 2 years in average). After the settlement, we assume that C_t is fixed and possibly null or negative.

In this application, we know that a change of the revision process happened at some point in the past. The question is then to determine whether or not this change impacted significantly the observed variations of claim amount over time. For that purpose, we consider the random variable $Z = sgn(C_1 - C_{0.5}) \log(1 + |C_1 - C_{0.5}|)$ that gives the variation of the claim amount between the 6th and the 12th month in log-scale, where $sgn(.)$ is the function that gives the sign of a real number taking respectively the values -1, 0 and 1 when this number is negative, null or positive. From a first analysis of the data, a Kolmogorov-Smirnov hypothesis test does not reject the assumption that observed realizations of Z before the change follow a finite parametric mixture with 12 components (Figure 3.2a).
For internal reasons, the insurance company is particularly interested in the 5th component of the mixture, highlighted in Figure 3.2a by a red arrow. This component represents slight decreases of claim amounts. We applied the WL and EWL tests to a sample of 15k claims where the first third of the sample is known to contain claims that are not impacted by the change of process. Under the null hypothesis that no change occurs in the 5th component of the mixture, both tests reject this hypothesis with a p-value below 10^{-4}. Figure 3.2b illustrates the underlying processes Λ_n and Λ_n^* and their respective thresholds: the change is significant and, according to both tests, it seems to occur from the time 1.07.
This conclusion allowed the insurance company to investigate further the quantification of the change.

This application shows that the WL and EWL tests can be used in the industry for the monitoring of changes, when they are unexpected but also to assess their significativity when they are known or suspected. For other topics that tackle change-point problems in non-life insurance, we refer for example to Dhaene et al. (2002), Andersen et al. (2009), Kwon and Vu (2017), Peštová and Pešta (2017) or Maciak et al. (2020).



(a) Empirical density of the variation in claims amount between the 6th and the 12th development months, before the change of process (7k obs).



(b) Log-ratio processes Λ_n and Λ_n^* of the WL and EWL tests for the detection of a change in the 5th component of the mixture (15k obs).

Figure 3.2 – Application of the WL and EWL tests to a change in the variation of incurred claims amount (Motor bodily injuries claims).

3.7 Conclusion

In this chapter, we consider a closed sample of independent random variables that follow a finite mixture distribution with parametric components. The sample might contain at most one change in the parameters of the first component. If there is a change, the r.v. are identically distributed before and after the change-point: the parameters which describe the distribution of the first component are different before and after the break-point while the other parameters of the mixture remain the same.
To test whether there is a change or not, we proposed two alternative tests (WLT and EWLT). Each test statistic is built upon a càd-làg process obtained from a likelihood ratio (see (3.5) and (3.38)). The specificity of these tests is that they can be computed using known inference algorithms. The first version uses weight functions to help the

likelihood ratio to *zoom* on the first component. In the second *extended* version, we added an adjustment that helps improve the type II error, as explained in Section 3.6.2.

With a technique from Davis et al. (1995), we derived in Theorems 3.17 and 3.18 the limit distribution of the test statistics under the null hypothesis in the form of a quadratic form of a multidimensional Brownian motion, with the help of a dedicated functional limit theorem. In particular, the proof is based on a functional delta method (Corollary 1.16) and multiple applications of the Extended Slutsky's Theorem 1.13 and the Continuous Mapping Theorem 1.11. We showed in Section 3.5 that validity conditions of the main result hold for univariate finite Gaussian mixtures within the framework of Hathaway (1985).

Numerical applications on simulated data for the Gaussian case showed that second version of the test outperforms significantly a benchmark test exposed in Csörgő and Horváth (1997) and defined by (3.41): the type II error is considerably reduced (when a change occurs in the first component) and the detection frequency remains low in most cases when a change occurs in another component.
Two issues of the benchmark test are that usual optimization algorithms have an unrealistic convergence run time for large samples, and that they fail to compute properly its statistic. However, in the case of simulated data, we assess that, without this computational issue, the benchmark test would have a lower type II error than our tests. Therefore the extended version of our tests remains so far the best candidate even if a dedicated algorithm for computing more robustly the statistic of the benchmark test would be an improvement.
In addition, in some cases, the three tests still detect a change when a change occurs in another component (i.e. not the first one). Extensions of our work could consider adding a penalization term to the likelihood ratio in order to improve this aspect.

We end the applications by a brief illustration of the proposed tests on variations of claim amounts for bodily injuries motor claims (real data), in the context of a change of process in the claims handling department of an insurance company. A change is detected in the fifth component of the Gaussian mixture with 12 components: the insurance company could therefore assess the change and investigate further its causes.

From the numerical applications, the WL and EWL tests are valid candidates when looking for a change in one component of a finite parametric mixture. In addition, the results obtained under the null hypothesis in Theorems 3.17 and 3.18 allow us to reduce significantly the calibration run time of the detection thresholds: the marginal run time of one simulation is divided by 10 000. Beyond these promising results, the possibilities for other techniques still exist and are worth to be explored.

3.8 Appendices

3.8.1 The constant $\boldsymbol{u}$

The constant $\boldsymbol{u} = \mathbb{E}_{H_0}\left[D_\theta(w \log f_1)(X_1, \boldsymbol{\theta})\right]$, defined in (3.8), plays a central role in the proof of Theorems 3.17 and 3.18 since most of the technical difficulties in Section 3.3 emerge only when $\boldsymbol{u} \neq 0$. This is the case in general, as illustrated in the following

numerical example.
With the notations of Section 3.5, we consider a numerical simulation for an univariate Gaussian mixture with 3 components defined by

$$\boldsymbol{\theta} := \Big((1/3, 1/3), (-1.00, 0.25), (0.00, 0.25), (1.00, 0.25)\Big).$$

For 10^5 simulations, the Monte-Carlo approximation of $\mathbb{E}_{H_0}\left[\frac{\partial}{\partial \mu_2}(w \log f_1)(X_1, \boldsymbol{\theta})\right]$ converges to a non null limit (Figure 3.3). This illustrates that $\boldsymbol{u}$ is not null in general.

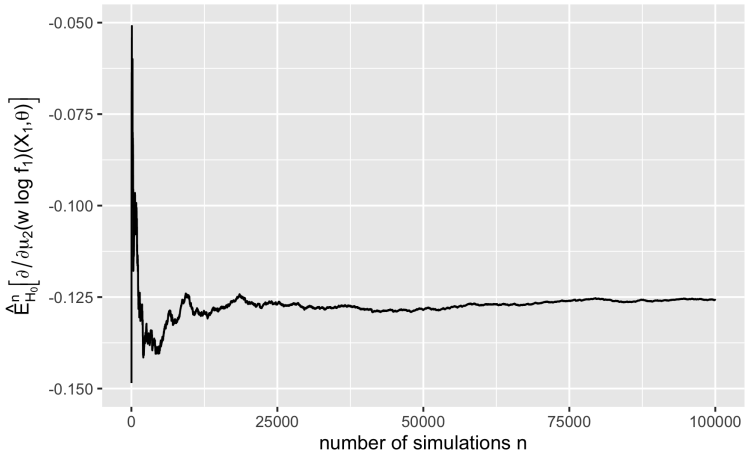


Figure 3.3 – Convergence of the Monte-Carlo approximation of $\mathbb{E}_{H_0}\left[\frac{\partial}{\partial \mu_2}(w \log f_1)(X_1, \boldsymbol{\theta})\right]$.

