Quelques contributions au traitement de données, de signaux et à ses applications

Nourddine Azzaoui

Document de synthèse

présenté pour obtenir

L'Habilitation à Diriger des Recherches

Mention Mathématiques Appliquées

École doctorale : Sciences Fondamentales

Quelques contributions au traitement de données, de signaux et à ses applications

Par : Nourddine Azzaoui

Parrain :

Arnaud Guillin	Université Clermont Auvergne

Rapporteurs :

Eric Moulines	Ecole Polytechnique
Michel Kieffer	Université de Paris-Saclay
Veronique Maume-Deschamps	Université de Lyon 1

Examinateurs :

Pierre Druilhet	Université Clermont Auvergne
Anne-Françoise Yao	Université Clermont Auvergne

Date de soutenance : 13 Avril 2022

Table des matières

Remerciements

En premier lieu, je voudrais exprimer toute ma reconnaissance aux professeur(e)s Véronique Maume-Deschamps, Michel Kieffer et Eric Moulines d'avoir acceptés de consacrer une partie de leur temps précieux pour juger ce travail. C'est pour moi un grand honneur et je les en remercie sincèrement. Je remercie Arnaud Guillin d'avoir accepté d'être mon responsable tutélaire et de présenter mes travaux de recherche aux différentes instances de l'Université. Je remercie également Anne-Françoise Yao et Pierre Druilhet d'avoir accepté de participer au jury de soutenance de mon HDR. Je voudrais exprimer ma reconnaissance à tous les membres (enseignants-chercheurs ou pas) du Laboratoire de Mathématiques Blaise Pascal pour le cadre de travail agréable et la bonne ambiance qu'ils ont su créer au sein du laboratoire. Je remercie enfin ma famille qui m'a apporté un soutien précieux.

Introduction

Ce document de synthèse résume mes travaux de recherche sur la modélisation et le traitement statistique de données issues de différents domaines d'applications. Mes activités de recherche s'inscrivent dans le cadre général du traitement statistique des données et du signal et de ses applications. Ma contribution dans ce domaine est multi-disciplinaire alliant à la fois un travail théorique avec souvent des finalités applicatives. La part importante de ces travaux a concerné l'analyse spectrale des processus α-stables et leurs applications en communication, en traitement de données et des signaux impulsifs. L'autre volet, qui n'est pas très éloigné de l'analyse spectrale a concerné l'analyse temps-fréquences de certains signaux non stationnaires et leurs applications ; en particulier en analyse de signaux physiologiques. La troisième partie est consacrée aux modèles de régressions, de sélection des variables ainsi que leurs applications en volcanologie entre autres. Pour simplifier la lecture, ce mémoire est organisé en deux tomes.

- Le Tome I a pour but de présenter, de manière succincte, mes travaux de recherche effectués durant ma carrière d'enseignant-chercheur. Dans cet exposé, nous insistons notamment sur la motivation et l'évolution de ces travaux en ne présentant que les résultats mathématiques clés. Pour garder une cohérence entre les sujets traités dans ce tome, certaines publications disciplinaires ne sont pas détaillées dans ce document mais peuvent être consultées dans le tome II.
- Le tome II, regroupe les textes originaux de mes publications dont la liste complète figure sur les pages 11–14. On y trouve également de manière détaillée les démonstrations des résultats obtenus. La structure des matières traitées ici suit la même organisation que dans le tome I.

Dans ce qui suit, nous donnons un résumé succinct des différents thèmes de recherches abordés dans ce tome :

Thème 1 : Modèles probabilistes appliqués en communications

Malgré le caractère mathématique abstrait des processus α-stables, mes travaux théoriques ont toujours été motivés par des applications en traitement du signal et en communications. Mes résultats les plus significatifs dans ce domaine ont suivi la chronologie suivante :

(1) Nous avons commencé par traiter le problème de la modélisation probabiliste du canal de

communication Ultra Large Bande (ULB) par des processus α-stables [71, 63, 44, 46, 26, 106].

(2) En se basant sur ces derniers résultats nous avons établi les fondements théoriques d'un récepteur de Cauchy optimal dans le contexte impulsif, multi-utilisateurs et multi-couches [48].

(3) Nous avons également généralisé ces résultats aux contextes mobiles et évolutifs [34].

(4) Récemment, nous avons revisité les résultats de Shannon pour établir des bornes de la capacité optimale d'un canal α-stable [18].

(5) Nous avons également étudié la structure de dépendance des interférences dans un milieu impulsif [27].

(6) Toujours en lien avec les modèles de canal, nous avons donné une estimation du débit maximum qu'on peut atteindre dans un canal impulsif [21].

Ces travaux ont motivé des développements de nouveaux résultats en cours visant à estimer, simuler et à caractériser les processus α-stables harmonisables non stationnaires [1, 2]; cette caractérisation permettra de mieux prendre en compte les communications en milieu impulsif et non stationnaire.

Thème 2 : Analyse Temps-Fréquences et applications

Non loin de mes travaux sur l'analyse spectrale des processus, je me suis intéressé à la Décomposition Modale Empirique (EMD) pour traiter des signaux multivariés. Ce travail qui revêt un caractère théorique et abstrait, nous a amené à donner une reformulation de l'EMD classique dans le cas unidimensionnel [58, 57]. Ce résultat intermédiaire nous a facilité la tâche de généralisation à des signaux vectoriels [58]. Partant de ces résultats, nous avons développé des applications dans le domaine médical [50, 55]. Dans le cadre de la thèse de Marta Campi, nous avons utilisé l'EMD pour l'extraction des *features* pour traiter des signaux de la parole et des applications en finance [5]. En gardant ce même esprit de technique d'analyse temps fréquence des signaux, nous avons donné des méthodes de classifications basées sur un algorithme de détection de rupture sur les énergies hautes et basses fréquences extraites à partir de la décomposition en ondelettes. Une application en médecine a été mise en place avec le CHU de Clermont-Ferrand [30].

Thème 3 : Modèles de régressions, sélection de variables et applications

En se basant sur la divergence symétrique de Kullback et les approches de maximum de vraisemblance et de la vraisemblance résiduelle, nous avons proposé deux nouveaux critères adaptés aux échantillons de petites tailles et ayant des erreurs corrélées. Ce résultat a été introduit pour remédier au biais négatif des critères classiques (AIC, KIC, ...) qui induisent une sur-parametrisation des modèles sélectionnés. Les performances de ces critères ont été examinées dans une large étude en simulations [33]. Par la suite, nous avons généralisé ce résultat aux problèmes de sélections de variables dans le cadre de régressions pénalisées de type LASSO. Ces résultats ont été appliqués à des problèmes de surveillance de panache de cendre en volcanologie où la taille des données est petite mais dont le nombre de covariables est élevé [25, 29, 12]. Nous avons considéré le problème de la régression en grande dimension et l'estimateurs des moindres carrés pénalisés [4]. Nous avons établi notamment des inégalités oracles en prédiction pour ces estimateurs lorsque les erreurs suivent une loi vérifiant une inégalité de Poincaré faible.

Dans le cadre de la thèse de P-A. Faye nous avons donné un algorithme de prédiction spatio-temporel utilisant des méthodes de Krigeage Bayésien couplé à l'information apportée par une boîte noire (Equation au dérivées partielles, schémas numériques …) [17]. Une technique de plan d'expériences a été utilisée pour le choix optimal du positionnement spatial des emplacements de mesures [14].

Transfert : Collaboration dans le cadre de projets industriels

Récemment, avec la tendance actuelle d'encourager l'interaction des mathématiques appliquées avec le monde industriel et professionnel, j'ai développé plusieurs projets avec des entreprises nationales ou internationales :

- **Le projet Eaugure :** L'objectif du projet est de développer des outils de surveillance en temps réel de la qualité des eaux de baignade en exploitant des données biologiques et météorologiques. La finalité de ce projet est la mise place d'un système prédictif, basé sur des alertes météorologiques permettant d'anticiper le risque de contamination bactérienne ainsi que les pollutions provenant par exemple du lessivage des sols... Le système élaboré permet également de mesurer la répartition spatiale des cyanobactéries sur le plan d'eau à surveiller. Nous avons mis au point des techniques de prévision spatio-temporelle pour évaluer la qualité de l'eau des lacs régionaux. Des expérimentations en conditions réelles sont en cours sur les plans d'eau de Cournon-d'Auvergne et d'Aydat.
- **Le projet Météo-Marketing :** Ce projet est en collaboration avec les entreprises *Phimeca* et *Périscope Créations* et pour lequel nous avons eu le soutien de la région Auvergne, l'AMIES et le fond européen FEDER. Nous avons développé des techniques innovantes dans le domaine du *météo marketing*. L'idée principale est de proposer un modèle de prévision statistique permettant d'évaluer l'impact de la météo sur le comportement de consommation digitale. Un produit à destinations des acteurs de l'*E-commerce* ou commerce traditionnel a été développé sous forme d'une API permettant de fournir en temps-réel l'impact des effets météo sur les ventes. Cette application est accessible sur le lien http://meteomarketing.com/.
- **Projet Vigires'eau :** La qualité et la sûreté des systèmes d'approvisionnement d'eau sont essentielles pour la santé publique. Il est donc nécessaire de prévenir toute intrusion dans ces systèmes, et de détecter en temps quasi réel l'introduction de pollution, qu'elle soit intentionnelle ou accidentelle. Dans le cadre du projet Vigirés'Eau en collaboration avec GDF - SUEZ, nous avons mis en place des algorithmes permettant la détection et la localisation d'intrus dans un réseau d'eau potable [40]. Ces résultats sont basés sur des approches et des modèles semi-paramétriques qui découlent naturellement des modèles hydrauliques. Ces approches combinent, d'une part, la simplicité des modèles paramétriques qui se prêtent facilement à une interprétation physique ; et d'autre part la flexibilité des modèles non paramétriques qui ont moins d'a priori sur le comportement du réseau.
- **Projet avec Bridgestone** : Sur le plan international, une collaboration avec le leader Japonais de pneumatique Bridgestone pour la détection en temps réel, et directement à travers le pneu, les conditions des routes [19]. Nous avons donné un algorithme d'extraction de *featrures* pour classer les conditions de la route en utilisant les signaux d'accéléromètres mesurés par des capteurs montés sur les pneus. Dans cette étude, la décomposition modale empirique a été utilisée pour extraire les caractéristiques des fréquences instantanées

et les relier aux différentes conditions de la route. Nous avons testé cet algorithme sur des données routières réelles dans différentes conditions climatiques. Il est à noter qu'à cause de la confidentialité de la plupart des sujets traités en collaboration industrielle, il n'a pas toujours été facile (voire impossible) de publier les résultats de ces travaux.

Avant de rentrer dans les détails de l'évolution de mes travaux de recherche, j'expose brièvement ici le contexte des évolutions de ces travaux. J'ai soutenu ma thèse de Doctorat en Mathématiques appliquées et statistiques, intitulée *Analyse et estimation spectrale des processus α-stables non stationnaires*, en décembre 2006 (sous la codirection de B. Schmitt et R. Sabre). Ma carrière d'enseignant chercheur a commencé en 2009-2010, quand j'ai été recruté en tant que maître de conférence à l'Université Clermont Auvergne (Anciennement Université Blaise Pascal). Ce poste a été créé pour renforcer l'équipe pédagogique en statistique, en particulier l'enseignement dans l'ancien Master Professionnel « statistique et traitement de données » où j'avais pris la responsabilité de plusieurs UE théoriques et appliquées. Depuis, septembre 2018 je suis responsable du Master de Mathématiques Appliquées, Statistique (MAS). Pendant la période 2012-2017, j'ai été élu au conseil du Laboratoire de Mathématiques Blaise Pascal (LMBP) UMR 6620. De 2017-2019 j'ai été élu au conseil national des universités CNU en section 26 : Mathématiques appliquées et applications des mathématiques. Ces tâches administratives et l'implication en enseignement ainsi que dans la vie locale et nationale a certainement eu une influence sur mes recherches mais malgré cela je pense avoir gardé une activité de recherche attestée par des publications régulières à la fois en mathématiques appliquées et en applications des mathématiques. J'ai participé également à plusieurs projets de recherche nationaux ou locaux, soit en tant que membre et/ou responsable : ANR Do-well be, projet Meteo-Marketing, projet Eaugure, Tellus CNRS, projet PEPS. J'ai également pris part activement, dans le cadre du Labex Clervolc et l'Isite Cap 20-25, où j'ai pu développer des collaborations fructueuses en interne à l'UCA notamment avec A. Guillin (LMBP) , O. Roche et M. Gouhier au Laboratoire Magma et Volcan (LMV). Mon implication dans ces projets s'est traduite par le co-encadrement de thèses ou de postdoc dont les détails sont donnés dans les **annexes**. Au niveau international, depuis 2012, je suis membre du Laboratoire Euro-Maghrébin de Mathématiques et de leurs Interactions où j'ai participé à plusieurs de ses rencontres. J'ai également été en 2015 à l'origine d'une convention de collaboration entre l'UCA et l'*Institute of Statistical Mathematics* (ISM-Tokyo) où je suis régulièrement invité par T. Matsui. Dans le cadre de ma collaboration avec G.W. Peters, j'ai été plusieurs fois invité l'*University Collège of London* (UCL) pour des séjours courts ce qui a abouti par la suite au co-encadrement de la thèse de M. Campi. En animation de la recherche à l'internationale j'ai organisé plusieurs sessions dans des congrès internationaux de premier plan, par exemple le Congrès Mondial de la Statistique ISI 2017, la conférence internationale de l'ERCIM on Computational and Methodological Statistics (CMStatistics) en 2016 à Séville et en 2018 à Pise. Je suis régulièrement sollicité pour faire des rapports sur des publications dans des revues comme par exemple IEEE Tran. on comm, IEEE Trans. Sig. Processing, JMVA, JASA, Pone...

Plus de détails sur le déroulement de ma carrière sont donnés de manière synthétique dans les **annexes**.

Tableau récapitulatif des productions scientifiques

Type du document	Publiés	Soumis	En préparation
Revues	14	4	2
Peer Reviewed Proceedings	12		1
Conférences nationales	3		
Livres et chapitres	1		
Rapports techniques	4		
Total	33	4	3

Certains travaux ont été réalisés en collaboration avec des collègues français et/ou étrangers, à savoir :

- Laurent Clavier de l'Université de Lille 1, IRCICA-IEMN,
- Pierre Druilhet LMBP de l'Université Clermont Auvergne,,
- Arnaud Guillin LMBP de l'Université Clermont Auvergne,
- Tomoko Matsui The Institute of Statistical Mathematics (Japan).
- Gareth W. Peters de University *College of London (UCL) and UCSB, Santa Barbara*,
- Olivier Roche de LMV de l'Université Clermont Auvergne,
- François Septier Université Bretagne Sud.
- Hichem Snoussi Université de Technologie de Troyes.
- Rachid Sabre Agro-Sup Dijon.

Articles soumis

[1] **N. AZZAOUI**, G. W. Peters, A. Guillin et M. Egan. *Spectral Characterization of the Non-Independent Increment Family of α-Stable Processes that Generalize Gaussian Process Models.* doi : 10.2139/ssrn.2892547.

[2] **N. AZZAOUI**, G. W. Peters, A. Guillin et M. Egan. *Supplement to : 'Spectral Characterization of the Family α-Stable Processes that Generalize Gaussian Process Models.'* doi : 10.2139/ssrn.2892570.

[3] O. Roche, C. D. Henry, **N. AZZAOUI** et A. Guillin. *Extremely long runout (>300 km) pyroclastic density currents*.

[4] A. Sy, C. Araujo-Bonjean, M.-E. Dury, **N. AZZAOUI** et A. Guillin. *An Extreme Value Mixture model to assess drought hazard in West Africa*. url : https://hal.uca.fr/hal-03297023/document.

Articles de revues

[1] T. Matsui, **N. AZZAOUI** et D. Murakami. « Analysis of COVID-19 evolution based on testing closeness of sequential data ». In : *Japanese Journal of Statistics and Data Science* (2022), p. 1-18. doi : 10.1007/s42081-021-00144-w.

[2] A. Tadini, **N. AZZAOUI**, O. Roche, P. Samaniego, B. Bernard, A. Bevilacqua, S. Hidalgo, A. Guillin et M. Gouhier. « Tephra fallout probabilistic hazard maps for Cotopaxi and Guagua Pichincha volcanoes (Ecuador) with uncertainty quantification ». In : *Journal of Geophysical Research : Solid Earth* 127.2 (2022). doi : 10.1029/2021JB022780.

[3] D. Abdillahi-Ali, **N. AZZAOUI**, A. Guillin, G. Le Mailloux et T. Matsui. « Penalized Least Square in Sparse Setting with Convex Penalty and Non Gaussian Errors ». In : *Acta Mathematica Scientia* 41.6 (2021), p. 2198-2216. doi : 10.1007/s10473-021-0624-0.

[4] M. Campi, G. W. Peters, **N. AZZAOUI** et T. Matsui. « Machine Learning Mitigants for Speech Based Cyber Risk ». In : *IEEE Access* 9 (2021), p. 136831-136860. doi : 10.1109/ACCESS.2021.3117080.

[5] O. Roche, **N. AZZAOUI** et A. Guillin. « Discharge rate of explosive volcanic eruption controls runout distance of pyroclastic density currents ». In : *Earth and Planetary Science Letters* 568 (2021), p. 117017. doi : 10.1016/j.epsl.2021.117017.

[6] A. Tadini, O. Roche, P. Samaniego, **N. AZZAOUI**, A. Bevilacqua, A. Guillin, M. Gouhier, B. Bernard, W. Aspinall, S. Hidalgo et al. « Eruption type probability and eruption source parameters at Cotopaxi and Guagua Pichincha volcanoes (Ecuador) with uncertainty quantification ». In : *Bulletin of Volcanology* 83.5 (2021), p. 1-25. doi : 10.1007/s00445-021-01458-z.

[7] A. Tadini, O. Roche, P. Samaniego, A. Guillin, **N. AZZAOUI**, M. Gouhier, M. de'Michieli Vitturi, F. Pardini, J. Eychenne, B. Bernard et al. « Quantifying the uncertainty of a coupled plume and tephra dispersal model : PLUME-MOM/HYSPLIT simulations applied to Andean volcanoes ». In : *Journal of Geophysical Research : Solid Earth* 125.2 (2020). doi : 10.1029/2019JB018390.

[8] M. Gouhier, J. Eychenne, **N. AZZAOUI**, A. Guillin, M. Deslandes, M. Poret, A. Costa et P. Husson. « Low efficiency of large volcanic eruptions in transporting very fine ash into the atmosphere ». In : *Nature : Scientific reports* 9.1 (2019), p. 1-12. doi : 10.1038/s41598-019-38595-7.

[9] M. L. de Freitas, M. Egan, L. Clavier, A. Goupil, G. W. Peters et **N. AZZAOUI**. « Capacity Bounds for Additive Symmetric α-Stable Noise Channels ». In : *IEEE Transactions on Information Theory* 63.8 (août 2017), p. 5115-5123. doi : 10.1109/TIT.2017.2676104.

[10] **N. AZZAOUI**, L. Clavier, A. Guillin et G. W. Peters. « Spectral Measures of α-Stable Distributions : An Overview and Natural Applications in Wireless Communications ». In : *Theoretical Aspects of Spatial-Temporal Modeling* (1er jan. 2015). doi : 10.1007/978-4-431-55336-6_3.

[11] **N. AZZAOUI**, A. Guillin, F. Dutheil, G. Boudet, A. Chamoux, C. Perrier, J. Schmidt et P. R. Bertrand. « Classifying heartrate by change detection and wavelet methods for emergency physicians ». In : *Essaim* 45 (2014), p. 48-57. doi : 10.1051/proc/201445005.

[12] B. Hafidi et **N. AZZAOUI**. « Criteria for longitudinal data model selection based on Kullback's symmetric divergence ». In : *Revue ARIMA* 15 (2012), p. 83-99. url : https://arima.episciences.org/1959.

[13] H. E. Ghannudi, L. Clavier, **N. AZZAOUI**, F. Septier et P. a. Rolland. « α-stable interference modeling and cauchy receiver for an IR-UWB Ad Hoc network ». In : *IEEE Transactions on Communications* 58.6 (juin 2010), p. 1748-1757. doi : 10.1109/TCOMM.2010.06.090074.

[14] **N. AZZAOUI** et L. Clavier. « Statistical channel model based on α-stable random processes and application to the 60 GHz ultra wide band channel ». In : *IEEE Transactions on Communications* 58.5 (mai 2010), p. 1457-1467. doi : 10.1109/TCOMM.2010.05.090069.

Conférences internationales (Peer reviewed)

[1] J. Hautot, C. Teulière et **N. AZZAOUI**. « Visual Radial Basis Q-Network ». In : *ICPRAI : International Conference on Pattern Recognition and Artificial Intelligence*. 2022.

[2] A. Tadini, O. Roche, P. Samaniego, A. Guillin, **N. AZZAOUI**, M. Gouhier, A. Bevilacqua, M. De'Michieli Vitturi, F. Pardini, B. Bernard et al. « Developing probabilistic tephra fallout hazard maps for Cotopaxi and Guagua Pichincha volcanoes, Ecuador, with uncertainty quantification ». In : t. 2019. 2019, p. V23I-0313. URL : lien.

[3] I. Keita, G. Takato, T. Matsui, G. W. Peters et **N. AZZAOUI**. « Feature Extraction Using Empirical Mode Decomposition for Tire Sensing ». In : *Infomation-Based Induction Sciences and Machine Learning (IBIS-ML2017)*. 2017, p. 315-320. URL : https://www.ieice.org/ken/paper/20171110nbZh/eng/.

[4] M. Gouhier, A. Guillin, **N. AZZAOUI**, J. Eychenne et S. Valade. « Source mass eruption rate retrieved from satellite-based data using statistical modelling ». In : *EGU General Assembly Conference Abstracts*. T. 17. 2015. URL : https://ui.adsabs.harvard.edu/abs/2015EGUGA..1710222G/abstract.

[5] X. Yan, L. Clavier, G. W. Peters, **N. AZZAOUI**, F. Septier et I. Nevat. « Skew-t copula for dependence modelling of impulsive (α-stable) interference ». In : *Proc. IEEE Int. Conf. Communications (ICC)*. Juin 2015, p. 4816-4821. DOI : 10.1109/ICC.2015.7249085.

[6] **N. AZZAOUI** et L. Clavier. « UWB channel modeling for objects evolving in impulsive environnements ». In : *Proc. IEEE Wireless Communications and Networking Conf. Workshops (WCNCW)*. Avr. 2012, p. 191-195. DOI : 10.1109/WCNCW.2012.6215488.

[7] L. Fillatre, P. Honeine, I. Nikiforov, C. Richard, H. Snoussi, **N. AZZAOUI**, B. Guépié, Z. Noumir, S. Deveughèle et H. Yin. « Vigires'eau : Surveillance en temps réel de la qualité de l'eau potable d'un réseau de distribution en vue de la détection d'intrusions ». In : *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'11),(ANR-CSOSG)*. 2011. URL : lien.

[8] A. Miraoui, H. Snoussi, J. Duchêne et **N. AZZAOUI**. « On the detection of elderly equilibrium degradation using multivariate-EMD ». In : *Proc. IEEE Globecom Workshops*. Déc. 2010, p. 2049-2053. DOI : 10.1109/GLOCOMW.2010.5700305.

[9] **N. AZZAOUI**, A. Miraoui, H. Snoussi et J. Duchene. « Empirical Mode Decomposition for vectorial bi-dimensional signals ». In : *Proc. Int. Workshop Multidimensional (nD) Systems*. Juin 2009, p. 1-4. DOI : 10.1109/NDS.2009.5191553.

[10] **N. AZZAOUI**, H. Snoussi et J. Duchene. « An enhanced empirical modal decomposition without sifting ». In : *Proc. IEEE/SP 15th Workshop Statistical Signal Processing*. Août 2009, p. 796-799. DOI : 10.1109/SSP.2009.5278474.

[11] **N. AZZAOUI** et L. Clavier. « An Impulse Response Model for the 60 Ghz Channel Based on Spectral Techniques of α-stable Processes ». In : *Proc. IEEE Int. Conf. Communications*. Juin 2007, p. 5040-5045. DOI : 10.1109/ICC.2007.832.

[12] **N. AZZAOUI**, L. Clavier et R. Sabre. « Path delay model based on α-stable distribution for the 60 GHz indoor channel ». In : *Proc. IEEE Global Telecommunications Conf. GLOBECOM '03*. T. 3. Déc. 2003, 1638-1643 vol.3. DOI : 10.1109/GLOCOM.2003.1258515.

Conférences nationales avec actes

[1] S. Coly, P. Druilhet, **N. AZZAOUI** et P. A. Faye. *Plans d'échantillonnage spatiaux dans un contexte de surveillance*. Sous la dir. de 50èmes Journées de Statistique. 2018. url : lien.

[2] L. Clavier et **N. AZZAOUI**. *Processus alpha-stables et communications numériques.* Sous la dir. de J. M. et Journée en l'honneur de Jacques Neveu. 2010. url : pdf.

[3] A. Miraoui, **N. AZZAOUI**, H. Snoussi et J. Duchêne. *Détection de dégradation de l'équilibre chez les personnes âgées*. Sous la dir. de C. rendu des Journées SFTAG. 2009. url : lien.

Rapports techniques

[1] S. Coly, **N. Azzaoui**, A. Guillin et A.-F. Yao. *Mise en place d'un système prédictif permettant d'anticiper le risque de contamination bactérienne des eaux douces.* Rapp. tech. Rapport technique du projet EAUGURE 2, 2020. url : lien.

[2] H. Font, J. Veslot, **N. Azzaoui**, A. Guillin et A.-F. Yao. *Methodes statistiques pour la prévisions de vente en fonction de la météo.* Rapp. tech. Rapport technique du projet Météo Marketing, 2018. url : lien.

[3] L. Clavier, **N. AZZAOUI** et W. Sawaya. *UWB and 60 GHz channel model as an α-stable random process*. Rapp. tech. Technical Report TD (08) 634, Lille, France, 2008. url : Lien.

Thèses et Mémoires

[1] **N. AZZAOUI**. « Analyse et Estimations Spectrales des Processus alpha-Stables non-Stationnaires ». Université de Bourgogne, 2006. url : lien.

Livres et chapitres

[1] **N. AZZAOUI**, A. Guillin, M. Gouhier, J. Eychenne et S. Valade. *Modélisation statistique pour la surveillance des éruptions volcaniques.* Sous la dir. de R. d'Auvergne. 2014, p. 153-170. url : lien.

Modèles en communication

La mesure et le traitement d'informations liés au confort, à la sécurité, aux communications ou généralement à une application spécifique, sont autant de challenges au coeur de l'évolution actuelle et future du monde industriel et de la recherche appliquée. En effet, la tendance des nouvelles technologies converge vers la mise en place de systèmes globaux permettant d'effectuer des tâches différentes selon les besoins et en s'adaptant aux contraintes économiques et écologiques... Nous assistons à la multiplication des réseaux de capteurs et d'objets communicants de natures hétérogènes qui sont les prémices d'une évolution vers ce qu'on appelle maintenant l'internet des objets. L'avènement de ces réseaux en même temps qu'une demande incessante en débit et d'une meilleure qualité de service ont augmenté la complexité des communications et mettent à rude épreuve les techniques et modèles classiques. En effet, ces derniers doivent s'adapter, d'une part, à l'évolution temporelle et spatiale et d'autre part, ils doivent prendre en compte les événements rares et imprévisibles qui peuvent avoir des conséquences désastreuses sur la prise de décision. C'est dans ce contexte que nous nous sommes focalisés sur le développement et l'approfondissement des modèles mathématiques nécessaires à la mise en place et à l'optimisation d'un système de communication autonome ayant la capacité d'évoluer en fonction de son environnement.

L'idée et les motivations principales sont inspirées du succès des interactions historiques entre les domaines des probabilités, les statistiques et le monde des communications, de la théorie de l'information et le traitement du signal. Nos travaux ont toujours eu deux aspects fondamentalement liés : dans un premier temps, une approche théorique, nécessaire à une bonne formalisation des problèmes et à la définition des solutions optimales ; dans un deuxième temps, l'utilisation de ces modèles pour la mise en place des réseaux futurs. Cette dernière tâche est loin d'être simple ; du fait de la complexité des systèmes étudiés et de la diversité des problèmes rencontrés. Les modèles théoriques cités ci-dessus ont été mis à contribution pour la modélisation des canaux de communication et de l'interférence dans un cadre impulsif. Nous nous sommes intéressés également à la modélisation de l'évolution spatiale et temporelle de ces modèles et à l'impact du bruit multi-utilisateurs inhérent à la densification des réseaux. En d'autres termes, quand un noeud transmet ou reçoit de l'information, comment cette information va être influencée par le canal (déformations dues à l'environnement) et par les interférences dues aux noeuds voisins et aux utilisateurs.

Les limites du canal à bruit additif gaussien proposées par Shannon en (1948) sont actuellement presque atteintes et on peut se poser la question de l'intérêt de poursuivre des recherches pour gagner d'infimes améliorations. L'évolution des réseaux cellulaires vers la 5G et l'arrivée de l'Internet des objets ont augmenté significativement la complexité des réseaux de communication. Ces derniers deviendront alors plus chaotiques et les apparitions des événements rares deviennent plus importantes. Cela conduit à une dégradation de la qualité des communications en raison du choix non approprié du modèle gaussien ou du second ordre en général. Pour remédier à ce problème, nous avons proposé des modèles basés sur des processus et variables α-stables, qui semblent plus appropriés à notre contexte. Nous avons revisité les résultats de Shannon pour établir des bornes de la capacité optimale d'un canal α-stable ainsi que le taux de débit maximum qu'on peut atteindre dans ce contexte impulsif.

En théorie des probabilités et des statistiques, les distributions α-stables représentent une classe importante de lois paramétriques généralisant ainsi les fameuses lois gaussiennes ($\alpha = 2$), de Cauchy ($\alpha = 1$) et les lois de Lévy ($\alpha = \frac{1}{2}$). Les propriétés fondamentales qui les rendent attrayantes pour des applications sont notamment :

a) La propriété de stabilité qui assure qu'une combinaison linéaire de variables aléatoires indépendantes de distribution stable est aussi une variable stable ; en conséquence, nous avons l'impressionnant résultat du théorème central limite généralisé qui assure qu'une distribution stable est la seule limite possible pour des sommes normalisées de variables indépendantes et identiquement distribuées (iid). Il faut noter que ce théorème central limite est la principale justification de l'utilisation de la distribution gaussienne dans la plupart des applications statistiques ; c'est-à-dire pour des variables aléatoires iid de variance finie, nous obtenons une convergence vers la distribution gaussienne qui est un cas particulier correspondant à $\alpha = 2$.

b) La deuxième propriété qui est de grande utilité dans nos travaux, est que les distributions stables, hormis le cas gaussien, sont dites à *queues lourdes*. D'un point de vue pratique, ceci se traduit par des probabilités plus élevées d'apparition des valeurs extrêmes et donne un aspect impulsif dans les réalisations de ces variables. Cette propriété est très importante dans la mesure où elle permet de modéliser les événements rares qui sont inhérents à la complexité de l'internet des objets.

Pour faciliter la lecture de ce document nous donnons un rappel succinct des différentes propriétés des distributions et processus α-stables que nous utiliserons par la suite. Un exposé plus détaillé peut être consulté dans le Tome 2 de ce document avec les références qui y sont citées.

Un vecteur aléatoire réel centré $X^d = (X_1, X_2, ..., X_d)$ est dit symétrique α-stable (SαS) si et seulement si sa fonction caractéristique est donnée par :

$$\phi(t_1, ..., t_d) = \exp\{-\int_{\mathbb{S}_d} \left|\sum_{i=1}^{d} t_i s_i\right|^{\alpha} d\Gamma(s_1, ..., s_d)\}, \tag{3.1}$$

Γ est une mesure symétrique unique définie sur la sphère unité $\mathbb{S}_d$ de $\mathbb{R}^d$ (voir[148]).

La définition d'un vecteur complexe SαS est naturellement généralisée en utilisant les parties réelles et imaginaires. D'où un vecteur aléatoire complexe, $X^d = (X_1, \dots, X_d)$ avec $X_i = X_{i,1} + \jmath X_{i,2}$, est S$\alpha$S si le vecteur réel $(X_{1,1}, X_{1,2}, ..., X_{d,1}, X_{d,2})$ est aussi SαS.

