

# THÈSE DE DOCTORAT EN CO-TUTELLE

présentée pour obtenir le titre de

DOCTEUR

**DE L'ÉCOLE SUPÉRIEURE DES COMMUNICATIONS DE TUNIS**

**Discipline : Technologies de l'Information et de la Communication**

&

**DE L'UNIVERSITÉ DE REIMS CHAMPAGNE ARDENNE**

**Discipline : Automatique, Signal, Productique, Robotique**

**Spécialité : Traitement du Signal**

par

**Nehla DEBBABI**

Sujet de la thèse :

**Approche algébrique et théorie des valeurs extrêmes pour la  
détection de ruptures : Application aux signaux biomédicaux**

Soutenue à Tunis le 14 Décembre 2015 devant le jury composé de :

|                     |   |                                                             |                               |
|---------------------|---|-------------------------------------------------------------|-------------------------------|
| M. Sofiane CHERIF   | : | Professeur à l'École Supérieure des Communications de Tunis | Président et Examineur        |
| Mme. Meriem JAÏDANE | : | Professeure à l'École Nationale d'Ingénieurs de Tunis       | Rapporteur                    |
| M. Cédric JOIN      | : | Maître de Conférences à l'Université de Lorraine, Nancy     | Rapporteur                    |
| Mme. Marie KRATZ    | : | Professeure à l'ESSEC Business School, Paris                | Co-encadrante et Examinatrice |
| M. Sadok EL ASMI    | : | Professeur à l'École Supérieure des Communications de Tunis | Directeur de thèse            |
| M. Mamadou MBOUP    | : | Professeur à l'Université de Reims Champagne Ardenne        | Directeur de thèse            |
| Mme. Julie CARREAU  | : | Ingénieur de recherche à l'Université de Montpellier        | Invitée                       |





*À mes chers parents Kabil et Amel.*  
*À mes frères Sadok, Mejdi et Haythem.*  
*À ma grand-mère Beya.*

# Remerciements

*"Soyons reconnaissants aux personnes qui nous donnent du bonheur ; elles sont les charmants jardiniers par qui nos âmes sont fleuries."*

Marcel Proust

Je commence par adresser mes vifs remerciements à mes directeurs de thèse Pr. Sadok EL ASMI et Pr. Mamadou MBOUP et à ma co-encadrante Pr. Marie KRATZ. Je vous suis reconnaissante pour le temps que vous avez consacré pour diriger mes travaux de thèse, pour vos remarques pertinentes et vos conseils avisés, aussi bien pour vos compétences et rigueur scientifique que vous m'avez fait bénéficier tout au long de mes études doctorales. Merci pour votre gentillesse et sympathie. Je suis très honorée que vous fassiez partie des personnes qui ont marqué ma vie.

Ensuite, je tiens à exprimer ma gratitude à Mme Meriem JAÏDANE, Professeure à l'École Nationale des Ingénieurs de Tunis, et M. Cédric JOIN, Maître de conférences HDR à l'Université de Lorraine, pour l'intérêt qu'ils ont porté à mon travail en s'engageant à être rapporteurs de ce mémoire de thèse.

Mes sincères remerciements s'adressent également à M. Sofiane CHERIF, Professeur à l'École Supérieure des Communications de Tunis, pour avoir accepté de présider le jury de ma thèse et d'examiner mon manuscrit de thèse. Je le remercie également pour ses qualités scientifiques et humaines qui ont fait de lui un bon Directeur du laboratoire COSIM de Sup'com dont je fais partie.

Je remercie particulièrement Mme Julie CARREAU, Ingénieur de recherche à l'Université de Montpellier, pour avoir accepté de faire partie du jury de ma thèse. Sa présence enrichira sans doute la discussion de mes travaux de thèse.

Mes plus profonds remerciements vont à mes parents qui n'ont jamais cessé de me soutenir et de m'encourager sur tous les plans pour achever ce projet de thèse. Papa, Maman, je vous en suis reconnaissante. Je remercie également mes frères pour leur présence à mes côtés durant mes années de thèse. Sans l'aide de ma famille, l'accomplissement de ce travail ne pourrait jamais voir le jour.

Mes remerciements vont également à tous mes amis, qu'ils soient en Tunisie, en France ou partout dans le monde, à tous les membres du laboratoire COSIM de Sup'com et du laboratoire

CReSTIC de l'URCA pour tous les moments agréables qu'on a passé ensemble. Merci pour les échanges scientifiques, culturels et linguistiques.

Une pensée à l'ensemble du personnel administratif de Sup'com et de son école doctorale EDTIC, et de l'URCA et de son école doctorale STS, pour leur serviabilité et pour l'effort qu'ils ont fait pour que ma thèse soit soutenue en temps voulu. Un grand merci à vous. Je n'oublierai pas de remercier aussi Mme Jihène OUERGHI, la secrétaire du laboratoire COSIM et Mme Ida LENCLUME, la secrétaire du laboratoire CReSTIC, pour leur gentillesse et disponibilité pour nous aider à effectuer les tâches administratives. Merci également pour tous les moments de joie partagés.

Enfin, ma reconnaissance va à tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail de thèse.

# Table des figures

|      |                                                                                                                                                                                                                                            |    |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1  | Exemples de caractérisation d'une rupture dans un signal gaussien simulé : (a) un changement de moyenne, (b) un changement de variance, (c) un changement transitoire abrupt, (d) un changement graduel, (f) un changement de fréquence. . | 13 |
| 1.2  | Exemple de décomposition d'un signal enregistré en composante déterministe (ruptures) et composante stochastique (bruit). Les points bleus associés au signal enregistré indiquent la présence de ruptures. . . . .                        | 19 |
| 1.3  | Schéma d'organisation de la thèse. . . . .                                                                                                                                                                                                 | 21 |
| 2.1  | Trois différents types de ruptures . . . . .                                                                                                                                                                                               | 26 |
| 2.2  | $\left  \frac{\lambda^-}{\lambda_1^+} \right $ et $\left  \frac{\lambda^-}{\lambda_2^+} \right $ pour $\nu \in \{3, \dots, 20\}$ , $\kappa \in \{2, 4, 8, 10\}$ et $M = 60$ . . . . .                                                      | 38 |
| 2.3  | La cumulative distribution function (cdf) de $\mathcal{D}_{4,n}^\omega$ vs la cdf de $\tilde{\mathcal{D}}_{4,n}^\omega$ pour des différentes valeurs de $\kappa$ et $\nu$ . . . . .                                                        | 40 |
| 2.4  | Potentiel d'action . . . . .                                                                                                                                                                                                               | 42 |
| 2.5  | 1s d'enregistrement d'activité neuronale du lobe antennaire d'un criquet . . . . .                                                                                                                                                         | 43 |
| 2.6  | Différentes formes de spikes . . . . .                                                                                                                                                                                                     | 44 |
| 2.7  | Représentation de 0.2s d'un signal neuronal généré pour SNR= 4 et FR= 30kHz. . . . .                                                                                                                                                       | 45 |
| 2.8  | Les courbes Receiver Operating Characteristic (ROC) de $\{\mathcal{D}_i\}_{i \in \{1, \dots, 4\}}$ . . . . .                                                                                                                               | 46 |
| 2.9  | Application de la fonction de décision $\mathcal{D}_4$ pour la détection des spikes de neurones. . . . .                                                                                                                                   | 48 |
| 2.10 | Les courbes ROC des trois méthodes de détection des spikes : la technique algébrique (►), la méthode Nonlinear Energy Operator (NEO) (●) et la méthode par transformation des ondelettes (■). . . . .                                      | 49 |
| 2.11 | pfa fixées vs pfa empiriques obtenues en appliquant le seuil déterminé dans la section 2.3.2 sur un signal neuronal généré pour $SNR = 4$ , $FR = 45$ , $\kappa = 3$ , $\mu = 7$ , et $M = 60$ . . . . .                                   | 50 |
| 2.12 | Signal ElectroCardioGramme (ECG) d'un état sain pour un seul cycle cardiaque . . . . .                                                                                                                                                     | 51 |
| 2.13 | Signal ECG d'un état sain . . . . .                                                                                                                                                                                                        | 52 |
| 2.14 | Interpolation des maximas par les splines cubiques et réduction des fausses alarmes : application sur le signal 108 de la bas MIT-BIH. Ici l'ordre de dérivation est 3. . . . .                                                            | 54 |
| 2.15 | Elimination de fausses alarmes . . . . .                                                                                                                                                                                                   | 55 |
| 2.16 | Application de l'approche algébrique pour la détection des pics R du signal ECG . . . . .                                                                                                                                                  | 57 |

|     |                                                                                                                                                                                                                                                               |     |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.1 | Représentation de la cdf et de la probability density function (pdf) de la loi Generalized Extreme Value (GEV) pour des différentes valeurs de $\xi$ : $\xi = 1 > 0$ : loi de Fréchet, $\xi = 0$ : loi de Gumbel, et $\xi = -1 < 0$ : loi de Weibull. . . . . | 62  |
| 3.2 | Représentation de la cdf et de la pdf de la Generalized Pareto Distribution (GPD) pour $u = 0$ , $\beta = 1$ et $\xi \in \{-0.2, 0, 0.5\}$ . . . . .                                                                                                          | 64  |
| 3.3 | Exemple de pdf hybride avec sa composante normale . . . . .                                                                                                                                                                                                   | 73  |
| 3.4 | Etude du comportement de la suite récurrente $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$ suivant la valeur initiale de $\tilde{u}^{(0)}$ . Exemple 1 pour $\theta = [0.4354, 0, 1]$ . . . . .                                        | 86  |
| 3.5 | Etude du comportement de la suite récurrente $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$ suivant la valeur initiale de $\tilde{u}^{(0)}$ . Exemple 2 pour $\theta = [4.0443, 3, 2]$ . . . . .                                        | 86  |
| 3.6 | Exemple de pdf hybride Rayleigh-GPD pour $\theta = [1, 1.3716]$ . . . . .                                                                                                                                                                                     | 87  |
| 3.7 | Exemple de pdf hybride Gauss-GPD pour $\theta = [0, 1, 0.9, 0.2]$ . . . . .                                                                                                                                                                                   | 93  |
| 3.8 | Exemple de pdf hybride GPD-Gauss-GPD pour $\theta = [-1.5, 1.5, 0, 1, 0.2, 0.5]$ . . . .                                                                                                                                                                      | 101 |



# Table des matières

|                                                                                   |           |
|-----------------------------------------------------------------------------------|-----------|
| <b>Table des figures</b>                                                          | <b>5</b>  |
| <b>Résumé</b>                                                                     | <b>10</b> |
| <b>Abstract</b>                                                                   | <b>11</b> |
| <b>1 Contexte général</b>                                                         | <b>12</b> |
| 1.1 Introduction . . . . .                                                        | 12        |
| 1.2 Exemples de méthodes existantes de détection des ruptures . . . . .           | 15        |
| 1.2.1 Changement dans les caractéristiques statistiques et spectrales du signal . | 15        |
| 1.2.2 Discontinuité dans le signal ou dans l'une de ses fonctionnalités . . . . . | 17        |
| 1.2.3 Événement rare . . . . .                                                    | 18        |
| 1.3 Plan de la thèse . . . . .                                                    | 18        |
| 1.3.1 Description du chapitre 2 . . . . .                                         | 20        |
| 1.3.2 Description du chapitre 3 . . . . .                                         | 22        |
| 1.4 Conclusion . . . . .                                                          | 23        |
| <b>2 Rupture : une discontinuité de première espèce</b>                           | <b>24</b> |
| 2.1 Introduction . . . . .                                                        | 24        |
| 2.2 Modélisation déterministe des ruptures . . . . .                              | 24        |
| 2.2.1 Approche algébrique pour la détection de ruptures : principes . . . . .     | 26        |
| 2.2.2 Approche algébrique pour la détection de ruptures : exemple . . . . .       | 27        |
| 2.2.3 Fonctions de décision . . . . .                                             | 30        |
| 2.2.4 Implémentation des différentes fonctions de décision . . . . .              | 32        |
| 2.3 Modélisation stochastique du bruit . . . . .                                  | 33        |
| 2.3.1 Loi de probabilité du bruit . . . . .                                       | 34        |
| 2.3.2 Sélection du seuil . . . . .                                                | 40        |
| 2.4 Applications . . . . .                                                        | 41        |
| 2.4.1 Détection des spikes dans les signaux neuronaux . . . . .                   | 41        |
| 2.4.2 Détection des ondes R dans les signaux Electrocardiogramme (ECG) . . .      | 50        |
| 2.5 Conclusion . . . . .                                                          | 57        |
| <b>3 Rupture : un évènement rare</b>                                              | <b>59</b> |
| 3.1 Introduction . . . . .                                                        | 59        |
| 3.2 La Théorie des Valeurs Extrêmes . . . . .                                     | 60        |
| 3.2.1 Méthode du Peak Over Threshold . . . . .                                    | 61        |

|                                                      |                                                                                                                                     |            |
|------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|------------|
| 3.3                                                  | Modélisation stochastique des ruptures : résumé du papier [Debbabi et al., 2012]                                                    | 65         |
| 3.4                                                  | Modélisation stochastique de toutes les données : ruptures et bruit . . . . .                                                       | 72         |
| 3.4.1                                                | Modèle hybride à deux composantes Gauss-GPD avec pondération uniforme : résumé du papier [Debbabi and Kratz, 2014] . . . . .        | 72         |
| 3.4.2                                                | Modèle hybride à deux composantes Rayleigh-GPD avec pondération uniforme : résumé du papier [Debbabi et al., 2014] . . . . .        | 87         |
| 3.4.3                                                | Modèle hybride à deux composantes Gauss-GPD avec pondération non-uniforme : résumé du papier [Debbabi et al., 2015a] . . . . .      | 93         |
| 3.4.4                                                | Modèle hybride à trois composantes GPD-Gauss-GPD avec pondération non-uniforme : résumé du papier [Debbabi et al., 2015b] . . . . . | 101        |
| 3.5                                                  | Conclusion . . . . .                                                                                                                | 107        |
| <b>Conclusion générale</b>                           |                                                                                                                                     | <b>108</b> |
| <b>A La variation de l'écart absolu à la médiane</b> |                                                                                                                                     | <b>111</b> |
| <b>Bibliographie</b>                                 |                                                                                                                                     | <b>112</b> |

# Liste des Acronymes

**BM** Block Maxima

**cdf** cumulative distribution function

**ECG** ElectroCardioGramme

**iid** indépendantes et identiquement distribuées

**GEV** Generalized Extreme Value

**GPD** Generalized Pareto Distribution

**NEO** Nonlinear Energy Operator

**pdf** probability density function

**POT** Peak Over Threshold

**ROC** Receiver Operating Characteristic

**RSB** Rapport Signal sur Bruit

**TCL** Théorème Central Limite

**TVE** Théorie des Valeurs Extrêmes

**v.a.** variable aléatoire

# Résumé

Ce travail développe des techniques non-supervisées de détection et de localisation en ligne de ruptures dans les signaux enregistrés dans un environnement bruité. Ces techniques reposent sur l'association d'une approche algébrique avec la Théorie des Valeurs Extrêmes (TVE). L'approche algébrique permet d'appréhender aisément les ruptures en les caractérisant en termes de distributions de Dirac retardées et leurs dérivées dont la manipulation est facile via le calcul opérationnel. Cette caractérisation algébrique, permettant d'exprimer explicitement les instants d'occurrences des ruptures, est complétée par une interprétation probabiliste en termes d'extrêmes : une rupture est un événement rare dont l'amplitude associée est relativement grande. Ces événements sont modélisés dans le cadre de la TVE, par une distribution de Pareto Généralisée. Plusieurs modèles hybrides sont proposés dans ce travail pour décrire à la fois le comportement moyen (bruit) et les comportements extrêmes (les ruptures) du signal après un traitement algébrique. Des algorithmes entièrement non-supervisés sont développés pour l'évaluation de ces modèles hybrides, contrairement aux techniques classiques utilisées pour les problèmes d'estimation en question qui sont heuristiques et manuelles. Les algorithmes de détection de ruptures développés dans cette thèse ont été validés sur des données générées, puis appliqués sur des données réelles provenant de différents phénomènes, où les informations à extraire sont traduites par l'apparition de ruptures.

# Abstract

This work develops non supervised techniques for on-line detection and location of change-points in noisy recorded signals. These techniques are based on the combination of an algebraic approach with the Extreme Value Theory (EVT). The algebraic approach offers an easy identification of the change-points. It characterizes them in terms of delayed Dirac distributions and their derivatives which are easily handled via operational calculus. This algebraic characterization, giving rise to an explicit expression of the change-points locations, is completed with a probabilistic interpretation in terms of extremes : a change point is seen as a rare and extreme event. Based on EVT, these events are modeled by a Generalized Pareto Distribution. Several hybrid multi-components models are proposed in this work, modeling at the same time the mean behavior (noise) and the extremes ones (change-points) of the signal after an algebraic processing. Non supervised algorithms are proposed to evaluate these hybrid models, avoiding the problems encountered with classical estimation methods which are graphical ad hoc ones. The change-points detection algorithms developed in this thesis are validated on generated data and then applied on real data, stemming from different phenomenons, where change-points represent the information to be extracted.

# Chapitre 1

## Contexte général

### 1.1 Introduction

La détection et la localisation des changements brusques dans les caractéristiques d'un système ou d'un signal est un problème particulièrement important, faisant, ces dernières décennies, l'objet de nombreuses recherches dans différents domaines. À titre d'exemples, nous mentionnons :

- *Economie / Finance* : détection de risques du marché [Zumbach et al., 1999], [Fliess et al., 2011a], [Broadwater and Chellappa, 2010], et détection de bulles financières [Watanabe et al., 2007].
- *Biomédecine* : détection de ruptures dans une séquence d'acide désoxyribonucléique (ADN) [Braun et al., 2000] et dans les signaux physiologiques (ECG [Pan and Tompkins, 1985], EEG [Sivasankari and Thanushkodi, 2009], EMG [Matei et al., 1997], signaux neuronaux [Mboup, 2012], etc).
- *Science de l'environnement* : détection de pannes ou de défauts dans les systèmes de dépollution biologique [Verdier, 2007], des tremblements de terre [Nikiforov and Tikhonov, 1986] et des tempêtes [Killick et al., 2010].
- *Télécommunication* : détection de ruptures dans les signaux GPS [Boutoille, 2007], et dans les spectres des signaux en radio cognitive [Ben Jemaa et al., 2012].
- *Politique / Contre-terrorisme* : détection d'attaques biologiques [Stoto et al., 2006]

Un changement brusque, que nous appellerons rupture ou encore *change-point* en anglais, représente souvent une information importante. Une rupture peut être synonyme de panne ou de dysfonctionnement d'un système, d'une anomalie, d'un risque, d'un sinistre, etc. Ainsi, une rupture peut changer de nature d'un problème à un autre. Il n'existe donc pas une unique façon de la définir et de la caractériser. Tout dépend du cadre du travail, du problème traité et de la manière dont nous la voyons. Nous allons voir par la suite que pour un même problème nous pouvons caractériser une rupture de deux ou plusieurs façons différentes.

Compte tenu de l'importance de l'information apportée par les ruptures, la détection de ces dernières est une étape indispensable en vue de l'analyse, la surveillance et le contrôle de nombreux systèmes. Dans ce travail, nous abordons ce sujet avec une modélisation

stochastique ou déterministe des signaux issus des systèmes étudiés ou des phénomènes observés.

Pour illustrer la variété d’approches pour la détection de ruptures, en se référant à la figure 1.1, nous considérons les exemples suivants de la littérature en fonction de la caractérisation de cette dernière.

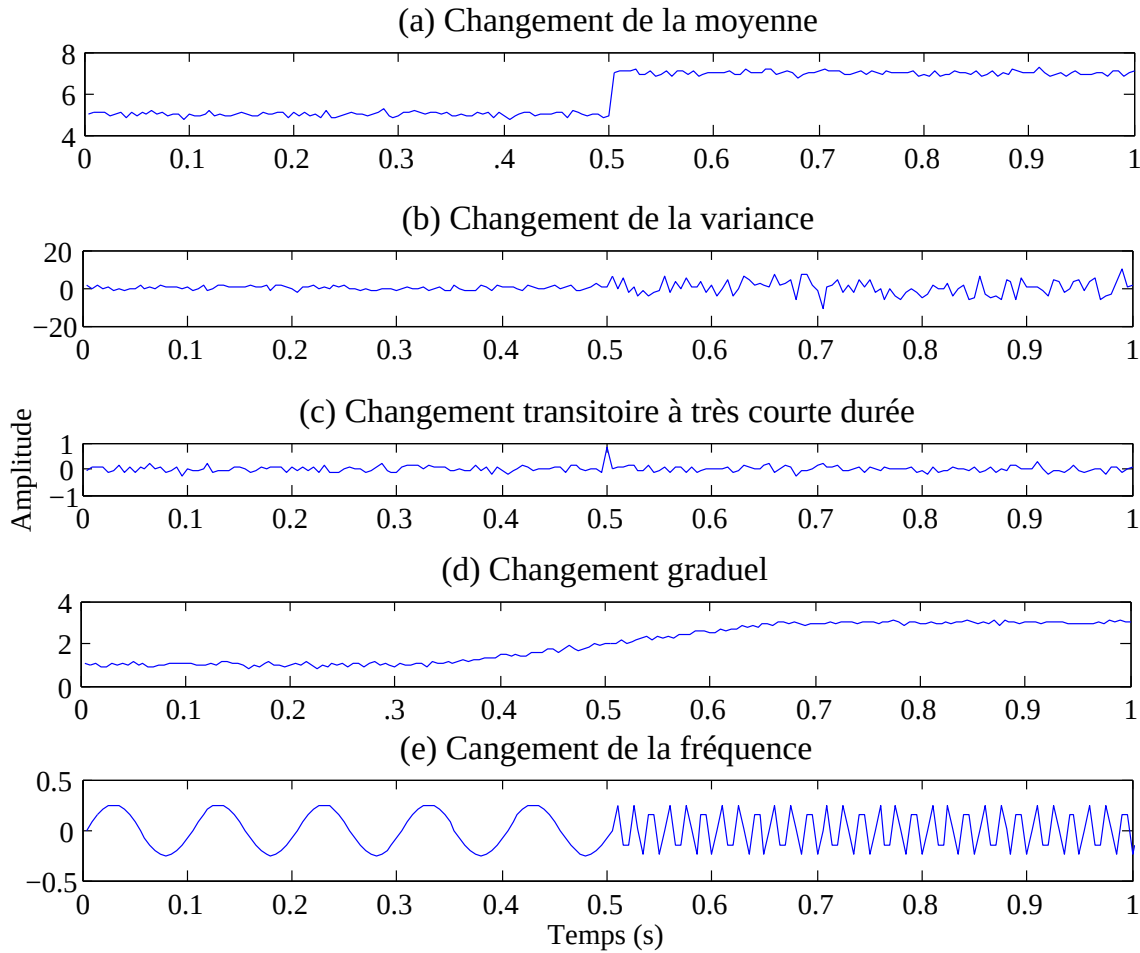


FIGURE 1.1 – Exemples de caractérisation d’une rupture dans un signal gaussien simulé : (a) un changement de moyenne, (b) un changement de variance, (c) un changement transitoire abrupt, (d) un changement graduel, (e) un changement de fréquence.

▷ Cadre stochastique :

- En choisissant un modèle stochastique pour un signal observé, nous pouvons définir la notion de rupture comme un changement brutal d’un paramètre statistique lié aux observations. Ce paramètre peut être la moyenne (signal (a)), la variance (signal (b)), etc. Plusieurs approches statistiques ont alors été développées pour déceler ces changements brutaux, basées par exemple sur les tests à hypothèses. Citons, entre autres, l’approche bayésienne [Adams and MacKay, 2007], [Mohammad-Djafari and Féron, 2006], [Moreno et al., 2005], la méthode de Shewhart [Shewhart, 1931], ou la méthode de la somme cumulée CUSUM [Page, 1954]. Dans un premier temps, ces méthodes

ont été appliquées en présence d'une seule rupture pour des observations supposées indépendantes. Par la suite, elles ont été généralisées pour le cas de données dépendantes [Lavielle and Teyssiere, 2007]. Elles ont été développées également pour le cas de multiples ruptures en utilisant par exemple la segmentation du signal binaire ou la sélection de modèles [Lavielle and Teyssiere, 2007].

- Alors que les signaux (a) et (b) représentaient des changements dans les caractéristiques statistiques du signal, le signal (e) représente un changement dans une caractéristique spectrale du signal. Une rupture peut aussi être envisagée comme un changement brusque de la fréquence à un instant donné. En appliquant l'une des méthodes d'analyse temps-fréquence des signaux ([Sejdić et al., 2009]), comme par exemple, la transformée de Gabor [Gabor, 1946], la transformée de Wigner-Ville [Wigner, 1932] ou encore la transformée en ondelettes [Meyer, 1992], ce changement de fréquence peut être détecté dans le domaine fréquentiel et localisé dans le domaine temporel ([Nenadic, 2005], [Li et al., 1995], ...). Il peut aussi être détecté dans le domaine fréquentiel en utilisant une approche algébrique comme dans les travaux de [Tiganj et al., 2012] et [Ben Jemaa et al., 2012].
- Une autre façon de caractériser une rupture est de considérer son occurrence comme un événement rare, *i.e.* un événement de probabilité d'apparition faible, et extrême, *i.e.* d'amplitude relativement très élevée. Cela nous conduira à considérer la TVE. Cette approche est utilisée pour l'étude des phénomènes dits 'anormaux' apparaissant dans de nombreux domaines, tels par exemple en actuariat et finance ([Embrechts et al., 1997] et [Carreau and Bengio, 2009b],...), en climatologie ([Naveau et al., 2005],[Davison, 2009]), et plus récemment en épidémiologie ([Guillou et al., 2014]).

▷ Cadre déterministe :

Enfin, une autre caractérisation possible est de considérer la rupture comme une discontinuité dans le signal (signaux (a) et (c)) ou dans l'une de ses fonctionnalités : la dérivée par exemple (signal (d)) ou encore dans le spectre du signal (signal(e)). Pour cette caractérisation nous proposons un modèle déterministe du signal, continu par morceaux, faisant apparaître les positions des ruptures. Une manière de faire apparaître des discontinuités est d'introduire les dérivées du signal. Plusieurs méthodes utilisant les dérivées ont été développées pour la détection de ce type de rupture. Nous trouvons par exemple les méthodes basées sur la sélection de modèles, où les discontinuités sont détectées dans les signaux résidus, [Laguna et al., 1989]. Nous citons également l'approche algébrique basée sur l'algèbre différentielle et le calcul opérationnel [Mboup, 2012], qui a été récemment appliquée sur différents jeux de données [Moussa Ali et al., 2011a], [Fliess et al., 2011a], [Mboup, 2012], [Ben Jemaa et al., 2012], etc.

Dans ce qui suit, nous présenterons quelques caractérisations de ruptures en donnant une description brève des principes des approches existantes pour les détecter.



## 1.2 Exemples de méthodes existantes de détection des ruptures

La définition et la caractérisation d'une rupture pour un problème bien défini constitue une étape primordiale pour la détection de changements brusques. Selon cette définition et cette caractérisation, plusieurs méthodes peuvent être développées. Nous présentons dans cette section l'idée générale de certaines en fonction de la caractérisation de la rupture. Dans ce qui suit, nous modélisons un signal échantillonné, représenté par des données, par une suite de variables aléatoires (v.a.)  $(X_n)_{n \in \mathbb{N}^*}$  indépendantes et identiquement distribuées (iid).

### 1.2.1 Changement dans les caractéristiques statistiques et spectrales du signal

Nous considérons dans ce paragraphe l'exemple de la rupture mesurée par un changement d'un paramètre statistique (la moyenne et/ou la variance) ou spectral (la fréquence).

#### 1.2.1.1 Changement dans les caractéristiques statistiques du signal

##### ► Changement de moyenne

Un changement brutal de moyenne admettant deux valeurs distinctes sur deux intervalles de temps disjoints peut être considéré comme une rupture. Désignons par  $\mu$  cette moyenne. Pour détecter le changement brusque de la valeur de  $\mu$ , une approche stochastique classique est le test d'hypothèses sur la moyenne. La première hypothèse appelée hypothèse nulle et notée  $\mathcal{H}_0$  correspond à une série  $(X_n)_{n \in \mathbb{N}^*}$  de moyenne constante. Tandis que la deuxième hypothèse dite alternative et notée  $\mathcal{H}_1$  traduit l'apparition d'une rupture. Comme la rupture représente un changement en moyenne, nous pouvons considérer que la série  $(X_n)_{n \in \mathbb{N}^*}$  n'est autre qu'une concaténation de deux sous-séries  $(X_{0,n})_{n \in \mathbb{N}^*}$  et  $(X_{1,n})_{n \in \mathbb{N}^*}$  de moyennes différentes respectives  $\mu_0$  et  $\mu_1$  dont le passage d'une sous-série à une autre traduit l'apparition d'une rupture.

Nous considérons donc le test à hypothèses suivant :

$$\begin{aligned} \mathcal{H}_0 : \quad \mu &= \mu_0 \\ \mathcal{H}_1 : \quad \mu &= \mu_1 \neq \mu_0 \end{aligned} \tag{1.1}$$

dont le rejet de l'hypothèse  $\mathcal{H}_0$  au risque  $\alpha$  implique la détection d'un changement en moyenne avec une probabilité  $1 - \alpha$ ,  $\alpha \in [0, 1]$ .

De nombreuses méthodes basées sur le principe du test à hypothèses ont été développées en utilisant le logarithme du rapport de vraisemblance. Nous pouvons citer par exemple l'approche bayésienne [Adams and MacKay, 2007], [Mohammad-Djafari and Féron, 2006], [Moreno et al., 2005]. Nous trouvons aussi la règle de Shewhart [Shewhart, 1931], la règle de la moyenne mobile [Page, 1954], aussi la règle de la somme cumulée CUSUM introduite par Page [Page, 1954] et qui est la plus utilisée récemment. Toutes ces méthodes ont été mises en place sur des données indépendantes, puis ont été généralisées au cas de données dépendantes

[Verdier, 2007].

Les limitations de ces méthodes résident dans la nécessité de la connaissance de la densité de probabilité sous chaque hypothèse. Pour palier cette limitation, certains auteurs ont proposé d'estimer le rapport de vraisemblance sans le passage par l'estimation des densités de probabilité sous chaque hypothèse [Kawahara and Sugiyama, 2009].

Pour détecter le changement brusque de la moyenne, nous pouvons aussi utiliser l'une des méthodes basées sur la sélection de modèles. Ces méthodes qui sont le plus souvent utilisées en automatiques pour détecter et localiser la défaillance d'un système, reposent sur la génération des signaux résidus, indicateurs de ruptures dans le fonctionnement du système. Les signaux résidus générés correspondent à la différence entre le comportement du système prédit par le modèle analytique et celui observé. Les méthodes basées sur la sélection de modèles nécessitent une connaissance vaste sur le système permettant sa modélisation mathématique. Nous citons par exemple les méthodes basées sur la génération de résidus par relation de parité, par le filtrage de Kalman ou par estimation paramétrique [Lavielle and Teyssiere, 2006], [Lavielle and Teyssiere, 2007].

#### ► Changement de variance

Nous considérons maintenant le cas d'une rupture mesurée par un changement net de variance. Ce type de ruptures se rencontre souvent dans le domaine de la finance pour mesurer le risque du marché en s'appuyant sur les données représentant les fluctuations du prix entre le vendeur et l'acheteur. Ces fluctuations augmentent lors d'une turbulence du marché. La détection de ces dernières est réalisée via le changement brusque de la variance de ces données appelée volatilité en finance. Cette caractérisation de la rupture a été traitée par Dacorogna et *al.* [Zumbach et al., 1999] et aussi par Fliess et *al.* [Fliess et al., 2011b] dans deux cadres différents. Le changement de variance peut indiquer l'apparition d'une rupture dans d'autres domaines. Citons, par exemple, les tremblements de terre, où Richter en 1958, pour construire son échelle, s'est basé sur la détection du changement de la variance représentant la somme de l'énergie libérée durant le tremblement de terre [Richter Charles, 1958]. Nous citons aussi le cas des orages, où la variance représente l'intensité de l'orage [Killick et al., 2010].

Comme pour le cas des changements de moyenne, pour détecter le changement de variance nous pouvons utiliser la sélection de modèles, ou encore le test à hypothèses en remplaçant la moyenne  $\mu$  par la variance  $\sigma^2$  [Killick et al., 2010].

Nous pouvons aussi utiliser la technique algébrique proposée par Fliess et *al.* [Fliess et al., 2011a]. Cette technique déterministe consiste à étudier le comportement de la volatilité pour détecter les ruptures. Nous reviendrons à cette approche en temps voulu pour un autre type de ruptures.

### 1.2.1.2 Changement dans les caractéristiques spectrales du signal

#### ► Changement de fréquence

Un changement brusque de la fréquence est une autre caractérisation possible d'une rupture. Ce problème est rencontré beaucoup dans le domaine biomédical pour détecter les phénomènes transitoires à courte durée dans les signaux physiologiques afin de diagnostiquer ou prédire certaines anomalies [Nenadic, 2005], [Li et al., 1995]. Une façon de détecter ces transitions est d'utiliser la transformée de Fourier. Cette méthode est efficace dans le cas où le signal est stationnaire car dans le cas contraire elle ne donne pas une information temporelle du changement de la fréquence détectée. Pour avoir accès à cette information, nous pouvons utiliser l'une des méthodes temps fréquence. À titre d'exemple, nous pouvons citer celle de la transformée de Fourier à court terme qui réalise une représentation fréquentielle du signal via une fenêtre d'analyse glissante reposant sur la stationnarité locale du signal. Nous trouvons aussi la distribution de Weigner Ville qui fournit une décomposition temps-fréquence sans aucune restriction sur les résolutions temporelles et fréquentielles. Nous trouvons également la transformée en ondelette qui permet la représentation du signal dans le plan temps-échelle [Nenadic, 2005], [Li et al., 1995].

### 1.2.2 Discontinuité dans le signal ou dans l'une de ses fonctionnalités

Une autre manière de caractériser une rupture est de la considérer comme une discontinuité dans le signal ou dans l'une de ses fonctionnalités (la dérivée, la volatilité [Fliess et al., 2011b], etc). Pour ce type de caractérisation nous proposons un modèle continu par morceaux faisant apparaître les points de discontinuité dans le signal. Une façon d'extraire ces discontinuités est de manipuler les dérivées du signal. Diverses méthodes employant la dérivation existent dans la littérature, nous citons par exemple la méthode de Pan et Tompkins [Pan and Tompkins, 1985] qui est basée sur l'application d'un filtre dérivateur, celle aussi basée sur la sélection de modèles où nous dérivons le signal résidus pour faire apparaître les discontinuités [Laguna et al., 1989]. Nous citons également une approche récemment appliquée dans différents domaines : il s'agit de l'approche algébrique qui est une approche déterministe non asymptotique basée sur l'algèbre différentielle et le calcul opérationnel [Mikusinski and Boehme, 1987]. Les bases et le cadre mathématique de cette approche algébrique proposée par Michel Fliess et Sira-Ramírez [Fliess and Sira-Ramírez, 2003] et dont l'idée de détection de ruptures en utilisant cette dernière a été proposée pour la première fois par Mamadou Mboup [Mboup, 2007], sont exposés dans [Fliess et al., 2003], [Fliess et al., 2004], [Fliess et al., 2008], [Mboup, 2009] et [Mboup et al., 2009]. Nous détaillerons ces principes dans le chapitre 2 mais donnons en les grandes lignes :

1. Dériver le signal, modélisé par un modèle régulier par morceaux, au sens de distribution.
2. Aller vers le domaine opérationnel [Mikusinski and Boehme, 1987] afin de manipuler facilement les dérivées (une dérivation, respectivement intégration, dans le domaine temporel correspond à une multiplication, respectivement division, dans le domaine opérationnel).
3. Trouver une équation différentielle en fonction des temps d'occurrence des ruptures.

4. Extraire un système linéaire dont les solutions représentent les positions des ruptures.
5. Retourner dans le domaine temporel pour résoudre le système linéaire trouvé.

Cette méthode a été développée par plusieurs auteurs pour la détection de ruptures dans le cadre général [Mboup et al., 2007], [Mboup et al., 2008], [Fliess et al., 2010], et aussi pour des applications dans différents domaines. Citons par exemple en :

- *Biomédical* : détection des spikes de neurones [Mboup, 2012], détection des pics R du signal ECG [Debbabi and El Asmi, 2012], [Rezk et al., 2011], [Rezk et al., 2012].
- *Automatique* : diagnostic des systèmes dynamiques linéaires [Moussa Ali et al., 2011a], [Moussa Ali et al., 2011b], [Moussa Ali, 2011], estimation des paramètres des systèmes linéaires stationnaires [Tian et al., 2008], [De Miras et al., 2013].
- *Finance* : étude de la volatilité [Fliess et al., 2011a].
- *Telecommunication* : détection du spectre en radio cognitive [Ben Jemaa et al., 2012].

### 1.2.3 Événement rare

Nous associons à l'apparition d'une rupture un événement rare dont la probabilité d'occurrence est très faible, mais l'amplitude est très forte. Pour détecter cette rupture, nous utiliserons la TVE. Cette méthode utilisée pour l'étude des phénomènes inhabituels, permet de déterminer la distribution des valeurs extrêmes sans la nécessité de connaître la distribution de toutes les observations. Pour ce faire, deux approches existent. L'approche classique où l'on s'intéresse aux lois limites des extrêmes (maximums ou minimums), et l'approche Peak Over Threshold (POT) où l'on considère plusieurs statistiques d'ordre les plus grandes ou les plus petites (et pas seulement les deux valeurs extrêmes). Ces deux approches se basent essentiellement sur deux théorèmes importants : Le théorème de Fisher et Tippet [Fisher and Tippet, 1928] qui permet de caractériser la loi de distribution des extrêmes selon trois lois possibles et le théorème de Balkema, de Haan [Balkema and de Haan, 1974] et Pickands [Pickands, 1975] qui donne une estimation de la queue de distribution à partir d'un seuil assez élevé. Ce théorème nous permet d'avoir une expression analytique de la fonction de répartition des observations au-delà d'un seuil à partir desquelles nous estimerons les paramètres.

Notons que la TVE est applicable dans le cas d'une distribution de loi continue. Nous en donnerons plus de détails dans le chapitre 3.

Après avoir donné une idée générale sur les différentes caractérisation d'une rupture ainsi que sur les principes des différentes méthodes existantes dans la littérature pour la détection de ruptures suivant leur caractérisation, nous allons présenter nos contributions dans ce contexte. *Comment avons-nous envisagé la rupture ? Quelles approches avons-nous proposé pour déceler les ruptures ?* Nous répondrons à ces questions en détail dans les chapitres 2 et 3.

## 1.3 Plan de la thèse

Pour organiser ce document, nous nous sommes basés sur deux règles principales.

**Règle 1 :** caractérisation de la rupture selon deux manières différentes :

1. Dans un cadre déterministe, nous l'avons envisagé comme une discontinuité dans une fonction à variation bornée.
2. Dans un cadre stochastique nous avons associé à l'apparition de la rupture un événement rare d'amplitude relativement extrême.

Suivant ces deux caractérisations de la rupture, nous avons divisé le reste du manuscrit en deux grandes parties où chaque partie représente un chapitre traitant une caractérisation différente. Bien évidemment, chaque chapitre met en évidence la méthode associée pour la détection de la rupture qu'il traite.

**Règle 2 :** modélisation des données : ruptures et / ou bruit.

Toutes les idées développées dans cette thèse reposent sur un enchaînement de raisonnement basé sur la manière d'exploiter le plus possible d'information sur les données pour améliorer la qualité de détection de ruptures. En effet, les signaux que nous traitons dans cette thèse sont des signaux bruités. Si nous désignons par  $y(t)$  un signal bruité, avec  $t \in \mathbb{R}_+$  représentant le temps, alors ce signal peut être exprimé comme la somme de deux composantes :

$$y(t) = x(t) + \omega(t),$$

où  $x$  est une composante déterministe contenant les ruptures et  $\omega$  est une composante aléatoire décrivant le bruit. Nous présentons un schéma illustratif de cette décomposition dans la figure ci-dessous. Nous notons que le signal enregistré représenté dans cette figure n'est qu'un simple exemple. Nous ne nous limitons pas à ce type de signaux.

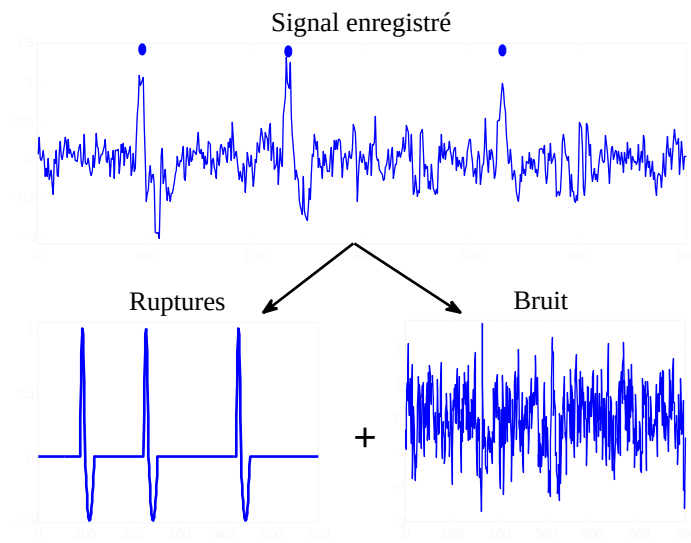


FIGURE 1.2 – Exemple de décomposition d'un signal enregistré en composante déterministe (ruptures) et composante stochastique (bruit). Les points bleus associés au signal enregistré indiquent la présence de ruptures.

La deuxième règle, suivant laquelle les deux chapitres sont à leur tour divisé en deux niveaux, repose alors sur la partie du signal que nous modélisons :  $x(t)$  et/ou  $\omega(t)$ .

Dans un premier temps, nous avons commencé par modéliser seulement les ruptures (la composante déterministe du signal :  $x(t)$ ) en leur associant un modèle déterministe, régulier, défini par morceaux, sans considérer aucun modèle du bruit perturbant ces dernières. Ensuite, pour avoir plus d'informations sur nos données, nous avons associé au bruit (composante aléatoire du signal  $\omega(t)$ ) un modèle probabiliste. À ce stade, les deux composantes sont modélisées séparément. La description de la modélisation de chaque composante constituera une sous-partie du chapitre 2.

Dans un second temps, afin d'améliorer la détection de ruptures, nous avons considéré les ruptures et le bruit : les deux composantes en même temps. Nous ne nous sommes intéressés au départ qu'à la modélisation des valeurs extrêmes du signal enregistré, en modélisant les excès au-delà d'un seuil par un modèle probabiliste. Ensuite, nous avons étendu la modélisation en considérant tout le signal enregistré, et pas seulement les extrêmes. Ces deux modélisations feront aussi l'objet de deux sous-parties du chapitre 3.

Pour résumer tout ce qui a été dit, nous représentons dans la figure 1.3 le schéma de l'organisation de la thèse, où pour chaque niveau nous nous focalisons sur la modélisation de la partie du signal représentée en rouge. L'enchaînement de raisonnement que nous avons suivi sera mieux détaillé dans ce qui suit et encore plus expliqué dans les chapitres eux mêmes.

### 1.3.1 Description du chapitre 2

Dans le chapitre 2, une rupture est vue comme une discontinuité de première espèce et d'un certain ordre dans une fonction à variation bornée en général, et dans un signal, de longueur de courbe représentative finie, faisant partie de la classe des fonctions à variation bornée, en particulier. Dans un premier niveau (voir figure 1.3, niveau 1), nous ne nous sommes intéressés qu'à la modélisation des ruptures. Nous n'avons alors considéré que la composante déterministe du signal à laquelle nous avons associé un modèle régulier par morceaux faisant apparaître les temps d'occurrence de ruptures. Pour mettre en évidence ces ruptures, nous avons adopté l'approche algébrique dont le principe est décrit et détaillé dans la première partie de ce deuxième chapitre. De cette approche algébrique, nous avons déduit différents détecteurs algébriques de ruptures. Ces détecteurs algébriques reposent sur des intégrales itérées. Ainsi, même si aucun modèle du bruit n'est explicitement incorporé dans la conception de ces détecteurs, les intégrales itérées, qui sont des filtres passe-bas, leurs confèrent une bonne robustesse au bruit.

La détection de rupture, peut néanmoins être améliorée en exploitant l'information que l'on peut avoir sur la composante aléatoire. Ainsi, dans un deuxième niveau, nous avons associé au bruit, issu du détecteur algébrique un modèle stochastique. La détermination de ce modèle est détaillée dans la troisième section de ce chapitre.

Pour mettre en pratique tout ce qui a été dit, la dernière partie du chapitre 2 a été consacrée

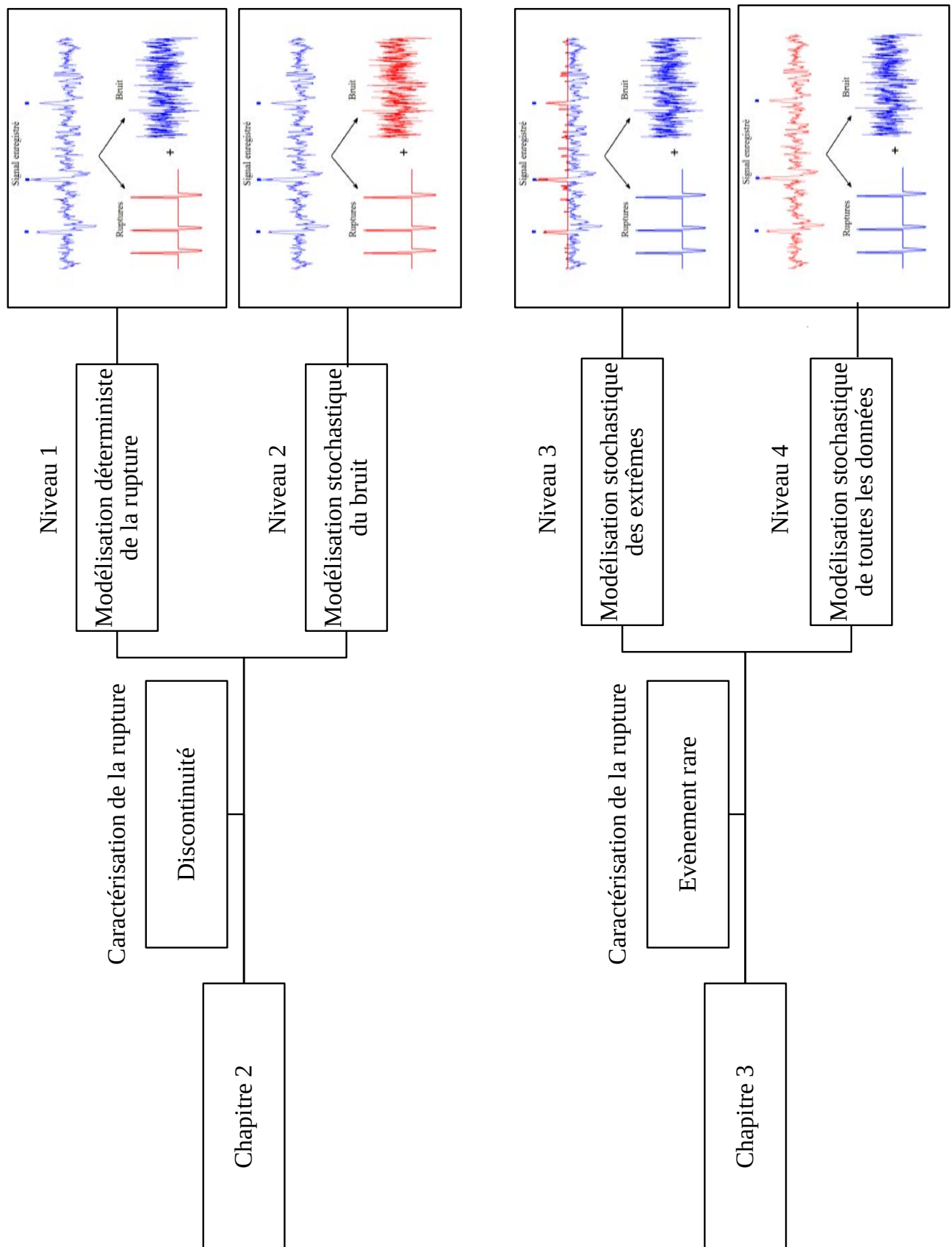


FIGURE 1.3 – Schéma d'organisation de la thèse.

au développement de deux applications. La première application concerne la détection de spikes de neurones ou encore la détection des potentiels d'action dans un signal neuronal. Pour cette application, nous avons d'abord comparé les différents détecteurs algébriques en se basant sur les courbes d'efficacité du récepteur, dites en anglais *Receiver Operating Characteristic Curves* (ROC). Ensuite, nous avons considéré celui qui donne une meilleure performance. Par la suite, nous avons appliqué ce dernier pour la détection des spikes de neurones en le comparant avec deux détecteurs existants dans la littérature. À la fin, nous avons exploité la connaissance du modèle probabiliste du bruit pour déterminer un seuil séparant les spikes des faux pics avec une probabilité de fausses alarmes fixée. Pour montrer encore les performances des détecteurs algébriques face au bruit, dans un autre contexte, nous avons appliqué l'un d'entre eux pour détecter les pics R dans un signal ElectroCardioGramme (ECG) en le comparant aussi à un autre détecteur classique. Les résultats de cette partie applicative ont été tous discutés et interprétés dans ce deuxième chapitre.

### 1.3.2 Description du chapitre 3

La rupture est vue différemment dans le troisième chapitre. Son occurrence est considérée comme un événement rare d'amplitude relativement extrême. Pour améliorer encore la détection de ruptures, nous nous sommes intéressés dans un troisième niveau à considérer toutes les données et à modéliser que les extrêmes vu l'importance de l'information qu'elles contiennent. Pour ce faire, nous nous sommes basés sur la TVE dont les principaux résultats sont introduits dans la première partie de ce chapitre. Plus particulièrement, nous nous sommes basés sur la méthode POT de la TVE qui, suivant le théorème de Balkema, de Haan et Pickands, modélise les excès au delà d'un certain seuil relativement élevé par une loi de Pareto généralisée, dite en anglais Generalized Pareto Distribution (GPD). Pour pouvoir bénéficier de ce superbe résultat, une première étape consiste à déterminer judicieusement le seuil de la POT. Tout d'abord et dans un travail préliminaire, nous avons déterminé ce seuil graphiquement en utilisant la méthode du '*mean excess plot*'. Une fois le seuil est déterminé, les excès au delà de ce dernier, contenant les ruptures, ont été modélisés par une GPD. Comme pour le niveau 2 du chapitre 2, la connaissance de la loi des extrêmes nous a permis de proposer un seuil séparant les ruptures du bruit avec une probabilité de fausses alarmes fixée, améliorant les résultats trouvés au niveau 2. Une application en neuroscience a été développée, mettant en évidence les résultats trouvés.

Dans un dernier niveau, niveau 4 de la figure 1.3, deux questions principales nous sont venues à l'esprit. La première est : *comment déterminer le seuil de la POT automatiquement sans passer par les méthodes graphiques ou semi-graphiques existantes dans la littérature ?* En fait, ces méthodes graphiques et semi-graphiques qui ont montré de bonnes performances dans le cas d'un traitement hors-ligne des données, deviennent inefficaces lorsque l'information à extraire doit apparaître en temps réel, comme par exemple pour le cas des données biomédicales. La deuxième question est : *Peut-on trouver un modèle probabiliste adéquat à nos données afin de mieux maîtriser le phénomène étudié ?* La réponse à ces deux questions a été donnée dans la dernière partie de la thèse. En effet, nous avons voulu garder l'avantage de la TVE qui nous donne la loi de probabilité des extrêmes, et étendre la modélisation à toutes les données en considérant aussi le comportement moyen de ces dernières. Ainsi, en se basant sur le Théorème



Central Limite (TCL) nous avons modélisé ce comportement moyen par une loi de probabilité gaussienne. Par conséquent, nous avons considéré une loi de probabilité multi-composantes. Cette loi de probabilité basée sur deux théorèmes limites : le TCL et le théorème de Balkema, de Haan et Pickands, se comporte comme une gaussienne en dessous d'une valeur et se comporte comme une GPD au delà de cette dernière. Cette valeur qui lie les deux composantes, représentant le point de jonction, définie en vrai le seuil indiquant la présence des extrêmes. Pour déterminer ce seuil automatiquement et répondre à la première question, nous avons proposé un algorithme itératif et non supervisé. La performance de cet algorithme a été testée sur des données simulées et évaluées en se basant sur les simulations de Monte Carlo. Une étude théorique sur l'existence de point stationnaire de l'algorithme proposé a été effectuée en s'appuyant sur l'un des théorèmes de point fixe : le théorème de Brouwer. À part la détermination judicieuse du seuil, cet algorithme offre un bon ajustement des données. Ainsi, nous répondons aux deux questions à la fois. Bien évidemment, le modèle à deux composantes : gaussienne pour le comportement moyen et GPD pour le comportement extrême, n'est pas applicable pour tout type de données. Il concerne seulement les données asymétriques à queue lourde. Pour élargir le domaine d'application des modèles hybrides, nous avons considéré, dans toute une dernière partie, différents modèles où à chaque fois le nombre de composantes, leurs natures et les coefficients associés changent. À chaque changement, nous effectuons les modifications nécessaires sur l'algorithme itératif pour ajuster le modèle considéré et déterminer le seuil ou les seuils indiquant la présence des extrêmes (maxima et/ou minima). Une application en neuroscience est donnée pour évaluer les performances de ces modèles hybrides.

## 1.4 Conclusion

Ce premier chapitre d'introduction a permis de présenter la notion de rupture avec ses différentes caractérisations dans différents contextes. Un état de l'art a été proposé, donnant une idée générale sur les principes des méthodes existantes dans la littérature traitant le problème de détection de ruptures pour chaque caractérisation. Le cadre du travail ainsi que le plan de la thèse ont été discutés à la fin de ce premier chapitre. Par la suite, nous présentons dans le deuxième chapitre la première caractérisation de la rupture considérée dans ce travail ainsi que tous les outils mathématiques utilisés pour la déceler. Aussi bien, nous détaillons dans le troisième chapitre la deuxième caractérisation de la rupture envisagée dans cette thèse tout en introduisant la TVE, le cadre mathématique permettant la détection de ce type de ruptures. Une conclusion générale des travaux effectués dans cette thèse est donnée à la fin de ce manuscrit, ouvrant des pistes de futures recherches.