3.8.2 Additional result

This appendix contains an additional result for quadratic forms.

Lemma 3.22. *Fix $d \geq 1$. If $x \in \mathbb{R}^d \mapsto q(x)$ is a quadratic form, then, for any $x, y \in \mathbb{R}^d$ and any real $s \neq 0$, the following equality holds:*

$$\frac{1}{s} q(x) + \frac{1}{1-s} q(y-x) - q(y) = \frac{q(x-sy)}{s(1-s)}.$$

Proof. The unique symmetric bilinear form [18] associated to q is the application

$$(x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto b_q(x, y) = \frac{1}{2}\Big(q(x+y) - q(x) - q(y)\Big).$$

By definition of b_q, we have that $q(x) = b_q(x, x)$ and $q(x+y) = q(x) + q(y) + 2\, b_q(x, y)$. Further, for any real s and any $x \in \mathbb{R}^d$, $q(sx) = s^2 q(x)$. It follows that

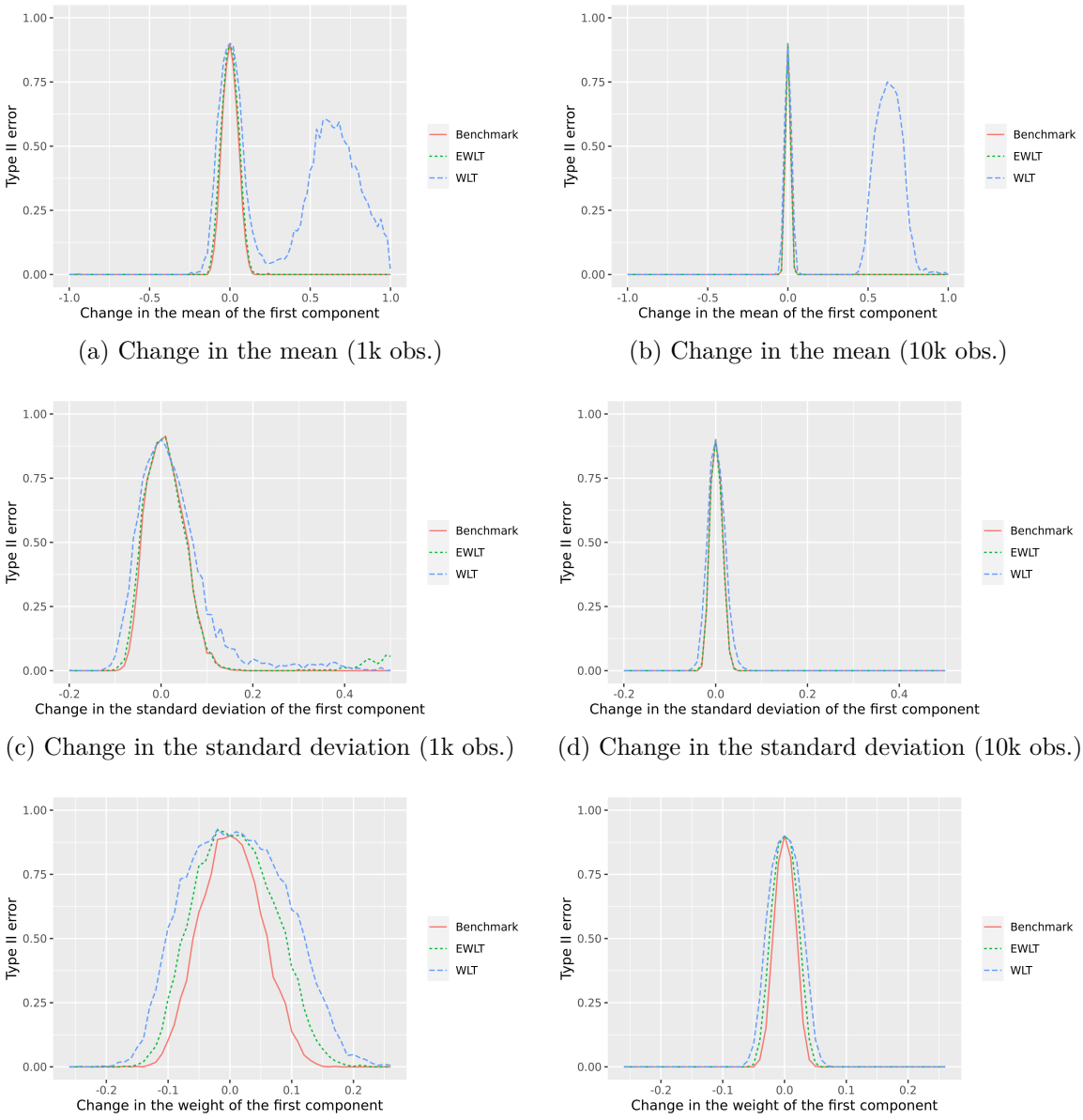
$$\begin{aligned}
&\frac{1}{s} q(x) + \frac{1}{1-s} q(y-x) - q(y) \\
=&\frac{(1-s)q(x) + s(q(x) + q(y) - 2\, b_q(x, y)) - s(1-s)q(y)}{s(1-s)} \\
=&\frac{q(x) - 2s\, b_q(x, y) + s^2 q(y)}{s(1-s)} = \frac{q(x) + 2\, b_q(x, -sy) + q(-sy)}{s(1-s)}.
\end{aligned}$$

The result follows. □

18. See for instance Section 41 in O'Meara (2000).

3.8.3 Additional illustrations

In this appendix, we provide additional illustrations of the numerical simulations exposed in Section 3.6.2.



(a) Change in the mean (1k obs.)

(b) Change in the mean (10k obs.)

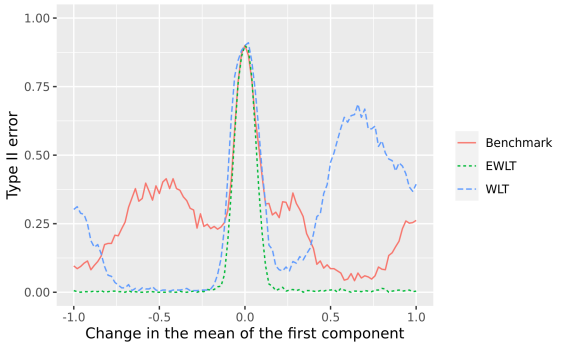
(c) Change in the standard deviation (1k obs.)

(d) Change in the standard deviation (10k obs.)

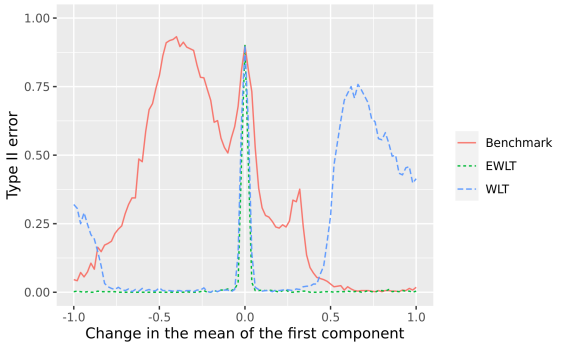
(e) Change in the weight parameter (1k obs.)