Pour généraliser la notion de covariance, Cambanis [168] a défini la covariation de deux composantes du vecteur SαS X^d par :

$$[X_i, X_k]_\alpha \triangleq \int_{\mathbb{S}_4} (s_1 + \jmath s_2)(s_3 + \jmath s_4)^{<\alpha-1>} d\Gamma, \tag{3.2}$$

où $z^{<\alpha>} = |z|^{\alpha-1}\overline{z}$, et $\overline{z}$ est le conjugué de z. Généralement, contrairement à la covariance des vecteurs et processus du second ordre, la covariation n'est ni symétrique ni bilinéaire. Cependant, elle a un rôle similaire dans de nombreuses situations pratiques [148]. Elle a les propriétés suivantes :

- Linéarité par rapport au composant gauche :
  - pour tout vecteur SαS (X_1, X_2, Y),
  $$[X_1 + X_2, Y]_\alpha = [X_1, Y]_\alpha + [X_2, Y]_\alpha. \tag{3.3}$$
  - pour tous scalaires complexes a et b :
  $$[aX, bY]_\alpha = ab^{<\alpha-1>}[X, Y]_\alpha. \tag{3.4}$$
- Si X et Y sont indépendants alors, $[X, Y]_\alpha = 0$.
- Si Y_1 et Y_2 sont indépendants alors nous avons l'additivité de la covariation :
$$[X, Y_1 + Y_2]_\alpha = [X, Y_1]_\alpha + [X, Y_2]_\alpha. \tag{3.5}$$
- Pour tout $1 < \alpha \leq 2$, l'application $X \longmapsto \|X\|_\alpha = ([X, X]_\alpha)^{\frac{1}{\alpha}}$ définit une norme sur l'espace vectoriel engendré par des variables aléatoires SαS, appelée norme de la covariation.
- La cinquième propriété donne $\mathbb{E}[|Z|]$ pour des variables aléatoires symétriques α-stables, à l'origine dues à Zolotarev [184].
  Soit $Z \sim S_\alpha(\gamma, 0, 0)$, avec $1 < \alpha \leq 2$. Alors,
  $$\mathbb{E}[|Z|] = \frac{2\Gamma\left(1 - \frac{1}{\alpha}\right)}{\pi}\gamma. \tag{3.6}$$
- Soit $Z \sim S_\alpha(\gamma, \beta, \delta)$, alors la densité de probabilité de Z peut être majorée par :
  $$p_Z(y) \leq \frac{\Gamma\left(\frac{1}{\alpha}\right)}{\gamma\alpha\pi}. \tag{3.7}$$
- Soit $Z \sim S_\alpha(\gamma, 0, \delta)$ avec $1 < \alpha \leq 2$. Alors, la densité de probabilité de Z vérifie le comportement asymptotique suivant :
  $$p_Z(z) \sim \frac{\alpha(1-\alpha)\gamma^\alpha}{\Gamma(2-\alpha)\cos\left(\frac{\pi\alpha}{2}\right)}|z|^{-\alpha-1} \text{ lorsque } |z| \to \infty. \tag{3.8}$$

Les démonstrations de ces propriétés peuvent être consultées dans [168, 147, 121].

Un processus stochastique réel ou complexe $\xi = (\xi_t, -\infty < t < \infty)$ est SαS si et seulement si tout sous-ensemble fini $(\xi_{t_1}, ..., \xi_{t_n})$ de ξ est un vecteur symétrique α-stable. De manière équivalente, toutes les combinaisons linéaires de ξ sont des variables aléatoires SαS [168].

3.1 Modélisation des canaux de communication ULB

Lorsque nous avons commencé à travailler sur ce sujet, les systèmes de communication Ultra Large Band (ULB) à des fréquences extrêmement élevées (60 GHz) étaient en cours de développement [103] et [78, 79, 72]. La modélisation des canaux de communication faisait face à de nouveaux défis liés au comportement modifié de la propagation des signaaux à ces fréquences extrêmes.

Plusieurs travaux sur la caractérisation des canaux ont été réalisés dans la littérature ; nous citons entre autres [134, 112, 111]. Ils sont principalement basés sur des approches déterministes ou des lancers de rayons *ray tracing*. Quand il s'agit de modèles statistiques, les références étaient rares, principalement à cause de la difficulté de modéliser la distribution des retards des trajets [1] [153, 149, 134, 115, 104]. Les modèles statistiques classiques sont basés sur les travaux de Bello [182] et généralement sur les propriétés de stationnarité (WSS=*Wide Sense stationarity*) et des "*diffuseurs non corrélés*" (US=*Uncorrelated scatterers*). Cependant, il a été montré récemment que les zones de stationnarité sont très limitées lorsque la bande passante est large et/ou quand la longueur d'onde est seulement de quelques millimètres [109, p. 88]. En outre, la propriété (US) n'est plus valable quand la résolution temporelle est très élevée voir[93]. Pour remédier à ces problèmes, de nouvelles approches sont nécessaires : par exemple, récemment un groupe préparant le standard IEEE 802.15.3a, propose un modèle du canal ULB sur la base des approches de Saleh et Valenzuela [164]. Si ces modèles se sont révélés adéquats, ils sont difficiles à utiliser pour l'analyse des performances théoriques et le développement des communications numériques adaptées ainsi que pour des solutions en traitement du signal [70, 76, 54].

3.1.1 Modèles statistiques classiques d'un canal de communication

L'idée naturelle consiste à représenter le canal comme une ligne à retard espacé aléatoirement. Dans ce cas, le canal peut être représenté par un filtre linéaire caractérisé par sa réponse impulsionnelle :

$$h(t) = \sum_{k=1}^{N} a_k \delta(t - \tau_k) e^{j\theta_k}. \tag{3.9}$$

Elle est exprimée en fonction de quatre variables aléatoires, a_k, τ_k, θ_k, N modélisant respectivement l'amplitude, le retard, la phase d'un trajet k et N le nombre total des trajets. Dans le modèle (3.9), la principale difficulté est de caractériser les retards des trajets et leurs amplitudes. Les phases sont généralement considérées comme des variables uniformes sur $[0, 2\pi]$ car la longueur du trajet est beaucoup plus grande que la longueur d'onde selon [109]. La discussion sur les retards et les amplitudes montre la difficulté de trouver le bon modèle pour une application donnée. Cette difficulté est encore plus grande lorsque la bande du canal devient plus large et lorsque la fréquence porteuse augmente.

- **Modèles classiques des retards :** Une première tentative de description des temps d'arrivée par une distribution de Poisson a donné des résultats décevants pour le canal 60 GHz [115]. Turin [176], observant que l'hypothèse de Poisson n'a pas été confirmée par l'expérience, a montré que, dans un environnement extérieur et à 488, 1280 et 2980 MHz, les temps

1. Il s'agit de la distribution du temps qu'un rayon met pour atteindre le récepteur au départ de l'émetteur en suivant un trajet donné.

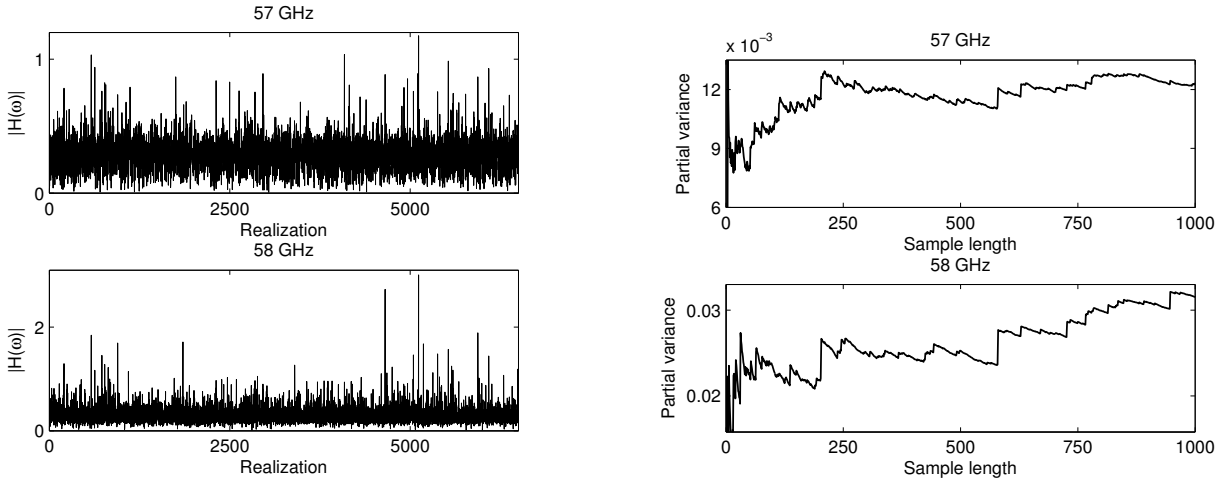
d'arrivée pouvaient être modélisés par un processus de Poisson modifié, où le paramètre de la loi de Poisson change quand un trajet arrive. Il a introduit, en quelque sorte, l'idée derrière la notion de *cluster* et a proposé un modèle qui a été connu plus tard comme le modèle $\Delta - K$ popularisé par Suzuki [171]. Son principe de base est que les trajets arrivent par groupe (cluster), ce qui a amené Saleh and Valenzuela [164] à montrer que le canal peut être modélisé par un mélange de deux distributions de Poisson ; le premier pour l'arrivée d'un cluster et le deuxième pour les arrivées à l'intérieur d'un cluster.

Dans [104], nous avons proposé un modèle basé sur des distributions α-stables pour la modélisation du retard à 60 GHz et quelques justifications théoriques et expérimentales du modèle ont été données.

- **Modèles des amplitudes :** D'un point de vue expérimental, les trajets détectés sont souvent considérés comme le résultat de la superposition de plusieurs rayons incidents. Une multitude de travaux ont été menés pour modéliser ce comportement ; par exemple, une analyse approfondie de la distribution des amplitudes résultantes a été réalisée dans [109, p. 118-128]. Lorsque le nombre de chemins fusionnés est assez grand, les distributions de Rayleigh, de Rice ou de Nakagami sont obtenues pour les amplitudes. Dans d'autres travaux, [163, 156] l'amplitude du multi-trajet a été modélisée par une distribution Log-normale. Molisch dans [56] a étudié la réponse impulsionnelle d'un canal ULB et en a conclu que les amplitudes ne suivent pas une distribution de Rayleigh.
- **Hypothèses classiques de stationnarité :** Jusqu'à présent et dans de nombreux travaux [110, 134, 114, 143, 117, 124, 107], des hypothèses judicieuses de stationnarité ont été adoptées pour modéliser les canaux de communication. Il apparaît cependant que ces hypothèses ne sont pas valables dans notre situation de canal ULB :
  - Dans le contexte des ondes millimétriques, une seule pièce de bureau présentera divers comportements et caractéristiques (voir figure 3.2). Dans les coins et près de l'émetteur, les réponses impulsionnelles présentent un profil de trajets multiples denses où les contributions dominantes proviennent de la réflexion spéculaire [134]. Les résultats de mesure dans [114] montrent une forte dépendance des paramètres statistiques utiles vis-à-vis de l'environnement de propagation. En outre, la zone locale où l'hypothèse WSS est vérifiée est réduite à quelques cm^2 en raison de la faible longueur d'onde.
  - En raison de l'augmentation de la bande passante, davantage de chemins peuvent être résolus. En conséquence, une corrélation entre les composants à trajets multiples peut apparaître et la propriété des diffuseurs non corrélés ne tient pas.

3.1.2 Notre solution : une nouvelle approche

Notre objectif est de donner une solution complète basée sur une modélisation à la fois justifiée théoriquement et validée par les observations. Pour atteindre cet objectif, un modèle basé sur des processus α-stable a été proposé et deux résultats permettant sa mise en place ont été établis ; le premier permet d'estimer efficacement les caractéristiques du canal, à savoir les densités spectrales et le deuxième permet de générer les réponses impulsionnelles du canal. L'originalité de nos travaux consiste à traiter le problème dans le domaine fréquentiel en s'intéressant à la fonction de transfert du canal. Afin d'expliquer la logique de notre approche, revenons à la fonction de transfert du canal $H(\omega)$ qui est naturellement exprimée comme la transformée de



(a) label 1 (b) label 2

FIGURE 3.1 – Réalisation de la fonction de transfert à des fréquences et des emplacements différents.

Fourier de la réponse impulsionnelle :

$$H(\omega) = \int e^{j\omega\tau} h(\tau) d\tau, \tag{3.10}$$

En raison des configurations aléatoires du milieu de transmission, nous considérons que les réponses impulsionnelles sont des trajectoires d'un processus stochastique $(h(\tau,.), \tau \in \mathbb{R})$. Dans ce contexte, la fonction de transfert résultante est également un processus harmonisable représenté par l'intégrale stochastique suivante :

$$H(\omega,.) = \int_{\mathbb{R}} e^{j\omega\tau} d\xi(\tau). \tag{3.11}$$

Le terme $d\xi(\tau)$ est une mesure aléatoire symétrique α-stable construite à partir du processus stochastique $h(\tau,.)$. Le choix d'un modèle α-stable n'a pas été fait au hasard mais il a été inspiré par les observations des données réelles et les mesures. La figure 3.1 montre le caractère impulsif pour deux exemples de $H(\omega,.)$ observés pour deux valeurs différentes de ω. Ce fait est confirmé par un test visuel de variance infinie à droite de la figure 3.1 pour plus détails voir par exemple [141, 26, p. 62-63].

Outre le fait que les processus α-stables sont de bons modèles pour des phénomènes impulsifs et à variance infinie [148], nous montrons qu'ils généralisent de manière naturelle les modèles classiques habituellement utilisés dans le contexte de transmissions classiques. Dans notre situation, l'un des avantages importants de notre approche est qu'il prend en compte les événements rares inhérents aux transmissions à de très hautes fréquences et/ou en milieux confinés. La contribution de notre travail repose sur la caractérisation du processus stochastique $H(\omega,.)$ donné dans (3.11) ce qui évite plusieurs étapes de pré-traitement des données. Nous montrons que dans le cadre d'une transmission ULB, la fonction de transfert aléatoire $H(\omega,.)$, peut être caractérisée par une mesure unique [2] μ définie sur $\mathbb{R}$.

2. Notre travail est inspiré de la terminologie des processus harmonisables où μ est appelée une mesure spectrale

3.1.3 Caractérisation du canal et estimation de ces paramètres

Caractérisation du canal : En se basant sur le fait que le processus ξ est à accroissements indépendants, la mesure aléatoire $d\xi$ est également à accroissements indépendants ; c'est-à-dire pour A et B tels que $A \cap B = \varnothing$, $d\xi(A)$ et $d\xi(B)$ sont indépendants. Dans ce cas, une mesure positive μ peut être construite comme suit :

$$\mu(A) = [d\xi(A), d\xi(A)]_\alpha = \|d\xi(A)\|_\alpha^\alpha. \tag{3.12}$$

A partir de la définition (3.12), nous pouvons montrer que $[d\xi(A), d\xi(B)]_\alpha = \mu(A \cap B)$ pour tous Boreliens A et B (voir[168], [84, 1, 26]). Ce résultat est généralisé à une représentation intégrale de la fonction de covariation du processus $H(\omega, .)$ Comme :

$$\begin{aligned} C(\omega, \omega') &\triangleq [H(\omega, .), H(\omega', .)]_\alpha \\ &= \int_{\mathbb{R}} e^{i((\omega-\omega')\lambda)} \mu\,(d(\lambda)). \end{aligned} \tag{3.13}$$

La preuve de ce résultat est détaillée dans [26]. On montre également que la mesure μ est unique et caractérise la loi du processus stochastique $H(\omega, .)$. En d'autres termes, le canal peut être identifié à travers la caractérisation de la mesure μ. Dans notre cas, cette dernière est à support continu ; il a alors une dérivée Radon-Nykodym ϕ, par rapport à la mesure de Lebesgue sur $\mathbb{R}$.

Résultats préliminaires pour l'estimation de ϕ : L'estimation de ϕ est basée sur les propriétés de la fonction caractéristique du processus harmonisable $H(w, .)$. Cette technique a été initialement suggérée par [167] pour des processus SαS harmonisables à temps continu. Nous l'avons adaptée à notre situation pour prendre en compte l'observation discrète et partielle de $H(\omega, .)$ voir par exemple [44, 26, 91, 142] pour plus de détails. L'idée est inspirée de la généralisation de l'analyse de Fourier en introduisant des noyaux généralisant ceux de Féjer par exemple à savoir les polynômes de Jackson définis par :

$$J^{(N)}(\lambda) = \frac{1}{q_{k,n}} \left(\frac{\sin(\frac{n\lambda}{2})}{\sin(\frac{\lambda}{2})} \right)^{2k} \quad \text{où} \quad q_{k,n} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\frac{\sin(\frac{n\lambda}{2})}{\sin(\frac{\lambda}{2})} \right)^{2k} d\lambda, \tag{3.14}$$

N est un nombre impair fixe tel que : $N = 2k(n-1)+1$, où $n \in \mathbb{N}$ et $k \in \mathbb{N} \cup \left\{\frac{1}{2}\right\}$; quand $k = \frac{1}{2}$, n est choisi comme un entier impair. On sait que dans ce cas il existe une fonction $\mathbf{j}_k$ satisfaisant la somme partielle suivante :

$$J^{(N)}(\lambda) = \sum_{m'=-k(n-1)}^{k(n-1)} \mathbf{j}_k(\frac{m'}{n}) \cos(\lambda m'). \tag{3.15}$$

Il est montré, par exemple dans [142] ou également [44, 26], que :

$$\mathbf{j}_k(u) = \begin{cases} \dfrac{Q_N^{(k)}(nu + k(n-1))}{Q_N^{(k)}(k(n-1))} & \text{if } nu \in \mathbb{Z} \\ 0 & \text{if } nu \notin \mathbb{Z} \end{cases} \tag{3.16}$$

[44, 84, 168]. Dans notre cas, il s'agit plutôt d'une quantité temporelle incluse dans le domaine du retard plutôt que dans le domaine fréquentiel. Nous utiliserons toutefois le terme mesure spectrale dans ce qui suit.

où $Q_N^{(k)}(\tau)$ est le nombre de combinaisons d'entiers $(t_1, t_2, \cdots, t_{2k})$ qui satisfont les conditions $\forall j \in \{1, 2, \cdots, 2k\} \quad 0 \leq t_j \leq n-1$ et $t_1 + t_2 + \cdots + t_{2k} = \tau$. On peut alors définir le noyau $J_N(\lambda)$, qui sera utilisé pour la construction du périodogramme $\hat{I}_N$ des $H(\omega, .)$ observés . Ce noyau vérifie la propriété importante :

$$|J_N(\lambda)|^\alpha = \left|A_N J^{(N)}(\lambda)\right|^\alpha \quad \text{où} \quad A_N = \left(\int_{-\pi}^{\pi} \left|J^{(N)}(\lambda)\right|^\alpha d\lambda\right)^{\frac{-1}{\alpha}}. \tag{3.17}$$

La constante A_N peut être explicitée comme :

$$A_N = \frac{\left(\frac{1}{2}\left(\frac{2}{\pi}\right)^{2k} + \frac{k}{2k-1}\right) n^{2k-1}}{\left(\left(\pi\left(\frac{2}{\pi}\right)^{2\alpha k} + \frac{2\pi\alpha k}{2\alpha k-1}\right) n^{2\alpha k-1}\right)^{\frac{1}{\alpha}}} \tag{3.18}$$

Estimation proprement dite de ϕ : Supposons que nous observons $H(\omega, .)$ en des fréquences angulaires équidistantes $\omega_i = i\frac{\Delta}{\omega}$, pour $i = 1, 2, ...N$ où $\frac{\Delta}{\omega}$ est choisie de façon à respecter le théorème d'échantillonnage. Un estimateur sans biais mais non consistant de ϕ peut être construit comme :

$$\hat{I}_N(\lambda) = C_{p,\alpha} |I_N(\lambda)|^p, \qquad 0 < p < \frac{\alpha}{2}. \tag{3.19}$$

Le terme $\mathrm{I}_N(\lambda)$ est le périodogramme modifié faisant intervenir les polynômes de Jackson $J_N(\lambda)$ suivant :

$$I_N(\lambda) = [\Delta\omega]^{\frac{1}{\alpha}} A_N \mathrm{Re}\left[\sum_{n'=-k(n-1)}^{k(n-1)} \mathbf{j}_k\left(\frac{n'}{n}\right) \exp\left\{-\jmath(n'\Delta\omega\lambda)\right\} H\left(n'\Delta\omega + k(n-1)\Delta\omega\right)\right]. \tag{3.20}$$

La constante de normalisation $C_{p,\alpha}$ est donnée par :

$$C_{p,\alpha} = \frac{D_p}{F_{p,\alpha}[c_\alpha]^{\frac{p}{\alpha}}}, \tag{3.21}$$

avec les constantes habituelles qu'on rencontre en théorie des processus et variables α-stables [148] données par :

$$c_\alpha = \frac{1}{\pi\alpha} \int_0^\pi |\cos(u)|^\alpha du, \tag{3.22}$$

$$D_p = \int_{-\infty}^{\infty} \frac{1 - \cos(u)}{|u|^{1+p}} du = \frac{2}{p}\Gamma(1-p)\cos\left(\pi\frac{p}{2}\right), \tag{3.23}$$

$$F_{p,\alpha} = \int_{-\infty}^{\infty} \frac{1 - e^{-|u|^a}}{|u|^{1+p}} du = 2\frac{\Gamma\left(1 - \frac{p}{\alpha}\right)}{\Gamma(1-p)\cos\left(\pi\frac{p}{2}\right)}. \tag{3.24}$$

Pour obtenir un estimateur consistant de $[\phi(\lambda)]^{\frac{p}{\alpha}}$, le périodogramme est lissé par un noyau de convolution définissant la fenêtre, $W_N(u) = M_N W(M_N u)$, où M_N vérifie $M_N \to +\infty$ et $\frac{M_N}{N} \to 0$ quand $N \to +\infty$. W est un noyau, pair de support $[-1, 1]$. Dans la partie applications et simulations, nous avons utilisé le noyau d'Epanechnikov avec $p = \frac{\alpha}{2.5}$ comme suggéré par exemple

dans [167]. Nous considérons par la suite le périodogramme lissé f_N défini par :

$$f_N(\lambda) = \int_{\mathbb{R}} W_N(\lambda - u)\hat{I}_N(u)du. \tag{3.25}$$

En s'inspirant des résultats de [142] ou plus généralement dans [91], nous avons montré que $f_N(\lambda)$ est un estimateur consistant et asymptotiquement sans biais de $[\phi(\lambda)]^{\frac{p}{\alpha}}$. Nous en déduisons donc un estimateur consistant de ϕ. Ce résultat est annoncé dans le théorème suivant :

Théorème 3.1.1 : Un estimateur asymptotiquement sans biais de $\phi(\lambda)$ est donné par :

$$\hat{\phi}(\lambda) = (f_N(\lambda))^{\frac{\alpha}{p}}. \tag{3.26}$$

Cet estimateur est consistant et sa vitesse de convergence est de l'ordre de $[\frac{M_N}{N}]^{\frac{\alpha}{p}}$

L'algorithme suivant résume la procédure que nous avons mise en place et appliquée pour estimer les paramètres du canal 60 Ghz.

Algorithm 1 : Estimation non paramétrique de $\hat{\phi}(\lambda)$ à partir des fonctions de transfert.

Input : Fonctions de transferts $H_\ell(\omega_i)_{i=1\dots N}$ mesurées aux emplacements $\ell = 1, \dots, L$.

1 Estimer l'indice de stabilité α à partir de $\{H_\ell(\omega)\}_{\ell=1,\dots,L}$ en utilisant [169].

2 Fixer $p = \frac{\alpha}{2.5}$ et calculer , c_α, D_p et $F_{p,\alpha}$ donnée dans (3.22), (3.23) et (3.24).

3 Calculer, $C_{p,\alpha}$ donnée par (3.21).

4 Calculer $\mathbf{j}_k(\frac{n'}{n})$ à partir de (3.15) et A_N à partir de (3.17).

5 **foreach** $\ell = 1 \dots L$ **do**

6 Calculer $I_{N,\ell}(\lambda)$ comme dans (3.20).

7 Calculer $\hat{I}_{N,\ell}(\lambda)$ en utilisant (3.19).

8 **end**

9 Calculer $\bar{\hat{I}}_N$, la moyenne de $\hat{I}_{N,\ell}(\lambda)$ sur $\ell = 1 \dots L$ à partir de (3.19).

10 Lisser $\bar{\hat{I}}_N(\lambda)$ pour obtenir $f_N(\lambda)$ comme dans (3.25).

Output : La densité spectrale $\hat{\phi}(\lambda) = (f_N(\lambda))^{\frac{\alpha}{p}}$.

Pour estimer l'indice de stabilité α, plusieurs méthodes ont été proposées dans la littérature ; nous citons entre autres [173, 166, 165, 178] qui ont introduit des techniques basées sur le maximum de vraisemblance ou se basant sur les quantiles. D'autres estimateurs utilisant la fonction caractéristique ont été introduits par [169]. Nous avons choisi cette dernière méthode car elle est facile à calculer et plus précise si les autres paramètres de la distribution α-stable ne sont pas connus *a priori*, voir par exemple [62].

3.1.4 Lien avec les modèles classiques et algorithmes de simulation.

Pendant nos travaux de modélisation des canaux de communication, nous avons voulu donner une solution générique et originale sans renier les modèles existants éprouvés en pratique et dans la littérature. Dans ce volet de ce travail, nous établissons l'existence d'un lien entre notre modèle et ceux communément utilisés, entre autre les normes actuelles IEEE 802.

Décomposition en séries et lien avec (3.9) : Un des résultats importants qui rend la mesure μ utile en pratique est le fait qu'elle peut être exprimée en terme de loi d'une variable aléatoire.

En effet, en la normalisant par sa masse totale $\mu(\mathbb{R})$ nous définissons $\hat{\mu} : A \longmapsto \frac{\mu(A)}{\mu(\mathbb{R})}$ une mesure de probabilité. Ainsi, par le théorème de représentation, il existe une variable aléatoire unique ϑ de distribution $\hat{\mu}$. Sur la base de l'estimation de la densité $\phi(\lambda)$ discutée plus haut, nous pouvons générer ϑ et l'utiliser pour la décomposition en série du processus harmonisable $H(W,.)$. L'idée est d'utiliser les séries de Lepage, pour plus de détails voir par exemple [170, 148, 84]

Théorème 3.1.2 : La fonction de transfert $H(\omega,.)$ peut être décomposée en série de type Lepage [148] comme :

$$H(\omega,.) \overset{d}{=} (\mu(\mathbb{R})c_\alpha)^{\frac{1}{\alpha}} \sum_{i=1}^{\infty} \gamma_i \Gamma_i^{-\frac{1}{\alpha}} e^{\jmath\omega\vartheta_i}(S_{i,1} + \jmath S_{i,2}). \tag{3.27}$$

L'égalité "$\overset{d}{=}$" signifie l'égalité dans la distribution.

- (γ_i) sont des copies i.i.d de variables aléatoires de Rademacher γ, i.e. $P(\gamma = -1) = P(\gamma = 1) = 1/2$,
- (Γ_i) sont les temps d'arrivée d'un processus de Poisson ; ils suivent une loi Gamma de moyenne i,
- $(\vartheta_i)_i$ sont des copies i.i.d de ϑ,
- $(S_{i,1}, S_{i,2})$ sont des copies i.i.d uniformément réparties sur le cercle unité. Nous pouvons les écrire en notation complexe : $S_{i,1} + \jmath S_{i,2} = e^{\jmath\theta_i}$.

En utilisant les propriétés de convergence en distribution, on peut intégrer (3.27) par rapport à ω de sorte que, pour tout t, on ait :

$$\int_{-\infty}^{\infty} H(\omega,.)e^{-\jmath\omega t}d\omega \overset{d}{=} (\mu(\mathbb{R})C_\alpha)^{\frac{1}{\alpha}} \sum_{i=1}^{\infty} \gamma_i \Gamma_i^{-\frac{1}{\alpha}} e^{\jmath\theta_i} \int_{-\infty}^{\infty} e^{-\jmath\omega(t-\vartheta_i)}d\omega. \tag{3.28}$$

En calculant l'intégrale dans la partie droite de (3.28) et en utilisant la transformée de Fourier inverse dans (3.11), nous pouvons écrire la réponse impulsionnelle aléatoire $h(t,.)$ Comme :

$$h(t,.) \overset{d}{=} (\mu(\mathbb{R})C_\alpha)^{\frac{1}{\alpha}} \sum_{i=1}^{\infty} \gamma_i \Gamma_i^{-\frac{1}{\alpha}} e^{\jmath\theta_i}\delta(t - \vartheta_i). \tag{3.29}$$

Nous pouvons remarquer que la réponse impulsionnelle donnée dans (3.29) a une structure similaire à la réponse impulsionnelle théorique d'un canal linéaire donnée dans (3.9). Il faut également noter que l'égalité de distribution (3.29) est une formule exacte de la réponse impulsionnelle aléatoire du canal.

Somme partielle de (3.29) : Les termes $_i$ sont les temps d'arrivée d'un processus de Poisson ; ils sont donc distribués selon une loi Gamma de moyenne i ; Quand i tend vers l'infini, $_i^{-1/\alpha}$ converge presque sûrement vers 0. La somme infinie (3.29) peut donc être tronquée à un ordre N que l'on choisit suffisamment grand pour avoir une approximation raisonnable en pratique, c'est-à-dire :

$$h(t,.) \overset{d}{=} (\mu(\mathbb{R})C_\alpha)^{\frac{1}{\alpha}} \sum_{i=1}^{N} \gamma_i \Gamma_i^{-\frac{1}{\alpha}} e^{\jmath\theta_i}\delta(t - \vartheta_i), \tag{3.30}$$

ou en d'autre terme, dans le domaine fréquentiel :

$$H_N(\omega) \overset{d}{=} (\mu(\mathbb{R})C_\alpha)^{\frac{1}{\alpha}} \sum_{i=1}^{N} \gamma_i \Gamma_i^{-\frac{1}{\alpha}} e^{\jmath\theta_i} e^{\jmath\omega\vartheta_i}, \tag{3.31}$$

Nous proposons une mesure de la qualité de l'approximation en calculant l'ordre N minimal tel que,

$$\mathbb{P}\left(\exists\omega \in \mathbb{R},\ \ |H_N(\omega) - H(\omega)| > \eta\right) < \varepsilon. \tag{3.32}$$

En d'autres termes, la probabilité, que l'écart entre le processus exact $H(\omega)$ et le processus approché $H_N(\omega)$ soit supérieure à η, doit être inférieure à un risque ε. Pour calculer N, nous avons besoin de la loi de probabilité de $|H_N(\omega) - H(\omega)|$. En suivant la technique proposée dans [139, p. 41-43], nous avons le résultat suivant :

Proposition 3.1.3 : pour tout $N > \frac{2}{\alpha}$ et $\eta > 0$, on a :

$$\mathbb{P}(|H(\omega) - H_N(\omega)| > \eta) \leq \frac{R_N(\alpha)}{\eta^2}, \tag{3.33}$$

où $R_N(\alpha) = \sum_{i=N+1}^{\infty} \frac{1}{[i - \frac{2}{\alpha}]^{\frac{2}{\alpha}}}$ est le reste de la série de Riemann.

L'inégalité (3.32) impose un test d'arrêt sur des fréquences infinies appartenant à un intervalle continu. Cependant, nous ne générons que des trajectoires discrètes du processus. Les fréquences angulaires d'observation sont fixes et notées $\omega_1, \dots, \omega_K$. Le test d'arrêt (3.32) sera naturellement remplacé par :

$$\mathbb{P}(\exists\omega_i \in \{\omega_1, \dots, \omega_K\},\ \ |H_N(\omega_i) - H(\omega_i)| > \eta) < \varepsilon. \tag{3.34}$$

Nous pouvons donc donner une borne supérieure pour le terme de gauche dans (3.34) de sorte que :

$$\mathbb{P}(\exists\omega_i \in \{\omega_1, \dots, \omega_K\},\ \ |H_N(\omega_i) - H(\omega_i)| > \eta) \leq \sum_{i=1}^{K} \mathbb{P}(|H_N(\omega_i) - H(\omega_i)| > \eta). \tag{3.35}$$

En utilisant (3.35) et (3.33), nous pouvons écrire :

$$\mathbb{P}(\exists\omega_i \in \{\omega_1, \dots, \omega_K\},\ \ |H_N(\omega_i) - H(\omega_i)| > \eta) \leq \frac{KR_N(\alpha)}{\eta^2}. \tag{3.36}$$

Enfin, pour avoir un écart maximum η avec un risque ε, on prend N tel que :

$$R_N(\alpha) \leq \frac{\eta^2\varepsilon}{K}. \tag{3.37}$$

En utilisant cette procédure nous avons établi un algorithme qui permet de générer des réponses impulsionnelles caractérisées par la mesure μ.