## Chapitre 2

# Rupture : une discontinuité de première espèce

### 2.1 Introduction

Une rupture est caractérisée dans ce chapitre par une discontinuité de première espèce et d'un certain ordre d'une fonction à variation bornée, en général, et d'un signal en particulier. Notre objectif principal est de détecter ces discontinuités dans un environnement bruité. Pour ce faire, l'approche adoptée dans ce chapitre est une approche algébrique. Basée sur l'algèbre différentielle et le calcul opérationnel de Mikusinski, cette technique permet d'exprimer explicitement les instants d'occurrence de ruptures et d'accentuer leur présence. Cependant, pour détecter les discontinuités, l'approche algébrique repose uniquement sur une modélisation déterministe des ruptures : aucun modèle du bruit, inhérent à tout signal, n'a été considéré dans la construction des détecteurs algébriques. Afin de pouvoir exploiter l'information que l'on peut avoir sur le bruit, nous associons, dans la deuxième partie de ce chapitre, à la modélisation déterministe des ruptures une modélisation stochastique du bruit. Nous déterminons la fonction de densité de probabilité du bruit après le traitement algébrique. Cette connaissance nous a permis de déterminer un seuil permettant de mieux distinguer les pics liés aux ruptures de ceux liés au bruit avec une probabilité inférieure à une probabilité donnée de fausses alarmes. Enfin, nous consacrons une troisième partie de ce chapitre pour souligner les performances des techniques algébriques pour la détection de ruptures ainsi que l'apport de la modélisation stochastique du bruit pour l'amélioration de la détection de ces dernières. Pour ce faire, nous développons deux applications biomédicales : en cardiologie et en neuroscience.

### 2.2 Modélisation déterministe des ruptures

Nous commençons par rappeler quelques définitions et notions topologiques que nous utilisons par la suite.

**Définition 2.1 (Discontinuité de première espèce)** *On dit que  $x_0$  est un point de discontinuité de première espèce d'une fonction  $f$  si les limites  $\lim_{x \rightarrow x_0^-} f(x)$  et  $\lim_{x \rightarrow x_0^+} f(x)$  existent et sont différentes.*

**Définition 2.2 (Discontinuité de première espèce d'ordre  $n$ )** Une discontinuité de première espèce de  $f$  en un point  $x_0$  est dite d'ordre  $n \in \mathbb{N}$  si  $\exists p \in \mathbb{N}$  tel que  $\forall p < n$ ,  $\lim_{x \rightarrow x_0^-} f^{(p)}(x) = \lim_{x \rightarrow x_0^+} f^{(p)}(x)$  et  $\forall p \geq n$ ,  $\lim_{x \rightarrow x_0^-} f^{(p)}(x) \neq \lim_{x \rightarrow x_0^+} f^{(p)}(x)$ , où l'exposant  $(p)$  désigne l'ordre de dérivation de  $f$ .

**Définition 2.3 (Fonction à variation bornée)** Soit la fonction  $f$  définie sur un intervalle compact de  $\mathbb{R}$ .  $f$  est dite à variation bornée si sa variation totale sur l'intervalle  $[a, b]$

$$V_a^b(f) = \sup_{\substack{a=t_0 < t_1 < \dots < t_n=b \\ n \geq 2}} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)| \quad (2.1)$$

est finie.

**Théorème 2.4** Toute fonction réelle régulière par morceaux sur un intervalle fermé, dont la longueur de sa courbe représentative est finie, est à variation bornée sur cet intervalle.

La preuve de ce théorème est détaillée dans [Howard, 2004], page 18, où  $f$  désigne un signal.

**Proposition 2.5 ([Choquet, 1969])** Toute fonction à variation bornée est réglée : tous ses points de discontinuité sont de première espèce.

**Lemme 2.6** L'ensemble de points de discontinuité d'une fonction à variation bornée est au plus dénombrable.

Voir [Surhone et al., 2010] pour plus de détails sur le lemme 2.6.

Dans un cadre général, nous nous intéressons dans cette thèse à la détection des ruptures dans les fonctions à variation bornée, décrites dans la définition 2.3. Plus précisément, dans les signaux dont la longueur de courbe représentative est finie, faisant partie, d'après le théorème 2.4, à la classe de ces fonctions à variation bornée. Les ruptures présentes dans ces fonctions sont caractérisées dans ce chapitre comme des discontinuités de ces dernières. D'après la proposition 2.5, ces discontinuités sont de première espèce (définition 2.1). Ainsi, tout au long de ce chapitre, une rupture désignera une discontinuité de première espèce et d'un certain ordre (définition 2.2) d'une fonction à variation bornée  $f : I \rightarrow \mathbb{R}$ , où  $I$  est un compact de  $\mathbb{R}$ . Pour se mettre dans le cadre du travail, et en se basant sur le théorème 2.4, la fonction  $f$  désignera dorénavant un signal non bruité modélisant un certain phénomène que nous nommons  $x$  de variable réelle représentant le temps. Dans le cas où le signal  $x$  admet plusieurs ruptures aux points  $(t_i)_{i \geq 0}$ , nous lui associons le modèle régulier par morceaux donné par (2.2), où les  $(f_i)_{i \geq 0} : \mathbb{R} \rightarrow \mathbb{R}$  sont des fonctions régulières :

$$x(t) = \sum_{i \geq 0} f_i(t - t_i)H(t - t_i), \quad (2.2)$$

où la variable  $t \in \mathbb{R}_+$  désigne le temps et la fonction  $H(\cdot)$  représente la fonction de Heaviside définie par :

$$H(t) = \begin{cases} 0, & \text{si } t < 0; \\ 1, & \text{si } t \geq 0 \end{cases}, \quad \forall t \in \mathbb{R}.$$

Nous illustrons dans la figure 2.1 trois exemples différents d'un signal  $x$  admettant une seule rupture, où l'ordre de discontinuité de la rupture et la nature des fonctions  $\{f_0, f_1\}$  varient d'un

exemple à un autre. En effet, l'exemple *A* de cette figure, pour lequel les fonctions  $\{f_0, f_1\}$  sont des fonctions sinusoïdales, nous montre une rupture d'ordre 0 *i.e.* une discontinuité du signal  $x$  lui même. Tandis que la rupture présente dans le signal affine par morceaux de l'exemple *B* est une rupture d'ordre 1 : c'est le signal dérivé  $\dot{x}$  qui admet une rupture. Enfin, la rupture d'ordre 2 donnée par l'exemple *C*, illustrant une fonction qui change de comportement brusquement : d'un comportement constant à un comportement parabolique, représente une rupture dans le signal  $\ddot{x}$ , la dérivée seconde de  $x$ .

À travers la figure 2.1, nous pouvons constater que l'ordre de la rupture et la nature des fonctions  $(f_i)_{i \geq 0}$  de (2.2) sont fortement liés au phénomène étudié.

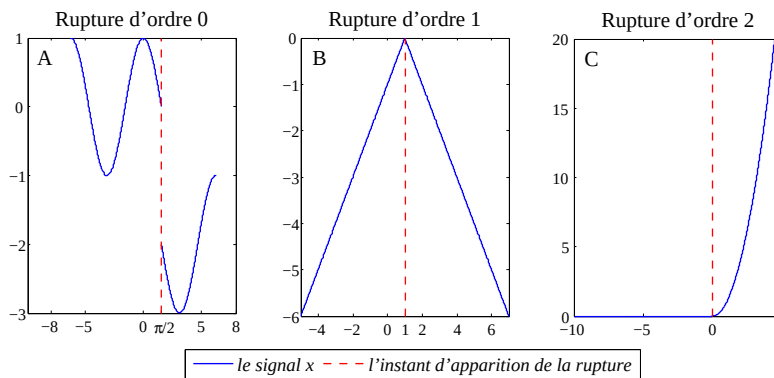


FIGURE 2.1 – Trois différents types de ruptures

Notre objectif principal est de détecter les ruptures du signal  $x$ , modélisé, dans un cadre général, par (2.2), et d'estimer leurs positions  $(t_i)_{i \geq 0}$ . Une façon de souligner la présence de telles ruptures est de dériver  $x$  au sens de distribution [Schwartz, 1998], ce qui constitue la première et la principale étape de l'approche algébrique que nous présentons ci-après.

### 2.2.1 Approche algébrique pour la détection de ruptures : principes

Dans cette section nous décrivons l'approche algébrique que nous utilisons pour détecter les ruptures présentes dans  $x$  et estimer leurs positions. Pour simplifier les choses, nous nous intéresserons plutôt à détecter la présence d'une rupture dans un intervalle glissant  $I_\tau^T$ , de taille  $T$  et d'origine  $\tau \in \mathbb{R}_+$ , où  $x$  admet au plus une rupture. La restriction de  $x$  sur cet intervalle, notée  $x_\tau$ , est exprimée comme suit :

$$x_\tau(t) = x(t + \tau)H(t), \quad \forall t \in [0, T] \quad (2.3)$$

Notons que d'après le lemme 2.6, l'ensemble des points de discontinuités de  $x$ , que nous notons  $D_{dis}(x)$ , est un ensemble au plus dénombrable appartenant à  $\mathbb{R}$ . Comme  $\mathbb{R}$  est un espace topologique séparé alors pour tout  $t_1, t_2 \in D_{dis}(x)$ , il existe un voisinage  $\mathcal{V}(t_1)$  de  $t_1$ , (resp.  $\mathcal{V}(t_2)$  de  $t_2$ ) dans  $\mathbb{R}$  tel que  $\mathcal{V}(t_1) \cap \mathcal{V}(t_2) = \emptyset$ . D'où l'existence de l'intervalle  $I_\tau^T$  contenant au plus une discontinuité dont la largeur  $T$  dépend de l'application et du phénomène étudié.

Comme  $x_\tau$  admet au plus une rupture dans  $I_\tau^T$ , il existe donc deux cas possibles :

- 1 :**  $x_\tau$  admet une et une seule rupture ou encore une discontinuité au point nommé  $t_\tau$  appartenant à  $]0, T[$ . Dans ce cas,  $x_\tau$  est régulier par morceaux. Il n'est donc pas dérivable au point  $t_\tau$ . Pour souligner la présence de cette discontinuité au point  $t_\tau$ , nous dérivons  $x_\tau$  au sens de distribution [Schwartz, 1998] dont l'expression de la  $N^{\text{ème}}$  dérivée est donnée par l'équation (2.4).

$$\frac{d^N x_\tau(t)}{dt^N} = [x_\tau^{(N)}](t) + \sum_{k=0}^{N-1} \mu_{N-k-1}^0 \delta(t)^{(k)} + \mu_{N-k-1} \delta(t - t_\tau)^{(k)}, \quad (2.4)$$

où  $\mu_{N-k-1}^0 = x_\tau^{(k)}(0)$  et  $\mu_{N-k-1} = x_\tau^{(k)}(t_\tau^+) - x_\tau^{(k)}(t_\tau^-)$ , la fonction  $\delta$  est la distribution de Dirac et  $[f]$  représente la partie régulière de  $f$ .

- 2 :**  $x_\tau$  n'a aucun point de discontinuité dans  $I_\tau^T$ . Il est donc lisse dans cet intervalle. Dans ce cas  $t_\tau = 0$ , désignant la discontinuité à l'origine induite par la présence de la fonction de Heaviside dans 2.3.

Nous mentionnons que l'équation (2.4) est l'élément clé de l'approche algébrique. À partir de cette équation, différents estimateurs algébriques peuvent être déduits pour la détection de ruptures. Comme le montre la figure 2.1, tout dépend de l'ordre de la rupture, de la nature des fonctions continues  $(f_i)_{i \geq 0}$  de (2.2), aussi bien de l'ordre de dérivation  $N$ . Par contre, la démarche adoptée par l'approche algébrique pour la détection de ruptures est quasiment la même. Elle est, en fait, basée sur trois grandes lignes : 1) exprimer l'équation (2.4) dans le domaine opérationnel [Mikusinski and Boehme, 1987] dans le but de manipuler aisément les fonctions de Dirac retardées, 2) écrire une équation différentielle faisant apparaître explicitement les temps de ruptures en la manipulant algébriquement pour annihiler les paramètres inconnus tels que les conditions initiales et les sauts du signal aux points de ruptures, 3) extraire un système linéaire dont la solution est la position de la rupture dans  $I_\tau^T$ . La résolution de ce dernier se fait par retour au domaine temporel en manipulant des convolutions de  $x_\tau$  par des filtres passe-bas. Toutes ces étapes seront détaillées dans ce qui suit sur un exemple particulier.

### 2.2.2 Approche algébrique pour la détection de ruptures : exemple

La méthode algébrique que nous présentons dans cette section est l'une des méthodes possibles que nous pouvons déduire de l'équation (2.4). Nous la développons pour la détection de ruptures d'ordre 0 dont le signal associé peut être modélisé par un modèle affine par morceaux : les fonctions  $(f_i)_{i \geq 0}$  de (2.2) sont des polynômes de degré inférieur ou égal à 1. Afin de simplifier les calculs, nous prenons  $N = 2$ . Ce choix permet d'éliminer la partie régulière de la dérivée dans l'équation (2.4). Ainsi, dans notre exemple, l'équation (2.4) devient comme suit :

$$\frac{dx_\tau^2(t)}{dt^2} = \mu_1^0 \delta(t) + \mu_1 \delta(t - t_\tau) + \mu_0^0 \dot{\delta}(t) + \mu_0 \dot{\delta}(t - t_\tau). \quad (2.5)$$

Afin de mieux manipuler les fonctions de Dirac retardées, nous exprimons (2.5) dans le domaine opérationnel [Mikusinski and Boehme, 1987]. L'équation (2.6) est ainsi obtenue.

$$s^2 \tilde{x}_\tau(s) - \mu_0^0 s - \mu_1^0 = (\mu_1 + \mu_0 s) e^{-t_\tau s}, \quad (2.6)$$

où  $\tilde{x}_\tau$  est l'analogue opérationnel de  $x_\tau$  et  $s$  désigne la variable opérationnelle. Sachant que les paramètres  $\mu_0$  et  $\mu_1$  et les conditions initiales  $\mu_0^0$  et  $\mu_1^0$ , sont des paramètres inconnus non pertinents en terme de détermination de l'instant d'occurrence de la rupture  $t_\tau$ . Il sera intéressant de les annihiler afin de simplifier le calcul. Pour ce faire, nous dérivons (2.6) deux fois. Nous obtenons alors la relation suivante :

$$\frac{d^2}{ds^2}(s^2\tilde{x}_\tau(s)) = \tilde{u}''(s), \quad (2.7)$$

où  $\tilde{u}(s) = (\mu_1 + \mu_0 s)e^{-t_\tau s}$  satisfait l'équation différentielle (2.8)

$$\tilde{u}''(s) + 2t_\tau\tilde{u}'(s) + t_\tau^2\tilde{u}(s) = 0. \quad (2.8)$$

les termes  $\tilde{u}'(s)$  et  $\tilde{u}(s)$  de (2.8) contiennent encore les paramètres inconnus et non pertinents  $\mu_0$  et  $\mu_1$ , pour les éliminer, nous dérivons cette dernière à l'ordre  $\kappa \geq 2$ . Ainsi, en se basant sur (2.7), nous obtenons l'équation différentielle suivante :

$$\frac{d^{\kappa+2}}{ds^{\kappa+2}}(s^2\tilde{x}_\tau(s)) + 2t_\tau \frac{d^{\kappa+1}}{ds^{\kappa+1}}(s^2\tilde{x}_\tau(s)) + t_\tau^2 \frac{d^\kappa}{ds^\kappa}(s^2\tilde{x}_\tau(s)) = 0, \quad \forall \kappa \geq 2. \quad (2.9)$$

Nous remarquons l'existence de multiplications de  $\tilde{x}_\tau$  par des  $s^2$  dans (2.9), correspondant à une double dérivation dans le domaine temporel ce qui n'est pas numériquement une opération robuste. On a donc intérêt à avoir que des intégrales dans le domaine temporel correspondant à des divisions par  $s$  dans le domaine opérationnel. Nous reviendrons sur ce point ultérieurement. Ainsi, nous divisons (2.9) par  $s^\nu$ ,  $\nu \geq 3$ . Ainsi,  $\forall \kappa \geq 2$ , nous obtenons l'équation du second ordre suivante :

$$\frac{1}{s^\nu} \frac{d^\kappa}{ds^\kappa}(s^2\tilde{x}_\tau(s))t_\tau^2 + \frac{1}{s^\nu} \frac{d^{i+\kappa}}{ds^{i+\kappa}}(s^2\tilde{x}_\tau(s))2t_\tau + \frac{1}{s^\nu} \frac{d^{\kappa+2}}{ds^{\kappa+2}}(s^2\tilde{x}_\tau(s)) = 0. \quad (2.10)$$

En prenant deux valeurs différentes de  $\kappa$ , par exemple  $\kappa$  et  $\kappa + 1$ ,  $\kappa \geq 2$ , nous obtenons un système linéaire de dimension 2 dont la solution est  $\Theta_\tau = [t_\tau^2, 2t_\tau]$ . À vrai dire, ce système n'est pas en fait un système d'équations linéaire naturel où les inconnus sont indépendants ou à la limite en relation linéaire. Ici, nous voyons bien que  $\Theta_\tau^1 = \frac{1}{4}\Theta_\tau^2$ , où  $\Theta_\tau^1$  et  $\Theta_\tau^2$  désignent respectivement la première et la deuxième composante du vecteur  $\Theta_\tau$ . Il s'agit bien d'une relation non linéaire rendant la détermination de la solution plus compliquée. Pour déceler ce problème et afin de figer la linéarité du système obtenu, nous ignorons dans la suite la dépendance entre  $t_\tau^2$  et  $2t_\tau$  *i.e.* entre  $\Theta_\tau^1$  et  $\Theta_\tau^2$ .

Il faut bien mentionner à ce stade que la méthode algébrique présentée dans cette section comme d'ailleurs toutes les techniques algébriques déduites de l'équation (2.4) sont bâties uniquement autour d'une modélisation déterministe de la rupture dans une situation idéaliste où le signal  $x$  est une fonction régulière par morceaux à variation bornée. Toutefois, cette situation n'est pas réaliste : les signaux traités sont en général affectés par du bruit provenant de plusieurs sources. Dans la suite, nous allons tenir compte de ce bruit en remplaçant  $x(t)$  par son homologue bruité  $y(t) = x(t) + \omega(t)$ , où  $\omega(t)$  désigne le bruit. Ainsi, en remplaçant  $x_\tau$  dans (2.10) par  $y_\tau$ , la restriction de  $y$  dans l'intervalle  $I_\tau^T$ , et en considérant cette équation pour  $\kappa$  et  $\kappa + 1$ ,  $\kappa \geq 2$ , comme décrit plus haut, nous aboutissons au système linéaire (2.11), dont la solution  $\tilde{\Theta}_\tau$

représente une estimation de  $\Theta_\tau$ .

$$\underbrace{\begin{bmatrix} \tilde{v}_\kappa^\nu(s, \tau) & \tilde{v}_{\kappa+1}^\nu(s, \tau) \\ \tilde{v}_{\kappa+1}^\nu(s, \tau) & \tilde{v}_{\kappa+2}^\nu(s, \tau) \end{bmatrix}}_{\tilde{P}_\kappa^\nu(s, \tau)} \underbrace{\begin{bmatrix} \tilde{t}_\tau^2 \\ 2\tilde{t}_\tau \end{bmatrix}}_{\tilde{\Theta}_\tau} = - \underbrace{\begin{bmatrix} \tilde{v}_{\kappa+2}^\nu(s, \tau) \\ \tilde{v}_{\kappa+3}^\nu(s, \tau) \end{bmatrix}}_{\tilde{Q}_\kappa^\nu(s, \tau)}, \quad (2.11)$$

où  $\forall \kappa \geq 2$ ,  $\tilde{v}_\kappa^\nu(s, \tau) = \frac{1}{s^\nu} \frac{d^\kappa}{ds^\kappa} (s^2 \tilde{y}_\tau(s))$ .

**Remarque 2.7** Comme il est dit plus haut, l'approche algébrique n'utilise aucun modèle du bruit dans sa construction. Elle est purement déterministe. Par contre, elle n'ignore pas la présence du bruit dans les signaux et elle en prend compte dans la conception des détecteurs algébriques. Nous pouvons voir cela au niveau de l'équation (2.9) où nous l'avons divisé par  $s^\nu$ ,  $\nu \geq 3$ , car une intégration temporelle atténue l'effet du bruit. Ce qui nous aidera à mieux détecter les ruptures.

Pour résoudre (2.11), nous exprimons les matrices  $\tilde{P}_\kappa^\nu(s, \tau)$  et  $\tilde{Q}_\kappa^\nu(s, \tau)$  dans le domaine temporel. Il suffit donc de déterminer l'expression de  $v_\kappa^\nu(t, \tau)$ , l'analogue temporel de  $\tilde{v}_\kappa^\nu(s, \tau)$ . Pour ce faire, notons par  $\tilde{f}$  (resp.  $\tilde{g}$ ) l'analogue opérationnel d'une fonction  $f$  (resp.  $g$ ) et  $n \in \mathbb{N}$ . Nous rappelons dans le tableau 2.1 les règles du calcul opérationnel utilisées pour la détermination de  $v_\kappa^\nu(t, \tau)$ .

TABLE 2.1 – Quelques règles usuelles du calcul opérationnel

| Domaine opérationnel | $\Leftrightarrow$                 |                                                                   | Domaine temporel |
|----------------------|-----------------------------------|-------------------------------------------------------------------|------------------|
| Multiplication       | $s^n \tilde{f}(s)$                | $\frac{d^n}{dt^n} f(t)$                                           | Dérivation       |
| Dérivation           | $\frac{d^n}{ds^n} \tilde{f}(s)$   | $(-1)^n t^n f(t)$                                                 | Multiplication   |
| Produit              | $\tilde{f}(s) \cdot \tilde{g}(s)$ | $(f * g)(t)$                                                      | Convolution      |
| Division             | $\frac{1}{s^n} \tilde{f}(s)$      | $\frac{1}{(n-1)!} \int_0^t (t-\lambda)^{n-1} f(\lambda) d\lambda$ | Intégration      |

Si nous avons  $\tilde{f}(s) = s^2 \tilde{y}_\tau(s)$  et  $\tilde{g}(s) = \frac{d^\kappa}{ds^\kappa}(\tilde{f}(s))$  alors d'après le tableau 2.1 des règles usuelles du calcul opérationnel  $f(t) = \ddot{y}_\tau(t)$  et  $g(t) = (-1)^\kappa t^\kappa f(t) = (-1)^\kappa t^\kappa \ddot{y}_\tau(t)$ . Si maintenant  $\tilde{h}(s) = \frac{1}{s^\nu} \tilde{g}(s)$ , son analogue opérationnel  $h(t)$  est exprimé comme suit :

$$h(t) = \frac{1}{(\nu-1)!} \int_0^t (t-\lambda)^{\nu-1} g(\lambda) d\lambda = \frac{(-1)^\kappa}{(\nu-1)!} \int_0^t (t-\lambda)^{\nu-1} \lambda^\kappa \ddot{y}_\tau(\lambda) d\lambda. \text{ Comme } \tilde{h}(s) \text{ n'est}$$

autre que  $\tilde{v}_\kappa^\nu(s, \tau)$ , par conséquent,  $v_\kappa^\nu(t, \tau) = h(t)$  c'est à dire :

$$v_\kappa^\nu(t, \tau) = \frac{(-1)^\kappa}{(\nu - 1)!} \int_0^t (t - \lambda)^{\nu-1} \lambda^\kappa \ddot{y}_\tau(\lambda) d\lambda, \forall \kappa \geq 2. \quad (2.12)$$

Une fois les valeurs de  $v_{\kappa+i}^\nu(t, \tau)$ ,  $i \in \{0, \dots, 3\}$ , sont déterminées, l'analogie temporel du système 2.11 sera le suivant :

$$\underbrace{\begin{bmatrix} v_\kappa^\nu(t, \tau) & v_{\kappa+1}^\nu(t, \tau) \\ v_{\kappa+1}^\nu(t, \tau) & v_{\kappa+2}^\nu(t, \tau) \end{bmatrix}}_{P_\kappa^\nu(t, \tau)} \begin{bmatrix} \tilde{t}_\tau^2 \\ 2\tilde{t}_\tau \end{bmatrix} = - \underbrace{\begin{bmatrix} v_{\kappa+2}^\nu(t, \tau) \\ v_{\kappa+3}^\nu(t, \tau) \end{bmatrix}}_{Q_\kappa^\nu(t, \tau)}. \quad (2.13)$$

Notons que le système (2.13) est défini en fonction de  $t$  : pour tout  $t \in [0, T]$ , nous obtenons une nouvelle estimation  $\tilde{\Theta}_\tau$  de  $\Theta_\tau$ . Comme les composantes de la matrice  $P_\kappa^\nu(t, \tau)$  sont exprimées sous forme d'intégrales sur  $[0, t]$ , ainsi, afin d'avoir plus d'information nous permettant d'obtenir une meilleure estimation de  $\Theta_\tau$ , nous allons travailler sur tout l'intervalle  $[0, T]$  en fixant par la suite  $t = T$ , la taille de l'intervalle  $I_\tau^T$ . Pour simplifier les notations,  $v_\kappa^\nu(\tau)$ ,  $P_\kappa^\nu(\tau)$  et  $Q_\kappa^\nu(\tau)$  désigneront dorénavant respectivement  $v_\kappa^\nu(T, \tau)$ ,  $P_\kappa^\nu(T, \tau)$  et  $Q_\kappa^\nu(T, \tau)$ . Nous allons voir dans la section suivante, qu'en se basant sur le système (2.13), il n'existe pas qu'une seule façon de détecter les ruptures. Différentes fonctions de décisions sont proposées dans la section 2.2.3.

### 2.2.3 Fonctions de décision

Le système (2.13) n'a pas toujours un sens. Il n'est pas défini si l'intervalle considéré ne contient pas une rupture. On ne peut donc pas parler de sa résolution sans s'assurer qu'il est bien défini. Nous verrons comment l'inspection des quantités  $v_\kappa^\nu(\tau)$  pourra permettre de détecter la présence ou non de ruptures et de les localiser. On obtient ainsi un moyen pour détecter les ruptures sans passer par la résolution directe du système (2.13). Pour ce faire, nous allons nous baser sur les informations données par les deux hypothèses ci-dessous, présentées dans le cadre idéaliste où le signal  $x$  n'est pas affecté par le bruit ou encore pour  $y = x$ , pour développer des fonctions de décisions indiquant la présence ou l'absence d'une rupture.

**Hypothèse nulle  $\mathcal{H}_0$**  : Absence de rupture dans  $I_\tau^T$ .

**Hypothèse alternatives  $\mathcal{H}_1$**  : Présence de rupture dans  $I_\tau^T$ .

Différentes fonctions de décision peuvent être extraites du système (2.13). Ces fonctions sont divisées en deux catégories : des fonctions de décisions basées sur l'hypothèse  $\mathcal{H}_0$  et d'autres basées sur l'hypothèse  $\mathcal{H}_1$ . Dans ce qui suit, nous présentons quelques unes.

#### 2.2.3.1 L'hypothèse nulle $\mathcal{H}_0$

Nous commençons par nous placer sous l'hypothèse nulle  $\mathcal{H}_0$  indiquant l'absence de rupture dans l'intervalle  $I_\tau^T$ . Pour arriver à construire des détecteurs algébriques sous  $\mathcal{H}_0$ , nous allons suivre l'enchaînement d'arguments suivant :



sous  $\mathcal{H}_0$ ,  $I_\tau^T$  ne contient pas de rupture  $\Rightarrow t_\tau = 0 \Rightarrow$  le signal  $x_\tau$  et sa dérivée sont lisses et ne font pas de sauts sur  $I_\tau^T \Rightarrow \mu_0 = \mu_1 = 0 \Rightarrow$  d'après (2.8),  $\forall \kappa \geq 2$ ,  $v_\kappa^\nu(\tau) = 0 \Rightarrow$  le système (2.13) est dégénéré :  $P_\kappa^\nu(\tau) = 0_{M_n(\mathbb{R})}$  et  $Q_\kappa^\nu = 0_{\mathbb{R}_2}$ .

De ces arguments il vient que si  $I_\tau^T$  ne contient pas de rupture alors  $\forall \kappa \geq 2$ ,  $v_\kappa^\nu(\tau) = 0$ ,  $\tau \in \mathbb{R}_+$ . Par conséquent, une rupture peut être détectée si l'un des  $v_\kappa^\nu(\tau)$ ,  $\forall \tau \in \mathbb{R}_+$ , est non nul. Toute une gamme de fonctions de décisions se base sur ce principe. Nous présentons ci-dessous quelques unes.

$$\mathcal{D}_1(\tau) = v_\kappa^\nu(\tau) \quad (2.14)$$

$$\mathcal{D}_2(\tau) = v_\kappa^\nu(\tau)^2 v_{\kappa+1}^\nu(\tau) v_{\kappa+2}^\nu(\tau) \quad (2.15)$$

$$\mathcal{D}_3(\tau) = v_\kappa^\nu(\tau) v_{\kappa+1}^\nu(\tau) \quad (2.16)$$

La fonction  $\mathcal{D}_2(\tau)$  désigne le produit des composantes de la matrice  $P_\kappa^\nu(\tau)$ . Si elle est nulle, elle indique, comme  $\mathcal{D}_3(\tau)$  d'ailleurs, que (2.13) est dégénéré et donc le système 2.13 n'est pas défini. Toutes ces fonctions de décision sont ainsi comparées à 0, et la décision suivante est prise.  $\forall i \in \{1, \dots, 3\}$ , si

- $\mathcal{D}_i(\tau) \neq 0$ , alors  $I_\tau^T$  contient une rupture,
- $\mathcal{D}_i(\tau) = 0$ , aucune rupture n'apparaît dans  $I_\tau^T$ .

Nous mentionnons que ce critère de décision indique juste la présence ou l'absence d'une rupture dans l'intervalle  $I_\tau^T$ . Il n'indique pas si une rupture existe alors elle est unique.

Nous rappelons que  $\mathcal{D}_1(\tau)$ ,  $\mathcal{D}_2(\tau)$  et  $\mathcal{D}_3(\tau)$  ne sont que trois fonctions de décisions possibles basées sur l'hypothèse  $\mathcal{H}_0$ . En raisonnant autrement, nous passons maintenant à voir quel raisonnement nous allons suivre sous l'hypothèse alternative  $\mathcal{H}_1$ .

### 2.2.3.2 L'hypothèse alternative $\mathcal{H}_1$

Comme pour l'hypothèse nulle, se placer sous l'hypothèse alternative  $\mathcal{H}_1$  implique certaines affirmations liées entre elles et qui nous mènent à la fin à déterminer une fonction de décision indiquant la présence ou l'absence d'une rupture. En effet, sous  $\mathcal{H}_1$ , on a :

1.  $I_\tau^T$  contient une rupture  $\Leftrightarrow$  le système 2.13 est non dégénéré, donc il est bien défini.
2. la rupture est unique dans  $I_\tau^T \Leftrightarrow$  le système 2.13 admet une solution unique  $\Leftrightarrow \det(P_\kappa^\nu(\tau)) \neq 0$ .
3. la solution est réelle (le temps)  $\Leftrightarrow$  la première équation de 2.13 qui est une équation du second degré admet des racines réelles  $\Leftrightarrow$  le discriminant réduit  $\Delta$  de cette équation est positif ou nul.

De ces trois points nous pouvons conclure que sous  $\mathcal{H}_1$ ,  $I_\tau^T$  contient une rupture  $\Leftrightarrow \Delta \geq 0$ . Comme  $\Delta = -\det(P_\kappa^\nu(\tau))$  et que  $\det(P_\kappa^\nu(\tau)) \neq 0$ , il s'ensuit que sous  $\mathcal{H}_1$ ,  $I_\tau^T$  contient une rupture  $\Leftrightarrow \Delta > 0$ . Par conséquent, nous allons considérer  $\Delta$  comme fonction de décision que nous notons  $\mathcal{D}_4$  et dont l'expression est donnée par :

$$\mathcal{D}_4(\tau) = v_{\kappa+1}^\nu(\tau)^2 - v_\kappa^\nu(\tau)v_{\kappa+2}^\nu(\tau), \quad \forall \tau \in \mathbb{R}. \quad (2.17)$$

À cette fonction de décision, nous associons la décision suivante :

- $\mathcal{D}_4(\tau) > 0$ , si  $I_\tau^T$  contient une rupture,
- $\mathcal{D}_4(\tau) \leq 0$ , si aucune rupture n'apparaît dans  $I_\tau^T$ .

Ainsi, une rupture est détectée si et seulement si  $\mathcal{D}_4(\tau) > 0$ .

Nous rappelons que les fonctions de décisions décrites dans cette section représentent quelques exemples de fonctions de décisions que nous pouvons avoir. L'objectif était de montrer que nous n'avons pas qu'une seule et aussi de montrer la logique et le raisonnement derrière chaque  $\mathcal{D}_i$ ,  $i \in \{1, \dots, 4\}$  qui diffère d'une fonction à une autre et d'une application à une autre. Nous rappelons aussi que toutes ces fonctions ont été développées dans un cadre idéaliste où le signal n'est pas perturbé par le bruit. Maintenant si nous déplaçons dans un environnement plus réaliste en considérant la présence du bruit. En appliquant les fonctions de décisions  $\mathcal{D}_i$ ,  $i \in \{1, \dots, 4\}$  nous allons avoir des pics indiquant la présence des ruptures mais aussi nous allons remarquer la présence de faux pics liés au bruit. Ainsi, pour prendre les décisions,  $\mathcal{D}_i$ ,  $i \in \{1, \dots, 4\}$  ne seront plus comparées à 0 mais à des seuils séparant les faux pics liés au bruit, que nous appelons les fausses alarmes, des pics indiquant la présence des ruptures. Nous discuterons la détermination de ces seuils pour  $\mathcal{D}_4$  dans la section 2.3.2. Maintenant, nous passons à décrire l'implémentation de  $\mathcal{D}_i$ ,  $i \in \{1, \dots, 4\}$  dans un environnement bruité, qui aura lieu dans le domaine temporel discret.

## 2.2.4 Implémentation des différentes fonctions de décision

Nous commençons cette section par rappeler que toutes les fonctions de décisions  $\mathcal{D}_i$ , pour  $i \in \{1, \dots, 4\}$  sont basées sur le calcul de  $v_\kappa^\nu(\tau)$ , donnée par (2.12) pour  $t = T$ , pour des différentes valeurs de  $\kappa$ ,  $\kappa \geq 2$ . La fonction  $v_\kappa^\nu(\cdot)$  est exprimée dans le domaine temporel continu. Pour l'implémenter, un passage au domaine temporel discret est ainsi nécessaire. Pour ce faire, nous effectuons d'abord les deux opérations ci-dessous sur  $v_\kappa^\nu(\cdot)$  afin de faciliter sa discrétisation.

1. Comme les seules données que nous avons sont celles du signal  $y$  et non pas de sa dérivée seconde, nous exprimons  $v_\kappa^\nu(\cdot)$  en fonction de  $y_\tau$  seulement en effectuant une double intégration par parties à (2.12). Comme résultat, nous obtenons l'équation suivante :

$$v_\kappa^\nu(\tau) = \frac{(-1)^\kappa}{(\nu-1)!} \int_0^T \frac{d^2}{d\lambda^2} \{(T-\lambda)^{\nu-1} \lambda^\kappa\} y_\tau(\lambda) d\lambda. \quad (2.18)$$

2. Nous redéfinissons (2.18) sur  $\mathbb{R}$  comme suit :

$$v_\kappa^\nu(\tau) = - \int_{\mathbb{R}} h_\kappa^\nu(T-\lambda) y_\tau(\lambda) d\lambda, \quad (2.19)$$

où

$$h_{\kappa}^{\nu}(\lambda) = \begin{cases} \frac{(-1)^{\kappa+1}}{(\nu-1)!} \frac{d^2}{d\lambda^2} \{\lambda^{\nu-1}(T-\lambda)^{\kappa}\}, & \text{si } \lambda \in [0, T]; \\ 0, & \text{sinon.} \end{cases}$$

En appliquant le changement de variables suivant à (2.19) :  $\lambda = T - \lambda$  et pour  $\tau' = T + \tau$ ,  $v_{\kappa}^{\nu}(\tau)$  est réécrite sous forme de produit de convolution de  $y$  avec  $h_{\kappa}^{\nu}$  comme suit :

$$v_{\kappa}^{\nu}(\tau) = \int_{\mathbb{R}} h_{\kappa}^{\nu}(\lambda) y(\tau' - \lambda) d\lambda. \quad (2.20)$$

Par conséquent, à travers (2.20), nous pouvons facilement constater que  $v_{\kappa}^{\nu}(\tau)$  n'est autre que la sortie d'un filtre linéaire de réponse impulsionnelle  $h_{\kappa}^{\nu}$ . Pour l'implémenter, nous nous déplaçons dans le domaine temporel discret et nous utilisons l'une des méthodes d'intégration numérique. En effet, considérons un échantillonnage uniforme du signal  $y$  avec une période d'échantillonnage  $T_s$ . Ainsi, nous obtenons les observations  $y_m = y(t_m)$ ,  $\forall m \geq 0$  et  $t_m = mT_s$ , désignant les temps d'observations. Si  $T$  représente  $M$  échantillons *i.e.*  $T = MT_s$  alors en utilisant la méthode des trapèzes d'intégration numérique, une approximation de  $v_{\kappa}^{\nu}(\tau)$  que nous noterons  $v_{\kappa,n}^{\nu}$  sera donnée par :

$$v_{\kappa,n}^{\nu} = \sum_{m=0}^M g_{\kappa,m}^{\nu} y_{n+M-m}, \forall \tau = nT_s \text{ and } \tau' = (n+M)T_s, n \in \mathbb{N}^* \quad (2.21)$$

où  $\forall 0 \leq m \leq M$ ,  $g_{\kappa,m}^{\nu} = W_m h_{\kappa}^{\nu}(t_m)$ . Avec  $W_m$  désigne les poids de la méthode des trapèzes :  $W_0 = W_M = 0.5$  et  $\forall 1 \leq m \leq M-1$ ,  $W_m = 1$ .

Ainsi,  $\forall \kappa \geq 2$ ,  $v_{\kappa,n}^{\nu}$  est implémentée comme la sortie d'un filtre linéaire à réponse impulsionnelle finie donnée par la séquence  $\{g_{\kappa,m}^{\nu}\}_{m=0}^M$ .

Par conséquent, comme les  $\{\mathcal{D}_i(\tau)\}_{i \in \{2, \dots, 4\}}$ ,  $\tau \geq 0$ , s'expriment en combinaison non-linéaire de  $v_{\kappa}^{\nu}(\tau)$  alors les  $\{\mathcal{D}_{i,n}\}_{i \in \{2, \dots, 4\}}$ , leurs analogues respectives dans le domaine temporel discret, sont des combinaisons non-linéaires de filtres linéaires.

Comme il est dit plus haut, les techniques algébriques de détection de ruptures, présentées jusqu'à ce niveau, sont basées uniquement sur la modélisation déterministe des ruptures. La connaissance à priori sur le bruit, inhérent aux signaux, a été totalement négligée. Dans le but d'améliorer la détection de ruptures, et de pouvoir distinguer entre les fausses alarmes et les vrais pics signalant la présence des ruptures, nous exploitons dans la section 2.3 l'information que l'on peut avoir sur le bruit.

## 2.3 Modélisation stochastique du bruit

Dans cette section, nous associons à la modélisation déterministe des ruptures une modélisation probabiliste du bruit représentant un phénomène aléatoire. Le bruit que nous modélisons est le bruit transformé après l'application des techniques algébriques. Nous nous intéressons dans ce qui suit à modéliser le bruit issu de la fonction de décision  $\mathcal{D}_4$ , où le bruit original discrétisé  $(\omega_n)_{n \in \mathbb{N}^*}$  est considéré gaussien de moyenne nulle et de variance notée  $\sigma_B^2$ . Le choix de la fonction de décision  $\mathcal{D}_4$  sera discuté dans la section 2.4, par contre le choix du

bruit gaussien est argumenté par le nombre énorme de phénomènes réels où le TCL [Feller, 1966] s'applique sur les données à traiter, modélisant le bruit sous-jacent par un processus gaussien [Mboup, 2008], [Digham et al., 2007], et [Ben Jemaa et al., 2012].

### 2.3.1 Loi de probabilité du bruit

Pour déterminer la loi de probabilité du bruit issu de la fonction de décision  $\mathcal{D}_4$ , nous appliquons  $\mathcal{D}_4$  sur le bruit seulement en considérant  $y(t) = \omega(t)$ . Pour ce faire, nous écrivons  $\mathcal{D}_{4,n}$  qui représente la sortie d'un filtre de Volterra [Mboup, 2008] admettant le bruit original discrétisé  $(\omega_n)_{n \in \mathbb{N}}$  comme entrée. Nous l'appelons  $\mathcal{D}_{4,n}^\omega$  relativement à  $\omega$ . Pour tout  $n \in \mathbb{N}^*$ ,  $\mathcal{D}_{4,n}^\omega$  est exprimé comme suit :

$$\begin{aligned} \mathcal{D}_{4,n}^\omega &= [B_n^t G_{\kappa+1}]^2 - [B_n^t G_\kappa][G_{\kappa+2} B_n] \\ &= B_n^t \{G_{\kappa+1} G_{\kappa+1}^t - \frac{1}{2} G_\kappa G_{\kappa+2}^t + G_{\kappa+2} G_\kappa^t\} B_n \\ &= B_n^t \mathcal{G} B_n \end{aligned} \quad (2.22)$$

où

$$\mathcal{G} = G_{\kappa+1} G_{\kappa+1}^t - \frac{1}{2} G_\kappa G_{\kappa+2}^t + G_{\kappa+2} G_\kappa^t, \quad (2.23)$$

$$B_n = \begin{bmatrix} \omega_n \\ \vdots \\ \omega_{n-M} \end{bmatrix} \text{ et } G_\kappa = \begin{bmatrix} g_{\kappa,0}^\nu \\ \vdots \\ g_{\kappa,n-M} \end{bmatrix}.$$

Déterminer la loi de probabilité de  $\mathcal{D}_{4,n}^\omega$  directement à partir de l'équation (2.22) semble être compliqué. Il faudra donc trouver une expression plus simple de (2.22) permettant la déduction de cette dernière. Pour ce faire, nous allons exprimer le bruit  $(B_n)_{n \in \mathbb{N}^*}$  dans la base formée de vecteurs propres de la matrice  $\mathcal{G}$  permettant d'exprimer  $\mathcal{D}_{4,n}^\omega$  comme une combinaison linéaire d'ordre 3 de loi Khi-deux de degré de liberté valant 1, notée  $\mathcal{X}_2(1)$ . En effet, nous commençons d'abord par présenter une propriété très importante de la matrice  $\mathcal{G}$  donnée par le lemme 2.8.

**Lemme 2.8** *Le spectre de  $\mathcal{G}$  est composé de deux valeurs propres positives  $\lambda_1^+ \geq \lambda_2^+ > 0$ , une valeur propre négative  $\lambda^-$  et  $M - 2$  valeurs propres nulles.*

Pour démontrer le lemme 2.8, nous avons besoin d'un deuxième lemme annoncé ci-dessous.

**Lemme 2.9** *Le spectre de la matrice  $\mathcal{A}$  donnée par :*

$$\mathcal{A} = \begin{bmatrix} \|G_{\kappa+1}\|^2 & G_{\kappa+1}^t G_\kappa & G_{\kappa+1}^t G_{\kappa+2} \\ -\frac{1}{2} G_{\kappa+2}^t G_{\kappa+1} & -\frac{1}{2} G_{\kappa+2}^t G_\kappa & -\frac{1}{2} \|G_{\kappa+2}\|^2 \\ -\frac{1}{2} G_\kappa^t G_{\kappa+2} & -\frac{1}{2} \|G_\kappa\|^2 & -\frac{1}{2} G_\kappa^t G_{\kappa+2} \end{bmatrix}$$

coïncide avec les trois valeurs propres non nulles de la matrice  $\mathcal{G}$  donnée par (2.23).

**Preuve 2.10 (Preuve du lemme 2.9)** *Soit  $\lambda$  une valeur propre non nulle de  $\mathcal{G}$  et  $U$  le vecteur propre qui lui est associé. Comme  $U$  appartient à l'image de  $\mathcal{G}$ , qui est  $\text{vect}\{G_{\kappa+1}, G_\kappa, G_{\kappa+2}\}$ ,*

alors il existe un unique vecteur  $V = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \in \mathcal{M}_{(3,1)}(\mathbb{R})$  tel que  $U = c_1 G_{\kappa+1} + c_2 G_{\kappa} + c_3 G_{\kappa+2} =$

$\mathcal{H}V$ , où  $\mathcal{H} = \begin{bmatrix} G_{\kappa+1} & G_{\kappa} & G_{\kappa+2} \end{bmatrix} \in \mathcal{M}_{(M+1,3)}(\mathbb{R})$ , et  $\mathcal{M}_{(l,c)}(\mathbb{R})$  désigne l'ensemble des matrices de  $l$  lignes et de  $c$  colonnes à coefficients dans  $\mathbb{R}$ . Ainsi, nous obtenons :

$$\mathcal{G}U = \mathcal{G}(c_1 G_{\kappa+1} + c_2 G_{\kappa} + c_3 G_{\kappa+2}) = \mathcal{G}\mathcal{H}V. \quad (2.24)$$

En remplaçant  $\mathcal{G}$  dans (2.24) par son expression donnée par (2.23) nous aurons :

$$\begin{aligned} \mathcal{G}U &= (G_{\kappa+1}G_{\kappa+1}^t - \frac{1}{2}G_{\kappa}G_{\kappa+2}^t + G_{\kappa+2}G_{\kappa}^t)(c_1 G_{\kappa+1} + c_2 G_{\kappa} + c_3 G_{\kappa+2}) \\ &= (c_1 \|G_{\kappa+1}\|^2 + c_2 G_{\kappa+1}^t G_{\kappa} + c_3 G_{\kappa+1}^t G_{\kappa+2})G_{\kappa+1} \\ &\quad - \frac{1}{2}(c_1 G_{\kappa+2}^t G_{\kappa+1} + c_2 G_{\kappa+2}^t G_{\kappa} + c_3 \|G_{\kappa+2}\|^2)G_{\kappa} \\ &\quad - \frac{1}{2}(c_1 G_{\kappa}^t G_{\kappa+2} + c_2 \|G_{\kappa}\|^2 + c_3 G_{\kappa}^t G_{\kappa+2})G_{\kappa+2} \end{aligned} \quad (2.25)$$

Nous remarquons que l'équation (2.25) peut s'écrire de la manière suivante :

$$\mathcal{G}U = \mathcal{H}AV. \quad (2.26)$$

D'autre part, nous savons déjà que

$$\mathcal{G}U = \lambda U = \lambda \mathcal{H}V = \mathcal{H}\lambda V. \quad (2.27)$$

Ainsi, de (2.26) et (2.27) nous tirons que  $\mathcal{H}AV = \mathcal{H}\lambda V$ . Les colonnes de la matrice  $\mathcal{H}$  étant linéairement indépendants par construction, il s'ensuit que  $AV = \lambda V$ . Par conséquent,  $\lambda$  est aussi une valeur propre de  $\mathcal{A}$ , qui est associée au vecteur propre  $V$ .

Nous rajoutons que l'unicité du vecteur  $V$  assure que pour chaque couple  $(U, \lambda)$  de vecteur et valeur propre de  $\mathcal{G}$ , où  $\lambda$  est non nulle, est associé un unique couple  $(V, \lambda)$  de vecteur et valeur propre de  $\mathcal{A}$ . Comme  $\mathcal{A}$  est de rang 3 alors le nombre de valeurs propres non nulles de  $\mathcal{G}$  est égal à 3 aussi. Autrement dit, les valeurs propres  $\lambda_1^+$ ,  $\lambda_2^+$  et  $\lambda^-$  de  $\mathcal{A}$  sont de multiplicité égale à 1.

Retournons maintenant à la preuve du lemme 2.8.

**Preuve 2.11 (Preuve du lemme 2.8)** D'après le lemme 2.9,  $\mathcal{G}$  admet trois valeurs propres non nulles. Montrons maintenant que ces dernières sont composées de deux valeurs propres positives et une négative. Pour ce faire, nous factorisons la matrice  $\mathcal{A}$  comme suit :

$$\mathcal{A} = \mathcal{D}\mathcal{H}^t\mathcal{H}$$

$$\text{où } \mathcal{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & 0 \end{bmatrix}.$$

Le déterminant de  $\mathcal{A}$  est alors :

$$\begin{aligned} \det(\mathcal{A}) &= \det(\mathcal{D}) \det(\mathcal{H}^t\mathcal{H}) \\ &= -\frac{1}{4} \det(\mathcal{H}^t\mathcal{H}) < 0 \end{aligned}$$

Comme le déterminant de  $\mathcal{A}$  est négatif, deux cas de figure sont possibles :  $\mathcal{A}$  admet une seule valeur propre négative ou toutes les valeurs propres de  $\mathcal{A}$  sont négatives. Nous montrons par l'absurde que le deuxième cas ne peut pas avoir lieu. En effet, si  $\mathcal{A}$  admet trois valeurs propres négatives, alors  $\forall v \in \mathbb{R}^{M+1}$ ,  $v^t \mathcal{G} v \leq 0$ , en particulier pour  $v \in \text{vect}\{G_{\kappa+1}, G_{\kappa}, G_{\kappa+2}\} \cap \text{vect}\{G_{\kappa}, G_{\kappa+2}\}^\perp$ , où  $E^\perp$  désigne l'ensemble orthogonal à l'ensemble  $E$ . Dans ce cas, on obtient :

$$\begin{aligned} v^t \mathcal{G} v &= v^t G_{\kappa+1} G_{\kappa+1}^t v \\ &= (G_{\kappa+1}^t v)^t G_{\kappa+1}^t v \\ &= \|G_{\kappa+1}^t v\|^2 \leq 0 \text{ absurde.} \end{aligned}$$

Nous pouvons ainsi conclure que  $\mathcal{G}$  admet deux valeurs propres positives  $\lambda_1^+$  et  $\lambda_2^+$ , une seule valeur propre négative  $\lambda^-$ , et les  $M - 2$  valeurs propres restantes sont nulles.

Le résultat obtenu est très important. Il nous facilitera les calculs : nous passons d'une dimension de  $M + 1$  à une dimension de 3. En effet, soient  $q_1^+$ ,  $q_2^+$  et  $q^-$  les vecteurs propres de  $\mathcal{G}$  associés respectivement aux valeurs propres  $\lambda_1^+$ ,  $\lambda_2^+$ , et  $\lambda^-$ . Comme  $\mathcal{G}$  est symétrique (il est facile à démontrer que  $\mathcal{G} = \mathcal{G}^t$ ) alors l'ensemble de ses vecteurs propres constitue une base orthonormée de  $\mathbb{R}^{M+1}$ . Ainsi, le bruit  $B_n$  peut être décomposé dans cette base comme suit :

$$B_n = a_n q_1^+ + b_n q_2^+ + c_n q^- + q^\perp(n), \quad (2.28)$$

où  $q^\perp(n)$  représente la somme pondérée de vecteurs propres associés aux  $M - 2$  valeurs propres nulles de  $\mathcal{G}$ . Les coefficients  $a_n$ ,  $b_n$ , et  $c_n \in \mathbb{R}$  sont obtenus en multipliant à chaque fois  $B_n^t$  par l'un des vecteurs propres  $q_1^+$ ,  $q_2^+$ , et  $q^-$ . Comme les vecteurs propres sont orthonormés, il s'ensuit que  $a_n = B_n^t q_1^+$ ,  $b_n = B_n^t q_2^+$ , et  $c_n = B_n^t q^-$ . En remplaçant  $B_n$  de l'équation (2.22) par son expression de (2.28),  $\mathcal{D}_{4,n}^\omega$  s'exprime ainsi comme suit :

$$\begin{aligned} \mathcal{D}_{4,n}^\omega &= \{a_n q_1^{+t} + b_n q_2^{+t} + c_n q^{-t} + q^{\perp(n)t}\} \mathcal{G} \{a_n q_1^+ + b_n q_2^+ + c_n q^- + q^\perp(n)\} \\ &= \{a_n q_1^{+t} + b_n q_2^{+t} + c_n q^{-t} + q^{\perp(n)t}\} \{a_n \lambda_1^+ q_1^+ + b_n \lambda_2^+ q_2^+ - c_n \eta^2 q^- + 0\} \end{aligned} \quad (2.29)$$

$$= \lambda_1^+ a_n^2 + \lambda_2^+ b_n^2 + \lambda^- c_n^2. \quad (2.30)$$

Notons que l'équation (2.29) est obtenue en remplaçant  $\mathcal{G} q_1^+$  par  $\lambda_1^+ q_1^+$  et de même pour les autres valeurs propres. L'équation (2.30) est obtenue en utilisant l'orthonormalité des vecteurs propres.

Avant de procéder à la détermination de la loi de probabilité de  $\mathcal{D}_{4,n}^\omega$ , nous rappelons d'abord quelques points en probabilité. Soient  $X$  et  $Y$  deux variable aléatoire (v.a.) indépendantes de fonctions de densité de probabilité dite en anglais 'probability density function' (pdf) respectives  $f_X$  et  $f_Y$ . On a alors :

1. Si la v.a.  $X$  suit la loi normale centrée réduite (de moyenne nulle et de variance unité) alors la v.a.  $X^2$  suit la loi  $\mathcal{X}_2(1)$ . Sa pdf  $f_{\mathcal{X}_2(1)}$  est donnée par :

$$f_{\mathcal{X}_2(1)}(x) = \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})} \frac{e^{-\frac{x}{2}}}{\sqrt{2}}, \quad \forall x \in \mathbb{R}_+, \quad (2.31)$$

où  $\Gamma(\cdot)$  désigne la fonction Gamma.

2. La pdf  $f_{cX}$  de la v.a.  $cX$ ,  $c \in \mathbb{R}$  est donnée par :

$$f_{cX}(x) = \frac{1}{|c|} f_X\left(\frac{x}{c}\right), \quad (2.32)$$

3. La pdf  $f_{X+Y}$  de la v.a.  $X + Y$ , est donnée par :

$$f_{X+Y}(x) = (f_X \star f_Y)(x). \quad (2.33)$$

Le symbole  $\star$  représente le produit de convolution.