(f) Change in the weight parameter (10k obs.)

Figure 3.4 – Type II error for a change in the first component, *theoretical* initialization.

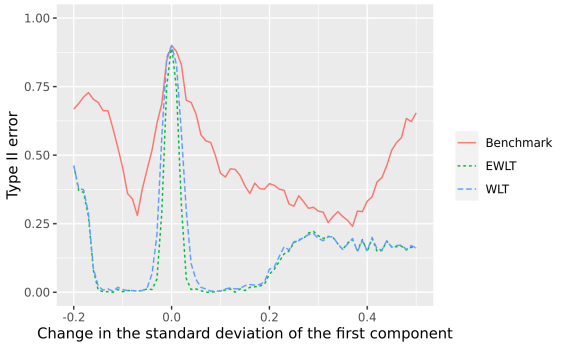


(a) Change in the mean (1k obs.)



(b) Change in the mean (10k obs.)

(c) Change in the standard deviation (1k obs.)

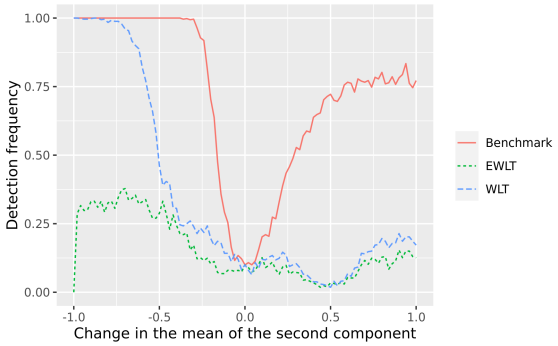


(d) Change in the standard deviation (10k obs.)

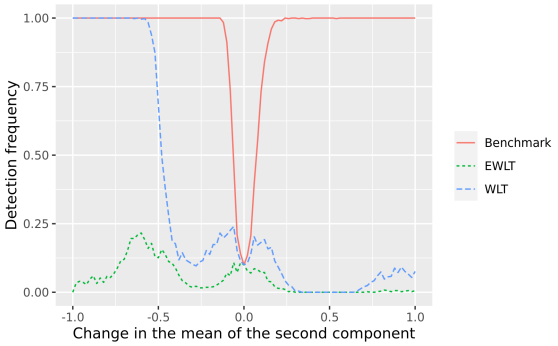
(e) Change in the weight parameter (1k obs.)

(f) Change in the weight parameter (10k obs.)

Figure 3.5 – Type II error for a change in the first component, *standard* initialization.

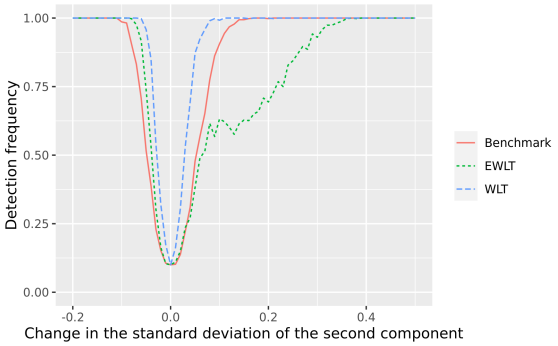


(a) Change in the mean (1k obs.)



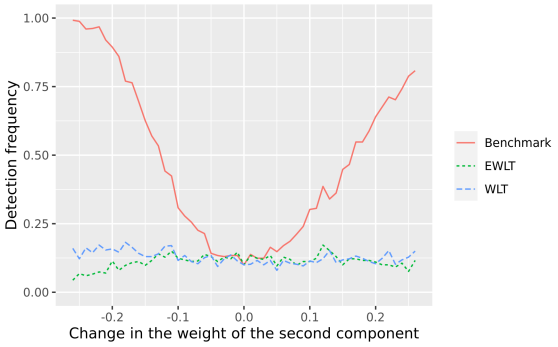
(b) Change in the mean (10k obs.)

(c) Change in the standard deviation (1k obs.)



(d) Change in the standard deviation (10k obs.)

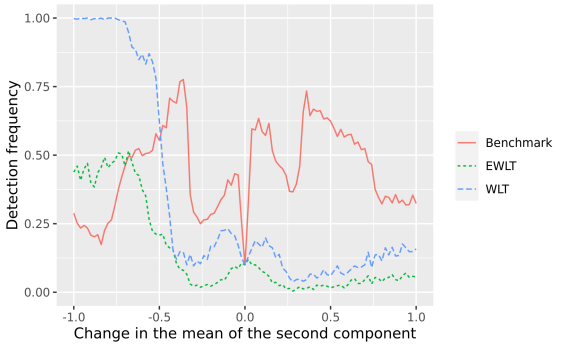
(e) Change in the weight parameter (1k obs.)



(f) Change in the weight parameter (10k obs.)

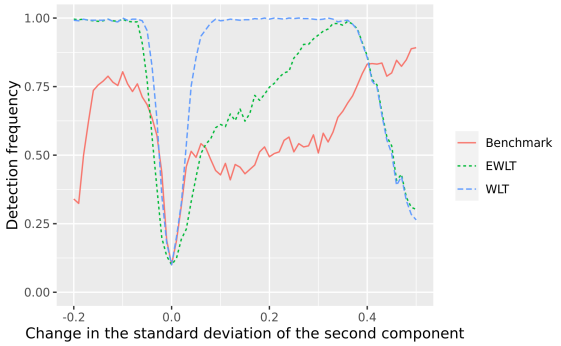
Figure 3.6 – Detection frequency for a change in the second component, *theoretical* initialization.

(a) Change in the mean (1k obs.)

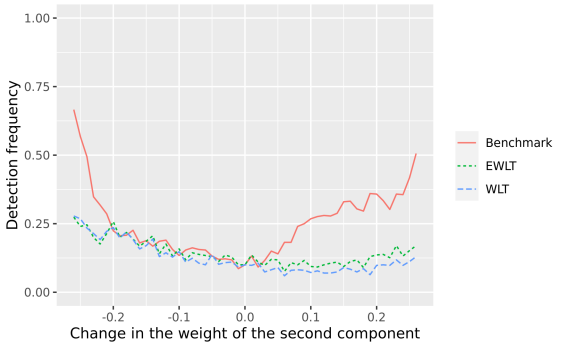


(b) Change in the mean (10k obs.)

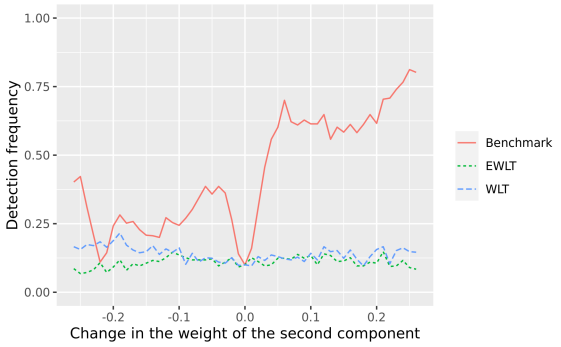
(c) Change in the standard deviation (1k obs.)