Nous avons adapté les réponses impulsionnelles simulées à la résolution temporelle limitée des observations expérimentales. En effet, si $I\Delta t$ la durée totale de la réponse impulsionnelle où Δt est le pas des observations, nous générons,

$$h(t) = \sum_{i=0}^{I-1} a_i \delta(i\Delta t), \tag{3.38}$$

Algorithm 2 : Générateur des réponses impulsionnelles d'un canal.

Input : La densité spectrale $\hat{\phi}(\lambda)$ et α estimés avec l'algorithme 1.

Calculer C_α et approcher la masse totale $\mu(\mathbb{R}) = \int \phi(\lambda) d\lambda$. ;

Trouver le seuil N comme dans (3.37). ;

do : ;

Générer $\Gamma_i = e_1 + \cdots + e_i$ où (e_i) sont des variables i.i.d. de loi exponentielle de moyenne 1. ;

Générer γ_i des copies i.i.d. de γ. ;

Générer ϑ_i à partir de μ . ;

Calculer la somme donnée dans (3.39). ;

end do : ;

Output : Une réponse impulsionnelle.

où a_i représente la superposition de tous les rayons qui arrivent entre deux instants successifs, elle est donnée par :

$$a_i = \sum_{j=1}^{N} \gamma_j \Gamma_j^{-\frac{1}{\alpha}} e^{j\theta_j} \mathbb{1}_{\vartheta_j \in [i\Delta t, (i+1)\Delta t[} \tag{3.39}$$

3.1.5 Validation du modèle et application au canal 60 GHz

Nous appliquons notre modèle théorique et les outils statistiques que nous avons développés aux fonctions de transferts mesurées dans une bande de fréquences entre 57 à 59 GHz. Nous montrons le bon comportement de notre modèle, sa capacité à représenter des situations très différentes et le bon ajustement entre les étalements des retards réels et simulés.

Configuration expérimentale des mesures : Un analyseur de réseau dédié, créé à l'IEMN [110], permet la vectorisation des fonctions de transferts à des fréquences sur la bande allant de 57 à 59 GHz. Ainsi la bande passante de 2 GHz est mesurée par pas de 1.25MHz. Cette plage de fréquence donne une résolution temporelle de 0.5ns sur un intervalle entre 0 et 800 ns. L'expérimentation a été effectuée dans l'environnement décrit dans la figure 3.2. Les mesures ont été effectuées sur 26 sites différents ; sur chaque site, le récepteur s'est déplacé sur plus de 250 positions sur une règle graduée par pas de 2mm qui est moins que la moitié de la longueur d'onde.

Application aux mesures et validation du modèle : Pour valider notre modèle, nous l'avons confronté aux réponses impulsionnelles observées dans le cas du canal 60 Ghz. Nous estimons d'abord l'indice de stabilité α ; on obtient $\alpha = 1.83$. Nous estimons par la suite la densité $\hat{\phi}$ de la mesure spectrale μ en utilisant l'algorithme 1. Dans la figure 3.3, nous présentons des réponses impulsionnelles simulées en utilisant l'algorithme 2. Pour chaque réponse impulsionnelle, nous avons utilisé (3.31) avec $N = 30000$ pour un risque $\eta = 0.05$ et une précision $\varepsilon = 0.01$.

La figure 3.3 montre la capacité de notre modèle à représenter des situations très différentes contrairement aux approches classiques. Une seule campagne de mesures suffit pour représenter le comportement du canal dans toute la pièce. Cela résulte du fait que la propriété de stationnarité au sens large du second ordre n'est pas requise pour notre modèle. Un autre fait important est que notre modèle est capable de représenter les temps d'arrivée des trajets avec une grande

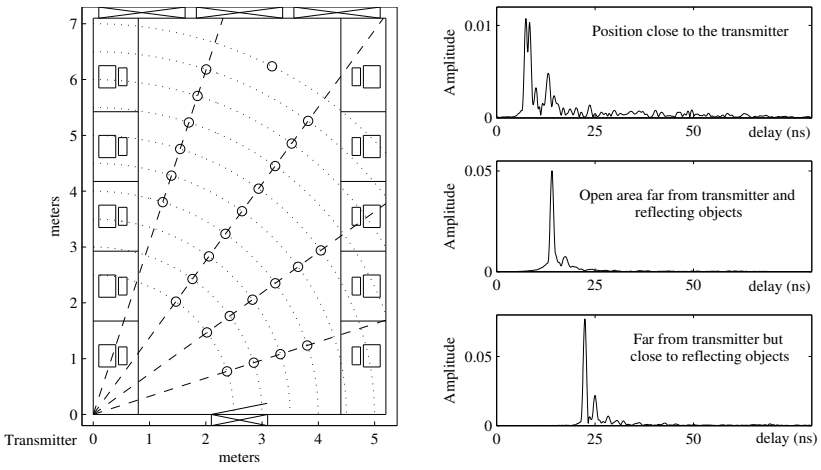


Figure 3.2 – Configurations des salles de mesures.

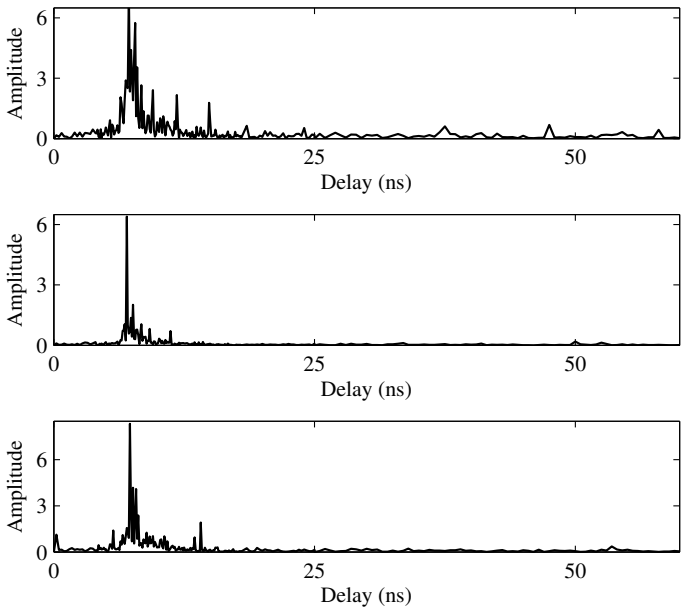


Figure 3.3 – Exemple de réponses impulsionnelles simulées.

précision. C'est une caractéristique importante lorsque des canaux à bande ultra large sont considérés.
Pour confirmer davantage la validité de notre modèle, nous avons comparé l'étalement des retards (*RMS Delay spread*) des mesures et des simulations ; ce paramètre, habituellement utilisé pour la validation des modèles de canaux, est donné par :

$$RMS = \sqrt{\frac{\sum t^2|h(t)|^2 - (\sum |th(t)|)^2}{\sum |h(t)|^2}}. \tag{3.40}$$

Nous présentons dans la figure 3.4 la densité de probabilité de l'étalement des retards des réponses impulsionnelles observées et de celles générées par notre modèle. Nous considérons deux situations : lorsque la mesure μ est estimée à partir de l'ensemble des 6500 mesures (cas 1) et lorsqu'elle est estimée à partir de 250 mesures choisies au hasard. Bien que nous puissions remarquer une différence dans les formes de densité, nous pouvons conclure que notre modèle représente efficacement la variabilité des réponses impulsionnelles réelles dans la pièce. On peut notamment remarquer que la queue des distributions, représentant des situations rares est très similaire. On voit aussi que 250 mesures sont suffisantes pour estimer la mesure μ et pour caractériser pleinement le canal. Enfin, nous montrons dans la figure 3.4 un exemple de densité de probabilité de la partie réelle, de la partie imaginaire et du module de $H(\omega_i)$ pour $\omega_i = 57GHz$ et $\omega_i = 58GHz$.

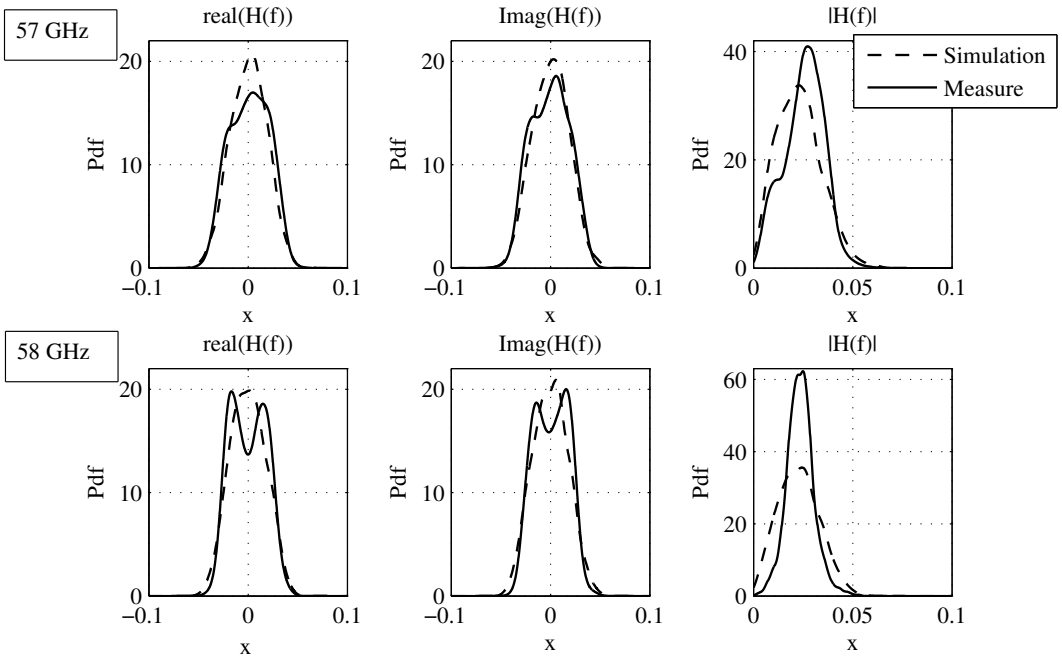


Figure 3.4 – Exemples de densité de probabilité $H(\omega)$ mesuré pour $\omega = 57$ GHz et $\omega = 58$ GHz.

Le fait de représenter la fonction de transfert par un processus aléatoire α-stable permet de s'affranchir de l'hypothèse de stationnarité WSS ou l'hypothèse de réflecteurs non corrélés au sens large. Avec une seule mesure positive, nous pouvons modéliser le comportement du canal dans tout l'environnement car les processus stables sont bien adaptés pour représenter des phénomènes très variables.

Des recherches supplémentaires restent encore à explorer notamment la généralisation de l'hypothèse d'accroissements indépendants au cas non indépendants. Cette généralisation est en cours de publication voir [1, 2] et permettra de prendre en compte des zones spatiales plus larges et de donner une représentation de l'évolution spatio-temporelle du canal.

3.2 Modélisation du canal dans un contexte évolutif ou mobile

Une difficulté essentielle de la modélisation statistique des canaux réside dans la capacité à représenter les évolutions spatio-temporelles. Cette variabilité est d'autant plus importante lorsqu'il s'agit de modéliser les communications dans des environnements en évolution rapide ; d'où la nécessité d'adapter les modèles existants à de telles situations. Un autre défi majeur est le fait que la spécification d'un tel modèle nécessite une vaste campagne de mesures et des traitements gourmands en temps de calcul. En raison de la complexité des modèles et de l'évolution rapide possible, cette tâche est quasi impossible pour une utilisation en temps réel ; en particulier, les noeuds ou capteurs de faible complexité. Dans ce contexte, nous avons introduit un modèle générique capable de s'adapter aux changements rapides de l'environnement tout en tenant compte de la dépendance spatiale de la fonction de transfert du canal et, d'autre part, de la variation du modèle en fonction de la structure de dépendance en fréquence. De ce dernier point de vue, de nombreux travaux ont décrit la dépendance en fréquence ; on cite par exemple [100] qui utilise un modèle autorégressif d'ordre deux pour la génération de la réponse en fréquence du canal ULB. Nous supposons également que le canal est représenté avec précision par un modèle α-stable comme nous l'avons proposé plus haut [44].

Pour reformuler le problème, nous supposons que le milieu de transmission est subdivisé en n régions R_x avec $x = 0, 1, \dots, n$. Nous considérons le processus stochastique $\mathcal{H}_x$ décrivant la fonction de transfert dans une région. Nous supposons que la dépendance spatiale a une mémoire de longueur p. Nous aurons donc les espérances conditionnelles suivantes :

$$\mathcal{H}_{x+1} = \mathbf{f}(\mathcal{H}_x, \dots, \mathcal{H}_{x-p}) + \Upsilon, \tag{3.41}$$

où $\mathbf{f}$ est une fonction inconnue et $\Upsilon = \mathcal{H}_{x+1} - \mathbb{E}\left(\mathcal{H}_{x+1} \mid \mathcal{H}_x, \dots, \mathcal{H}_{x-p}\right)$ tous deux ne dépendent pas des emplacements ou de la zone spatiale. L'équation (3.41) peut être vue comme un modèle *boîte noire* qui ne suppose aucun *a priori* sur l'environnement. Afin de comprendre la dépendance en fréquence, nous présentons le modèle des fonctions de transfert observées à $w_1, \dots, w_m$, comme suit :

$$\text{pos x+1} \begin{cases} \mathcal{H}_{x+1}(w_1) = \mathbf{f}(\mathcal{H}_x(w_1), \dots, \mathcal{H}_{x-p}(w_1)) + \Upsilon_x(w_1) \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ \mathcal{H}_{x+1}(w_m) = \mathbf{f}(\mathcal{H}_x(w_m), , \dots, \mathcal{H}_{x-p}(w_n)) + \Upsilon_x(w_m) \end{cases}$$

Puisque la fonction $\mathbf{f}$ a capturé la structure de dépendance spatiale, il est donc plus logique de supposer que la structure de dépendance de Υ est liée seulement aux fréquences. Inspirés des travaux de [100], nous postulons que pour une position fixe, la dépendance en fréquence de Υ peut être décrite par un schéma autorégressif.

$$\Upsilon_x(w_i) = \sum_{k=1}^{q} \theta_k \Upsilon_x(w_{i-k}) + \varepsilon_{x,i} \tag{3.42}$$

Afin de prendre en compte la nature impulsive, nous supposons que les erreurs $\varepsilon_{x,i}$ sont des

bruits blancs α-stables.

La principale difficulté dans le modèle boîte noire (3.41), est le *fléau de la dimension* qui survient lors de l'estimation de la fonction non paramétrique $\mathbf{f}$ lorsque p devient grand. Afin de surmonter un tel inconvénient, nous supposerons qu'au lieu de $(\mathcal{H}_x, \dots, \mathcal{H}_{x-p})$, la fonction de transfert $\mathcal{H}_{x+1}$ dépend d'une combinaison linéaire de ce vecteur . Cela réduit $\mathbf{f}$ à une fonction d'une seule variable ce qui évitera le problème de la dimension. Cela nous a conduit à proposer le modèle suivant :

$$\mathcal{H}_{x+1}(w_i) = \mathbf{f}(\sum_{k=0}^{p} \eta_k \mathcal{H}_{x-k}(\omega_i)) + \sum_{k=1}^{q} \theta_k \Upsilon_x(w_{i-k}) + \varepsilon_{x,i} \tag{3.43}$$

Pour résumer, $(\mathbf{f}, \eta)$ représentent la dépendance spatiale, le coefficient θ décrit la dépendance linéaire en fréquence et $\varepsilon_{x,i}$ les innovations aléatoires et impulsives du processus $\mathcal{H}_x$. Ce modèle est très général et permettra d'utiliser des outils statistiques connus dans la littérature sous le nom de modèle *semi-paramétrique* partiellement linéaire. Ils ont été largement étudiés dans la littérature; nous citons entre autres [130], [133], [127], [102] ... Pour la compatibilité avec le formalisme semi-paramétrique nous réécrivons de manière équivalente l'équation (3.43) comme suit : on commence par prendre $N = nm$ où n est le nombre de positions et m est le nombre de fréquences observées. Pour chaque $x = 1, \dots, n-1$, pour chaque $i = 1, \dots, m$ on prend $s = (x-1)m + i$ et on note :

$$\begin{aligned}
Y_s &= \mathcal{H}_{x+1}(w_i), \\
U_s &= [\mathcal{H}_x(w_i), \dots, \mathcal{H}_{x-p}(w_i)]^\dagger, \\
V_s &= [\Upsilon_x(w_{i-1}), \dots, \Upsilon_x(w_{i-q})\,]^\dagger, \\
X_s &= [U_s^\dagger, V_s^\dagger]^\dagger, \\
\varepsilon_s &= \varepsilon_{x,i}
\end{aligned}$$

où † indique la transposition d'une matrice. Cela conduira aux notations habituelles d'un modèle semi-paramétrique *SIM* (Single Index Model) :

$$Y_s = X_s^\dagger \boldsymbol{\theta} + \mathbf{f}(X_s^\dagger \eta) + \varepsilon_s \tag{3.44}$$

où $\boldsymbol{\theta} = (\mathbf{0}_{p+1}, \quad \theta_1, \dots, \theta_q)$ et $\eta = (\eta_0, \dots, \eta_p, \quad \mathbf{0}_q)$ et $\mathbf{0}_p$ est le vecteur nul de ^{p}.

Les motivations classiques de l'utilisation de ce genre de modèle (3.44) sont détaillées entre autres dans [127], [102], [67]. Dans notre contexte, de nombreuses indications suggèrent l'utilisation du modèle (3.44); par exemple les figures Fig.3.5 et Fig.3.6 illustrent les structures semi-linéaires et autoregressives de la dépendance spatiale et fréquentielle. Le schéma autorégressif a été inspiré de ce qui se faisait dans la littérature [100]. La théorie des modèles semi-paramétriques a été largement étudiée dans le cas du modèle gaussien ou de second ordre en général. Quand on avait abordé ce travail, aucun résultat n'était connu dans le cas de bruits à variance infinie ou en particulier α-stables. Dans les paragraphes qui suivent nous détaillons une procédure d'estimation et de caractérisation de ces modèles dans le cas α-stable.

3.2.1 Procédure d'estimation et de caractérisation du modèle évolutif

Nous nous intéressons ici au modèle (3.44) où $\boldsymbol{\theta}$ et η sont des vecteurs de paramètres inconnus. Pour des raisons de simplicité, nous supposons que la fonction $\mathbf{f}(\cdot)$ est deux fois différentiable. Les variables aléatoires $\{\varepsilon_s\}$ sont une suite d'erreurs avec $\mathbb{E}[\varepsilon_s \mid X_s] = 0$. Un grand nombre de travaux de la littérature ont discuté ces modèles, notamment lorsque les erreurs sont des variables

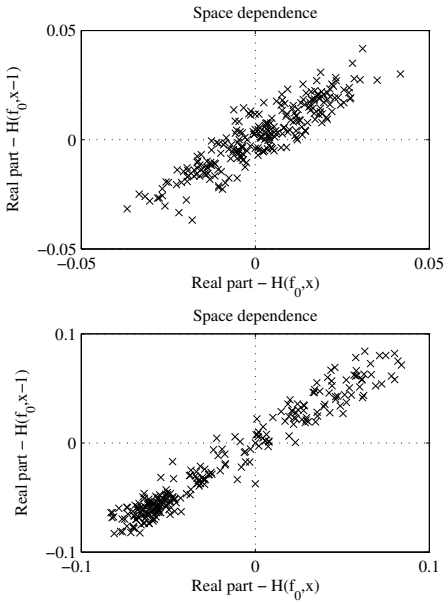


FIGURE 3.5 – Le pas de fréquence est de 5 MHz.

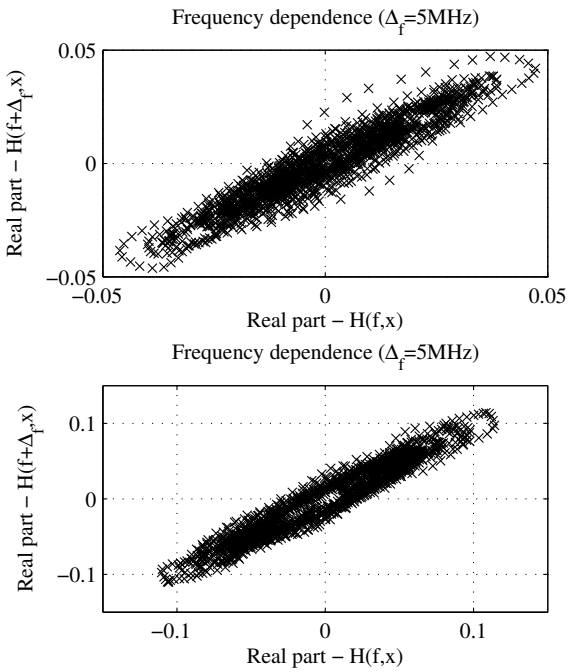


FIGURE 3.6 – 2.5mm distance entre x and x+1

aléatoires ayant des moments de second ordre ; pour plus de détails voir par exemple [127] et [67]. Nous supposons que les ε_s sont i.i.d. symétriques α-stables ayant un paramètre d'échelle fixe σ. Afin d'estimer les paramètres inconnus η, θ et la fonction $\mathbf{f}$, nous introduisons la notation suivante :

$$\mathbf{f}_{1\eta}(u) = \mathbb{E}[Y_s \mid X_s^\dagger \eta = u], \qquad \text{et} \qquad \mathbf{f}_{2\eta}(u) = \mathbb{E}[X_s \mid X_s^\dagger \eta = u], \tag{3.45}$$

Cette décomposition est inspirée de la décomposition de l'espérance conditionnelle $\mathbf{f}(u) = \mathbb{E}[Y_s - X_s^\dagger \theta \mid X_s^\dagger \eta = u]$ qui mène facilement à la formule :

$$\mathbf{f}(u) = \mathbf{f}_{1\eta}(u) - (\mathbf{f}_{2\eta}(u))^\dagger \theta \tag{3.46}$$

À partir des espérances conditionnelles (3.45) nous proposons une estimation à noyau de type Nadarya-Watson [180, 181]. Pour η donné, nous estimons d'abord $\mathbf{f}_{1\eta}(\cdot)$ et $\mathbf{f}_{2\eta}(\cdot)$ par :

$$\hat{\mathbf{f}}_{1\eta}(u) = \frac{\sum_{s=1}^{N} K_h(X_s^\dagger \eta - u) Y_s}{\sum_{s=1}^{N} K_h(X_s^\dagger \eta - u)} \qquad \text{et} \qquad \hat{\mathbf{f}}_{2\eta}(u) = \frac{\sum_{s=1}^{N} K_h(X_s^\dagger \eta - u) X_s}{\sum_{s=1}^{N} K_h(X_s^\dagger \eta - u)}$$

Avec $K_h = \frac{1}{h} K(\frac{x}{h})$ et K est un noyau : c'est-à-dire que K est une densité de probabilité paire non négative (par exemple le noyau Gaussien $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$). L'estimation complète de la fonction $\mathbf{f}$ a donc besoin de l'estimation des paramètres η et θ. Pour cela, nous utilisons une méthode de

type moindre carrés. D'abord définissons les notations suivantes :

$$\overline{Y}_{s\eta} = Y_s - \hat{\mathbf{f}}_{1\eta}(X_s^\dagger \eta) \quad , \quad \overline{X}_{s\eta} = X_s - \hat{\mathbf{f}}_{2\eta}(X_s^\dagger \eta), \quad \text{et} \quad S_N(\theta, \eta; h) = \sum_{s=1}^{N} (\overline{Y}_{s\eta} - \overline{X}_{s\eta}^\dagger \theta)^2$$

L'estimation consiste à minimiser $S_N(\theta, \eta, h)$ en fonction (θ, η, h). On remarque tout d'abord que, pour un (η, h) fixe, l'estimateur des moindres carrés de θ peut être déduit à l'aide des techniques classiques des moindres carrés ; il est donné par :

$$\hat{\theta}(\eta, h) = \left(\sum_{s=1}^{N} \overline{X}_{s\eta} \overline{X}_{s\eta}^\dagger \right)^+ \sum_{s=1}^{N} \overline{X}_{s\eta} \overline{Y}_{s\eta}, \tag{3.47}$$

où $(\,.\,)^+$ désigne le pseudo-inverse d'une matrice. On estime ensuite (η, h) par $(\hat{\eta}, \hat{h})$ en minimisant,

$$\hat{S}_N(\eta, h) = \sum_{s=1}^{N} (\overline{Y}_{s\eta} - \overline{X}_{s\eta}^\dagger \hat{\theta}(\eta, h))^2 \tag{3.48}$$

Inspiré de l'équation (3.46), nous proposons l'estimateur non paramétrique de $\mathbf{f}(\cdot)$:

$$\hat{\mathbf{f}}(u) = \hat{\mathbf{f}}_{1\hat{\eta}}(u) - (\hat{\mathbf{f}}_{2\hat{\eta}}(u))^\dagger \hat{\theta}(\hat{\eta}, \hat{h}). \tag{3.49}$$

Lorsque le paramètre d'échelle σ est inconnu, il peut être estimé par la méthode des moments fractionnaires comme suit :

$$\hat{\sigma} = C_\alpha(\rho) \left(\frac{1}{N} \sum_{s=1}^{N} \left| \overline{Y}_{s\hat{\eta}} - \overline{X}_{s\hat{\eta}}^\dagger \hat{\theta}(\hat{\eta}, \hat{h}) \right|^\rho \right)^{\frac{1}{\rho}} \quad \text{avec} \quad C_\alpha(\rho) = \left(\frac{\alpha \sqrt{\pi}\, \Gamma(\frac{-\rho}{2})}{2^{\rho+1}\, \Gamma(\frac{1+\rho}{2})\, \Gamma(\frac{-\rho}{\alpha})} \right)^{\frac{1}{\rho}}, \tag{3.50}$$

où $1 < \rho < \alpha$ et $C_\alpha(\rho)$ est une constante universelle ne dépendant que de α et de ρ [148, 84].

Le mode de convergence et la consistance de ces estimateurs ont été étudiés dans le cas du second ordre, le lecteur trouvera un aperçu détaillé dans la littérature voir par exemple [67].

3.2.2 Procédure de tests de validité

Dans cette section, nous présentons des techniques statistiques pour tester l'adéquation des modèles semi-paramétriques de type (3.44). Nous procéderons par analogie au contexte de second ordre. Ainsi, nous nous concentrons sur la fonction moyenne conditionnelle définie pour $u \in {}^{p+q+1}$ par :

$$m(u) = \mathbb{E}[Y_s \mid X_s = u)]$$

Nous cherchons d'abord à vérifier l'existence d'une structure de dépendance spatiale puis déterminer les zones spatiales à l'intérieur desquelles le modèle que nous proposons est stationnaire. Nous construisons donc deux tests permettant de vérifier cela.

- Dans un premier temps, nous construisons un test permettant de vérifier si les évolutions des fonctions de transfert peuvent être réduites à l'évolution spatiale, sans tenir compte de la dépendance fréquentielle. Nous cherchons donc à tester l'hypothèse nulle suivante :

$$\begin{aligned} (\mathbb{H}_0) \quad &: \quad m(x) = \mathbf{f}(x^\dagger \eta), \\ (\mathbb{H}_1) \quad &: \quad m(x) = \mathbf{f}(x^\dagger \eta) + \Delta(x) \text{ for all } x \in {}^{p+1}, \end{aligned}$$

  où $\mathbf{f}(\cdot)$ est une fonction inconnue sur $\mathbb{R}$, η est un vecteur de paramètres inconnus et Δ est toute fonction régulière définie sur ${}^{p+1}$. Sous l'hypothèse nulle $(\mathbb{H}_0)$ nous avons donné des

techniques d'estimation de η et $\mathbf{f}(\cdot)$ discutées plus haut (section 3.2.1); nous aurons les estimations simplifiées :

$$\hat{\mathbf{f}}(X_s^\dagger\eta) = \frac{\sum_{t=1}^{N} Y_t K_h((X_s - X_t)^\dagger\eta)}{\sum_{u=1}^{N} K_h(X_s - X_u)^\dagger\eta)}, \qquad \hat{\eta} = \arg\min_{(\eta,h)} \sum_{s=1}^{N}(Y_s - \hat{\mathbf{f}}(X_s^\dagger\eta))^2 \quad \text{et} \quad \hat{\sigma} = C_\alpha(\rho)\left(\frac{1}{N}\sum_{s=1}^{N}|\hat{\varepsilon}_s|^\rho\right)^{\frac{1}{\rho}}$$

Les erreurs $(\varepsilon_s)_s$ sont supposées avoir une distribution symétrique α-stable et sont estimées par $\hat{\varepsilon}_s = Y_s - \hat{\mathbf{f}}(X_s^\dagger\hat{\eta})$. Par analogie avec le cas du second ordre et pour la compatibilité avec la variance infinie, nous avons suggéré la statistique de test suivante :

$$\mathcal{L} = \frac{\sum_{s=1}^{N}|\hat{Y}_s|}{\hat{\sigma}} \tag{3.51}$$

De nombreuses améliorations sont encore nécessaires dans le cas des variables α-stables, notamment la distribution asymptotique et la robustesse de la statistique $\mathcal{L}$. Pour les tests d'hypothèses dans le cas des modèles semi-paramétriques du second ordre, le lecteur trouvera une riche littérature dans [136], [133], [95].

- **Test de la linéarité partielle du modèle** : Il s'agit d'une extension naturelle du modèle (SIM : *Single Index Model*) aux situations où le terme de la régression contient une composante linéaire. Le test suivant évalue si la dépendance à l'espace et à la fréquence est suffisante ($\mathbb{H}_0$) ou s'il serait plus précis de prendre en compte d'autres phénomènes, par exemple une dépendance non linéaire en fréquence. Cela revient au test d'hypothèses suivant :

$$\begin{aligned}(\mathbb{H}_0) \quad &: \quad m(x) = x^\dagger\boldsymbol{\theta} + \mathbf{f}(x^\dagger\eta)\\ (\mathbb{H}_1) \quad &: \quad m(x) = x^\dagger\boldsymbol{\theta} + \mathbf{f}(x^\dagger\eta) + \Delta(x),\end{aligned}$$

pour tout $x \in {}^{p+q+1}$. Ce problème de test a été étudié pour le cas du second ordre par [127], [102]. Avec le même raisonnement que dans le modèle *single index* et en utilisant les techniques d'estimation sous l'hypothèse ($\mathbb{H}_0$) présentée dans la section 3.2.1, nous proposons la statistique de test $\mathcal{L}$ donnée dans (3.51)avec $\widehat{\widetilde{Y}}_s = Y_s - X_s^\dagger\hat{\boldsymbol{\theta}} - \hat{\mathbf{f}}(X_s^\dagger\hat{\eta})$,. Les quantités $\hat{\boldsymbol{\theta}}, \hat{\eta}$, $\hat{\mathbf{f}}(\cdot)$ et $\hat{\sigma}$ étant respectivement les estimateurs donnés dans (3.47), (3.48), (3.49) et (3.50).

L'un des principaux problèmes pour évaluer la compatibilité des modèles est d'examiner la puissance du test et de niveau :

$$\begin{aligned}\alpha &= \mathbb{P}(\mathcal{L} > l_\alpha \mid \text{ tel que } \mathbb{H}_0 \text{ " vraie" })\\ \beta &= \mathbb{P}(\mathcal{L} > l_\alpha \mid \text{ tel que } \mathbb{H}_1 \text{ " vraie" })\end{aligned}$$

où l_α sera tabulée à partir de la distribution de $\mathcal{L}$. Pour cette application, des approximations Monte-Carlo ont été réalisées mais il convient de noter qu'une étude approfondie de ses propriétés asymptotiques et sa robustesse est nécessaire.