Nous rappelons aussi que  $(a_n = B_n^t q_1^+)_{n \in \mathbb{N}}$ ,  $(b_n = B_n^t q_2^+)_{n \in \mathbb{N}}$  et  $(c_n = B_n^t q^-)_{n \in \mathbb{N}}$ . Comme  $(B_n)_{n \in \mathbb{N}}$  est un bruit gaussien de moyenne nulle et de variance  $\sigma_B^2$ , et  $q_1^+$ ,  $q_2^+$  et  $q^-$  sont trois vecteurs orthonormaux, donc de norme unité, alors  $(a_n)_{n \in \mathbb{N}}$ ,  $(b_n)_{n \in \mathbb{N}}$  et  $(c_n)_{n \in \mathbb{N}}$  sont aussi de moyenne nulle et de même variance  $\sigma_B^2$  que  $(B_n)_{n \in \mathbb{N}}$ . Il s'agit d'une combinaison linéaire de v.a. gaussiennes dont les coefficients sont de normes unités. Ainsi, en réécrivant l'expression de  $\mathcal{D}_{4,n}^\omega$ , donnée par (2.30), comme suit :

$$\begin{aligned} \mathcal{D}_{4,n}^\omega &= (\lambda_1^+ \sigma_B^2) \left(\frac{a_n}{\sigma_B}\right)^2 + (\lambda_2^+ \sigma_B^2) \left(\frac{b_n}{\sigma_B}\right)^2 + (\lambda^- \sigma_B^2) \left(\frac{c_n}{\sigma_B}\right)^2 \\ &= (\lambda_1^+ \sigma_B^2) \tilde{a}_n^2 + (\lambda_2^+ \sigma_B^2) \tilde{b}_n^2 + (\lambda^- \sigma_B^2) \tilde{c}_n^2. \end{aligned}$$

nous exprimons  $\mathcal{D}_{4,n}^\omega$  en fonction des *vas*  $(\tilde{a}_n = \frac{a_n}{\sigma_B})_{n \in \mathbb{N}}$ ,  $(\tilde{b}_n = \frac{b_n}{\sigma_B})_{n \in \mathbb{N}}$  et  $(\tilde{c}_n = \frac{c_n}{\sigma_B})_{n \in \mathbb{N}}$  qui sont des *vas* gaussiennes de moyenne nulle et variance unité. Par conséquent,  $(\tilde{a}_n^2)_{n \in \mathbb{N}}$ ,  $(\tilde{b}_n^2)_{n \in \mathbb{N}}$  et  $(\tilde{c}_n^2)_{n \in \mathbb{N}}$  suivent la loi  $\mathcal{X}_2(1)$  dont la pdf est donnée par (2.31). Ainsi,  $\mathcal{D}_{4,n}^\omega$  s'exprime comme étant une combinaison linéaire de  $\mathcal{X}_2(1)$  dont les coefficients sont les valeurs propres de la matrice  $\mathcal{A}$  multipliés par la variance du bruit  $\sigma_B^2$ .

La détermination de la loi de probabilité d'une combinaison de  $\mathcal{X}_2(1)$  lorsque tous les coefficients sont positifs est faisable. Il suffit juste d'appliquer les règles de probabilités énoncées plus haut. Par contre, dans le cas contraire où au moins un des coefficients est négatif, elle devient impossible. En effet, appliquons (2.32) pour un  $c < 0$  et  $X$  suivant une  $\mathcal{X}_2(1)$ . La pdf résultante est la suivante :

$$f_{c\mathcal{X}_2(1)}(x) = \frac{1}{|c|} \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})} \frac{e^{-\frac{x}{2c}}}{\sqrt{2}} \quad (2.34)$$

Or pour  $x \in \mathbb{R}_+$  et  $c < 0$ , nous aurons  $\frac{x}{c} \in \mathbb{R}_-$ , n'appartenant pas au domaine de définition de la loi  $\mathcal{X}_2(1)$  qui est  $\mathbb{R}_+$ . Ce problème est rencontré dans notre cas :  $\lambda^- < 0$ . Par conséquent, la détermination de la loi de  $\mathcal{D}_{4,n}^\omega$  n'est pas possible. Comme une alternative, nous proposons une approximation de  $\mathcal{D}_{4,n}^\omega$  dont la détermination de loi est réalisable. L'approximation la plus simple est de ne considérer que les termes liés à  $\lambda_1^+$  et  $\lambda_2^+$  et de négliger celui lié à  $\lambda^-$  dans (2.30). Afin d'avoir une idée sur l'erreur que nous pouvons avoir en négligeant ce terme, nous traçons dans la figure 2.2 le rapport  $|\frac{\lambda^-}{\lambda_1^+}|$  et  $|\frac{\lambda^-}{\lambda_2^+}|$  pour des différentes valeurs de  $\nu > 2$ ,  $\kappa \geq 2$  et pour  $M = 60$ .

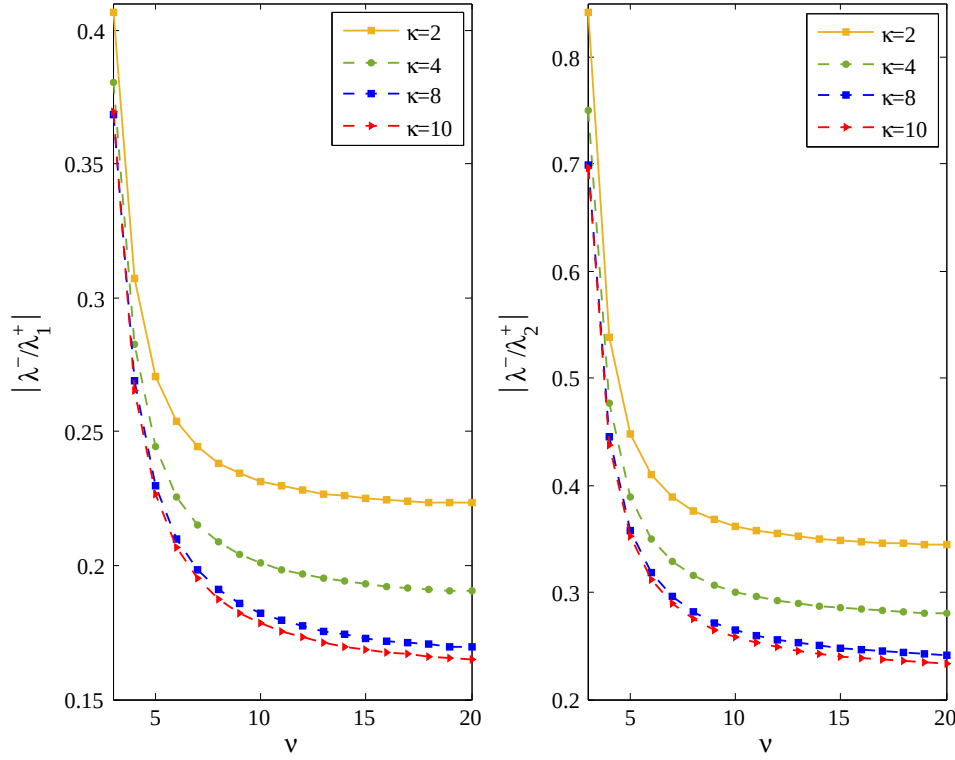


FIGURE 2.2 –  $|\frac{\lambda^-}{\lambda_1^+}|$  et  $|\frac{\lambda^-}{\lambda_2^+}|$  pour  $\nu \in \{3, \dots, 20\}$ ,  $\kappa \in \{2, 4, 8, 10\}$  et  $M = 60$ .

Nous voyons à travers la figure 2.2 pour toutes les valeurs de  $\kappa$  et  $\nu$  les quantités  $|\frac{\lambda^-}{\lambda_1^+}|$  et  $|\frac{\lambda^-}{\lambda_2^+}|$  sont inférieures à 1, ce qui nous montre, que le terme  $\lambda^- c_n^2$  de (2.30) a moins de poids devant les termes  $\lambda_1^+ a_n^2$  et  $\lambda_2^+ b_n^2$ . Plus ces deux quantités s'approchent de 0, plus nous aurons moins d'erreur en négligeant le terme associé à  $\lambda^-$ . Nous remarquons à travers les courbes de la figure 2.2 que pour des valeurs de  $\kappa \geq 4$  et  $\nu \geq 5$  nous obtenons  $|\frac{\lambda^-}{\lambda_1^+}| < 0.2$  et  $|\frac{\lambda^-}{\lambda_2^+}| < 0.4$ . Ceci, nous donnera une idée sur les valeurs de  $\kappa$  et  $\nu$  à choisir pour avoir une bonne approximation de  $\mathcal{D}_{4,n}^\omega$  que nous notons  $\tilde{\mathcal{D}}_{4,n}^\omega$  dont l'expression est donnée par (2.35) et (2.36).

$$\tilde{\mathcal{D}}_{4,n}^\omega = \lambda_1^+ a_n^2 + \lambda_2^+ b_n^2 \quad (2.35)$$

$$= (\lambda_1^+ \sigma_B^2) \tilde{a}_n^2 + (\lambda_2^+ \sigma_B^2) \tilde{b}_n^2. \quad (2.36)$$

$\tilde{\mathcal{D}}_{4,n}^\omega$  est une combinaison de  $\mathcal{X}_2(1)$  de coefficients positifs. Sa fonction de densité de probabilité est donnée par le lemme 2.12.

**Lemme 2.12** *La fonction de densité de probabilité de  $\tilde{\mathcal{D}}_{4,n}^\omega$  est donnée par :*

$$f_{\tilde{\mathcal{D}}_{4,n}^\omega}(x) = \frac{1}{2\sqrt{\lambda_1^+ \lambda_2^+}} e^{-p \frac{x}{\sigma_B}} \mathcal{I}_0\left(q \frac{x}{\sigma_B}\right) \quad (2.37)$$

où  $p = \frac{\lambda_1^+ + \lambda_2^+}{4\lambda_1^+ \lambda_2^+}$ ,  $q = \frac{\lambda_1^+ - \lambda_2^+}{4\lambda_1^+ \lambda_2^+}$ , et  $\mathcal{I}_0$  désigne la fonction de Bessel modifiée de premier espèce et d'ordre 0.



**Preuve 2.13 (Preuve du lemme 2.12)** D'après l'équation (2.32), les pdfs des vas  $(\lambda_1^+ \sigma_B^2 a_n)_{n \in \mathbb{N}}$  et  $(\lambda_2^+ \sigma_B^2 b_n)_{n \in \mathbb{N}}$  sont données respectivement par  $f_{\lambda_1^+}$  et  $f_{\lambda_2^+}$ ,  $\forall x \in \mathbb{R}_+$ , comme suit :

$$f_{\lambda_1^+}(x) = \frac{1}{|\lambda_1^+ \sigma_B^2|} f_{\mathcal{X}_2}\left(\frac{x}{\lambda_1^+ \sigma_B^2}\right) \quad (2.38)$$

$$f_{\lambda_2^+}(x) = \frac{1}{|\lambda_2^+ \sigma_B^2|} f_{\mathcal{X}_2}\left(\frac{x}{\lambda_2^+ \sigma_B^2}\right) \quad (2.39)$$

$$(2.40)$$

La pdf finale de  $\tilde{\mathcal{D}}_{4,n}^\omega$  est alors, d'après (2.33), le produit de convolution entre  $f_{\lambda_1^+}$  et  $f_{\lambda_2^+}$ . Pour simplifier les calculs, nous nous déplaçons au domaine opérationnel [Schwartz, 1998] où le produit de convolution est transformé en produit usuel comme décrit dans le tableau 2.1. Pour ce faire, nous donnons ci-dessous les expressions des fonctions  $\tilde{f}_{\mathcal{X}_2(1)}$ ,  $\tilde{f}_{\lambda_1^+}$ , et  $\tilde{f}_{\lambda_2^+}$  les transformées respectives de  $f_{\mathcal{X}_2(1)}$ ,  $f_{\lambda_1^+}$ , et  $f_{\lambda_2^+}$  dans le domaine opérationnel [Abramowitz and Stegun, 1972].

$$\tilde{f}_{\mathcal{X}_2(1)}(s) = \frac{1}{\sqrt{2}} \frac{1}{\sqrt{s + \frac{1}{2}}}, \quad (2.41)$$

$$\tilde{f}_{\lambda_1^+}(s) = \frac{1}{\sqrt{2}} \frac{1}{\sqrt{\lambda_1^+ \sigma_B^2 s + \frac{1}{2}}}, \quad (2.42)$$

$$\tilde{f}_{\lambda_2^+}(s) = \frac{1}{\sqrt{2}} \frac{1}{\sqrt{\lambda_2^+ \sigma_B^2 s + \frac{1}{2}}}. \quad (2.43)$$

L'analogie opérationnel  $\tilde{f}_{\tilde{\mathcal{D}}_{4,n}^\omega}$  de  $f_{\tilde{\mathcal{D}}_{4,n}^\omega}$  est alors exprimé par :

$$\begin{aligned} \tilde{f}_{\tilde{\mathcal{D}}_{4,n}^\omega}(s) &= \tilde{f}_{\lambda_1^+}(s) \tilde{f}_{\lambda_2^+}(s), \\ &= \frac{1}{\sqrt{2}} \frac{1}{\sqrt{\lambda_1^+ \sigma_B^2 s + \frac{1}{2}}} \frac{1}{\sqrt{2}} \frac{1}{\sqrt{\lambda_2^+ \sigma_B^2 s + \frac{1}{2}}}, \\ &= \frac{1}{2\sqrt{\lambda_1^+ \lambda_2^+}} \frac{1}{\sqrt{\sigma_B^2 s + \frac{1}{2\lambda_1^+}} \sqrt{\sigma_B^2 s + \frac{1}{2\lambda_2^+}}}. \end{aligned}$$

la pdf  $f_{\tilde{\mathcal{D}}_{4,n}^\omega}$ , donnée par (2.44), est par conséquent obtenue par retour au domaine temporel en appliquant la transformée opérationnel inverse sur  $\tilde{f}_{\tilde{\mathcal{D}}_{4,n}^\omega}$ .

$$f_{\tilde{\mathcal{D}}_{4,n}^\omega}(x) = \frac{1}{2\sqrt{\lambda_1^+ \lambda_2^+}} e^{-p \frac{x}{\sigma_B}} \mathcal{I}_0\left(q \frac{x}{\sigma_B}\right). \quad (2.44)$$

Pour des différentes valeurs de  $\kappa$  et  $\nu$ , nous traçons dans la figure 2.3 la cdf empirique de  $\mathcal{D}_{4,n}^\omega$  (courbes en bleu), la cdf empirique de  $\tilde{\mathcal{D}}_{4,n}^\omega$  (courbes en noir) et la cdf analytique de  $\tilde{\mathcal{D}}_{4,n}^\omega$  (courbes en rouge, ligne discontinue), correspondantes à un bruit simulé gaussien de moyenne nulle et de variance unité. À travers les quatre cas considérés dans cette figure, nous voyons que :

1. les courbes en rouge et en noir qui représentent respectivement les cdfs théoriques et empiriques de  $\tilde{\mathcal{D}}_{4,n}^\omega$  pour chaque couple  $(\kappa, \nu)$  sont confondues. Ce qui montre que la loi trouvée, donnée par la pdf de (2.44), correspond bien à la loi de probabilité de  $\tilde{\mathcal{D}}_{4,n}^\omega$ .

2. plus  $\kappa$  et  $\nu$  sont grandes plus la courbe en noir correspondante à la cdf empirique de  $\tilde{\mathcal{D}}_{4,n}^\omega$  s'approche de la courbe en bleu correspondante à la cdf empirique de  $\mathcal{D}_{4,n}^\omega$  et plus nous aurons une bonne approximation de  $\mathcal{D}_{4,n}^\omega$ . Ce qui appuie ce qui a été donné et discuté à travers la courbe 2.2.

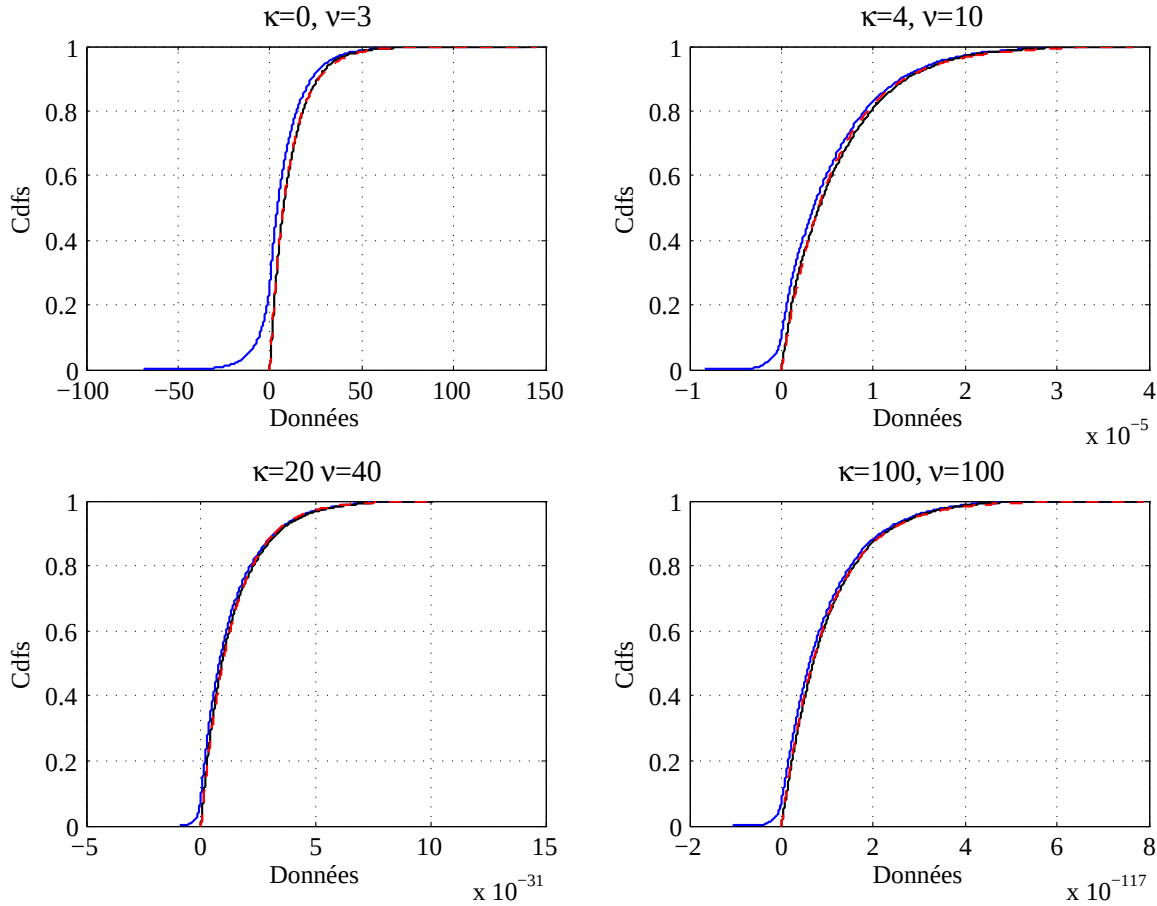


FIGURE 2.3 – La cdf de  $\mathcal{D}_{4,n}^\omega$  vs la cdf de  $\tilde{\mathcal{D}}_{4,n}^\omega$  pour des différentes valeurs de  $\kappa$  et  $\nu$

Une fois la loi de probabilité de  $\tilde{\mathcal{D}}_{4,n}^\omega$  est déterminée, elle sera par la suite utilisée pour la sélection du seuil séparant les fausses alarmes des vrais pics avec une probabilité de fausses alarmes (pfa) donnée. Ceci fera l'objet de la section 2.3.2.

### 2.3.2 Sélection du seuil

Comme nous avons déterminé une loi approximative du bruit issu de la fonction de décision  $\mathcal{D}_4$ , nous allons rester sur le même exemple pour déterminer un seuil de décision, que nous appelons  $\gamma_{\mathcal{D}_4}$ , assurant une pfa inférieure à une pfa donnée et appelée  $p_{\gamma_{\mathcal{D}_4}}$ , *i.e.* la probabilité de détecter une fausse alarme au delà de  $\gamma_{\mathcal{D}_4}$  soit inférieure à  $p_{\gamma_{\mathcal{D}_4}}$ . Nous développons cette partie pour les signaux ayant le modèle suivant :

$$y(t) = \begin{cases} \omega(t), & \text{sous } \mathcal{H}_0; \\ x(t) + \omega(t), & \text{sous } \mathcal{H}_1, \end{cases} \quad \forall t \in \mathbb{R}_+,$$

où  $x$  ici contient que les ruptures. Nous notons que même si le signal original ne suit pas ce modèle : sous  $\mathcal{H}_0$  il n'est pas réduit au bruit mais une composante régulière est rajoutée au

bruit, cela ne posera pas de problème car dans ce cas la version du signal après le traitement algébrique sera réduite au bruit sous  $\mathcal{H}_0$  car la composante régulière sera annulée par les dérivations algébriques (retour à l'exemple de la section 2.2.2).

Comme le signal  $y$  est bruité, le test d'hypothèses de la section 2.2.3 suivant

$$\mathcal{D}_{4,n} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\leq}} 0, \forall \tau \in \mathbb{R}_+, \quad (2.45)$$

sera remplacé par celui de (2.46).

$$\mathcal{D}_{4,n} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\leq}} \gamma_{\mathcal{D}_4}, \forall \tau \in \mathbb{R}_+. \quad (2.46)$$

Nous cherchons à déterminer  $\gamma_{\mathcal{D}_4}$  assurant une  $\text{pfa} \leq p_{\gamma_{\mathcal{D}_4}}$ . Pour ce faire, nous définissons d'abord qu'est ce qu'une fausse alarme. Une fausse alarme apparaît lorsque l'amplitude du signal  $y$  dépasse le seuil  $\gamma_{\mathcal{D}_4}$  et que nous sommes sous l'hypothèse  $\mathcal{H}_0$ , où nous avons que le bruit. Ceci est traduit mathématiquement par l'événement suivant :

$$\mathcal{E}_{\mathcal{D}_{4,n}}^{\mathcal{H}_0} = \{\mathcal{D}_{4,n} > \gamma_{\mathcal{D}_4} \mid \mathcal{H}_0\}.$$

Pour déterminer  $\gamma_{\mathcal{D}_4}$  avec lequel cet événement apparaît avec une probabilité inférieure à  $\gamma_{\mathcal{D}_4}$  il faut déterminer la loi de probabilité de  $\mathcal{E}_{\mathcal{D}_{4,n}}^{\mathcal{H}_0}$ . En effet, Comme sous  $\mathcal{H}_0$  nous avons que le bruit, cet événement sera équivalent au suivant :

$$\mathcal{E}_{\mathcal{D}_{4,n}}^{\mathcal{H}_0} = \{\mathcal{D}_{4,n}^\omega > \gamma_{\mathcal{D}_4}\},$$

dont la détermination de la loi de probabilité a été discutée dans la section 2.3.1. Sachant que d'après (2.30)  $\mathcal{D}_{4,n}^\omega = \tilde{\mathcal{D}}_{4,n}^\omega - \lambda^- c_n^2$ , il s'ensuit que :

$$\mathbb{P}(\mathcal{D}_{4,n}^\omega > \gamma_{\mathcal{D}_4}) < \mathbb{P}(\tilde{\mathcal{D}}_{4,n}^\omega > \gamma_{\mathcal{D}_4}),$$

où  $\mathbb{P}$  désigne la probabilité. Par conséquent, déterminer  $\gamma_{\mathcal{D}_4}$  tel que  $\mathbb{P}(\tilde{\mathcal{D}}_{4,n}^\omega > \gamma_{\mathcal{D}_4}) < p_{\gamma_{\mathcal{D}_4}}$  assurera que  $\mathbb{P}(\mathcal{D}_{4,n}^\omega > \gamma_{\mathcal{D}_4}) < p_{\gamma_{\mathcal{D}_4}}$ . Pour ce faire, il faut juste inverser la fonction de répartition associée à la pdf  $f_{\tilde{\mathcal{D}}_{4,n}^\omega}$ .

Dans la section suivante, nous discutons tous les résultats de simulations de tout ce qui a été développé dans ce chapitre appliqués sur deux types de signaux différents.

## 2.4 Applications

Pour mieux voir l'efficacité des techniques algébriques, décrites dans la section 2.2, en terme de fiabilité de détection et de robustesse au bruit, et pour mieux étudier leurs performances, nous développons dans ce qui suit deux applications où l'information à extraire est caractérisée par l'apparition d'une discontinuité dans le signal à traiter qui sont des signaux à variation bornée.

### 2.4.1 Détection des spikes dans les signaux neuronaux

Nous étudions dans cette section le problème de détection de spikes de neurones. Nous commençons par décrire le signal neuronal.

### 2.4.1.1 Description des signaux neuronaux

Le signal neuronal traduit l'activité électrique du neurone. Il est composé d'un train de potentiels d'action. Chaque potentiel d'action (PA), dit aussi spike, représente le changement rapide de polarité de la cellule nerveuse, suite à la transmission d'un influx nerveux d'un neurone à un autre à travers son axone. Ce changement brusque de polarité, suivi d'une période réfractaire où la cellule nerveuse entre dans une phase d'inhibition, aucun autre PA ne peut être émis par le même neurone, dure en moyenne 3ms. Nous représentons la morphologie d'un PA dans la figure 2.4.

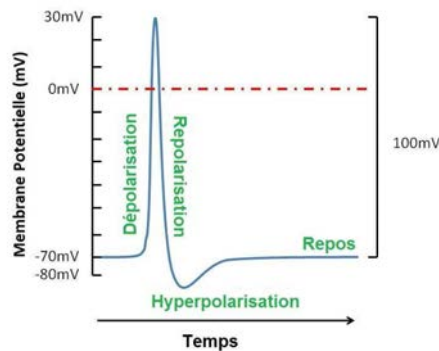


FIGURE 2.4 – Potentiel d'action

Deux différents types d'enregistrement du signal neuronal existent [Dayan and Abbott, 2005]. Le premier est l'enregistrement intracellulaire, pour lequel une micro-électrode est injectée à l'intérieur de l'axone du neurone. Ce type d'enregistrement donne comme résultat un signal clair et net où l'information neuronale est facile à extraire. Par contre, l'effectuer demeure une tâche difficile, vu la taille microscopique du neurone. Pour cette raison, l'enregistrement le plus utilisé est l'enregistrement extra-cellulaire pour lequel la micro-électrode est placée dans le milieu extérieur de l'axone du neurone. Le problème rencontré avec ce type d'enregistrement est que comme le système nerveux est composé d'un nombre relativement énorme de neurones (des milliers de neurones pour les insectes *vs* des milliards de neurones pour l'Homme), la micro-électrode placée dans le milieu extérieur enregistre les activités neuronales de tous les neurones qui l'entourent. Le signal enregistré est ainsi la sommation de toutes ces activités neuronales. Nous représentons dans la figure 2.5, 1s d'enregistrement mesuré à partir du lobe antennaire d'un criquet [Pouzat et al., 2002] et filtré entre 300Hz et 5kHz. Dans cette figure, tous les pics représentent des PAs.

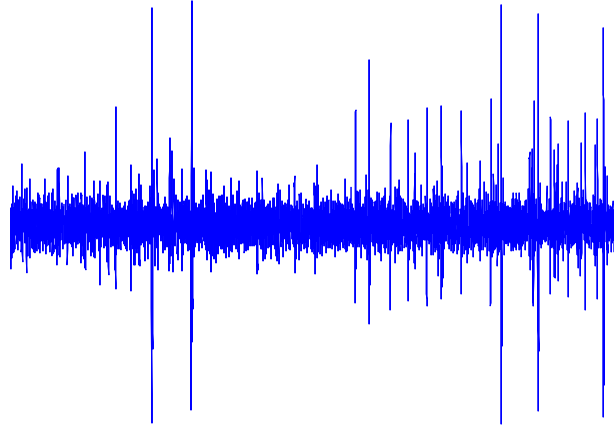


FIGURE 2.5 – 1s d'enregistrement d'activité neuronale du lobe antennaire d'un criquet

Vu que l'extraction de l'information neuronale constitue une étape primordiale pour le décodage de l'information neuronale [Rieke et al., 1999], nous nous intéressons dans ce qui suit à détecter cette information. Plus précisément à détecter l'apparition des spikes, indiquant la réception et l'émission d'une information, dans le signal neuronal. Pour ce faire, nous allons appliquer les techniques algébriques.

#### 2.4.1.2 Approche algébrique pour la détection de spikes de neurones

Vu la durée très rapide du spike, quelques millisecondes, nous caractérisons son apparition comme une discontinuité de première espèce dans le signal neuronal au point admettant l'amplitude la plus élevée. Ainsi, nous transformons le problème de détection de spikes de neurones au problèmes de détection des ruptures dans le signal neuronal. Pour ce faire, et pour pouvoir appliquer les techniques algébriques, nous avons besoin de définir l'ordre de la rupture, la nature des fonctions continues  $(f_i)_{i \geq 0}$  de l'équation (2.2), l'ordre de dérivation  $N$  de (2.4) et la taille de l'intervalle  $I_\tau^T$ . En se basant sur la morphologie du PA, donnée par la figure 2.4, nous avons associé l'ordre 0 à la discontinuité, nous avons aussi choisi de modéliser le signal neuronal par des polynômes d'ordre inférieur ou égal à 1. Pour se débarrasser de la partie régulière de la dérivée de (2.4), nous avons pris  $N = 2$ . Enfin, comme taille de  $I_\tau^T$  nous avons pris  $T$  égal à la durée d'un PA, y compris la période réfractaire. Nous sommes sûrs que durant cette durée au plus un PA peut apparaître. Le cas de  $N = 2$  a été détaillé dans ce chapitre, avec différentes fonctions de décisions qui ont été développées. Afin de voir laquelle entre elles est la plus performante dans notre cas, nous allons étudier les performances des 5 fonctions de décisions décrites dans 2.2.3.

#### Étude de performances des fonctions de décision pour $N = 2$

En traçant les courbes d'efficacité du récepteur, dite en anglais *Receiver Operating Characteristic curves*, ou encore les courbe ROC, nous étudions dans cette section les performances des fonctions de décisions décrites dans la section 2.2.3 en terme de probabilité de détection correcte  $P_{dc}$  et de probabilité de fausses alarmes  $P_{fa}$ . Il s'agit de tracer les courbes de  $P_{dc}$  vs.  $P_{fa}$ . Pour ce faire, une première étape consiste à simuler des signaux neuronaux artificiels auxquels nous rajoutons du bruit avec des différentes valeurs de Rapport Signal sur Bruit (Rapport Signal sur Bruit (RSB)). La génération de ces signaux est décrite ci-dessous.

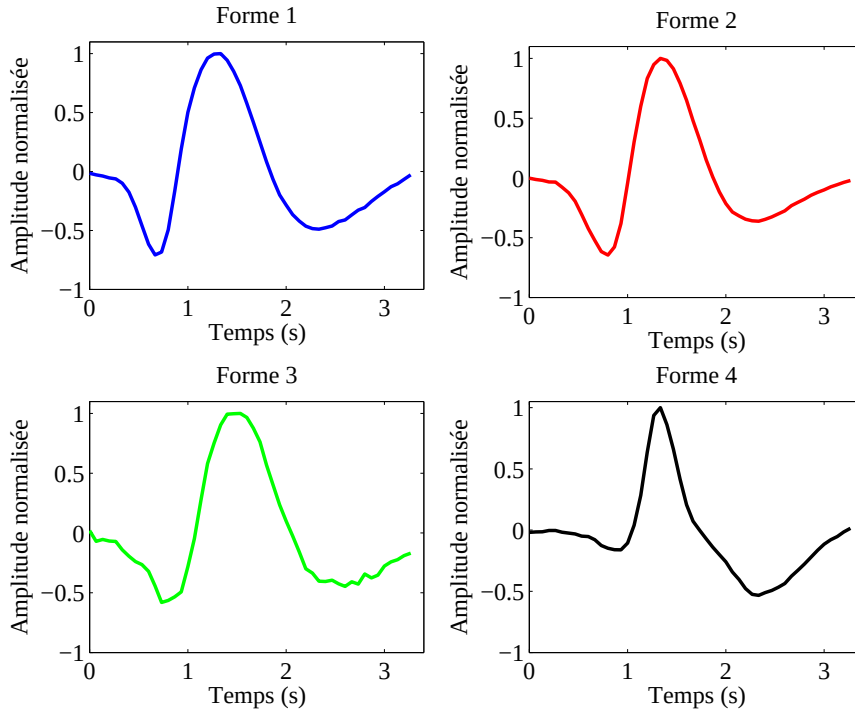


FIGURE 2.6 – Différentes formes de spikes

### Description des signaux générés

Pour se rapprocher le plus de la réalité nous nous sommes basés sur des données neuronales réelles pour générer d'autres synthétiques. Les données neuronales réelles utilisées représentent 4 enregistrements de 20s chacun, échantillonnés à 15kHz, filtrés entre 300Hz et 5000Hz, et mesurés à partir du lobe antennaire d'un criquet [Pouzat et al., 2002]. De ces 4 enregistrements 256 spikes ont été extraits et regroupés en 4 classes. Le centroïde,  $S_i$ ,  $1 \leq i \leq 4$ , de chaque classe a été ensuite considéré comme un modèle de spike et utilisé dans la génération des signaux synthétiques. Pour tout  $1 \leq i \leq 4$ ,  $S_i$  a été normalisé *i.e.*  $\|S_i\|_\infty = 1$ , et est de durée 3.33ms. Nous représentons dans la figure 2.6, les quatres formes obtenues. Le bruit de fond restant dans les 4 enregistrement après l'extraction des 256 spikes a été rassemblé et utilisé comme le bruit neuronal. Il est composé de 285000 échantillons.

Une fois les 4 modèles de spikes et le bruit neuronal sont obtenus, nous procédons à la génération des signaux artificiels qui sont tous composés de  $10^4$  échantillons, équivalent à une durée de 0.66s. La procédure de génération de ces derniers est basée sur les étapes suivantes :

1. Détermination des instants d'apparition de spikes suivant un processus de Poisson, dont le paramètre désigne le taux d'émission de PAs qu'on notera FR (Firing Rate).
2. Répétition de l'étape 1. jusqu'à l'obtention d'une période réfractaire de 2ms après chaque spike.
3. Association aléatoire des modèles de spikes de la figure 2.6 aux instants générés dans les étapes 1. et 2.

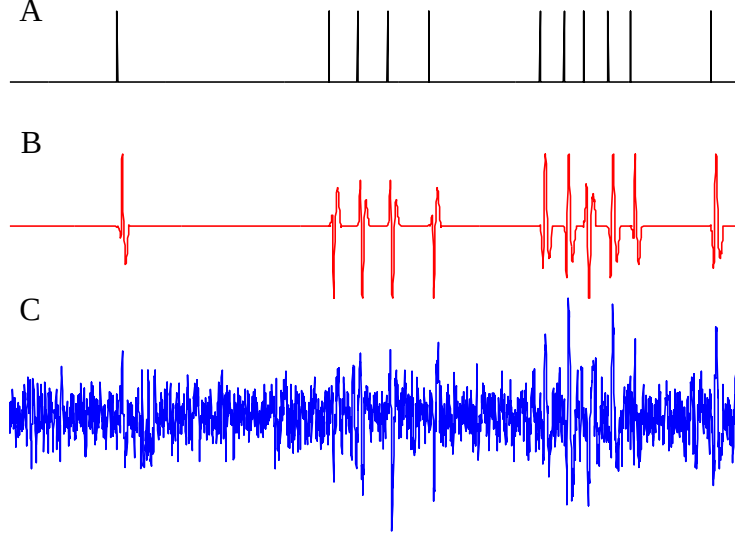


FIGURE 2.7 – Représentation de 0.2s d'un signal neuronal généré pour SNR= 4 et FR= 30kHz.

4. Addition d'un segment, sélectionné aléatoirement, de  $10^4$  échantillons du bruit neuronal avec une variance  $\sigma_B^2$  déterminée suivant une valeur donnée du RSB qui est défini comme dans [Nenadic, 2005] par :

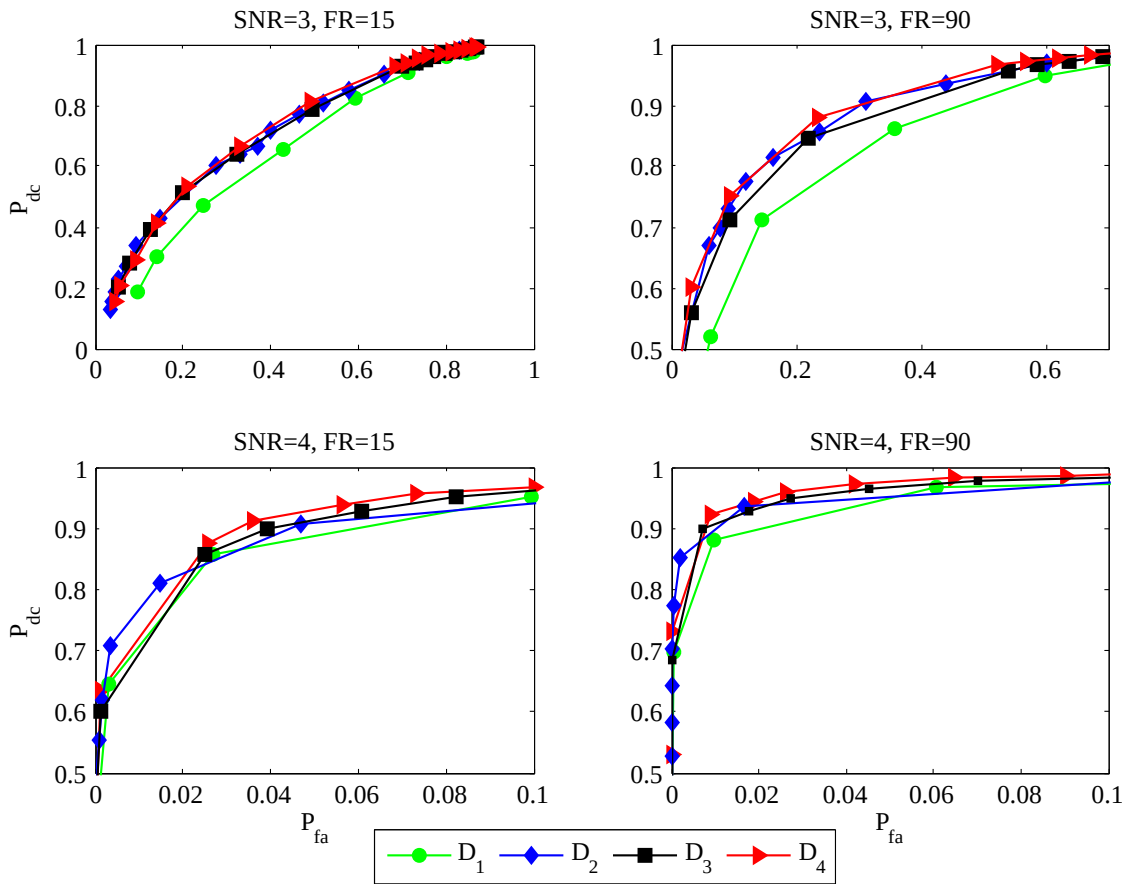
$$RSB = \frac{\|S_i\|_\infty}{\sigma_B} = \frac{1}{\sigma_B}, \quad 1 \leq i \leq 4. \quad (2.47)$$

Nous illustrons ces 4 étapes dans la figure 2.7, où nous considérons 0.2s de signal généré pour un SNR= 4 et FR= 30kHz. Dans cette figure, le signal A représente les instants d'apparition de spikes obtenus après la première et la deuxième étape. Le signal B traduit l'étape 3., où à chaque instant d'apparition de spike, une forme aléatoire de spike est associée. Enfin, le signal C désigne le signal neuronal bruité généré.

Après avoir généré les signaux neuronaux artificiels, nous passons maintenant à tracer les courbes ROC.

### Les courbes ROC

Pour des couples différents (SNR,FR) fixés, nous traçons les courbes ROC des fonctions de décisions  $\{\mathcal{D}_i\}_{i \in \{1, \dots, 4\}}$ . Pour chaque couple (SNR,FR), sur  $10^3$  simulations, nous déterminons  $\tilde{P}_{dc}^{sl}$  et  $\tilde{P}_{fa}^{sl}$ , les moyennes respectives des probabilités de détections correctes et de fausses alarmes empiriques pour les données au delà d'un certain seuil  $sl$ . les probabilités empiriques  $\tilde{P}_{dc}^{sl}$  et  $\tilde{P}_{fa}^{sl}$  sont déterminées comme dans [Nenadic, 2005]. Pour SNR=  $\{3, 4\}$  et pour FR=  $\{15, 90\}$ , nous traçons les résultats obtenus dans la figure 2.8.

FIGURE 2.8 – Les courbes ROC de  $\{\mathcal{D}_i\}_{i \in \{1, \dots, 4\}}$ .

À travers les courbes ROC de la figure 2.8, nous remarquons que les trois fonctions de décisions  $\{\mathcal{D}_i\}_{i \in \{2,3,4\}}$  ont des performances comparables ce qui montre que le choix de la fonction de décision n'influence pas beaucoup sur la qualité de détection. Par contre, les courbes relatives à  $\mathcal{D}_1$  (courbes avec des cercles remplis en vert ●) montrent des performances moins bonnes. Ceci s'explique par le fait que  $\mathcal{D}_1$  est la seule fonction de décision parmi les quatre qui est implémentée comme la sortie d'un filtre linéaire. En effet,  $\{\mathcal{D}_\kappa\}_{\kappa \in \{2,3,4\}}$  sont implémentées comme les sorties des filtres non linéaires. La multiplication des  $v_{\kappa,n}^\nu$ ,  $\kappa \geq 2$ , entre eux permet de mieux réduire l'amplitude du bruit et accentuer celle de spikes. Si par exemple une fausse alarmes est détectée par un  $v_{\kappa_1,n}^\nu$  et non par  $v_{\kappa_2,n}^\nu$ , pour  $\kappa_1 \neq \kappa_2$ ,  $\kappa_1, \kappa_2 \geq 2$ , alors elle ne sera pas détectée par le produit  $v_{\kappa_1,n}^\nu v_{\kappa_2,n}^\nu$ . Ce qui n'est pas le cas avec  $\mathcal{D}_1$ .

Nous voyons également dans cette figure que pour un  $SNR = 4$ , nous obtenons des probabilités de détections correctes élevées ( $P_{dc} > 0.9$ ) à partir des probabilités de fausses alarmes très faible ( $P_{fa} < 0.05$ ) ce qui montre la robustesse de ces techniques algébriques face au bruit.

Pour un niveau de bruit très élevé ( $SNR = 3$ ), ces méthodes arrivent quand même à détecter les spikes avec des  $P_{cd} > 0.7$  pour des  $P_{fa} < 0.5$  (sous-figure en haut à gauche).



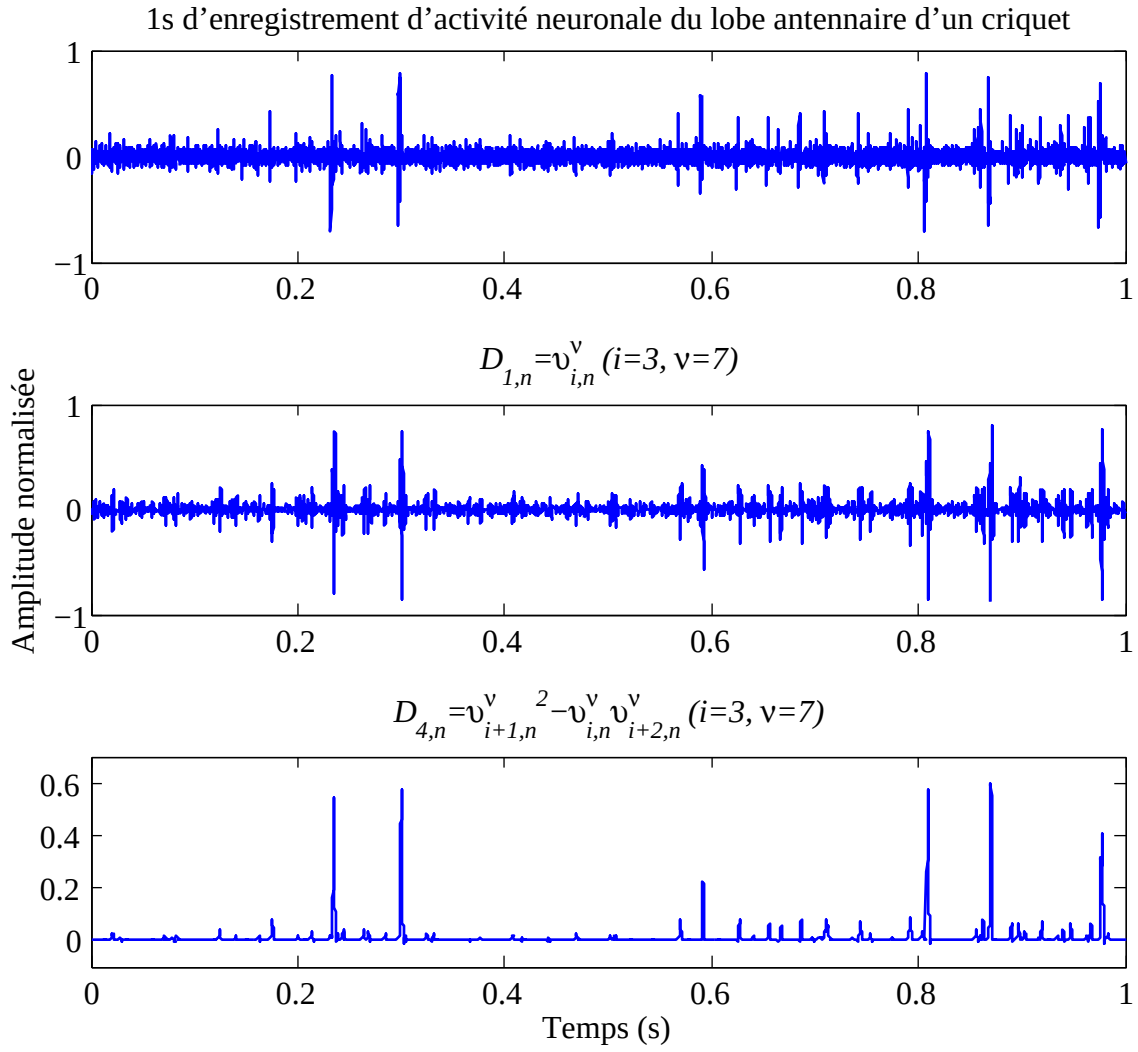
Nous mentionnons que les performances de ces techniques augmentent lorsque les FR augmentent ce qui est illustré dans la figure 2.8.

Nous remarquons que la performance du détecteur  $\mathcal{D}_4$  (courbes avec des triangles en rouge ►) est légèrement meilleures par rapport aux autres puisque ses courbes ROC sont au-dessus des autres courbes. Pour cette raison, nous allons l'adopter pour la détection des spikes de neurones.

Nous illustrons dans la figure 2.9 l'application de la fonction de décision  $\mathcal{D}_4$  sur 1s d'un signal neuronal réel, donné par la figure 2.5, pour la détection des spikes de neurones. Nous voyons bien qu'avec  $\mathcal{D}_{4,n}$  (le troisième signal de la figure) nous obtenons une réduction remarquable de l'amplitude du bruit et un rehaussement important de celle des spikes permettant de distinguer facilement entre spikes et fausses alarmes. Pour argumenter ce qui a été dit sur la fonction de décision  $\mathcal{D}_1$  plus haut, nous traçons dans la sous-figure au milieu  $\mathcal{D}_{1,n}$ . Une différence considérable entre les résultats donnés par  $\mathcal{D}_1$  et  $\mathcal{D}_4$  met en évidence l'utilisation des filtres non-linéaires.

Nous avons jusqu'à présent étudié les performances des fonctions de décisions développées dans ce chapitre en les comparant entre elles. En se basant sur les courbes ROC, nous avons trouvé que la fonction de décision  $\mathcal{D}_4$  offre de meilleure performance en terme de probabilité de bonne détection et probabilité de fausses alarmes. Nous allons par la suite comparer la performance de cette dernière avec celles de deux méthodes existantes dans la littérature qui ont montré dans différents domaines une fiabilité de détection. Il s'agit de la méthode de la transformation en ondelettes [Nenadic, 2005] et la méthode non linéaire de détection d'énergie, dite en anglais Nonlinear Energy Operator (NEO) [Kim and Kim, 2000]. Nous traçons les courbes ROC des trois méthodes dans la figure 2.10 pour des différentes valeurs de SNR et FR.

Nous remarquons à travers la figure 2.10 que les courbes relatives à la technique algébriques (triangle rouge ►) sont presque toujours en dessus des deux autres pour les SNR= 3.5 et SNR= 4, montrant que cette technique offre plus de fiabilité en terme de détection par rapport à la méthode des ondelettes et celle de NEO. Par contre dans un environnement très bruité (SNR= 3) elle garde de performance meilleure lorsque  $P_{fa} > 0.5$ .

FIGURE 2.9 – Application de la fonction de décision  $\mathcal{D}_4$  pour la détection des spikes de neurones.

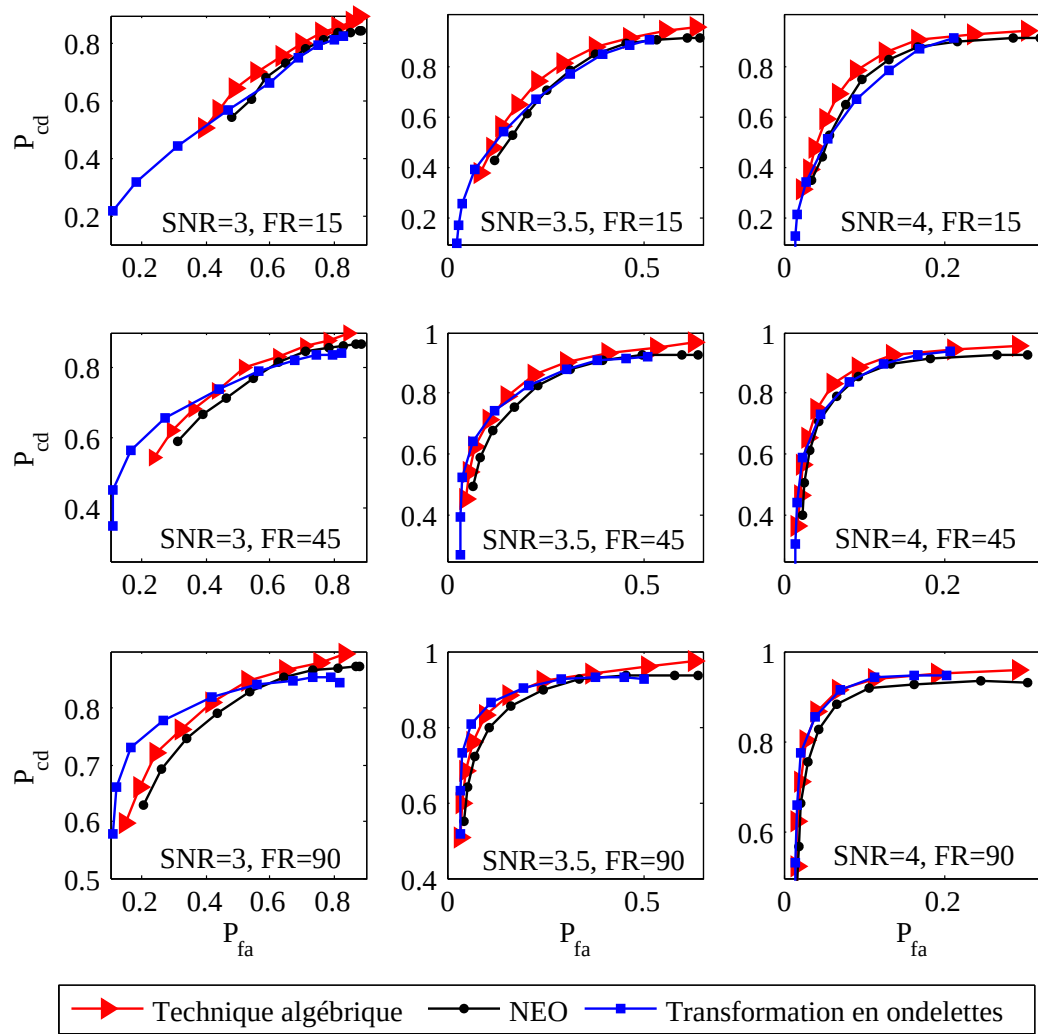


FIGURE 2.10 – Les courbes ROC des trois méthodes de détection des spikes : la technique algébrique ( $\blacktriangleright$ ), la méthode NEO ( $\bullet$ ) et la méthode par transformation des ondelettes ( $\blacksquare$ ).

### Exemples de sélection de seuil

Pour tester la performance du seuil distinguant entre les ruptures et les fausses alarmes avec une pfa inférieure à une pfa donnée, nous nous sommes basés sur les simulations de Monte Carlo. En effet, nous avons fixé un vecteur de pfas :  $[0.02, 0.04, 0.06, 0.08, 0.1]$ . Pour chaque valeur de pfa fixée, nous avons généré 500 signaux neuronaux. Pour chaque signal, nous avons déterminé le seuil  $\gamma_{\mathcal{D}_4}$  de la section 2.3.2 et déterminé la pfa empirique associée. Nous avons à la fin comparé la moyenne des pfas empiriques, calculé comme dans [Nenadic, 2005], sur 500 simulations avec la pfa fixée. Nous illustrons les résultats obtenus pour  $SNR = 4$ ,  $FR = 45$ ,  $\kappa = 3$ ,  $\mu = 7$ , et  $M = 60$  dans la figure 2.4.1.2. Nous voyons à travers cette figure que les pfas empiriques sont toujours inférieures à celle fixées ce qui appuie la partie théorique sur la sélection du seuil (section 2.3.2). D'où on voit l'apport de la modélisation stochastique du bruit pour la détection de ruptures.

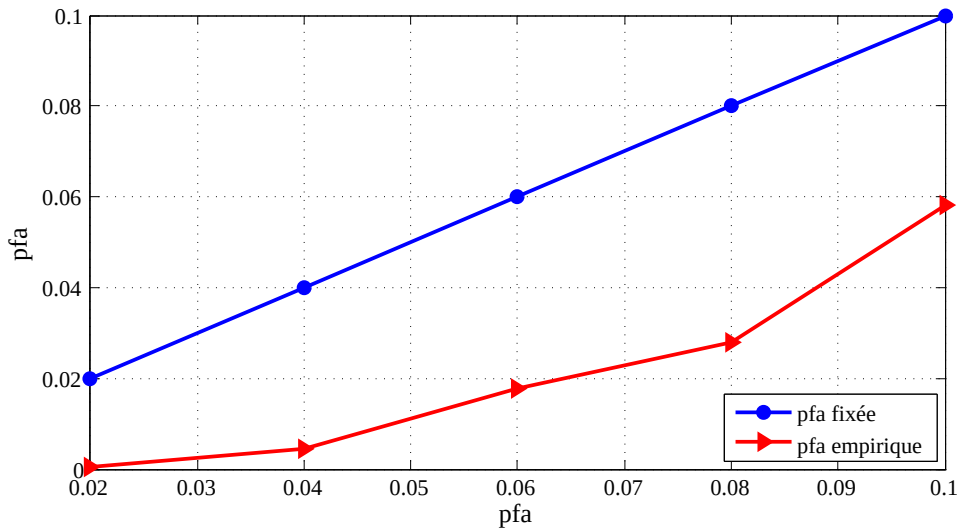


FIGURE 2.11 – pfa fixées *vs* pfa empiriques obtenues en appliquant le seuil déterminé dans la section 2.3.2 sur un signal neuronal généré pour  $SNR = 4$ ,  $FR = 45$ ,  $\kappa = 3$ ,  $\mu = 7$ , et  $M = 60$ .

Il faut mentionner que la performance de la sélection du seuil dépend du choix des paramètres  $\kappa$  et  $\nu$  et de la valeur de  $SNR$ .