(d) Change in the standard deviation (10k obs.)

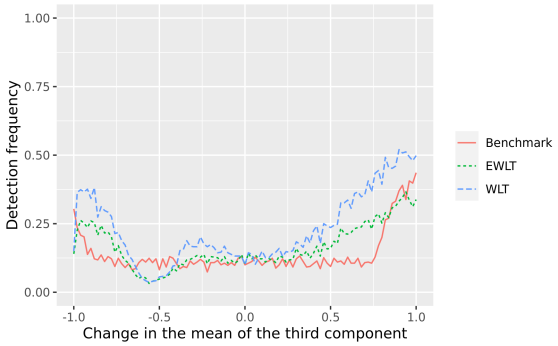


(e) Change in the weight parameter (1k obs.)



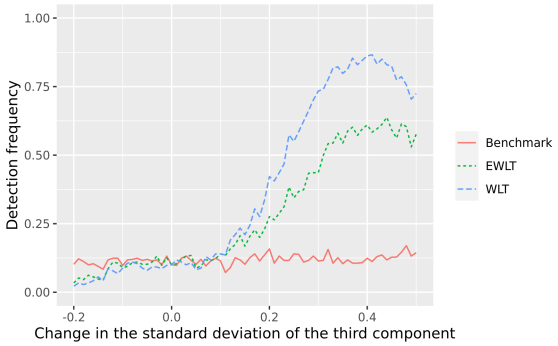
(f) Change in the weight parameter (10k obs.)

Figure 3.7 – Detection frequency for a change in the second component, *standard* initialization.

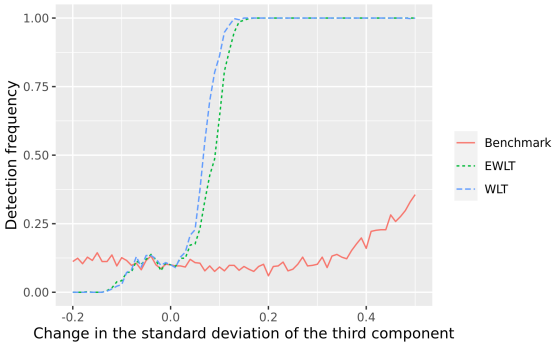


(a) Change in the mean (1k obs.)

(b) Change in the mean (10k obs.)

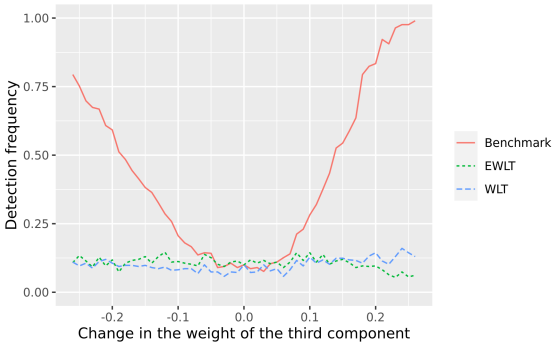


(c) Change in the standard deviation (1k obs.)



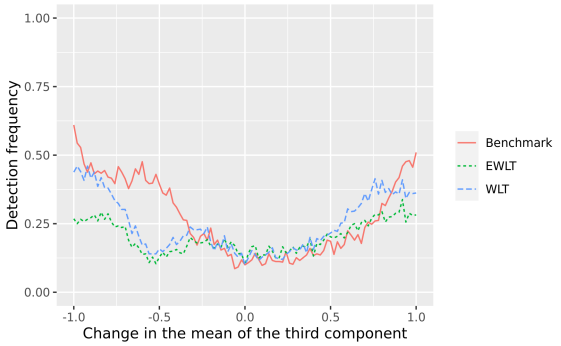
(d) Change in the standard deviation (10k obs.)

(e) Change in the weight parameter (1k obs.)



(f) Change in the weight parameter (10k obs.)

Figure 3.8 – Detection frequency for a change in the third component, *theoretical* initialization.



(a) Change in the mean (1k obs.)

(b) Change in the mean (10k obs.)

(c) Change in the standard deviation (1k obs.)

(d) Change in the standard deviation (10k obs.)

(e) Change in the weight parameter (1k obs.)

(f) Change in the weight parameter (10k obs.)

Figure 3.9 – Detection frequency for a change in the third component, *standard* initialization.