3.3 Modélisation de l'interférence et récepteurs optimaux

Les réseaux de capteurs et l'internet des objets (IoT) sont devenus l'une des solutions majeures pour relever, rassembler et analyser des données. La plupart des solutions classiques dans ce domaine se basent sur les statistiques de second ordre et le théorème central limite qui mènent souvent à des modèles Gaussiens. Ces derniers ne conviennent pas dans un environnement impulsif où l'influence des objets interférants est considérable. Ceci est fortement limitant quand les événements rares sont ceux que nous voulons représenter. Ces considérations mènent naturellement aux distributions α-stables pour remplacer les modèles gaussiens. Dans ce contexte, le bruit est une composante importante dans les communications et rend la question des interférences de plus en plus cruciale. Généralement, l'interférence liée aux muti-utilisateurs peut être considérée naturellement comme une superposition de variables aléatoires indépendantes et identiquement distribuées. Avec la densification des réseaux, le nombre d'utilisateurs devient plus grand, par conséquent, en cas de variance finie (signaux non impulsifs), le théorème central limite suggère une approximation par des distributions gaussiennes qui donnent d'ailleurs souvent des résultats précis. Toutefois, la configuration *adhoc* induit naturellement l'apparition de signaux impulsifs et modifie la loi perturbation liée aux accès multiples MAI (*Multiple Acces Interference*). Dans ce qui suit, nous discutons la modélisation des interférences avec des distributions α-stables ; d'autre travaux ont été consacrés à cette question dans [155, 141, 60, 51, 52, 49, 42].

3.3.1 Modélisation des interférences d'accès multiples

Nous considérons un système ULB et une configuration *ad hoc*. Nous supposons que le récepteur est situé au centre d'un cercle C de rayon R. Soit N la variable aléatoire représentant le nombre d'utilisateurs actifs et interférants à l'intérieur du cercle C.

Au niveau de l'émetteur : Chaque utilisateur k envoie un signal $S^{(k)}(t)$ répété N_s fois générant une trame de longueur $N_s T_s$. Quand la méthode de transmission est, par exemple, de type TH-PPM (*Time Hoping*, *Pulse Position Modulation*), les transmissions sont supposées être asynchrones. Si T_s et T_c représentent, respectivement la durée de la trame envoyée et le temps du traitement de celle-ci, le signal transmis $S^{(k)}(t)$ par l'utilisateur k est donné par :

$$S^{(k)}(t) = \sum_{j=-\infty}^{+\infty} \sqrt{E^{(k)}} w\left(t - jT_s - c_j^{(k)} T_c - a_j^{(k)} \varepsilon - \delta^{(k)}\right), \tag{3.52}$$

- Le terme $E^{(k)}$ représente l'énergie de l'impulsion transmise par l'utilisateur k.
- Le terme $w(t) = \sum_{j=1}^{N_s} w_p\left(t - jT_s - c_j^{(k)} T_c\right)$ est l'impulsion émise par unité d'énergie ; elle peut être vue comme N_s répétitions d'une impulsion de base $w_p(t)$.
- Afin de permettre au canal d'être partagé par de nombreux utilisateurs sans trop de collisions, chaque utilisateur se voit attribuer un décalage temporel $c_j^{(k)}$.
- Le terme ϵ est le décalage de base introduit par la PPM et les $a_j^{(k)}$ sont les valeurs binaires codées par la j^{eme} pulsation.
- La valeur $\delta^{(k)}$ représente le décalage ou retard de transmission entre l'utilisateur k et le

récepteur.

Plus de détails sont disponibles dans notre article original [49]. Nous supposons que la communication souhaitée et celle correspondant à l'utilisateur 0 qui envoie un signal utile $S^{(0)}(t)$. Les autres utilisateurs sont considérés comme interférants.

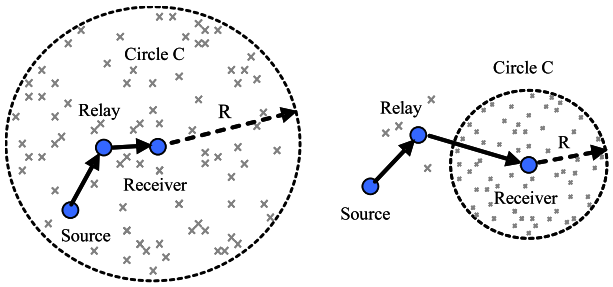


FIGURE 3.7 – Exemple de scénarios d'interférences

Au niveau du récepteur : Le signal reçu est corrélé avec une *forme d'onde de référence* $m(t)$, pour plus de détails voir notre article original [49]. Nous nous sommes intéressés donc à la variable aléatoire :

$$Z = \int_0^{N_S T_S} \sum_{k=0}^{N} \left(S^{(k)}(t) * h^{(k)}(t) + n(t) \right) m(t) dt, \tag{3.53}$$

où $*$ représente la convolution, $h^{(k)}$ est la réponse impulsionnelle du canal et $n(t)$ est un bruit blanc gaussien.

Modèle pour la variable Z : La variable Z peut être écrite comme somme de 4 variables [113] décrivant : la contribution du signal utile, l'effet de l'interférence liée au multi-trajet, l'interférence liée au bruit multi-utilisateur et un bruit blanc Gaussien. Pour simplifier et sans perdre de généralité, nous supposons que la composante muti-trajet est négligeable. Dans ce cas l'équation (3.53) devient :

$$Z = \sum_{k=0}^{N} \gamma_k \int_0^{N_s T_s} \left(S^{(k)}(t - \tau_k) + n(t) \right) m(t) \, dt \tag{3.54}$$

où γ_k est l'atténuation du canal et τ_k le retard de propagation. En enlevant le terme du bruit $n(t)$ et le signal souhaité $S^0(.)$ dans l'équation (3.54), on obtient la variable aléatoire décrivant l'interférence multi-utilisateur que l'on peut écrire :

$$Z_u = \sum_{k=1}^{N} \gamma_k \psi_k. \tag{3.55}$$

- Les $(\psi_k)_{k=1,\cdots,N}$ sont des variables aléatoires décrivant la corrélation entre les signaux des interférences et le signal de référence $m(t)$.
- Les $(\gamma_k)_{k=1,\cdots,N}$ sont des variables aléatoires positives, i.i.d. décrivant l'atténuation due au milieu de transmission.

Supposons maintenant que nous soyons placés dans un trafic dense où N est très grand. Une approche intuitive consisterait à utiliser le théorème central limite classique et approcher Z par

des variables Gaussiennes. Il a été montré, par exemple dans [80], que cette approximation gaussienne n'est pas complètement vraie dans notre contexte. Par ailleurs, les termes γ_k peuvent avoir une nature impulsive car l'atténuation ici est relative au lien souhaité γ_0. En effet, si le lien souhaité est très éloigné mais que celui qui interfère est proche, les valeurs relatives observées de γ_k peuvent être "très grandes" [3]. Ces événements rares sont très importants dans notre contexte et donnent un caractère impulsif aux atténuations. Le théorème central limite généralisé peut être utilisé (voir [141, p. 22] ou [148, p. 9]) pour aboutir à des distributions ayant un domaine d'attraction α-stable. Dans [49], nous nous sommes intéressés à l'impact de la configuration *ad hoc* sur la distribution du bruit multi-utilisateur dans un système IR-ULB. Le cadre général classique est proposé dans [60] et une application à la radio cognitive avec une loi α-stable tronquée est discutée dans [42].

Une nouvelle approche : Nous supposons que le nombre d'interférants actifs suit un processus de Poisson d'intensité λ. Dans ce cas, le nombre d'interférants actifs à l'intérieur du cercle C est donné par :

$$\mathbb{P}(\kappa = k) = \frac{e^{-\lambda\pi R^2}\left(\lambda\pi R^2\right)^k}{k!} \tag{3.56}$$

λ est le nombre moyen d'interférants par unité de surface(ou de volume) ; ce paramètre est lié à la structure du réseau. Cette idée a été également utilisée dans [155, 85]. L'hypothèse de distribution poissonnienne est valable pour de nombreuses solutions et couches physiques différentes (DS-CDMA, FH-CDMA) [155] ou TH-PPM-UWB [49]. Nous avons commencé par montrer le résultat suivant :

Proposition 1 : La variable aléatoire Z_u décrivant l'interférence due aux utilisateurs multiples peut être approchée par une variable α-stable de paramètres vérifiant les propriétés suivantes :

- L'indice de stabilité dépend du coefficient d'atténuation de la propagation dans le milieu de transmission a. il est égal à $\alpha = \frac{4}{a}$.
- le paramètre d'échelle dépend de la nature de l'impulsion envoyée et de la répartition des interférants.

$$\sigma = -\lambda\pi \int_0^{+\infty} \frac{d\phi_\psi}{du}(u)\, u^{-\frac{4}{a}} du \tag{3.57}$$

La démonstration complète de ce résultat est détaillée dans [49]. Nous en donnons ici les grandes lignes. D'abord, en utilisant la répartition poissonnienne des interférants, nous montrons que la log-fonction caractéristique de Z_u est donnée par :

$$\varphi_{Z_u}(\omega) = \log\left(\phi_{Z_u}\right) = \lim_{R\to+\infty} \lambda\pi R^2 \left(E\left[e^{j\omega\gamma\psi}\right] - 1\right) \tag{3.58}$$

Pour un rayon de cercle R donné et en conditionnant par rapport à γ on a :

$$E\left[e^{j\omega\gamma\psi}\right] = \int_{R^{-\frac{a}{2}}}^{+\infty} E\left[e^{j\omega\gamma\psi}\,|\gamma = x\right] f_\gamma(x)dx = \int_{R^{-\frac{a}{2}}}^{+\infty} \phi_\psi(\omega x)\, \frac{4x^{-\frac{4}{a}-1}}{aR^2} dx \tag{3.59}$$

3. Pour mieux illustrer la possibilité d'avoir des variances infinies, considérons une communication entre deux noeuds lointaines, si à coté, un troisième noeud ad-hoc vient d'être activé il va percevoir des signaux de très haute amplitude (car à cause de la distance, la liaison est au départ réglée pour une puissance élevée d'émission et de réception).

Nous montrons par la suite que :

$$\varphi_{Z_u}(\omega) = \lim_{R\to+\infty} \lambda\pi R^2 \left(\phi_\psi\left(\omega R^{-\frac{a}{2}}\right) - 1\right) + \lambda\pi\omega^{\frac{4}{a}} \int_0^{+\infty} \frac{d\phi_\psi}{du}(u)\, u^{-\frac{4}{a}} du \tag{3.60}$$

Pour conclure, nous utilisons l'architecture du réseau qui permet de supposer l'isotropie de la distribution de ψ, ce qui implique que $\phi_\psi(\omega)$ peut s'écrire comme $\phi_\psi(|\omega|)$. En conséquence, nous pouvons montrer que $\lim_{R\to+\infty} \lambda\pi R^2 \left(\phi_\psi\left(\omega R^{-\frac{a}{2}}\right) - 1\right) = 0$ et que :

$$\varphi_{Z_u}(\omega) = \lambda\pi|\omega|^{\frac{4}{a}} \int_0^{+\infty} \frac{d\phi_{\psi_0}}{du}(u)\, u^{-\frac{4}{a}} du \tag{3.61}$$

Dans (3.61), la log-fonction caractéristique peut s'écrire :

$$\varphi_{Z_u}(\omega) = -\sigma|\omega|^{\frac{4}{a}} \qquad \text{avec} \qquad \sigma = -\lambda\pi \int_0^{+\infty} \frac{d\phi_{\psi_0}}{du}(u)\, u^{-\frac{4}{a}} du \tag{3.62}$$

3.3.2 Récepteur optimal - évaluation des performances

Le problème de construction de récepteurs optimaux revient au problème statistique de test d'hypothèse suivant :

$$\begin{cases} H_0 : x(k) = s_0(k) + n_g(k) + n_\alpha(k), \quad k = 1, 2, \dots, N \\ \\ H_1 : x(k) = s_1(k) + n_g(k) + n_\alpha(k), \quad k = 1, 2, \dots, N \end{cases} \tag{3.63}$$

où $s_i(.)$, $i = 0, 1$, est l'un des deux signaux obtenus après corrélation du signal (3.52); ils correspondent aux valeurs binaires choisies. Le terme $n_\alpha(.)$ est un bruit blanc impulsif symétrique α-stable (SαS) de paramètre d'échelle γ et $n_g(.)$ un bruit blanc gaussien de variance σ^2. Les deux bruits, gaussiens et impulsifs, sont indépendants l'un de l'autre et du signal transmis. En se basant sur les résultats de la section précédente, la fonction caractéristique du bruit additif total résultant est :

$$\phi_X(\omega) = \exp(-\gamma|\omega|^\alpha - \sigma^2\omega^2) \tag{3.64}$$

La densité de probabilité f_X de X peut être approchée numériquement via la transformée de Fourier inverse de la fonction caractéristique Φ_X.
Pour décider entre les deux hypothèses H_0 et H_1, un test de rapport de vraisemblance mène au calcul de la statistique de test suivante :

$$\Lambda = \sum_{k=1}^{N} \log\left\{\frac{f_X(x(k) - s_1(k))}{f_X(x(k) - s_0(k))}\right\}, \tag{3.65}$$

et la compare à un seuil prédéfini η. Lorsque $\Lambda \geq \eta$, le récepteur décide que $s_1(.)$ a été envoyé, autrement c'est $s_0(.)$ qui a été envoyé. En communication, on s'intéresse notamment à la moyenne de la probabilité d'erreur binaire BER (*Bits Error Rate*); elle est donnée par :

$$P_e = \int_{-\infty}^{+\infty} \mathbb{P}\left(Z < -x \,\middle|\, a_0^{(0)} = 0, Z_n\right) f_X(x) dx \tag{3.66}$$

Pour N assez grand, et par un argument du théorème central limite évoqué par exemple dans [144, 145] la statistique Λ peut être approchée par une distribution gaussienne; dans ce cas, la

probabilité d'erreur moyenne est approchée par :

$$P_e = \text{erfc}\left(\frac{\mu_0}{\sqrt{2\sigma_0^2}}\right), \tag{3.67}$$

où erfc(.) est la fonction d'erreur complémentaire. Les deux paramètres μ_0 et σ_0^2 sont la moyenne et la variance de Λ sous l'hypothèse que que le signal s_0 a été émis. Nous montrons qu'on peut les écrire sous la forme :

$$\mu_0 = \sum_{k=1}^{N} \int_{-\infty}^{+\infty} f_X(\xi - s_0(k)) \log\left\{\frac{f_X(\xi - s_1(k))}{f_X(\xi - s_0(k))}\right\} d\xi, \tag{3.68}$$

$$\sigma_0^2 = \sum_{k=1}^{N} \int_{-\infty}^{+\infty} f_X(\xi - s_0(k)) \log^2\left\{\frac{f_X(\xi - s_1(k))}{f_X(\xi - s_0(k))}\right\} d\xi - \frac{\mu_0^2}{N}, \tag{3.69}$$

Comme expliqué dans [144] l'expression de la probabilité d'erreur n'est valable qu'asymptotiquement. En plus, dans le cas général de distribution α-stable, la densité de probabilité f_X n'a pas de forme explicite ; c'est pour cette raison que nous ne nous sommes intéressés qu'au récepteur de Cauchy dans [49].

3.3.3 Illustration sur un exemple.

Nous considérons un système IR-UWB et une configuration *ad hoc*. La sortie du récepteur Z_d est donnée dans l'équation (3.53) ; nous ne considérons que les canaux à trajet unique correspondant au LOS (*Line Of Sight*) [4]. Dans ce cas, l'atténuation du canal est donnée par $\gamma_i \propto d^{-a/2}$ où a est le paramètre d'atténuation. Un résumé des paramètres utilisés en simulation est présenté dans le tableau 3.1. Dans cette configuration, nous montrons que la fonction log-caractéristique de Z

Table 3.1 – Paramètres utilisés pour la simulation.

Paramètre	Valeur
durée de la trame T_s	10ns
Durée de l'impulsion	0.3ns
Le décalage PPM ϵ	0.3ns

peut s'écrire :

$$\begin{aligned}\varphi_Z(\omega) &= \frac{|Nq}{R^2}|\omega|^{\frac{4}{a}} \int_0^{+\infty} \frac{d\phi_\psi}{du}(u)\, u^{-\frac{4}{a}} du \\ &= \frac{|Nq}{R^2}|\omega|^{\frac{4}{a}} F,\end{aligned} \tag{3.70}$$

où $\phi_\psi(\omega)$ est la fonction caractéristique de $\psi(\omega)$ (définie dans (3.55)). La valeur F est indépendante de ω de sorte que Z est une variable aléatoire symétrique α -stable avec les paramètres $\alpha = \frac{4}{a}$ et $\sigma = -\left(|Nq/R^2\right)F$ (les deux paramètres restants sont nuls). Puisque ψ possède des moments

4. Ceci n'est bien sûr pas très réaliste mais sa simplicité permet de vérifier la validité des paramètres théoriques que nous avons démontrée dans [49].

finis, l'intégrale pour calculer F existe lorsque a est plus grand que 2, ce qui est le cas dans la plupart des situations pratiques.

Pour évaluer σ, nous devons calculer F dans (3.70). Elle peut être obtenue analytiquement lorsque $\omega_p(t)$ est une impulsion rectangulaire ; dans ce cas, la fonction caractéristique de ψ est égale à $\phi_\psi(\omega) = \frac{\sin(\omega)}{\omega}$. Un simple calcul montre que F est donnée par :

$$F = \frac{2^{-1-\frac{4}{a}}\sqrt{\pi}\Gamma\left(\frac{-2}{a}\right)}{\Gamma\left(\frac{1}{2}+\frac{2}{a}\right)} - \frac{2^{-2-\frac{4}{a}}\sqrt{\pi}\Gamma\left(\frac{-2}{a}\right)}{\Gamma\left(\frac{3}{2}+\frac{2}{a}\right)} \tag{3.71}$$

Enfin, P_e est calculé en utilisant l'intégration numérique de :

$$P_e = \int_{-\infty}^{+\infty} \mathbb{P}\left(Z < -x \,\middle|\, a_0^{(0)} = 0, Z_n\right) f_X(x)dx \tag{3.72}$$

La densité de probabilité de Z n'est pas explicitement connue ; nous calculons donc P_e de manière semi-analytique via des simulations Monte-Carlo. Dans la figure 3.8 nous représentons des exemples de courbes du BER, en fonction du rapport signal à bruit, pour plusieurs configurations (différentes valeurs du coefficient d'atténuation a, le nombre moyen d'utilisateurs par unité de surface λ et la taille de la zone considérée R). Nous montrons un bon ajustement entre le BER

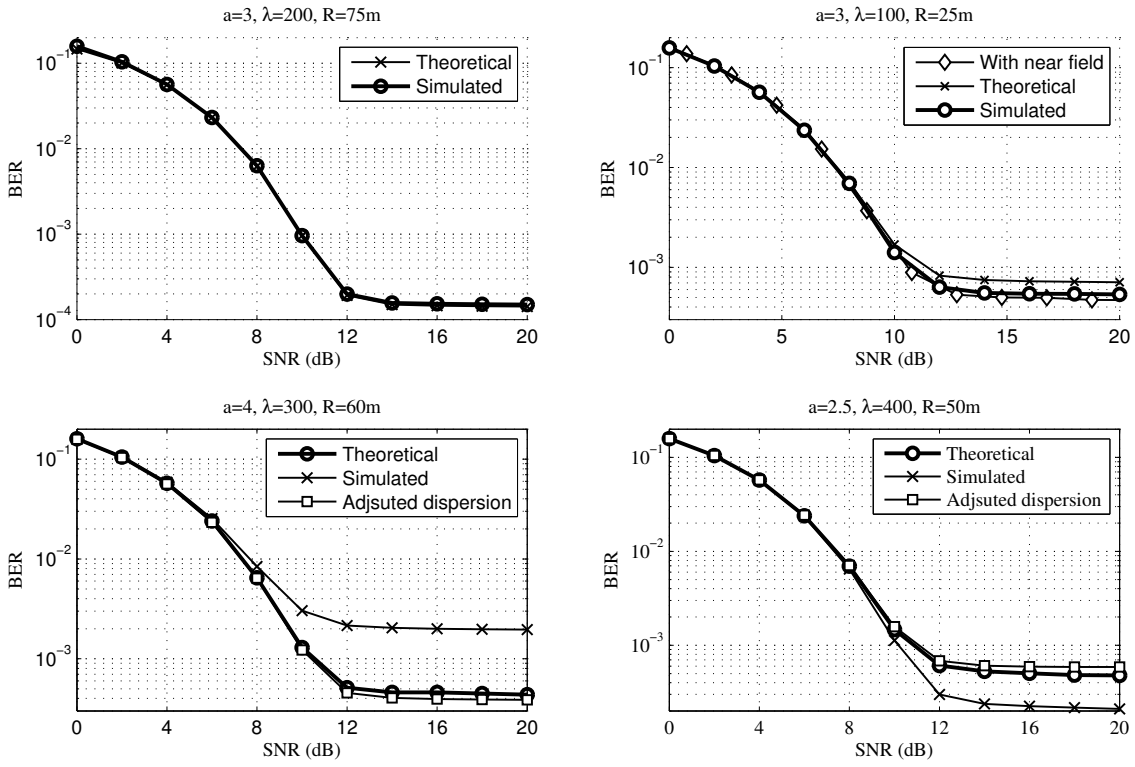


FIGURE 3.8 – Comparaison du BER calculé à partir de nos résultats théoriques et une approximation semi-analytique dans le cas d'une impulsion rectangulaire et plusieurs configurations.

semi-analytique et le système simulé dans le cas d'une impulsion rectangulaire. Cependant, cet

ajustement n'est pas toujours parfait et certaines erreurs peuvent parfois être remarquées. Dans de tels cas, nous ajustons la dispersion de la distribution et obtenons un ajustement précis entre les courbes. Un comportement similaire est obtenu avec d'autres formes d'impulsions mais la dispersion doit être estimée car nous n'avons pas d'expression analytique pour F. D'autres exemples et discussions sont détaillés dans [49, 68]

3.4 Capacité optimale du canal α-stable

Nous venons de voir dans la section précédente que le bruit impulsif représente une caractéristique importante des systèmes de communications modernes [45, 53]. Les premiers modèles ont été introduits, dans le cadre de la physique statistique, pour l'interférence dans les réseaux sans fil [172, 126, 87, 31, 24]. Une autre approche conduit au canal à bruit additif α-stable ($AS\alpha SN$: *Additive symmetric α-stable Noise*) donné par

$$Y = X + Z, \tag{3.73}$$

Comme nous l'avons montré dans la section précédente, les canaux $AS\alpha SN$ apparaissent naturellement dans les réseaux sans fil avec interférence [53, 21]. Cette dernière survient lorsque l'ensemble des dispositifs de transmissions actifs dans un champ spatial de Poisson change rapidement ; ce qui signifie que l'interférence peut être caractérisée par une série de Lepage [147] et implique à son tour que les statistiques d'interférence sont α-stables.

La difficulté de caractériser la capacité des canaux $AS\alpha SN$, à la manière de Shannon [183, 185] est due en partie au fait qu'une contrainte de puissance $\mathbb{E}[X^2] \leq P$ est généralement imposée. Contrairement au cas Gaussien, le moment d'ordre 2 des distributions α-stables est infini pour $\alpha < 2$. Pour surmonter cette difficulté, nous remplaçons la contrainte d'énergie par une contrainte sur le moment d'ordre 1. $\mathbb{E}[|X|] \leq c, \quad c > 0$. L'avantage d'imposer cette contrainte est que l'on peut établir les bornes supérieures et inférieures de la capacité du canal $AS\alpha SN$ avec α dans $(1, 2]$.

Dans les paragraphes qui suivent, nous donnons la formulation du problème d'optimisation de la capacité. Nous montrons, par la suite, l'existence et l'unicité de la distribution optimale solution de ce problème. Nous donnons explicitement des bornes supérieures et inférieures de la capacité. Dans la section 3.4.2, nous donnons une approximation numérique de cette capacité en utilisant l'algorithme Blahut-Arimoto.

3.4.1 Problème d'optimisation de la capacité et résultats préliminaires.

Soit $\mathcal{B}(\mathbb{R})$ la tribu borélienne sur $\mathbb{R}$ et $\mathcal{P}$ l'ensemble des mesures de probabilités définies sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ muni de la topologie faible. Nous notons $\Lambda(c)$ l'ensemble des mesures de probabilité μ de variables aléatoires qu'on notera X et qui ont un moment d'ordre 1 fini. c'est à dire :

$$\Lambda(c) = \left\{ \mu \in \mathcal{P}, \quad \text{telle que} \quad \mathbb{E}_\mu[|X|] \leq c \right\}$$

Quand elles existent, nous désignerons par p_X la densité de X et par $p_Y(\cdot; \mu)$ la densité de probabilité de Y, qui est paramétrée par la mesure de probabilité d'entrée μ. Nous définissons la capacité du canal $AS\alpha SN$ comme solution du problème de maximisation suivant :

$$C = \max \left\{ I(X, Y), \text{ tel que } \mu \in \Lambda(c) \right\}, \tag{3.74}$$

où $I(X; Y)$ est l'information mutuelle du canal dans (3.73), et μ est la mesure de probabilité de X.

Existence et unicité : L'idée est de montrer d'abord que ce problème de maximisation (3.74) admet une solution unique μ^* telle que :

$$\mu^* = \underset{\mu \in \Lambda(c)}{\operatorname{argmax}} \{ I(X,Y) \} \tag{3.75}$$

Théorème 3.4.1 : Pour le canal ASαSN de (3.73), nous montrons les résultats préliminaires suivants :

(1) L'information mutuelle $I(X,Y)$ est continue sur $\Lambda(c)$.

(2) L'ensemble des mesures de probabilité $\Lambda(c)$ est compact pour la topologie faible. De plus, la mesure de probabilité de la capacité optimale μ^* existe.

(3) La mesure de probabilité de la capacité optimale μ^* est unique dans $\Lambda(c)$.

Démonstration. La démonstration détaillée de ces résultats se trouvent dans notre papier [18]. Nous en donnons ici les grandes lignes. Sans perdre de généralité, nous supposons que les densités p_X et p_Y existent

(1) Soit $Y_k = X_k + Z$ où $Z \sim S_\alpha(\gamma, 0, 0)$ de densité de probabilité notée P_Z et de fonction caractéristique notée $\phi_Z(.)$. Il suffit de montrer que pour toute suite $(\mu_k)_k$ de $\Lambda(c)$ qui converge faiblement vers μ, l'information mutuelle correspondante $I(X_k, Y_k)$, converge vers $I(X,Y)$. Nous utilisons d'abord la relation entre l'information mutuelle et l'entropie que l'on peut écrire dans notre cas comme suit :

$$I(X_k, X_k + Z) = h(X_k + Z) - h(Z). \tag{3.76}$$

Notons d'abord que l'entropie $h(Z)$ est bornée [5]. En appliquant [36, Lemma 3] et le fait que $\mathbb{E}_{\mu_k}[|X_k|] \leq c$ pour chaque $\mu_k \in \Lambda(c)$, nous montrons d'abord que $I(X_k, X_k + Z)$ est aussi bornée. Nous utilisons le fait que $p_{Y_k}(x) = \int_{-\infty}^{\infty} p_Z(x-y)\mu_k(dy)$ et donc pour montrer que $\lim_{k\to\infty} h(X_k + Z) = h(X + Z)$, il suffit de montrer que :

$$\lim_{k\to\infty} \int_{-\infty}^{\infty} p_Z(x-y)\mu_k(dy) = \int_{-\infty}^{\infty} p_Z(x-y)\mu(dy). \tag{3.77}$$

Ceci découle de la définition de la convergence faible.

(2) Pour montrer la compacité de $\Lambda(c)$ dans $\mathcal{P}$, il suffit de montrer que $\Lambda(c)$ est fermé borné (théorème de Prokhorov [119]). En utilisant l'inégalité de Markov et le fait que $\mathbb{E}_\mu[|X|] \leq c$, remarquons d'abord, que pour chaque $\epsilon > 0$, il existe $a_\epsilon > 0$ tel que pour tout $\mu \in \Lambda(c)$,

$$\Pr(|X| \geq a_\epsilon) \leq \frac{\mathbb{E}_\mu[|X|]}{a_\epsilon} \leq \frac{c}{a_\epsilon} < \epsilon \tag{3.78}$$

Notons $\mathcal{K}_\epsilon = [-a_\epsilon, a_\epsilon]$, alors $\mathcal{K}_\epsilon$ est compact sur $\mathbb{R}$ et $\mu(\mathcal{K}_\epsilon) \geq 1 - \epsilon$ pour tout $\mu \in \Lambda(c)$. Par conséquent, $\Lambda(c)$ est borné.

Soit maintenant $(\mu_n)_n$ une suite de $\Lambda(c)$ qui converge faiblement vers μ_0. En prenant par exemple $f(x) = |x|$ et en y appliquant le théorème de Portmanteau pour la convergence

5. Ceci est obtenu en appliquant la Propriété 3.7.

faible [119] nous avons,

$$\mathbb{E}_{\mu_0}[|X|] = \int_{-\infty}^{\infty} f(x)\mu_0(dx) \leq \liminf_{n\to\infty} \int_{-\infty}^{\infty} f(x)\mu_n(dx) \leq c. \tag{3.79}$$

Ceci signifie que $\mu_0 \in \Lambda(c)$ ce qui implique que $\Lambda(c)$ est fermé. Par conséquent il est compact.

En utilisant la continuité de l'information mutuelle $I(X,Y)$ sur $\Lambda(c)$ et le théorème des valeurs extrêmes, la mesure de probabilité de la capacité optimale μ^* existe.

(3) Supposons qu'il existe deux solutions μ_0, μ_1 du problème (3.75). D'après la définition de la capacité donnée dans (3.74) cela implique qu'implicitement on a : $p_Y(.,\mu_0) = p_Y(.,\mu_1)$. En utilisant l'indépendance des entrées X et du bruit Z dans (3.73), les deux fonctions caractéristiques vérifient, $\phi_Z(t)\phi_{\mu_0}(t) = \phi_Z(t)\phi_{\mu_1}(t)$ pour tout $t \in \mathbb{R}$. Puisque la fonction caractéristique $\phi_t(t)$ du bruit α-stable est strictement non-nulle, ceci implique que $\mu_0 = \mu_1$. □

Bornes inférieures et supérieures de la capacité : Si on note $W(\cdot \mid x)$ la loi conditionnelle du canal sachant $X = x$ c'est-à-dire correspondante à la variable aléatoire $Z = x + N$, $x \in \mathbb{R}$. Soit $R(\cdot)$ une mesure de probabilité absolument continue sur $\mathbb{R}$, avec p_R sa densité de probabilité. Puisque l'alphabet de l'entrée et de la sortie est l'ensemble $\mathbb{R}$, qui est séparable, nous avons donc par application de [105, Théorème 5.1] :

$$C \leq \int_{-\infty}^{\infty} D(W(\cdot|x) \,||\, R(\cdot))d\mu^*, \tag{3.80}$$

où $D(\cdot\,||\,\cdot)$ est la divergence Kullback-Leibler. Puisque $W(\cdot|x)$ et R admettent des densités par rapport à la mesure de Lebesgue, on peut écrire,

$$C \leq \mathbb{E}_{\mu^*}\left[\int_{-\infty}^{\infty} p_Z(y)\log_2\left(\frac{p_Z(y)}{p_R(y)}\right)dy\right]. \tag{3.81}$$

Les distributions P_R que nous utiliserons pour établir les bornes supérieures de la capacité sont données par

(i) La *distribution de Laplace*, de densité de probabilité :

$$p_{R_S}(y) = \frac{\lambda}{2}\exp(-\lambda|x|), \tag{3.82}$$

où $\lambda > 0$ est un paramètre à choisir.

(ii) La *distribution polynomiale*, avec la fonction de densité

$$p_{R_P}(x) = \begin{cases} \frac{c_{x_0}}{1+|x|}, & |x| \leq x_0 \\ \frac{c_{x_0}}{|x|}, & |x| > x_0, \end{cases} \tag{3.83}$$

où $x_0 > 0$ et c_{x_0} est choisi de sorte qu'il normalise la fonction de densité.

Le choix de ces deux distributions est inspiré par les propriétés des moments fractionnaires ainsi que le comportement asymptotique des distributions symétriques α-stables.