#### 2.4.2 Détection des ondes R dans les signaux Electrocardiogramme (ECG)

Dans cette section, nous appliquons l'une des approches algébriques basée sur l'équation (2.4) de la section 2.2 sur un autre jeu de données afin de détecter les ruptures qu'il contient : il s'agit du signal ElectroCardioGramme (ECG) que nous décrivons ci-dessous.

### 2.4.2.1 Description des signaux ECG

Le signal ECG est le signal traduisant l'activité électrique du cœur. Pour chaque cycle cardiaque, ce signal est composé d'une séquence d'ondes nommées : l'onde P, le complexe QRS, et l'onde T [Hertault, 1985], dont la durée moyenne est de 0.8s. Nous représentons sa morphologie d'un état sain dans la figure 2.12.

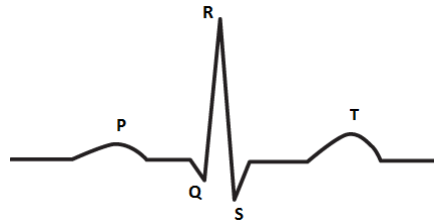


FIGURE 2.12 – Signal ECG d'un état sain pour un seul cycle cardiaque

Le signal ECG est enregistré à travers des électrodes placées dans des zones bien définies du corps [Hertault, 1985]. Lors de l'enregistrement, le signal ECG est généralement contaminé par du bruit de différentes origines. Ce bruit peut être d'origine environnementale comme les interférences du réseaux. Il peut aussi être d'origine physiologique tels que la dérive de la ligne de base, constituant la ligne isoélectrique de l'activité cardiaque, et le mouvement du patient comme la toux et l'éternuement qui peuvent modifier la position des électrodes. Aussi, le mouvement des muscles qui est un bruit d'origine biologique pouvant perturber le signal ECG. La présence de ce bruit dans les signaux ECG rend l'extraction des informations cliniques difficile ce qui peut engendrer, dans certains cas, des diagnostics erronés des arythmies cardiaques.

Nous représentons dans la figure 2.13 quatre exemples de signaux ECG réels, représentant des morphologies et des anomalies différentes. Ces signaux sont obtenus de la base de données MIT BIH des arythmies cardiaques [MIT, 1980].

Dans ces signaux nous nous intéressons à détecter les complexes QRS. En effet, le complexe QRS du signal ECG représente la partie la plus significative du signal. Il traduit la dépolarisation des ventricules admettant la majorité de la masse du cœur. Plus particulièrement, nous nous focalisons sur la détection des ondes R qui ont l'amplitude la plus élevée. La détection de cette dernière permet principalement le diagnostic de certaines anomalies, mais aussi elle permet d'identifier les cycles cardiaques et de déterminer leur fréquence.

À première vue, d'après la figure 2.13, nous pouvons avoir l'impression que la détection des pics R est facile puisque l'onde R admet l'amplitude la plus élevée et que la détection des ondes R peut s'effectuer par un simple seuillage du signal ECG. Toutefois, ceci reste vrai dans le cas où le signal est parfaitement enregistré et qu'il représente l'activité électrique d'un état sain. Dans le cas contraire, représentant la majorité des signaux enregistrés, la détection des ondes R devient compliquée. En effet, certaines anomalies sont représentées par l'apparition des complexes QRS renversés qui ne peuvent pas être détectés par un simple seuillage, comme

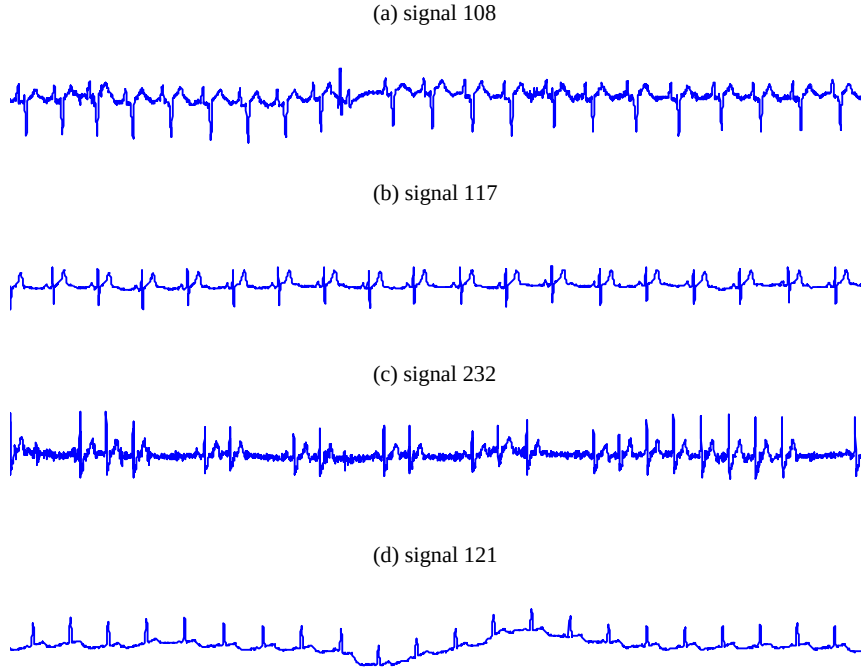


FIGURE 2.13 – Signal ECG d'un état sain

le cas du signal 108 de la base MIT-BIH (signal (a) de la figure 2.13). Aussi, certaines maladies reposent sur l'apparition d'une onde T d'amplitude comparable à celle de l'onde R qui peuvent être détectées à la place de ces dernières comme dans le cas du signal (b) de la figure 2.13. Le signal (c) de la même figure montre un cycle cardiaque non-uniforme alors que le signal (d) souligne la déformation de la ligne de base où l'application d'un simple seuillage augmente la probabilité de louper les pics R. Tous ces problèmes montrent le besoin d'un traitement plus sophistiqué des signaux ECG. Pour ce faire, nous avons appliqué une approche algébrique pour détecter les pics R que nous avons considéré comme des ruptures dans le signal ECG [Debbabi and El Asmi, 2012]. Nous donnons plus de détails dans la section suivante.

#### 2.4.2.2 Approche algébrique pour la détection des pics R dans le signal ECG

Sachant que le complexe QRS dure en moyenne 0.08s, nous avons caractérisé ce changement brusque de polarité des ventricules par une discontinuité de première espèce et d'ordre 0. Pour les détecter, nous nous sommes basés sur l'équation (2.4) de la section 2.2. Pour des fonctions régulières constantes, les  $(f_i)_{i \geq 0}$  de (2.2), pour  $N = 1$ , l'ordre de dérivation de (2.4) et pour un  $T$ , la largeur de  $I_\tau^T$ , qui vaut la durée moyenne d'un cycle cardiaque que nous la noterons  $T_{ECG} = 0.8s$ , nous avons construit un détecteur algébrique [Debbabi and El Asmi, 2012] qui permet de détecter et localiser les pics R. En effet, comme  $N = 1$ , le système d'équations différentielles 2.11 sera alors équivalent à une seule équation affine en  $\tilde{t}_\tau$ , l'estimée de la position du pic R  $t_\tau$ . Nous donnons l'expression de cette équation dans le domaine opérationnel :

$$\tilde{v}_{\kappa+1,\tau}^{\nu,1}(s) + \tilde{t}_\tau \tilde{v}_\kappa^{\nu,1}(s, \tau) = 0 \quad (2.48)$$

où  $\forall \tau \in I_\tau^T$ ,  $\kappa > 0$ ,  $\nu > 1$ ,  $\tilde{v}_\kappa^{\nu,1}(s, \tau) = \frac{1}{s^\nu} \frac{d^\kappa}{ds^\kappa}(s \tilde{y}_\tau(s))$ . L'analogie opérationnel  $\tilde{y}_\tau$  du signal  $y_\tau$  représente dans cette application le signal ECG. Notons que l'exposant 1 de la fonction  $\tilde{v}_\kappa^{\nu,1}(\cdot)$  désigne ici la valeur de  $N$  qui vaut 1. Comme il s'agit du temps qui est positif, pour que  $\tilde{t}_\tau$  soit positif, il faut que les coefficients de l'équation (2.48) soient de signes opposés. Pour cette raison, nous avons choisi la fonction de décision suivante :

$$\mathcal{D}_{ECG}(\tau) = -v_{\kappa+1}^{\nu,1}(\tau)v_\kappa^{\nu,1}(\tau), \quad (2.49)$$

où  $v_\kappa^{\nu,1}(\tau)$  est l'analogie temporel de  $\tilde{v}_\kappa^{\nu,1}(s, \tau)$  évalué à l'instant  $T$ .

À cette fonction de décision nous associons le critère de décision suivant :

- $\mathcal{D}_{ECG}(\tau) \geq 0$ , alors  $I_\tau^T$  contient un pic R,
- $\mathcal{D}_{ECG}(\tau) = 0$ , aucun pic R n'apparaît dans  $I_\tau^T$ .

Puisque le signal ECG est contaminé par le bruit,  $\mathcal{D}_{ECG}(\tau)$  peut être positive mais indiquant une fausse alarme. Pour distinguer entre les pics R et ceux liés au bruit, contrairement à ce qui a été développé pour les signaux neuronaux, nous avons proposé la règle de décision présentée ci-dessous.

### 2.4.2.3 Règle de décision

Nous proposons dans cette section un algorithme qui permet de réduire le nombre de fausses alarmes et de bien les séparer des pics R. Il est basé sur trois étapes principales :

#### **Étape 1 : Interpolation polynomiale**

Comme le détecteur  $\mathcal{D}_{ECG}$  est le résultat de plusieurs dérivations, beaucoup de fluctuations apparaît alors dans le signal ECG traité par ce détecteur. Ainsi, pour réduire le nombre de pics liés à ces fluctuations, nous avons interpolé, par des fonctions splines cubiques, tous les maxima positifs de  $\mathcal{D}_{ECG}$ . Un exemple sur la base MIT-BIH est donnée par la figure 2.14 où le signal bleu représente la fonction de décision  $\mathcal{D}_{ECG}$ , et les étoiles bleues tous les maxima correspondants. Dans cette figure, seulement une partie de  $\mathcal{D}_{ECG}$  est représentée, contenant un seul vrai pic R (carré vert). Nous pouvons très bien remarquer à travers la figure 2.14 le nombre de fausses alarmes que nous pouvons détecter pour un seul pic R. La courbe rouge en gras désigne la fonction d'interpolation. Il est clair que le nombre de pics détectés avec cette fonction (losanges rouges) est beaucoup plus réduit que celui de  $\mathcal{D}_{ECG}$ . Nous considérons dans la suite que les pics obtenus après l'interpolation. Éliminer les fausses alarmes de ces derniers fera l'objet de l'étape 2.

#### **Étape 2 : Elimination des fausses alarmes**

Soit  $t_R$  la position d'un pic R détecté. Comme un cycle cardiaque ne peut contenir qu'un seul pic R, alors aucun autre ne pourra être détecté dans un intervalle centré en  $t_R$  et de largeur  $2d_{ECG}$ , où  $d_{ECG}$  désigne la durée d'un cycle cardiaque. Comme cette dernière n'est pas connue : elle varie d'une personne à une autre et même pour la même personne, la durée du cycle cardiaque

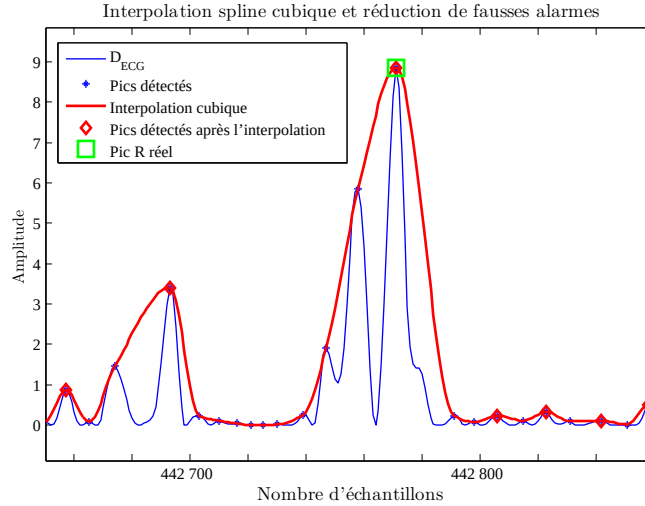


FIGURE 2.14 – Interpolation des maxima par les splines cubiques et réduction des fausses alarmes : application sur le signal 108 de la bas MIT-BIH. Ici l'ordre de dérivation est 3.

dépend de son activité physique, nous la remplaçons par  $\alpha T_{ECG}$ ,  $\alpha \in (0, 1]$ . Ce paramètre  $\alpha$  permet de considérer les cycles cardiaques dont la durée est inférieure à  $T_{ECG}$ . Par conséquent, si un pic est détecté, il sera le seul détecté dans un intervalle de largeur  $2\alpha T_{ECG}$ . Nous allons ainsi se baser sur ce fait pour éliminer les fausses alarmes. Pour ce faire nous utilisons l'algorithme suivant :

1. Déterminer le maximum du signal. Puisque l'onde R a l'amplitude la plus élevée dans le signal. Nous sommes alors sûrs que le maximum représentera un pic R.
2. Garder le pic R détecté.
3. Éliminer toutes les composantes du signal trouvées dans l'intervalle centré à la position du maximum trouvé dans le point 1. et de largeur  $2\alpha T_{ECG}$ .
4. Répéter les points 1. et 2. sur le même signal jusqu'à ce que toutes les composantes de ce derniers deviennent nulles.

Nous illustrons les principes de cet algorithme dans la figure 2.15 où la courbe rouge désigne la fonction d'interpolation et les losanges rouges sont les pics qu'ils lui correspondent. Les deux carrés vert sont les vrais pics R. Les deux parties hachurées désignent la partie à éliminer dans un intervalle de largeur  $2\alpha T_{ECG}$ . Comme Il est montré dans cette figure, Les fausses alarmes ont été presque éliminées. Par contre dans certains cas comme celui de la figure 2.15, entre deux pics R successifs, nous pouvons détecter une fausse alarme. Ceci est dû au fait que le signal ECG est pseudo-périodique et que pour deux pics R successifs, les deux intervalles d'élimination des fausses alarmes peuvent étre disjoints. Pour déceler ce problème, nous passons à l'étape de vérification qui constitue la dernière étape de la règle de décision.

### Étape 3 : Vérification

Pour vérifier qu'aucune fausse alarme n'est détectée, nous effectuons un test sur tous les pics détectés après l'étape 2 de la règle de décision. En effet, soit  $\{pic(i)\}_{0 \leq i \leq n_{pd}}$ , tous les pics



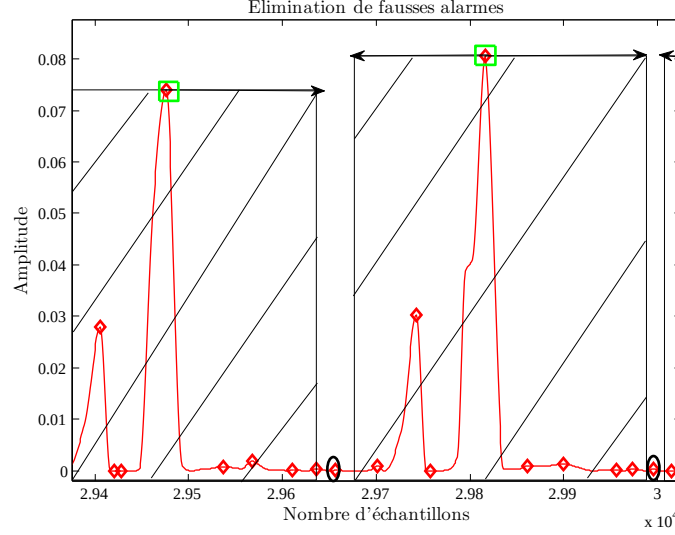


FIGURE 2.15 – Elimination de fausses alarmes

détectés. Le nombre total de ces pics est  $n_{pd}$ . Pour tout  $i \in \{6, \dots, n_{pd} - 5\}$ , nous comparons  $pic(i)$  à  $m_i^-$ , la moyenne des 5 pics détectés précédant  $pic(i)$ , et à  $m_i^+$ , la moyenne des 5 pics détectés après  $pic(i)$ . Les expressions de  $m_i^-$  et  $m_i^+$ ,  $\forall i \in \{6, \dots, n_{pd} - 5\}$ , sont données par :

$$m_i^- = \frac{1}{5} \sum_{k=i-5}^{i-1} pic(k),$$

$$m_i^+ = \frac{1}{5} \sum_{k=i+1}^{i+5} pic(k).$$

La décision finale est ainsi prise comme suit :  $\forall i \in \{6, \dots, n_{pd} - 5\}$

- si  $pic(i) < \beta m_i^-$  et  $pic(i) > \beta m_i^+$ ,  $\beta \in (0, 1]$ , alors  $pic(i)$  est une fausse alarme, donc nous l'ignorons.
- Dans le cas contraire,  $pic(i)$  est retenu comme un pic R.

Ici,  $\beta$  désigne un facteur régulateur. Il permet de régler l'amplitude des pics R puisqu'ils peuvent varier d'un signal à un autre et même dans le même signal d'un cycle à un autre.

Nous représentons quelques résultats de l'application du détecteur  $\mathcal{D}_{ECG}$  pour détecter les pic R présents dans le signal ECG.

#### 2.4.2.4 Résultats

Pour étudier la performance du détecteur  $\mathcal{D}_{ECG}$ , nous l'avons appliqué sur des signaux issus de la base MIT-BIH [MIT, 1980]. Cette base contient 48 enregistrements annotés obtenus de 47 sujets. 60% des enregistrements ont été mesurés sur des sujets hospitalisés. Chaque signal de cette base est échantillonné à 360 Hz et dure environ 30 minutes. Pour ce faire, nous nous

| Enregistrement | Nombre de battements | Méthode             | $n_{dr}$ | $n_{df}$ | $S_e$ (%) | $S_p$ (%) |
|----------------|----------------------|---------------------|----------|----------|-----------|-----------|
| 104            | 2229                 | $\mathcal{D}_{ECG}$ | 0        | 2        | 100       | 99,91     |
|                |                      | TOC                 | 0        | 2        | 100       | 99.91     |
| 116            | 2412                 | $\mathcal{D}_{ECG}$ | 16       | 0        | 99.34     | 100       |
|                |                      | TOC                 | 16       | 1        | 99.34     | 99.96     |
| 203            | 2980                 | $\mathcal{D}_{ECG}$ | 11       | 2        | 99.63     | 99.93     |
|                |                      | TOC                 | 15       | 3        | 99.50     | 99.90     |
| 213            | 3251                 | $\mathcal{D}_{ECG}$ | 3        | 0        | 99.91     | 100       |
|                |                      | TOC                 | 3        | 1        | 99.91     | 99.97     |
| 221            | 3427                 | $\mathcal{D}_{ECG}$ | 1        | 0        | 99.96     | 100       |
|                |                      | TOC                 | 3        | 3        | 99.88     | 99.91     |

TABLE 2.2 – Résultats de l’application de  $\mathcal{D}_{ECG}$  sur 5 signaux de la base de données MIT-BIH présentant des morphologies complexes pour la détection des pics R : Comparaison avec la méthode TOC

sommes basés sur deux critères de performances : la spécificité et la sensibilité du détecteur, dont les expressions sont données par (2.50) et (2.51).

$$S_p = \frac{n_{dc}}{n_{dc} + n_{df}}, \quad (2.50)$$

$$S_e = \frac{n_{dc}}{n_{dc} + n_{dr}}, \quad (2.51)$$

où  $n_{dc}$ ,  $n_{df}$ , et  $n_{dr}$  désignent respectivement le nombre de détections correctes, le nombre de détections fausses et le nombre de détections ratées. Notons que  $S_p$  et  $S_e$  correspondent respectivement à la probabilité de détection correcte et la probabilité de fausses alarmes. Nous allons les étudier ci-après en terme de pourcentage, en les multipliant pas 100.

Nous représentons dans le tableau 2.2, les spécificités et la sensibilités du détecteur algébrique  $\mathcal{D}_{ECG}$ , obtenues pour 5 signaux de la base MIT-BIH. Nous avons choisi de représenter les signaux 104, 116, 203, 213, et 221 représentant une fréquence de bruit élevée et des morphologies complexes. Les résultats obtenus sont aussi comparés dans le même tableaux par ceux obtenus par la méthode la plus utilisée dans la littérature pour la détection des pics R dans les signaux ECG : Il s’agit de la méthode de la Transformée en Odelettes Continues (TOC) [Li et al., 1995, Yoon et al., 2010]. Comme il est montré dans le tableau 2.2,  $\mathcal{D}_{ECG}$  améliore la détection de pic R en terme de bonne détection et de fausses alarmes. En effet, pour l’enregistrement 104 les performances des deux méthodes sont pareils. Pour les enregistrements 203 et 213, les deux méthodes ratent le même nombre de pics R, par contre le  $\mathcal{D}_{ECG}$  ne détecte aucune fausse alarme. Pour les signaux 213 et 221, le  $\mathcal{D}_{ECG}$  réduit le nombre de fausses alarmes et le nombre de pics ratés par rapport à la TOC.

Dans la figure 2.16, nous illustrons l’application de  $\mathcal{D}_{ECG}$  sur le signal 108. Seulement une partie de ce signal est représentée. Le signal 108 a la particularité d’avoir une morphologie compliquée et d’être le signal le plus bruité dans la base de données MIT-BIH. Il a une ligne de base ondulée, des complexes QRS qui sont parfois normaux et parfois renversés (voir l’échelle

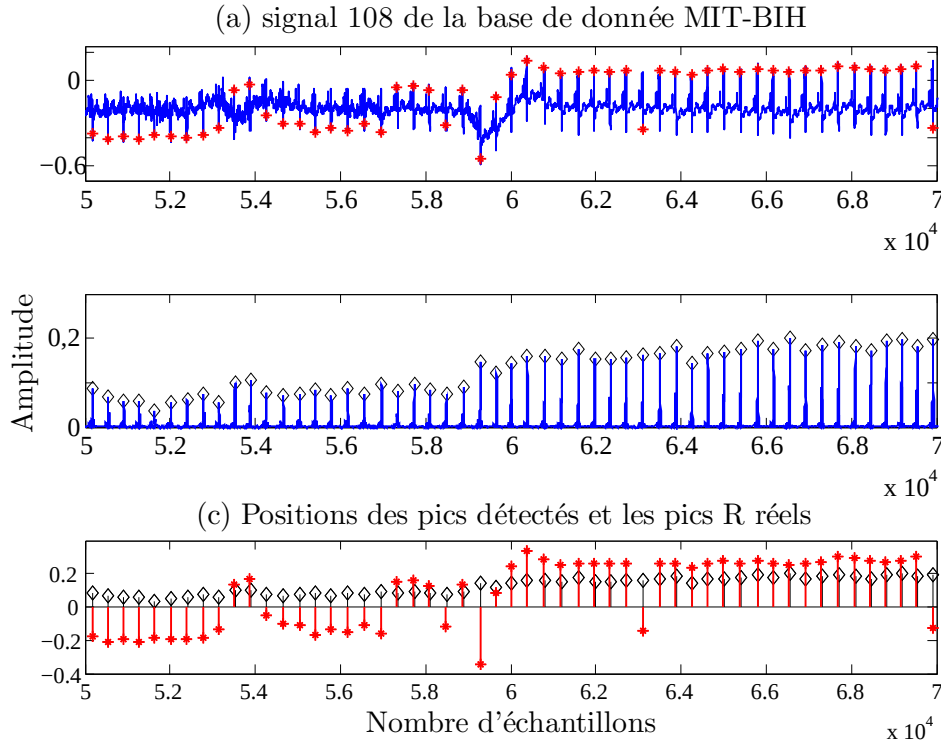


FIGURE 2.16 – Application de l'approche algébrique pour la détection des pics R du signal ECG

des abscisses du signal (a) entre  $6.2 \cdot 10^4$  et  $6.4 \cdot 10^4$ ), des amplitudes de pics R variées, etc. Dans cette figure, les étoiles rouges désignent les vrais pics R alors que les losanges noirs désignent les pics R détectés avec le  $\mathcal{D}_{ECG}$ . Le signal (b) illustrant le résultat de l'application de  $\mathcal{D}_{ECG}$  sur le signal 108 montre bien qu'à part la bonne détection, le détecteur algébrique permet de résoudre l'ondulation de la ligne de base qui est claire dans le signal (a). La comparaison entre les pics réels et ceux détectés est représentés sur le signal (c) où on remarque bien une bonne détection sur cette partie du signal malgré la complexité de sa morphologie. À part la bonne performance du détecteur algébrique, le tableau 2.2 et la figure 2.4.2.4 soulignent aussi la bonne performance de la règle de décision qui a joué un rôle important pour la réduction de fausses alarmes. Notons qu'en se basant sur les simulations, le  $\alpha$  de l'étape 2 de la règle de décision a été prise égale à 0.25. Alors que le  $\beta$  de l'étape 3 de vérification donne de meilleurs résultats s'il est pris dans l'intervalle  $[0.3, 0.4]$ .

## 2.5 Conclusion

Ce chapitre a été divisé en deux grandes parties théoriques validées par les simulations. Basée sur une modélisation déterministe des ruptures tout en ignorant la présence du bruit dans les signaux que nous traitons, la première partie a développé l'approche algébrique utilisée pour la détection des ruptures. Nous avons vu à travers les simulations que cette approche offre de bonne performance en terme de détection de ruptures. Dans le but d'améliorer encore la détection de ruptures, la deuxième partie de ce chapitre a été consacrée à l'exploitation de l'information que nous pouvons avoir sur le bruit. En associant une modélisation stochastique au bruit, nous

avons estimé sa loi de probabilité. Nous avons pu ainsi déterminer un seuil au-delà duquel une fausse alarme apparaît avec une probabilité inférieure à une valeur donnée. Pour aller plus loin dans la détection des ruptures, nous allons maintenant garder la modélisation déterministe de la rupture en lui associant une modélisation stochastique de cette dernière. En effet, nous associons à l'apparition d'une rupture un événement rare d'amplitude relativement extrême. À travers la TVE, nous allons nous intéresser par la suite à étudier le comportement extrême de nos signaux. Cette partie fera l'objet du troisième chapitre.

## Chapitre 3

# Rupture : un évènement rare

### 3.1 Introduction

Dans ce chapitre, la notion de rupture est abordée différemment. Son apparition est considérée comme un événement rare apparaissant avec une faible probabilité d'occurrence et une forte, voire extrême, amplitude. Avec cette nouvelle caractérisation de la rupture, nous transformons le problème de détection de ruptures en un problème de détermination des valeurs extrêmes associées à des événements rares dans les signaux traités par l'approche algébrique. En effet, cette fois-ci, nous ignorons le comportement moyen des données, représentant le bruit, et nous ne nous intéressons qu'au comportement extrême modélisant les ruptures. En se basant sur le théorème de Balkema, de Haan et Pickands de la TVE, nous approchons la loi de probabilité des extrêmes au delà d'un certain seuil, par la loi de Pareto généralisée dont le paramètre de forme est fortement lié à la valeur du seuil. D'où l'importance d'une détermination judicieuse de ce seuil. Dans un premier temps, nous déterminons ce dernier graphiquement via la méthode du '*mean excess plot*'. Les résultats obtenus sur les signaux neuronaux, montrent bien que la connaissance de la loi des extrêmes permet mieux la distinction entre les fausses alarmes et les ruptures, rendant la détection de ruptures plus fiable. Afin d'automatiser et améliorer cette méthode, nous proposons une détermination non supervisée du seuil, en développant un algorithme convergeant vers un point stationnaire, le seuil. Cette détermination non supervisée du seuil est en fait basée sur la considération de toutes les données : comportements moyen et extrêmes ou encore bruit et ruptures. En effet, pour aller plus loin dans la détection des ruptures, les résultats que nous avons obtenu sur les extrêmes nous ont motivé à avoir une vue d'ensemble sur la distribution de probabilité de tout le signal en modélisant le comportement moyen des données (le bruit) par une loi gaussienne via le TCL et le comportement extrême (les ruptures) par une Pareto généralisée via le théorème de Pickands. Les pdfs de ces deux lois de probabilité sont liées entre elles par un point de jonction représentant le seuil indiquant la présence des extrêmes. Une pdf hybride à deux composantes est ainsi obtenue, modélisant les ruptures avec le bruit. Contrairement à ce qui existe dans la littérature, cette pdf hybride ne permet pas seulement l'ajustement de données mais aussi d'indiquer à quel niveau on a des ruptures, dont la signification diffère selon l'application. En se basant sur ce modèle hybride et sur l'algorithme de la détermination non-supervisée du seuil, nous présentons dans la dernière partie de ce chapitre, différents modèles probabilistes multi-composantes développés pour différentes applications.

### 3.2 La Théorie des Valeurs Extrêmes

Commençons ce dernier chapitre par introduire l'univers probabiliste dans lequel nous allons baigner. Il s'agit d'une théorie qui vient compléter la théorie statistique classique où l'on s'intéresse à étudier le comportement moyen d'un échantillon de v.a. et où le TCL [Feller, 1966] joue un rôle primordial. Lorsque l'échantillon de v.a. étudié contient des observations de grandes valeurs, le TCL ainsi que les outils statistiques traditionnels ne sont plus adaptés et présentent peu d'intérêt pour l'étude du comportement extrême de cet échantillon. C'est ainsi qu'intervient la TVE que nous présentons dans ce chapitre. Bâtie autour d'une analogie avec le TCL, la TVE offre un terrain favorable pour l'étude et la caractérisation du comportement des valeurs extrêmes ou encore des quantiles extrêmes. D'après les définitions 3.3 et 3.4, que nous présentons par la suite, les quantiles extrêmes se situent dans la queue de la distribution. C'est à ces événements rares que nous nous intéressons et aux valeurs extrêmes associées que nous désirons modéliser. Deux approches ont été développées par la TVE dont les résultats sont d'importance majeure dans ce contexte. La première approche, dite classique, concerne la modélisation des du maximum ou du minimum d'une suite de v.a., par une loi limite appartenant à la famille des lois des extrêmes : Weibull, Gumbel et Fréchet et ce *via* le théorème de Fisher et Tippett 3.6 [Fisher and Tippett, 1928]. La seconde approche concerne la modélisation des excès au-delà d'un certain seuil, appelée en anglais POT, par une loi de Pareto généralisée *via* le théorème de Balkeman, de Haan et Pickands 3.10 [Pickands, 1975]. Dans cette thèse, nous utilisons la deuxième méthode du POT pour modéliser les valeurs extrêmes au-delà d'un seuil. Nous présentons cette méthode dans la section 3.2.1 tout en introduisant les deux théorèmes sur lesquels se base la TVE. Rappelons dans un premier temps quelques notions de base.

**Définition 3.1 (*n*-échantillon)** *On appelle  $n$ -échantillon d'un processus stochastique  $X$ , un échantillon de  $n$  v.a.  $(X_1, X_2, \dots, X_n)$  issues de  $X$ , indépendantes et de même loi que  $X$ . On dit alors que l'échantillon est indépendant et identiquement distribué et on le note iid.*

Nous notons que certains ouvrages définissent un  $n$ -échantillon dans un cadre plus général où les v.a. peuvent être dépendantes. Dans cette thèse, un  $n$ -échantillon est défini comme dans 3.1.

**Définition 3.2 (Statistiques d'ordre)** *Soit  $(X_1, X_2, \dots, X_n)$  un  $n$ -échantillon d'un processus stochastique continu  $X = (X_t)_{t \in \mathbb{R}}$ , défini sur l'espace probabilisé  $(\Omega, \mathcal{A}, \mathbb{P})$ . On appelle la  $i^{\text{ème}}$  statistique d'ordre de ce  $n$ -échantillon,  $1 \leq i \leq n$ , la statistique*

$$\begin{aligned} X_{(i)} : \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto X_{(i)}(\omega) = X_S(\omega)e_i \end{aligned}$$

où  $e_i$  désigne le  $i^{\text{ème}}$  vecteur de la base canonique de  $\mathbb{R}^n$  et  $X_S(\omega)$  est le vecteur d'observations  $[X_1(\omega), X_2(\omega), \dots, X_n(\omega)]$  trié dans l'ordre croissant.

La première statistique d'ordre est ainsi  $X_{(1)}(\omega) = \min([X_1(\omega), X_2(\omega), \dots, X_n(\omega)])$ , la  $n^{\text{ème}}$  statistique d'ordre  $X_{(n)}(\omega) = \max([X_1(\omega), X_2(\omega), \dots, X_n(\omega)])$  et on a

$$X_{(1)}(\omega) \leq X_{(2)}(\omega) \leq \dots \leq X_{(n)}(\omega), \quad \forall \omega \in \Omega.$$

**Définition 3.3 (Quantile d'ordre  $\alpha$ )** Soit  $F$  une fonction de répartition d'une certaine loi de probabilité. On définit le quantile d'ordre  $\alpha$ , noté  $q_\alpha$ , comme suit :

1. Si  $F$  est continue et strictement croissante, alors  $q_\alpha$  est donné par  $q_\alpha = F^{-1}(\alpha)$ , où  $F^{-1}$  est l'inverse de  $F$  ( $q_\alpha$  est la racine unique de l'équation  $F(x) = \alpha$ ).
2. Si  $F$  est la fonction de répartition d'une loi discrète ou que  $F$  est continue mais pas strictement croissante alors  $q_\alpha = \inf\{x : F(x) \geq \alpha\}$

**Définition 3.4 (Quantile extrême)** Soit  $(\alpha_n)_{n \in \mathbb{N}^*} \in ]0, 1[$  une suite croissante. Un quantile  $q_{\alpha_n}$ , d'ordre  $\alpha_n$  est dit extrême lorsque  $\alpha_n \rightarrow 1$  quand  $n \rightarrow \infty$ .

**Remarque 3.5** Dans certains ouvrages, les quantiles d'ordre  $\alpha$  sont définis en fonction de  $\bar{F} = 1 - F$  désignant la fonction de survie relative à la fonction de répartition  $F$ .

### 3.2.1 Méthode du Peak Over Threshold

Soit  $(X_n)_{n \in \mathbb{N}^*}$  une suite de v.a. iid issues de  $X \in (\Omega, \mathcal{A}, \mathbb{P})$  ou encore un  $n$ -échantillon de  $X$  de fonction de répartition  $F$ . Nous désignons par  $M_1 = X_1$ , et  $M_n = \max\{X_1, X_2, \dots, X_n\}$ , pour  $n \geq 2$ . Nous exprimons dans (3.1) la fonction de répartition de  $M_n$ .

$$\begin{aligned} \mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \mathbb{P}(X_2 \leq x) \dots \mathbb{P}(X_n \leq x) \\ &= F^n(x) \end{aligned} \tag{3.1}$$

Nous notons que ce résultat est obtenu grâce au fait que  $X_1, X_2, \dots, X_n$  sont des v.a. iid.

Nous regardons la distribution asymptotique de  $M_n$  en faisant tendre  $n$  vers l'infini, nous obtenons la loi limite dégénérée suivante :

$$\lim_{n \rightarrow \infty} F^n(x) = \begin{cases} 0, & \text{si } F(x) < 1; \\ 1, & \text{si } F(x) = 1. \end{cases} \tag{3.2}$$

Pour avoir une loi limite non dégénérée pour le maximum  $M_n$ , d'une façon analogue au TCL donnant la loi asymptotique de la moyenne normalisée de  $n$  v.a., l'idée de Fisher et Tippett en 1928 a été de normaliser aussi  $M_n$  en essayant une normalisation linéaire et ainsi de donner lieu à un résultat fondamental de la TVE qui est le théorème de Fisher-Tippett, énoncé ci-dessous.

**Théorème 3.6 ([Fisher and Tippett, 1928])** Soit  $(X_n)_{n \in \mathbb{N}^*}$  une suite de v.a. iid de fonction de répartition  $F$  et  $(M_n = \max\{X_1, X_2, \dots, X_n\})_{n \in \mathbb{N}^*}$ . S'il existe deux suites normalisantes  $(a_n)_{n \in \mathbb{N}^*} \in \mathbb{R}_+^*$  et  $(b_n)_{n \in \mathbb{N}^*} \in \mathbb{R}$  et une fonction de distribution limite non-dégénérée  $G$  tel que

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n \rightarrow \infty} F^n(a_n x + b_n) = G(x), \tag{3.3}$$

alors  $G(x) = G_{EV}(x; \xi)$  à une translation et changement d'échelle près, où  $G_{EV}$  est la fonction de répartition de la distribution généralisée des valeurs extrêmes définie dans 3.7.

On dit alors que  $F$  appartient au domaine d'attraction de  $G$ .

**Définition 3.7 (Distribution généralisée des valeurs extrêmes (GEV))** La distribution généralisée des valeurs extrêmes dite en anglais 'Generalized Extreme Value' (GEV) est définie par sa fonction de densité de probabilité  $g_{EV}$  et sa fonction de répartition  $G_{EV}$  données respectivement par (3.4) et (3.5).

$$g_{EV}(x; \xi) = u(x)^{\xi+1} \exp(-u(x)), \quad (3.4)$$

et

$$G_{EV}(x; \xi) = \exp(-u(x)), \quad (3.5)$$

où

$$u(x) = \begin{cases} (1 + \xi x)^{-\frac{1}{\xi}}, & \text{si } \xi \neq 0 \\ \exp(-x), & \text{si } \xi < 0 \end{cases}, \quad \forall x \in \mathcal{D}_{EV}(\xi) = \begin{cases} (-\infty, -\frac{1}{\xi}], & \text{si } \xi < 0 \\ \mathbb{R}, & \text{si } \xi = 0 \\ [-\frac{1}{\xi}, \infty), & \text{si } \xi > 0 \end{cases},$$

et  $\xi$  désigne le paramètre de forme ou encore l'indice de queue ou des valeurs extrêmes.

**Remarque 3.8** La distribution GEV, suivant la valeur de l'indice de queue  $\xi$ , correspond à l'une des lois extrêmes suivantes :

- la loi de Weibull si  $\xi < 0$  : distribution à queue droite bornée.
- la loi Gumbel si  $\xi = 0$  : distribution à queue droite légère et à décroissance exponentielle relativement rapide.
- la loi de Fréchet si  $\xi > 0$  : distribution à queue droite lourde et à décroissance lente.

Un exemple illustrant ces trois cas de figure est donné par la figure 3.1.

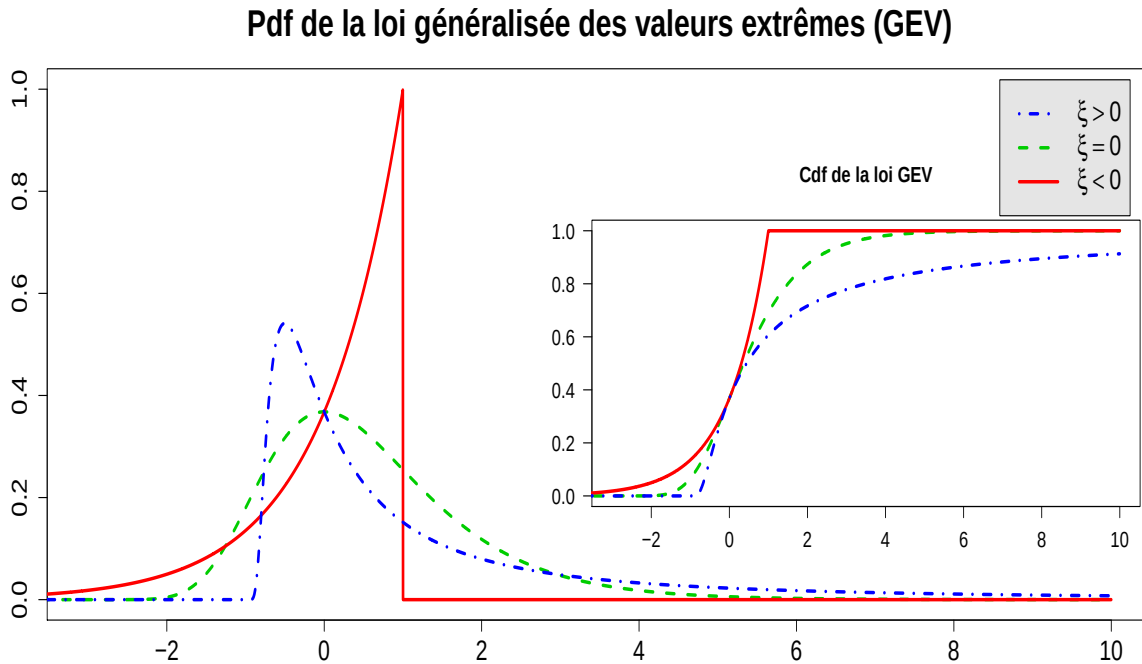


FIGURE 3.1 – Représentation de la cdf et de la pdf de la loi GEV pour des différentes valeurs de  $\xi$  :  $\xi = 1 > 0$  : loi de Fréchet,  $\xi = 0$  : loi de Gumbel, et  $\xi = -1 < 0$  : loi de Weibull.



Nous concluons que le théorème 3.6 permet de connaître la loi du maximum  $M_n$  indépendamment de la nature de la loi sous-jacente de fonction de répartition  $F$ . Pour en savoir plus, nous invitons le lecteur à se référer à [Embrechts et al., 1997].

Dans le cas où les v.a. sont dépendantes, Gumbel a introduit en 1958 la méthode du Block Maxima [Gumbel, 1958] afin de se mettre dans les conditions du théorème de 3.6. Cette méthode consiste en fait à diviser les données en plusieurs blocs de taille assez grande, générant ainsi des v.a. indépendantes, et ensuite de considérer le maximum de chaque bloc. Les maxima ainsi obtenus représenteront un échantillon de  $M_n$  dont la loi est approchée *via* le théorème 3.6 par l'une des lois des extrêmes : Weibull, Gumbel ou Fréchet. Cette méthode permet d'avoir une idée du comportement du maximum mais il est plus intéressant et pertinent de recueillir l'information à partir non pas d'un seul point mais de plusieurs extrêmes. C'est ce qui est proposé dans la méthode du dépassement du seuil ou la POT qui consiste à modéliser les excès au delà d'un certain seuil. Pour décrire cette méthode nous commençons par définir la fonction de répartition conditionnelle  $F_u$  des observations au-delà d'un seuil  $u$ . En effet, soit  $X_1, \dots, X_n$  un  $n$ -échantillon de fonction de répartition  $F$ . Considérons les  $N_u$  statistiques d'ordre (définition 3.2) dépassant  $u$ . Nous définissons alors de nouvelles v.a.  $Y_i = (X_i - u | X_i > u)$ ,  $i \in \{1, \dots, N_u\}$ , décrivant les excès au-delà de  $u < x_m$ , où  $x_m = \sup\{x \in \mathbb{R}_+, F(x) < 1\}$  dont les statistiques d'ordre supérieur sont  $Y_{(i)} = (X_{(n-N_u+i)} - u | X_{(n-N_u+i)} > u)$ ,  $i \in \{1, \dots, N_u\}$ . Nous cherchons alors à estimer  $F_u$ , la loi du processus  $Y$  qui décrit ces nouvelles *va*. En effet, pour tout  $y \in \mathbb{R}_+$ ,

$$\begin{aligned} F_u(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X - u \leq y | X > u) \\ &= \frac{\mathbb{P}(X \leq u + y, X > u)}{\mathbb{P}(X > u)} \\ &= \frac{F(u + y) - F(u)}{1 - F(u)}. \end{aligned} \quad (3.6)$$

Le deuxième théorème fondamental de la TVE, celui de Balkema, de Haan [Balkema and de Haan, 1974] et Pickands [Pickands, 1975], nous dit que la fonction de répartition conditionnelle  $F_u$  exprimée dans (3.6) peut être approchée par la loi de Pareto généralisée définie dans 3.9. Nous énonçons ce théorème dans 3.10.

**Définition 3.9 (Distribution de Pareto généralisée (GPD))** *La distribution de Pareto généralisée dite en anglais 'Generalized Pareto Distribution' (GPD) est définie par sa fonction de densité de probabilité  $g$  et sa fonction de répartition  $G$  données respectivement par (3.7) et (3.8).*

$$g(x; \xi, \beta) = \begin{cases} \frac{1}{\beta} (1 + \frac{\xi}{\beta} x)^{-1-\frac{1}{\xi}}, & \text{si } \xi \neq 0 \\ \frac{1}{\beta} \exp(-\frac{x}{\beta}), & \text{si } \xi = 0 \end{cases}, \quad \forall x \in \mathcal{D}(\xi, \beta), \quad (3.7)$$

et

$$G(x; \xi, \beta) = \begin{cases} 1 - (1 + \frac{\xi}{\beta} x)^{-\frac{1}{\xi}}, & \text{si } \xi \neq 0 \\ 1 - \exp(-\frac{x}{\beta}), & \text{si } \xi = 0 \end{cases}, \quad \forall x \in \mathcal{D}(\xi, \beta), \quad (3.8)$$

où  $\xi$  et  $\beta$  désignent respectivement l'indice de queue et le paramètre d'échelle et

$$\mathcal{D}(\xi, \beta) = \begin{cases} [0, \infty), & \text{si } \xi \geq 0 \\ [0, -\frac{\beta}{\xi}], & \text{si } \xi < 0 \end{cases}$$

**Théorème 3.10** ([Balkema and de Haan, 1974] et [Pickands, 1975]) Une fonction de répartition  $F$  appartient au domaine d'attraction de  $G_{EV}$  si et seulement si il existe  $\beta \in \mathbb{R}_+^*$  et  $\xi \in \mathbb{R}$  tel que la loi des excès au delà d'un seuil  $u$ , assez élevé, peut être approchée uniformément par une loi de Pareto généralisée i.e.

$$\lim_{u \rightarrow x_m} \sup_{0 \leq y \leq x_m - u} |F_u(y) - G(y; \xi, \beta)| = 0, \quad (3.9)$$

**Remarque 3.11** Le paramètre de forme  $\xi$  qui contrôle l'épaisseur de la queue de distribution est le même pour la GPD et la GEV. Il suit les mêmes règles définies dans la remarque 3.8. Nous représentons dans la figure 3.2 les pdf et les cdf de la GPD pour des différentes valeurs de  $\xi$  montrant la différence de l'épaisseur de la queue de distribution :  $\xi = -0.2$  : distribution à support borné,  $\xi = 0$  : distribution à queue légère, et  $\xi = 0.5$  : distribution à queue lourde.

**Pdf de la loi de Pareto généralisée (GPD)**

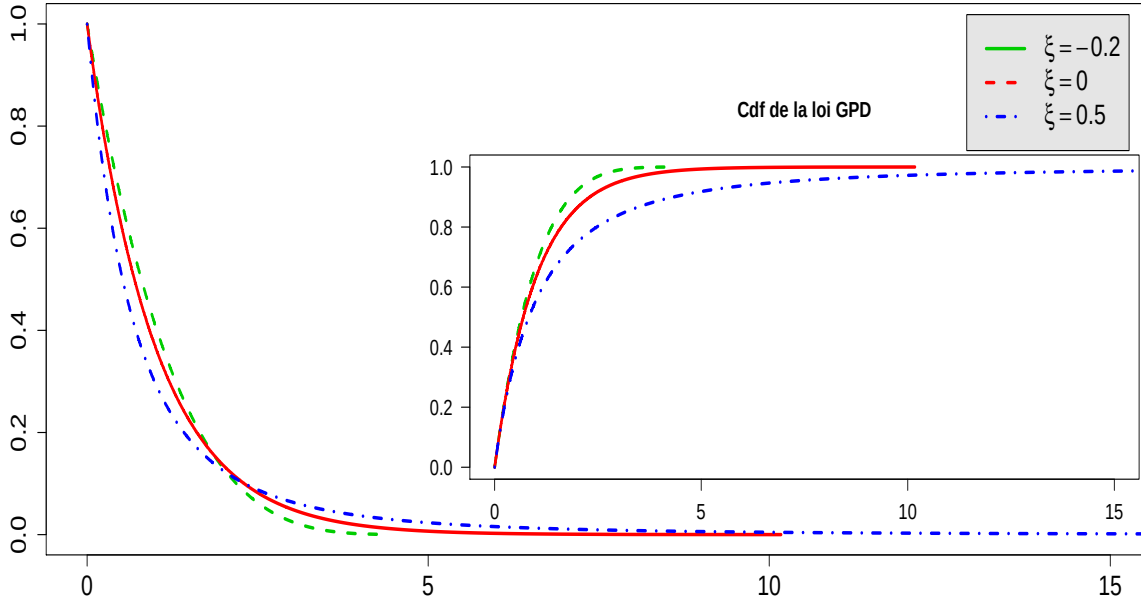


FIGURE 3.2 – Représentation de la cdf et de la pdf de la GPD pour  $u = 0$ ,  $\beta = 1$  et  $\xi \in \{-0.2, 0, 0.5\}$ .

Nous allons dans la section suivante mettre en pratique la méthode du POT pour modéliser les ruptures présentes dans les données que nous traitons. Pour ce faire, nous allons donner d'abord, dans la section 3.3, un résumé du travail [Debbabi et al., 2012] publié dans la conférence internationale *20<sup>th</sup> European Signal Processing Conference (EUSIPCO)* à ce propos. Ensuite, nous présentons le papier [Debbabi et al., 2012].

Nous mentionnons que nous avons préféré employer la POT dans cette thèse et non pas la Block Maxima (BM) pour éviter la perte d'information causée par cette dernière en ne considérant qu'un seul extrême, le maximum.

### 3.3 Modélisation stochastique des ruptures : résumé du papier [Debbabi et al., 2012]

Nous caractérisons, dans ce chapitre, l'apparition d'une rupture par un événement rare dont les observations associées sont de valeurs extrêmes. Nous rappelons que dans le chapitre 2, nous nous sommes intéressés seulement aux ruptures, caractérisées dans un cadre déterministe comme des discontinuités dans les signaux traités, sans considérer aucun modèle du bruit affectant ces dernières. Ensuite, pour mieux détecter les ruptures nous avons associé au bruit une modélisation stochastique. L'information que nous avons pu avoir en modélisant le bruit nous a permis de mieux séparer les fausses alarmes des ruptures.

Arrivés à ce stade, nous avons raisonné autrement : pourquoi s'intéresser au bruit (comportement moyen des données), alors que l'information que nous voulons extraire est située au niveau de la queue de distribution des données traitées (comportement extrême des données) ? D'où l'idée de modéliser les excès au-delà d'un certain seuil, contenant les ruptures, par une GPD via la méthode du POT de la TVE. En effet, tout au long de ce chapitre, un signal représenté par des données est modélisé par un  $n$ -échantillon  $(X_1, X_2, \dots, X_n)$  d'un processus continu  $X$  défini sur l'espace probabilisé  $(\Omega, \mathcal{A}, \mathbb{P})$ . Soit maintenant  $Y_i = (X_i - u | X_i > u)$ ,  $i \in \{1, \dots, N_u\}$ ,  $N_u$  étant le nombre d'excès au-delà du seuil  $u$ . D'après le théorème 3.9 la loi de probabilité de  $Y$  peut être approchée par une GPD d'indice de queue  $\xi$  et de paramètre de forme  $\beta$ .

La connaissance de la loi des extrêmes nous a incité à améliorer encore la détection de ruptures. Pour ce faire, comme pour la section 2.3.2 du chapitre 2, nous avons proposé de déterminer un seuil  $\eta$  permettant la détection de ruptures avec une pfa donnée  $p_\eta$ . En revanche, dans cette section nous définissons différemment une fausse alarme. En effet, nous modélisons le temps d'occurrence de ruptures par un processus de Poisson d'intensité  $\lambda$ . Par conséquent, le processus  $T_e$ , décrivant le temps écoulé entre deux ruptures successives, suit la loi exponentielle du même paramètre  $\lambda$ . Ainsi, nous définissons une fausse alarme par l'événement d'être dans la queue de distribution (valeurs extrêmes) et que le temps écoulé entre deux ruptures successives soit inférieure à  $t_{min} = \min\{t_i - t_{i-1}, i \geq 1\}$ , où  $t_i$  désigne le temps d'occurrence de la  $i^{\text{ème}}$  rupture. Autrement dit, nous voulons déterminer un seuil  $\eta$ , qui est déjà supérieur à  $u$ , tel que la probabilité de le dépasser et d'avoir un temps écoulé entre deux ruptures successives inférieur à  $t_m$  soit égale à  $p_\eta$ . Ceci peut être traduit mathématiquement comme suit :

$$p_\eta = \mathbb{P}(Y > \eta, T < t_{min}). \quad (3.10)$$

En supposant l'indépendance entre l'amplitude de la rupture et le temps de son occurrence (en général c'est le cas) l'équation (3.10) se simplifie en (3.11).

$$p_\eta = \mathbb{P}(Y > \eta) \mathbb{P}(T < t_{min}). \quad (3.11)$$

Sachant que nous connaissons la loi de  $Y$  et de  $T$  :  $Y$  suit la loi GPD et  $T$  suit la loi exponentielle, nous pouvons ainsi facilement déterminer le seuil  $\eta$  qui est donné par :

$$\eta = \frac{\hat{\beta}}{\hat{\xi}} \left[ \left( \frac{p_\eta}{1 - e^{-\hat{\lambda} t_{min}}} \right) - 1 \right], \quad (3.12)$$

où  $\hat{\beta}$ ,  $\hat{\xi}$  et  $\hat{\lambda}$  désignent respectivement les estimées de  $\beta$ ,  $\xi$ , et  $\lambda$ .

Cette méthode est bien développée et détaillée dans l'article [Debbabi et al., 2012]. Dans ce papier, que nous exposons dans ce qui suit, nous avons appliqué le seuil de (3.12) pour améliorer la détection de spikes de neurones, où  $X$  représente le signal neuronal traité par l'approche algébrique, et  $t_{min}$  est prise égale à la période réfractaire. Le paramètre de la loi exponentielle a été approché par l'inverse de la moyenne empirique, représentant l'inverse de l'estimateur de moment d'ordre 1 qui est l'estimateur le plus simple. C'est aussi l'estimateur de moments qui a été utilisé pour déterminer les paramètres de la loi GPD [Hosking and Wallis, 1987b]. Le seuil  $u$  satisfaisant le théorème 3.9 a été déterminé graphiquement par la méthode du *mean excess plot* [Demichel et al., 2011], [Ghosh and Resnick, 2010], qui est la méthode la plus utilisée dans la littérature. En effet, cette méthode consiste à tracer la courbe  $e(\alpha)$  de la moyenne des excès au-delà du seuil  $\alpha$ , dont l'expression est la suivante :

$$e(\alpha) = E(X - \alpha | X > \alpha), \quad \alpha \in \mathbb{R} \quad (3.13)$$

Lorsque  $X$  suit la loi GPD de paramètres  $\xi$  et  $\beta$ , l'équation (3.13) sera sous la forme :

$$e(\alpha) = \frac{\xi}{1 - \xi} \alpha + \frac{\beta}{1 - \xi} \quad (3.14)$$

Dans ce cas,  $e(\alpha)$  est une fonction linéaire en  $\alpha$ . Donc pour estimer le seuil  $u$ , nous traçons la fonction de moyenne des excès empirique et nous repérons la valeur de  $u$  à partir de laquelle la courbe devient linéaire. Cette méthode permet de bien déterminer le seuil  $u$  indiquant la présence des extrêmes ou encore à partir duquel les données suivent une GPD mais il faut mentionner que c'est une méthode graphique qui peut mettre du temps pour déterminer le seuil ce qui présente sa limitation.