3.9 Glossary of notations

$\hat{A}$	$= -\frac{1}{n}\sum_{i=1}^{n}\left(D^2_\theta(\log f)(X_i,\boldsymbol{\theta}) + \frac{1}{2}\sum_{l=1}^{d}\left(\hat{\theta}_l - \boldsymbol{\theta}_l\right) D^3_\theta(\log f)(X_i,\theta')_{l,.,.}\right)$ defined in (3.13).
$\hat{A}_{0,s}$ or $\hat{A}_{s,1}$	$\hat{A}_{s_1,s_2} = -\frac{1}{\lfloor s_2 n\rfloor - \lfloor s_1 n\rfloor}\sum_{i=\lfloor s_1 n\rfloor+1}^{\lfloor s_2 n\rfloor}\left(D^2_\theta(\log f)(X_i,\boldsymbol{\theta})\right.$ $\left.+\frac{1}{2}\sum_{l=1}^{d}(\hat{\theta}_{s_1,s_2;l} - \boldsymbol{\theta}_l) D^3_\theta(\log f)(X_i,\theta'_{s_1,s_2})_{l,.,.}\right)$ defined in (3.14).
$\boldsymbol{\beta}$	$= \mathbb{E}_{H_0}\left[\log f_1(Y,\boldsymbol{\lambda}_1)\right]$, with Y a r.v. with density $f_1(.,\boldsymbol{\lambda}_1)$. See Section 3.4.
$c_{s,n}$	$= \sum_{i=1}^{\lfloor sn\rfloor} w(X_i,\hat{\theta}_{0,s}) + \sum_{i=\lfloor sn\rfloor+1}^{n} w(X_i,\hat{\theta}_{s,1})$, defined in (3.37).
d	dimension of the parameter set Θ, see Section 3.2.1.
$\mathbb{D}_{[\bar{s},1]}$	$= \mathbb{D}([\bar{s},1],(\mathbb{R}^d)^2 \times gl_d(\mathbb{R}))$. Skorokhod metric space of càd-làg paths, see Lemma 3.12.
$f_k(.,\lambda_k)$	density function of the k-th component, see Section 3.2.1.
$g(.,.,.;.,.)$	map defined in (3.26) for $(\iota,u,I;A,J) \in (\mathbb{R}^d)^2 \times (gl_d(\mathbb{R}))^2 \times \mathbb{R}^{d\times d\times d}$ by $(\iota,u,I;A,J) \mapsto \left(\iota,u,I - \frac{1}{2}\sum_{l=1}^{d}(\iota^T(A^T)_{.,l})J_{.,.,l}\right)$.
$\boldsymbol{g}(.,.,.)$	linear map defined in (3.25) for $(\iota,u,I) \in (\mathbb{R}^d)^2 \times gl_d(\mathbb{R})$ by $(\iota,u,I) \mapsto \left(\iota,u,-\boldsymbol{I}^{-1}\boldsymbol{(}I - \frac{1}{2}\sum_{l=1}^{d}(\iota^T({\boldsymbol{I}^{-1}}^T)_{.,l})\boldsymbol{J}_{.,.,l}\boldsymbol{)}\boldsymbol{I}^{-1}\right)$.
$\hat{\iota}_{0,s}$	$= \frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor} D_\theta(\log f)(X_i,\boldsymbol{\theta})$, defined in (3.19).
$\hat{I}_{0,s}$	$= -\frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor} D^2_\theta(\log f)(X_i,\boldsymbol{\theta})$, defined in (3.19).
$\boldsymbol{I}$	$= -\mathbb{E}_{H_0}\left[D^2_\theta(\log f)(X_1,\boldsymbol{\theta})\right]$. Fisher information matrix defined in (3.1).
$\hat{J}_{0,s}$	$= \frac{1}{\lfloor sn\rfloor}\sum_{i=1}^{\lfloor sn\rfloor} D^3_\theta(\log f)(X_i,\theta'_{0,s})$ defined in Lemma 3.13.
$\boldsymbol{J}$	$= \mathbb{E}_{H_0}\left[D^3_\theta(\log f)(X_1,\boldsymbol{\theta})\right]$ defined in (3.22).
λ_k	density function parameter of the k-th component, see Section 3.2.1.
$\hat{\lambda}_1$	see $\hat{\theta}$.
$\hat{\lambda}_{0,s,1}$ or $\hat{\lambda}_{s,1,1}$	see $\hat{\theta}_{0,s}$ or $\hat{\theta}_{s,1}$.
Λ_n	$= (\Lambda_{s,n})_{s\in[\bar{s},1-\bar{s}]}$. Detection process of the WL test, defined in (3.5).
$\Lambda_{s,n}$	$= Q^1_{s,n} + Q^2_{s,n} - Q^1_{1,n}$ defined in (3.5) (see also (3.9)).
Λ^*_n	$= (\Lambda^*_{s,n})_{s\in[\bar{s},1-\bar{s}]}$. Detection process of the EWL test, defined in (3.38).
$\Lambda^*_{s,n}$	$= \frac{c_{1,n}}{c_{s,n}}\Lambda_{s,n} + \left(\frac{c_{1,n}}{c_{s,n}} - 1\right)\sum_{i=1}^{n} w(X_i,\hat{\theta})\log f_1(X_i,\hat{\lambda}_1)$ defined in (3.38).
m	number of components in the mixture, see Section 3.2.1.
p_k	weight of the k-th component in the mixture, see Section 3.2.1.
$\Phi(.)$	map defined in (3.28) by $(x_s)_{s\in[\bar{s},1]} \mapsto (\varphi(x_s))_{s\in[\bar{s},1]}$.
$q(.,.,.;.,.)$	map defined in (3.32) by $(\iota,u,I;A,J) \mapsto u^T A\iota + \iota^T A^T JA\iota + \boldsymbol{u}^T I\iota$.
$\boldsymbol{q}(.)$	map defined in (3.33) by $z \in \mathbb{R}^{2d+d^2} \mapsto q(\boldsymbol{g}(\Sigma z);\boldsymbol{I}^{-1},\boldsymbol{U})$.
$\boldsymbol{q}^*(.)$	map defined in (3.39) for $z \in \mathbb{R}^{3d+d^2}$ by $z \mapsto q^*(\boldsymbol{g}^*(\Sigma^* z);\boldsymbol{I}^{-1},\boldsymbol{U},\boldsymbol{V})$.
Q^1_n	$= (Q^1_{s,n})_{s\in[\bar{s},1]}$ defined in (3.10).
$Q^1_{s,n}$	$= \sum_{i=1}^{\lfloor sn\rfloor}\left(w(X_i,\hat{\theta}_{0,s})\log f_1(X_i,\hat{\lambda}_{0,s,1}) - w(X_i,\boldsymbol{\theta})\log f_1(X_i,\boldsymbol{\lambda}_1)\right)$ $-\boldsymbol{u}^T\boldsymbol{I}^{-1}\sum_{i=1}^{\lfloor sn\rfloor} D_\theta(\log f)(X_i,\boldsymbol{\theta})$ defined in (3.10).
Q^2_n	$= (Q^2_{s,n})_{s\in[\bar{s},1-\bar{s}]}$ defined in (3.10).
$Q^2_{s,n}$	$= \sum_{i=\lfloor sn\rfloor+1}^{n}\left(w(X_i,\hat{\theta}_{s,1})\log f_1(X_i,\hat{\lambda}_{s,1,1}) - w(X_i,\boldsymbol{\theta})\log f_1(X_i,\boldsymbol{\lambda}_1)\right)$ $-\boldsymbol{u}^T\boldsymbol{I}^{-1}\sum_{i=\lfloor sn\rfloor+1}^{n} D_\theta(\log f)(X_i,\boldsymbol{\theta})$ defined in (3.10).
$\bar{s}$	$\in (0,0.5)$. We assume that the change-point is in $[\bar{s},1-\bar{s}]$ if it exists (Assumption 3.1).
S_n	$= \sup_{s\in[\bar{s},1-\bar{s}]}\Lambda_{s,n}$ defined in (3.6). Detection statistic of the WL test.
S^*_n	$= \sup_{s\in[\bar{s},1-\bar{s}]}\Lambda^*_{s,n}$ defined in Section 3.4. Detection statistic of the EWL test.
Σ	unique positive semi-definite square root of the covariance matrix of the i.i.d. terms in the average $\hat{\xi}_{0,1}$, see Lemma 3.12.