Théorème 3.4.2 : Considérons le canal $AS\alpha SN$ donné dans (3.74) où $Z \sim S_\alpha(\gamma_Z, 0, 0)$ et $1 < \alpha < 2$. Les bornes supérieures et inférieures de la capacité peuvent être données de manière explicite :

(1) En prenant X ayant une distribution α-stable, nous montrons que la capacité optimale, solution de (3.74), est minorée par :

$$C \geq \frac{1}{\alpha}\log_2\left(1 + M_\alpha\left(\frac{c}{\gamma_Z}\right)^\alpha\right), \text{ où } M_\alpha = \left(\frac{\pi}{2\Gamma\left(1-\frac{1}{\alpha}\right)}\right)^\alpha \tag{3.84}$$

(2) En prenant X ayant une distribution de Laplace (3.82), nous montrons qu'elle est majorée par,

$$C \leq \log_2\left(\frac{2\Gamma\left(\frac{1}{\alpha}\right)}{\lambda\gamma_N\alpha\pi}\right) + (\log 2)^{-1}\lambda\left(\frac{2\gamma_N\Gamma\left(1-\frac{1}{\alpha}\right)}{\pi} + c\right). \tag{3.85}$$

(3) Si nous utilisons la distribution p_{R_p} donnée dans (3.83) avec $x_0 > 1$, dans ce cas la capacité admet la borne supérieure approximative suivante :

$$C \lessapprox \log\left(\frac{C_\alpha x_0^{-\alpha-1}}{c_{x_0}}(1 + \mathbb{E}_Z[|Z|] + c) + \frac{\Gamma(1/\alpha)}{\alpha\pi\gamma c_{x_0}}[\mathbb{E}_N[|N|] + c]\right). \tag{3.86}$$

Démonstration. *(1)* Considérons le cas particulier où, $X \sim S_\alpha(\gamma_X, 0, 0)$ avec $\mathbb{E}|X| = c$. D'après la propriété de stabilité $Y = X + Z \sim S_\alpha(\gamma_Y, 0, 0)$ avec $\gamma_Y = (\gamma_X^\alpha + \gamma^\alpha)^{\frac{1}{\alpha}}$. Dans ce cas on peut écrire, $Y \overset{d}{=} \gamma_Y U$ et $Z \overset{d}{=} \gamma_Z U$ où $U \sim S_\alpha(1, 0, 0)$. En utilisant les propriétés de l'entropie nous montrons que :

$$\begin{aligned} I(X,Y) &= h(Y) - h(Y|X) = h(\gamma_Y U) - h(\gamma U) \\ &= h(U) + \log_2(\gamma_Y) - h(U) - \log_2(\gamma_Z) \\ &= \log_2\left(\frac{(\gamma_X^\alpha + \gamma_Z^\alpha)^{\frac{1}{\alpha}}}{\gamma_Z}\right) = \frac{1}{\alpha}\log_2\left(1 + \frac{\gamma_X^\alpha}{\gamma_Z^\alpha}\right) \end{aligned}$$

En utilisant la propriété 3.6 et le fait que $\mathbb{E}[|X|] = c$.

$$I(X,Y) = \frac{1}{\alpha}\log_2\left(1 + \left(\frac{c\pi}{2\gamma_Z\Gamma\left(1-\frac{1}{\alpha}\right)}\right)^\alpha\right), \tag{3.87}$$

ce qui implique la minoration (3.84).

(2) D'après l'inégalité (3.80), la capacité peut être majorée par,

$$C \leq (\log 2)^{-1}\mathbb{E}_{\mu^*}\left[\int_{-\infty}^{\infty} p_Z(y)\log\left(\frac{p_Z(y)}{p_{R_S}(y)}\right)dy\right]. \tag{3.88}$$

Par application de la Propriété 3.7, nous obtenons facilement :

$$\begin{aligned} C &\leq (\log 2)^{-1}\log\left(\frac{2\Gamma\left(\frac{1}{\alpha}\right)}{\lambda\gamma_N\alpha\pi}\right) + (\log 2)^{-1}\lambda\mathbb{E}_{\mu^*}\left[\int_{-\infty}^{\infty} p_Z(y)|y|dy\right] \\ &= \log_2\left(\frac{2\Gamma\left(\frac{1}{\alpha}\right)}{\lambda\gamma_Z\alpha\pi}\right) + (\log 2)^{-1}\lambda\mathbb{E}_{\mu^*}\left[\int_{-\infty}^{\infty} p_N(y)|y + X|dy\right]. \end{aligned} \tag{3.89}$$

L'application de l'inégalité triangulaire donne,

$$C \leq \log_2\left(\frac{2\Gamma\left(\frac{1}{\alpha}\right)}{\lambda\gamma_Z\alpha\pi}\right) + (\log 2)^{-1}\lambda\left(\mathbb{E}[|Z|] + \mathbb{E}_{\mu^*}[|X|]\right)$$
$$\leq \log_2\left(\frac{2\Gamma\left(\frac{1}{\alpha}\right)}{\lambda\gamma_Z\alpha\pi}\right) + (\log 2)^{-1}\lambda\left(\frac{2\gamma_Z\Gamma\left(1-\frac{1}{\alpha}\right)}{\pi} + c\right), \tag{3.90}$$

qui découle de la Propriété 3.6.

(3) Nous appliquons l'inégalité de Jensen deux fois sur (3.80), nous avons :

$$C \leq \mathbb{E}_{\mu^*}\left[\int_{-\infty}^{\infty} p_Z(y)\log\left(\frac{p_Z(y)}{p_{R_p}(y)}\right)dy\right] \leq \log\left(\mathbb{E}_{\mu^*}\left[\int_{-\infty}^{\infty} p_Z(y)\frac{p_Z(y)}{p_{R_p}(y)}dy\right]\right), \tag{3.91}$$

Comme $x_0 > 1$ et en distinguant les deux cas où $|y| > x_0$ et $|y| <= x_0$, nous utilisons le comportement asymptotique de la queue de distributions donné par la propriété (3.8), on obtient :

$$C \lessapprox \log\left(\mathbb{E}_{\mu^*}\left[\int_{|z|>x_0}\frac{p_Z(z)C_\alpha|z|^{-\alpha-1}}{p_R(z+x)}dz\right] + \frac{\Gamma(1/\alpha)}{\alpha\pi\gamma c_{x_0}}\left[\mathbb{E}_Z[|Z|] + \mathbb{E}_{\mu^*}[|X|]\right]\right). \tag{3.92}$$

Notons que,

$$\mathbb{E}_{\mu^*}\left[\int_{|z|>x_0}\frac{p_Z(z)C_\alpha|z|^{-\alpha-1}}{p_R(z+x)}dz\right] = \mathbb{E}_{\mu^*}\left[\int_{|z|>x_0}\frac{p_Z(z)C_\alpha(1+|z+x|)|z|^{-\alpha-1}}{c_{x_0}}dz\right]$$
$$\leq \mathbb{E}_{\mu^*}\left[\int_{|z|>x_0}\frac{p_Z(z)C_\alpha(1+|z+x|)x_0^{-\alpha-1}}{c_{x_0}}dz\right]$$
$$\leq \frac{C_\alpha x_0^{-\alpha-1}}{c_{x_0}}\left(1+\mathbb{E}_Z[|Z|] + \mathbb{E}_{\mu^*}[|X|]\right). \tag{3.93}$$

Ce qui donne la borne approximative de la capacité optimale. Le détail et le calcul technique peuvent être consulté dans notre papier original [18].

□

Notons que quand $x_0 \to \infty$, la borne approximative dans (3.86) converge vers la borne supérieure. Ce résultat est dû au fait que nous avons établi l'approximation en utilisant le comportement asymptotique de la queue de distribution α-stable (3.8). D'autres méthodes alternatives permettent d'établir les bornes de la capacité en solvant le problème dual de la capacité [105]. En particulier, dans le contexte de canaux à bruit gaussien, la capacité peut être majorée par

$$C \leq \min_{\gamma\geq 0}\max_{x\in\mathbb{R}}[D(W(\cdot|x)\,||\,R(\cdot)) + \gamma(c - \mathbb{E}[|X|])]. \tag{3.94}$$

Dans la section suivante, nous étudions numériquement la capacité du canal $AS\alpha SN$ via l'algorithme Arimoto-Blahut, qui présente l'avantage de converger vers la capacité optimale lorsque l'approximation de la loi du canal converge vers une loi stable.

3.4.2 Simulations numériques et validation

Dans cette section, nous considérons une approximation numérique de la capacité en adaptant l'algorithme de Blahut-Arimoto au problème d'optimisation de la capacité dans (3.74). Contrairement à [43], notre étude numérique considère la contrainte $\mathbb{E}[|X|] \leq c$, plutôt qu'une contrainte de puissance ou sur le moment de second ordre.

Algorithm 3 : Calcul de la capacité approximative en (3.95).

Input : Initialiser $r^{(0)}(x) = \dfrac{1}{|S_X|}$, $C_0 = 0,\ C_{-1} = -2\epsilon$.

1 **do :** Tant que $C_n - C_{n-1} > \epsilon$;

2 Calculer $Q^{(n)}(x|y) = \dfrac{r^{(n-1)}(x)P(y|x)}{\sum_{x=1}^{m} r^{(n-1)}(x)P(y|x)}$ et

$$C_n = \sum_{x=1}^{m}\sum_{y=1}^{n} r^{(n-1)}(x)P(y|x)\log_2\left(\frac{Q^{(n)}(x|y)}{r^{(n-1)}(x)}\right);$$

3 Résoudre ν tel que, $\sum_{x=1}^{m}\left(1-\dfrac{|x|}{c}\right)e^{\nu|x|}\prod_{y=1}^{n}Q^{(n)}(x|y)^{P(y|x)} = 0.$;

4 Mettre à jour, $r^{(n)}(x) = \dfrac{e^{\nu|x|}\prod_{y=1}^{n}Q^{(n)}(x|y)^{P(y|x)}}{\sum_{x'=1}^{m}e^{\nu|x'|}\prod_{y=1}^{n}Q^{(n)}(x'|y)^{P(y|x')}}$;

Output : La capacité optimale approximative C_n.

Algorithme d'approximation numérique de la capacité : L'algorithme Blahut-Arimoto fournit un moyen d'approcher numériquement la capacité d'un canal discret à support borné. Pour approcher (3.74), nous donnons une modification de l'algorithme présenté dans [175, Section IV] pour tenir compte de la spécificité des distributions α-stables et de la contrainte sur le moment d'ordre 1. Dans le cas du canal $AS\alpha SN$, le support de la densité du bruit n'est ni discret ni borné. L'idée est d'approcher le canal (3.74) par :

$$Y_a = X_{X_{\max}} + Z_{Z_{\max}}, \tag{3.95}$$

où $Z_{Z_{\max}}$ est obtenue en discrétisant la densité du bruit α-stable Z. La loi du canal est désignée dans ce cas par $Q(\cdot|x)$. Nous notons S_X et S_Z les supports respectifs de la variable aléatoire $X_{X_{\max}}$ et de $Z_{Z_{\max}}$. Le problème d'optimisation de la capacité du canal (3.95) avec la contrainte $\mathbb{E}[|X_{X_{\max}}|] \leq c$ est alors de chercher :

$$C_{approx} = \max_{\mathbf{q}\in\mathcal{Q}}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n} q(x_i)Q(y_j|x_i)\log_2\left(\frac{Q(y_j|x_i)q(x_j)}{p(y_i)q(x_j)}\right),\ \text{tel que},\ \sum_{i=1}^{m}|x_i|q(x_i) \leq c\right\}, \tag{3.96}$$

où $m = |S_X|$, $n = |S_Z|$, et $\mathcal{Q}$ est la famille des probabilités discrètes de support S_X. Pour obtenir la solution, C_{approx}, de (3.96), l'idée clé due à Blahut et Arimoto [175, 174] est que,

$$C_{approx} = \max_{\mathbf{q}\in\mathcal{Q}} \left\{ \max_{\Phi} J(\mathbf{q},), \text{ tel que }, \sum_{i=1}^{m} |x_i| q(x_i) \leq c \right\} \tag{3.97}$$

où

$$J(\mathbf{q}, \Phi) = \sum_{i=1}^{m} \sum_{j=1}^{n} q(x_i) Q(y_j|x_i) \log_2 \left(\frac{\Phi(x_i|y_j)}{q(x_i)} \right) \tag{3.98}$$

et Φ est une matrice de probabilités de transition arbitraire $m \times n$. La capacité approximative est alors obtenue en alternant les deux problèmes de maximisation, ce qui conduit à l'algorithme 3.

Il a été montré, voir par exemple [16], que l'algorithme de Blahut-Arimoto converge vers la capacité lorsque la distribution discrétisée converge, au sens la variation totale, vers la loi du canal α-stable. Plus précisément, il existe un $M > 0$ tel que,

$$|C^* - C_{approx}| \leq M \|p_{Y_a} - p_Y\|_{TV} + o(\|p_{Y_a} - p_Y\|_{TV})|, \tag{3.99}$$

où

$$\|p_{Y_a} - p_Y\|_{TV} = \frac{1}{2} \int_{\mathbb{R}} |p_{Y_a}(x) - p_Y(x)| dx. \tag{3.100}$$

Par conséquent, l'approximation de l'algorithme Blahut-Arimoto converge vers la capacité du canal. Nous comparons nos bornes avec l'approximation numérique obtenue à l'aide de l'algorithme de Blahut-Arimoto. Tout d'abord, nous étudions le comportement de la capacité approximative via l'algorithme de Blahut-Arimoto en fonction du rapport $\frac{c}{\gamma_Z}$. Ce dernier est analogue au rapport signal sur bruit (SNR : *Signal-to-Noise Ratio*) que l'on trouve dans la borne de la capacité d'un canal gaussien avec une contrainte de puissance. Des discussions détaillées sur les résultats des simulations et la validité de nos résultats sont données dans [18].

La prolifération des capteurs et le développement de la technologie d'acquisition de données ont permis de recueillir de grandes masses de données menant à ce qu'on a appelé, il y a quelques années déjà, le *Big Data*. La très grande dimension des données qui en résulte pose un défi majeur pour le traitement et l'analyse de ces données. En effet, pour une modélisation statistique fiable, il est primordial d'avoir une bonne qualité de prédiction combinée à une représentation parcimonieuse. La statistique en grande dimension englobe les modèles de régression, de classification supervisée ou non, la modélisation graphique, les tests multiples où le nombre d'hypothèses considérées est supérieur à la taille de l'échantillon. La régression, en particulier, constitue la base des techniques d'apprentissage automatique ou de ce que l'on appelle actuellement l'intelligence artificielle. L'intérêt de l'utilisation d'une modélisation par la régression est d'une part *descriptif* car elle permet de comprendre et d'expliquer le lien statistique entre des variables d'intérêt et d'autres variables explicatives D'autre part, elle a un rôle *prédictif* car elle permet d'inférer ou de tirer des conclusions sur des situations qui n'ont pas encore été observées. Considérons un échantillon $(y_i, x_i)_{i=1,\dots,n}$, où y_i est l'observation d'une variable d'intérêt y sur un individu i et $x_i = \left(x_{i1}, \dots, x_{ip}\right)^t \in \mathbb{R}^p$ est un vecteur d'observations des variables explicatives correspondantes à y_i. Le modèle d'observation de la régression s'écrit, pour tout $i = 1, \dots, n$:

$$y_i = f(x_i) + \xi_i$$

où $f : \mathbb{R}^k \longrightarrow \mathbb{R}$ est une fonction inconnue et $(\xi_i)_{i=1,\dots,n} \in \mathbb{R}$ représentent les bruits du modèle. Ce modèle peut s'écrire sous la forme vectorielle :

$$Y = \mathbf{f} + \xi, \tag{4.1}$$

où $Y = (y_1, \dots, y_n)^t$, $\mathbf{f} = (f(x_1), \dots, f(x_n))^t$ et $\xi = (\xi_1, \dots, \xi_n)^t$. Les différents domaines où interviennent la régression et ses multiples variantes sont considérables. Les plus populaires sont :

- La régression linéaire qui correspond au cas où $x \in \mathbb{R}^p$ et $f(x_i) = \langle x_i, \beta \rangle$ avec $\beta \in \mathbb{R}^p$ et $x_i = \left(x_i^1, \dots, x_i^p\right)^t$. Quand il existe une structure de groupement des variables comme par exemple en analyse de la variance à plusieurs facteurs, la fonction de régression peut s'écrire $f(x_i) = \sum_{k=1}^{M} \sum_{j \in G_k} \beta_{j,k} x_i^j$ avec $G_1, \dots, G_M$ est une partition de $\{1, \dots, p\}$.
- La régression non paramétrique ou l'estimation par projection ; il s'agit du cas où $f(.)$ est projetée sur un espace de fonctions de base comme par exemple des bases d'ondelettes, une base de Fourier ou des splines.
- Le modèle de régression additif correspond au cas où $f(x) = \sum_{k \in K} f_k(x_k)$ avec f_k des fonctions régulières. Les modèles semi-paramétriques et les modèles GAM (*Generalised Additive Models*) sont des exemples de modèles additifs.

Nous considérons le sous-espace linéaire $\mathbf{f}_\beta = \mathbb{X}\beta$ où $\mathbb{X} \in \mathbb{R}^{n\times p}$ est une matrice à p colonnes et $\beta = (\beta_1, \dots, \beta_p)^t \in \mathbb{R}^p$ un vecteur de p paramètres pour approcher le vecteur moyen inconnu $\mathbf{f}$. Le sous-espace linéaire $\mathbf{f}_\beta = \mathbb{X}\beta$ s'écrit d'un point de vue général :

$$\mathbf{f}_\beta = \mathbb{X}\beta = \left(\sum_{j=1}^{p} \beta_j f_j(x_1), \dots, \sum_{j=1}^{p} \beta_j f_j(x_n) \right)^t \tag{4.2}$$

où $f_j : \mathbb{R}^k \longrightarrow \mathbb{R}$ sont des fonctions fixes de $x_i \in \mathbb{R}^k$. Nous avons abordé ces problèmes de régressions de deux points de vue différents mais qui peuvent parfois être complémentaires. Il s'agit de distinguer le cas classique où nous disposons des échantillons de très grandes tailles et le cas où les phénomènes étudiés ne permettent pas de collecter une grande masse de données comme en volcanologie par exemple.

- Sélection de modèles en petites dimensions : Les critères de sélection de modèle jouent un rôle important dans la théorie de l'information et l'analyse statistique des données, en particulier dans les modèles de régression, les modèles de mélange et l'analyse des séries chronologiques ... Le critère le plus connu est celui d'Akaike (*AIC : Akaike Information Criterion*) qui a été conçu comme un estimateur sans biais de la divergence de Kullback entre le vrai modèle qui a réellement généré les données et une approximation statistique de celui-ci. Nous nous sommes posés la question de la sélection des modèles dans le cas des échantillons de petite taille. En donnant une version corrigée, nous avons adapté les critères de sélection habituels pour les rendre compatibles avec ce cas.
- Régressions *sparses* en grandes dimensions : Nous nous sommes intéressés au cas où $p >> n$ dans le modèle (4.1). Les techniques classiques ne permettent pas d'estimer un modèle ayant plus de paramètres que d'observations disponibles. Néanmoins, en cas de parcimonie (*sparsité*), c'est-à-dire que seulement un petit nombre de coefficients β_j sont non-nuls, diverses méthodes ont été proposées pour estimer efficacement $\mathbf{f}$. Par exemple, les performances des estimateurs des moindres carrés pénalisés notamment le *lasso* et *group-lasso* ont été établies sous forme d'inégalités oracles dans le cadre gaussien. Nous avons donné des résultats équivalents dans le cas de distributions à queues lourdes ou plus généralement des distributions vérifiant des inégalités de Poincaré faibles dite également inégalité de *gap spectral*.
- Sélection des variables et application en volcanologie : Il s'agit d'une application typique de cohabitation entre données de grandes dimensions (données météo, images satellites...) et des échantillons de très petites tailles comme les prélèvements au sol, les données physico-chimiques... Dans le cadre de la surveillance des nuages de cendres, la taille des échantillons des dépôts au sol n'est pas assez large car ils requièrent des prélèvement *Insitu*. En même temps, nous voulons expliquer le débit d'éruption (MER : *Mass Eruption Rate*) en fonction du dépôt au sol et de plusieurs paramètres physico-chimiques liés à une éruption volcanique comme la hauteur, l'acidité, le type de volcan, les conditions météorologiques lors de l'éruption, ainsi que les données satellites. Pour trouver une solution qui permette la gestion d'une crise volcanique en temps réel, il est nécessaire d'adapter les techniques de sélection de variables pour donner un modèle adéquat faisant intervenir des variables explicatives que l'on peut observer en temps réel. Nous avons donné un algorithme de

sélection de variables dans le cadre d'un mélange de régressions gaussiennes. D'autres applications permettant la quantification de l'incertitude sur les aléas volcaniques ont été développées dans [13, 7, 3, 6]

4.1 Critères de sélection de modèles en petits échantillons

Le critère le plus connu AIC a été conçu comme un estimateur sans biais d'une variante de la divergence dirigée de Kullback également connue sous le nom d'information de Kullback-Leibler. Cette mesure de séparation entre deux modèles statistiques est asymétrique et a tendance à choisir un modèle surparamétré lorsque la taille de l'échantillon est petite ou que le nombre de paramètres est relativement grand par rapport à la taille de l'échantillon. Afin de surmonter ce problème, de nombreuses versions corrigées de l'AIC ont été proposées dans différentes situations particulières, voir par exemple,[159] [157], [146, 132]. D'autre part, la divergence symétrique de Kullback a été utilisée pour établir des versions corrigées des critères de sélection de modèles [97] [82]. Quand il s'agit de données longitudinales ou dépendantes, l'AICc ne peut pas être directement appliqué à la sélection de modèles pour les données avec des erreurs corrélées.

4.1.1 Critères de sélection corrigés

On se place dans le cadre du modèle (4.1) pour données longitudinales ; c'est-à-dire quand y_{ij} est une mesure de la variable d'intérêt sur l'individu $i \in \{1, \dots, m\}$ à des instants t_{ij}. L'analyse des données longitudinales a suscité un grand intérêt dans les domaines des essais cliniques, de l'épidémiologie, de l'agriculture et de la médecine au cours de la dernière décennie. Ces données apparaissent lorsque des mesures répétées sont obtenues pour un individu et lorsque plusieurs variables sont observées à des instants différents. Les mesures successives pour chaque individu ont tendance à être corrélées. Il est donc nécessaire de prendre en compte cette liaison afin de produire le modèle adéquat. Une synthèse des aspects théoriques et appliqués, des détails de la structure du modèle et de l'estimation des paramètres de l'analyse longitudinale est présentée dans [108], [92], [98], [140] et [152]. Nous désignons $Y_i = (Y_{i1}..., Y_{in_i})^t$ le vecteur de dimension $(n_i \times 1)$ des observations répétées sur l'individu i. Pour simplifier, nous supposons que le nombre de répétitions est le même pour tous les individus $n_i = n$. Supposons que le vrai modèle qui a généré les données s'écrit sous la forme matricielle (4.1) comme :

$$Y = X_0\beta_0 + \varepsilon_0, \tag{4.3}$$

où $\varepsilon_0 \sim \mathcal{N}(0, \sigma_0^2 V_0)$ avec V_0 une matrice diagonale en blocs de $(n \times n)$ matrices Σ_0. En pratique, nous ne connaissons pas le vrai modèle et nous ajustons donc les données avec un modèle candidat que l'on peut écrire sous la forme matricielle :

$$Y = X\beta + \varepsilon, \tag{4.4}$$

où Y est le vecteur des observations de la variable d'intérêt, X la matrice de design ou de variables explicatives et $\varepsilon \sim \mathcal{N}(0, \sigma^2 V)$ l'erreur du modèle. La matrice V est diagonale en m blocs de matrices Σ. Pour simplifier, nous notons, $\theta_0 = (\beta_0, \sigma_0, \Sigma_0)$ les paramètres inconnus du vrai modèle et $\theta = (\beta, \sigma, \Sigma)$ les paramètres du modèle candidat. On notera également $N = n.m$.

Notion de vraisemblance résiduelle : On note $f(Y|\theta_0)$ et $f(Y|\theta)$ les densités de probabilités respectives du modèle qui a généré les données et du modèle candidat. Dans ce cas, la log-vraisemblance

pour le modèle candidat est donnée par :

$$\log f(Y|\theta) = -\frac{1}{2}\left\{N\log\sigma^2 + m\log|\Sigma| + (Y-X\beta)^t V^{-1}(Y-X\beta)/\sigma^2\right\} \tag{4.5}$$

De même, la fonction de log-vraisemblance du vrai modèle s'obtient en remplaçant θ, σ, β, Σ et V dans l'équation (4.5) par θ_0, σ_0, β_0, Σ_0 et V_0 respectivement. La méthode du maximum de vraisemblance restreinte (ou résiduelle) a été introduite par [177] ; elle permet d'estimer les composantes de la matrice de variance dans un modèle linéaire généralisé. En adoptant les résultats de [108] ou [158], et en omettant les termes constants, la vraisemblance restreinte pour le modèle candidat (4.4) est définie par :

$$\log f_r(Y|\theta) = -\frac{1}{2}\left\{(N-p)\log\sigma^2 + m\log|\Sigma| + \log|X^tV^{-1}X| + Y^tAY/\sigma^2\right\} \tag{4.6}$$

où la matrice $A = V^{-1} - V^{-1}X(X^tV^{-1}X)X^tV^{-1}$. Dans ce cas, l'estimateur du maximum de vraisemblance résiduelle de (β, σ^2, V) est donné par :

$$\hat\beta = (X^t\hat V^{-1}X)^{-1}X^t\hat V^{-1}Y \quad \text{et} \quad \hat\sigma^2 = \frac{(Y-X\beta)^t\hat V^{-1}(Y-X\beta)}{(N-p)}.$$

Le terme $\hat V^{-1}$ est obtenu en maximisant par rapport V, la fonction $\ell_r(V) = \log f_r(Y|\theta) - \frac{1}{2}\log|X^tV^{-1}X|$. De même, la fonction de vraisemblance restreinte pour le vrai modèle (4.3), peut être obtenue en remplaçant σ, β, Σ, X, p et V dans l'équation (4.6) par σ_0, β_0, Σ_0, X_0, p_0 et V_0 respectivement.

Mesures de séparation entre modèles : Nous nous intéressons ici aux distances qui permettent d'évaluer la proximité entre les modèles. La plus connue est la distance dirigée de Kulback-Leibler que l'on peut écrire dans notre contexte sous la forme :

$$d(\theta_0, \theta) = \boldsymbol{E}_{\theta_0}\{-2\log f(Y|\theta)\}$$

où $\boldsymbol{E}_{\theta_0}$ désigne l'espérance par rapport à $f(Y|\theta_0)$ du vrai modèle. L'AIC et ses versions corrigées sont des estimateurs de $\boldsymbol{E}_{\theta_0}\{-2\log f(Y|\hat\theta)\}$. Comme précisé plus haut, cette mesure est dissymétrique et présente l'inconvénient d'un choix privilégiant un biais de sur-paramétrisation.

- Une mesure de séparation entre le modèle générateur et un modèle candidat est donnée par la divergence symétrique de Kullback [179], définie par :

$$J(\theta_0, \theta) = \{d(\theta_0, \theta) - d(\theta_0, \theta_0)\} + \{d(\theta, \theta_0) - d(\theta, \theta)\} \tag{4.7}$$

  Dans notre contexte, $d(\theta_0, \theta_0)$ ne dépend pas de θ ; comme le but est de discriminer entre différents modèles candidats, nous proposons la version,

$$K(\theta_0, \theta) = d(\theta_0, \theta) + \{d(\theta, \theta_0) - d(\theta, \theta)\} \tag{4.8}$$

  comme substitut de $J(\theta_0, \theta)$. Dans le cadre des échantillons de grandes tailles, un estimateur asymptotiquement sans biais de,

$$\Omega(\theta_0) = \boldsymbol{E}_{\theta_0}\{K(\theta_0, \hat\theta)\} \tag{4.9}$$

  a été introduit par [120]. Cet estimateur est donné par le critère de sélection de modèles suivant :

$$KIC = N\log\hat\sigma^2 + m\log|\hat\Sigma| + 3(p+1) \tag{4.10}$$

- Pour s'adapter au cas de la vraisemblance (4.6), une mesure utile de l'écart entre la fonction

de log-vraisemblance résiduelle des modèles candidats et des vrais modèles est définie par,

$$\Omega_r(\theta_0) = \boldsymbol{E}_{\theta_0}\{d_r(\theta_0, \hat{\theta}) + d_r(\hat{\theta}, \theta_0) - d_r(\hat{\theta}, \hat{\theta})\}, \tag{4.11}$$

où $d_r(\theta_0, \theta) = \boldsymbol{E}_{\theta_0}\{-2 \log f_r(Y|\theta)\}$.

Pour construire des versions corrigées de ces critères, nous supposons d'abord que la famille des modèles candidats contient le vrai modèle. Cette hypothèse forte est utilisée dans la constriction du KIC par [120] et de ses versions modifiées [77, 97, 81, 82]. Dans le cas particulier de la régression, cela revient à prendre $p > p_0$ et à réarranger les colonnes de X de sorte que $X_0\beta_0 = X\beta^*$, où $\beta^* = (\beta_0^t, \beta_1^t)^t$ et β_1 est un vecteur de zéros $(p - p_0) \times 1$. Sous cette hypothèse, nous avons établi les critères suivants :

Proposition 4.1.1 : En supposant que le vrai modèle est inclus dans la famille des modèles candidats, nous avons les critères de sélection de modèles corrigés suivant :

(1) Une version corrigée de l'AIC est donnée dans notre cas par,

$$AICc = N \log \hat{\sigma}^2 + m \log |\hat{\Sigma}| + 2\frac{N(p+1)}{N-p-2} \tag{4.12}$$

est un estimateur approché de $d(\theta_0, \theta)$.

(2) Le critère défini par,

$$KICc = N \log \hat{\sigma}^2 + m \log |\hat{\Sigma}| + \frac{(p+1)(3N-p-2)}{N-p-2} \tag{4.13}$$

est un estimateur approché de $\Omega(\theta_0)$

(3) De même le critère

$$MRC_{sd} = (N-p) \log \hat{\sigma}^2 + m \log |\hat{\Sigma}| + p \log(N) + \frac{(N-p)^2}{N-p-2} + (N-p)\{\log(\frac{2}{N-p}) + \Psi(\frac{N-p}{2} \tag{4.14}$$

est un estimateur approché de $\Omega_r(\theta_0)$. La fonction $\Psi(.) = \frac{\Gamma'(.)}{\Gamma(.)}$ est la fonction digamma.