Les résultats obtenus en appliquant le seuil (3.12) pour la détections des spikes de neurones ont été comparés avec ceux obtenus en utilisant la méthode du deuxième chapitre où la connaissance de la loi du bruit a été nécessaire pour sélectionner le seuil (voir section 2.3.2 du chapitre 2). Une amélioration remarquable en terme de détection de fausses alarmes a été obtenue.

# COMBINING ALGEBRAIC APPROACH WITH EXTREME VALUE THEORY FOR SPIKE DETECTION

Nehla DEBBABI<sup>1,2</sup>, Marie KRATZ<sup>3</sup>, Mamadou MBOUP<sup>2</sup> and Sadok EL ASMI<sup>1</sup>

<sup>1</sup> Research Lab. COSIM, Engineering College of Communications, University of Carthage

<sup>2</sup> CReSTIC, Université de Reims Champagne Ardenne

<sup>3</sup> ESSEC Business School Paris

{nehla.debbabi,elasmi}@supcom.rnu.tn, kratz@essec.edu, mamadou.mbou@univ-reims.fr

## ABSTRACT

This paper uses the Extreme Value Theory (EVT) for threshold selection in a previously proposed algebraic spike detection method. The algebraic method characterizes the occurrence of a spike by an irregularity in the neural signal and devises a nonlinear (Volterra) filter which enhances the presence of such irregularities. These appear as (positive) high amplitude pulses in the output signal. The pulses are isolated. We then interpret the occurrence of a spike as a rare and extreme event that we model in the framework of EVT. With this model, we derive an explicit expression of the decision threshold corresponding to a given probability of false-alarm. Simulation results show that the empirical probability of false alarm is close to the predicted one by applying the derived theoretical threshold.

**Index Terms**— Neural spike detection, algebraic approach, Extreme Value Theory, Generalized Pareto distribution, Mean Excess Plot.

## 1. INTRODUCTION

The neural signal consists of a sequence of action potentials (APs) *viz* spikes representing the electrical activity of the neurons. An AP represents the polarity inversion across the neuron membrane, it has an amplitude of 100 *mv* and a duration of about 2 to 3 *ms*. This duration includes a refractory period [1] measuring roughly 1 *ms* during which a second AP cannot be initiated. Spike detection is the first mandatory step in any information processing of neural recording. It is therefore an important task in neuroscience. When the signal is recorded extracellularly, it is inevitably corrupted by several kinds of noises. The most significant one originates from the activities of remote neurons with respect to the electrode. In many situations, this background noise appears with a low Signal to Noise Ratio (SNR) making the problem of spike detection challenging. The spike detection is a classical problem which has been developed by different methods. We refer to the Nonlinear Energy Operator (NEO) method [2], as well as the Wavelet Transform one [3]. We cite likewise the algebraic

approach detailed in [4], in which we are interested in this work. Using operational calculus and algebraic derivatives, this method defines a decision function for spike detection which corresponds to a Volterra filtering of the neural signal reducing the noise and highlighting spikes. A way to detect spikes is then devised by comparing the output signal with a threshold. The threshold is determined in [4] for a given probability of false alarm (pfa) after estimating the unknown cumulative distribution function (cdf) of the noise presented in the output of the Volterra filter. However, the obtained results using the determined threshold are not completely satisfactory. In order to obtain a more reliable detection, we propose in this paper a new spike decision threshold by combining the algebraic approach with the Extreme Value Theory (EVT) [5]. We characterize the spike occurrence as an extreme and rare event. Thereby, we transform the problem of spike detection into extreme value determination. Hence, we formalize a new way to determine the probability of false alarm by estimating the distribution of extremes, *i.e.*, spikes instead of estimating the unknown entire distribution of the decision function. Using the Pickands theorem [6], the extremes distribution is approximated by a Generalized Pareto Distribution (GPD). To carry out the performances of the proposed method, using simulated signals, we compare the obtained results to those achieved using the algebraic approach [4] for a fixed pfa. Our method improves the quality of the detection in terms of false alarm, providing more realistic results. The validation of the proposed method using generated signals proves its usefulness when applied to real data. The remainder of this paper is organized as follows. Section 2 develops the proposed threshold determination. Simulation results are discussed in Section 3. Finally, conclusions are drawn in section 4.

## 2. DESCRIPTION OF THE PROPOSED THRESHOLD

### 2.1. Algebraic approach

In this section we give a brief recall of the algebraic approach [4] to spike detection. A spike is the result of a sudden electrical discharge across the neural membrane. Its occurrence

is interpreted in [4] as an irregularity in the neural signal, say  $y(t)$ . We consider that the spikes are isolated so that there is at most one spike in each interval  $[\tau, \tau + T]$  of a given width  $T$ , representing, *e.g.*, the refractory period. Then by setting  $y_\tau(t) = y(\tau + t)$ ,  $t \in [0, T]$  and by ignoring the noise, we have  $\frac{d^n}{dt^n} y_\tau(t) = [y_\tau^{(n)}](t) + \sum_{k=1}^n \mu_{n-k} \delta(t - t_\tau)^{(k-1)}$

where  $[y_\tau^{(n)}]$  represents the regular part of the  $n^{th}$  signal derivative, the superscript  $(k)$  stands for differentiation of order  $k$ ,  $t_\tau$  is the possible irregularity point in  $[\tau, \tau + T]$ ,  $\mu_k = y_\tau^{(k)}(t_\tau+) - y_\tau^{(k)}(t_\tau-)$  and  $\delta$  is the Dirac delta function. Using the algebraic framework of [7], it has been shown in [4] that the change-point (spiking instant)  $t_\tau$  can be characterized by

$$a_0(\tau)t_\tau^2 - 2a_1(\tau)t_\tau + a_2(\tau) = 0, \quad (1)$$

where each coefficient  $a_i(\tau)$ ,  $i \in \{0, 1, 2\}$ , is in the form  $a_i(\tau) = \int_0^T p_i(t)y(t + \tau)dt$  for some polynomial  $p_i$ . A key point in [4] is that this representation is valid only when  $y$  has a change-point in  $[\tau, \tau + T]$ , otherwise the coefficients  $a_i(\tau)$  vanish (in the noise free case). This observation plus the presence of noise led the authors of [4] to consider the following statistical hypothesis testing problem

$$\mathcal{H}_0 : X(\tau) \leq \gamma \quad \text{vs} \quad \mathcal{H}_1 : X(\tau) > \gamma$$

where the decision function  $X(\tau)$  defined by the discriminant of (1), as in

$$X(\tau) = a_1(\tau)^2 - a_0(\tau)a_2(\tau). \quad (2)$$

Now, one can infer from the form of the coefficients  $a_i(\tau)$ , that  $X$  reads as a Volterra filtered version of the neural signal  $y$  (see [4] for more details). A spike occurrence is then meant when  $X(\tau)$  is above a given threshold  $\gamma$  (under the alternative hypothesis  $\mathcal{H}_1$ ). Henceforth, we model the decision function as a continuous process  $X = (X_t)_{t \in \mathbb{R}^+}$  on some probability space  $(\Omega, \mathcal{A}, \mathbb{P})$  and consider a  $n$ -sample  $(X_i)_{i=1, \dots, n}$ . To determine the threshold  $\gamma$  for a given pfa  $p$ , we need to determine the expression of the probability of false alarm. In [4], it is determined as

$$p = \mathbb{P}(X > \gamma \mid \mathcal{H}_0) \quad (3)$$

Since the statistical characteristics of  $X$  are unknown, the determination of the probability (3) becomes difficult. To overcome this issue, the algebraic approach proceeds as follows: knowing that the signal under the null hypothesis  $\mathcal{H}_0$  is reduced to noise, the probability (3) is calculated by estimating the noise cdf of the output Volterra filter. But as the determined probability is the noise probability, the threshold  $\gamma$  will depend only on the noise variance and not on signal characteristics. That causes unsatisfactory results. Hence, to improve the spike detection and to make it more reliable, we propose to combine the algebraic method with an EVT approach.

## 2.2. Threshold determination method for a fixed level of false alarm

In this paragraph, we present a new way to determine a decision threshold for a given pfa which does not require anymore to estimate the cdf of  $X$ . We consider the spike occurrence in the neural signal as a rare extreme event. Therefore, instead of estimating the entire distribution of  $X$ , we will restrict our attention to the estimation of the distribution of the extremes (spikes) above some high threshold  $u$  (to be determined), *i.e.* the tail distribution of  $X$ . So let  $Y_u$  define the process of exceedances of  $X$  over a given threshold  $u$

$$Y_u(j) = (X_j - u \mid X_j > u), \quad \text{for } 1 \leq j \leq N_u,$$

where  $N_u$  is the number of exceedances above  $u$ . We will first focus on the estimation of the distribution of  $Y_u$ , then we will define a decision criterion for the detection of a spike. The criterion will be related on the exceedance of  $Y_u$  over a second threshold  $\eta$ , which will be determined using the following modeling of the appearance of a false alarm

### False alarm modelling

We model the number of spike occurrences in a finite time interval of length  $t$  by a Poisson process  $\{N(t); t \geq 0\}$  with intensity  $\lambda$ . Therefore, the waiting time, say  $T$ , between two successive spikes obeys the exponential distribution with the same parameter  $\lambda$ . Let  $(T_i; 1 \leq i \leq k)$  be a  $k$ -sample. Unlike in [4], we choose here another way to define a false alarm. We denote by  $r_p$  the refractory period following each AP of a given neuron. Knowing that in this period no spike can be fired by the same neuron, a false alarm will be meant when the excess over  $u$  overtakes the threshold  $\eta$  when the condition  $\{T < r_p\}$  is realized. Hence, the probability of false alarm is defined as

$$p = \mathbb{P}(Y_u > \eta, T < r_p) \quad (4)$$

The amplitude of a spike is not related to its time occurrence. Therefore, we can assume the independence between  $Y_u$  and  $T$ . Thus (4) becomes

$$p = \mathbb{P}(Y_u > \eta) \mathbb{P}(T < r_p) \quad (5)$$

### GPD fit

To determine  $\eta$ , we need to estimate the distribution of  $Y_u$ . This distribution, named the excess distribution function, is given by

$$F_u(x) = \mathbb{P}(Y_u \leq x) = \mathbb{P}(X - u \leq x \mid X > u), \quad \forall x \geq 0.$$

Using the Pickands theorem [6],  $F_u$  can be approximated for a high threshold  $u$  by a Generalized Pareto Distribution (GPD)  $G_{\xi, \sigma}$  with shape parameter  $\xi$  and scale parameter  $\sigma$ :

$$F_u \approx G_{\xi, \sigma}$$

with

$$G_{\xi, \sigma}(x) = \begin{cases} 1 - (1 + \frac{\xi}{\sigma}x)^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ 1 - \exp(-\frac{x}{\sigma}), & \text{if } \xi = 0 \end{cases} \quad x \in \mathcal{D}(\xi, \sigma)$$

where

$$\mathcal{D}(\xi, \sigma) = \begin{cases} [0, \infty), & \text{if } \xi \geq 0 \\ [0, -\frac{\sigma}{\xi}], & \text{if } \xi < 0 \end{cases}$$

### Mean Excess Plot

It remains to identify a threshold  $u$  above which this approximation holds. For this purpose, we use a graphical tool called the Mean Excess Plot (MEP) [8, 9] which helps to find an appropriate  $u$  above which the exceedances follow a GPD. Recall that the mean excess function of a random variable  $X$  is defined by  $e(u) = E(X - u | X > u)$ . When  $X$  follows a GPD, the mean excess function is linear in  $u$  and expressed as

$$e(u) = \frac{\sigma}{1 + \xi} + \frac{\xi}{1 + \xi}u$$

Hence, to determine  $u$ , we draw the empirical mean excess function defined by

$$e_n(u) = \frac{\sum_{i=1}^n (X_i - u) \mathbb{I}_{(X_i > u)}}{\sum_{i=1}^n \mathbb{I}_{(X_i > u)}}, \quad u \geq 0 \quad (6)$$

and we select the value  $u$  from which the behavior of (6) is linear. Here  $\mathbb{I}_S$  denotes the indicator function of the set  $S$ .

### Parameters estimation

Once the threshold  $u$  is determined, the shape and the scale parameters of the approximating GPD are estimated, using for instance the moments method [10], and are given, respectively, by

$$\hat{\xi} = \frac{1}{2} \left( 1 - \frac{\overline{Y_u}^2}{S_{Y_u}^2} \right) \quad \text{and} \quad \hat{\sigma} = \overline{Y_u} \left( \frac{1}{2} + \frac{\overline{Y_u}^2}{S_{Y_u}^2} \right),$$

where  $\overline{Y_u}$  and  $S_{Y_u}^2$  are the sample mean and variance of the exceedances expressed as

$$\overline{Y_u} = \frac{1}{N_u} \sum_{i=1}^{N_u} Y_u(i) \quad \text{and} \quad S_{Y_u}^2 = \frac{1}{N_u - 1} \sum_{i=1}^{N_u} (Y_u(i) - \overline{Y_u})^2.$$

If the condition  $\xi < 1/4$  is satisfied, it can be shown by standard methods (see e.g. [11]) that  $\hat{\sigma}$  and  $\hat{\xi}$  are asymptotically normally distributed with covariance matrix  $A$  satisfying

$$N_u A \sim \Gamma = \frac{(1 - \xi)^2}{(1 - 2\xi)(1 - 3\xi)(1 - 4\xi)} (a_{ij})_{1 \leq i, j \leq 2},$$

$$\begin{aligned} \text{with } a_{11} &= 2\sigma^2(1 - 6\xi + 12\xi^2), \\ a_{22} &= (1 - 2\xi)^2(1 - \xi + 6\xi^2), \\ a_{12} &= a_{21} = \sigma(1 - 2\xi)(1 - 4\xi - 12\xi^2). \end{aligned}$$

In this case, a confidence interval with asymptotic confidence level  $\alpha$  can be deduced

$$L \leq \begin{pmatrix} \sigma \\ \xi \end{pmatrix} \leq U$$

where the lower and the upper bounds denoted by  $L$  and  $U$  respectively of the confidence interval are given by

$$\begin{aligned} L &= \begin{pmatrix} \hat{\sigma} \\ \hat{\xi} \end{pmatrix} + \left( \frac{1}{N_u} \Gamma \right)^{1/2} \begin{pmatrix} q((1 - \alpha)/2) \\ q((1 - \alpha)/2) \end{pmatrix} \\ U &= \begin{pmatrix} \hat{\sigma} \\ \hat{\xi} \end{pmatrix} + \left( \frac{1}{N_u} \Gamma \right)^{1/2} \begin{pmatrix} q((1 + \alpha)/2) \\ q((1 + \alpha)/2) \end{pmatrix} \end{aligned}$$

Here,  $q(\kappa)$  is the standard normal distribution quantile of order  $\kappa$ .

As well, we determine the parameter  $\lambda$  of the exponential distribution using the moments method. It is expressed as  $\hat{\lambda} = \overline{T}^{-1}$  with the following asymptotic confidence interval for a confidence level  $\alpha$

$$\left[ \overline{T} - \sqrt{\frac{S_T^2}{k}} q(\alpha/2) \right]^{-1} < \lambda < \left[ \overline{T} - \sqrt{\frac{S_T^2}{k}} q(1 - \alpha/2) \right]^{-1}$$

where  $\overline{T}$  and  $S_T^2 = \frac{1}{k-1} \sum_{i=1}^k (T_i - \overline{T})^2$  are respectively the sample mean and variance of  $T$ .

### Application for probability of false alarm determination

At this stage, we know that  $Y_u$  follows approximately a GPD with parameters  $\hat{\xi}$  and  $\hat{\sigma}$ , and that  $T$  obeys an exponential law with parameter  $\hat{\lambda}$ . Coming back to (5), we obtain

$$\begin{aligned} p &= (1 - G_{\hat{\xi}, \hat{\sigma}}(\eta)) \mathbb{P}(T < r_p) \\ &= \left( 1 + \frac{\hat{\xi}}{\hat{\sigma}} \eta \right)^{-\frac{1}{\hat{\xi}}} (1 - e^{-\hat{\lambda} r_p}) \end{aligned}$$

from which we deduce the threshold  $\eta$ , expressed as

$$\eta = \frac{\hat{\sigma}}{\hat{\xi}} \left[ \left( \frac{p}{1 - e^{-\hat{\lambda} r_p}} \right)^{-\hat{\xi}} - 1 \right]$$

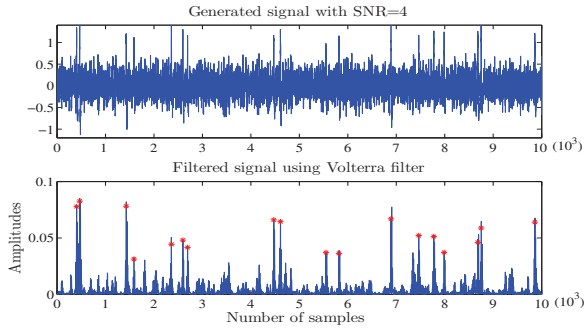
Recall that  $Y_u$  models the exceedances of  $X$  above  $u$ . Thus the threshold  $\eta$  is obtained relatively to  $u$  and not to zero. Consequently, the final threshold  $\mathcal{T}$  above which the false alarms appear with a probability  $p$  is given by

$$\mathcal{T} = \eta + u \quad (7)$$



### 3. APPLICATION OF THE PROPOSED THRESHOLD ON SIMULATED SIGNAL

To rigorously test the performance of the proposed threshold, we carry out simulations on synthesized signal. According to the AP morphology, we simulate a spike template. Using this template, we generate a spike train with maximal amplitude equal to  $M$ , describing a Poisson process and ensuring a refractory period of 2 ms as a duration after each synthesized AP. To make the simulation more realistic, we corrupt the generated signal by a noise with a fixed SNR defined as:  $SNR = \frac{M}{\sigma}$ , where  $\sigma$  is the standard deviation of the background noise. The used noise is extracted from real neural recording with high SNR [12]. Fig.1. illustrates the Volterra filtering of a generated neural signal with SNR=4. As it is noticed in this figure, using the algebraic approach, real spikes (red \*) are successfully detected. But due to noise effect, an important number of false alarms are detected as well. Hence, the need of decision threshold determination.

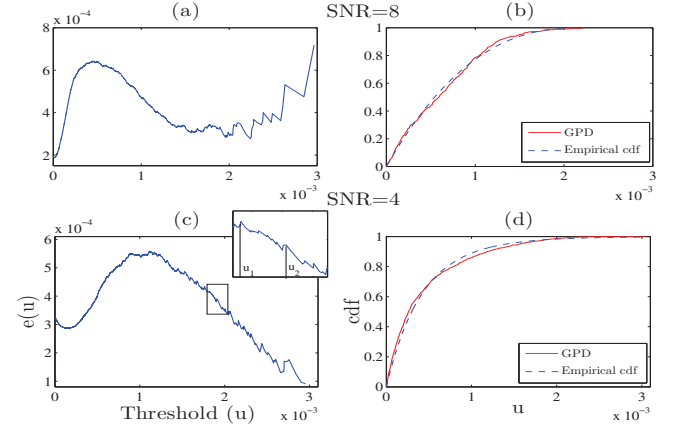


**Fig. 1.** Upper graph: generated neural signal with SNR=4. Lower graph: corresponding filtered signal using the algebraic approach where the real spikes are marked by red stars

#### 3.1. MEP for GPD estimation for a certain threshold $u$

Fig. 2. depicts the use of the MEP method for threshold determination using two generated neural signals with different SNR. For SNR=4, the graph (c) shows the MEP where a zoomed part is given in the right top. Through this zoom we can notice that the MEP admits an oscillatory curve containing several linear behaviors. This implies that the choice of the threshold is not unique: it can be  $u_1$  or  $u_2$  or some other threshold. To ensure the best fit of GPD we choose the threshold that minimizes the distance  $d$  between the GPD and the empirical cdf of exceedances according to the infinity norm. Using the optimal threshold, the plot of these two distributions are shown in the graph (d) where we can see that the GPD seems to fit pretty well the empirical distribution. The analogous results for SNR=8 are presented in (a) and (b). Some results of the threshold choice in terms of distance be-

tween the estimated GPD and the corresponding empirical cdf are illustrated in the Table 1, where the optimal ones are given in the gray boxes.



**Fig. 2.** MEP for GPD estimation

**Table 1.** Threshold determination

| Thresholds                         | $\alpha$ s.t. $u = q_n(\alpha)$ | $N_u$ | $\ d\ _\infty$ |
|------------------------------------|---------------------------------|-------|----------------|
| <b>Generated signal with SNR=8</b> |                                 |       |                |
| $2.4 \cdot 10^{-4}$                | 84, 34%                         | 1292  | 0.0832         |
| $4.94 \cdot 10^{-4}$               | 90.86%                          | 753   | 0.0372         |
| $7.57 \cdot 10^{-4}$               | 93.50%                          | 536   | 0.0542         |
| <b>Generated signal with SNR=4</b> |                                 |       |                |
| $5.27 \cdot 10^{-4}$               | 84.77%                          | 1330  | 0.058          |
| $5.82 \cdot 10^{-4}$               | 87.49%                          | 1093  | 0.0479         |
| $7.57 \cdot 10^{-4}$               | 92.21%                          | 680   | 0.06           |

The estimates of the parameters of the GPD and the exponential distributions are given with their confidence intervals (CI), for both signals, in Table 2.

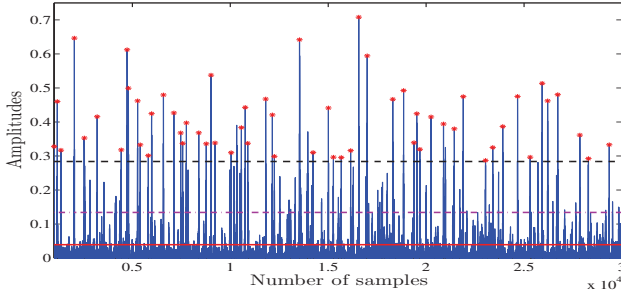
**Table 2.** Simulation parameters of GPD and exponential distribution

| Parameters                         | Generated neural signals     |                              |
|------------------------------------|------------------------------|------------------------------|
|                                    | SNR=8                        | SNR=4                        |
| Samples                            | 8247                         | 8738                         |
| $u$                                | $4.943 \cdot 10^{-4}$        | $5.823 \cdot 10^{-4}$        |
| $\alpha$ s.t.<br>$u = q_n(\alpha)$ | 90.86%                       | 87.49%                       |
| $N_u$                              | 753                          | 1093                         |
| $\hat{\xi}$                        | -0.406                       | 0.075                        |
| $CI(\xi)$                          | $[-0.409, -0.403]$           | $[0.073, 0.077]$             |
| $\hat{\sigma}$                     | $9 \cdot 10^{-4}$            | $4.17 \cdot 10^{-4}$         |
| $CI(\sigma)$                       | $[8.65, 9.35] \cdot 10^{-4}$ | $[3.77, 4.57] \cdot 10^{-4}$ |
| $\hat{\lambda}$                    | 0.0037                       | 0.0054                       |
| $CI(\lambda)$                      | $[0.0035, 0.004]$            | $[0.0051, 0.0057]$           |

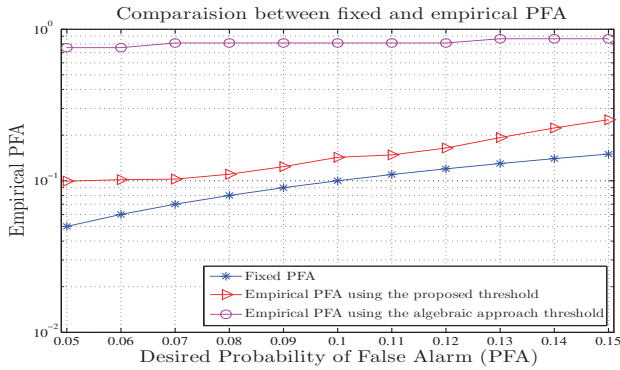


### 3.2. Interpretation and comparison

Using a generated neural signal with SNR=8, and for  $p = 0.1$ , we show in Fig. 3 the improvement of spike detection results using the proposed threshold. Assuming that the neural noise is Gaussian distributed, one can show [4] that  $X$  in (2) may be approximated by the sum of two  $\chi^2$  variables. Using this approximation, we deduce from (3), the threshold  $\gamma$ . The horizontal red solid line represents this value of  $\gamma$ . Next, we compute the empirical pfa, denoted by  $\text{epfa}_\gamma$ , according to  $\gamma$ . We obtain  $\text{epfa}_\gamma = 0.811$ : This is much higher than the prescribed value  $p = 0.1$ , showing that the assumed model of  $X$  under  $\mathcal{H}_0$  is not adequate. For the same prescribed pfa  $p = 0.1$ , the black dashed line represents the proposed threshold  $\mathcal{T}$  in (7). This threshold leads to a more realistic empirical pfa  $\text{epfa}_\mathcal{T} = 0.14$  which is very close to the fixed value.



**Fig. 3.** Spike detection using the proposed threshold ( black dashed line ) compared to the algebraic approach threshold (red solid line) for  $p = 0.1$ . The purple point dashed line represents the threshold obtained using the MEP



**Fig. 4.** Comparison between fixed pfa (blue \*) and empirical pfa using the proposed threshold (red  $\triangleright$ ) and the threshold described in [4] (purple  $\circ$ ).

To highlight furthermore the performance of the proposed threshold, we calculate  $\text{epfa}_\gamma$  and  $\text{epfa}_\mathcal{T}$  for different values of pfa:  $p$  varies from 0.05 to 0.15. Note that the maximum value of  $p$  is obtained when  $\eta = 0$ , i.e.  $\mathbb{P}(Y_u > \eta) = 1$ . Therefore, from equation (5) the maximum value of  $p$  is reached

for  $\mathbb{P}(T \leq r_p)$ . Simulation results are illustrated in the Fig. 4. We notice that the curve describing  $\text{epfa}_\mathcal{T}$  (red  $\triangleright$ ) is closer to the prescribed pfa (blue \*), while the difference between this latter and the curve of  $\text{epfa}_\gamma$  (purple  $\circ$ ) is obvious. This improvement is due to the fact that using the proposed threshold, there is no need to go through the estimation of the unknown distribution of the decision function  $X$ . However, based on EVT, we have determined the distribution of extreme values representing spikes which is known based on the Pickands theorem. There is a minor difference between real and empirical probability of false alarm of order of  $10^{-2}$  induced by the estimation of GPD and exponential parameters.

### 4. CONCLUSION

This paper develops a new spike decision threshold for the algebraic approach based on the extreme value theory. Applying the algebraic approach for detection and using the extreme value theory for making decision offers a robust and efficient spike detection. Comparing the obtained results with those achieved using the algebraic approach threshold, demonstrates the performance of the proposed threshold in terms of false alarm, giving results more realistic in accordance with what is predicted in theory.

### 5. REFERENCES

- [1] B. S. Gutkin, B. Ermentrout, and M. Rudolph, "Spike generating dynamics and the conditions of spike-time precision in cortical neurons," *J. of Comput. Neuroscience*, vol. 15(1), pp. 91–103, 2003.
- [2] K.H.Kim and S.J.Kim, "Neural spike sorting under nearly 0 db signal-to-noise ratio using nonlinear energy operator and artificial neural network classifier," *Biomedical Engineering, IEEE Transactions*, vol. 47, pp. 1406–1411, 10, Oct. 2000.
- [3] Z. Nenadic, "Spike detection using the continuous wavelet transform," *IEEE Trans. biomedical engineering*, vol. 52, no. 1, pp. 74–87, 2005.
- [4] Z. Tiganj M. Mboup, "A Volterra filter for neuronal spike detection, preprint version, "first version available on HAL INRIA at: <http://hal.inria.fr/docs/00/67/66/06/PDF/spike-HAL.pdf>.
- [5] T. Mikosch P. Embrechts, C. Klppelberg, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, 1997.
- [6] J. Pickands, "Statistical inference using extreme order statistics," *Annals of Statistics*, vol. 3, pp. 119–131, 1975.
- [7] M. Fliess and H. Sira-Ramírez, "An algebraic framework for linear identification," *ESAIM: COCV*, vol. 9, pp. 151–168, 2003.
- [8] S. Ghosh and S. Resnick, "A discussion on mean excess plots," *Stochastic Process and their Applications*, vol. 3, pp. 1492–1517, 2010.
- [9] Y. Demichel, A. Estrade, M. Kratz, and G. Samorodnitsky, "How fast can the chord-length distribution decay," *Advances in Applied Probability*, vol. 43, pp. 504–523, 2011.
- [10] J. Hosking and J. Wallis, "Parameter and quantile estimation for the generalized pareto distribution function," *Technometric*, vol. 29 (3), pp. 339–349, 1987.
- [11] Rao. C. R, *linear statistical inference and its applications*, New York: John Wiley, 1973.
- [12] Christophe Pouzat, *Methods and Models in Neurophysics. Les Houches 2003 Summer School.*, chapter Technique(s) for Spike - Sorting, pp. 729–786, Elsevier, 2005, Available from: <http://fr.arxiv.org/abs/q-bio.QM/0405012>.

### 3.4 Modélisation stochastique de toutes les données : ruptures et bruit

Pour aller plus loin dans la modélisation des données contenant des ruptures que nous traitons, nous allons dans une dernière étape considérer tout ce que nous avons développé jusque là. Nous allons considérer toutes les données : le bruit et les ruptures. Nous allons séparer ces deux composantes et modéliser chacune par une loi différente, en se basant sur deux théorèmes fondamentaux en probabilité. Nous modélisons ainsi les données par un modèle hybride défini par morceaux, où sur chaque morceau nous avons une loi différente. Ce modèle hybride est caractérisé par sa pdf  $h$ , donnée par (3.15), que nous définissons dans le cas général où le nombre de composantes  $n_c$  est strictement supérieur à 1. Soit

$$h(x; \theta) = \sum_{i=1}^{n_c} \gamma_i f_i(x; \theta_i) \mathbb{1}_{[x_{i-1}, x_i]}, \quad (3.15)$$

où chaque  $\gamma_i$ ,  $i = 1, \dots, n_c$ , représente le poids associé à la pdf  $f_i$ ,  $\{[x_{i-1}, x_i], i = 1, \dots, n_c\}$  est une partition de  $\mathbb{R}$  et  $\theta$  est le vecteur de paramètres du modèle contenant tous les paramètres  $\theta_i$  des pdfs  $f_i$ . Nous notons que pour ce modèle général, le nombre et la nature des composantes changent selon le type de données à modéliser [Furrer and Katz, 2008], [Mandava et al., 2011], [Kollu et al., 2012], [Scollnik, 2007] et [Knecht and Küttel, 2003].

Dans cette dernière partie de la thèse, nous allons en premier lieu considérer le modèle hybride à deux composantes où  $f_1$  et  $f_2$  désignent respectivement la pdf gaussienne et la pdf de la GPD. En effet, nous modélisons le bruit, représentant le comportement moyen des données par une gaussienne *via* le TCL, comme dans la section 2.3 du premier chapitre et modélisons les ruptures, représentant le comportement extrême, par une GPD *via* le théorème de Balkema, de Haan et Pickands [Pickands, 1975], comme dans la section 3.3 de ce chapitre. Ce modèle sera discuté dans la section 3.4.1, pour une pondération uniforme *i.e.*  $\gamma_1 = \gamma_2$ , qui va être sous forme d'un résumé du papier [Debbabi and Kratz, 2014], publié à la conférence internationale 2014 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

#### 3.4.1 Modèle hybride à deux composantes Gauss-GPD avec pondération uniforme : résumé du papier [Debbabi and Kratz, 2014]

Nous nous sommes intéressés dans l'étude [Debbabi and Kratz, 2014] à modéliser les données asymétriques à queues lourdes. Nous avons ainsi considéré les données où les ruptures n'apparaissent que dans la queue droite de la distribution que nous avons modélisé par le modèle hybride suivant :

$$h(x; \theta) = \begin{cases} \gamma f(x; \mu, \sigma), & \text{si } x \leq u \\ \gamma g(x - u; \xi, \beta), & \text{si } x > u \end{cases} \quad (3.16)$$

où  $f$  désigne la pdf gaussienne donnée par (3.17) de moyenne  $\mu \in \mathbb{R}$  et d'écart type  $\sigma \in \mathbb{R}_+^*$ .

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad \forall x \in \mathbb{R} \quad (3.17)$$

$g$  représente la pdf de la GPD donnée par (3.7) d'indice de queue  $\xi$  et de paramètre d'échelle  $\beta$ ,  $\theta = [u, \mu, \sigma]$ ,  $\gamma$  est le facteur de pondération uniforme et  $u$  désigne en même temps le point

de jonction des deux pdfs, le seuil indiquant la présence des extrêmes.

Comme dans [Carreau and Bengio, 2009b], nous avons imposé une condition de classe  $\mathcal{C}^1$  au point de jonction  $u$ . Cette condition permet non seulement d’avoir une pdf continue et lisse mais aussi de réduire le nombre de paramètres à estimer de 6 à 4. Un quatrième paramètre peut être aussi exprimé en fonction des 3 qui restent en imposant à  $h$  d’être une pdf *i.e.* d’intégrale unité. Donc, à la fin, nous nous trouvons avec seulement trois paramètres inconnus. C’est pour cela que  $\theta$  est composé de seulement 3 paramètres.

Un exemple de la pdf hybride de (3.16) est illustré dans la figure 3.3 pour  $\theta = [0.4354, 0, 1]$  où nous voyons bien la différence entre cette dernière et sa composante normale au niveau de la queue droite de la distribution.

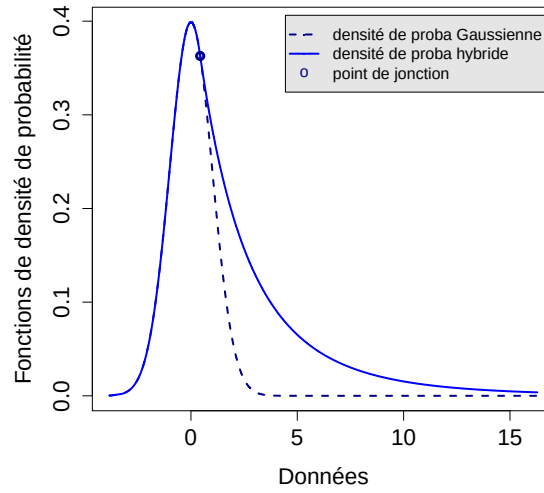


FIGURE 3.3 – Exemple de pdf hybride avec sa composante normale

Il faut mentionner à ce niveau que le modèle de Pareto hybride défini par (3.16) a été introduit par Carreau et Bengio dans [Carreau and Bengio, 2009b]. Les auteurs de ce travail ont utilisé la distribution qu’ils ont proposé dans un modèle non-paramétrique de mélange de lois et ce, dans le but de bien modéliser les données asymétriques à queues lourdes, ce que les lois de mélange classiques (gaussien par exemple) dépassent. Dans cette partie de la thèse, nous considérons ce modèle, nommé G-GPD dans [Debbabi and Kratz, 2014], pour une utilisation directe et non pas dans un mélange de lois.

Les avantages de ce modèle hybride sont multiples. En effet, il ne permet pas seulement de bien ajuster les données comme pour les méthodes classiques d’estimation de densité de probabilité telles que la méthode des noyaux [T. Buch-larsen and Bolancé, 2005] ou la méthode des mélanges de lois [Kollu et al., 2012], mais aussi de pouvoir les interpréter puisque au-delà de  $u$  nous avons les extrêmes indiquant les ruptures, qui, sans oublier, peuvent nous signaler différentes informations suivant le phénomène étudié. Aussi, la relation entre les paramètres du modèle nous permet d’exprimer explicitement le seuil  $u$  en fonction des autres paramètres du modèle. Ainsi, nous évitons les problèmes rencontrés lors de la détermination du seuil satisfaisant le théorème de Balkema, de Haan et Pickands avec les méthodes existantes dans

la littératures qui sont graphiques comme le cas par exemple de la méthode du *mean excess plot* que nous avons utilisé dans le papier [Debbabi et al., 2012] ou la méthode d'estimateur de Hill [Hill, 1975] ou encore la méthode QQ-estimateur [Kratz and Resnick, 1996]. Cette détermination graphique du seuil nécessite une intervention humaine ce qui met du temps et peut causer dans certain cas de sérieux problèmes comme par exemple pour le traitement des signaux biomédicaux où l'information doit être extraite en temps réel. Pour éviter ces problèmes, nous avons proposé, dans l'étude ci-dessous [Debbabi and Kratz, 2014], de déterminer ce seuil d'une manière non supervisée pour le modèle décrit dans (3.16).

Nous avons proposé dans [Debbabi and Kratz, 2014] un algorithme itératif qui commence par un choix d'un premier seuil initial et qui converge vers un point stationnaire qui est le seuil  $u$ . Chaque itération de l'algorithme est composée de deux étapes principales : la première étape consiste à estimer les paramètres de la loi gaussienne minimisant l'erreur quadratique entre la fonction de répartition hybride, admettant le seuil trouvé dans l'itération précédente, et la fonction de répartition empirique. Alors que la deuxième étape consiste à déterminer le nouveau seuil minimisant l'erreur quadratique entre la fonction de répartition hybride, admettant les paramètres de la gaussienne trouvés, et celle empirique. Pour chaque itération, les paramètres de la GPD et le facteur de pondération sont déterminés implicitement via les relations trouvées par les conditions imposées à la pdf  $h$ . L'algorithme proposé est rapide à converger. Il permet non seulement une détermination non supervisée du seuil au delà duquel les observations suivent une GPD mais aussi de bien estimer les paramètres de la loi normale ainsi que ceux de la GPD. En se basant sur les simulations de Monte Carlo et en utilisant la divergence de Kullback Liebler comme critère d'évaluation, on a montré que l'estimation des paramètres du modèle hybride en utilisant notre algorithme a amélioré les résultats obtenus en utilisant la méthode proposée dans [Carreau and Bengio, 2009b] qui est aussi automatique et qui consiste à estimer tous les paramètres du modèle hybride ensemble *via* l'estimateur du maximum de vraisemblance.

Un autre avantage du modèle hybride (3.16) et de l'algorithme proposé est qu'ils permettent d'estimer les paramètres de la loi GPD même lorsque  $\xi \geq 1$  ce qui n'est pas le cas avec les méthodes existantes d'estimation de paramètres de la GPD basées sur le calcul des moments d'ordre supérieur [Hosking and Wallis, 1987a].

Notons que pour [Debbabi and Kratz, 2014],  $h(x; u, \mu, \sigma, \xi, \beta)$  désigne bien  $h(x; \theta)$  de (3.16).

## A NEW UNSUPERVISED THRESHOLD DETERMINATION FOR HYBRID MODELS

Nehla DEBBABI<sup>1,2</sup> and Marie KRATZ<sup>3</sup><sup>1</sup> CReSTIC, Faculté des Sciences Exactes et Naturelles, Université de Reims Champagne Ardenne<sup>2</sup> Research Lab. COSIM, Engineering College of Communications of Tunis, University of Carthage<sup>3,\*</sup> ESSEC Business School, CREAR Risk research center

nehla.debbabi@supcom.tn, kratz@essec.edu

## ABSTRACT

A Gauss-GPD hybrid model that links a Gaussian distribution to a Generalized Pareto Distribution (GPD) is considered for asymmetric heavy tailed data. The paper proposes a new unsupervised iterative algorithm to find successively the junction point between the two distributions and to estimate the hybrid model parameters. Simulation results show that this method provides a reliable position for the junction point, as well as an accurate estimation of the GPD parameters, which improves results when compared with other methods. Another advantage of this approach is that it can be adapted to any hybrid model.

**Index Terms**— Heavy-tailed data modelling, Hybrid density estimation, Extreme Value Theory (EVT), Generalized Pareto distribution (GPD), Unsupervised algorithm.

## 1. INTRODUCTION

Modelling non-homogeneous and multi-component data is a problem that challenges scientific researchers in several fields [1, 2, 3, 4, 5, 6]. In general, it is not possible to find a simple and closed form probabilistic model to describe such data. Therefore, it seems natural to consider non-parametric approaches, such as *e.g.* kernel density estimation ones [7] or non-parametric Bayesian methods [8, 9] to name a few. However, when the multiple components are separable, parametric modelling becomes again tractable. Several hybrid models have been proposed in such context, combining two or more densities [1, 2, 3, 4, 5]. In this work, we are interested in the hybrid model that links a Gaussian distribution to a Generalized Pareto Distribution (GPD), noted G-GPD, to model asymmetric heavy tailed data, as *e.g.* in [3]. Why such a choice of modelling? Since we are interested in fitting the whole distribution underlying asymmetric heavy tailed data, the idea is to consider both the mean and tail behaviors, and to use limit theorems for each one, as suggested and developed analytically in [10]. On one hand, the Central Limit Theorem (CLT) justifies to introduce a Gaussian distribution for the mean behavior, on the other hand Pickand's

theorem [11] tells us that the tail of the distribution may be evaluated through a GPD above a high threshold, using the Peak Over Threshold (POT) method of the Extreme Value Theory (EVT). Hence, if the G-GPD modelling may appear restrictive at first glance, it is in fact quite general since it could apply to any underlying distribution.

This modelling might be useful in various fields, under the presence of asymmetric heavy tailed data, as for instance in (re)insurance when dealing with asymmetric heavy-tailed claims, or in finance [3, 12]. Other examples can be given in signal processing, when considering the spike detection in neural signals in biomedicine [13], or the energy detection for unknown signals over a fading channels in telecommunication [14], or bearing defect early detection in vibratory signals in machine diagnostics [15].

The main issue with this modelling is the determination of the junction point (or threshold) between the distributions, and the estimation of the different parameters. Much literature in EVT has been dedicated to the determination and estimation of the threshold above which the observations can be modeled by a GPD, via standard POT methods (mean excess plot method, Hill estimator, QQ-estimator; see *e.g.* [13, 16, 17, 18, 19, 20]). The difficulty faced when applying these methods is that they are graphical ad hoc approaches. The offline solution of those methods represents an important disadvantage especially when complexity burden and/or delay processing are critical. To overcome this difficulty, especially in practice, unsupervised methods have been developed (*e.g.* [2, 3]), in particular in [3] for the G-GPD model.

In this paper we propose a new unsupervised approach for the selection of the junction point, as well as for the estimation of the G-GPD parameters. It starts by enforcing the continuity and the differentiability of the G-GPD probability density function (pdf) at the junction point, as in [3], but then proceeds in an iterative way to determine successive thresholds and parameters of the two distributions that are involved. This approach does not depend on the choice of initial parameters; the algorithm starts with a selection of an arbitrary value of the threshold and stops, when it converges, to a fixed threshold with respect to some given error. We are going to see that it provides good results, improving the estimation of the

\* M. Kratz is also member of MAP5, UMR8145, Univ. Paris Descartes

model when compared with other methods, in particular with [3]. Note that another great benefit of our method is that it can be adapted to any other hybrid model, hence can resolve some non-homogenous data modeling problems.

The remainder of this paper is organized as follows. In section 2 we describe our iterative algorithm for the unsupervised threshold determination. Simulation results are presented in section 3. A comparison of our algorithm with the one presented in [3], as well as with standard methods of GPD parameters estimation, is discussed in the same section. Conclusions follow in the last section.

## 2. DESCRIPTION OF THE ITERATIVE APPROACH FOR UNSUPERVISED THRESHOLD SELECTION

Let us introduce the G-GPD hybrid model that we will consider, linking a Gaussian distribution and a GPD at a junction point denoted by  $u$ . It is defined by the pdf:

$$h(x; u, \mu, \sigma, \xi, \beta) = \begin{cases} \gamma f(x; \mu, \sigma), & \text{if } x \leq u \\ \gamma g(x - u; \xi, \beta), & \text{if } x > u \end{cases}$$

where  $\mu \in \mathbb{R}$  and  $\sigma \in \mathbb{R}_+$  are the mean and the standard deviation, respectively, of the Gaussian pdf  $f$  expressed as

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad \forall x \in \mathbb{R}$$

while  $\xi$  and  $\beta$  represent the tail index and the scale parameter respectively of the GPD pdf  $g$  defined by

$$g(x; \xi, \beta) = \begin{cases} \frac{1}{\beta} \left(1 + \frac{\xi}{\beta} x\right)^{-1-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right), & \text{if } \xi = 0 \end{cases} \quad \forall x \in \mathcal{D}(\xi, \beta)$$

where

$$\mathcal{D}(\xi, \beta) = \begin{cases} [0, \infty), & \text{if } \xi \geq 0 \\ [0, -\frac{\beta}{\xi}], & \text{if } \xi < 0 \end{cases}$$

The regulator factor  $\gamma$  ensures that the hybrid  $h$  is a pdf, i.e.

$$\int_{\mathbb{R}} h(x; u, \mu, \sigma, \xi, \beta) dx = 1.$$

It is expressed as a function of the Gaussian cumulative distribution function (cdf)  $F$ , with parameters  $\mu$  and  $\sigma$ , evaluated at  $u$  as

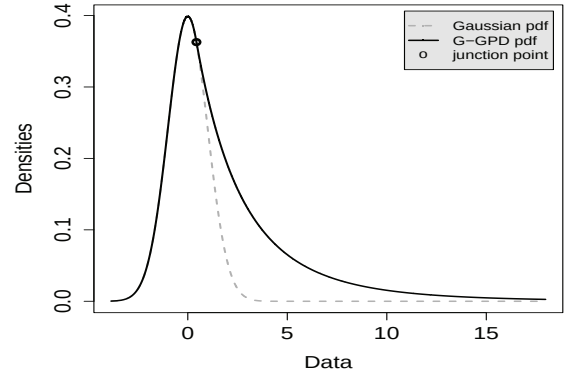
$$\gamma = \frac{1}{F(u; \mu, \sigma) + 1}.$$

The corresponding G-GPD cdf  $H$  is then expressed as

$$H(x; u, \mu, \sigma, \xi, \beta) = \begin{cases} \gamma F(x; \mu, \sigma), & \text{if } x \leq u \\ \gamma (F(u; \mu, \sigma) + G(x - u; \xi, \beta)), & \text{if } x > u \end{cases}$$

where  $G$  denotes the GPD cdf with parameters  $\xi$  and  $\beta$ .

In order to obtain a smooth pdf, the continuity and the differentiability of the G-GPD pdf at  $u$  are enforced, as in e.g.



**Fig. 1.** Gaussian pdf (dotted curve) with parameters  $\mu = 0$  and  $\sigma = 1$  and G-GPD pdf (continuous curve) with parameters  $\mu = 0, \sigma = 1, u = 0.4354, \xi = 0.2$  and  $\beta = 2.7558$ .

[3] i.e.  $f(u; \mu, \sigma) = g(0; \xi, \beta)$  and  $f'(u; \mu, \sigma) = g'(0; \xi, \beta)$ . From these relations a deterministic  $\xi$  and  $\beta$  are deduced as functions of the three remaining parameters  $u, \mu$  and  $\sigma$ , namely as

$$\beta_u = f(u; \mu, \sigma)^{-1}. \quad (1)$$

$$\xi_u = \beta \frac{(u - \mu)}{\sigma^2} - 1. \quad (2)$$

An example of the G-GPD pdf is illustrated in Fig. 1, where the difference between this latter and the Gaussian distribution is clearly noticeable.

The main idea of our method, after enforcing the continuity and the differentiability of the G-GPD pdf at the junction point, is to proceed in an iterative way. Each iteration is composed of two main steps. The first one consists in fitting the Gaussian distribution, minimizing the Squared Error (SE) between the G-GPD cdf obtained when considering the threshold determined in the previous iteration, and the empirical G-GPD one, using the Levenberg Marquardt algorithm [21, 22]. Once the Gaussian parameters  $\mu$  and  $\sigma$  are determined, the second step aims at determining the next threshold by minimizing the SE between the G-GPD cdf, admitting these Gaussian parameters, and the empirical G-GPD one, using again the Levenberg Marquardt algorithm. Note that the GPD parameters are implicitly determined using the equations given for the junction point.

### 2.1. Construction of our iterative algorithm

First, we consider an  $n$ -sample  $X = (x_i)_{1 \leq i \leq n}$  with a G-GPD as a parent distribution.

**Step 1:** We start our algorithm by determining the empirical



G-GPD cdf  $H_n$  of the n-sample  $X$  defined by

$$H_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{x_i \leq x\}}, \quad \forall x \in \mathbb{R},$$

where 1 denotes the indicator function.

**Step 2:** We choose an arbitrary initial threshold equal to a certain quantile of  $X$ , say  $u_0$ , e.g.  $u_0 = q_{50\%}$ , the 50% quantile. Then, we evaluate the mean  $\mu_0$  and the standard deviation  $\sigma_0$  of the Gaussian distribution of the first iteration by minimizing the SE between the G-GPD cdf  $H$ , when considering the threshold  $u_0$ , and the empirical one  $H_n$  using the Levenberg-Marquardt method [21, 22]. As a matter of fact, we chose to use this method since it is a robust one for nonlinear system resolution based on a minimization in the least squares sense. Hence, the estimates  $\hat{\mu}_0$  and  $\hat{\sigma}_0$  of the parameters of the Gaussian distribution of the first iteration are obtained as a solution of the minimization problem:

$$(\hat{\mu}_0, \hat{\sigma}_0) = \underset{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+}{\operatorname{argmin}} \|H(X; u_0, \mu, \sigma, \xi_{u_0}, \beta_{u_0}) - H_n(X)\|_2^2,$$

where  $\xi_{u_0}$  and  $\beta_{u_0}$  are the GPD parameters obtained using equations (1) and (2) and  $\|\cdot\|_2$  denotes the Euclidean norm.

**Step 3:** In the previous step we fitted the Gaussian part of the distribution, in this step we will fit the GPD one. To do so, we keep the threshold  $\hat{u}_1$  that minimizes the SE between the G-GPD cdf, using the estimated Gaussian parameters  $\hat{\mu}_0$  and  $\hat{\sigma}_0$ , and the empirical one using as well the Levenberg-Marquardt method. In other words,  $\hat{u}_1$  is the solution of the minimization problem:

$$\hat{u}_1 = \underset{u \geq 0}{\operatorname{argmin}} \|H(X; u, \hat{\mu}_0, \hat{\sigma}_0, \xi_u, \beta_u) - H_n(X)\|_2^2,$$

where  $\forall u \geq 0$ ,  $\xi_u$  and  $\beta_u$  are the GPD parameters relative to  $u$ , obtained using equations (1) and (2).

**Step 4:** At this stage, we proceed iteratively to fit the normal distribution and the GPD, simultaneously like we did in step 2 and 3. It means, for each iteration, the estimated Gaussian parameters are used to find the junction point while this latter will be used to evaluate the normal parameters of the next iteration, and so on. This can be written mathematically as follows:  $\forall k \geq 1$

- $(\hat{\mu}_k, \hat{\sigma}_k) \leftarrow \underset{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+}{\operatorname{argmin}} \|H(X; \hat{u}_k, \mu, \sigma, \xi_{\hat{u}_k}, \beta_{\hat{u}_k}) - H_n(X)\|_2^2$
- $\hat{u}_{k+1} = \underset{u \geq 0}{\operatorname{argmin}} \|H(X; u, \hat{\mu}_k, \hat{\sigma}_k, \xi_u, \beta_u) - H_n(X)\|_2^2$
- The algorithm stops when the following condition is satisfied

$$\|u_{k+1} - u_k\|_2^2 < \epsilon$$

with  $\epsilon > 0$ , small enough.

Hereafter, we present the pseudocode associated to this algorithm.

## 2.2. Pseudocode of the proposed algorithm

**Algorithm 1** Iterative algorithm for unsupervised threshold determination and G-GPD parameters estimation

Initialization of  $u_0, \epsilon$

- 1: Determination of the empirical G-GPD cdf  $H_n$  of the n-sample  $X = (x_i)_{1 \leq i \leq n}$
- 2: Estimation of  $\hat{\mu}_0$  and  $\hat{\sigma}_0$  when considering  $u_0$ ;

$$(\hat{\mu}_0, \hat{\sigma}_0) \leftarrow \underset{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+}{\operatorname{argmin}} \|H(X; u_0, \mu, \sigma, \xi_{u_0}, \beta_{u_0}) - H_n(X)\|_2^2$$

$(\beta_{u_0}, \xi_{u_0})$  are satisfying (1) and (2).

- 3: Determination of the first threshold  $\hat{u}_1$ , using the fixed Gaussian parameters  $(\hat{\mu}_0, \hat{\sigma}_0)$ ;

$$\hat{u}_1 \leftarrow \underset{u \geq 0}{\operatorname{argmin}} \|H(X; u, \hat{\mu}_0, \hat{\sigma}_0, \xi_u, \beta_u) - H_n(X)\|_2^2$$

- 4: Iterative process;

$k \leftarrow 1$

while  $\|\hat{u}_k - \hat{u}_{k-1}\|_2^2 \geq \epsilon$

$$(\hat{\mu}_k, \hat{\sigma}_k) \leftarrow \underset{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+}{\operatorname{argmin}} \|H(X; \hat{u}_k, \mu, \sigma, \xi_{\hat{u}_k}, \beta_{\hat{u}_k}) - H_n(X)\|_2^2$$

$$\hat{u}_{k+1} \leftarrow \underset{u \geq 0}{\operatorname{argmin}} \|H(X; u, \hat{\mu}_k, \hat{\sigma}_k, \xi_u, \beta_u) - H_n(X)\|_2^2$$

$k \leftarrow k + 1$

end while

return  $(u_{k+1}, \mu_k, \sigma_k, \xi_{u_{k+1}}, \beta_{u_{k+1}})$

## 3. SIMULATION RESULTS AND DISCUSSION

To assess the performance of our algorithm in terms of accurate estimation of the G-GPD model, we build on Monte Carlo simulations. In order to do this, we simulate 100 data sets of length  $n = 10^5$ , denoted by  $X^k = (x_i^k)_{1 \leq i \leq n}$ ,  $\forall 1 \leq k \leq 100$ , following the G-GPD distribution for a given fixed parameters  $\theta = (u, \mu, \sigma, \xi_u, \beta_u)$ . For each data set, we determine the estimate  $\hat{\theta}^k = (\hat{u}^k, \hat{\mu}^k, \hat{\sigma}^k, \xi_{\hat{u}^k}, \beta_{\hat{u}^k})$  of  $\theta$  using our algorithm. Then, we calculate the Kullback-Leibler (K-L) divergence, measuring the difference between the G-GPD pdf with parameter  $\theta$  and the corresponding estimated one. It is defined as follows:

$$D_{KL}(h_k \| \hat{h}_k) = \frac{1}{n} \sum_{i=1}^n h_k(x_i^k) \log \left( \frac{h_k(x_i^k)}{\hat{h}_k(x_i^k)} \right),$$

where  $h_k(X^k) = h(X^k; u, \mu, \sigma, \xi_u, \beta_u)$

and  $\hat{h}_k(X^k) = h(X^k; \hat{u}^k, \hat{\mu}^k, \hat{\sigma}^k, \xi_{\hat{u}^k}, \beta_{\hat{u}^k})$ ,  $\forall 1 \leq k \leq 100$ .