$\hat{\theta}$	$= (\hat{p}_1, \ldots, \hat{p}_{m-1}, \hat{\lambda}_1, \ldots, \hat{\lambda}_m)$. One consistent sequence of solutions of the likelihood equations over the sample $X_1, \ldots, X_n$, see Section 3.2.1.
$\hat{\theta}_{0,s}$ or $\hat{\theta}_{s,1}$	$\hat{\theta}_{s_1,s_2} = (\hat{p}_{s_1,s_2,1}, \ldots, \hat{p}_{s_1,s_2,m-1}, \hat{\lambda}_{s_1,s_2,1}, \ldots, \hat{\lambda}_{s_1,s_2,m})$. One consistent sequence of solutions of the likelihood equations over the sample $X_{\lfloor s_1 n \rfloor + 1}, \ldots, X_{\lfloor s_2 n \rfloor}$, see Section 3.2.1.
$\boldsymbol{\theta}$	true parameter under the null hypothesis, see Section 3.2.1.
θ'	point on the segment $[\hat{\theta}, \boldsymbol{\theta}]$, see (3.12).
$\theta'_{0,s}$ or $\theta'_{s,1}$	θ'_{s_1,s_2} is a point on the segment $[\hat{\theta}_{s_1,s_2}, \boldsymbol{\theta}]$, see (3.14).
Θ_0	parameter set for the weights of the mixture components, see Section 3.2.1.
Θ'	convex subset of Θ for which Assumption 3.5 holds.
$\hat{u}_{0,s}$	$= \frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta (w \log f_1)(X_i, \boldsymbol{\theta})$, defined in (3.19).
$\boldsymbol{u}$	$= \mathbb{E}_{H_0} \left[D_\theta (w \log f_1)(X_1, \boldsymbol{\theta}) \right]$ defined in (3.8).
$\hat{U}_{0,s}$	$= \frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D^2_\theta (w \log f_1)(X_i, \theta'_{0,s})$, see proof of Theorem 3.16.
$\boldsymbol{U}$	$= \mathbb{E}_{H_0} \left[D^2_\theta (w \log f_1)(X_1, \boldsymbol{\theta}) \right]$ defined in (3.34).
$\hat{v}_{0,s}$	$= \frac{1}{\lfloor sn \rfloor} \sum_{i=1}^{\lfloor sn \rfloor} D_\theta w(X_i, \boldsymbol{\theta})$, see Section 3.4.
$\boldsymbol{v}$	$= \mathbb{E}_{H_0} \left[D_\theta w(X_1, \boldsymbol{\theta}) \right]$, see Section 3.4.
$\boldsymbol{V}$	$= \mathbb{E}_{H_0} \left[D^2_\theta w(X_1, \boldsymbol{\theta}) \right]$, see Section 3.4.
$w(x, \theta)$	$= (p_1 f_1(x, \lambda_1))/(f(x, \theta))$ defined in (3.2).
$w \log f_1$	application defined in (3.4) by $(x, \theta) \mapsto w(x, \theta) \log f_1(x, \lambda_1)$.
$\hat{\xi}_{0,s}$	$= (\hat{\iota}_{0,s}, \hat{u}_{0,s} - \boldsymbol{u}, \hat{I}_{0,s})$, defined in (3.19).
$\boldsymbol{\xi}$	$= (0, 0, \boldsymbol{I})$ defined in (3.21).
$\boldsymbol{\xi}_\cdot$	constant process s.t. $\boldsymbol{\xi}_s = \boldsymbol{\xi}$ for all $s \in [\bar{s}, 1]$, see proof of Theorem 3.14.
$\hat{\xi}'_{0,s}$	$= (\hat{\iota}_{0,s}, \hat{u}_{0,s} - \boldsymbol{u}, \hat{A}^{-1}_{0,s})$ defined in (3.24).
$\boldsymbol{\xi}'$	$= (0, 0, \boldsymbol{I}^{-1})$, see Theorem 3.14.
$\hat{\xi}''_{0,\cdot}$	$= (\hat{\xi}''_{0,s})_{s \in [\bar{s}, 1]}$, see proof of Theorem 3.14.
$\hat{\xi}''_{0,s}$	$g(\hat{\xi}_{0,s}; \hat{A}^{-1}_{0,s}, \hat{J}_{0,s})$ if (3.27) holds, $\boldsymbol{\xi}$ otherwise. See proof of Theorem 3.14.
$\hat{\xi}^*_{0,s}$	$= (\hat{\iota}_{0,s}, \hat{u}_{0,s} - \boldsymbol{u}, \hat{v}_{0,s} - \boldsymbol{v}, \hat{I}_{0,s})$, see Section 3.4.
$\boldsymbol{\xi}^*$	$= (0, 0, 0, \boldsymbol{I})$, see Section 3.4.

Bibliography

A., D., Habart, M., Rainer, C., and Sow, A. (2018). Exploring the longevity risk using statistical tools derived from the Shiryaev-Roberts procedure. *Eur. Actuar. J.*, 8(1):27–51.

Abraham, R., Marsden, J. E., and Ratiu, T. (1988). *Manifolds, tensor analysis, and applications*, volume 75 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition.

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *J. Bus. Econom. Statist.*, 25(2):177–190.

Andersen, T. G., Davis, R. A., Kreiß, J.-P., and Mikosch, T. V. (2009). *Handbook of financial time series*. Springer-Verlag Berlin Heidelberg.

Andrews, D. W. K. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62(6):1383–1414.

Barber, D., Cemgil, A. T., and Chiappa, S., editors (2011). *Bayesian time series models*. Cambridge University Press, Cambridge.

Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*. Prentice Hall Information and System Sciences Series. Prentice Hall, Inc., Englewood Cliffs, NJ.

Basu, A., Shioya, H., and Park, C. (2011). *Statistical inference*, volume 120 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL. The minimum distance approach.

Beibel, M. (1996). A note on Ritov's Bayes approach to the minimax property of the cusum procedure. *Ann. Statist.*, 24(4):1804–1812.

Benaglia, T., Chauveau, D., Hunter, D., and Young, D. (2009). mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software*, 32.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition. A Wiley-Interscience Publication.

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems*, volume 243 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. Y. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208.

Caselli, G., Vallin, J., Vaupel, J. W., and Yashin, A. (1987). Age-specific mortality trends in france and italy since 1900: period and cohort effects. *European Journal of Population/Revue européenne de démographie*, 3(1):33–60.

Chatterjee, A. (2012). Detection of change points: a survey of methodologies. *Adv. Appl. Stat.*, 27(2):131–165.

Chen, J. (2017). Consistency of the MLE under mixture models. *Statist. Sci.*, 32(1):47–63.

Chen, J. and Gupta, A. K. (2012). *Parametric statistical change point analysis*. Birkhäuser/Springer, New York, second edition. With applications to genetics, medicine, and finance.

Coleman, R. (2012). *Calculus on normed vector spaces*. Universitext. Springer, New York.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J.

Croix, J.-C., Planchet, F., and Thérond, P.-E. (2015). Mortality: a statistical approach to detect model misspecification. *Bulletin Français d'Actuariat*, 15:75–112.

Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. With a foreword by David Kendall.

Cutler, D. and Meara, E. (2001). Changes in the age distribution of mortality over the 20th century. Technical report, National bureau of economic research.

Davis, R. A., Huang, D. W., and Yao, Y.-C. (1995). Testing for a change in the parameter values and order of an autoregressive model. *Ann. Statist.*, 23(1):282–304.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474.

De Jong, P. and Boyle, P. P. (1983). Monitoring mortality: A state-space approach. *Journal of Econometrics*, 23(1):131–146.

Dehling, H., Franke, B., Kott, T., and Kulperger, R. (2014). Change point testing for the drift parameters of a periodic mean reversion process. *Stat. Inference Stoch. Process.*, 17(1):1–18.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38. With discussion.

Dhaene, J., Dhaene, G., Beirlant, J., Claeskens, G., Croux, C., Degryse, H., Dewachter, H., Gijbels, I., Goovaerts, M., Hubert, M., Roodhooft, F., Schouten, W., and Willekens, M. (2002). Managing uncertainty: Financial, actuarial and statistical modeling. *Review of Business and Economics*, L:23–48.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.*, 42:204–223.

Didou, M. (2011). Amélioration de la mortalité : tendances passées et projections en france, en espagne, au royaume-uni et aux etats-unis. *Mémoire de l'Institut des Actuaires, ISFA*.

Doob, J. L. (1935). The limiting distributions of certain statistics. *Ann. Math. Statist.*, 6(3):160–169.

Dudley, R. M. and Norvaiša, R. (2011). *Concrete functional calculus*. Springer Monographs in Mathematics. Springer, New York.