Démonstration. Les détails de la preuve de ces résultats se trouvent dans notre papier original [34]. Nous donnons ici les idées principales des démonstrations :

(1) - (2) En utilisant les équations (4.9) et (4.8), on peut écrire,

$$\Omega(\theta_0) = \boldsymbol{E}_{\theta_0}\{d(\theta_0, \hat{\theta}) + d(\hat{\theta}, \theta_0) - d(\hat{\theta}, \hat{\theta})\}. \tag{4.15}$$

D'abord nous montrons que,

$$\begin{aligned} d(\theta_0, \hat{\theta}) &= \boldsymbol{E}_{\theta_0}\left\{N \log \sigma^2 + m \log |\Sigma| + \frac{(Y - X\beta)^t V^{-1} (Y - X\beta)}{\sigma^2}\right\}_{\theta=\hat{\theta}} \\ &= N \log \hat{\sigma}^2 + m \log |\hat{\Sigma}| + \frac{\sigma_0^2}{\sigma^2} tr(\hat{V}^{-1} V_0) + \frac{(\hat{\beta} - \beta^*)^t X^t \hat{V}^{-1} X (\hat{\beta} - \beta^*)}{\hat{\sigma}^2} \end{aligned} \tag{4.16}$$

où tr désigne la trace. Avec le même raisonnement nous montrons que :

$$d(\hat{\theta}, \theta_0) = N \log \sigma_0^2 + m \log |\Sigma_0| + \frac{\hat{\sigma}^2}{\sigma_0^2} tr(\hat{V} V_0^{-1}) + \frac{(\hat{\beta} - \beta^*)^t X^t V_0^{-1} X (\hat{\beta} - \beta^*)}{\sigma_0^2} \tag{4.17}$$

et que :

$$d(\hat{\theta},\hat{\theta}) = N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + N \tag{4.18}$$

En remplaçant (4.16), (4.17) et (4.18) dans (4.15), nous obtenons :

$$\Omega(\theta_0) = E_{\theta_0}\left\{N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + \frac{\sigma_0^2}{\sigma^2}tr(\hat{V}^{-1}V_0) + \frac{(\hat{\beta}-\beta^*)^tX^t\hat{V}^{-1}X(\hat{\beta}-\beta^*)}{\hat{\sigma}^2}\right\} \tag{4.19}$$

$$+E_{\theta_0}\left\{N\log\frac{\sigma_0^2}{\hat{\sigma}^2} + m\log\frac{|\Sigma_0|}{|\hat{\Sigma}|} + \frac{\hat{\sigma}^2}{\sigma_0^2}tr(\hat{V}V_0^{-1}) + \frac{(\hat{\beta}-\beta^*)^tX^tV_0^{-1}X(\hat{\beta}-\beta^*)}{\sigma_0^2} - N\right\} \tag{4.20}$$

En utilisant le fait que $\hat{V}$ est un estimateur consistant de V_0, et en utilisant le même raisonnement que dans [77], on peut écrire $\hat{V} = V_0 + o_p(1)$. Ce qui implique que,

$$AICc = N\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + 2\frac{N(p+1)}{N-p-2} \tag{4.21}$$

est un estimateur sans biais du premier terme de l'équation (4.19). En utilisant cette approximation et le fait que $\hat{\beta}-\beta^* = (X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1}\epsilon_0$, on peut approcher le terme de l'équation (4.20).

$$E_{\theta_0}\left\{m\log\frac{|\Sigma_0|}{|\hat{\Sigma}|}\right\} \approx 0 \tag{4.22}$$

Le troisième terme de (4.20) s'écrit donc comme :

$$\begin{aligned} E_{\theta_0}\left\{\frac{\hat{\sigma}^2}{\sigma_0^2}tr(\hat{V}V_0^{-1})\right\} &= E_{\theta_0}\left\{\frac{\epsilon_0^t(\hat{V}^{-1}-\hat{V}^{-1}X(X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1})\epsilon_0}{\sigma_0^2}\right\} \\ &\approx E_{\theta_0}\left\{\frac{\epsilon_0^t(V_0^{-1}-V_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1})\epsilon_0}{\sigma_0^2}\right\} = N-p \end{aligned} \tag{4.23}$$

En outre, le quatrième terme de (4.20) devient,

$$\begin{aligned} E_{\theta_0}\left\{\frac{(\hat{\beta}-\beta^*)^tX^tV_0^{-1}X(\hat{\beta}-\beta^*)}{\sigma_0^2}\right\} &\approx E_{\theta_0}\left\{\frac{\epsilon_0^tV_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1}\epsilon_0}{\sigma_0^2}\right\} \\ &= tr(X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1}) = p \end{aligned} \tag{4.24}$$

Nous utilisons le fait que $N(\hat{\sigma}^2/\sigma_0^2)$ suit une loi de khi-deux à $(N-p)$ degrés de libertés. Le même raisonnement que dans [131, Lemma 3], permet d'écrire :

$$E_{\theta_0}\left\{N\log\frac{\sigma_0^2}{\hat{\sigma}^2}\right\} = (p+1) + o(1). \tag{4.25}$$

En utilisant (4.12) et en remplaçant (4.25), (4.22), (4.23) et (4.23) dans (4.20), on obtient les deux résultats (1) et (2) de la proposition.

(3) Nous commençons par montrer que l'équation (4.11) s'écrit :

$$\Omega_r(\theta_0) = E_{\theta_0}\left\{(N-p)\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + \log|X^t\hat{V}^{-1}X| + \frac{\sigma_0^2}{\sigma^2}tr(\hat{A}^{-1}V_0)\right\} \tag{4.26}$$

$$+E_{\theta_0}\left\{(N-p)\log\frac{\sigma_0^2}{\hat{\sigma}^2} + m\log\frac{|\Sigma_0|}{|\hat{\Sigma}|} + \log\frac{|X^tV_0^{-1}X|}{|X^t\hat{V}^{-1}X|} - tr(\hat{V}\hat{A}) + \frac{\hat{\sigma}^2}{\sigma_0^2}tr(\hat{V}A_0)\right\} \tag{4.27}$$

En adaptant les résultats de [77], nous pouvons approximer l'équation (4.26) par

$$MRC = (N-p)\log\hat{\sigma}^2 + m\log|\hat{\Sigma}| + p\log(N) + \frac{(N-p)^2}{N-p-2} \tag{4.28}$$

De même que pour (2), nous utilisons le fait que,

$$\boldsymbol{E}_{\theta_0}\Big\{m\log\frac{|\Sigma_0|}{|\hat{\Sigma}|}\Big\} \approx 0 \quad \text{et} \quad \boldsymbol{E}_{\theta_0}\Big\{\log|X^tV_0^{-1}X| - \log|X^t\hat{V}^{-1}X|\Big\} \approx 0$$

et d'un autre côté on montre que,

$$\begin{aligned}
&\boldsymbol{E}_{\theta_0}\left\{\frac{\hat{\sigma}^2}{\sigma_0^2}tr(\hat{V}A_0)\right\}\\
=\ &\boldsymbol{E}_{\theta_0}\left\{\frac{\epsilon_0^t(\hat{V}^{-1}-\hat{V}^{-1}X(X^t\hat{V}^{-1}X)^{-1}X^t\hat{V}^{-1})\epsilon_0}{\sigma_0^2}tr\{(V_0^{-1}-V_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1})\hat{V}\}\right\}\\
\approx\ &\boldsymbol{E}_{\theta_0}\left\{\frac{\epsilon_0^t(V_0^{-1}-V_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1})\epsilon_0}{\sigma_0^2}tr\{(V_0^{-1}-V_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1})V_0\}\right\}\\
=\ &(N-p)\boldsymbol{E}_{\theta_0}\left\{\frac{\epsilon_0^t(V_0^{-1}-V_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1})\epsilon_0}{\sigma_0^2}\right\} = N-p
\end{aligned} \tag{4.29}$$

De plus, nous avons,

$$\begin{aligned}
\boldsymbol{E}_{\theta_0}\Big\{tr(\hat{A}\hat{V})\Big\} &\approx \boldsymbol{E}_{\theta_0}\Big\{tr(V_0^{-1}-V_0^{-1}X(X^tV_0^{-1}X)^{-1}X^tV_0^{-1})V_0)\Big\}\\
&= N-p
\end{aligned} \tag{4.30}$$

On sait que pour une loi de khi-deux χ_ν de ν degrés de libertés nous avons $\boldsymbol{E}(\chi_\nu) = \log(2) + \Psi(\nu/2)$ et comme $(N-p)(\hat{\sigma}^2/\sigma_0^2)$ suit une loi de khi-deux à $(N-p)$ degrés de libertés, nous avons :

$$\boldsymbol{E}_{\theta_0}\Big\{(N-p)\log\frac{\hat{\sigma}^2}{\sigma_0^2}\Big\} = (N-p)\{\log\frac{2}{N-p} + \Psi(\frac{N-p}{2})\} \tag{4.31}$$

En combinant (4.28), (4.31), (4.29), (4.30) et (4.27), Nous avons le résultat 3 de la proposition.

□

Notons que tous les critères corrigés proposés ici convergent vers les critères habituels quand la taille de l'échantillon devient suffisamment grande ; c'est-à-dire quand N tend vers l'infini. Dans [33], nous donnons un tableau détaillant des résultats de simulation comparant les performances de nos critères avec les différents critères de la littérature.

4.2 Régressions pénalisées avec des erreurs non gaussiennes

On considère le modèle de régression sous forme vectoriel suivant :

$$Y = \mathbf{f} + \xi \tag{4.32}$$

où $\mathbf{f} \in \mathbb{R}^n$ est un vecteur moyen déterministe inconnu à estimer et $\xi \in \mathbb{R}^n$ un vecteur aléatoire des erreurs du modèle. On suppose que ξ a une mesure de probabilité qui satisfait une inégalité

de Poincaré faible. Selon la définition et les notations données dans [8, 89], une mesure de probabilité μ satisfait une inégalité de Poincaré faible s'il existe une fonction $\gamma : [0, +\infty) \longrightarrow \mathbb{R}^+$ telle que toute fonction locale $h : M \longrightarrow \mathbb{R}$ satisfait pour tout $s > 0$ l'inégalité

$$Var_\mu(h) \leq \gamma(s) \int |\nabla h|^2 \, d\mu + s \, Osc(h)^2, \tag{4.33}$$

où $Osc(h) = \sup h - \inf h$ est l'oscillation totale de la fonction h.

On s'intéresse au problème d'estimation de $\mathbf{f}$ par $\mathbb{X}\hat{\beta}$ dans un espace de Hilbert $\mathcal{H}$ muni d'un produit scalaire $\langle \cdot, \cdot \rangle$ et la norme correspondante $\|\cdot\|_{\mathcal{H}}$. Nous nous sommes intéressés à l'étude des performances de l'estimateur des moindres carrés pénalisés avec une pénalité convexe :

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{B}} \|\mathbf{Y} - \mathbb{X}\beta\|_n^2 + F(\beta) \tag{4.34}$$

où $\mathbb{X} : \mathcal{H} \longrightarrow \mathbb{R}^n$ est un opérateur linéaire,l'application $F : \mathcal{H} \longrightarrow \mathbb{R}$ est une pénalité convexe et $\mathbb{B}$ un sous-ensemble de $\mathcal{H}$ convexe et fermé par rapport à la norme $\|\cdot\|_{\mathcal{H}}$. La norme empirique $\|\cdot\|_n$ est définie par, $\|u\|_n^2 = \frac{1}{n}\sum_{i=1}^n u_i^2$. Notre contribution principale est d'établir des inégalités oracles dans un cadre général où le bruit des observations est issu de mesures de probabilité qui satisfont une inégalité de Poincaré faible (inégalité de gap spectral). Nous avons notamment montré que dans ce cas, le profil des inégalités de concentration dépend de la dimension de l'espace sous-jacent. Dans le cas gaussien, les inégalités de concentration habituellement établies sont indépendantes de la dimension. Notre résultat est basé sur une notion de *compatibilité* liée à la constante suivante :

$$\overline{\mu}_{c_0}(\beta) = \inf\left\{\mu > 0 : \frac{1}{c_0} \times \frac{c_0 \left|\mathcal{P}_\beta u\right|_1 - \left|\mathcal{P}_\beta^\perp u\right|_1}{\|\mathbb{X}u\|_n} \leq \mu, \ \forall u \in \mathbb{R}^p : \left|\mathcal{P}_\beta^\perp u\right|_1 \leq c_0 \left|\mathcal{P}_\beta u\right|_1\right\}. \tag{4.35}$$

Il s'agit d'une version modifiée de la constante de *compatibilité* introduite dans [15, 22] donnée par :

$$\mu_{c_0}(\beta) = \inf\left\{\mu > 0 : \left\|\mathcal{P}_\beta u\right\| \leq \mu \|\mathbb{X}u\|_n, \ \forall u : \left\|\mathcal{P}_\beta^\perp u\right\| \leq c_0 \left\|\mathcal{P}_\beta u\right\|\right\}$$

où $\mathcal{P}_\beta$ est un opérateur associé à $\mathbb{A} = span\{\beta\}$ et qui vérifie la condition de *décomposabilité* (4.36) suivante :

$$\mathcal{P}_{\mathbb{A}} A = A, \ \forall A \in \mathbb{A} \quad \text{et} \quad \|A\| + \left\|\mathcal{P}_A^\perp B\right\| = \|B + \mathcal{P}_{\mathbb{A}}(A - B)\|, \ \forall A \in \mathbb{A}, \ \forall B \in \mathcal{H}. \tag{4.36}$$

L'exemple classique d'un opérateur vérifiant cette condition de décomposabilité est la projection orthogonale.

Pour toute application positive et homogène $h : \mathcal{H} \to [0, +\infty[$ et pour tout $\tau > 0$ on note l'événement suivant :

$$\Omega = \left\{\sup_{u \in \mathcal{H} : h(u) \leq 1} \frac{1}{n} \xi^T \mathbb{X} u \leq \tau\right\}, \tag{4.37}$$

Notre résultat principal a été établi dans le cas où la pénalité $F(\beta)$ est proportionnelle à une norme de régularisation c'est-à-dire $F(\beta) = \lambda \|\beta\|$. Ce résultat est résumé dans le théorème suivant :

Théorème 4.2.1 : Soit $\mathbb{P}$ une loi de probabilité des erreurs vérifiant l'inégalité du gap spectral (4.33) avec la fonction γ on suppose que $\mathcal{P}_\beta$ vérifie la condition de décomposabilité (4.36). Supposons que $\mathbb{P}(\Omega) \geq \frac{1}{2}$ et soit $k \geq 1, c \geq 0$ et $\lambda \geq 2(\tau + k)$. Dans ce cas, l'estimateur $\hat{\beta}$, défini dans

(4.34) vérifie, avec une probabilité d'au moins $1 - 3\Theta(kc\sqrt{n})$, l'inégalité suivante :

$$\left\|\mathbb{X}\hat{\beta} - f\right\|_n^2 \leq \inf_{\beta\in\mathbb{A}}\left[\|\mathbb{X}\beta - f\|_n^2 + \frac{(\lambda + 2(\tau+k))^2}{4}\overline{\mu}_{c_0}^2(\beta)\right] + \frac{(\lambda + 2(\tau+k))^2 c^2}{4} \tag{4.38}$$

où $\Theta(u) = \inf\left\{s \in (0,1/4); \exp\left(\frac{-u}{4\sqrt{\gamma(s)}}\right) \leq s\right\}$ converge vers 0 quand u tend vers l'infini.

Démonstration. Pour ne pas encombrer ce document avec plus de détails techniques, le lecteur pourra consulter les démonstrations détaillées de ces résultats dans notre papier original [4]. L'idée générale est basée sur les deux résultats intermédiaires démontrés dans [4, Proposition 2.1] et [4, Proposition 2.2] :

(1) Nous commençons par montrer que, pour tout $\beta \in \mathbb{B}$ et pour tout $\mathbf{f} \in \mathbb{R}^n$, l'estimateur $\hat{\beta}$ vérifie :

$$\left\|\mathbb{X}\hat{\beta} - \mathbf{f}\right\|_n^2 - \|\mathbb{X}\beta - \mathbf{f}\|_n^2 \leq \frac{2}{n}\xi^T\mathbb{X}(\hat{\beta} - \beta) + F(\beta) - F(\hat{\beta}) - \left\|\mathbb{X}(\hat{\beta} - \beta)\right\|_n^2 \tag{4.39}$$

(2) Nous utilisons les résultats concernant l'inégalité de Poincaré faible (4.33) détaillés dans [88] ; nous montrons, que pour tous $k \geq 1, c \geq 0$ et $s \in (0,1/4)$, nous avons,

$$\mathbb{P}\left(\forall u \in \mathcal{H} : \frac{1}{n}\xi^T\mathbb{X}u \leq (\tau + k)\max\left(h(u), c\left\|\mathbb{X}u\right\|_n\right)\right) \geq 1 - 3\Theta(kc\sqrt{n}) \tag{4.40}$$

L'utilisation de l'inégalité (4.40) et les caractéristiques de la constante de compatibilité (4.35) nous donnent l'inégalité (4.38). □

Nous illustrons notre résultat principal avec deux exemples de distributions de probabilité satisfaisant l'inégalité de Poincaré faible (4.33) :

✠ **Exemple de distribution à queue lourde :** Nous considérons la distribution $\mathbb{P}$ dans $\mathbb{R}^n$ la mesure produit de,

$$d\mu_\alpha(t) = \frac{\alpha(1+|t|)^{-1-\alpha}}{2}dt \text{ for } \alpha > 2. \tag{4.41}$$

Cette mesure vérifie l'inégalité du gap spectral (4.33)avec

$$\gamma(s) = c_\alpha\left(\frac{s}{n}\right)^{-2/\alpha}, \quad \text{pour } s \in (0, \frac{1}{4})$$

Plus de discussions et de détails sont donnés dans [88, example 3.5]. Soit maintenant $h : \mathcal{H} \to [0, +\infty[$ une fonction positive et homogène et $\tau > 0$. Alors, il existe des constantes $t_0(\alpha) > e$ et $C(\alpha)$ telles que $\frac{tc\sqrt{n}}{n^{\frac{1}{\alpha}}} \geq t_0(\alpha)$. En appliquant le résultat (2) pour $c \geq 0$, on a :

$$\mathbb{P}\left(\forall u \in \mathcal{H} : \frac{1}{n}\xi^T\mathbb{X}u \leq (\tau + t)\max\left(h(u), c\left\|\mathbb{X}u\right\|_n\right)\right) \geq 1 - \frac{1}{2}C(\alpha)\left(\frac{\log(\frac{tc\sqrt{n}}{n^{\frac{1}{\alpha}}})}{\frac{tc\sqrt{n}}{n^{\frac{1}{\alpha}}}}\right)^\alpha.$$

Nous déduisons alors, en appliquant le théorème 4.2.1 que l'estimateur $\hat{\beta}$ (4.34) satisfait

avec une probabilité d'au moins $1 - \frac{1}{2}C(\alpha)\left(\frac{\log(\frac{tc\sqrt{n}}{\frac{1}{n^{\alpha}}})}{\frac{tc\sqrt{n}}{\frac{1}{n^{\alpha}}}}\right)^{\alpha}$

$$\left\|\mathbb{X}\hat{\beta} - f\right\|_n^2 \leq \inf_{\beta \in A}\left[\|\mathbb{X}\beta - f\|_n^2 + \frac{(\lambda + 2(\tau + t))^2}{4}\mu_{c_0}^2(\beta)\right] + \frac{(\lambda + 2(\tau + t))^2 c^2}{4} \tag{4.42}$$

- **Exemple de distribution sous-exponentielle :** Nous considérons ici $\mathbb{P}$ comme le produit de la mesure de probabilité,

$$d\nu_r = d_r e^{-|t|^r} dt, \tag{4.43}$$

où l'exposant r est dans $(0, 1)$, et la quantité positive $d_r = \frac{2\Gamma(\frac{1}{r})}{r}$ est une constante de normalisation telle que $\int d_r e^{-|t|^r} dt = 1$. Cette mesure vérifie (4.33) avec :

$$\gamma(s) = k_r\left(\log(\frac{2n}{s})\right)^{(2/r)-s}, \ s \in (0, 1/4)$$

Pour plus de détails, voir par exemple : [88, Exemple 5.4]. De même nous montrons que, pour $k \geq 0$, $c \geq 0$ et $r \in (0, 1)$, nous avons la borne explicitée dans l'équation suivante :

$$\mathbb{P}\left(\forall u \in \mathcal{H} : \frac{1}{n}\xi^T \mathbb{X}u \leq (\tau + k)\max\left(h(u), c\left\|\mathbb{X}u\right\|_n\right)\right) \geq 1 - 5\exp\left(\frac{-c_r k c\sqrt{n}}{\max\left((kc\sqrt{n})^r, \log n\right)^{1/r-1}}\right)$$

où c_r est une constante qui dépend uniquement de r. En particulier, pour un ϵ fixé, un grand n, et k vérifiant $k \geq c_r(\log\frac{10}{\epsilon})(\log n)^{\frac{1}{r}-1}$, on a :

$$\mathbb{P}\left(\forall u \in \mathcal{H} : \frac{1}{n}\xi^T \mathbb{X}u \leq \left(\tau + \frac{k}{c\sqrt{n}}\right)\max\left(h(u), c\left\|\mathbb{X}u\right\|_n\right)\right) \geq 1 - \frac{\epsilon}{2}$$

Dans cet exemple, l'estimateur $\hat{\beta}$ donné par le problème de minimisation (4.34) satisfait avec une probabilité d'au moins $1 - 5\exp\left(\frac{-c_r k c\sqrt{n}}{\max\left((kc\sqrt{n})^r, \log n\right)^{1/r-1}}\right)$,

$$\left\|\mathbb{X}\hat{\beta} - f\right\|_n^2 \leq \inf_{\beta \in A}\left[\|\mathbb{X}\beta - f\|_n^2 + \frac{(\lambda + 2(\tau + k))^2}{4}\mu_{c_0}^2(\beta)\right] + \frac{(\lambda + 2(\tau + k))^2 c^2}{4}. \tag{4.44}$$

En particulier, pour $\lambda \geq 2(\tau + \frac{k}{c\sqrt{n}})$, alors avec une probabilité d'au moins $1 - \frac{\epsilon}{2}$,

$$\left\|\mathbb{X}\hat{\beta} - f\right\|_n^2 \leq \inf_{\beta \in A}\left[\|\mathbb{X}\beta - f\|_n^2 + \frac{(\lambda + 2\left(\tau + \frac{k}{c\sqrt{n}}\right))^2}{4}\mu_{c_0}^2(\beta)\right] + \frac{(\lambda + 2(\tau + \frac{k}{c\sqrt{n}}))^2 c^2}{4}$$

Nous avons appliqué nos résultats au cas classique du Lasso, et du group Lasso où $\mathcal{H} = \mathbb{R}^p$ muni de la norme euclidienne $|\cdot|_2$.

(1) L'estimateur *Lasso* pour lequel :

- La norme de régularisation $\|\cdot\|$ est la norme ℓ_1 sur $\mathbb{R}^p$, $|\beta|_1 = \sum_{j=1}^p \left|\beta_j\right|$ sur $\mathbb{R}^p$.
- L'opérateur linéaire $\mathcal{P}_\beta$ est celui de la projection orthogonale sur le sous-espace engendré par $\left\{e_j : j \in Supp(\beta)\right\}$ où $(e_j)_{1 \leq j \leq p}$ est la base canonique de $\mathbb{R}^p$

(2) L'estimateur *Group-lasso* :

- La norme de régularisation $\|\cdot\|$ définie comme,

$$\|\beta\| = |\beta|_{2,1} = \sum_{k=1}^{M} \left(\sum_{j \in G_k} \beta_j^2 \right)^{1/2} = \sum_{k=1}^{M} \left| \beta_{G_k} \right|_2$$

- L'opérateur linéaire $\mathcal{P}_\beta$ est défini comme la projection orthogonale sur le sous-espace engendré par $\left\{ e_j : j \in \underset{k \in Supp_G(\beta)}{\cup} G_k \right\}$ où $G_1, \dots, G_M$ est une partition de l'ensemble d'indices $\{1, \dots, p\}$ et $Supp_G(\beta) = \left\{ k \in \{1, \dots, M\} : \beta_{G_k} \neq 0 \right\}$ le support par groupe du vecteur β et M le nombre de groupes.

Les inégalités oracles de l'erreur de prédiction pour le lasso et group lasso sont établies sous la condition de compatibilité (4.35) qui n'est pas en général donnée de manière explicite. Dans le cas particulier du Lasso et avec les deux exemples traités plus haut, nous avons établi une borne supérieure explicite de la constante de compatibilité pour le lasso :

- En fonction de la constante s de parcimonie (*sparcité*) c'est-à-dire une constante s telle que : $|Supp(\beta)| \leq s$.
- En fonction de la corrélation maximale entre les colonnes de la matrice de design $\mathbb{X}$:

$$\kappa = \sup_{1 \leq i \neq j \leq p} \frac{\left| \left\langle \mathbb{X} e_i, \mathbb{X} e_j \right\rangle \right|}{n}$$

4.3 Modèles de régressions et applications en volcanologie

Les nuages de cendres émis lors des éruptions volcaniques représentent un risque majeur susceptible d'avoir des conséquences dramatiques sur l'environnement et les personnes. Les effets néfastes de ces particules peuvent parfois durer des années et affecter de manière significative l'équilibre climatique régional et mondial. L'un des défis majeurs des volcanologues et des autorités civiles, est d'évaluer l'ampleur d'une éruption volcanique et ses conséquences sur les populations et les biens.

En particulier, nous cherchons à prévoir l'emplacement et la concentration des particules dans les nuages de cendres pendant les heures qui suivent une éruption volcanique, mais également quantifier les dépôts au sol.
Les techniques actuelles sont basées sur des modèles de transport et de dispersion des cendres (*VATD : Volcanic Ash Transport and Dispersion*) qui exploitent des paramètres en entrée tels que la hauteur du panache, la répartition en taille du téphra et le taux de l'éruption des masses (*MER : Masse Eruption Rate*). Ce dernier est un paramètre clé car il contrôle directement la quantité de cendres injectées dans l'atmosphère. Cependant, en mode temps réel,

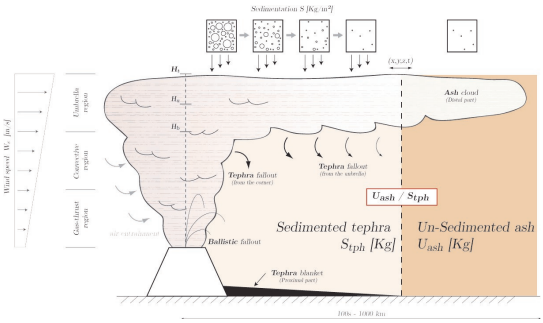


FIGURE 4.1 – Illustrations des mécanisme de dispersion sédimentation.

le MER est très difficile à obtenir et le seul moyen actuel qui permet de l'évaluer de manière assez précise, est de passer par des prospections et des études détaillées de dépôt au sol ; cette méthode peut prendre des mois de collecte et de traitements ; par conséquent elle ne correspond pas aux exigences du temps réel en cas de crise volcanique. Le seul modèle utilisé actuellement est basé généralement sur des lois empiriques permettant de déterminer le MER en fonction de la hauteur du panache ; celui-ci a l'avantage d'être mesuré assez facilement (par radars météorologiques au sol , aéronefs...). Cependant, cette technique peut être difficile à appliquer notamment à cause des incertitudes sur la hauteur ; par exemple, en raison de mauvaises conditions météorologiques, de l'existence de vents de travers... Pour pallier le manque de certitude sur la hauteur, d'autres paramètres doivent être rajoutés. Ces derniers doivent être disponibles en temps réel pour pouvoir être exploités en cas de crise volcanique. Au cours des deux dernières décennies, le développement des techniques satellitaires a permis des mesures plus fréquentes et plus précises concernant les éruptions.

4.3.1 Lien entre les dépôts et mesures satellite

Nous nous intéressons au téphra provenant d'un panache lors d'une éruption explosive et nous distinguerons les parts sédimentées et non sédimentées. La fraction sédimentée s'étend sur des distances variables (allant de 5 à 150 km) en fonction de l'importance de l'éruption [162, 154, 128] , mais cela reste "proche" pour une observation satellite. La fraction non sédimentée correspond à de fines particules (plus petites que $30\mu m$) et peuvent être transportées sur des distances beaucoup plus importantes (de l'ordre du millier de kilomètres) et restant dans l'atmosphère pendant plusieurs jours [128, 59].

Le déplacement et la dispersion des nuages de cendres volcaniques sont principalement contrôlés par le mécanisme de sédimentation des particules, basé principalement sur la taille des grains et leur densité [128]. Le but est donc d'établir l'éventuelle existence d'une relation entre la masse de téphra sédimenté et les cendres non sédimentées, sur la base d'éruptions volcaniques bien documentées. Si une telle relation existe, des estimations de la masse de cendre non sédimentée, basées sur des observations satellites, pourraient être utilisées pour évaluer la masse de téphra sédimenté et le MER, en sommant les parts sédimentées et non sédimentées et en tenant compte de la durée d'émission.

Généralement, le MER n'est estimé qu' a posteriori par une étude fine des dépôts, sans tenir compte de la part de cendres.

Mesures des données satellites : Les satellites météorologiques sont utilisés depuis longtemps pour la détection et le suivi des cendres volcaniques. Par exemple, les propriétés de transmission des ondes infrarouges thermiques des nuages de cendres ont été utilisées très tôt pour discriminer et caractériser les cendres [161, 160].

Une simple méthode est basée sur des gradients de brillance de température, utilisant l'extinction progressive des cendres entre

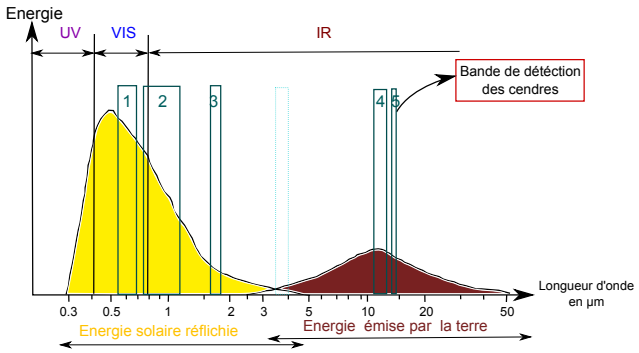


FIGURE 4.2 – Longueur d'onde et détection des cendres

11 μm et 12 μm. Pour une complète caractérisation, on utilise généralement un modèle de transfert radiatif (RTM) et une formulation de la diffusion électromagnétique de Mie, a été donnée par [150]. Cela permet de quantifier tous les paramètres importants, comme le rayon effectif ou la masse des cendres volcaniques dans le nuage de cendres.

Mesures des données au sol : Les études au sol de chutes de téphra permettent d'estimer le volume ou la masse (e.g. [74, 39]). Elles sont sujettes à une grande sensibilité aux divers paramètres : la loi de variation empirique, l'abondance des données sur champ [90, 32]. Il a été choisi de ne considérer que les variables qui peuvent être disponibles en temps réel pendant une crise. Ainsi, plusieurs données caractéristiques des volcans ont été utilisées pour l'élaboration du modèle :

- P1 : présence ou non, sur une échelle allant de 0 à 2, de nuage ou de pluie,
- P2 : type de dépôt au sol (catégorie allant de 0 à 2),
- P3 : caractéristique de perte du terrain (catégorie allant de 0 à 3),
- P4 : phréatomagmatisme (catégorie allant de 0 à 3),
- P5 : composition du magma (catégorie allant de 0 à 3),
- H : hauteur du panache (en km),
- S_2 : masse estimée de téphra sédimenté via les données satellite,
- S_1 : masse totale de téphra sédimenté mesuré au sol (variable d'intérêt).

Notre objectif est d'agréger les données satellites à divers paramètres accessibles décrits ci-dessus (acidité, connaissance du magma,...) pour obtenir une prévision de la masse au sol la plus précise possible et qui pourrait servir à une alternative crédible à la formulation empirique basée sur la seule hauteur. Nous n'avons que peu de données et nous ne pouvons nous permettre de conserver trop de paramètres car il est important de pouvoir apporter une nouvelle formule physiquement explicable, facile à implémenter, avec des paramètres de site récupérables sans trop de difficultés. Pour cela, nous avons utilisé des techniques fines de choix de modèles et de sélection de variables à même d'optimiser avec parcimonie les paramètres d'entrée.

4.3.2 Méthodologie statistique : choix et sélection de variables

L'idée consiste à considérer un mélange de K régressions qu'on peut formaliser par :

$$f(x, y, \Phi) = \sum_{k=1}^{K} \pi_k f(y, x^t \beta_k, \sigma_k), \tag{4.45}$$

où, $f(y, x^t\beta_k, \sigma_k)$ désigne la densité d'une loi normale de moyenne $x^t\beta_k$ et de variance σ_k^2 ; les coefficients $0 < \pi_k < 1$ vérifient $\sum_{k=1}^{K} \pi_k = 1$; tous ces paramètres seront regroupés dans un vecteur que nous noterons $\Phi = (\pi_k, \beta_k, \sigma_k)_{k=1\ldots,K}$.

Nous nous intéressons ici au cas où la taille des échantillons est relativement petite et/ou le nombre de variables explicatives est élevé. Pour s'adapter à ce genre de situations, les régressions de type *Ridge* ou *Lasso*, ont été introduites pour contribuer au regroupement des variables ou pour réduire la variance des résidus. Cette idée a été introduite par[138] puis améliorée [96] ; cela consiste à régulariser l'estimation des paramètres par l'introduction d'un terme de pénalité dans le problème d'estimation de maximum de vraisemblance comme nous l'avons détaillé dans la section précédente. Le choix de la pénalisation permet donc de mettre l'accent soit sur la sélection

ou bien sur la réduction de la variance. Ces techniques n'ont pas cessé de susciter l'intérêt de la communauté statistique. Plusieurs améliorations théoriques et numériques ont été introduites dans la littérature : nous citons entre autres, [47, 41, 35].
La limitation de ces deux techniques réside dans le fait qu'elles sont soit spécialisées dans la sélection (regroupement des variables) ou soit dans la réduction des erreurs quadratiques. L'idée est donc d'introduire un nouveau terme de pénalisation qui permettra d'une part de favoriser le regroupement des prédicteurs afin de déterminer le nombre de classes K dans le problème de mélange de régressions (5.1) ; d'autre part, il permet d'améliorer l'estimation des paramètres affectés à chaque classe. Pour atteindre cet objectif, au lieu d'utiliser un *Lasso* classique, nous introduisons dans la fonction de pénalisation la différence entre les paramètres de chaque classe. En effet, considérons $(y_1, \dots, y_n)^t$ un vecteur d'observations d'une variable d'intérêt Y et $(x_1^t, \dots, x_n^t)^t$ une matrice d'observations des variables explicatives X ; les x_i sont des vecteurs de $\mathbb{R}^p$. Le problème de sélection et d'estimation revient donc à maximiser la fonction log-vraisemblance pénalisée suivante :

$$\ell(\Phi, Y, X) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f(y, x^t\beta_k, \sigma_k)\right) - \lambda\, \mathbf{Pe}(\beta_k); \tag{4.46}$$

Le terme de pénalisation **Pe(.)** est donné par :

$$\mathbf{Pe}(\beta_k) = \alpha \sum_{j=1}^{p} |\beta_{kj}| + (1-\alpha) \sum_{j=2}^{p} \sum_{l=1}^{j-1} |\beta_{kj} - \beta_{kl}|, \tag{4.47}$$

où $0 < \alpha < 1$. Il s'agit ici d'une pénalité du type *pairwise fused lasso*[41]. Le problème de maximisation de la fonction (4.46) est difficile à résoudre directement et de manière analytique ; nous avons donc choisi d'estimer les paramètres en maximisant de manière itérative.