The closest to 0 the K-L divergence is, the most accurate evaluation of the G-GPD model we obtain. To test the reliability of the G-GPD model parameters estimation, we choose the average of  $D_{KL}(h_k \parallel \hat{h}_k)$ ,  $1 \leq k \leq 100$ , as evaluation criterion. Table 1 gives the average K-L divergence obtained for three different examples of simulated data sets. The first G-GPD model is simulated with  $\xi = 0.2$ ,  $\mu = 0$  and  $\sigma = 1$ , the second one with  $\xi = 1$ ,  $\mu = 2$  and  $\sigma = 1$  while the third with  $\xi = 0.5$ ,  $\mu = 1$  and  $\sigma = 1$ . We mention that for each case the exact values of  $\beta$  and  $u$  are determined using the obtained relations between the model parameters at the junction point. Table 1 exhibits the performance of our algorithm, the average K-L divergence being less than  $10^{-3}$ . Comparing the results obtained with our method and Method [3], we see that the G-GPD model evaluation is enhanced using our algorithm.

Note that, based on simulations, our algorithm converges regardless the choice of the initial value of threshold. It is obvious that the convergence will be slower if the chosen value of initial threshold is far from the real junction point. For the simulations, we chose  $u_0 = H_n^{-1}(0.5)$  and  $\epsilon = 10^{-12}$  as initial parameters where  $H_n^{-1}$  denotes the inverse empirical cdf.

Besides the reliable determination of the threshold, our algorithm shows a good performance in terms of the GPD parameters estimation. To highlight this point, we compare the obtained results using our method with standard statistical methods [23, 24]. Several methods in literature have been developed for the estimation of the GPD parameters, among which the MOments Method (MOM) based on the use of the first and the second moment of the GPD. Knowing that the GPD distribution admits a finite moment of order  $k$  if and only if the tail index  $\xi < \frac{1}{k}$ , the performance of the MOM decreases when  $\xi \geq \frac{1}{2}$ . The Probability Weighted Moments (PWM) method, has then been proposed in [23] extending the MOM when  $\xi < 1$ , but the estimation of the GPD parameters is still unreliable for  $\xi \geq 1$ . Another statistical method, well used in literature, is the Maximum Likelihood Method (MLM) [23]. The limitation of this method is that it can returns a local maximum instead of the global one. To overcome the encountered problem, Zhang has proposed a specific method in [24] for any value of  $\xi$  based on a Bayesian analysis. Nevertheless, it remains an ad-hoc method.

We show in Table 2 that our method for G-GPD parameters estimation goes over those encountered problems. Indeed, when revisiting Examples 2 and 3, we can compare in Table 2 the mean value of the estimated GPD tail parameter  $\xi_u^k$ ,  $\forall 1 \leq k \leq 100$ , as well as the mean value of the GPD shape parameter  $\beta_u^k$ ,  $\forall 1 \leq k \leq 100$ , when using our algorithm and existing methods. We can easily infer the reliable estimation of the GPD parameters obtained via our algorithm, when computing the mean squared error (mse) between the exact and the estimated parameters via the different methods.

|            | Example 1              | Example 2              | Example 3              |
|------------|------------------------|------------------------|------------------------|
| Method [3] | $1.6664 \cdot 10^{-4}$ | $8.9004 \cdot 10^{-5}$ | $5.3537 \cdot 10^{-5}$ |
| Our meth.  | $1.6591 \cdot 10^{-4}$ | $4.3661 \cdot 10^{-5}$ | $3.2163 \cdot 10^{-5}$ |

**Table 1.** Comparison of the average K-L divergence obtained using our method and the one presented in [3] for three different examples: Example 1 for  $\xi = 0.2$ ,  $\mu = 0$  and  $\sigma = 1$ , Example 2 for  $\xi = 1$ ,  $\mu = 2$  and  $\sigma = 1$  while the third with  $\xi = 0.5$ ,  $\mu = 1$  and  $\sigma = 1$ .

|                                                                                                           | Our meth.            | MOM     | PWM     | MLM     | Zhang   |
|-----------------------------------------------------------------------------------------------------------|----------------------|---------|---------|---------|---------|
| <b>Generated data sets with <math>\xi = 1</math>, <math>\mu = 2</math>, and <math>\sigma = 1</math></b>   |                      |         |         |         |         |
| $\xi = 1$                                                                                                 | 0.9992               | 0.4992  | 0.835   | 0.5007  | 0.5007  |
| mse( $\xi$ ) ( $10^{-3}$ )                                                                                | 0.2347               | 250.74  | 30.2165 | 249.327 | 249.313 |
| $\beta = 3.0905$                                                                                          | 3.0917               | 25.0006 | 6.6813  | 8.2346  | 8.2347  |
| mse( $\beta$ )                                                                                            | $0.24 \cdot 10^{-3}$ | 877.28  | 12.991  | 26.464  | 26.465  |
| <b>Generated data sets with <math>\xi = 0.5</math>, <math>\mu = 1</math>, and <math>\sigma = 1</math></b> |                      |         |         |         |         |
| $\xi = 0.5$                                                                                               | 0.4991               | 0.4191  | 0.105   | 0.1876  | 0.1877  |
| mse( $\xi$ ) ( $10^{-3}$ )                                                                                | 0.1346               | 7.5503  | 156.55  | 97.609  | 97.532  |
| $\beta = 2.8727$                                                                                          | 2.8735               | 4.2269  | 5.5153  | 5.6406  | 5.636   |
| mse( $\beta$ )                                                                                            | $0.16 \cdot 10^{-3}$ | 1.8692  | 13.275  | 7.6647  | 7.6367  |

**Table 2.** Comparison of the mean estimates of the GPD parameters using our method and existing ones, obtained from 100 generated data sets for two different cases,  $(\xi, \mu, \sigma) = (1, 2, 1)$  and when  $(\xi, \mu, \sigma) = (0.5, 1, 1)$ , respectively.

#### 4. CONCLUSION

This paper proposes a new numerical method, with an iterative algorithm for unsupervised threshold detection for a Gauss-GPD hybrid model that links a Gaussian distribution to a GPD. Simulation results show that our method enhances the evaluation of the junction point as well as the estimation of the two distributions parameters, in particular compared with the method proposed in [3]. Besides the accurate determination of the threshold (or junction point), we obtain a better estimation of the GPD parameters than with standard methods. We also avoid the issue of hand supervising. As a follow up, it would be interesting to explore a new hybrid model with two junction points introducing an intermediate behavior between the Gaussian distribution and the GPD.

#### REFERENCES

- [1] A. K. Mandava, L. Shahram, and E. E. Regentova, "Reliability assessment of microarray data using fuzzy classification methods: A comparative study," in *Advances in Computing and Communications*. 2011, vol. 190 of *Communications in Computer and Information Science*, pp. 351–360, Springer Berlin Heidelberg.



- [2] M. Knecht and S. Küttel, "The czeledin distribution function," in *XXXIV ASTIN Colloquium, Berlin*, August 2003.
- [3] J. Carreau and Y. Bengio, "A hybrid pareto model for asymmetric fat-tailed data: the univariate case," *Extremes*, vol. 1, pp. 53–76, 2009.
- [4] W. Gehin, "Modélisation des queues de distribution des rendements des actif financiers," M.S. thesis, EURIA, Brest, 2011.
- [5] V. P. Singh C. Li and A. K. Mishra, "Simulation of the entire range of daily precipitation using a hybrid probability distribution," *WATER RESOURCES RESEARCH*, vol. 48, pp. 1–17, 2012.
- [6] M. E. Kuhl and P.S. Bhairgond, "Nonparametric estimation of nonhomogeneous poisson processes using wavelets," in *Simulation Conference, 2000. Proceedings. Winter, 2000*, vol. 1, pp. 562–571 vol.1.
- [7] K. JooSeuk and D.S Clayton, "Robust kernel density estimation," *Journal of Machine Learning Research*, vol. 13, pp. 2529–2565, 2012.
- [8] S. G. Walker, P. Damien, P. W. Laud, and A. F. M. Smith, "Bayesian nonparametric inference for random distributions and related functions," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 61, no. 3, pp. 485–527, 1999.
- [9] P. Orbanz and Y.W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*. Springer, 2010.
- [10] M. Kratz, "There is a var beyond usual approximations," 2013, preprint arXiv:1311.0270.
- [11] J. Pickands, "Statistical inference using extreme order statistics," *Annals of Statistics*, vol. 3, pp. 119–131, 1975.
- [12] J. Carreau and Y. Bengio, "A hybrid pareto mixture for conditional asymmetric fat-tailed distributions," *Neural Networks, IEEE Transactions on*, vol. 20, no. 7, pp. 1087–1101, 2009.
- [13] N. Debbabi, M. Kratz, M. Mboup, and S. El Asmi, "Combining algebraic approach with extreme value theory for spike detection," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 1836–1840.
- [14] F. F. Digham, M. S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," *Communications, IEEE Transactions on*, vol. 55, no. 1, pp. 21–24, 2007.
- [15] Y. Yang, D. Yu, and J. Cheng, "A roller bearing fault diagnosis method based on {EMD} energy entropy and {ANN}," *Journal of Sound and Vibration*, vol. 294, no. 12, pp. 269 – 277, 2006.
- [16] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, 1997.
- [17] S. Tanaka, K. Takara, A. Snorrason, H. Finnsdottir, and E. M. Moss, "A study on threshold selection in pot analysis of extreme floods," *IAHS-AISH publication*, , no. 271, pp. 299–304, 2002, eng.
- [18] B. M. Hill, "A simple general approach to inference about the tail of a distribution," *The Annals of Statistics*, vol. 3, pp. 1163–1174, 1975.
- [19] M. Kratz and S. I. Resnick, "The qq-estimator and heavy tails," *Communications in Statistics. Stochastic Models*, vol. 12, no. 4, pp. 699–724, 1996.
- [20] Y. Demichel, A. Estrade, M. Kratz, and G. Samorodnitsky, "How fast can the chord length distribution decay?," *Advances in Applied Probability*, vol. 43, pp. 504–523, 2011.
- [21] K. Levenberg, "A method for the solution of certain nonlonear problems in least squares quart," *Applied Math*, vol. 2, pp. 164–168, 1944.
- [22] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, pp. 431–441, 1963.
- [23] J. R. M. Hosking and J. R. Wallis, "Parameter and quantile estimation for the generalized pareto distribution," *Technometrics*, vol. 29, no. 3, pp. 339–349, 1987.
- [24] J. Zhang and M. A. Stephens, "A new and efficient estimation method for the generalized pareto distribution," *Technometrics*, vol. 51, no. 3, pp. 316–325, 2009.

Dans une étape plus avancée, nous avons démontré la convergence de l'algorithme en montrant qu'un point stationnaire de l'algorithme n'est autre qu'un point fixe d'une certaine fonctionnelle utilisée pour chaque itération. Ainsi, en utilisant le théorème de point fixe de Brouwer [Kakutani et al., 1941], nous avons pu démontrer l'existence d'un point fixe de cette fonctionnelle. Nous détaillons la preuve de la convergence ci-après.

### Convergence de l'algorithme proposé :

Nous commençons par rappeler l'algorithme ci-dessous :

---

#### **Algorithm 1** Algorithme itératif pour la détermination non supervisée du seuil

---

- 1: Initialisation de  $\tilde{u}^{(0)}$  et  $\epsilon$ .
- 2: Détermination de la fonction de répartition empirique  $H_n$  du n-échantillon  $X = (X_i)_{1 \leq i \leq n}$ , donnée par :

$$H_n(X_i) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}_{\{X_j \leq X_i\}}.$$

- 3: Procédure itérative

$k \leftarrow 0$

**Répéter**

$k \leftarrow k + 1$

**Étape 1 :** Détermination de  $\tilde{\mu}^{(k)}$  et  $\tilde{\sigma}^{(k)}$ , les estimateurs, respectivement, de  $\mu$  et de  $\sigma$  :

$$\tilde{p}^{(k)} = [\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}] \leftarrow \underset{(\mu, \sigma) \in \mathbf{R} \times \mathbf{R}_+^*}{\operatorname{argmin}} \left\| H(X; \theta \mid \tilde{u}^{(k-1)}) - \mathbf{H}_n(X) \right\|_2^2$$

où  $\theta \mid \tilde{u}^{(k-1)}$  désigne  $\theta$  pour  $u = \tilde{u}^{(k-1)}$  fixé et  $\mathbf{H}_n(X) = [H_n(X_1), \dots, H_n(X_n)]^t$ .

**Étape 2 :** Détermination de  $\tilde{u}^{(k)}$ , l'estimateur de  $u$ , comme suit :

$$\tilde{u}^{(k)} \leftarrow \underset{u \in \mathbf{R}_+}{\operatorname{argmin}} \left\| H(X; \theta \mid \tilde{p}^{(k)}) - \mathbf{H}_n(X) \right\|_2^2$$

où  $\theta \mid \tilde{p}^{(k)}$  désigne  $\theta$  pour  $[\mu, \sigma] = \tilde{p}^{(k)}$  fixé.

**Jusqu'à**  $|\tilde{u}^{(k)} - \tilde{u}^{(k-1)}| < \epsilon$  **ou**  $k > n_{\max}$

**Retour**  $\theta^{(k)} = [\tilde{u}^{(k)}, \tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}]$

---

Nous mentionnons que  $H(x; \theta)$  désigne la cdf relative à  $h(x; \theta)$  de (3.16).

Une première étape vers l'étude de la convergence de l'algorithme 1 est de s'intéresser à l'existence d'un point stationnaire de ce dernier. Ce qui constituera le contenu de la section suivante.

#### 3.4.1.1 Existence d'un point stationnaire de l'algorithme

Nous commençons cette section par définir mathématiquement et judicieusement le cadre du travail. En effet, pour une réalisation donnée  $X = (X_i)_{1 \leq i \leq n}$ , nous considérons les deux fonctions

$\varphi_p$  et  $\varphi_u$  définies par :

$$\begin{aligned}\varphi_p : \mathcal{D}_u &\rightarrow \mathcal{D}_p \\ u &\mapsto \varphi_p(u; X) = \underset{p \in \mathcal{D}_p}{\operatorname{argmin}} \sum_u(p; X),\end{aligned}$$

et

$$\begin{aligned}\varphi_u : \mathcal{D}_p &\rightarrow \mathcal{D}_u \\ p &\mapsto \varphi_u(p; X) = \underset{u \in \mathcal{D}_u}{\operatorname{argmin}} \sum_p(u; X),\end{aligned}$$

où  $\mathcal{D}_u$  et  $\mathcal{D}_p$  désignent respectivement les domaines de définition de  $u$  et  $p = [\mu, \sigma]$ . Les deux fonctions  $\sum_u$  et  $\sum_p$  sont définies par :

$$\begin{aligned}\sum_u : \mathcal{D}_p &\rightarrow \mathbb{R} \\ p &\mapsto \sum_u(p; X) = \sum_{i=1}^n \left( H(X_i; \theta | p) - H_n(X_i) \right)^2,\end{aligned}$$

et

$$\begin{aligned}\sum_p : \mathcal{D}_u &\rightarrow \mathbb{R} \\ u &\mapsto \sum_p(u; X) = \sum_{i=1}^n \left( H(X_i; \theta | p) - H_n(X_i) \right)^2,\end{aligned}$$

où  $\theta | u$  (resp.  $\theta | p$ ) représente  $\theta$  pour une valeur de  $u$  donnée (resp. pour une valeur de  $p$  donnée).

Nous mentionnons que  $\sum_u$  et  $\sum_p$  sont deux fonctions convexes. Ainsi, elles admettent chacune un unique minimum. Cela montre que les fonctions  $\phi_p$  et  $\phi_u$  respectivement associées définissent bien deux applications. Comme  $\sum_u$  et  $\sum_p$  sont définies d'une manière très implicite, leur convexité a été déduite *via* les simulations.

Dans ce qui suit,  $\tilde{a}^{(k)}$  désignera l'estimée de la variable  $a$  de l'itération  $k$ .

En utilisant les fonctions  $\varphi_u$  et  $\varphi_p$ , pour un  $\tilde{u}^{(0)}$  donné, la première itération de l'algorithme 1 sera donnée par la relation suivante :

$$\begin{cases} \tilde{p}^{(1)} = \varphi_p(\tilde{u}^{(0)}; X) \\ \tilde{u}^{(1)} = \varphi_u(\tilde{p}^{(1)}; X) = \varphi_u(\varphi_p(\tilde{u}^{(0)}; X); X). \end{cases} \quad (\text{R1})$$

Si maintenant nous considérons la fonctionnelle  $\phi$  définie par :

$$\begin{aligned}\phi : \mathcal{D}_u &\rightarrow \mathcal{D}_u \\ u &\mapsto \phi(u; X) = \varphi_u(\varphi_p(u; X); X),\end{aligned}$$

alors la relation (R1) sera équivalente à (R2), donnée par :

$$\tilde{u}^{(1)} = \phi(\tilde{u}^{(0)}; X). \quad (\text{R2})$$

Si nous utilisons, dans la relation (R1),  $\tilde{u}^{(k-1)}$  comme donnée, pour  $k \geq 1$ , il vient que :

$$\tilde{u}^{(k)} = \phi(\tilde{u}^{(k-1)}; X). \quad (\text{R3})$$

Ainsi, en utilisant (R3), l'algorithme 1 sera équivalent à celui donné ci-dessous :

---

**Algorithm 2** Algorithme itératif pour la détermination non supervisée du seuil : version 2
 

---

- 1: Initialisation de  $\tilde{u}^{(0)}$  et  $\epsilon$ .
- 2: Détermination de la fonction de répartition empirique  $H_n$ .
- 3: Procédure itérative
 

$k \leftarrow 0$   
**Répéter**  
 $k \leftarrow k + 1$

$$\tilde{u}^{(k)} = \phi(\tilde{u}^{(k-1)}; X)$$

**Jusqu'à**  $|\tilde{u}^{(k)} - \tilde{u}^{(k-1)}| < \epsilon$  **ou**  $k > n_{max}$

**Retour**  $\tilde{u}^{(k)}$

---

Une façon de démontrer l'existence d'un point stationnaire de l'algorithme 2 est de démontrer l'existence de point fixe de la fonctionnelle  $\phi$ . Pour ce faire, nous allons nous baser sur le théorème du point fixe. Plusieurs versions de ce théorème ont été développées dans la littérature, telles celle de Banach [Zeyada et al., 2007], la version de Markov-Kakutani [Werner, 1993] ou encore celle de Schauder [Niculescu et al., 2009], etc. Dans nos travaux, nous allons choisir la version de Brouwer [Kakutani et al., 1941] rappelée dans le théorème 3.12.

**Théorème 3.12 (Théorème de Brouwer [Kakutani et al., 1941])**

*Toute application continue d'une boule fermée d'un espace Euclidien dans elle même admet un point fixe.*

Ainsi, d'après le théorème 3.12, pour que la fonctionnelle  $\phi$  admette au moins un point fixe, il faut que les deux conditions suivantes soient satisfaites :

( $\mathcal{C}_1$ ) :  $\mathcal{D}_u$  représente une boule fermée d'un espace Euclidien.

( $\mathcal{C}_2$ ) :  $\phi$  est continue sur  $\mathcal{D}_u$ .

En fait, la condition ( $\mathcal{C}_1$ ) est déjà satisfaite :  $\mathcal{D}_u$  est une boule fermée de  $\mathbb{R}$  qui est un espace Euclidien. Il s'agit bien d'un intervalle fermé borné de  $\mathbb{R}$ . En effet, comme  $u \in \mathbb{R}_+$ , alors  $u$  est toujours minoré par 0. Il est aussi toujours majoré par le maximum des observations. Il vient que  $\mathcal{D}_u = [0, \max(X)]$  qui est bien un fermé borné de  $\mathbb{R}$ .

Nous passons maintenant à démontrer que  $\phi$  est continue sur  $\mathcal{D}_u$ . Pour ce faire, nous avons besoin des deux théorèmes énoncés ci-dessous.

**Théorème 3.13 (Théorème de Heine-Cantor [Rudin, 1964])** *Si  $f$  est une fonction continue sur un compact  $K$ , alors elle est uniformément continue sur  $K$ .*

**Théorème 3.14 ([Ramsay et al., 2007])** *Soient  $\mathcal{X}$  et  $\mathcal{Y}$  deux espaces métriques où  $\mathcal{X}$  est un fermé borné et*

$$\begin{aligned} g : \mathcal{X} \times \mathcal{Y} &\rightarrow \mathbb{R} \\ (x, \alpha) &\rightarrow g(x, \alpha) \end{aligned}$$

une fonction uniformément continue en  $x$  et  $\alpha$  tel que

$$x(\alpha) = \underset{x \in \mathcal{X}}{\operatorname{argmin}} g(x, \alpha)$$

est bien définie,  $\forall \alpha \in \mathcal{Y}$  alors la fonction  $x(\alpha) : \mathcal{Y} \rightarrow \mathcal{X}$  est continue.

Comme  $\phi$  est composée de deux fonctions, pour démontrer qu'elle est continue sur  $\mathcal{D}_u$ , il faut montrer que chacune des fonctions  $\varphi_u$  et  $\varphi_p$  est continue sur son domaine de définition. Nous allons seulement discuter la continuité de  $\varphi_p$ . La continuité de  $\varphi_u$  est obtenue en suivant la même logique que pour  $\varphi_p$ .

### Continuité de $\varphi_p$ :

En se basant sur le théorème 3.14, pour démontrer que  $\varphi_p$  est continue, il faut montrer que la fonction  $\sum_u$  est uniformément continue sur  $\mathcal{D}_p$  et que  $\mathcal{D}_p$  est un compact.

Nous rappelons que pour une réalisation donnée, la moyenne et la variance de la loi normale sont finies et bornées. En effet, nous avons

$$\begin{aligned} \min(X) &\leq \mu \leq \max(X), \\ \frac{MAD}{\sqrt{2}} &\leq \sigma \leq 2MAD, \end{aligned} \tag{3.18}$$

où  $MAD$  désigne la 'Median Absolute Deviation' ou encore la variation de l'écart absolu à la médiane. Voir annexe A pour plus de détails.

Nous mentionnons que le membre à droite de l'inégalité (3.18) est déduit de l'équation (A.4) de l'annexe A alors que le membre gauche de la même inégalité est obtenu en utilisant l'inégalité de Bienaymé-Tchebychev rappelé dans le théorème 3.15 pour  $\alpha = MAD$ . Notons que  $MAD$  est bien positive ce qui permet l'utilisation de ce théorème.

**Théorème 3.15 (Inégalité de Bienaymé-Tchebychev)** Soit la va  $X$  de moyenne  $\mu$  et de variance finie  $\sigma^2$ . Alors pour tout  $\alpha > 0$ ,

$$\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}.$$

Nous concluons alors que  $\mathcal{D}_p \subseteq [\min(X), \max(X)] \times [\frac{MAD}{\sqrt{2}}, 2MAD]$ . En considérant  $\mathcal{D}_p = [\min(X), \max(X)] \times [\frac{MAD}{\sqrt{2}}, 2MAD]$ , nous avons bien un compact de  $\mathbb{R}^2$  : le produit cartésien de deux compacts de  $\mathbb{R}$ .

Maintenant, il nous reste à montrer que  $\sum_u$  est uniformément continue sur  $\mathcal{D}_p$ . Comme  $\mathcal{D}_p$  est un compact de  $\mathbb{R}^2$  alors il suffira, d'après le théorème 3.13, de montrer que  $\sum_u$  est continue pour déduire sa continuité uniforme. En effet, comme  $\sum_u(p; X) = \sum_{i=1}^n (H(X_i; \theta | u) - H_n(X_i))^2$  alors pour étudier sa continuité, il faut étudier celle de la cdf hybride par rapport à  $p$ . Comme  $H$  est composée de deux fonctions, nous allons étudier la continuité de chacune par rapport à  $p$ . Nous commençons par la cdf gaussienne et nous considérons  $p$  et  $q = p + \delta p$  deux éléments de

$\mathcal{D}_p$  tel que  $\|\delta p\| \rightarrow 0$ , où  $\delta p = \begin{bmatrix} \delta\mu \\ \delta\sigma \end{bmatrix}$  est une petite perturbation de  $p$  et  $\|\cdot\|$  est une norme sur  $\mathbb{R}^2$ . En écrivant le développement limité d'ordre 1 de  $F$  autour de  $q$  nous obtenons pour tout  $x \in \mathbb{R}$ , fixé, l'expression suivante :

$$F(p; x) = F(q; x) + \langle \nabla F(q; x), \delta p \rangle + \theta(\|\delta p\|), \quad (3.19)$$

où  $\theta(\|\delta p\|) \rightarrow 0$  lorsque  $\|\delta p\| \rightarrow 0$ ,  $\langle \cdot \rangle$  est le produit scalaire usuel de  $\mathbb{R}^2$  et l'opérateur  $\nabla F$  désigne le gradient de  $F$  donné par :

$$\nabla F(q; x) = \begin{bmatrix} \frac{\partial F(p; x)}{\partial \mu} \\ \frac{\partial F(p; x)}{\partial \sigma} \end{bmatrix} = \begin{bmatrix} -f(x; p) \\ \frac{\mu - u}{\sigma} f(x; p) \end{bmatrix}.$$

De l'équation (3.19), nous obtenons ainsi

$$\begin{aligned} |F(p; x) - F(q; x)| &= |\langle \nabla F(q; x), \delta p \rangle + \theta(\|\delta p\|)| \\ &\leq |\langle \nabla F(q; x), \delta p \rangle| + |\theta(\|\delta p\|)| \\ &\leq |f(x; p)| \left| -\delta\mu + \frac{\mu - u}{\sigma} \delta\sigma \right| + |\theta(\|\delta p\|)| \end{aligned} \quad (3.20)$$

Comme  $\|\delta p\| \rightarrow 0$  et que  $\theta(\|\delta p\|) \rightarrow 0$  lorsque  $\|\delta p\| \rightarrow 0$  le membre de droite de (3.20) tendra aussi vers 0. Ainsi, nous avons montré que

$$\forall p, q \in \mathcal{D}_p \text{ tel que } \|p - q\| \rightarrow 0 \text{ alors } |F(p; x) - F(q; x)| \rightarrow 0.$$

D'où la continuité de  $F$  par rapport à  $p$ . De même, nous montrons la continuité de la cdf  $G$  de Pareto généralisé par rapport à ses paramètres  $\xi$  et  $\beta$ . Nous rappelons que dans notre cas,  $\xi$  et  $\beta$  s'expriment en fonction de  $p$  comme suit :

$$\begin{aligned} \beta(p) &= \frac{1}{f(p; x)} \\ \xi(p) &= \frac{u - \mu}{\sigma} \beta - 1 \end{aligned}$$

qui sont deux fonctions continues par rapport à  $p$ .  $G$  est ainsi continue par rapport à  $p$  comme composée de fonctions continues par rapport à  $p$ .

Finalement, nous pouvons déduire la continuité de la fonction  $\sum_u(p; X) = \sum_{i=1}^n (H(X_i; \theta | u) - H_n(X_i))^2$  sur  $\mathcal{D}_p$  comme composée, somme et produit de fonctions continues sur  $\mathcal{D}_p$ . Par conséquent, d'après le théorème 3.13, cette fonction est uniformément continue sur  $\mathcal{D}_p$ . Ce qui nous donne la continuité de la fonction  $\varphi_p$  sur  $\mathcal{D}_u$  d'après le théorème 3.14.

### Continuité de $\varphi_u$ :

Comme il est dit plus haut, nous n'allons pas détailler la continuité de  $\varphi_u$  sur  $\mathcal{D}_p$ . Elle en fait démontrée en procédant de la même manière que pour  $\varphi_p$ . Nous montrons d'abord que la fonction  $\sum_p$  est uniformément continue sur  $\mathcal{D}_u$  qui est un compact de  $\mathbb{R}$ . Ensuite, nous

déduisons la continuité de  $\varphi_u$  sur  $\mathcal{D}_p$  via le théorème 3.14.

Nous pouvons conclure à la fin de cette section que la fonctionnelle  $\phi$  est continue sur  $\mathcal{D}_u$  comme composée de deux fonctions continues :  $\varphi_p$  et  $\varphi_u$ . D'où l'existence d'au moins un point fixe par  $\phi$  d'après le théorème 3.12 de Brouwer. Ce qui montre l'existence d'au moins un point stationnaire de l'algorithme. Maintenant, pour s'assurer de la convergence vers ce dernier, une deuxième étape consiste à étudier l'unicité du point fixe et sa nature : s'il est attractif ou répulsif. Nous aimerions bien avoir un unique point fixe attractif pour conclure la convergence de l'algorithme vers ce point. Ceci peut être assuré lorsque la fonctionnelle  $\phi$  est une contraction, ou encore une fonction  $k$ -Lipschitzienne avec  $k \in ]0, 1[$  (théorème du point fixe du Banach [Zeyada et al., 2007]). Comme  $\phi$  est définie d'une manière très implicite, ses propriétés analytiques sont inconnues. Il est toutefois difficile de montrer qu'elle est contractante. C'est pourquoi nous allons étudier la convergence de l'algorithme empiriquement via des simulations. En effet, dans les figures 3.4 et 3.5 nous traçons différentes suites récurrentes  $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$  suivant la valeur initiale  $\tilde{u}^{(0)}$  de  $u$ . Pour chaque figure, nous avons généré des données suivant la loi hybride pour un vecteur de paramètres  $\theta$  fixé. Ensuite, sur ces données nous avons appliqué l'algorithme 1 pour des différentes valeurs initiales de  $u$ . Pour chaque valeur  $\tilde{u}^{(0)}$ , nous avons tracé la suite récurrente  $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$  associée. La figure 3.4 représente l'exemple de  $\theta = [0.4354, 0, 1]$  avec  $\xi = 0.2$  et  $\beta = 2.7558$ . Nous remarquons à travers cette figure que indépendamment du choix de  $\tilde{u}^{(0)}$ , la suite récurrente converge vers la valeur exacte de  $u$ , représentée par la ligne horizontale en noir, qu'on notera  $u^*$  et qui vaut 0.4354. Nous pouvons ainsi tirer la conclusion suivante :

1. Si  $\tilde{u}^{(0)} \leq u^*$ , alors la suite récurrente associée  $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$  est croissante. Prenons l'exemple de  $\tilde{u}^{(0)} = q_{0.35}$  : courbe des cercles gris ou encore l'exemple de  $\tilde{u}^{(0)} = q_{0.375}$  : la courbe des triangles rouges dirigés vers le haut.
2. Si  $\tilde{u}^{(0)} \geq u^*$ , alors la suite récurrente associée  $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$  est décroissante. Prenons l'exemple de  $\tilde{u}^{(0)} = q_{0.45}$  : courbe des losanges bleus ou encore l'exemple de  $\tilde{u}^{(0)} = q_{0.475}$  : la courbe des triangles en rose dirigés vers le bas.

Par conséquent, d'après la figure 3.4, indépendamment du choix de la valeur initiale  $\tilde{u}^{(0)}$ , la suite  $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$  est monotone sur  $\mathcal{D}_p$  et converge vers un point stationnaire qui est  $u^*$ .

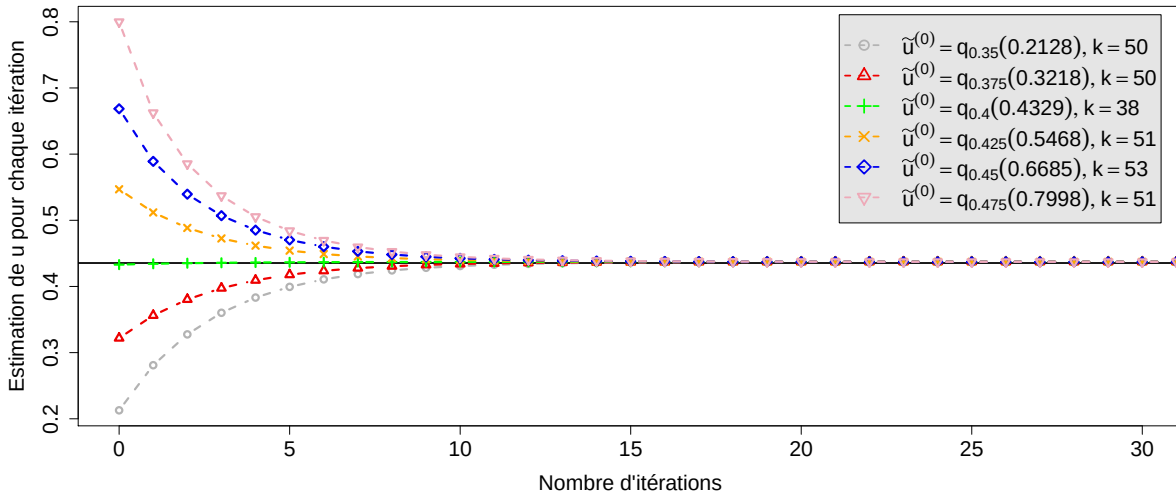
**Exemple 1 de la convergence de l'algorithme vers le point stationnaire désiré ( $u=0.4354$ )**

FIGURE 3.4 – Etude du comportement de la suite récurrente  $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$  suivant la valeur initiale de  $\tilde{u}^{(0)}$ . Exemple 1 pour  $\theta = [0.4354, 0, 1]$

Nous pouvons voir aussi à travers la figure 3.4 et la figure 3.5, qui d'ailleurs étudie la convergence de l'algorithme pour un autre exemple, que lorsque la valeur initiale s'approche de la valeur exacte, l'algorithme met moins d'itérations pour converger (courbe des + verts pour l'exemple 1 et courbe des croix oranges de l'exemple 2).

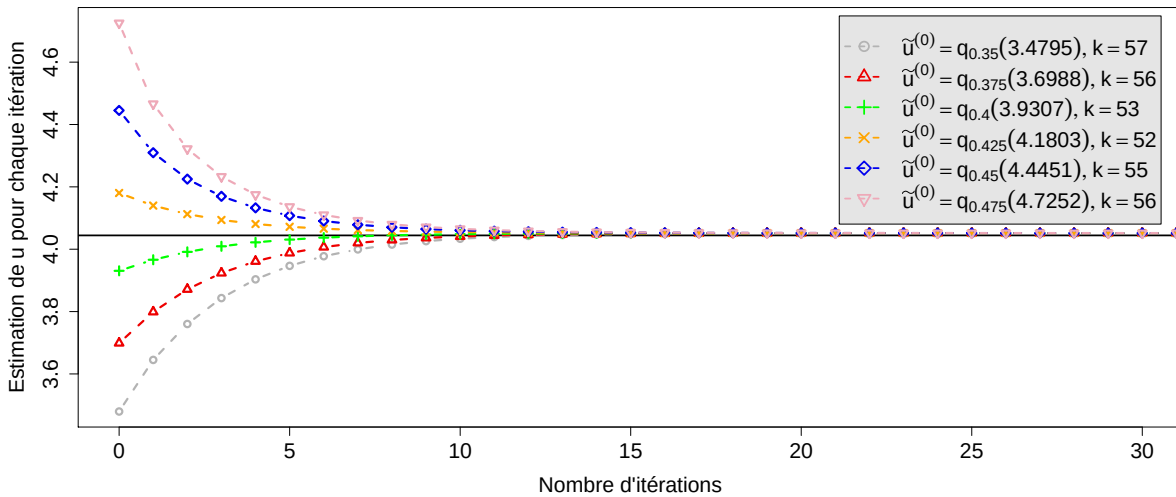
**Exemple 2 de la convergence de l'algorithme vers le point stationnaire désiré ( $u=4.0443$ )**

FIGURE 3.5 – Etude du comportement de la suite récurrente  $\{\tilde{u}^{(k+1)} = \phi(\tilde{u}^{(k)})\}_{k \in \mathbb{N}^*}$  suivant la valeur initiale de  $\tilde{u}^{(0)}$ . Exemple 2 pour  $\theta = [4.0443, 3, 2]$

Nous avons démontré mathématiquement l'existence de point stationnaire de l'algorithme proposé et nous avons montré *via* les simulations son unicité. Ainsi, nous pouvons conclure à la convergence de l'algorithme proposé.



### 3.4.2 Modèle hybride à deux composantes Rayleigh-GPD avec pondération uniforme : résumé du papier [Debbabi et al., 2014]

Pour montrer que le modèle (3.15) peut modéliser de nombreux exemples de données, nous allons rester dans le même principe du modèle (3.16) mais cette fois-ci en remplaçant la pdf gaussienne par une Rayleigh pour modéliser l'évanouissement, dit aussi fading en anglais, du signal reçu à travers un canal de transmission sans fil [Debbabi et al., 2014]. Ce travail, que nous exposons par la suite, a été publié dans la conférence internationale *World Congress On E-Learning, Education and Computer Science (WCEECS2014)*. Le modèle considéré dans ce papier, qu'on le nomme Rayleigh-GPD, est défini par la pdf hybride suivante :

$$h(x; \theta) = \begin{cases} \gamma r(x; \sigma), & \text{si } x \leq u \\ \gamma g(x - u; \xi, \beta), & \text{si } x > u \end{cases} \quad (3.21)$$

où  $r$  désigne la pdf de loi de Rayleigh de paramètre d'échelle  $\sigma \in \mathbb{R}_+^*$  donnée par :

$$r(x; \sigma) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad \forall x \in \mathbb{R}_+. \quad (3.22)$$

Le paramètre  $\gamma$  représente le poids uniforme,  $\xi$  et  $\beta$  sont les paramètres de la GPD et  $u$  désigne le seuil. Dans ce modèle  $\theta = [\sigma, u]$ . Pour  $\theta = [1, 1.3716]$ , nous traçons dans la figure 3.6, la pdf hybride Rayleigh-GPD correspondante ainsi que sa composante Rayleigh, où nous remarquons une bonne différence entre ces deux dernières à partir du seuil  $u$ , donné par le point de jonction.

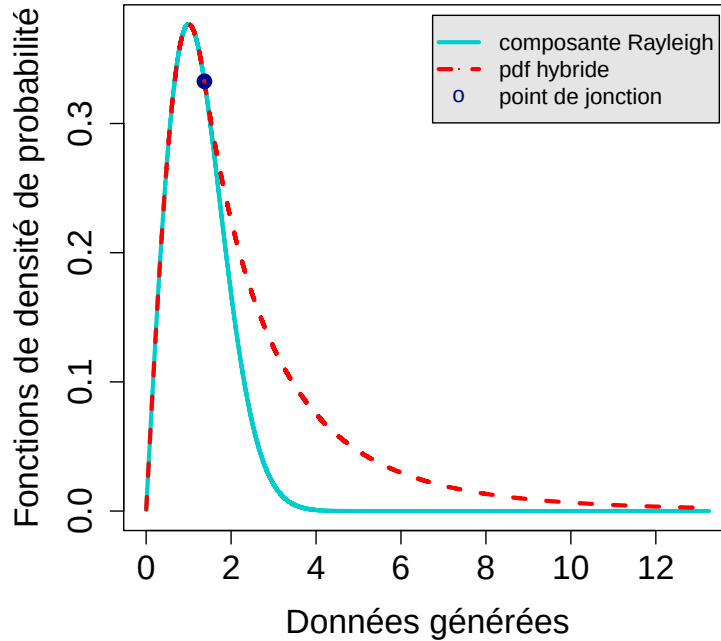


FIGURE 3.6 – Exemple de pdf hybride Rayleigh-GPD pour  $\theta = [1, 1.3716]$ .

Comme pour [Carreau and Bengio, 2009b] et [Debbabi and Kratz, 2014], la continuité et la dérivabilité de la fonction de densité de probabilité de la loi hybride Rayleigh-GPD ont été imposées. Ainsi, des relations entre les paramètres des deux lois ont été trouvées permettant de

réduire le nombre de paramètres à estimer de 5 à 2 paramètres qui sont le paramètre d'échelle de la loi Rayleigh  $\sigma$  et le seuil  $u$ . Comme nous avons que deux paramètres à estimer, nous l'avons estimé directement en minimisant la moyenne quadratique entre la cdf Rayleigh-Gauss et la cdf empirique. Une fois  $\sigma$  et  $u$  sont estimés, les estimations de  $\gamma$ ,  $\beta$  et  $\xi$  sont déterminées à travers les relations trouvées par les contraintes imposées à la pdf  $h$ .

Dans le papier [Debbabi et al., 2014], la performance du modèle Rayleigh-GPD a été testée en se basant sur les simulations de Monte Carlo et en utilisant le critère du logarithme du rapport de vraisemblance moyen.

Nous notons que la pdf  $h(x; t, \sigma, \xi, \beta)$  de ce papier désigne bien la pdf Rayleigh-Gauss  $h(x; \theta)$  où  $t = u$ .

## EXTREME VALUE THEORY FOR WIRELESS FADING CHANNEL MODELING

*Nehla DEBBABI<sup>1,2</sup>, Sadok EL ASMI<sup>1</sup> and Hichem ARFA<sup>3</sup>*

<sup>1</sup> Research Lab. COSIM, Engineering College of Communications of Tunis, University of Carthage

<sup>2</sup> CReSTIC, Faculty of Exact and Natural Sciences, University of Reims Champagne Ardenne

<sup>3</sup> Research Lab. INNOV'COM, Engineering College of Communications of Tunis, University of Carthage  
nehla.debbabi@supcom.tn, elasmis@supcom.rnu.tn, hichem.arfa@isima.rnu.tn

### ABSTRACT

This paper proposes a new distribution for wireless fading channel modeling. The proposed distribution, named by R-GPD, consists on the connection of Rayleigh distribution, modeling the mean behavior of the channel by virtue of the Central Limit Theorem (CLT), and a Generalized Pareto Distribution (GPD), modeling the asymptotic behavior of the envelope fading, according to Extreme Value Theory (EVT). In the case when the strength of the transmitted signal decreases over time, the proposed hybrid distribution offers better modeling of the envelope fading behavior than the classical Rayleigh one. To estimate the parameters of the proposed distribution, a new technique based on a minimization in the least squares sense is suggested. According to Monte Carlo simulations, the performance of the new technique for the R-GPD parameters estimation follows the average log-likelihood ratio criterion.

**Index Terms**— Extreme Value Theory, Deep fading channel, Generalized Pareto Distribution, Rayleigh distribution

### 1. INTRODUCTION

The presence of fading in the wireless channel requires an understanding of the statistical proprieties of the envelope of the received signal, which is essential in performance analysis of contemporary wireless communication systems, such as Long Term Evolution (LTE) system. When the observed fading is a short term fading, which consists on a fast amplitude fluctuations in the received signal [1], it is often modeled by a Rayleigh probability density function (pdf) by virtue of the Central Limit Theorem (CLT) [2]. Indeed, the received signal is simply the summation of propagated signals via multiple paths in the channel in terms of inphase and quadrature. When the number of multiple paths is large, the received signal can be a Gaussian complex signal. Therefore, the envelope of the fading is Rayleigh distributed [3]. The Rayleigh distribution has shown a good performances in terms of well-fitting the envelope of the short term fading

[4]. Unfortunately, this distribution misestimates the channel when the signal strength decreases over time, involving a deep fading and a considered outage rate. Such scenario can be treated as a reduction of the number of received signals, which do not allow the application of the CLT, implying the inadequate use of the Rayleigh distribution. In such situation, the occurrence of a deep fading can be considered as extreme rare event, since the deep fading appears with a low probability and relatively high amplitude. However, it will be illustrated by the right tail of the envelope fading distribution. The misestimating of the distribution tail of the envelope fading using the Rayleigh distribution motivates the idea of separating the mean behavior of the channel from the asymptotic one. Since the encountered problem appears in the tail of the distribution, we propose in this paper a new distribution that takes into account the asymptotic behavior of the envelope fading, described via the Extreme Value Theory (EVT) [5]. As a matter of fact, according to the Pickands theorem [6], the Peak Over Threshold (POT) method of the EVT shows that the exceedances above a high threshold follow a Generalized Pareto distribution (GPD), regardless the underlying distribution. Based on that result, we introduce in this work a new hybrid pdf, that we denote by R-GPD, linking two distributions; a Rayleigh to a GPD at a junction point (or threshold).

Note that several hybrid pdfs have been proposed in literature for a better modeling of various type of non-homogenous data in different contexts, such as *e.g.* in medicine to model the DNA microarray [7], in finance to model the profit and loss of a portfolio [8] or in insurance to model the claims [9] to name a few.

To make the proposed pdf smooth, as in [9, 10], we enforced the continuity of the R-GPD pdf and of its derivative at the junction point. Therefore, we obtain a relation between the junction point and the parameters of the GPD and the Rayleigh distributions. Besides having a smooth pdf, the imposed conditions at the junction point reduce the number of the R-GPD parameters to be estimated. Indeed, the R-GPD model have 4 unknown parameters; the scale parameter of the Rayleigh distribution, the tail index and the scale parameter

of the GPD and the junction point or the threshold above which the data follow a GPD. The knowledge of two parameters allows the determination of the other two remainders. Much work have been done in literature for the estimation of the threshold of the POT method [5, 11, 12, 13, 14]. Unfortunately, these methods are a hand tuning methods which requires the human intervention. Unlike the existing methods, in this work, we determine the junction point automatically, which represents a great novelty in the EVT. In fact, to estimate the R-GPD parameters, we propose in this paper a new technique providing a satisfactory results. It consists on the estimation of the threshold and the scale parameters of the Rayleigh distribution via a minimization in the least squares sens. To resolve the optimization problem we built on the Levenberg-Marquardt method [15, 16]. Once these two parameters are estimated, we determine the GPD parameters ensuring the continuity and the differentiability of the R-GPD at the junction point. To carry out the readability of the proposed technique for parameters estimation, we based on on Monte Carlo simulations where the used evaluation criterion is the average log-likelihood ratio.

We mention that in this work, only the analytical study of the proposed R-GPD distribution for the envelope fading modeling is developed.

The remainder of this paper is organized as follows. In section 2 we detail the proposed hybrid distribution for fading channel modeling. The proposed technique for R-GPD parameters estimation is given in section 3. Numerical results are discussed in section 4. And finally, conclusions are drawn in the last section.

## 2. DESCRIPTION OF THE PROPOSED FADING CHANNEL MODEL

We describe in this section the proposed hybrid model for fading channel modeling. This model, denoted by *R-GPD*, links a Rayleigh distribution to a Generalized Pareto Distribution (GPD) at a junction point  $t$ . It is then defined by the following probability density function (pdf)

$$h(x; t, \sigma, \xi, \beta) = \begin{cases} \alpha r(x; \sigma), & \text{if } x \leq t \\ \alpha g(x - t; \xi, \beta), & \text{if } x > t \end{cases}$$

where  $r$  represents the Rayleigh pdf expressed as a function of the scale parameter  $\sigma \in \mathbb{R}_+^*$  as follows

$$r(x; \sigma) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad \forall x \in \mathbb{R}^+$$

while  $g$  is the GPD pdf admitting  $\xi \in \mathbb{R}$  and  $\beta \in \mathbb{R}_+^*$  as tail index and scale parameter, respectively, and expressed as

$$g(x; \xi, \beta) = \begin{cases} \frac{1}{\beta} (1 + \frac{\xi}{\beta} x)^{-1-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right), & \text{if } \xi = 0 \end{cases} \quad \forall x \in \mathcal{D}(\xi, \sigma)$$

where

$$\mathcal{D}(\xi, \sigma) = \begin{cases} [0, \infty), & \text{if } \xi \geq 0 \\ [0, -\frac{\beta}{\xi}], & \text{if } \xi < 0 \end{cases}$$

Here, the regular parameter  $\alpha$  is introduced to ensure that  $h$  is a pdf. It will be determined in section 2.1

By integrating the hybrid pdf  $h$ , we can infer to the cumulative distribution function (cdf) of the *R-GPD* model expressed as

$$H(x; t, \sigma, \xi, \beta) = \begin{cases} \alpha R(x; \sigma), & \text{if } x \leq t \\ \alpha (R(t; \sigma) + G(x - t; \xi, \beta)), & \text{if } x > t \end{cases}$$

where  $R$  denotes the Rayleigh cdf while  $G$  represents the GPD one.

### 2.1. Determination of the regulator parameter

To ensure that  $h$  defines a pdf, it must integrates to 1 *i.e.*

$$\int_{\mathbb{R}} h(x; t, \sigma, \xi, \beta) dx = 1. \quad (1)$$

Equation (1) is equivalent to the following one

$$\int_{-\infty}^t \alpha r(x; \sigma) dx + \int_t^{+\infty} \alpha g(x - t; \xi, \beta) dx = 1. \quad (2)$$

By applying the following substitution  $\lambda \leftarrow x - t$  in the second integral of equation (2), we obtain

$$\alpha R(t; \sigma) + \underbrace{\alpha \int_0^{+\infty} g(\lambda; \xi, \beta) d\lambda}_{=1} = 1. \quad (3)$$

Consequently, from (3) we can deduce the following expression of  $\alpha$

$$\alpha = \frac{1}{R(t; \sigma) + 1}$$

### 2.2. Determination of the GPD parameters

As in [10], in order to obtain a continuous hybrid pdf, the continuity and the differentiability of the *R-GPD* pdf are enforced at the junction point  $t$ . This, is mathematically given by the following two equations

$$\begin{cases} r(t; \sigma) = g(0; \xi, \beta) \\ r'(t; \sigma) = g'(0; \xi, \beta) \end{cases} \quad (4)$$

The system of equations (4) is equivalent to the following one

$$\begin{cases} r(t; \sigma) = \frac{1}{\beta} \\ \frac{1}{t} - \frac{t}{\sigma^2} = -\frac{1+\xi}{\beta} \end{cases} \quad (5)$$

Consequently, a deterministic expressions of  $\xi_t$  and  $\beta_t$  are then obtained and expressed as

$$\begin{cases} \beta_t = r(t; \sigma)^{-1} \\ \xi_t = \beta(\frac{-1}{t} + \frac{t}{\sigma^2}) - 1 \end{cases} \quad (6)$$

Hence, once the scale parameter of the Rayleigh distribution and the junction point we can evaluate the GPD part of the R-GPD model using equations of the system (6). Here, the GPD parameters  $\xi_t$  and  $\beta_t$  are indexed by  $t$  to mention that they depend on the value of the threshold  $t$ . If we change the threshold, the observations above this latter change and then we obtain a different GPD with different parameters. In the next section, we will describe how we estimate the parameters of the R-GPD pdf.

### 3. PROPOSED TECHNIQUE FOR R-GPD MODEL EVALUATION

To evaluate the R-GPD model, we propose in this section a technique that provides an accurate estimation of the R-GPD parameters. Our technique is simple and reliable. It consists on two main steps which we describe hereafter.

To start, let us consider an  $n$ -sample  $X = (x_i)_{1 \leq i \leq n}$  following the R-GPD distribution. To evaluate the proposed model, we proceed as follows

1. The first step consists on the estimation of the threshold or the junction point  $t$  and the scale parameter of the Rayleigh distribution  $\sigma$ . In order to achieve that, we keep the two values  $\hat{t}$  and  $\hat{\sigma}$  minimizing the distance between the Rayleigh cdf and the empirical one, denoted by  $H_n$  and given by the equation (8), in the least squares sense. Hence, the vector  $(\hat{t}, \hat{\sigma})$  is the solution of the following optimization problem

$$(\hat{t}, \hat{\sigma}) = \underset{\substack{\sigma \in \mathbb{R}_+^* \\ t \in \mathbb{R}_+^*}}{\operatorname{argmin}} \|H(X; t, \sigma, \xi_t, \beta_t) - H_n(X)\|_2^2, \quad (7)$$

where

$$H_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{x_i \leq x\}}, \quad \forall x \in \mathbb{R}. \quad (8)$$

Here 1 represents the indicator function.

To resolve (7), we use the Levenberg-Marquardt method detailed in [15, 16], which is a robust method for the resolution of nonlinear least squares optimization problem, based on a combination of the Gauss-Newton and the gradient algorithms.

2. Once  $\hat{t}$  and  $\hat{\sigma}$  are estimated, we determine the estimates of the GPD parameters  $\hat{\xi}_{\hat{t}}$  and  $\hat{\beta}_{\hat{t}}$  using equations (6) as

follows

$$\begin{cases} \beta_{\hat{t}} = r(\hat{t}; \hat{\sigma})^{-1} \\ \xi_{\hat{t}} = \beta_{\hat{t}}(\frac{-1}{\hat{t}} + \frac{\hat{t}}{\hat{\sigma}^2}) - 1 \end{cases} \quad (9)$$

The proposed technique is validated via Monte Carlo simulations detailed in the next section.

## 4. NUMERICAL RESULTS

Based on Monte-Carlo simulations, we highlight the accuracy of the R-GPD parameters estimation, using the proposed technique, in terms of average log-likelihood ratio minimization. To do this, for a fixed sequence of parameters  $\theta = (\sigma, t, \xi_t, \beta_t)$ , we generate 100 random sets of data following the R-GPD distribution admitting those parameters. We denote by  $X^p = (x_i^p)_{1 \leq i \leq n}$ , the  $p^{th}$  set of data, where  $n = 10^5$  represents the number of samples. For  $1 \leq p \leq 100$ , we estimate the parameters of the R-GPD pdf,  $\hat{\theta} = (\hat{\sigma}^p, \hat{t}^p, \hat{\xi}_t^p, \hat{\beta}_t^p)$ , relative to  $X^p$ . Thereafter, we calculate the log-likelihood ratio  $r_p$  between the fixed and the estimated pdfs as follows

$$r_p = \frac{1}{n} \sum_{i=1}^n \log\left(\frac{h(x_i^p; t, \sigma, \xi, \beta)}{h(x_i^p; \hat{t}^p, \hat{\sigma}^p, \hat{\xi}_t^p, \hat{\beta}_t^p)}\right), \quad \forall 1 \leq p \leq 100$$

Once the  $r_p$ ,  $1 \leq p \leq 100$  are computed, we can then determine the average log-likelihood ratio, denoted by  $R$ , which represents our evaluation criterion. It is expressed as

$$R = \frac{1}{p} \sum_{i=1}^p r_p$$

It is important to note that the log-likelihood ratio defines a distance on  $\mathbb{R}^p$ . Hence,  $\forall 1 \leq p \leq 100$ , the more close to 0 the  $r_p$  is, the more accurate the estimation of the R-GPD parameters is.

We illustrate in the Table 1 the simulation results obtained for three different examples: for  $\theta_1 = (1, 1.3716, 0.2, 1.8676)$ ,  $\theta_2 = (2, 2.5696, -0.1, 3.5534)$  and  $\theta_3 = (0.5, 0.7249, 0.5, 0.9865)$ . For each example, we represent in this table, the estimated parameters of the corresponding pdf. As it is clearly noticeable in Table 1, the absolute error between the fixed and the estimated parameters is of order  $10^{-4}$  which proves the reliability of the proposed technique of R-GPD parameters estimation. The accuracy of the parameters estimation is as well, highlighted via the average log-likelihood ratio. For the three simulated cases, this average is of order  $10^{-6}$  involving a well evaluation of the model.