Durrett, R. (2010). *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition.

El Karoui, N., Loisel, S., and Salhi, Y. (2017). Minimax optimality in robust detection of a disorder time in doubly-stochastic Poisson processes. *Ann. Appl. Probab.*, 27(4):2515–2538.

Everitt, B. S. and Hand, D. J. (1981). *Finite mixture distributions*. Chapman & Hall, London-New York. Monographs on Applied Probability and Statistics.

Feinberg, E. A. and Shiryaev, A. N. (2006). Quickest detection of drift change for Brownian motion in generalized Bayesian and minimax settings. *Statist. Decisions*, 24(4):445–470.

Feng, Z. D. and McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(3):609–617.

Foster, D. P. and George, E. I. (1993). Estimation up to a change-point. *Ann. Statist.*, 21(2):625–644.

Fotopoulos, S. B., Jandhyala, V. K., and Khapalova, E. (2010). Exact asymptotic distribution of change-point mle for change in the mean of Gaussian sequences. *Ann. Appl. Stat.*, 4(2):1081–1104.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):495–580. With 32 discussions by 47 authors and a rejoinder by the authors.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer, New York.

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors (2019). *Handbook of mixture analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL.

Gandy, A., Jensen, U., and Lütkebohmert, C. (2005). A cox model with a change-point applied to an actuarial problem. *Brazilian Journal of Probability and Statistics*, pages 93–109.

Ganji, M. and Mostafayi, R. (2019). Bayes and non-Bayes estimation of change point in nonstandard mixture inverse Weibull distribution. *J. Stat. Theory Appl.*, 18(1):79–86.

Giordani, P. and Kohn, R. (2008). Efficient Bayesian inference for multiple change-point and mixture innovation models. *J. Bus. Econom. Statist.*, 26(1):66–77.

Girshick, M. A. and Rubin, H. (1952). A Bayes approach to a quality control model. *Ann. Math. Statistics*, 23:114–125.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64(2):413–430.

Hathaway, R. J. (1983). *Constrained maximum likelihood estimation for a mixture of m univariate normal distributions*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.), Rice University.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.*, 13(2):795–800.

Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *Canad. J. Statist.*, 30(3):347–371.

Karr, A. F. (1993). *Probability*. Springer Texts in Statistics. Springer-Verlag, New York.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27:887–906.

Killick, R. and Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19.

Knaus, J., Porzelius, C., Binder, H., and Schwarzer, G. (2009). Easier parallel computing in R with snowfall and sfCluster. *The R Journal*, 1(1):54–59.

Kotz, S., Johnson, N. L., and Read, C. B., editors (1985). *Encyclopedia of statistical sciences. Vol. 5*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York. Lindeberg condition to multitrait-multimethod matrices.

Kwon, H. S. and Vu, U. Q. (2017). Consideration of a structural-change point in the chain-ladder method. *Communications for Statistical Applications and Methods*, 24(3):211–226.

Lachos Dávila, V. H., Cabral, C. R. B., and Zeller, C. B. (2018). *Finite mixture of skewed distributions*. SpringerBriefs in Statistics. Springer, Cham. SpringerBriefs in Statistics—ABE.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Li, Q., Guo, F., Kim, I., Klauer, S. G., and Simons-Morton, B. G. (2018). A Bayesian finite mixture change-point model for assessing the risk of novice teenage drivers. *J. Appl. Stat.*, 45(4):604–625.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.*, 42:1897–1908.

Maciak, M., Pešta, M., and Peštová, B. (2020). Changepoint in dependent and non-stationary panels. *Statist. Papers*, 61(4):1385–1407.

Mäkeläinen, T., Schmidt, K., and Styan, G. P. H. (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *Ann. Statist.*, 9(4):758–767.

Matthews, D. E., Farewell, V. T., and Pyke, R. (1985). Asymptotic score-statistic processes and tests for constant hazard against a change-point alternative. *Ann. Statist.*, 13(2):583–591.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

Mei, Y., Han, S. W., and Tsui, K.-L. (2011). Early detection of a change in Poisson rate after accounting for population size effects. *Statist. Sinica*, 21(2):597–624.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Ann. Statist.*, 14(4):1379–1387.

Moustakides, G. V. (2004). Optimality of the CUSUM procedure in continuous time. *Ann. Statist.*, 32(1):302–315.

Mouyopa Djitta, L. (2015). Etude de l'aggravation du risque dépendance dans un contexte de réassurance. *Mémoire de l'Institut des Actuaires, ISUP*.

Norden, R. H. (1973). A survey of maximum likelihood estimation. II. *Internat. Statist. Rev.*, 41(1):39–58.

Olivieri, A., Pitacco, E., et al. (2002). Inference about mortality improvements in life annuity portfolios. In *27th International Congress of Actuaries, Cancun, Mexico*.

O'Meara, O. T. (2000). *Introduction to quadratic forms*. Classics in Mathematics. Springer-Verlag, Berlin. Reprint of the 1973 edition.

Oueslati, A. and Lopez, O. (2013). A proportional hazards regression model with change-points in the baseline function. *Lifetime Data Anal.*, 19(1):59–78.

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

Pandya, M. and Jadav, P. V. (2009). Bayes estimation of change point in non-standard mixture distribution of inverse Weibull with unknown proportions. *J. Indian Statist. Assoc.*, 47(1):41–62.

Pandya, M. and Jadav, P. V. (2010). Bayes estimation of change point in non standard mixture of left-truncated exponential with unknown proportions. *Comm. Statist. Theory Methods*, 39(15):2725–2742.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London Series A*, 185:71–110.

Peštová, B. and Pešta, M. (2017). Change point estimation in panel data without boundary issue. *Risks*, 5(1):7.

Planchet, F. and Tomas, J. (2014). Construction et validation des références de mortalité de place. *Institut des Actuaires, Note de travail*, II(1291-11,v1.4).

Pollak, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.*, 13(1):206–227.

Pollak, M. (2009). The shiryaev–roberts changepoint detection procedure in retrospect—theory and practice. In *Proceedings of the 2nd International Workshop on Sequential Methodologies, University of Technology of Troyes, Troyes, France*.

Pollak, M. and Tartakovsky, A. G. (2009). Optimality properties of the Shiryaev-Roberts procedure. *Statist. Sinica*, 19(4):1729–1739.

Polunchenko, A. S. and Tartakovsky, A. G. (2010). On optimality of the Shiryaev-Roberts procedure for detecting a change in distribution. *Ann. Statist.*, 38(6):3445–3457.

Pons, O. (2009). *Estimation et tests dans les modèles de mélanges de lois et de ruptures*. Hermes Science Publications/Lavoisier, Paris.

Pons, O. (2018). *Estimations and tests in change-point models*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ.

Poor, H. V. and Hadjiliadis, O. (2009). *Quickest detection*. Cambridge University Press, Cambridge.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.*, 9(1):225–228.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, 26(2):195–239.

Rhodes, T. E. and Freitas, S. A. (2004). Advanced statistical analysis of mortality. *AFIR papers. Boston Colloquia. Westwood: MIB inc.*

Roberts, S. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250.

Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics*, 8:411–430.