En considérant la structure du terme de pénalité donné dans (4.47) et que nous l'appliquons à notre modèle de mélange (5.1), l'estimateur du vecteur paramètre Φ sera obtenu par maximisation $\ell(\Phi, Y, X)$. L'équation (4.46) étant difficile à maximiser directement, nous utilisons un algorithme de type *EM : Expectation Maximisation* [1]. Pour atteindre cet objectif, nous introduisons d'abord une variable auxiliaire Z ; il s'agit d'une matrice de dimension $n \times K$, tel que z_{ik} est égal à 1 si y_i est générée à partir de la k^{eme} composante du mélange, elle est égale à 0 sinon. Dans ce cas, la log-vraisemblance des données complètes, tenant compte des observations de X, Y et Z, est donnée par la formule suivante :

$$\ell_c(\Phi,\ Y,\ Z,\ X) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik}\{\log \pi_k + \log[f_k(y_i,\ x_i^t\beta_k,\ \sigma_k)]\} \tag{4.48}$$

1. L'algorithme EM a été introduit dans les années 70. il est inspiré des techniques bayesiennes et des calculs de probabilité a posteriori. Cet algorithme se déroule de manière itérative ; chaque itération comporte deux étapes :
   - L'étape E (*Expectation*) : elle consiste à calculer l'espérance de la vraisemblance en tenant compte des variables observées et des paramètres d'initialisation,
   - L'étape M (*Maximisation*) : Elle consiste à estimer les paramètres en maximisant la vraisemblance trouvée à l'étape E.

On utilise ensuite les paramètres trouvés en M comme valeurs initiales pour l'itération suivante et ainsi de suite jusqu'à la convergence de l'algorithme.

où, avec $Y = (y_1, \ldots, y_n)^t$, X est une matrice $n \times p$ telle que $X = (x_1, \ldots, x_n)^t$,

$$\log[f_k((y_i, x_i^t\beta_k, \sigma_k)] = -\frac{\log(2\pi) + \log(\sigma_k^2)}{2} - \frac{(y_i - x_i^t\beta_k)^2}{2\sigma_k^2}.$$

Dans les cas classiques où on connaît à l'avance l'appartenance des observations aux classes, l'opérateur que l'algorithme *EM* doit maximiser itérativement est défini par :

$$\mathcal{Q}(\Phi, \Phi^{(m)}) = \mathbb{E}(\ell_c(\Phi, Y, Z, X)|Y, \Phi^{(m)}) \tag{4.49}$$

où $\Phi^{(m)}$ est la valeur courante à l'itération m et ℓ_c est donnée par l'équation (4.46). Pour adapter l'algorithme à notre problème, où ni le nombre de classes n'est connu, ni leur répartition, nous proposons de maximiser l'opérateur pénalisé suivant, au lieu de l'opérateur (4.49) :

$$\mathcal{Q}_p(\Phi, \Phi^{(m)}) = \mathbb{E}(\ell_c(\Phi, Y, Z, X)|Y, \Phi^{(m)}) - \lambda \sum_{k=1}^{K} \mathbf{Pe}(\beta_k) \tag{4.50}$$

Le terme de pénalité **Pe** est défini dans (4.47).

L'étape E :

Au niveau de l'itération m, nous disposons d'un vecteur $\Phi^{(m)}$ en entrée. Dans cette étape, on calcule la fonction Q_p en remplaçant z_{ik} par $\tau_{ik} = \mathbb{E}(z_{ik}|y_i)$. En utilisant une inversion Bayesienne, il est facile de voir que :

$$\tau_{ik}^{(m+1)} = \frac{\pi_k^{(m)} f_k(y_i, x_i^t\beta_k^{(m)}, \sigma_k^{(m)})}{\sum_{k=1}^{K} \pi_k^{(m)} f_k(y_i, x_i^t\beta_k^{(m)}, \sigma_k^{(m)})}$$

Ce qui permet de déduire la formule (4.50) de manière explicite ;

$$\mathcal{Q}_p(\Phi, \Phi^{(m)}) = \sum_{k=1}^{K}\sum_{i=1}^{n} \tau_{ik}^{(m+1)}\{\log \pi_k + \log[f_k(y_i, x_i^t\beta_k, \sigma_k)]\} - \lambda \sum_{1\le k\le K} \mathbf{Pe}(\beta_k) \tag{4.51}$$

avec $\mathbf{Pe}(\beta_k)$ est donnée dans (4.47). L'opérateur Q_p est maintenant prêt pour être utilisé par l'étape M pour estimer les paramètres du modèle par maximum de vraissemblance.

Etape M :

Dans cette étape, le but est de maximiser $Q_p(\Phi, \Phi^{(m)})$ par rapport $(\pi_k, \beta_k, \sigma_k)_k$. En dérivant (4.51) il est facile de voir que :

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(m+1)}}{n}$$

Pour $k = 1, \ldots, K$, nous avons

$$\sigma_k^{(m+1)} = \frac{(\tilde{Y}_k^{(m)} - \tilde{X}_k^{(m)}\beta_k^{(m)})^t(\tilde{Y}_k^{(m)} - \tilde{X}_k^{(m)}\beta_k^{(m)})}{\mathbf{tr}(W_k^{(m)})}, \tag{4.52}$$

avec $W_k^{(m)} = \mathrm{diag}(\tau_k^{(m)})$, $\tau_k^{(m)} = (\tau_{1k}^{(m)}, \ldots, \tau_{nk}^{(m)})^t$, $\tilde{X}_k^{(m)} = [\tilde{W}^{(m)}]^{1/2}X$ et $\tilde{Y}_k^{(m)} = [\tilde{W}^{(m)}]^{1/2}Y$.

Nous notons $\beta = (\beta_1, \ldots, \beta_K)$ et $\beta_k = (\beta_{k1}, \ldots., \beta_{kp})^t$. En utilisant les techniques classiques de la régression linéaire nous voyons facilement que l'estimation de β à l'itération m revient à la

minimisation de :

$$\hat{\beta}^{(m+1)} = \arg\min_{\beta} \sum_{k=1}^{K} n_k \log(\sigma_k^{(m+1)}) + \frac{1}{2} \sum_{k=1}^{K} \frac{\|\tilde{Y}_k^{(m)} - \tilde{X}_k^{(m)}\beta_k\|^2}{\left[\sigma_k^{(m+1)}\right]^2}$$
$$+ \lambda \sum_{1\le k\le K} \{\alpha \sum_{1\le j\le p} |\beta_{kj}| + (1-\alpha) \sum_{j=2}^{p} \sum_{l=1}^{j-1} |\beta_{kj} - \beta_{kl}|\}$$

Ce problème peut être décomposé en K problèmes d'optimisation. C'est-à-dire, pour $k = 1, \dots, K$:

$$\hat{\beta}_k^{(m+1)} = \arg\min_{\beta_k} n_k \log(\sigma_k^{(m+1)}) + \frac{1}{2} \frac{\|\tilde{Y}_k^{(m)} - \tilde{X}_k^{(m)}\beta_k\|^2}{\left[\sigma_k^{(m+1)}\right]^2} \tag{4.53}$$
$$+ \quad \lambda\{\alpha \sum_{1\le j\le p} |\beta_{kj}| + (1-\alpha) \sum_{j=2}^{p} \sum_{l=1}^{j-1} |\beta_{kj} - \beta_{kl}|\}$$

Or le premier terme ne dépend pas de β_k car, dans (4.52), $\sigma_k^{(m+1)}$ ne dépend que des paramètres d'entrée de l'étape M, donc (4.53) devient,

$$\hat{\beta}_k^{(m+1)} = \arg\min_{\beta_k} \frac{1}{2} \frac{\|\tilde{Y}_k^{(m)} - \tilde{X}_k^{(m)}\beta_k\|^2}{\sigma_k^2(m+1)} + \lambda\{\alpha \sum_{1\le j\le p} |\beta_{kj}| + (1-\alpha) \sum_{j=2}^{p} \sum_{l=1}^{j-1} |\beta_{kj} - \beta_{kl}|\}$$

Ce qui peut être simplifié en prenant $\lambda_1 = 2\lambda[\sigma_k^{(m+1)}]^2$, on aura donc :

$$\hat{\beta}_k^{(m+1)} = \arg\min_{\beta_k} \|\tilde{Y}_k^{(m)} - \tilde{X}_k^{(m)}\beta_k\|^2 + \lambda_1\{\alpha \sum_{1\le j\le p} |\beta_{kj}| + (1-\alpha) \sum_{j=2}^{p} \sum_{l=1}^{j-1} |\beta_{kj} - \beta_{kl}|\}. \tag{4.54}$$

Nous montrons que ce problème est équivalent à un problème de Lasso classique. Ce résultat représente un intérêt remarquable notamment du point de vue computationnelle car il permet de recycler les algorithmes de Lasso existants dans la littérature. La technique consiste seulement à reparamétrer le problème en réalisant la transformation suivante (voir [41]) :

$$\theta_{j0}^{(k)} = \beta_{kj} \qquad \text{pour } 1 \le j \le p \tag{4.55}$$
$$\theta_{jl}^{(k)} = \beta_{kj} - \beta_{kl} \qquad \text{pour } 1 \le l < j \le p. \tag{4.56}$$

D'abord on voit bien que $\theta_{jl}^{(k)}$ doit vérifier :

$$\theta_{jl}^{(k)} = \theta_{j0}^{(k)} - \theta_{l0}^{(k)} \qquad \text{pour} \qquad 1 \le l < j \le p \tag{4.57}$$

Notons pour chaque classe k le nouveau vecteur de paramètres défini par, $\theta^{(k)} = (\theta_{10}^{(k)}, \dots, \theta_{p0}^{(k)}, \theta_{21}^{(k)}, \dots, \theta_{p(p-1)}^{(k)})$ et considérons $\mathcal{I} = \{(i, j) \quad | \quad 0 \le j < i \le p\}$ l'ensemble des indices des composantes de $\theta^{(k)}$ vérifiant la contrainte (4.57). Le problème d'optimisation (4.54) devient donc :

$$\hat{\theta}^{(k)^{(m+1)}} = \arg\min_{\theta^{(k)}} \|\tilde{Z}_k^{(m)} - \tilde{D}_k^{(m)}\theta^{(k)}\|^2 + \lambda_1\{\alpha \sum_{1\le j\le p} |\theta_{j0}^{(k)}| + (1-\alpha) \sum_{j=2}^{p} \sum_{l=1}^{j-1} |\theta_{jl}^{(k)}|\}$$

avec $\tilde{Z}_k^{(m)} = (\tilde{Y}_k^{(m)^t}, 0_{\tilde{p}}^t)^t$ et $\tilde{D}_k^{(m)} = (\tilde{X}_k \,|\, 0_{n\times(\tilde{p})})$ où $\tilde{p} = C_p^2 = \frac{p(p-1)}{2}$. Ce qui permet d'écrire la nouvelle

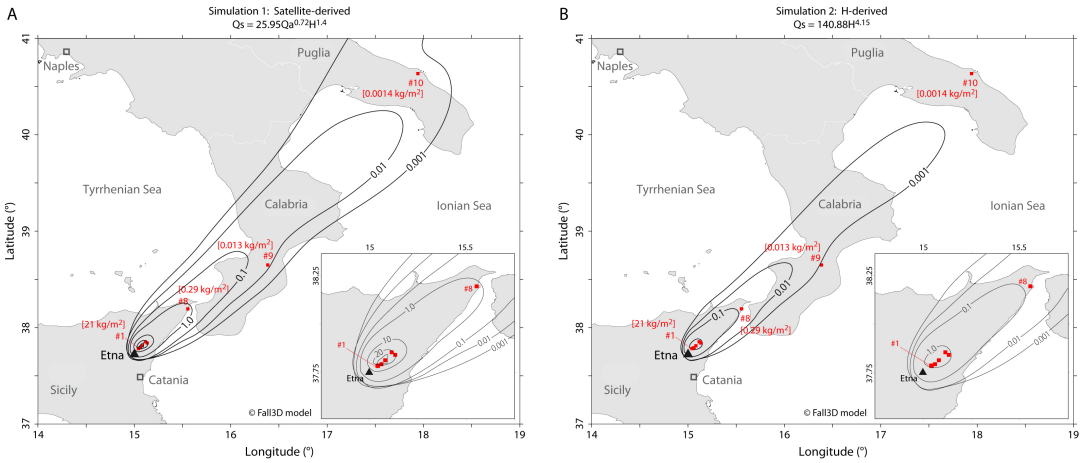


FIGURE 4.3 – Exemple simulations de dispersion de cendres basée sur notre modèle à droite et le modèle classique à gauche.

paramétrisation comme :

$$\hat{\theta}^{(k)^{(m+1)}} = \arg\min_{\theta^{(k)}} \|\tilde{Z}_k - \tilde{\tilde{D}}_k^{(m)} \tilde{\theta}^{(k)}\|^2 + \lambda_1 \sum_{t \in \mathcal{I}} |\tilde{\theta}_t^{(k)}|\}$$

avec $\tilde{\tilde{D}}_k^{(m)} = \tilde{D}_k^{(m)} \text{diag}(\alpha, \ldots \alpha, (1-\alpha), \ldots, (1-\alpha))^{-1}$.

4.3.3 Exemples d'applications en volcanologie

Prédiction du MER en vue de simulations : L'application de la méthodologie décrite plus haut a permis de déterminer les covariables permettant de prédire de manière optimale le MER. Nous avons ainsi pu montrer que, parmi toutes les variables potentiellement explicatives, le MER peut s'expliquer par la hauteur. En plus de cela, l'introduction des données satellites permet d'améliorer significativement le modèle actuellement utilisé. En outre, nous montrons également que les paramètres de ce modèle dépendent grandement de la composition chimique du magma. Pour illustrer ces résultats, la figure 4.3 donne un exemple de simulations de retombées de téphra pendant l'éruption du volcan Etna du 23 février 2013. Le modèle de dispersion/sédimentation que nous avons utilisé pour les simulations est FALL3D qui représente maintenant un standard utilisé à l'INGV (*Istituto Nazionale di Geofisica e Vulcanologia*) en Italie et à travers le monde. Dans cette figure, les dépôts de retombées de téphra modélisés sont représentés sous la forme de courbes de niveau d'isomasse (lignes noires). Ces contours reflètent la *charge de téphra calculée* en kg/m2 à partir du modèle. Aux différents emplacements matérialisés par des carrés rouges, des prélèvements ont été réalisés sur le terrain et ont permis d'obtenir la *charge de téphra réellement mesurée* au sol. Contrairement au modèle actuel, notre proposition permet d'estimer plus précisément le MER et par conséquent les dépôts au sol . D'autres modèles de simulation comme le modèle MOCAGE du VAAC de Toulouse ont été testé sur d'autres volcans comme l'éruption plinienne du Kelut du 13 février 2014; plus de discussions sont détaillées dans [12].

Développement de cartes d'aléas volcaniques : Ce travail a été réalisé en collaboration avec le Laboratoire Magmas et Volcans (LMV). L'objectif est de développer des ensembles de cartes probabilistes des aléas volcaniques tenant compte de la multitude des sources d'incertitudes. Ces

dernières sont liées à la sensibilité du modèle numérique utilisé pour les simulations, au comportement futur des volcans, et à l'incertitude sur la variation des paramètres physiques de la source éruptive. Nous avons obtenu trois résultats principaux décrits dans trois articles :

- Le premier [10] a permis de quantifier l'incertitude et la sensibilité des deux modèles numériques PLUME-MoM et HYSPLIT, qui simulent d'une part les mécanismes de production de panaches volcaniques, et d'autre part le transport et la sédimentation des cendres. Nos résultats ont été testés en confrontant notre modèle aux données observées sur le terrain de quatre éruptions des volcans andins (en Équateur et au Chili) d'amplitudes et de styles éruptifs différents. Des indicateurs quantifiant l'ampleur de la sous-estimation ou de la sur-estimation du modèle ont été ainsi évalués [10].
- Dans un deuxième papier [7], nous avons donné une quantification de la probabilité d'occurrence d'un type d'éruption et des incertitudes sur les paramètres d'entrée. Cette évaluation a été effectuée lors d'une session d'élicitation à laquelle ont participé 20 experts. Cette session a permis de déterminer :
  - les probabilités relatives pour différents styles éruptifs, pour la prochaine éruption ou bien pour au moins un type d'éruption dans les 100 prochaines années,
  - les gammes d'incertitude, pour chaque type d'éruption, sur la masse totale des dépôts de retombées, sur la durée totale de l'éruption ainsi que la hauteur moyenne du panache.
- Le but du troisième article [3] est de développer des cartes d'aléa probabilistes du point de vue de l'accumulation ou du dépôt des cendres au sol, pour les deux volcans (Cotopaxi et Guagua Pichincha) en Équateur. Pour la production de ces cartes, les paramètres d'entrée du modèle (y compris les conditions atmosphériques), les approximations physiques des paramètres d'entrée du modèle numériques ainsi que les probabilités d'occurrence de différents types d'éruptions dans des délais spécifiques sont parmi les sources d'incertitude les plus importantes. Nous utilisons les modèles PLUME-MoM/HYSPLIT couplés à une procédure de quantification de l'incertitude pour produire :
  - des cartes, à l'usage des autorités Equatoriennes (aviation, ministère de l'agriculture, ministère du logement), qui indiquent les seuils et les différentes probabilités d'accumulation des cendres en cas d'éruption ; voir figure 4.4 ;
  - des cartes, à l'usage de l'IRD en Équateur, qui indiquent une probabilité des différents seuils d'accumulation de cendres et des courbes d'aléa pour 10 sites sensibles de la ville de Quito ; Figure 4.5.

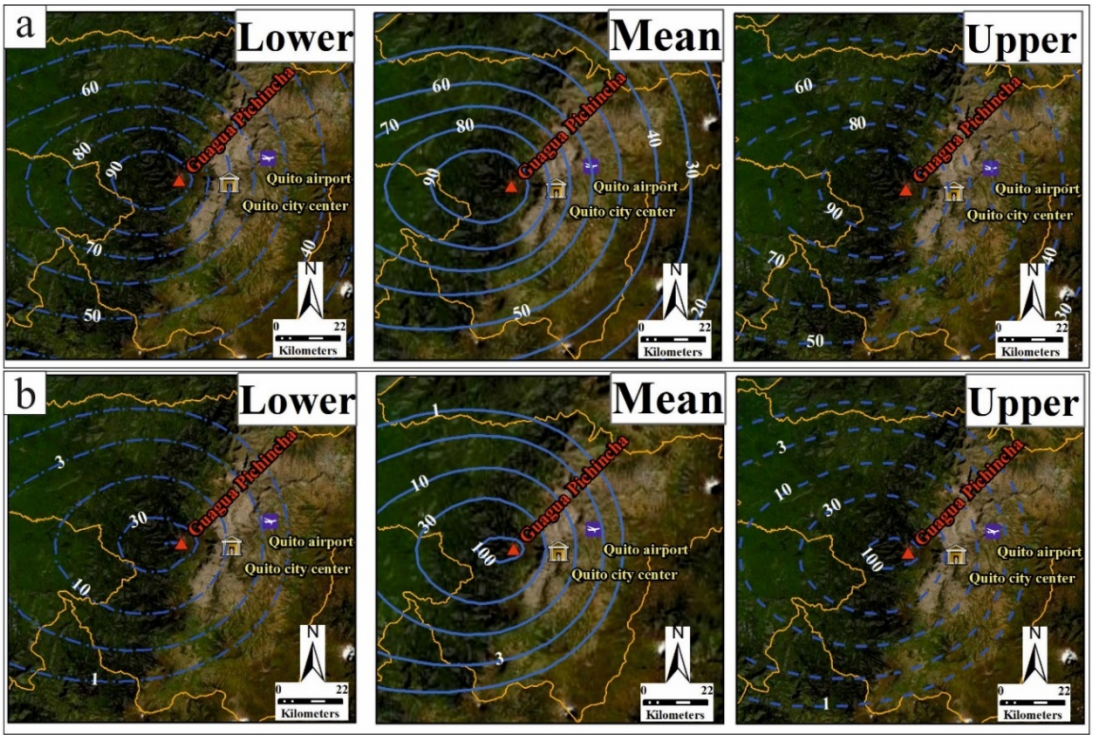


FIGURE 4.4 – Exemples de cartes probabilistes pour le Volcan Guagua Pichincha (Equateur). a) Seuil d'accumulation de 1 mm de cendres, pour différentes probabilités (lignes bleues) ; b) Seuils d'accumulation de cendres (1, 3, 10, 30, 100 mm) pour une probabilité fixe (95%). Les termes désignent le résultat (Mean) et l'incertitude (Lower, Upper) du modèle.

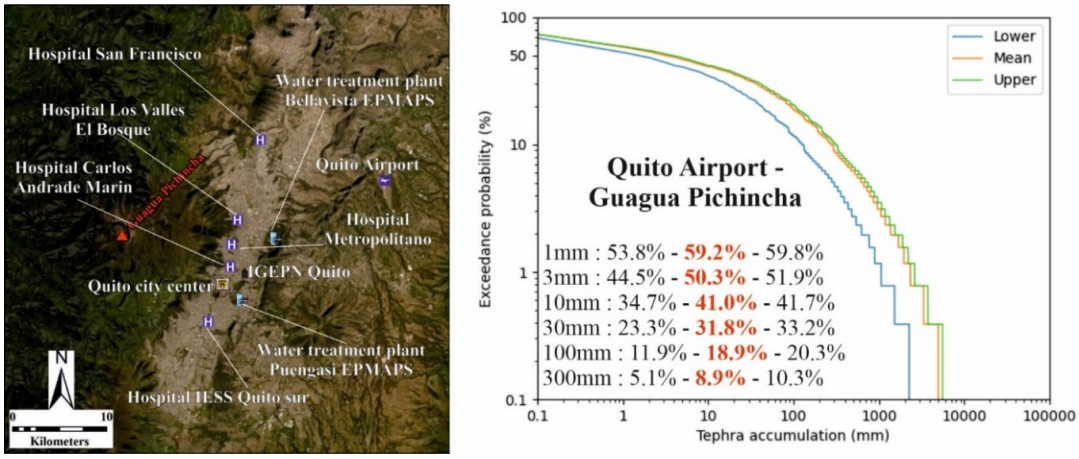


FIGURE 4.5 – Gauche : sites sensibles de la ville de Quito pour lesquels a été produite une courbe de probabilité d'aléa pour chaque volcan. Droite : exemple de courbe d'aléa pour l'aéroport de Quito et le volcan Guagua Pichincha (droite). Les nombres désignent l'épaisseur d'accumulation des cendres (1-300 mm) et les pourcentages de probabilités (valeurs moyenne, minimale et maximale)

Analyse temps-fréquences

La représentation temporelle d'un signal est fondamentale pour détecter les variations d'un processus, mais pour comprendre en profondeur ses caractéristiques, une autre représentation joue également un rôle central ; il s'agit de ses caractéristiques fréquentielles. Au cours des cinquante dernières années de recherche en statistique, en traitement du signal et en probabilité appliquée, une riche littérature a été produite pour étudier les propriétés et la relation entre l'analyse du signal dans le domaine temporel et dans le domaine fréquentiel. Récemment, on a assisté à une résurgence de ces idées dans la littérature sur l'apprentissage automatique, avec de nouvelles applications. Divers outils permettant de passer d'un domaine à l'autre ont été introduits et étudiés dans la littérature. Les techniques les plus connues sont celles basées sur l'analyse de Fourier et l'analyse multi-résolution ou d'ondelettes ainsi que leurs variantes...

Récemment, avec le besoin de développer des méthodes utilisant seulement des données observées, le concept de décomposition modale empirique (EMD) a été introduit pour la première fois par [129]. Le principe essentiel de l'EMD est de décomposer les signaux en une somme de certaines fonctions oscillatoires appelées modes propres ou modes intrinsèques (IMF : *Intrinsic Mode Functions*). Par rapport aux ondelettes ou à l'analyse de Fourier, une IMF représente un mode oscillatoire plus général qu'une harmonique simple. En effet, contrairement à ces derniers, une IMF peut avoir une amplitude et une fréquence variable dans le temps. Comparée aux décompositions traditionnelles de Fourier et d'ondelettes, l'EMD présente plusieurs avantages. Premièrement, outre le fait qu'elle est intuitive et peut être simplement mise en oeuvre, elle est adaptative et peut être facilement interprétée physiquement. Deuxièmement, comme discuté par [129, 137], elle est plus appropriée pour traiter des données provenant de systèmes non linéaires et/ou non stationnaires. Enfin, contrairement à la décomposition en ondelettes, l'EMD ne nécessite pas de déterminer une fonction de base a-priori. Notre contribution dans ce domaine d'analyse temps-fréquences est résumée dans les points suivants :

- Détection de composantes fréquentielles pour données physiologiques : Le but de ce travail est d'exploiter le comportement du rythme cardiaque pour détecter des situations de stress dans différentes populations : médecins urgentistes, sportifs amateurs, comportements animaliers... La régulation du rythme cardiaque résulte de deux systèmes antagonistes : le système sympathique et le système parasympathique qui correspondent à deux

bandes de fréquences différentes. Ainsi, tout déséquilibre dans ces deux systèmes se traduit par une variation de la modulation de la fréquence cardiaque. Cette alternance entre équilibre et déséquilibre (en l'occurrence ici une grande variabilité de la fréquence) est considérée comme un indicateur de bonne santé. Une diminution de cette variabilité est directement liée au stress, à la fatigue et à la diminution des performances physiques. L'étude dynamique du rythme cardiaque est faite par la combinaison de deux méthodes statistiques : dans un premier temps, utiliser des ondelettes pour l'extraction de l'énergie spectrale dans les bandes de hautes/basses fréquences correspondantes respectivement aux deux systèmes parasympathique et orthosympathique. Dans un deuxième temps, nous utilisons des techniques de classification non supervisées pour élaborer une typologie de l'activité cardiaque en distinguant des groupes homogènes ou des profils d'états reflétant l'état physiologique d'un individu.

- Reformulation de l'EMD : Dans un premier temps, nous avons cherché à expliciter une formulation théorique de cette décomposition. Cela nous a amené à définir ce que nous avons appelé IMF élémentaire (EIMF : *Elementary Intrinsic Mode Functions*) ; cette définition utilise les propriétés de convexité et une notion de symétrie bien explicite. Contrairement aux extremas, la notion de convexité est facilement généralisable au cas des signaux vectoriels et multivariés. En utilisant les propriétés de convexité et de symétrie des EIMF, nous avons proposé une décomposition modale empirique dont les composantes sont des solutions d'équations fonctionnelles simples. Nous montrons, par la suite, la convergence et l'unicité de cette décomposition sous certaines conditions de régularité. Pour rendre cette technique exploitable dans la pratique, nous construisons un algorithme de type spline permettant d'approcher les composantes EIMF de notre décomposition modale empirique. L'avantage de notre méthode réside principalement dans le fait qu'elle ne fait pas appel à une procédure de tamisage et donc elle ne requiert pas un test d'arrêt ; ce point était problématique dans l'EMD classique. La décomposition proposée converge après un nombre déterminé d'itérations.
- Apprentissage automatique avec l'EMD et application en *cyber-risk* : Nous avons proposé, dans [5], une nouvelle solution d'apprentissage automatique pour l'ASV (*Automtic Speaker Verification*) basée sur l'extraction de caractéristiques fréquentielles des signaux vocaux. Notre proposition combine à la fois les techniques spectrales classiques et la décomposition EMD décrite plus haut. L'idée est d'utiliser les IMFs ainsi que leurs fréquences instantanées comme des signature permettant de distinguer une voix humaine authentique d'une voix artificielle répliquée.

5.1 Extraction des composantes fréquentielles liées au stress

La miniaturisation électronique permet de produire des dispositifs de mesures physiologiques de plus en plus précis. Le développement d'algorithmes plus fiables permettent à ces appareils d'être manipulés sans l'aide d'un professionnel de santé. La mesure la plus populaire est la fréquence cardiaque qui est affichée par les montres ou intégrée dans certains smartphones. Ces données, une fois analysées, contiennent une mine d'informations sur l'état de santé d'un patient. Il y a encore quelques années, l'analyse de ces signaux était généralement faite grâce à l'oeil expérimenté des cardiologues. Ces derniers s'intéressent à l'étude des signaux cardiaques dans deux bandes de fréquences : les bandes orthosympathiques et parasympathiques, c'est-à-dire les

bandes de fréquences (0.04 Hz, 0.15Hz)et (0.15 Hz, 0.5Hz) respectivement. La définition de ces bandes et leurs liens avec le système nerveux central ont été établis dans [135]. Elles sont basées sur le fait que l'énergie spectrale de ces bandes est un indicateur pertinent du niveau de stress d'un individu. En effet, le couple parasympathique/orthosympathique est souvent comparé à un couple de freinage/accélération ; voir par exemple [116]. Au repos, il y a un effet de freinage permanent sur la fréquence cardiaque alors que toute sollicitation du système cardio-vasculaire produit initialement une réduction de frein parasympathique suivie d'une implication progressive du système sympathique. Ces mécanismes sont très intéressants à surveiller dans de nombreuses maladies (insuffisance cardiaque, troubles du rythme...). Dans le domaine de la physiologie, ces données sont cruciales pour mesurer le niveau de vigilance ; en particulier le risque de s'endormir en conduisant, le niveau de stress induit par l'activité physique ou le niveau de stress perçu. Les modèles fractals ont été utilisés en cardiologie après les travaux de [123], qui ont appliqué l'analyse du spectre multifractal pour modéliser les séries RR et classer les individus en fonction de ce spectre. Nous utilisons des ondelettes pour extraire les composantes d'énergie associées aux hautes et basses fréquences. Nous utilisons les profils d'énergie ainsi extraits pour les relier à l'état d'un patient en utilisant des techniques de classification ; plus de détails techniques se trouvent dans notre papier [30].

Modélisation aléatoire de la fréquence cardiaque : Compte-tenu de la grande variabilité des tâches aléatoires et des situations auxquelles un individu peut être confronté, les battements du coeur peuvent être modélisés par un processus aléatoire non stationnaire. Pour des raisons biologiques évidentes, nous supposons que la fréquence cardiaque est comprise entre 20 et 250 battements par minute. Dans ce travail, nous nous intéressons plutôt aux intervalles RR, c'est-à-dire aux durées entre deux pics de type R, voir figure 5.1.