**Table 1.** R-GPD model evaluation for different examples

| Params                                                | $\sigma$ | $t$    | $\xi$   | $\beta$ |
|-------------------------------------------------------|----------|--------|---------|---------|
| <b>Example 1</b>                                      |          |        |         |         |
| Fixed                                                 | 1        | 1.3716 | 0.2     | 1.8676  |
| Estimated                                             | 1.0002   | 1.3718 | 0.1996  | 1.8679  |
| Absolute error ( $10^{-4}$ )                          | 2.2409   | 2.0996 | 3.2967  | 3.0825  |
| Average log-likelihood ratio: $-2.6537 \cdot 10^{-7}$ |          |        |         |         |
| <b>Example 2</b>                                      |          |        |         |         |
| Fixed                                                 | 2        | 2.5696 | -0.1    | 3.5534  |
| Estimated                                             | 1.9993   | 2.5691 | -0.0994 | 3.5525  |
| Absolute error ( $10^{-4}$ )                          | 6.8718   | 5.7202 | 5.2720  | 9.3274  |
| Average log-likelihood ratio: $9.5291 \cdot 10^{-6}$  |          |        |         |         |
| <b>Example 3</b>                                      |          |        |         |         |
| Fixed                                                 | 0.5      | 0.7249 | 0.5     | 0.9865  |
| Estimated                                             | 0.5000   | 0.7250 | 0.5004  | 0.9866  |
| Absolute error ( $10^{-4}$ )                          | 0.1153   | 0.6962 | 4.3905  | 1.0486  |
| Average log-likelihood ratio: $-1.2376 \cdot 10^{-6}$ |          |        |         |         |

## 5. CONCLUSIONS

This paper develops a new distribution for wireless fading channel modeling. The main contribution of the proposed distribution, named R-GPD, is to consider both the mean and the asymptotic behaviors of the channel. The R-GPD pdf represents a junction between a Rayleigh pdf and a GPD one. In order to estimate the parameters of this pdf, a new technique, based on the use of the Levenberg-Marquardt has been proposed. To carry out the performance of this technique, the average log-likelihood criterion has been used. As a follow up, it would be interesting to apply our method on real deep fading context.

## 6. REFERENCES

- [1] M. Schwartz, *Mobile Wireless Communications*, Cambridge University Press, 2005.
- [2] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*, 2002.
- [3] P. Mohana Shankar, *Fading and shadowing in wireless systems*, Springer, 2012.
- [4] A. Gaya and D.C. Panda, "Modeling, calculating and capacity enhancement of a rayleigh fading mimo channel," in *Applied Electromagnetics Conference (AEMC), 2011 IEEE*, 2011, pp. 1–3.
- [5] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, 1997.
- [6] J. Pickands, "Statistical inference using extreme order statistics," *Annals of Statistics*, vol. 3, pp. 119–131, 1975.
- [7] A. K. Mandava, L. Shahram, and E. E. Regentova, "Reliability assessment of microarray data using fuzzy classification methods: A comparative study," in *Advances in Computing and Communications*. 2011, vol. 190 of *Communications in Computer and Information Science*, pp. 351–360, Springer Berlin Heidelberg.
- [8] M. Knecht and S. Küttel, "The czeledin distribution function," in *XXXIV ASTIN Colloquium, Berlin*, August 2003.
- [9] J. Carreau and Y. Bengio, "A hybrid pareto mixture for conditional asymmetric fat-tailed distributions," *Neural Networks, IEEE Transactions on*, vol. 20, no. 7, pp. 1087–1101, 2009.
- [10] J. Carreau and Y. Bengio, "A hybrid pareto model for asymmetric fat-tailed data: the univariate case," *Extremes*, vol. 1, pp. 53–76, 2009.
- [11] N. Debbabi, M. Kratz, M. Mboup, and S. El Asmi, "Combining algebraic approach with extreme value theory for spike detection," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 1836–1840.
- [12] S. Tanaka, K. Takara, A. Snorrason, H. Finnsdottir, and E. M. Moss, "A study on threshold selection in pot analysis of extreme floods," *IAHS-AISH publication*, , no. 271, pp. 299–304, 2002, eng.
- [13] B. M. Hill, "A simple general approach to inference about the tail of a distribution," *The Annals of Statistics*, vol. 3, pp. 1163–1174, 1975.
- [14] Y. Demichel, A. Estrade, M. Kratz, and G. Samorodnitsky, "How fast can the chord length distribution decay?," *Advances in Applied Probability*, vol. 43, pp. 504–523, 2011.
- [15] K. Levenberg, "A method for the solution of certain nonlinear problems in least squares quart," *Applied Math*, vol. 2, pp. 164–168, 1944.
- [16] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, pp. 431–441, 1963.

Comme dans la recherche, nous voulons toujours améliorer les choses, nous allons considérer par la suite le modèle (3.16) avec une pondération non-uniforme, afin d'avoir plus de degrés de liberté, mais en contrepartie un paramètre de plus à estimer. Dans la section 3.4.3, nous donnons un résumé de ce travail [Debbabi et al., 2015a] qui a été publié dans la conférence internationale *2015 International Conference on Communications and Networking (ComNet)*.

### 3.4.3 Modèle hybride à deux composantes Gauss-GPD avec pondération non-uniforme : résumé du papier [Debbabi et al., 2015a]

Le modèle que nous avons considéré dans [Debbabi et al., 2015a] est donnée par la pdf suivante :

$$h(x; \theta) = \begin{cases} \gamma_1 f(x; \mu, \sigma), & \text{si } x \leq u \\ \gamma_2 g(x - u; \xi, \beta), & \text{si } x > u \end{cases} \quad (3.23)$$

où ici  $\theta = [\mu, \sigma, \gamma_1, \gamma_2]$  et les poids  $\gamma_1$  et  $\gamma_2$  sont de valeurs différentes.

Nous représentons dans la figure 3.8 la pdf  $h$ , donnée par (3.23) pour un  $\theta = [0, 1, 0.9, 0.2]$ .

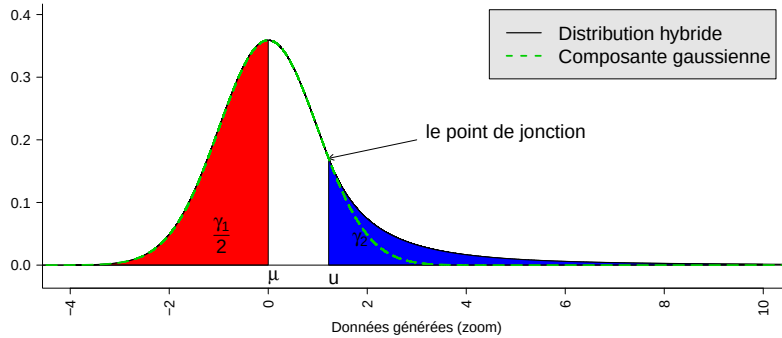


FIGURE 3.7 – Exemple de pdf hybride Gauss-GPD pour  $\theta = [0, 1, 0.9, 0.2]$ .

La motivation derrière la proposition de ce modèle est de mieux manipuler les queues de distributions, puisque c'est cette partie de la distribution qui va nous donner plus d'information sur les ruptures. En effet, si nous regardons la cdf correspondante au modèle 3.23 donnée par :

$$H(x; \theta) = \begin{cases} \gamma_1 F(x; \mu, \sigma), & \text{si } x \leq u_\theta \\ \gamma_1 F(u_\theta; \mu, \sigma) + \gamma_2 G(x - u_\theta; \xi_\theta, \beta_\theta), & \text{si } x > u_\theta \end{cases} \quad (3.24)$$

où  $F$  et  $G$  désignent respectivement la cdf de la gaussienne et de la GPD, en imposant  $\int_{\mathbb{R}} h(x; \theta) = 1$  comme pour [Carreau and Bengio, 2009b], [Debbabi and Kratz, 2014] et [Debbabi et al., 2014], nous obtenons la relation suivante :

$$\gamma_1 F(u_\theta; \mu, \sigma) + \gamma_2 = 1 \quad (3.25)$$

Nous allons montrer en utilisant (3.24) et (3.25) la différence entre le modèle (3.16) avec pondération uniforme et le modèle (3.23) où  $\gamma_1 \neq \gamma_2$ . Pour le cas où  $\gamma = \gamma_1 = \gamma_2$ , nous

pouvons déduire de (3.25) que  $\gamma \in [\frac{1}{2}, 1]$ . Pour ce cas, le seuil  $u$  est toujours situé entre le premier quartile et la médiane de la distribution hybride et très proche de la moyenne gaussienne. En effet, comme  $\mu \leq u$  nous avons :

$$\frac{\gamma}{2} = H(\mu; \theta) \leq H(u; \theta) = 1 - \gamma \quad (3.26)$$

où le membre de droite de (3.26) est déduit de (3.25). Comme  $\gamma \in [\frac{1}{2}, 1]$  alors (3.26) sera équivalente à (3.27).

$$\frac{1}{4} \leq H(\mu; \theta) \leq H(u; \theta) \leq \frac{1}{2} \quad (3.27)$$

Par conséquent,

$$q_{0.25} \leq \mu \leq u \leq q_{0.5}. \quad (3.28)$$

Cette contrainte, liée à la pondération uniforme, limite l'utilisation du modèle hybride pour  $\gamma_1 = \gamma_2$ . Ainsi, il n'est applicable que sur les données pour lesquelles le seuil est très proche de la moyenne gaussienne ou encore lorsque la probabilité d'être dans la queue de distribution,  $1 - H(u; \theta) = \gamma$ , est entre  $\frac{1}{2}$  et 1 ce qui est un peu paradoxal vu qu'il s'agit d'événements rares.

Cette limitation du seuil est surmontée avec la pondération non-uniforme. En effet, comme nous travaillons avec des données asymétriques à queue droite lourde alors le mode de la densité hybride (3.23), représentant la moyenne gaussienne, est inférieur à sa médiane  $M$  [Willemse, 2009]. Ainsi, nous obtenons :

$$0 \leq \frac{\gamma_1}{2} = H(\mu; \theta) \leq H(M; \theta) = \frac{1}{2}.$$

Il s'ensuit que  $0 \leq \gamma_1 \leq 1$ . Ainsi, d'après l'équation (3.25), il vient que  $0 \leq \gamma_2 \leq 1$ . D'autre part, nous avons  $H(u; \theta) = 1 - \gamma_2$  donc  $\gamma_2$  n'est autre que la probabilité de dépassement du seuil  $u$ . Comme,  $\gamma_2 \in [0, 1]$ , cela implique que

$$q_0 \leq \mu \leq u \leq q_1. \quad (3.29)$$

L'équation (3.29) montre bien que le seuil  $u$  peut être n'importe quelle valeur de  $\mathbb{R}$  pour une pondération non-uniforme ce qui n'est pas le cas lorsque  $\gamma_1 = \gamma_2$ . Un autre avantage du modèle (3.23) est le fait que  $\gamma_2$  représente la probabilité de dépassement du seuil (l'aire bleue de la figure 3.8). Cela offre une meilleure maîtrise de la queue de distribution. Ce travail est bien détaillé dans le papier [Debbabi et al., 2015a] donné dans la page suivante.

Comme nous avons rajouté un paramètre de plus, bien évidemment l'estimation des paramètres sera plus compliquée que pour le cas de la pondération uniforme. Pour avoir une estimation fiable de ces derniers, nous avons adapté l'algorithme 1 au nouveau modèle. En effet, l'algorithme adapté démarre en se basant sur des valeurs initiales bien étudiées et s'arrête lorsqu'une condition d'arrêt que nous fixons est satisfaite. Chaque itération  $k$  est composée de deux étapes principales :

1. évaluation de la gaussienne : les paramètres  $\gamma_1$ ,  $\mu$  et  $\sigma$  sont estimés en minimisant la moyenne quadratique entre la cdf hybride, en considérant l'estimée de  $\gamma_2$  de l'itération  $k - 1$ , et la cdf empirique.



2. évaluation de la GPD : le paramètre  $\gamma_2$  est de nouveau estimé en minimisant également la moyenne quadratique entre la cdf hybride, en considérant cette fois-ci les estimées de  $\gamma_1$ ,  $\mu$  et  $\sigma$  déterminées dans la première étape de cette itération  $k$ , et la cdf empirique.

Nous notons que les paramètres restants :  $\beta$ ,  $\xi$  et  $u$  sont estimés implicitement *via* les relations obtenues par les contraintes imposées sur la pdf  $h$ .

Dans ce même papier, nous avons testé la performance de l'algorithme en utilisant les simulations de Monte Carlo. Comme critère d'évaluation, nous avons considéré le logarithme du rapport de vraisemblance moyen. Ensuite, nous avons mis en pratique le modèle hybride à pondération non-uniforme en l'appliquant sur des données neuronales tout en faisant une comparaison entre les modèles (3.16) et (3.23). À travers cette comparaison, nous avons pu bien voir que le modèle (3.16) n'est applicable que pour le cas où les données présentent un seuil proche de la moyenne gaussienne comme pour le cas des données assurantielles danoises [Carreau and Bengio, 2009b], [Embrechts et al., 1997].

# A HYBRID DISTRIBUTION FOR HIGHLY RIGHT SKEWED DATA

Nehla DEBBABI <sup>1,2</sup>, Marie KRATZ <sup>3</sup>, and SADOK EL ASMI <sup>1</sup>

<sup>1</sup> Research Lab. COSIM, Engineering College of Communications, University of Carthage

<sup>2</sup> CReSTIC, Faculty of Exact and Natural Sciences, University of Reims Champagne Ardenne

<sup>3,\*</sup> ESSEC Business School, CREAR Risk research center

nehla.debbabi@supcom.tn, kratz@essec.edu, elasmi@supcom.rnu.tn

## ABSTRACT

A non-uniform weighted two-components distribution is proposed in the present study for highly right skewed data modeling. We consider a G-GPD model that links a Gaussian distribution to a Generalized Pareto Distribution (GPD) at a junction point, with different weights for each component. It improves a G-GPD model with uniform weights that had been introduced in a preliminary study (see [1]). An iterative algorithm for parameters estimation is then provided, offering an accurate estimation of the Gaussian and GPD parameters, a judicious weighting of the model as well as a reliable position of the junction point, determined successfully in an unsupervised way. The performance of the iterative algorithm and the underlying new distribution, as compared with the existing G-GPD model, is studied on generated data and then on real extracellular neural recordings.

**Index Terms**— Algorithm, Extreme Value Theory (EVT), Gauss-Generalized Pareto distribution (G-GPD), Hybrid models, Right skewed data, Levenberg Marquardt algorithm

## 1. INTRODUCTION

A large number of real phenomenons are observed through a non-homogeneous and multi-components data, containing dissimilar constituents [2–8]. A substantial specific case of this data is given by a highly right skewed data, where, mostly, extreme values point out the presence of relevant information. This type of data may arise in various applications under different contexts, such as in telecommunication for energy of unknown signals over fading channel [9], in insurance when dealing with right skewed claims [10], or in hydrology for observed daily precipitation [11], to name a few. Recent years has gained an increasing interest in modeling such type of data. For this purpose, several models have been proposed in the literature. We cite for instance the non parametric models [1, 12] which provide a good fit of data, but without highlighting the presence of extremes. The Peak Over Threshold (POT) method of the Extreme Value Theory (EVT) is a good alternative for extremes behavior modeling.

Indeed, via the Pickands theorem [13], the extremes above a certain threshold are modeled by a Generalized Pareto Distribution (GPD) [10]. Nonetheless, the POT method allows to model only extremes and not the entire distribution. To go even further in the modeling of highly right skewed data, several hybrid two-components distributions have been developed, taking into account both mean and extremes behaviors of data [1, 4, 8, 14–16].

In this study, we are interested in one of these hybrid or composite models which links a Gaussian distribution to a GPD, noted G-GPD, at a junction point (or threshold) determined when imposing a condition of class  $\mathcal{C}^1$ , as e.g. in [1] and in [17]. Our preliminary work [17] focused on developing an unsupervised iterative and convergent algorithm for the selection of the junction point and for the estimation of the G-GPD parameters, improving already the estimation of the model when compared with other methods, in particular with [1]. Nevertheless, the two components of the distribution were weighted uniformly, with the consequence of forcing the threshold to be located between the first quartile and the median of the data. Hence, for the complete study developed in this paper, we remove this constraint considering a non-uniform weighting and adapt the algorithm developed in [17] to be able to propose a performing G-GPD model, with a better handling of the tail distribution, and an unsupervised way of determining the threshold.

In Section 2, we describe our G-GPD model, where the difference between uniform and non-uniform weighting of the two components is underlined. The development of our algorithm for the model evaluation is carried out in Section 3. This algorithm is iterative and for each iteration the parameters of the Gaussian distribution and the GPD are estimated simultaneously via numerical optimizations in least squares sense, solved using the Levenberg Marquardt (LM) method [18, 19]. Besides the reliable estimation of both components parameters and the judicious weighting of the model, the main advantage of the proposed algorithm is that it allows to estimate successfully such a model in an unsupervised way, unlike standard methods determining the threshold for the esti-

\* M. Kratz is also member of MAP5, UMR8145, Univ. Paris Descartes

mation of the tail of the distribution, which are non automatic (see e.g. [10, 20–23]). Based on Monte Carlo simulations and through the average of the relative log-likelihood criterion, we highlight in Section 4 the effectiveness of the proposed model and algorithm of parameters estimation. We also compare our model with the G-GPD one with uniform weighting [17] on real extracellular neural recordings [24]. Better performance is obtained using our model. Finally, conclusions are drawn in the last section.

## 2. HYBRID MODEL FOR HIGHLY RIGHT SKEWED DATA

We consider a non-uniform weighted G-GPD model. It connects a Gaussian probability density function (pdf) to a GPD one. Unlike for the G-GPD [1, 17], the considered weighting for both components is non-uniform *i.e.*  $\gamma_1 \neq \gamma_2$ , where  $\gamma_1, \gamma_2 \in \mathbf{R}_+$  denote the weights. The proposed hybrid pdf  $h$  is then expressed as

$$h(x; \theta) = \gamma_1 f(x; \mu, \sigma) \mathbb{1}_{\{x \leq u_\theta\}} + \gamma_2 g(x - u_\theta; \xi_\theta, \beta_\theta) \mathbb{1}_{\{x > u_\theta\}}$$

where  $\mathbb{1}$  is the indicator function,  $\theta = [\mu, \sigma, \gamma_1, \gamma_2]$ ,  $\mu$  and  $\sigma$  represent the mean and the standard deviation of the Gaussian pdf  $f$ , expressed in (1), respectively,  $\xi_\theta$  and  $\beta_\theta$  are the index tail and the scale parameter of the GPD pdf  $g$ , given by (2), respectively, and finally,  $u_\theta \geq \mu$  denotes the junction point.

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \forall (x, \mu, \sigma) \in \mathbf{R} \times \mathbf{R} \times \mathbf{R}_+^* \quad (1)$$

$$g(x; \xi, \beta) = \begin{cases} \frac{1}{\beta} (1 + \frac{\xi}{\beta} x)^{-1-\frac{1}{\xi}}, & \text{if } \xi \neq 0 \\ \frac{1}{\beta} e^{-\frac{x}{\beta}}, & \text{if } \xi = 0 \end{cases}, \quad (2)$$

$\forall x \geq 0$  if  $\xi \geq 0$ , and  $x \in [0, -\frac{\beta}{\xi}]$  if  $\xi < 0$ , ( $\xi \in \mathbf{R}, \beta \in \mathbf{R}_+^*$ ).

The corresponding cumulative distribution function (cdf)  $H$  is given by

$$H(x; \theta) = \gamma_1 F(x; \mu, \sigma) \mathbb{1}_{\{x \leq u_\theta\}} + \{\gamma_1 F(u_\theta; \mu, \sigma) + \gamma_2 G(x - u_\theta; \xi_\theta, \beta_\theta)\} \mathbb{1}_{\{x > u_\theta\}}$$

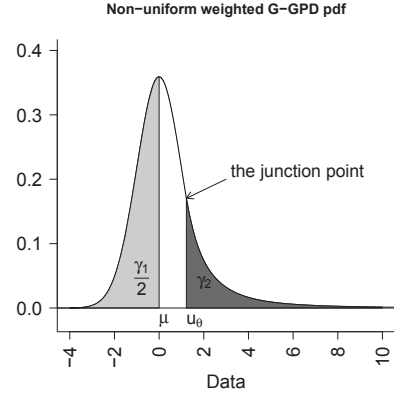
where  $F$  and  $G$  are the Gaussian and GPD cdf, respectively.

In order to obtain a smooth pdf and to reduce the number of unknown parameters, we enforce  $h$  to be a pdf of class  $C^1$ . Three equations interconnecting all model parameters are then generated, decreasing the number of unknown parameters from 7 to 4. Hence,  $h$  can be parameterized by only  $\theta$ , where  $u_\theta, \xi_\theta$  and  $\beta_\theta$  are expressed explicitly in terms of  $\theta$ . The first equation, given by (3), is obtained when forcing  $\int_{\mathbf{R}} h(x; \theta) dx = 1$ , *i.e.*

$$\gamma_1 F(u_\theta; \mu, \sigma) + \gamma_2 = 1 \quad (3)$$

from which we deduce that

$$u_\theta = F^{-1} \left( \frac{1 - \gamma_2}{\gamma_1}; \mu, \sigma \right). \quad (4)$$



**Fig. 1.** The hybrid pdf as a function of  $\gamma_1$  and  $\gamma_2$  with  $\theta = [0, 1, 0.9, 0.2]$ .

Next, the imposition of the continuity and the differentiability of  $h$  at  $u_\theta$  gives rise to the following expressions of  $\xi_\theta$  and  $\beta_\theta$

$$\beta_\theta = \frac{\gamma_2}{\gamma_1 f(u_\theta; \mu, \sigma)}; \quad \xi_\theta = \beta_\theta \frac{u_\theta - \mu}{\sigma^2} - 1. \quad (5)$$

We mention that  $(\gamma_1, \gamma_2) \in [0, 1]^2$ . Indeed, since the skewness of the distribution is positive, the mode of  $h$ , equal to the Gaussian mean, is below its median  $M$  [25]. It implies that  $0 \leq \frac{\gamma_1}{2} = H(\mu; \theta) \leq H(M; \theta) = \frac{1}{2}$ . Consequently,

$$0 \leq \gamma_1 \leq 1. \quad (6)$$

Now, from (3) and (6), we notice that  $0 \leq \gamma_2 \leq 1$ , and  $1 - \gamma_1 \leq \gamma_2$ .

An example of the hybrid pdf  $h$  is drawn in Fig.1, illustrating the physical signification of the weights. The parameter  $\gamma_2$ , representing the dark gray area, is nothing else than the probability of exceeding the threshold  $u_\theta$  while,  $\gamma_1$  representing two times the light gray area, is a regulatory parameter ensuring that  $h$  integrates to 1.

**Remark.** When considering uniform weights  $\gamma = \gamma_1 = \gamma_2$  (as in [1, 17]), the threshold will be located between the first quartile and the median of the G-GPD and close to the Gaussian mean. Indeed, as  $\mu \leq u_\theta$ , we have

$$\frac{\gamma}{2} = H(\mu; \theta) \leq H(u_\theta; \theta) = 1 - \gamma \quad (7)$$

where the RHS of (7) is deduced from (3). We can also easily infer from (3) that  $\gamma \in [\frac{1}{2}, 1]$ . Hence, we can write

$$\frac{1}{4} \leq H(\mu; \theta) \leq H(u_\theta; \theta) \leq \frac{1}{2} = H(M; \theta)$$

which implies that

$$q_{0.25} \leq \mu \leq u_\theta \leq q_{0.5} = M, \quad (8)$$

where  $q_\alpha$  represents the quantile of order  $\alpha$  associated to  $H$ .

When  $\gamma_1 \neq \gamma_2$ , we easily deduce that the junction point of the model is free of constraints and can be any quantile. This is due to  $\gamma_2$  that plays a fundamental role in the handling of the tail distribution. The price to be paid is that the parameters estimation becomes more challenging with seven parameters to be estimated and only three known equations. To overcome such a problem, we propose, in section 3, an iterative algorithm for model evaluation inspired from the one detailed in the preliminary study [17], that we give here for completeness.

### 3. EVALUATION OF THE HYBRID MODEL

For an accurate evaluation of the hybrid model, we propose an unsupervised iterative algorithm that splits the problem of parameters estimation into two principal steps. It evaluates the Gaussian part and the GPD one simultaneously. The algorithm starts by fixing the initial parameters  $\mu^{(0)}$ ,  $\sigma^{(0)}$ ,  $\gamma_1^{(0)}$ ,  $u_\theta^{(0)}$  and  $\varepsilon$  and stops when a convergence criterion is satisfied. In the following, we detail the algorithm where  $X = (X_i)_{1 \leq i \leq n}$  is an  $n$ -sample with a non-uniform weighted G-GPD as parent distribution, and  $\tilde{a}^{(k)}$  denotes the estimate of the parameter  $a$  of the  $k^{th}$  iteration.

**Iteration 1:**

#### Step 1: Evaluation of the Gaussian component

Using only data below  $u_\theta^{(0)}$ , the initial parameter relative to  $u_\theta$ , we determine  $\tilde{\mu}^{(1)}$ ,  $\tilde{\sigma}^{(1)}$  and  $\tilde{\gamma}_1^{(1)}$  minimizing the Squared Error (SE) between the hybrid cdf, when considering  $u_\theta^{(0)}$ , and the empirical cdf  $H_n$  expressed as:

$$H_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}, \quad \forall x \in \mathbb{R}$$

This amounts to the following minimizing problem:

$$(\tilde{\mu}^{(1)}, \tilde{\sigma}^{(1)}, \tilde{\gamma}_1^{(1)}) \leftarrow \underset{\substack{(\mu, \sigma) \in \mathbf{R} \times \mathbf{R}_+ \\ 0 \leq \gamma_1 \leq 1 \\ X_i \leq u_\theta^{(0)}}}{\operatorname{argmin}} \sum_{i=1}^n (\gamma_1 F(X_i; \mu, \sigma) - H_n(X_i))^2$$

which is numerically resolved using the LM algorithm [18, 19], whose initial parameters are  $\mu^{(0)}$ ,  $\sigma^{(0)}$  and  $\gamma_1^{(0)}$  corresponding to  $\mu$ ,  $\sigma$  and  $\gamma_1$  respectively.

#### Step 2: Evaluation of the GPD component

The problem that may be encountered at this stage is that when  $\gamma_2$  and  $n$  are small enough, only few observations exceed  $u_\theta$ . This may lead to unsatisfactory results when evaluating the GPD component. To overcome such a problem, we generate a sequence of synthetic increasing data  $(Y_i)_{1 \leq i \leq p}$ , with a logarithmic step, increasing the number of points above  $u_\theta$ . For any  $0 \leq i \leq p$ ,  $Y_i$  is expressed as

$$Y_i = \min(X) + (\max(X) - \min(X)) \log_{10} \left( 1 + \frac{9(i-1)}{p-1} \right).$$

Thereafter, we determine  $\tilde{\gamma}_2^{(1)}$  minimizing the SE between the hybrid cdf, when considering  $\tilde{\mu}^{(1)}$ ,  $\tilde{\sigma}^{(1)}$ , and  $\tilde{\gamma}_1^{(1)}$ , and the empirical one evaluated on the new data set, as follows:

$$\tilde{\gamma}_2^{(1)} \leftarrow \underset{1 - \tilde{\gamma}_1^{(1)} \leq \gamma_2 \leq 1}{\operatorname{argmin}} \sum_{i=1}^p \left( H(Y_i; [\tilde{\mu}^{(1)}, \tilde{\sigma}^{(1)}, \tilde{\gamma}_1^{(1)}, \gamma_2]) - H_n(Y_i) \right)^2$$

This optimization problem is as well numerically resolved using the LM method, where the initial parameter  $\gamma_2^{(0)}$  is chosen as the middle of the interval  $[1 - \tilde{\gamma}_1^{(1)}, 1]$ .

Steps 1 and 2 are carried out for each iteration with an update of parameters until the convergence of the algorithm. For any  $k > 1$ , the instructions of the  $k^{th}$  iteration are described hereafter.

**Iteration  $k$ :**

#### Step 1: Evaluation of the Gaussian component

$$(\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\gamma}_1^{(k)}) \leftarrow \underset{\substack{(\mu, \sigma) \in \mathbf{R} \times \mathbf{R}_+ \\ 0 \leq \gamma_1 \leq 1 \\ X_i \leq u_{\tilde{\theta}^{(k-1)}}}}{\operatorname{argmin}} \sum_{i=1}^n (\gamma_1 F(X_i; \mu, \sigma) - H_n(X_i))^2. \quad (9)$$

The initial parameters of the LM method for the resolution of (9) are updated as follows:  $\mu^{(0)} = \tilde{\mu}^{(k-1)}$ ,  $\sigma^{(0)} = \tilde{\sigma}^{(k-1)}$  and  $\gamma_1^{(0)} = \tilde{\gamma}_1^{(k-1)}$ .

#### Step 2: Evaluation of the GPD component

$$\tilde{\gamma}_2^{(k)} \leftarrow \underset{1 - \tilde{\gamma}_1^{(k)} \leq \gamma_2 \leq 1}{\operatorname{argmin}} \sum_{i=1}^p \left( H(Y_i; [\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\gamma}_1^{(k)}, \gamma_2]) - H_n(Y_i) \right)^2. \quad (10)$$

Thereafter, we update  $\gamma_2^{(0)}$  for the LM method, used to solve (10). It is chosen as the middle of the interval  $[1 - \tilde{\gamma}_1^{(k)}, 1]$ .

**Convergence criterion:**

The algorithm iterates until it satisfies the following stop condition  $|u_{\tilde{\theta}^{(k)}} - u_{\tilde{\theta}^{(k-1)}}| < \varepsilon$ , where  $\varepsilon$  is small enough. We mention that when  $\varepsilon$  is chosen too small, the algorithm does not stop since the distance between two successive thresholds will never be less than  $\varepsilon$ . For that, we add a supplementary condition to stop the algorithm in that case: when  $k = 50$ . This choice of  $k$  is arbitrary. To get more precision, one can increase it.

We mention that once  $\theta$  is estimated the estimates of  $u_\theta$ ,  $\beta_\theta$  and  $\xi_\theta$  can be computed using relations (4) and (5).

## 4. NUMERICAL RESULTS AND DISCUSSION

### 4.1. Performance of the iterative algorithm

This section discusses the performance of the unsupervised iterative algorithm for the hybrid model parameters estimation, in terms of accurate and reliable fitting of highly right skewed data. To do so, we built on Monte Carlo simulations. We consider  $\{X^q = (X_p^q)_{1 \leq p \leq n}\}_{1 \leq q \leq m}$  and  $\{Y^q = (Y_p^q)_{1 \leq p \leq l}\}_{1 \leq q \leq m}$ , an  $m$  training and test sets, respectively (named also in broadcasting as in-sample and out-of-samples [26], respectively), with non-uniform weighted G-GPD as parent distribution admitting a fixed vector of parameters  $\theta$ . All training sets are of length  $n$  while all test sets are of length  $l$ . On each training set  $X^q$ ,  $1 \leq q \leq m$ , we determine  $\hat{\theta}^q$ , the estimate of  $\theta$ , using our algorithm. Thereafter, we compare the hybrid pdf  $h$  according to the fixed  $\theta$  to the corresponding estimated one  $\tilde{h}$  using  $\hat{\theta}^q$  on each test set  $Y^q$ . To do so, we compute the average of the log-likelihood function  $\mathcal{D}$ , over  $m$  simulations, between  $h(Y^q; \hat{\theta}^q)$  and  $h(Y^q; \theta)$ , whose formula is given by

$$\mathcal{D} = \frac{1}{ml} \sum_{q=1}^m \sum_{p=1}^l \log \left( \frac{h(Y_p^q; \theta)}{\tilde{h}(Y_p^q; \hat{\theta}^q)} \right). \quad (11)$$

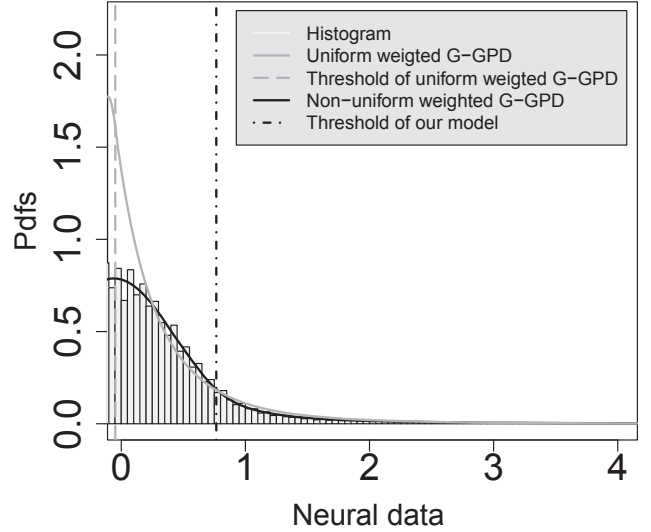
It is obvious that smaller the value of  $\mathcal{D}$  is, more trustworthy our algorithm is. For  $m = 100$ ,  $n = l = 10^5$ , we illustrate in Table 1. the average of log-likelihood function  $\mathcal{D}$  obtained for three different examples:  $\theta_1 = [2, 0.5, 0.9, 0.2]$ ,  $\theta_2 = [0, 1, 0.99, 0.05]$  and  $\theta_3 = [-1, 2, 0.9, 0.2]$ . As it is noticeable in this table, the obtained  $\mathcal{D}$  for the three examples is of order of  $10^{-5}$ , which points out an acceptable similarity between the exact hybrid model and the estimated one using our algorithm.

**Table 1.** Average of the relative log-likelihood  $\mathcal{D}$  of the hybrid model for three different examples

| Examples      | $\theta_1$             | $\theta_2$             | $\theta_3$             |
|---------------|------------------------|------------------------|------------------------|
| $\mathcal{D}$ | $5.4948 \cdot 10^{-5}$ | $8.6263 \cdot 10^{-5}$ | $4.8653 \cdot 10^{-5}$ |

### 4.2. Application on real data

Once the performance of the algorithm is validated on generated data, we apply it on real data and compare it to the G-GPD model with uniform weighting in terms of goodness-of-fit. The data of interest is a real extracellular neural recording of 20 s, available in [27] and measured on the antennal lobe of an adult locust [24]. The information to be extracted from this data lies on the extreme behaviors (left and right) of the data. Since the neural data can be considered as symmetric, it is sufficient to evaluate the right tail of the distribution using observations above the mean (the left one can simply be deduced by symmetry). Calibrating the hybrid model on the data set, we obtain the following estimates for  $\theta$  and the



**Fig. 2.** Pdf estimation of neural data using G-GPD models: with uniform and non-uniform weighting

threshold  $u$ , namely  $\tilde{\theta} = [-0.0543, 0.4828, 0.9539, 0.0887]$  and  $\tilde{u} = 0.7656$ , respectively. The obtained results are illustrated in Fig. 2. As shown in this figure, the neural data distribution is well estimated using the hybrid model. It appears also clearly that it performs much better than the one using uniform weighting. The latter exhibits a threshold very close to the Gaussian mean (whereas it has been introduced to separate mean and tail behaviors). This highlights the benefit of using a non-uniform weighting in the model.

## 5. CONCLUSION

In this paper a non-uniform weighted G-GPD model is proposed for highly right skewed data. An iterative algorithm is then developed to estimate parameters and weights of the hybrid model and to find successfully the junction point. A comparative study on neural data shows that this model outperforms the one with uniform weighting. A natural extension of this model would be to consider different left and right tails.

## REFERENCES

- [1] J. Carreau and Y. Bengio, "A hybrid pareto model for asymmetric fat-tailed data: the univariate case," *Extremes*, vol. 1, pp. 53–76, 2009.
- [2] P. Sermpezis and T. Spyropoulos, "Modelling and analysis of communication traffic heterogeneity in opportunistic networks," *Mobile Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.



- [3] M. Rangaswamy, J. H. Michels, and B. Himed, "Statistical analysis of the non-homogeneity detector for {STAP} applications," *Digital Signal Processing*, vol. 14, no. 3, pp. 253 – 267, 2004.
- [4] V. P. Singh C. Li and A. K. Mishra, "Simulation of the entire range of daily precipitation using a hybrid probability distribution," *Water Resources Research*, vol. 48, pp. 1–17, 2012.
- [5] R. Kollu, S. Rayapudi, S. Narasimham, and K. Pakkurthi, "Mixture probability distribution functions to model wind speed distributions," *International Journal of Energy and Environmental Engineering*, vol. 3, no. 1, pp. 27, 2012.
- [6] F. Rossi, M. Fiorentino, and P. Versace, "Two-component extreme value distribution for flood frequency analysis," *Water Resources Research*, vol. 20, no. 7, pp. 847–856, 1984.
- [7] F. Lagona, A. Maruotti, and M. Picone, "A non-homogeneous hidden markov model for the analysis of multi-pollutant exceedances data," April 2011.
- [8] A. K. Mandava, L. Shahram, and E. E. Regentova, "Reliability assessment of microarray data using fuzzy classification methods: A comparative study," in *Advances in Computing and Communications*. 2011, vol. 190 of *Communications in Computer and Information Science*, pp. 351–360, Springer Berlin Heidelberg.
- [9] F. F. Digham, M. S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," *Communications, IEEE Transactions on*, vol. 55, no. 1, pp. 21–24, 2007.
- [10] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, 1997.
- [11] E. M. Furrer and R. W. Katz, "Improving the simulation of extreme precipitation events by stochastic weather generators," *Water Resources Research*, vol. 44, no. 12, 2008.
- [12] M. Guillén T. Buch-larsen, J. P. Nielsen and C. Bolancé, "Kernel density estimation for heavy-tailed distributions using the champnowne transformation," *Statistics*, vol. 39, no. 6, pp. 503–516, 2005.
- [13] J. Pickands, "Statistical inference using extreme order statistics," *Annals of Statistics*, vol. 3, pp. 119–131, 1975.
- [14] M. Knecht and S. Küttel, "The czeledin distribution function," in *XXXIV ASTIN Colloquium, Berlin*, August 2003.
- [15] S. Nadarajah and S.A.A. Bakar, "New composite models for the danish fire insurance data," *Scandinavian Actuarial Journal*, vol. 2014, no. 2, pp. 180–187, 2014.
- [16] M. Kratz, "Normex, a new method for evaluating the distribution of aggregated heavy tailed risks," *Extremes*, vol. 17, no. 4, pp. 661–691, 2014.
- [17] N. Debbabi and M. Kratz, "A new unsupervised threshold determination for hybrid models," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3440–3444.
- [18] K. Levenberg, "A method for the solution of certain nonlinear problems in least squares quart," *Applied Math*, vol. 2, pp. 164–168, 1944.
- [19] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, pp. 431–441, 1963.
- [20] N. Debbabi, M. Kratz, M. Mboup, and S. El Asmi, "Combining algebraic approach with extreme value theory for spike detection," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 1836–1840.
- [21] B. M. Hill, "A simple general approach to inference about the tail of a distribution," *The Annals of Statistics*, vol. 3, pp. 1163–1174, 1975.
- [22] M. Kratz and S. I. Resnick, "The qq-estimator and heavy tails," *Communications in Statistics. Stochastic Models*, vol. 12, no. 4, pp. 699–724, 1996.
- [23] Y. Demichel, A. Estrade, M. Kratz, and G. Samorodnitsky, "How fast can the chord length distribution decay?," *Advances in Applied Probability*, vol. 43, pp. 504–523, 2011.
- [24] O. Mazar C. Pouzat and G. Laurent, "Using noise signature to optimize spike-sorting and to assess neuronal classification quality," *Journal of neuroscience methods*, vol. 122, no. 1, pp. 43–57, 2002.
- [25] I. Willemse, *Statistical Methods and Calculation Skills*, Juta and Company Ltd, 2009.
- [26] A. Inoue and L. Kilian, "In-sample or out-of-sample tests of predictability: Which one should we use?," *Econometric Reviews*, vol. 23, no. 4, pp. 371–402, 2005.
- [27] C. Pouzat, "Tetrode recording from the antennal lobe of a locust (*schistocerca americana*)," jan 2015.

Toutes ces études ont été effectuées en ne tenant compte que de la queue de distribution à droite. Il est naturel à présent de généraliser ces études en considérant les deux queues de distribution, à gauche et à droite. C'est le cas par exemple des données neuronales. De par la nature de ces données, nous considérerons un modèle hybride à deux queues GPD que nous introduisant dans la dernière partie de ce chapitre.

#### 3.4.4 Modèle hybride à trois composantes GPD-Gauss-GPD avec pondération non-uniforme : résumé du papier [Debbabi et al., 2015b]

Nous donnons dans cette section une introduction au travail [Debbabi et al., 2015b] publié dans la conférence francophone *Groupe d'Etudes du Traitement du Signal et des Images, (GRETSI20015)* que nous présentons dans la page suivante. Dans ce travail, nous avons considéré le modèle hybride à trois composantes suivant :

$$h(x; \theta) = \begin{cases} \gamma_1 g(u_1 - x; \xi_1, \beta_1), & \text{si } x \leq u_1 \\ \gamma_2 f(x; \mu, \sigma), & \text{si } u_1 < x \leq u_2 \\ \gamma_3 g(x - u_2; \xi_2, \beta_2), & \text{si } x > u_2 \end{cases} \quad (3.30)$$

Ici  $\theta = [u_1, u_2, \mu, \sigma, \xi_1, \xi_2]$ ,  $\gamma_1$ ,  $\gamma_2$  et  $\gamma_3$  représentent les poids et  $u_1$  et  $u_2$  sont les deux points de jonction. les deux couples  $(\xi_1, \beta_1)$  et  $(\xi_2, \beta_2)$  représentent respectivement les paramètres de la GPD de la queue gauche et de la queue droite.

Nous illustrons un exemple de la pdf hybride (3.30) pour  $\theta = [-1.5, 1.5, 0, 1, 0.2, 0.5]$ . Nous voyons bien à travers cette figure que les poids  $\gamma_1$  et  $\gamma_3$  jouent le rôle respectivement de la probabilité d'être dans la queue gauche (aire rouge) et la probabilité d'être dans la queue droite (aire bleue).

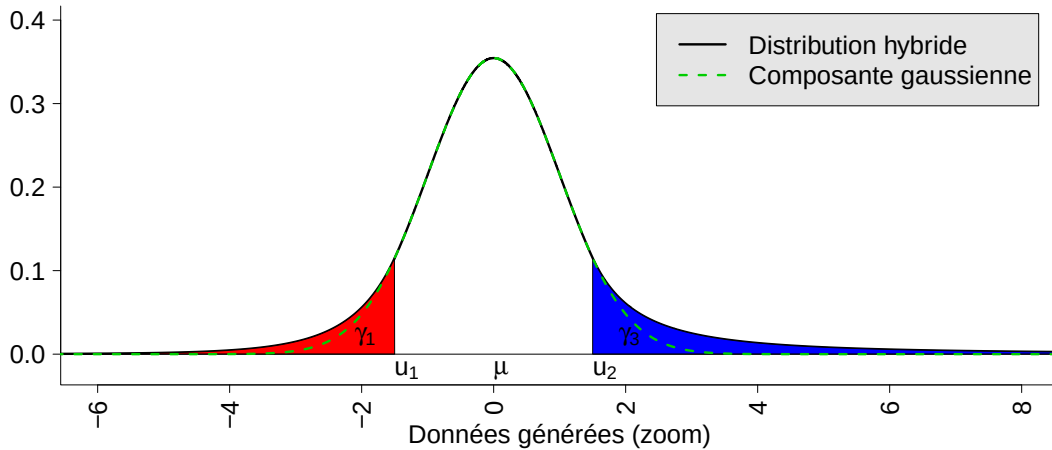


FIGURE 3.8 – Exemple de pdf hybride GPD-Gauss-GPD pour  $\theta = [-1.5, 1.5, 0, 1, 0.2, 0.5]$ .

Nous notons que ce modèle à été proposé dans [Gehin, 2011] mais avec une pondération uniforme. Pour la même raison donnée pour le modèle à deux composantes, nous proposons ici ce modèle pour des poids de valeurs différentes. Ce modèle, décrit dans le papier [Debbabi et al., 2015b], présenté dans la page suivante, admet 11 paramètres. Comme pour

les autres modèles que nous avons développés, nous imposons une condition de classe  $\mathcal{C}^1$  aux points de jonction  $u_1$  et  $u_2$ . Ainsi, avec l'équation  $\int_{\mathbb{R}} h(x; \theta) = 1$  nous réduisons le nombre de paramètres à estimer à 6 ce qui reste quand même un nombre élevé rendant leur estimation plus difficile. Encore une fois, nous allons nous servir de notre fameux algorithme en l'adaptant au modèle (3.30) pour estimer les paramètres de ce dernier. Une application de ce modèle est illustrée dans l'étude ci-dessous [Debbabi et al., 2015b] où nous voyons bien l'ajustement judicieux des données neuronales par le modèle hybride à trois composantes.



# Distribution hybride pour la modélisation de données à deux queues lourdes: Application sur les données neuronales

Nehla DEBBABI<sup>1,2</sup>, Marie KRATZ<sup>3</sup>, Sadok EL ASMI<sup>2</sup>, Mamadou MBOUP<sup>1,4</sup>

<sup>1</sup>CRESTIC, Faculté des Sciences Exactes et Naturelles, Université de Reims Champagne Ardenne, France

<sup>2</sup>Laboratoire COSIM, École Supérieure des Communications de Tunis, Université de Carthage, Tunisie

<sup>3</sup>ESSEC Business School, CEAR risk research center, Cergy-Pontoise, France

<sup>4</sup>Non-A, Inria Lille Nord Europe

Nehla.Debbabi@supcom.tn, Kratz@essec.edu  
Elasmi@supcom.rnu.tn, Mamadou.Mboup@univ-reims.fr

**Résumé** – Ce travail propose un modèle hybride pour modéliser des données à deux queues lourdes. Le modèle proposé est une distribution à trois composantes pondérées : une distribution Gaussienne, pour modéliser le comportement moyen des données, liée à deux distributions de Pareto généralisées pour modéliser les comportements extrêmes. Un algorithme itératif et non supervisé est ensuite proposé pour une estimation fiable des points de jonctions entre les trois distributions, les paramètres de ces dernières ainsi que les poids affectés à chaque composante du modèle hybride. Une application sur des données neuronales réelles issues d'un enregistrement extracellulaire, est développée pour évaluer les performances du modèle proposé, comparé à la distribution normale.

**Abstract** – A new hybrid model for two heavy tailed data modelling is proposed in this study. The proposed model is a weighted three-components distribution: a Gaussian distribution, to model the mean behavior of the data, linked to two generalized Pareto distributions, modelling the extreme ones. An unsupervised iterative algorithm is then developed to estimate accurately the junction points between the three distributions, the parameters of these latter as well as the weights of the hybrid model. An application on real extracellular neural recordings is developed to evaluate the performance of the proposed hybrid model, compared to the normal distribution.

## 1 Introduction

Les données d'observation de phénomènes complexes sont généralement non-homogènes et multi-composantes [1, 2, 3]. Trouver un modèle adéquat offrant le meilleur ajustement de ce type de données est un sujet d'importance majeure, qui intéresse les chercheurs depuis des décennies [4, 5]. Dans le présent travail, nous abordons ce sujet pour un cas particulier de données à deux queues lourdes. Ce type de données est rencontré dans différents domaines où les queues de distribution, contenant des valeurs extrêmes, portent souvent une information importante. À titre d'exemple, on cite le cas des rendements de l'indice S&P 500 en finance [6], où les valeurs extrêmes indiquent un risque du marché, ou encore le cas des données neuronales issues d'un enregistrement extracellulaire en neuroscience [7], où les extrêmes indiquent la manifestation de Potentiels d'Action (ou spikes). Modéliser ce type de données directement par une gaussienne ne suffit pas : les deux queues sont sous-estimées [6, 8]. Les méthodes non paramétriques représentent une bonne alternative mais ne donnent aucune information sur les extrêmes. Nous privilégions pour cette raison les méthodes paramétriques comme le mélange de noyaux gaussiens, la distribution q-gaussienne ou encore la distribution de

Pareto généralisée (GPD). Notons que notre objectif n'est pas seulement d'ajuster les données par une distribution adéquate mais c'est aussi, et surtout, de pouvoir interpréter les extrêmes. Une façon de le faire est d'exploiter le théorème de Pickands [9] de la Théorie des Valeurs Extrêmes (TVE) en modélisant les extrêmes au delà d'un seuil bien choisi par une GPD [10, 11]. Cependant, et contrairement aux autres méthodes, la TVE ne permet de bien estimer que les extrêmes mais pas toute la distribution. Pour garder l'avantage de la TVE mais avec une modélisation complète, nous proposons dans ce travail de séparer le comportement moyen des extrêmes. En effet, comme dans [6], nous proposons de modéliser le comportement moyen des données par une gaussienne *via* le Théorème Central Limite (TCL), et les comportements extrêmes par deux GPD différentes *via* le théorème de Pickands [9]. Pour un meilleur ajustement des données, et pour plus de souplesse au niveau des queues, contrairement à [6], nous attribuons un poids différent à chaque morceau. Les trois morceaux sont connectés entre eux en deux points de jonctions, où on impose des conditions de classe  $\mathcal{C}^1$ . Ces conditions permettent non seulement d'avoir une fonction de densité de probabilité (pdf) lisse mais aussi de réduire le nombre de paramètres à estimer. Pour une estimation fiable des paramètres du modèle hybride ainsi obtenu, nous

proposons un algorithme itératif non supervisé, basé sur le principe de l'algorithme introduit dans [12]. Pour chaque itération, l'algorithme proposé estime les paramètres du modèle hybride en résolvant des problèmes d'optimisation au sens des moindres carrés en utilisant la méthode de Levenberg Marquardt [13, 14]. Nous mentionnons qu'à part l'estimation fiable des paramètres du modèle, l'algorithme est non supervisé permettant d'éviter les problèmes rencontrés par les méthodes existantes de détermination du seuil satisfaisant le théorème de Pickands, qui sont graphiques et non automatiques [11, 10, 15]. Pour évaluer les performances du modèle proposé, une application sur des signaux neuronaux issus d'un enregistrement extracellulaire est développée. Les résultats obtenus montrent que le modèle hybride fournit un bon ajustement des données neuronales, comparé à la distribution normale.

Le reste du papier est organisé comme suit : La première section décrit le modèle hybride proposé. Le pseudo-code de l'algorithme itératif est donné dans la section 3. Les résultats de simulations sont discutés dans la section 4 et la conclusion de ce travail est donnée à la fin.

## 2 Modèle hybride proposé

La distribution hybride proposée dans ce travail est une distribution gaussienne dont les queues, gauche et droite, ont été remplacées par deux GPD. La densité de probabilité correspondante est ainsi la fonction par morceaux définie par :

$$h(x; \theta) = \gamma_1 g(u_1 - x; \xi_1, \beta_1) \mathbb{1}_{\{x \leq u_1\}} + \gamma_2 f(x; \mu, \sigma) \mathbb{1}_{\{x \in ]u_1, u_2\}} + \gamma_3 g(x - u_2; \xi_2, \beta_2) \mathbb{1}_{\{x > u_2\}}. \quad (1)$$

où  $\mathbb{1}_{\{\cdot\}}$  est la fonction indicatrice,  $u_1$  et  $u_2$  représentent les points de jonction ou encore les seuils à partir desquels la distribution change de comportement et  $\theta = [u_1, u_2, \mu, \sigma, \xi_1, \xi_2]$ . Les paramètres  $\gamma_1$ ,  $\gamma_2$  et  $\gamma_3$  désignent les poids et  $g$  est la pdf de la GPD, donnée par (2) de paramètres  $\xi_1$  et  $\beta_1$  pour la queue gauche (resp.  $\xi_2$  et  $\beta_2$  pour la queue droite). Enfin,  $f(\cdot; \mu, \sigma)$  est la densité de la loi normale de moyenne  $\mu$  et d'écart type  $\sigma$ .

$$g(x; \xi, \beta) = \begin{cases} \frac{1}{\beta} (1 + \frac{\xi}{\beta} x)^{-1-\frac{1}{\xi}}, & \text{si } \xi \neq 0 \\ \frac{1}{\beta} e^{-\frac{x}{\beta}}, & \text{si } \xi = 0 \end{cases}, \quad (2)$$

$$x \in \begin{cases} [0, \infty), & \text{si } \xi \geq 0 \\ [0, -\frac{\beta}{\xi}], & \text{si } \xi < 0 \end{cases}, \text{ et } (\xi, \beta) \in \mathbb{R} \times \mathbb{R}_+^*.$$

où  $\xi$  désigne l'indice de queue et  $\beta$  le paramètre d'échelle. Puisque  $h$  est une pdf, sa masse totale doit être égale à 1. Cette contrainte se traduit par la relation

$$\gamma_1 + \gamma_2 [F(u_2; \mu, \sigma) - F(u_1; \mu, \sigma)] + \gamma_3 = 1, \quad (3)$$

où  $F$  désigne la fonction de répartition (cdf) de la loi normale. Dans le but d'avoir une pdf lisse, on impose des conditions de continuité et de dérivabilité aux points de jonction. Ces conditions permettent non seulement de réduire le nombre de paramètres à estimer mais aussi, une fois les seuils sont déterminés, de pouvoir les affiner empiriquement (voir section 4.2).

Les relations ainsi obtenues aux points  $u_1$  et  $u_2$  avec (3) permettent d'exprimer explicitement les paramètres  $\beta_1$ ,  $\beta_2$ ,  $\gamma_1$ ,  $\gamma_2$  et  $\gamma_3$  en fonction du reste des paramètres comme suit

$$(E) \begin{cases} \beta_1 = -\frac{(1 + \xi_1)\sigma^2}{u_1 - \mu}; & \beta_2 = \frac{(1 + \xi_2)\sigma^2}{u_2 - \mu} \\ \gamma_2 = [\beta_1 f(u_1; \mu, \sigma) + \beta_2 f(u_2; \mu, \sigma) + \int_{u_1}^{u_2} f(x; \mu, \sigma) dx]^{-1} \\ \gamma_1 = \beta_1 \gamma_2 f(u_1; \mu, \sigma); & \gamma_3 = \beta_2 \gamma_2 f(u_2; \mu, \sigma) \end{cases}$$

La cdf de la loi hybride est donnée par :

$$H(x; \theta) = \gamma_1 [1 - G(u_1 - x; \xi_1, \beta_1)] \mathbb{1}_{\{x \leq u_1\}} + [\gamma_1 + \gamma_2 [F(x; \mu, \sigma) - F(u_1; \mu, \sigma)]] \mathbb{1}_{\{x \in ]u_1, u_2\}} + [1 - \gamma_3 [1 - G(x - u_2; \xi_2, \beta_2)]] \mathbb{1}_{\{x > u_2\}} \quad (4)$$

où  $G$  représente la cdf de la GPD.