Robine, J.-M., Cheung, S. L. K., Le Roy, S., Van Oyen, H., Griffiths, C., Michel, J.-P., and Herrmann, F. R. (2008). Death toll exceeded 70,000 in europe during the summer of 2003. *Comptes Rendus Biologies*, 331(2):171 – 178.

Servier, A. (2010). Etude de la stabilité et de la fiabilité des données nécessaires à la construction de tables d'expérience. *Mémoire de l'Institut des Actuaires, ISUP*.

Shiryaev, A. N. (1961). The problem of quickest detection of a violation of stationary behavior. *Dokl. Akad. Nauk SSSR*, 138:1039–1042.

Shiryaev, A. N. (1963). Optimal methods in quickest detection problems. *Teor. Verojatnost. i Primenen.*, 8:26–51.

Shiryaev, A. N. (1996). Minimax optimality of the method of cumulative sums (CUSUM) in the continuous time case. *Uspekhi Mat. Nauk*, 51(4(310)):173–174.

Shiryayev, A. N. (1978). *Optimal stopping rules*. Springer-Verlag, New York-Heidelberg. Translated from the Russian by A. B. Aries, Applications of Mathematics, Vol. 8.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Slutsky, E. (1925). Uber stochastische asymptoten und grenzwerte. *Metron*, 5(3):3–89.

Song, Z., Liu, Y., Li, Z., and Zhang, J. (2018). A weighted likelihood ratio test-based chart for monitoring process mean and variability. *J. Stat. Comput. Simul.*, 88(7):1415–1436.

Spreeuwenberg, P., Kroneman, M., and Paget, J. (2018). Reassessing the Global Mortality Burden of the 1918 Influenza Pandemic. *American Journal of Epidemiology*, 187(12):2561–2567.

Tanaka, K. (2009). Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scale parameters. *Scand. J. Stat.*, 36(1):171–184.

Tartakovsky, A., Nikiforov, I., and Basseville, M. (2015). *Sequential analysis*, volume 136 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL. Hypothesis testing and changepoint detection.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Tomas, J. and Planchet, F. (2014). Constructing entity specific projected mortality table: adjustment to a reference. *Eur. Actuar. J.*, 4(2):247–279.

Tomas, J. and Planchet, F. (2015). Prospective mortality tables: taking heterogeneity into account. *Insurance Math. Econom.*, 63:169–190.

Trefethen, L. N. and Bau, III, D. (1997). *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Vallin, J. and Meslé, F. (2010). Espérance de vie : peut-on gagner trois mois par an indéfiniment ? *Population et Sociétés*, 473.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics*, 20:595–601.

Wilson, R. C., Nassar, M. R., and Gold, J. I. (2013). A mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLoS Comput. Biol.*, 9(7):e1003150, 18.

Wu, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1):95–103.

Wu, Y. (2005). *Inference for change-point and post-change means after a CUSUM test*, volume 180 of *Lecture Notes in Statistics*. Springer, New York.

Wu, Y. (2015). Estimation of change-point and post-change means by adaptive cusum procedures. *Conference Paper: Joint Statistical Meeting*.

Wu, Y. (2016a). Inference after truncated one-sided sequential test. *Comm. Statist. Theory Methods*, 45(10):3076–3094.

Wu, Y. (2016b). Inference for post-change parameters after sequential CUSUM test under AR(1) model. *J. Statist. Plann. Inference*, 168:52–67.

Zou, C., Fang, X., and Zhai, G. (2015). Ovarian cancer screening based on mixture change-point model. *J. Syst. Sci. Complex.*, 28(2):471–488.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov models for time series*, volume 110 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL. An introduction using R.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Université de Bretagne Occidentale

Titre : Études de détection de rupture : procédure en ligne pour des modèles discrets de Poisson et test hors ligne pour des mélanges paramétriques. Application à des problèmes issus de l'assurance.

Mots-clés : Détection de rupture, mélange paramétrique, théorèmes limites faibles de séquences dépendantes, test de ratio de vraisemblance pondérée, estimateur séquentiel convergent, procédure de Shiryaev-Roberts, mortalité nationale française, application à des données d'assurance.

Résumé : Cette thèse traite de deux études distinctes de techniques de détection de rupture.
Le premier problème concerne l'estimation séquentielle du paramètre post-changement dans le cas d'une séquence discrète de variables aléatoires de Poisson. Nous introduisons un estimateur alternatif qui est basé sur la statistique de la procédure de Shiryaev-Roberts. Nous montrons qu'il s'agit d'un estimateur convergent. Les applications numériques indiquent que, juste après le changement, il a un biais et une variance significativement plus faibles que ceux de l'estimateur du maximum de vraisemblance, avec des propriétés asymptotiques similaires. Des applications à des données réelles d'assurance sont présentées, pour des cas où le changement est évident ou non.
Dans la seconde étude, nous introduisons deux tests d'hypothèse qui permettent de détecter une rupture dans la première composante d'un mélange fini de lois paramétriques, pour un échantillon où au plus un changement se produit. Chaque test repose sur un ratio de vraisemblance pondérée qui est calculable à l'aide d'algorithmes d'estimation standards. Avec une technique issue de Davis et al. (1995), nous obtenons sous l'hypothèse nulle les lois limites des statistiques de test comme des formes quadratiques d'un mouvement Brownien, à l'aide d'un théorème limite fonctionnel dédié. Nous montrons que les conditions de validité du résultat limite sont valides pour le cas Gaussien dans le cadre donné par Hathaway (1985). Des applications numériques sur des données simulées illustrent les avantages des tests alternatifs comparés à un test standard de référence. Une application illustrative sur des données réelles d'assurance non-vie est donnée pour les tests alternatifs.

Title: Studies of change-point detection: on-line scheme for discrete Poisson models and off-line test for parametric mixtures. Application to insurance problems.

Keywords: Change-point detection, parametric mixture, weak limit theorems for dependent sequences, weighted likelihood quotient test, consistent sequential estimator, Shiryaev-Roberts scheme, French mortality, applications to insurance data.

Abstract: This thesis concerns two distinct studies of change-point detection techniques.
The first problem deals with the sequential estimation of the post-change distribution when observing an on-line sequence of Poisson random variables. We introduce an alternative estimator based on the detection statistic of the Shiryaev-Roberts procedure. We show that it is a consistent estimator for the true post-change parameter. Numerical simulations indicate that, compared to the usual Maximum Likelihood Estimator, it has significantly reduced bias and variance just after the change, with identical asymptotic properties. Applications on real data from the insurance industry are provided, in a context where the change is obvious or not.
In the second study, we introduce two alternative hypothesis tests that aim to detect a change-point in the first component of a finite parametric mixture, for a closed sample where at most one change occurs. They are based on weighted likelihood ratios that can be computed with standard inference algorithms. With a technique from Davis et al. (1995), we derive the limit distribution of their statistics under the null hypothesis in the form of quadratic forms of multidimensional Brownian motions, with the help of a dedicated functional limit theorem. We show that validity conditions of the main result hold for univariate finite Gaussian mixtures within the framework of Hathaway (1985). Numerical applications on simulated data illustrate the advantage of the alternative tests compared to a standard benchmark test. An application to Property and Casualty insurance real data is provided for the alternative tests.