Nous mentionnons qu'en se basant sur (3) et (4), il est facile de démontrer que  $(\gamma_1, \gamma_3) \in [0, 1]^2$  et que  $\gamma_2 \geq 0$ .

Un exemple de la pdf hybride est donné par la figure 1. Cette figure illustre bien la différence entre les queues de la pdf hybride (en continu) et celles de sa composante normale (en pointillé) ainsi que la signification physique des coefficients de pondération. Comme il est montré dans cette dernière, le paramètre  $\gamma_1$  (resp.  $\gamma_3$ ), représentant l'aire gris-clair (resp. l'aire gris-foncé), n'est autre que la probabilité que la population soit en dessous du seuil  $u_1$  (resp. au delà du seuil  $u_2$ ). En revanche,  $\gamma_2$ , représentant deux fois l'aire blanche, est un paramètre régulateur assurant que  $h$  est une pdf.

Notons que le modèle hybride proposé généralise celui introduit dans [6]. En effet, ce dernier utilise la même pondération  $\gamma$  pour les trois distributions *c-à-d*  $\gamma_1 = \gamma_2 = \gamma_3 = \gamma$ . Dans ce cas,  $H(u_1; \theta) = \gamma$  et  $H(u_2; \theta) = 1 - \gamma$  ce qui implique que les deux seuils doivent être symétriques par rapport à la médiane pour pouvoir appliquer ce modèle. Une autre limitation de la pondération uniforme est que d'après (3),  $\frac{1}{3} \leq \gamma \leq \frac{2}{3}$  *c-à-d* la probabilité de se trouver dans les queues doit appartenir à  $[\frac{1}{3}, \frac{2}{3}]$ , sinon le modèle ne s'applique plus. Toutes ces limitations sont surmontées avec la pondérations non uniforme, considérée dans ce travail, où les seuils sont sans contraintes.

Afin d'évaluer le modèle hybride, nous proposons dans la section suivante un algorithme itératif non supervisé qui donne une estimation de tous les paramètres du modèle.

## 3 Algorithme itératif pour l'évaluation du modèle hybride proposé

L'algorithme proposé est inspiré de celui décrit dans [12]. Il est non supervisé et itératif. Pour chaque itération les paramètres du modèle sont estimés en deux étapes : 1) estimer au sens des moindres carrés les paramètres de la gaussienne et des deux indices de queue des GPD en considérant les deux seuils

trouvés dans l'itération précédente, 2) fixer les paramètres trouvés dans 1) et estimer au sens des moindres carrés les deux seuils. L'algorithme proposé, dont le pseudo-code est donné par l'algorithme 1, commence par un choix arbitraire de  $u_1$  et  $u_2$  et itère jusqu'à ce qu'une précision satisfaisante ou un nombre maximal  $k_{max}$  d'itérations soit atteint. Une fois  $\theta$  est estimée, le reste des paramètres est déterminé via le système (E).

---

**Algorithme 1** Pseudo-code de l'algorithme itératif proposé

---

- 1: Détermination de la fonction de répartition empirique  $H_n$  du n-échantillon  $X = (X_i)_{1 \leq i \leq n}$

$$H_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}, \quad \forall x \in \mathbb{R}.$$

- 2: Fixer les paramètres initiaux  $\tilde{u}_1^{(0)}, \tilde{u}_2^{(0)}$ , et  $\epsilon$   
 3: Procédure itérative

$k \leftarrow 1$

**tant que**  $((\tilde{u}_2^{(k)} - \tilde{u}_2^{(k-1)})^2 + (\tilde{u}_1^{(k)} - \tilde{u}_1^{(k-1)})^2) \geq \epsilon$  et  $k < k_{max}$

- a. Détermination de  $\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\xi}_1^{(k)}$  et  $\tilde{\xi}_2^{(k)}$  les estimateurs, respectivement, de  $\mu, \sigma, \xi_1$  et  $\xi_2$

$$(\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\xi}_1^{(k)}, \tilde{\xi}_2^{(k)}) \leftarrow \underset{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+}{\underset{(\xi_1, \xi_2) \in \mathbb{R}^2}{\operatorname{argmin}}} \|H(X; \theta_1^{(k)}) - H_n(X)\|_2^2$$

$$\text{où } \theta_1^{(k)} = [u_1^{(k-1)}, u_2^{(k-1)}, \mu, \sigma, \xi_1, \xi_2].$$

- b. Détermination de  $\tilde{u}_1^{(k)}$  et  $\tilde{u}_2^{(k)}$  les estimateurs, respectivement, de  $u_1$  et  $u_2$

$$(\tilde{u}_1^{(k)}, \tilde{u}_2^{(k)}) \leftarrow \underset{(u_1, u_2) \in \mathbb{R} \times \mathbb{R}_+}{\operatorname{argmin}} \|H(X; \theta_2^{(k)}) - H_n(X)\|_2^2$$

$$\text{où } \theta_2^{(k)} = [u_1, u_2, \tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\xi}_1^{(k)}, \tilde{\xi}_2^{(k)}]$$

$k \leftarrow k + 1$

**fin**

**retour**  $\theta^{(k)} = [\tilde{u}_1^{(k)}, \tilde{u}_2^{(k)}, \tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{\xi}_1^{(k)}, \tilde{\xi}_2^{(k)}]$

---

Notons que tous les problèmes d'optimisation de cet algorithme sont résolus par la méthode de Levenberg Marquardt [13, 14].

## 4 Résultats numériques et discussions

Dans cette section, nous étudions les performances de l'algorithme et du modèle hybride proposés, en terme de fiabilité de l'estimation des paramètres, du bon ajustement des données ainsi que la séparation du comportement moyen des extrêmes.

### 4.1 Étude de performances de l'algorithme itératif

Les performances de l'algorithme sont étudiées à partir de simulations de Monte Carlo. En effet, nous simulons  $p = 100$

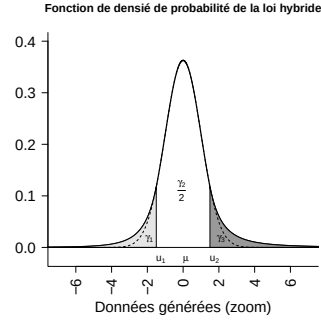


FIGURE 1 – Fonction de densité de probabilité hybride pour  $\theta = [-1.5, 1.5, 0, 1, 0.2, 0.5]$ .

séquences de test  $\{X^{(k)} = (X_i^{(k)})_{1 \leq i \leq n}\}_{k=1}^p$  (resp. d'apprentissage  $\{Y^{(k)} = (Y_i^{(k)})_{1 \leq i \leq n}\}_{k=1}^p$ ) de taille  $n = 10^5$  et suivant la loi hybride de paramètre  $\theta$  fixé. Pour tout  $k \in \{1, \dots, p\}$ , nous déterminons  $\tilde{\theta}^{(k)}$ , l'estimateur de  $\theta$  par l'algorithme itératif sur  $\{Y^{(k)}\}$ . Ensuite, nous déterminons sur les séquences de test le logarithme du rapport de vraisemblance moyen,  $\mathcal{C}$  :

$$\mathcal{C} = \frac{1}{np} \sum_{k=1}^p \sum_{i=1}^n \log \left( \frac{h(X_i^{(k)}; \theta)}{h(X_i^{(k)}; \tilde{\theta}^{(k)})} \right) \quad (5)$$

Pour un exemple de  $\theta = [-1, 1, 0, 1, 0.2, 0.2]$ , les résultats obtenus pour  $n = 10^5$  sont consignés dans le tableau 1. Ces résultats montrent un absence de biais asymptotique. La très faible valeur correspondante de  $\mathcal{C} = 6.11 \cdot 10^{-5}$  suggère des performances intéressantes.

TABLE 1 – Estimation des paramètres du modèle hybride

| $\theta$                               | -1    | 1    | 0     | 1    | 0.2  | 0.2  |
|----------------------------------------|-------|------|-------|------|------|------|
| $\mu_{\theta_j}$                       | -1.01 | 1.01 | -0.00 | 1.00 | 0.20 | 0.20 |
| $\sigma_{\theta_j}^2 (\times 10^{-4})$ | 9.82  | 7.45 | 0.53  | 0.82 | 1.91 | 1.56 |

où  $\mu_{\theta_j}$  et  $\sigma_{\theta_j}^2$  désignent respectivement la moyenne et la variance empirique des estimations de  $\theta_j$ , la composante  $j$  de  $\theta$ ,  $j \in \{1, \dots, 6\}$ .

### 4.2 Étude de performance du modèle hybride

Motivés par le problème de détection de spikes de neurones, nous avons choisi d'appliquer le modèle hybride sur des données neuronales réelles issues d'un enregistrement extracellulaire. Ces données représentent 4 enregistrements de 20 s chacun, filtrés entre 300 Hz et 5000 Hz, et mesurés à partir du lobe antennaire d'un criquet [16]. Nous présentons dans la figure 2, graphe A, 1 s du premier enregistrement où les pics représentent des spikes. L'histogramme relatif à ces données est tracé dans le graphe B, avec la pdf hybride ajustée (ligne rouge continue), dont l'estimée de  $\theta$  est donnée par  $\tilde{\theta} = [-0.7792, 0.7241, -0.0288, 0.4636, 0.5082, 0.6251]$ , ainsi

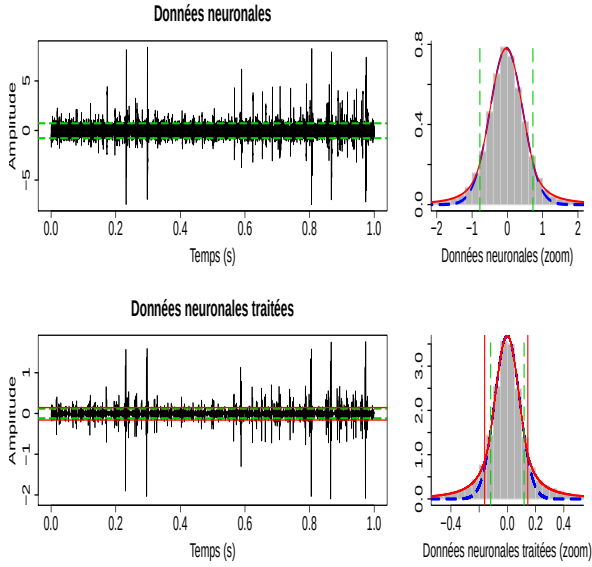


FIGURE 2 – Estimation de la pdf par la loi hybride et la loi normale sur des données neuronales réelles

que sa composante normale (ligne bleue discontinue). Nous remarquons une différence significative entre les deux pdfs au niveau des queues de distributions montrant que les données sont mieux ajustées avec le modèle hybride qu'avec la normale. De plus, les deux seuils du modèle hybride trouvés (représentés par des droites vertes discontinues dans tous les graphes de la figure 2) montrent bien qu'ils séparent les spikes du bruit. Comme il est montré dans le graphe A, le signal neuronal est bruité. Afin de réduire la puissance de ce bruit et rendre les spikes plus visibles, nous le filtrons par un filtre dérivateur, qui est un filtre algébrique linéaire [7]. La linéarité du filtre conserve la gaussianité du comportement moyen ce qui permettra d'ajuster aussi les signaux filtrés. Le graphe C montre le résultat du filtrage du signal représenté dans A. L'ajustement de ce dernier est illustré dans le graphe D. À travers les graphes C et D, nous remarquons un réhaussement des spikes par rapport au bruit ainsi qu'un bon ajustement de données (la différence entre la pdf hybride et la normale est devenue plus nette).

La figure 2 montre que le modèle hybride ajuste bien les données et sépare bien les spikes du bruit. Par contre, nous remarquons dans les graphes A et C la présence d'une zone intermédiaire entre les spikes et le bruit, où la distinction entre ces deux derniers est difficile. Afin d'avoir plus d'extrêmes que du bruit, nous proposons d'affiner les seuils d'une façon empirique, en tolérant une marge d'erreur  $d$  définie à partir de (6). Ceci est possible puisque  $h$  est lisse.

$$Err(x) = |h(x; \theta) - \gamma_2 f(x; \mu, \sigma)|, \quad \forall x \leq u_1 \text{ et } x \geq u_2. \quad (6)$$

Les seuils affinés,  $\tilde{u}_1$  et  $\tilde{u}_2$ , relatifs respectivement à  $u_1$  et  $u_2$ , sont déterminés par :

$$\begin{aligned} \tilde{u}_1 &= \min_{x \leq u_1} |u_1 - \{x/Err(x) \geq d\}| \\ \tilde{u}_2 &= \min_{x \geq u_2} |u_2 - \{x/Err(x) \geq d\}| \end{aligned}$$

Pour  $d = \frac{1}{4} \max(Err)$ , nous traçons dans le graphe C et D les deux seuils  $\tilde{u}_1$  et  $\tilde{u}_2$  (droites rouges continues) relatifs aux estimateurs de  $u_1$  et  $u_2$ . Nous voyons bien que les deux seuils affinés ne changent pas trop l'allure de la pdf par contre ils permettent d'éliminer plus de bruit dans le signal neuronal.

## 5 Conclusion

Dans ce travail, nous avons proposé un modèle hybride pour des données à deux queues lourdes. Pour évaluer ce modèle, un algorithme itératif non supervisé a été développé et testé sur des données simulées. Les performances du modèle hybride ont été étudiées sur des données neuronales réelles, brutes et filtrées. Le modèle proposé a montré des résultats satisfaisants en terme d'ajustement et de séparation du comportement moyen, représentant le bruit, des extrêmes. Pour séparer encore davantage les deux comportements, un affinement empirique des deux points de jonction a été proposé pour réduire la zone intermédiaire, transitoire entre les comportements moyen et extrêmes. Une étude théorique de ce comportement intermédiaire ainsi qu'une étude de la convergence de l'algorithme feront l'objet de développements futurs.

## Références

- [1] A. K. Mandava, L. Shahram, and E. E. Regentova, "Reliability assessment of microarray data using fuzzy classification methods : A comparative study," in *Advances in Computing and Communications*, 2011, vol. 190.
- [2] P. Sermpezis and T. Spyropoulos, "Modelling and analysis of communication traffic heterogeneity in opportunistic networks," *Mobile Computing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.
- [3] V. P. Singh C. Li and A. K. Mishra, "Simulation of the entire range of daily precipitation using a hybrid probability distribution," *Water resources research*, vol. 48, pp. 1–17, 2012.
- [4] M.S. Sambridge and W. Compston, "Mixture modeling of multi-component data sets with application to ion-probe zircon ages," *Earth and Planetary Science Letters*, vol. 128, no. 3–4, pp. 373 – 390, 1994.
- [5] M. Rangaswamy, J. H. Michels, and B. Himed, "Statistical analysis of the non-homogeneity detector for {STAP} applications," *Digital Signal Processing*, vol. 14, no. 3, pp. 253 – 267, 2004.
- [6] W. Gehin, "Modélisation des queues de distribution des rendements des actif financiers," M.S. thesis, EURIA, Brest, 2011.
- [7] M. Mboup, "Neural spike detection and localisation via volterra filtering," in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, Sept 2012, pp. 1–5.
- [8] Zoran Tiganj and Mamadou Mboup, "A non-parametric method for automatic neural spike clustering based on the non-uniform distribution of the data," *Journal of neural engineering*, vol. 8, no. 6, pp. 066014, 2011.
- [9] J. Pickands, "Statistical inference using extreme order statistics," *Annals of Statistics*, vol. 3, pp. 119–131, 1975.
- [10] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, 1997.
- [11] N. Debbabi, M. Kratz, M. Mboup, and S. El Asmi, "Combining algebraic approach with extreme value theory for spike detection," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 1836–1840.
- [12] N. Debbabi and M. Kratz, "A new unsupervised threshold determination for hybrid models," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3440–3444.
- [13] K. Levenberg, "A method for the solution of certain nonlinear problems in least squares quart," *Applied Math*, vol. 2, pp. 164–168, 1944.
- [14] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, pp. 431–441, 1963.
- [15] Y. Demichel, A. Estrade, M. Kratz, and G. Samorodnitsky, "How fast can the chord length distribution decay ?," *Advances in Applied Probability*, vol. 43, pp. 504–523, 2011.
- [16] O. Mazor C. Pouzat and G. Laurent, "Using noise signature to optimize spike-sorting and to assess neuronal classification quality," *Journal of neuroscience methods*, vol. 122, no. 1, pp. 43–57, 2002.

### 3.5 Conclusion

Nous arrivons maintenant à la fin de ce chapitre qui était divisé en trois principales parties. La première partie était dédiée à l'introduction de la TVE, l'approche sur laquelle nous nous sommes basés tout au long de ce chapitre pour modéliser les ruptures dont l'apparition a été caractérisée par un événement rare lié à une valeur extrême. La deuxième partie de ce chapitre a été consacrée à la modélisation de ces événements rares en se basant sur la méthode du POT de la TVE qui nous permet d'approcher la loi des excès au-delà d'un seuil par une GPD. Ce seuil a été déterminé dans cette partie graphiquement via la méthode du '*mean excess plot*'. Enfin, la dernière partie a rassemblé tout ce que nous avons développé dans cette thèse, en modélisant toutes les données comprenant le bruit et les ruptures par des lois hybrides séparant ces deux composantes. Ces lois hybrides ont permis non seulement le bon ajustement des données mais aussi nous ont aidé à déterminer le seuil de la POT d'une manière automatique non-supervisée. Nous avons vu à travers ce chapitre l'évolution de l'extraction de l'information en modélisant à chaque fois une partie des données. Ces techniques n'ont pas encore été appliquées sur des données traitées par l'approche algébrique où le comportement moyen des données traitées par cette dernière n'est pas gaussien comme pour le cas où le bruit présente une somme pondérée de khi-deux. Ainsi, nous envisageons dans les travaux futurs de proposer des lois hybrides pour un comportement moyen non gaussien. Aussi bien, nous envisageons de considérer le comportement intermédiaire apparant entre le comportement moyen et extrêmes dans le but d'améliorer encore et encore la détection des ruptures.

# Conclusion générale

L'objectif de cette thèse était de développer des méthodes non-supervisées de détection et de localisation en ligne des ruptures dans les signaux enregistrés dans un environnement bruité. Afin d'assurer une détection fiable de ruptures, l'idée était de proposer des méthodes qui exploitent au maximum les informations que l'on peut avoir sur les deux aspects du signal enregistré : déterministe et aléatoire. Pour répondre à cet objectif, notre démarche était de caractériser, dans un cadre déterministe, la rupture comme une discontinuité dans le signal, ensuite, de compléter cette caractérisation par une interprétation probabiliste en terme d'extrêmes dans un cadre stochastique : une rupture est un événement rare et extrême.

En se basant sur ces deux caractérisations différentes de la rupture, nous avons développé différentes méthodes non-supervisées de détection et de localisation en-ligne de ruptures. L'évolution de ces méthodes était basée sur l'amélioration de la détection tout en incorporant à chaque fois une nouvelle information sur les données traitées. En effet, nous avons décomposé le signal bruité à traiter  $y(t)$ ,  $t \in \mathbb{R}_+$ , en deux composantes :  $y(t) = x(t) + \omega(t)$ , où  $x(t)$  désigne la composante déterministe du signal contenant les ruptures et  $\omega(t)$  représente sa composante aléatoire décrivant le bruit. Nous nous sommes ensuite focalisés sur la modélisation de  $x(t)$  et/ou  $\omega(t)$ .

Dans un premier temps et dans un cadre déterministe, nous avons envisagé la rupture comme une discontinuité dans le signal. Nous ne nous sommes ensuite intéressés qu'à la modélisation des ruptures ou encore de la composante déterministe du signal. Nous avons ainsi associé à la composante  $x(t)$  un modèle régulier par morceaux faisant apparaître les instants d'occurrence des ruptures. Pour souligner la présence des discontinuités dans le signal, nous avons adopté une approche algébrique permettant d'exprimer explicitement les instants d'occurrence des ruptures via l'algèbre différentielle et le calcul opérationnel. De cette approche algébrique nous avons déduit différents détecteurs algébriques basés sur l'utilisation d'intégrales itérées. Ces détecteurs algébriques sont implémentés comme la sortie d'une combinaison (linéaire ou non-linéaire) de filtres passe-bas. Bien que l'approche algébrique est une approche purement déterministe : aucun modèle du bruit n'a été utilisé dans sa construction, l'utilisation des filtres passe-bas dans ses détecteurs lui a offert une bonne robustesse face au bruit. Pour mettre en évidence les performances de ces détecteurs algébriques, nous les avons comparés entre eux en se basant sur les courbes ROC, ensuite, nous avons appliqué celui qui offre la meilleure performance en terme de probabilité de détection et probabilité de fausses alarmes pour détecter les spikes de neurones. Une deuxième application d'un détecteur algébrique pour la détection de pics R dans les signaux ECG a aussi fait preuve de la robustesse de l'approche déterministe utilisée.

Arrivons à ce niveau, nous avons pensé à améliorer la détection de rupture tout en considérant l'information que l'on peut avoir sur la composante aléatoire du signal issu d'un traitement algébrique. Nous avons ainsi associé à ce dernier un modèle stochastique. Cette connaissance de loi du bruit nous a amené à déterminer un seuil assurant une détection de ruptures avec une probabilité de fausses alarmes fixée, améliorant ainsi les résultats de détection. La modélisation séparée des deux composantes du signal (ruptures et bruit) nous a motivé à voir ce qui se passe en modélisant ces deux dernières en même temps.

Dans un second temps et dans un cadre stochastique, nous avons donné une nouvelle caractérisation à la rupture. Nous l'avons caractérisé comme un événement rare, et extrême. Tout au début, nous nous sommes intéressés à modéliser les extrêmes vu l'importance de l'information qu'elles fournissent. Pour ce faire, nous nous sommes basés sur la méthode du POT de la TVE pour modéliser les excès au delà d'un seuil, représentant les ruptures, par une loi de Pareto généralisée. Dans une première application de la méthode POT, nous avons déterminé le seuil, indiquant la présence des extrêmes, graphiquement en utilisant la méthode du *mean excess plot*. La connaissance de la loi des extrêmes nous a permis de proposer un deuxième seuil séparant les ruptures des fausses alarmes avec une probabilité donnée. Les résultats obtenus en combinant l'approche algébrique et la TVE : modéliser les extrêmes du signal après un traitement algébrique, ont montré une amélioration de détection lors d'une application pour la détection de spikes de neurones.

La connaissance de la loi des extrêmes nous a permis de mieux cerner les ruptures ce qui nous a poussé à étendre la modélisation à toutes les données en considérant les comportements moyen et extrêmes. Nous avons ainsi proposé différents modèles hybrides composés par deux ou plusieurs composantes différentes. Le premier modèle que nous avons considéré est un modèle destiné aux données asymétriques à queue lourde. Il est nommé Gauss-GPD et il lie une distribution gaussienne à une distribution de Pareto généralisée en un point de jonction avec une pondération uniforme des deux lois. En effet, le comportement moyen des données a été modélisé *via* le TCL alors que le comportement extrême des données a été modélisé par une GPD *via* la TVE. Pour ajuster ce modèle hybride, nous avons proposé un algorithme itératif non-supervisé permettant d'estimer judicieusement les paramètres du modèle. L'avantage de cet algorithme est double : il permet non seulement de bien ajuster les données et de bien les interpréter (les extrêmes au-delà du point de jonction indiquent la présence d'une information importante) mais aussi de déterminer le seuil indiquant la présence des extrêmes automatiquement contrairement aux méthodes classiques qui sont heuristiques et graphiques. À ce niveau, les performances du modèle hybride ainsi que de l'algorithme ont été testées sur des données générées.

Pour aller plus loin dans la modélisation de toutes les données, nous avons proposé d'améliorer le modèle Gauss-GPD en lui associant une pondération non-uniforme des deux composantes. Cette pondération non-uniforme a offert une meilleure manipulation de la queue de distribution. Ensuite, nous avons étendu ce modèle à un modèle nommé GPD-Gauss-GPD qui lie une gaussienne à deux GPD modélisant la queue droite et gauche de la distribution

avec une pondération non-uniforme. Ce modèle a été appliqué sur les données neuronales originales et après un traitement algébrique. Dans les deux cas, le bon ajustement des données a montré une séparation nette des ruptures du bruit. Nous avons aussi développé un modèle qui connecte une loi Rayleigh à une loi GPD avec une application en télécommunication. Bien évidemment, lorsque le nombre de composantes augmente dans les modèles hybrides, le nombre de paramètres à estimer augmente aussi et l'estimation de ces derniers devient plus compliquée. Ce problème est surmonté avec l'algorithme itératif proposé dans ce travail que nous l'avons adapté pour chaque modèle proposé.

Les résultats obtenus par les méthodes proposées dans cette thèse ouvrent des perspectives intéressantes pour la détection de ruptures ainsi que pour la modélisation des données non-homogènes et multi-composantes :

- Les modèles hybrides proposés dans cette thèse considèrent un comportement moyen gaussien. Il sera intéressant de proposer d'autres modèles hybrides modélisant des données ayant un comportement moyen non gaussien. Comme, par exemple, pour le cas du signal issu du filtre non-linéaire algébrique modélisé par une combinaison de lois Khi-deux.
- Vu l'importance de l'information donnée par les extrêmes, il sera bénéfique d'appliquer les modèles hybrides proposés dans cette thèse sur d'autres jeux de données : en télécommunication par exemple ou encore sur des données assurantielles et financières.
- Le passage brusque d'un comportement moyen à un comportement extrême nous incite à étudier le comportement intermédiaire entre ces deux derniers et ce en proposant des modèles hybrides qui prennent en considération ce comportement intermédiaire.



## Annexe A

# La variation de l'écart absolu à la médiane

Soit  $X$  une *va* de loi gaussienne de moyenne  $\mu$  et d'écart type  $\sigma$ . Nous soulignons dans cet annexe la relation entre  $\sigma$  et la variation de l'écart absolu à la médiane, dite en anglais *Median Absolute Deviation* (MAD) [Nenadic, 2005]. Nous rappelons que la MAD de  $X$  est définie comme étant la médiane de la *va*  $Y = |X - \mu|$ . La relation entre  $\sigma$  et la MAD de  $X$  est déduite à partir de la cdf de  $Y$ . En effet, nous avons

$$\begin{aligned}\mathbb{P}(Y \leq y) = \mathbb{P}(|X - \mu| \leq y) &= \begin{cases} \mathbb{P}(-y \leq X - \mu \leq y), & \text{si } y \geq 0; \\ 0, & \text{si } y < 0. \end{cases} \\ &= \begin{cases} \int_{-y}^y f(\lambda; 0, \sigma) d\lambda, & \text{si } y \geq 0; \\ 0, & \text{si } y < 0. \end{cases} \\ &= \begin{cases} 2[F(y; 0, \sigma) - F(0; 0, \sigma)], & \text{si } y \geq 0; \\ 0, & \text{si } y < 0. \end{cases}\end{aligned}$$

En particulier pour  $y = M$ , la médiane de  $Y$ , représentant aussi la MAD de  $X$ , nous obtenons

$$\mathbb{P}(Y \leq M) = 2[F(M; 0, \sigma) - F(0; 0, \sigma)]. \quad (\text{A.1})$$

D'autre part, nous avons

$$\mathbb{P}(Y \leq M) = \frac{1}{2}. \quad (\text{A.2})$$

En combinant (A.1) et (A.2) il vient que

$$\begin{aligned}\frac{1}{2} &= 2[F(M; 0, \sigma) - F(0; 0, \sigma)], \\ &= \text{erf}\left(\frac{M}{\sigma\sqrt{2}}\right).\end{aligned} \quad (\text{A.3})$$

De l'équation (A.3), nous pouvons tirer facilement la relation entre  $\sigma$  et  $M$  donnée par :

$$\sigma = \frac{M}{\sqrt{2}\text{erf}^{-1}(\frac{1}{2})} \approx \frac{M}{0.6745}, \quad (\text{A.4})$$

où  $\text{erf}$  est la fonction d'erreur et  $\text{erf}^{-1}$  est son inverse.

Notons que  $\sqrt{2}\text{erf}^{-1}(\frac{1}{2}) \approx 0.6745$ .

# Bibliographie

- [MIT, 1980] (1980). *MIT/BIH arrhythmia database-Tape directory and format specification*. Document BMEC TR00, Mass. Inst. Technol., Cambridge, 1980. Database is available from Bioengineering Division KB-26, Beth-Israel Hospital, 330 Brookline Avenue, Boston, MA 02215.
- [Abramowitz and Stegun, 1972] Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions : with Formulas, Graphs, and Mathematical Tables*. Dover books on mathematics. Dover, New York.
- [Adams and MacKay, 2007] Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv :0710.3742*.
- [Balkema and de Haan, 1974] Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5) :pp. 792–804.
- [Ben Jemaa et al., 2012] Ben Jemaa, A., Turki, M., and Guibene, W. (2012). Enhanced energy detector via algebraic approach for spectrum sensing in cognitive radio networks. In *Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), 2012 7th International ICST Conference on*, pages 113–117.
- [Boutoille, 2007] Boutoille, S. (2007). *Systèmes de fusion pour la segmentation hors-ligne de signaux GPS multi-porteuses*. Theses, Université du Littoral Côte d’Opale.
- [Braun et al., 2000] Braun, J., Braun, R., and Müller, H.-G. (2000). Multiple changepoint fitting via quasilielihood, with application to dna sequence segmentation. *Biometrika*, 87(2) :301–314.
- [Broadwater and Chellappa, 2010] Broadwater, J. and Chellappa, R. (2010). Adaptive threshold estimation via extreme value theory. *Signal Processing, IEEE Transactions on*, 58(2) :490–500.
- [Carreau and Bengio, 2009a] Carreau, J. and Bengio, Y. (2009a). A hybrid pareto mixture for conditional asymmetric fat-tailed distributions. *Neural Networks, IEEE Transactions on*, 20(7) :1087–1101.
- [Carreau and Bengio, 2009b] Carreau, J. and Bengio, Y. (2009b). A hybrid pareto model for asymmetric fat-tailed data : the univariate case. *Extremes*, 1 :53–76.
- [Choquet, 1969] Choquet, G. (1969). *Cours d’analyse : Tome II. Topologie*. Masson et Cie.
- [Davison, 2009] Davison, A. (2009). Spatial extremes and applications : Discussion outcomes.
- [Dayan and Abbott, 2005] Dayan, P. and Abbott, L. F. (2005). *Theoretical Neuroscience : Computational and Mathematical Modeling of Neural Systems*. The MIT Press.

- [De Miras et al., 2013] De Miras, J., Join, C., Fliess, M., Riachy, S., and Bonnet, S. (2013). Active magnetic bearing : A new step for model-free control. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 7449–7454.
- [Debbabi and El Asmi, 2012] Debbabi, N. and El Asmi, S. (2012). Algebraic approach for r-peak detection in the electrocardiogram (ecg) signal. In *Systems and Computer Science (ICSCS), 2012 1st International Conference on*, pages 1–5.
- [Debbabi et al., 2014] Debbabi, N., El Asmi, S., and Arfa, H. (2014). Extreme value theory for wireless fading channel modeling. In *Proc. of World Congress On E-Learning, Education and Computer Science*, pages 13–16.
- [Debbabi and Kratz, 2014] Debbabi, N. and Kratz, M. (2014). A new unsupervised threshold determination for hybrid models. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3440–3444.
- [Debbabi et al., 2015a] Debbabi, N., Kratz, M., and El Asmi, S. (2015a). A hybrid distribution for highly right skewed data. In *2015 International Conference on Communications and Networking (ComNet)*.
- [Debbabi et al., 2015b] Debbabi, N., Kratz, M., El Asmi, S., and Mboup, M. (2015b). Distribution hybride pour la modélisation de données à deux queues lourdes : Application sur les données neuronales. In *Groupe d’Etudes du Traitement du Signal et des Images, GRETSI*.
- [Debbabi et al., 2012] Debbabi, N., Kratz, M., Mboup, M., and Asmi, S. E. (2012). Combining algebraic approach with extreme value theory for spike detection. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1836–1840.
- [Demichel et al., 2011] Demichel, Y., Estrade, A., Kratz, M., and Samorodnitsky, G. (2011). How fast can the chord length distribution decay? *Advances in Applied Probability*, 43 :504–523.
- [Digham et al., 2007] Digham, F. F., Alouini, M. S., and Simon, M. K. (2007). On the energy detection of unknown signals over fading channels. *Communications, IEEE Transactions on*, 55(1) :21–24.
- [Embrechts et al., 1997] Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag.
- [Feller, 1966] Feller, W. (1966). *An Introduction to Probability Theory and its Applications*. Wiley.
- [Fisher and Tippett, 1928] Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge Univ Press.
- [Fliess et al., 2011a] Fliess, M., Join, C., and Hatt, F. (2011a). Is a probabilistic modeling really useful in financial engineering? In *Conférence Méditerranéenne sur l’Ingénierie Sûre des Systèmes Complexes, MISC 2011*, Agadir, Morocco.
- [Fliess et al., 2011b] Fliess, M., Join, C., and Hatt, F. (2011b). Volatility made observable at last. *CoRR*, abs/1102.0683.
- [Fliess et al., 2010] Fliess, M., Join, C., and Mboup, M. (2010). Algebraic change-point detection. *Appl. Algebra Eng. Commun. Comput.*, 21(2) :131–143.

- [Fliess et al., 2004] Fliess, M., Join, C., and Sira-Ramírez, H. (2004). Robust residual generation for linear fault diagnosis : an algebraic setting with examples. *International Journal of Control*, 77(14) :1223–1242.
- [Fliess et al., 2008] Fliess, M., Join, C., and Sira-Ramirez, H. (2008). Non-linear estimation is easy. *International Journal of Modelling, Identification and Control*, 4(1) :12–27.
- [Fliess et al., 2003] Fliess, M., Mboup, M., Mounier, H., and Sira-Ramirez, H. (2003). Questioning some paradigms of signal processing via concrete examples. In H. Sira-Ramirez, G. S.-N., editor, *Algebraic Methods in Flatness, Signal Processing and State Estimation*, Algebraic Methods in Flatness, Signal Processing and State Estimation, pages pp. 1–21, Mexico. Editorial Lagares.
- [Fliess and Sira-Ramírez, 2003] Fliess, M. and Sira-Ramírez, H. (2003). An algebraic framework for linear identification. *ESAIM : Control, Optimisation and Calculus of Variations*, 9 :151–168.
- [Furrer and Katz, 2008] Furrer, E. M. and Katz, R. W. (2008). Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, 44(12).
- [Gabor, 1946] Gabor, D. (1946). Theory of communication. part 1 : The analysis of information. *Journal of the Institution of Electrical Engineers-Part III : Radio and Communication Engineering*, 93(26) :429–441.
- [Gaya and Panda, 2011] Gaya, A. and Panda, D. (2011). Modeling, calculating and capacity enhancement of a rayleigh fading mimo channel. In *Applied Electromagnetics Conference (AEMC), 2011 IEEE*, pages 1–3.
- [Gehin, 2011] Gehin, W. (2011). Modélisation des queues de distribution des rendements des actif financiers. Master’s thesis, EURIA, Brest.
- [Ghosh and Resnick, 2010] Ghosh, S. and Resnick, S. (2010). A discussion on mean excess plots. *Stochastic Process and their Applications*, 3 :1492–1517.
- [Guillou et al., 2014] Guillou, A., Kratz, M., and Strat, Y. L. (2014). An extreme value theory approach for the early detection of time clusters. a simulation-based assessment and an illustration to the surveillance of salmonella. *Statistics in medicine*, 33(28) :5015–5027.
- [Gumbel, 1958] Gumbel, E. (1958). Statistics of extremes. 1958. *Columbia Univ. press, New York*.
- [Gutkin et al., 2003] Gutkin, B. S., Ermentrout, B., and Rudolph, M. (2003). Spike generating dynamics and the conditions of spike-time precision in cortical neurons. *J. of Comput. Neuroscience*, 15(1) :91–103.
- [Hertault, 1985] Hertault, J. (1985). *Comprendre l’ECG*.
- [Hill, 1975] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3 :1163–1174.
- [Hosking and Wallis, 1987a] Hosking, J. and Wallis, J. (1987a). Parameter and quantile estimation for the generalized pareto distribution funcrion. *Technometric*, 29 (3) :339–349.
- [Hosking and Wallis, 1987b] Hosking, J. R. M. and Wallis, J. R. (1987b). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3) :339–349.

- [Howard, 2004] Howard, R. M. (2004). *Principles of random signal analysis and low noise design : The power spectral density and its applications*. John Wiley & Sons.
- [Inoue and Kilian, 2005] Inoue, A. and Kilian, L. (2005). In-sample or out-of-sample tests of predictability : Which one should we use? *Econometric Reviews*, 23(4) :371–402.
- [JooSeuk and Clayton, 2012] JooSeuk, K. and Clayton, D. (2012). Robust kernel density estimation. *Journal of Machine Learning Research*, 13 :2529–2565.
- [Kakutani et al., 1941] Kakutani, S. et al. (1941). *A generalization of Brouwer’s fixed point theorem*. Duke University Press.
- [Kawahara and Sugiyama, 2009] Kawahara, Y. and Sugiyama, M. (2009). Change-point detection in time-series data by direct density-ratio estimation. In *SDM*, volume 9, pages 389–400. SIAM.
- [Killick et al., 2010] Killick, R., Eckley, I. A., Ewans, K., and Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37(13) :1120 – 1126.
- [Kim and Kim, 2000] Kim, K. H. and Kim, S. J. (2000). Neural spike sorting under nearly 0-db signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier. *Biomedical Engineering, IEEE Transactions on*, 47(10) :1406 –1411.
- [Knecht and Küttel, 2003] Knecht, M. and Küttel, S. (2003). The czeledin distribution function. In *XXXIV ASTIN Colloquium, Berlin*.
- [Kollu et al., 2012] Kollu, R., Rayapudi, S., Narasimham, S., and Pakkurthi, K. (2012). Mixture probability distribution functions to model wind speed distributions. *International Journal of Energy and Environmental Engineering*, 3(1) :27.
- [Kratz, 2013] Kratz, M. (2013). There is a var beyond usual approximations. preprint arXiv :1311.0270.
- [Kratz, 2014] Kratz, M. (2014). Normex, a new method for evaluating the distribution of aggregated heavy tailed risks. *Extremes*, (4). special issue on *Extremes in Finance* (Guess Ed. P. Embrechts) : 661-691.
- [Kratz and Resnick, 1996] Kratz, M. and Resnick, S. I. (1996). The qq-estimator and heavy tails. *Communications in Statistics. Stochastic Models*, 12(4) :699–724.
- [Kuhl and Bhairgond, 2000] Kuhl, M. E. and Bhairgond, P. (2000). Nonparametric estimation of nonhomogeneous poisson processes using wavelets. In *Simulation Conference, 2000. Proceedings. Winter*, volume 1, pages 562–571 vol.1.
- [Lagona et al., 2011] Lagona, F., Maruotti, A., and Picone, M. (2011). A non-homogeneous hidden markov model for the analysis of multi-pollutant exceedances data.
- [Laguna et al., 1989] Laguna, P., Caminal, P., Thakor, N., and Jane, R. (1989). Adaptive qrs shape estimation using hermite model. In *Engineering in Medicine and Biology Society, 1989. Images of the Twenty-First Century., Proceedings of the Annual International Conference of the IEEE Engineering in*, pages 683–684 vol.2.
- [Lavielle and Teyssiere, 2006] Lavielle, M. and Teyssiere, G. (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3) :287–306.

- [Lavielle and Teyssiere, 2007] Lavielle, M. and Teyssiere, G. (2007). Adaptive detection of multiple change-points in asset price volatility. In *Long memory in economics*, pages 129–156. Springer.
- [Levenberg, 1944] Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares quart. *Applied Math*, 2 :164–168.
- [Li et al., 2012] Li, C., Singh, V. P., and Mishra, A. K. (2012). Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water resources research*, 48 :1–17.
- [Li et al., 1995] Li, C., Zheng, C., and Tai, C. (1995). Detection of ecg characteristic points using wavelet transforms. *Biomedical Engineering, IEEE Transactions on*, 42(1) :21–28.
- [Mandava et al., 2011] Mandava, A. K., Shahram, L., and Regentova, E. E. (2011). Reliability assessment of microarray data using fuzzy classification methods : A comparative study. In *Advances in Computing and Communications*, volume 190.
- [Marquardt, 1963] Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11 :431–441.
- [Matei et al., 1997] Matei, C. M., Duchêne, J., and Nikiforov, I. (1997). Détection de ruptures dans les signaux électromyographiques. In *Groupe d’Etudes du Traitement du Signal et des Images, GRETSI*.
- [Mboup, 2007] Mboup, M. (2007). Parameter estimation via differential algebra and operational calculus. Research report. pp.40. HAL<inria-00138294>.
- [Mboup, 2008] Mboup, M. (2008). A Volterra filter for neuronal spike detection. Research report. <https://hal.inria.fr/inria-00347048>.
- [Mboup, 2009] Mboup, M. (2009). Parameter estimation for signals described by differential equations. *Applicable Analysis*, 88(1) :29–52.
- [Mboup, 2012] Mboup, M. (2012). Neural spike detection and localisation via volterra filtering. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–5.
- [Mboup et al., 2007] Mboup, M., Join, C., and Fliess, M. (2007). A revised look at numerical differentiation with an application to nonlinear feedback control. In *Control Automation, 2007. MED ’07. Mediterranean Conference on*, pages 1–6.
- [Mboup et al., 2008] Mboup, M., Join, C., and Fliess, M. (2008). An on-line change-point detection method. In *Control and Automation, 2008 16th Mediterranean Conference on*, pages 1290–1295.
- [Mboup et al., 2009] Mboup, M., Join, C., and Fliess, M. (2009). Numerical differentiation with annihilators in noisy environment. *Numerical algorithms*, 50(4) :439–467.
- [Meyer, 1992] Meyer, Y. (1992). Wavelets and operators, volume 37 of cambridge studies in advanced mathematics.
- [Mikusinski and Boehme, 1987] Mikusinski, J. and Boehme, T. K. (1987). *Operational Calculus*, volume 2. PWN Varsovie Oxford University Press, Oxford.
- [Mohammad-Djafari and Féron, 2006] Mohammad-Djafari, A. and Féron, O. (2006). Bayesian approach to change points detection in time series. *International Journal of Imaging Systems and Technology*, 16(5) :215–221.

- [Moreno et al., 2005] Moreno, E., Casella, G., and Garcia-Ferrer, A. (2005). An objective bayesian analysis of the change point problem. *Stochastic Environmental Research and Risk Assessment*, 19(3) :191–204.
- [Moussa Ali, 2011] Moussa Ali, A. (2011). *Fault diagnosis without a priori model*. Theses, Université Henri Poincaré - Nancy I.
- [Moussa Ali et al., 2011a] Moussa Ali, A., Join, C., and Hamelin, F. (2011a). Diagnostic des systèmes dynamiques linéaires sans modèle explicite. *Journal Européen des Systèmes Automatisés*, 45(Spécial : Nouvelles tendances en automatique) :339–362.
- [Moussa Ali et al., 2011b] Moussa Ali, A., Join, C., and Hamelin, F. (2011b). A robust algebraic approach to fault diagnosis of uncertain linear systems. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 1557–1562.
- [Nadarajah and Bakar, 2014] Nadarajah, S. and Bakar, S. (2014). New composite models for the danish fire insurance data. *Scandinavian Actuarial Journal*, 2014(2) :180–187.
- [Naveau et al., 2005] Naveau, P., Nogaj, M., Ammann, C., Yiou, P., Cooley, D., and Jomelli, V. (2005). Statistical methods for the analysis of climate extremes. *Comptes Rendus Geoscience*, 337(10) :1013–1022.
- [Nenadic, 2005] Nenadic, Z. (2005). Spike detection using the continuous wavelet transform. *IEEE Trans. biomedical engineering*, 52, no. 1 :74–87.
- [Niculescu et al., 2009] Niculescu, C. P., Roventa, I., and Lau, A. T. M. (2009). Schauder fixed point theorem in spaces with global nonpositive curvature. *Fixed Point Theory and Applications*, 2009 :20.
- [Nikiforov and Tikhonov, 1986] Nikiforov, I. and Tikhonov, I. (1986). Application of change detection theory to seismic signal processing. In Basseville, M. and Benveniste, A., editors, *Detection of Abrupt Changes in Signals and Dynamical Systems*, volume 77 of *Lecture Notes in Control and Information Sciences*, pages 355–373. Springer Berlin Heidelberg.
- [Orbanz and Teh, 2010] Orbanz, P. and Teh, Y. (2010). Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer.
- [Page, 1954] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2) :pp. 100–115.
- [Pan and Tompkins, 1985] Pan, J. and Tompkins, W. J. (1985). A real-time qrs detection algorithm. *Biomedical Engineering, IEEE Transactions on*, BME-32(3) :230–236.
- [Papoulis and Pillai, 2002] Papoulis, A. and Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*.
- [Pickands, 1975] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3 :119–131.
- [Pouzat, 2005] Pouzat, C. (2005). *Methods and Models in Neurophysics. Les Houches 2003 Summer School*, chapter Technique(s) for Spike - Sorting, pages 729–786. Elsevier. Available from : <http://fr.arxiv.org/abs/q-bio.QM/0405012>.
- [Pouzat, 2015] Pouzat, C. (2015). Tetrode recording from the antennal lobe of a locust (*schistocerca americana*).

- [Pouzat et al., 2002] Pouzat, C., Mazor, O., , and Laurent, G. (2002). Using noise signature to optimize spike-sorting and to assess neuronal classification quality. *Journal of neuroscience methods*, 122(1) :43–57.
- [Ramsay et al., 2007] Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations : a generalized smoothing approach. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 69(5) :741–796.
- [Rangaswamy et al., 2004] Rangaswamy, M., Michels, J. H., and Himed, B. (2004). Statistical analysis of the non-homogeneity detector for {STAP} applications. *Digital Signal Processing*, 14(3) :253 – 267.
- [Rao, 1973] Rao, C. R. (1973). *linear statistical inference and its applications*. New York : John Wiley.
- [Rezk et al., 2012] Rezk, S., Join, C., and Asmi, S. (2012). Inter-beat (r-r) intervals analysis using a new time delay estimation technique. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 929–933.
- [Rezk et al., 2011] Rezk, S., Join, C., and El Asmi, S. (2011). An algebraic derivative-based method for r wave detection. In *Signal Processing Conference, 2011 19th European*, pages 1578–1582.
- [Richter Charles, 1958] Richter Charles, F. (1958). *Elementary seismology*. San Francisco and London : WH Freeman and Company.
- [Rieke et al., 1999] Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1999). *Spikes : Exploring the Neural Code*. MIT Press, Cambridge, MA, USA.
- [Rossi et al., 1984] Rossi, F., Fiorentino, M., and Versace, P. (1984). Two-component extreme value distribution for flood frequency analysis. *Water Resources Research*, 20(7) :847–856.
- [Rudin, 1964] Rudin, W. (1964). *Principles of mathematical analysis*, volume 3. McGraw-Hill New York.
- [Sambridge and Compston, 1994] Sambridge, M. and Compston, W. (1994). Mixture modeling of multi-component data sets with application to ion-probe zircon ages. *Earth and Planetary Science Letters*, 128(3) :373–390.
- [Schwartz, 1998] Schwartz, L. (1998). *Théorie des distributions*. Hermann.
- [Schwartz, 2005] Schwartz, M. (2005). *Mobile Wireless Communications*. Cambridge University Press.
- [Scollnik, 2007] Scollnik, D. P. M. (2007). On composite lognormal-pareto models. *Scandinavian Actuarial Journal*, 2007(1) :20–33.
- [Sejdić et al., 2009] Sejdić, E., Djurović, I., and Jiang, J. (2009). Time-frequency feature representation using energy concentration : An overview of recent advances. *Digital Signal Processing*, 19(1) :153 – 183.
- [Sermpezis and Spyropoulos, 2015] Sermpezis, P. and Spyropoulos, T. (2015). Modelling and analysis of communication traffic heterogeneity in opportunistic networks. *Mobile Computing, IEEE Transactions on*, PP(99) :1–1.
- [Shankar, 2012] Shankar, P. M. (2012). *Fading and shadowing in wireless systems*. Springer.



- [Shewhart, 1931] Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. ASQ Quality Press.
- [Sivasankari and Thanushkodi, 2009] Sivasankari, N. and Thanushkodi, D. K. (2009). Automated epileptic seizure detection in eeg signals using fastica and neural network. *International Journal of Advances in Soft Computing and Its Applications*, 1(2).
- [Stoto et al., 2006] Stoto, M., Fricker, Ronald D., J., Jain, A., Diamond, A., Davies-Cole, J., Glymph, C., Kidane, G., Lum, G., Jones, L., Dehan, K., and Yuan, C. (2006). Evaluating statistical methods for syndromic surveillance. In Wilson, A., Wilson, G., and Olwell, D., editors, *Statistical Methods in Counterterrorism*, pages 141–172. Springer New York.
- [Surhone et al., 2010] Surhone, L., Timplendon, M., and Marseken, S. (2010). *Froda's Theorem*. VDM Publishing.
- [T. Buch-larsen and Bolancé, 2005] T. Buch-larsen, J. P. Nielsen, M. G. and Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, 39(6) :503–516.
- [Tanaka et al., 2002] Tanaka, S., Takara, K., Snorrason, A., Finnsdottir, H., and Moss, E. M. (2002). A study on threshold selection in pot analysis of extreme floods. *IAHS-AISH publication*, (271) :299–304. eng.
- [Tian et al., 2008] Tian, Y., Floquet, T., and Perruquetti, W. (2008). Estimation des paramètres des systèmes linéaires : approche algébrique. In *CIFA08*, Bucarest, Romania.
- [Tiganj and Mboup, 2011] Tiganj, Z. and Mboup, M. (2011). A non-parametric method for automatic neural spike clustering based on the non-uniform distribution of the data. *Journal of neural engineering*, 8(6) :066014.
- [Tiganj et al., 2012] Tiganj, Z., Mboup, M., Chevallier, S., and Kalunga, E. (2012). Online frequency band estimation and change-point detection. In *Systems and Computer Science (ICSCS), 2012 1st International Conference on*, pages 1–6. IEEE.
- [Verdier, 2007] Verdier, G. (2007). *Détection statistique de rupture de modèle dans les systèmes dynamiques. Application à la supervision de procédés de dépollution biologique (Thèse de doctorat, Université Montpellier 2 (Sciences et Techniques), Montpellier, FRA)*. PhD thesis.
- [Walker et al., 1999] Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(3) :485–527.
- [Watanabe et al., 2007] Watanabe, K., Takayasu, H., and Takayasu, M. (2007). A mathematical definition of the financial bubbles and crashes. *Physica A : Statistical Mechanics and its Applications*.
- [Werner, 1993] Werner, D. (1993). A proof of the markov-kakutani fixed point theorem via the hahn-banach theorem. *Extracta Math.*, 8 :37–38.
- [Wigner, 1932] Wigner, E. (1932). On the quantum correction for thermodynamic equilibrium. *Physical Review*, 40(5) :749.
- [Willemse, 2009] Willemse, I. (2009). *Statistical methods and calculation skills*. Juta and Company Ltd.

- [Yang et al., 2006] Yang, Y., Yu, D., and Cheng, J. (2006). A roller bearing fault diagnosis method based on {EMD} energy entropy and {ANN}. *Journal of Sound and Vibration*, 294(1–2) :269 – 277.
- [Yoon et al., 2010] Yoon, U.-J., Hwang, I.-S., Noh, Y.-S., Chung, I.-C., and Yoon, H.-R. (2010). Comparison of cwt with dwt for detecting qrs complex on wearable ecg recorder. In *Wavelet Analysis and Pattern Recognition (ICWAPR), 2010 International Conference on*, pages 300–303.
- [Zeyada et al., 2007] Zeyada, F., Moubarak, M., and Soliman, A. (2007). A generalization of generalized banach fixed-point theorem in a weak left small self-distance space. *Miskolc Mathematical Notes*, 8(99) :105.
- [Zumbach et al., 1999] Zumbach, G. O., Dacorogna, M. M., Olsen, J. L., and Olsen, R. B. (1999). Introducing a scale of market shocks.

# **Titre : Approche algébrique et théorie des valeurs extrêmes pour la détection de ruptures : Application aux signaux biomédicaux**

**Disciplines : Technologies de l'Information et de la Communication & Automatique, Signal, Productique, Robotique**

Ce travail développe des techniques non-supervisées de détection et de localisation en ligne de ruptures dans les signaux enregistrés dans un environnement bruité. Ces techniques reposent sur l'association d'une approche algébrique avec la TVE. L'approche algébrique permet d'appréhender aisément les ruptures en les caractérisant en termes de distributions de Dirac retardées et leurs dérivées dont la manipulation est facile via le calcul opérationnel. Cette caractérisation algébrique, permettant d'exprimer explicitement les instants d'occurrences des ruptures, est complétée par une interprétation probabiliste en termes d'extrêmes : une rupture est un événement rare dont l'amplitude associée est relativement grande. Ces événements sont modélisés dans le cadre de la TVE, par une distribution de Pareto Généralisée. Plusieurs modèles hybrides sont proposés dans ce travail pour décrire à la fois le comportement moyen (bruit) et les comportements extrêmes (les ruptures) du signal après un traitement algébrique. Des algorithmes entièrement non-supervisés sont développés pour l'évaluation de ces modèles hybrides, contrairement aux techniques classiques utilisées pour les problèmes d'estimation en question qui sont heuristiques et manuelles. Les algorithmes de détection de ruptures développés dans cette thèse ont été validés sur des données générées, puis appliqués sur des données réelles provenant de différents phénomènes, où les informations à extraire sont traduites par l'apparition de ruptures.

**Mots clés :** Algèbre différentielle, Algorithmes, Détection de ruptures, Estimation, Modélisation, Signaux biomédicaux, Théorie des valeurs extrêmes.

## **Title : Algebraic approach and extreme value theory for change-point detection : Application to biomedical signals**

This work develops non supervised techniques for on-line detection and location of change-points in noisy recorded signals. These techniques are based on the combination of an algebraic approach with the Extreme Value Theory (EVT). The algebraic approach offers an easy identification of the change-points. It characterizes them in terms of delayed Dirac distributions and their derivatives which are easily handled via operational calculus. This algebraic characterization, giving rise to an explicit expression of the change-points locations, is completed with a probabilistic interpretation in terms of extremes : a change point is seen as a rare and extreme event. Based on EVT, these events are modeled by a Generalized Pareto Distribution. Several hybrid multi-components models are proposed in this work, modeling at the same time the mean behavior (noise) and the extremes ones (change-points) of the signal after an algebraic processing. Non supervised algorithms are proposed to evaluate these hybrid models, avoiding the problems encountered with classical estimation methods which are graphical ad hoc ones. The change-points detection algorithms developed in this thesis are validated on generated data and then applied on real data, stemming from different phenomenons, where change-points represent the information to be extracted.

**Key words :** Algorithms, Biomedical signals, Change-point detection, Differential algebra, Estimation, Extreme Value Theory , Modeling.