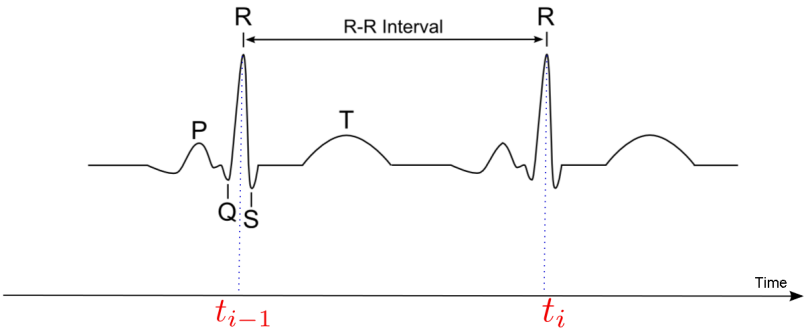


Figure 5.1 – Illustration des mesures du signal RR

Nous considérons alors le processus $X_i = t_i - t_{i-1}$ que nous supposons compris entre $\frac{60}{250}s < X_t < \frac{60}{20}s$, les t_i ce sont les instants correspondants au pic R voir figure 5.1 . Il est naturel de considérer que l'activité d'un individu est constituée d'une alternance de repos et d'excitations plus ou moins intenses. Il est donc tout à fait raisonnable de considérer que le processus X_t est stationnaire par morceaux, que ce soit dans le domaine temporel ou fréquentiel voir par exemple [118]. En d'autres termes, pour tout $t \in \mathbb{R}$, le processus X_t peut être représenté de la manière suivante :

$$X(t) = \mu(t) + \int_{\mathbb{R}} e^{it\xi}\sqrt{f(t,\,\xi)}dW(\xi). \tag{5.1}$$

Les composantes de l'équation (5.1) sont :

- L'application $\xi \mapsto f(t,\ \xi)$ est une fonction paire, positive et constante par morceaux, c'est-à-dire qu'il existe une partition $\tau_1,\ \ldots,\ \tau_K$ telle que $f(t, \xi) = f_k(\xi)$ pour $t \in [\tau_i, \tau_{i+1}[$.
- La fonction $t \mapsto \mu(t)$ est également constante par morceaux pour éventuellement une autre partition $\tilde{\tau}_1,\ \ldots,\ \tilde{\tau}_L$ avec $\mu(t) = \mu_\ell$ pour tout $t \in [\tilde{\tau}_\ell, \tilde{\tau}_{\ell+1}[$.
- $dW(\xi)$ est une mesure de Wiener bien choisie de façon à ce que le processus X_t soit réel, [37] ou [66].

Selon les recommandations du Groupe de travail de *l'European Soc. Cardiology et du North American Soc. of Pacing and Electrophysiology*, nous utilisons les notations suivantes :

- On notera la bande de fréquence orthosympathique par, $LF = [\omega_1,\ \omega_2] = (0.04Hz,\ 0.15Hz)$;
- La bande $HF = [\omega_2,\ \omega_3] = (0.15Hz,\ 0.5Hz)$ désigne la bande parasympathique.

Extraction des énergies HF et LF : Afin d'extraire efficacement les énergies correspondantes aux bandes de fréquences HF et LF on utilise des techniques inspirées par [66, 61]. Nous donnons ici une brève description de cette idée ; à l'aide d'un choix approprié d'une base d'ondelettes, on extrait les énergies associées aux bandes LF et HF et localisées autour d'un instant b . Cela revient à calculer les modules des coefficients d'ondelettes $|W_i(b)|^2$ pour $i = 1, 2$, avec la formule :

$$W_i(b) = \int_{\mathbb{R}} \psi_i(t-b)X(t)dt, \tag{5.2}$$

les ψ_1 et ψ_2 sont deux ondelettes dont les supports, dans le domaine fréquentiel, sont respectivement les bandes LF et HF.

Choix des ondelettes ψ_1 et ψ_2 : Dans l'idéal, on utiliserait deux filtres ψ_1 et ψ_2 à support compact, dont les transformées de Fourier auraient un support inclus dans les bandes HF et LF; évidemment ceci est impossible voir par exemple [125]. Par conséquent, le mieux que l'on puisse faire est de choisir entre un filtre à support compact exclusivement dans le domaine fréquentiel ou temporel. Le prix à payer pour la compacité du support dans un domaine est la perte de la compacité du support dans l'autre. Dans la pratique, la plupart des filtres utilisés décroissent rapidement vers 0 au voisinage de l'infini. Pour évaluer l'effet de la perte de compacité, [37] ont introduit la notion de ρ-pseudo compacité du support.

Définition 5.1.1 : Un filtre ψ possède un ρ-pseudo support compact noté I si et seulement si :

$$\frac{\int_I |\psi(t)|^2 dt}{\int_{\mathbb{R}} |\psi(t)|^2 dt} = \rho$$

Cela signifie que la valeur de $1 - \rho$ mesure la perte d'énergie si nous forçons le support à être compact (on tronque la fonction à un support compact). Pour extraire l'énergie correspondante à une bande de fréquences données, il suffit de choisir les ψ_i dans (5.2) ayant un ρ-pseudo support inclus dans cette bande et de prendre une valeur de ρ proche de 1. La proposition suivante donne une méthode générique permettant de trouver de tels supports voir [37].

Proposition 5.1.1 : Soit ψ un filtre à support compact dans le domaine temporel $[L_1,\ L_2]$ et ayant un ρ-pseudo support $[_1,\ _2]$ dans le domaine fréquentiel. Considérons l'application $\psi_1(t) = \mu \times e^{i\eta t}\psi(\lambda t)$ et notons $[\omega_1,\ \omega_2]$ une bande de fréquences arbitraire. Si on choisit λ et η telles que :

$$\lambda = \frac{\omega_2 - \omega_1}{\Lambda_2 - \Lambda_1} \quad , \quad \eta = \frac{\omega_1 + \omega_2}{2} - (\omega_2 - \omega_1)\frac{\Lambda_2 + \Lambda_1}{\Lambda_2 - \Lambda_1}$$

Dans ce cas la fonction ψ_1 a un ρ-pseudo support dans le domaine temporel donné par :

$$\left[\frac{2-1}{\omega_2 - \omega_1}L_1, \frac{2-1}{\omega_2 - \omega_1}L_2\right]$$

Démonstration. Il suffit de remarquer que $\hat{\psi}_1(\xi) = \mu \times \hat{\psi}(\frac{\xi - \eta}{\lambda})$, et déduire que le ρ-pseudo support de ψ_1 égale à $\eta + \lambda$ ρ-pseudo support de ψ, voir par exemple [37, 30]. □

Cette dernière proposition permet de définir différents choix possibles des filtres ψ_1 et ψ_2. Pour des raisons de simplicité de calcul, nous avons utilisé dans [30] une ondelette mère de Gabor définie par

$$\psi(t) = e^{i\eta t}g(t) \qquad \text{où} \qquad g(t) = \frac{1}{(\sigma^2\pi)^{1/4}}e^{-\frac{t^2}{2\sigma^2}} \tag{5.3}$$

Pour plus de détails, voir par exemple [125]. L'avantage de cette ondelette est qu'elle a le même ρ-pseudo support que ce soit dans le domaine temporel ou fréquentiel ; par exemple pour $[-L,\ L] = [-3.5, 3.5]$ on trouve $\rho = 0.9995$. En passant, par transformation de Fourier, au domaine fréquentiel, nous avons,

$$\hat{\psi}(t) = \hat{g}(\xi - \eta),\ \hat{g}(\xi) = (4\pi\sigma^2)^{1/4}e^{\frac{-\sigma^2\xi^2}{2}} \tag{5.4}$$

En utilisant la proposition 5.1.1, nous adaptons l'ondelette de Gabor pour qu'elle ait un ρ-pseudo support à l'intérieur des bandes de fréquence LF et HF. On obtient donc les deux ondelettes de Gabor définies par (5.3) avec le choix suivant pour les paramètres :

$$\eta_1 = \frac{\omega_1 + \omega_2}{2} \qquad \text{et} \qquad \sigma_1 = \frac{2L}{\omega_2 - \omega_1} \tag{5.5}$$

$$\eta_2 = \frac{\omega_2 + \omega_3}{2} \qquad \text{et} \qquad \sigma_2 = \frac{2L}{\omega_3 - \omega_2} \tag{5.6}$$

De plus $|\rho$ **pseudo Supp** $\psi_1| = \dfrac{4L^2}{\omega_2 - \omega_1}$ et $|\rho$ **pseudo Supp** $\psi_2| = \dfrac{4L^2}{\omega_3 - \omega_2}$. Pour $L = 3.5$ on aura $\rho = 0.9995$ ce qui se traduit par une perte d'énergie de 5.10^{-4}. L'utilisation de l'ondelette Gabor est également plus efficace en terme de temps de calcul.

Construction des *features* pour la classification : Pour l'extraction des composantes ou *features* qui serviront à la classification, nous utilisons des techniques de détection de ruptures. Ces dernières ont été largement étudiées dans la littérature ; nous citons entre autres [151]. Les méthodes classiques sont généralement basées sur des techniques de moindres carrés pénalisés qui souffrent d'une complexité de calcul, ce qui les rend difficiles à mettre en place pour une application en temps réel. Pour surmonter ce problème, les auteurs dans [38] ont introduit une technique plus rapide pouvant être implémentée en ligne et en temps *quasi*-réel. L'idée principale de cette méthode est basée sur deux étapes ; la première consiste à détecter les changements

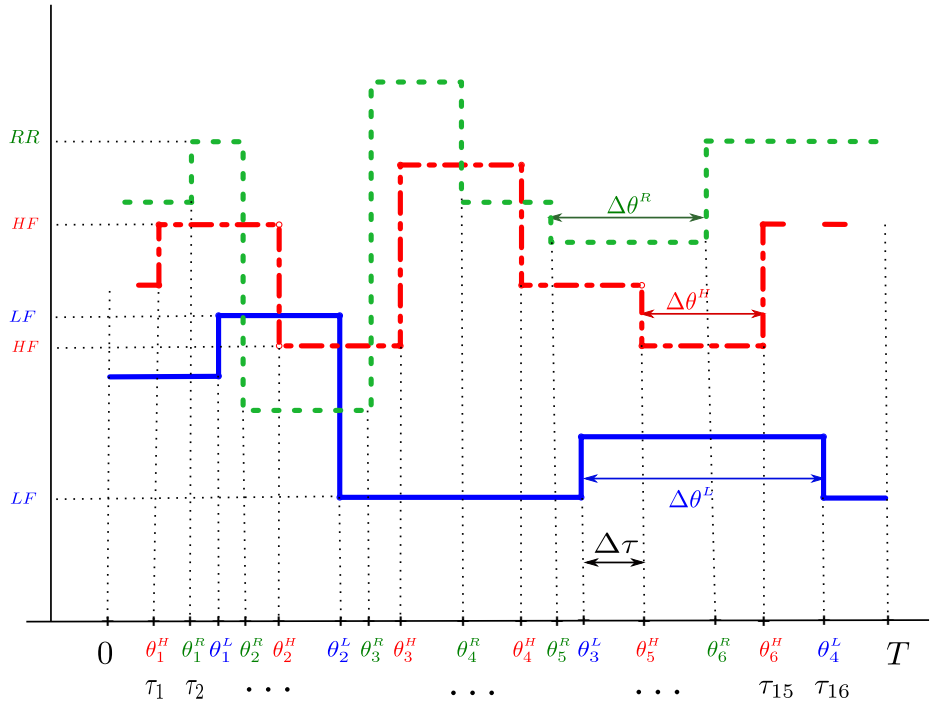


FIGURE 5.2 – Illustration sur la façon dont nous construisons les variables $\Delta\tau, \Delta\theta$ et $\Delta\delta$

sans se soucier des fausses alarmes, la seconde consiste à effectuer un test statistique, avec une p-value optimisée, pour garder ou éliminer ces fausses alarmes. Cette méthode a été appliquée sur les signaux bruts RR, les énergies HF et LF. Ces dernières ont été calculées à partir de la formule (5.2) où ψ_1 et ψ_2 sont choisies avec les paramètres donnés dans (5.5) et (5.6). Pour illustrer cela, notons $\mathcal{T} = \tau_1 < \tau_2 < \ldots < \tau_n$ (resp. $\Theta = \theta_1 < \theta_2 < \ldots < \theta_p$), les instants de ruptures ou de changements du niveau d'énergie HF (resp. énergie LF). Nous rassemblons ces deux séquences dans la famille $\mathcal{T} \cup \Theta$ et nous réarrangeons ces éléments en les notant, $\delta_1 < \delta_2 < \ldots < \delta_M$. Une illustration est donnée dans la figure 5.2.
À partir des instants de rupture nous construisons des variables décrivant l'état du système nerveux central :

- la variable $\Delta\tau_i = \tau_i - \tau_{i-1}$ représente la durée d'un niveau de l'énergie HF. Cette durée peut être considérée comme la durée où seul le système parasympathique (freinage) est activé et qu'il a atteint un régime fixe HF_i pendant cette durée.

- la variable $\Delta\theta_j = \theta_j - \theta_{j-1}$ représente la durée d'un niveau de l'énergie LF. Cette durée peut être considérée comme la période où seul le système sympathique (accélération) est en action et a établi un régime fixe HF_j.

- la variable $\Delta\delta = \delta_i - \delta_{i-1}$ donne la durée du changement inter énergies HF et LF; elle représente les périodes de passage d'un système à l'autre.

Pour un individu donné, nous construisons le tableau suivant qui décrit les différents états de son activité cardiaque. Dans ce cas, pour chacun des M états, en plus des variables définies plus haut, nous ajoutons également les moyennes et les variances des niveaux d'énergies HF, LF, et le RR.

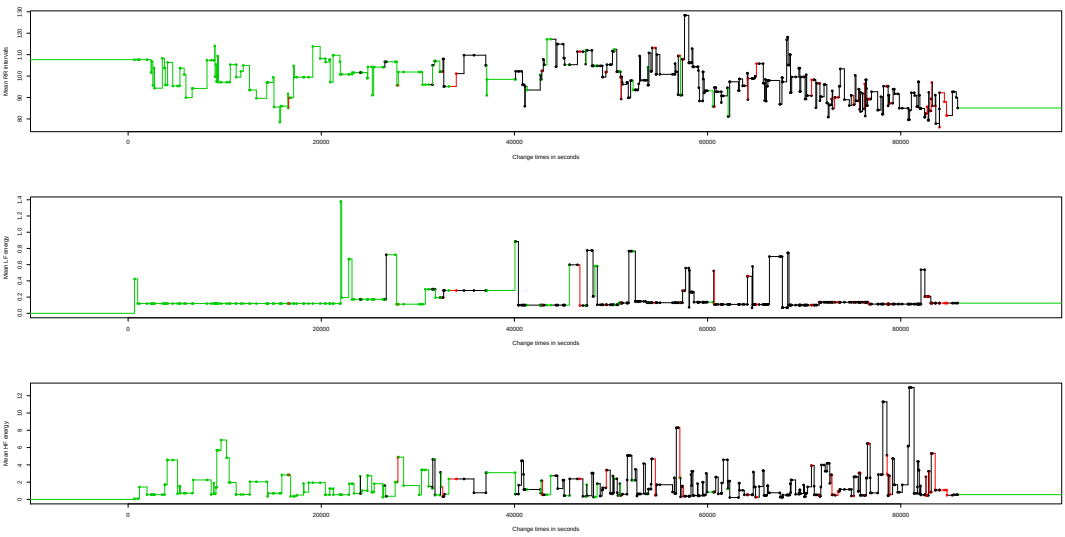


FIGURE 5.3 – Exemple de classification sur un médecin urgentiste pendant sa période de garde

Etats	$\Delta\tau$	$\Delta\theta$	$\Delta\delta$	HF	σ_H^2	LF	σ_L^2
1	$\tau_1 - \tau_0$	$\theta_i - \theta_{i-1}$	$\delta_i - \delta_{i-1}$	HF_1	σ_H^2	LF_1	σ_L^2
	☒	☒	☒	☒	☒	☒	☒
M-1	$\tau_M - \tau_{M-1}$	$\theta_i - \theta_{i-1}$	$\delta_i - \delta_{i-1}$	HF_{M-1}	σ_H^2	LF_{M-1}	σ_L^2
M	$T - \tau_M$	$\theta_i - \theta_{i-1}$	$\delta_i - \delta_{i-1}$	HF_M	σ_H^2	LF_M	σ_L^2

Une méthode de classification non supervisée, appliquée sur les composantes du tableau ci-dessus, a permis de construire des variables discriminantes qui reflètent la façon dont la fréquence cardiaque est modulée par l'activité d'un individu. Un exemple d'application a été testé sur les enregistrements cardiaques d'une cohorte de médecins urgentistes pendant leurs périodes de garde. La figure 5.3 donne une illustration du pouvoir discriminant des *features* que nous avons construit. Des discussions et des commentaires plus détaillées sont présentés dans [30].

5.2 Décomposition modale empirique et applications

Pour mieux comprendre le principe et les motivations de la décomposition EMD, nous rappelons la transformation de Hilbert et la notion de fréquence instantanée. Pour un signal réel $x(t)$ sa transformée de Hilbert est définie par :

$$y(t) = \frac{1}{\pi} p.v. \int_{-\infty}^{+\infty} \frac{x(\tau)}{t-\tau} d\tau$$

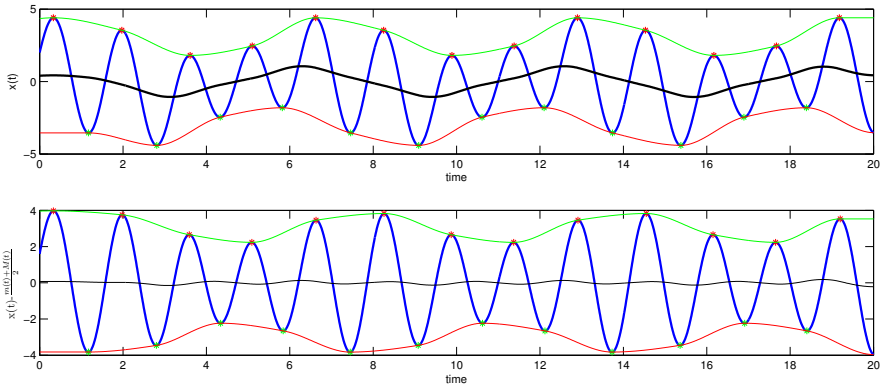


FIGURE 5.4 – Illustration de la procédure de tamisage

où $p.v.$ désigne la valeur principale de Cauchy. Le signal analytique correspondant à $x(t)$ est ainsi défini par : $z(t) = x(t) + \jmath y(t)$ qui peut être également exprimé comme $z(t) = a(t)\exp\{\jmath\theta(t)\}$, où $a(t) = \sqrt{x^2(t) + y^2(t)}$ est l'amplitude de $z(t)$ et $\theta(t) = \arctan\frac{y(t)}{x(t)}$ sa phase instantanée. La fréquence instantanée $f(t)$ peut être obtenue par différentiation :

$$f(t) = \frac{1}{2\pi}\frac{\partial\theta(t)}{\partial t}.$$

L'interprétation physique de la notion de fréquence instantanée est bien acceptée quand le signal $z(t)$ est presque circulaire dans le domaine complexe. Elle mesure dans ce cas, la vitesse de rotation du signal complexe. Par contre, quand le signal $z(t)$ ne vérifie pas cette propriété, la fréquence instantanée perd sa signification pratique. Dans ce cas, pour redéfinir la fréquence instantanée, la technique de l'EMD consiste à décomposer le signal $x(t)$ en composantes circulaires (modes propres ou (IMF)) dont la fréquence instantanée peut être vue comme une vitesse de rotation. En fait, les IMFs sont des fonctions localement symétriques ; ce qui fait que leurs fréquences instantanées ont une interprétation physique.
Les composantes IMF sont obtenues par l'algorithme de *tamisage* qui consiste en une extraction itérative des composantes oscillant plus ou moins rapidement [129]. La décomposition modale empirique d'un signal $x(t)$ est basée sur le processus de tamisage (dit aussi de criblage) qui est décrit par les étapes suivantes :

(1) Identifier tous les extrêma locaux de $x(t)$.

(2) Calculer l'enveloppe supérieure $M(t)$ en reliant tous les maxima locaux par une spline cubique. Répéter la procédure pour les minima locaux afin de produire l'enveloppe inférieure $m(t)$. On s'assure que les enveloppes supérieures et inférieures couvrent tout le signal.

(3) Actualiser les données en calculant la différence entre les données et la moyenne des enveloppes supérieures et inférieures $x(t) \leftarrow x(t) - \frac{M(t)+m(t)}{2}$..

(4) Répéter les étapes (1), (2) et (3) jusqu'à obtenir une IMF, $c(t)$.

(5) Mettre à jour les données initiales en soustrayant l'IMF obtenue en 4, $x(t) \leftarrow x(t) - c(t)$.

(6) Répéter (1)-(5) jusqu'à l'obtention d'une tendance $r(t)$; une courbe avec au plus un extremum.

Une illustration de la procédure du tamisage est donnée dans la figure 5.4. Ainsi, après avoir réalisé la décomposition EMD, la transformation de Hilbert de chaque IMF est calculée. Dans ce

cas, le signal original $x(t)$ peut être exprimé en une décomposition de type Fourier comme :

$$x(t) = Re\left\{\sum_{k=1}^{K+1} a_k(t)\exp\{\jmath \int 2\pi f_k(t)dt\}\right\},$$

Cette décomposition est proposée dans [129], elle est connue sous le nom de transformation de Hilbert- Huang (HHT). Pour plus de détails concernant ces approches nous citons entre autres, [129], [99, 101, 83, 73, 86, 94, 69], [64], [65].

5.2.1 EMD unidimensionnelle

L'un des problèmes de la technique de tamisage est qu'après l'opération de moyennage des enveloppes, la fonction résultante peut ne pas être une IMF. Ce comportement indésirable est atténué par la répétition de la procédure de *tamisage* mais cette solution a de nombreux inconvénients : d'une part, nous ne disposons d'aucun critère théorique pour la procédure d'arrêt. D'autre part, nous n'avons aucune garantie que cette procédure converge vers la décomposition souhaitée. Bien que la décomposition EMD soit largement utilisée dans la pratique, elle souffre de certains inconvénients liés à son caractère empirique. En effet, cette décomposition est définie à partir d'un algorithme itératif, adaptatif et intuitif. Elle n'est pas basée sur une formulation mathématique rigoureuse ; on ignore de ce fait toutes les propriétés concernant la convergence de l'algorithme. Notre idée est d'introduire une nouvelle formulation d'une IMF élémentaire basée sur le changement de convexité du signal plutôt que de calculer les enveloppes moyennes à partir des seuls extrema.

Définition 5.2.1 : Une fonction $x(t)$ est une IMF élémentaire si et seulement si elle vérifie :

- La fonction $x(t)$ a au moins trois points d'inflexion et que ces points sont aussi les points de passages par zéro. [1].
- Pour trois inflexions consécutives, la fonction est symétrique par rapport au point d'inflexion central.
- Une tendance est toute fonction qui a au plus un changement de convexité.

Il faut noter que les IMF élémentaires sont des IMF au sens classique de Huang ou l'algorithme du tamisage, ce qui n'est pas en contradiction avec la théorie classique. Afin de décomposer le signal en EIMF, la procédure consiste à identifier les points pour lesquels la convexité change ; ces derniers donnent une idée de la nature oscillatoire du signal étudié.

Décomposition dans le cas univarié : Nous commençons par étudier la décomposition pour une oscillation locale simple où un seul changement des convexités est observé (cas de trois inflexions consécutives). Nous montrons ensuite que la décomposition est unique si on impose que le reste de la décomposition est une tendance. Enfin, nous traitons la construction des décompositions pour un signal général en utilisant uniquement les points de changement de convexité. Le résultat suivant montre l'existence et la structure théorique d'une EIMF sur un intervalle local $[a, b]$.

1. Une EIMF est une fonction qui commence toujours par zéro ou qui est un décalage temporel d'une telle fonction ; par exemple le sinus et le cosinus sont des EIMF (le cosinus est le sinus décalé de $\pi/2$). De plus, une fonction sera considérée comme une EIMF si l'intervalle de temps des observations est suffisamment grand pour voir que cette fonction est oscillante (par exemple le cosinus sera considéré comme une tendance sur l'intervalle $[0, 2\pi]$).

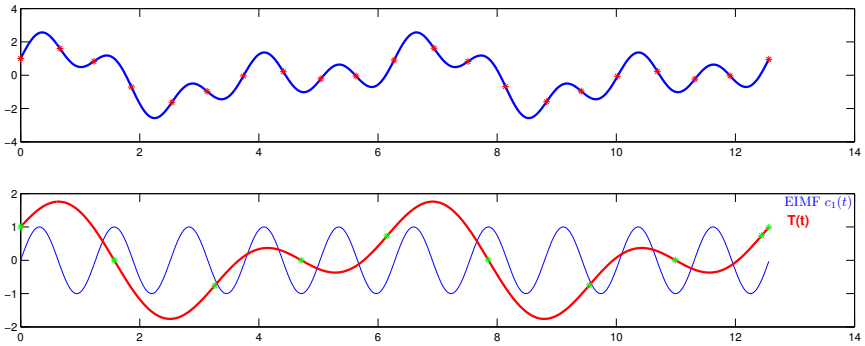


FIGURE 5.5 – Illustration de l'extraction de la première EIMF $c(t)$.

Lemme 5.2.1 : Soit $x(t)$ une fonction assez régulière dont la courbe a un changement de convexité sur l'intervalle $[a,b]$.

(1) Le signal $x(t)$ admet la décomposition $x(t) = c(t) + T(t)$ où $c(t)$ est une EIMF sur $[a,b]$. Dans ce cas, il existe une fonction impaire φ telle que :

$$c(t) = \frac{x(t) - x(a+b-t)}{2} - \varphi(2t - (a+b)),$$

où φ vérifie $\varphi(2t-(a+b)) = \frac{T(t)-T(a+b-t)}{2}$.

(2) Le couple (φ, T) de la décomposition donnée en (1) est unique dans le sens où la fonction impaire φ est la seule telle que T ne change pas le type de sa convexité sur l'intervalle $[a,b]$.

Cette propriété d'unicité prouvée dans le dernier lemme joue un rôle fondamental car la fonction résultante T ne change pas de convexité sur les intervalles locaux. Ceci implique que nous réduisons les points caractéristiques (points d'inflexion dans ce cas) qui servent à l'extraction des EIMF. Ainsi, si la technique du lemme 5.2.1 est appliquée à des intervalles locaux successifs, cette propriété de convexité inchangée forcera la convergence de la procédure d'extraction des EIMF car cela réduit la rapidité des oscillations de la tendance résultante T. La proposition suivante détaille ce résultat :

Proposition 5.2.2 : Soit $x(t)$ une fonction régulière réelle définie sur un intervalle $[A,B]$ avec $m = 2\ell - 1$ points d'inflexion $I_1(a_1, x(a_1)), \dots, I_m(a_m, x(a_m))$. Nous désignons par $I_0(A, x(A))$ et $I_{m+1}(B, x(B))$, le premier et le dernier point du signal. Il existe un famille unique de fonctions impaires $(\varphi_1, \dots, \varphi_\ell)$ telles que la fonction $x(t)$ a la décomposition suivante :

$$x(t) = c(t) + T(t)$$

où c est une EIMF donnée par : $c(t) = \sum_{i=1}^{\ell} \frac{x(t) - x(a_{2i-1} + a_{2i+1} - t)}{2} - \varphi_i(2t - (a_{2i-1} + a_{2i+1}))$, φ_i sont des fonctions impaires et T est une fonction ayant au plus ℓ points d'inflexion.

Une application de la procédure d'extraction décrite dans la proposition 5.2.2 est illustrée dans la figure 5.5. On voit bien que la tendance $(T(.)$ ainsi extraite possède un nombre de changement de convexité plus petit que celui de $c(.)$.

Comme dans l'algorithme de tamisage, la dernière proposition suggère un moyen de décomposer le signal en une somme de EIMF. En effet, cette décomposition est obtenue en soustrayant la

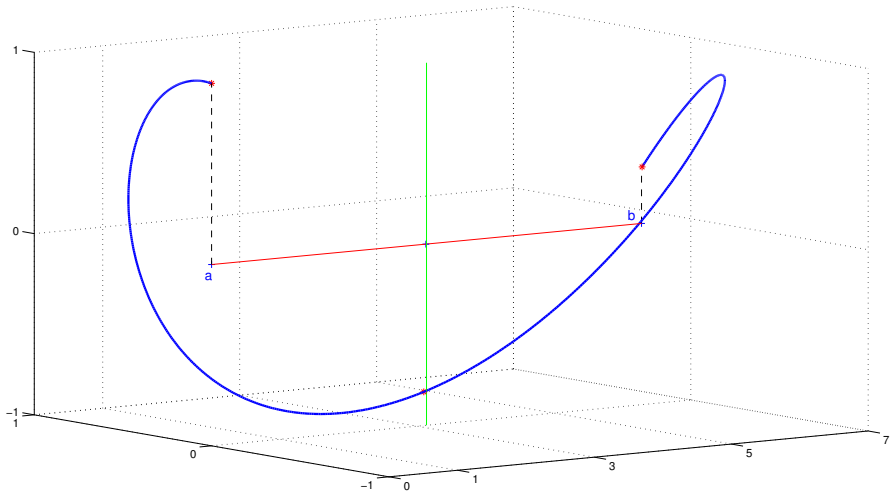


FIGURE 5.6 – Illustration d'une oscillation locale sur l'intervalle $[a, b]$

fonction $c(t)$ et en répétant la procédure d'extraction sur la fonction résultante T jusqu'à obtenir une tendance finale. Une autre propriété importante est que, pendant le processus d'extraction, à chaque itération, nous réduisons les points d'inflexion d'au moins la moitié. Cette réduction des points caractéristiques, assure la convergence de la décomposition.

5.2.2 Décomposition des signaux vectoriels

Dans cette partie, nous nous sommes intéressés à la décomposition modale empirique (EMD) des fonctions vectorielles bivariées. Comme nous l'avons mentionné dans le cas unidimensionnel, nous utilisons la notion de changement de convexité du signal au lieu de la variation au niveau de ses extrema. Ces derniers n'ont pas de signification naturelle pour des fonctions vectorielles complexes. Le signal est ainsi décomposé en trois catégories de composantes :

(1) Les IMF rotatives ayant des courbes non planaires mais qui changent de convexité plus ou moins rapidement.

(2) Les IMF oscillantes ayant des courbes planaires qui changent de convexité. Ces composantes correspondent aux EIMF du cas unidimensionnel.

(3) Les tendances qui sont des courbes qui ne changent pas de convexité.

En utilisant le caractère itératif et adaptatif de l'EMD classique, nous introduisons une nouvelle technique qui permet d'extraire récursivement les composantes du signal. La nouveauté de notre technique est qu'elle ne passe pas par la procédure de tamisage et par conséquent elle permet d'éviter ses inconvénients. Ces résultats présentent une alternative pour remédier aux problèmes évoqués par [73] et [75]. Pour donner une idée de notre technique, considérons une oscillation locale d'une fonction $\vec{f}(t) = (x(t), y(t))$ sur $[a, b]$, voir Fig. 5.6.

L'idée est de chercher une fonction $\vec{T}(t) = (T_1(t), T_2(t))$ telle que $\vec{f}(t) - \vec{T}(t) = \vec{c}(t)$ soit une IMF ; c'est-à-dire symétrique par rapport à la droite passante par $(\frac{a+b}{2}, 0, 0)$ et parallèle au vecteur de

référence $\vec{k}$. Dans ce cas nous montrons que :

$$c_1(t) = \frac{x(t) - x(a+b-t)}{2} - \varphi(2t-(a+b)), \tag{5.7}$$

$$c_2(t) = \frac{y(t) + y(a+b-t)}{2} - \psi(2t-(a+b)). \tag{5.8}$$

où φ et ψ sont respectivement impaires et paires. Nous montrons également que cette décomposition est unique si on impose à $\vec{T}$ de tourner ou osciller moins rapidement que $\vec{c}$ sur l'intervalle $[a,b]$. Cette unicité et le fait que la fonction $\vec{T}$ oscille moins rapidement que l'IMF extraite $\vec{c}$ assurent la convergence de notre technique après un nombre fini d'itérations. Cela suggère une décomposition en ligne et de façon itérative de tout signal régulier. Nous montrons le résultat général suivant :

Théorème 5.2.3 : Soit $\vec{f}(t) = (x(t), y(t))$ une fonction régulière ayant $m = 2\ell - 1$ points d'inflexion $a_1, \dots, a_m$. Il existe une famille de fonctions impaires $(\varphi_1, \dots, \varphi_\ell)$ et de fonctions $(\psi_1, \dots, \psi_\ell)$ paires telles que $\vec{f}(t) = \vec{c}(t) + \vec{T}(t)$ où

$$c_1(t) = \sum_{i=1}^{\ell} \left(\frac{x(t) - x(a_{2i-1} + a_{2i+1} - t)}{2} - \varphi_i(t - \mathbf{m}_i)\right) \mathbb{1}_{[a_{2i-1}, a_{2i+1}]}$$

$$c_2(t) = \sum_{i=1}^{\ell} \left(\frac{y(t) + y(a_{2i-1} + a_{2i+1} - t)}{2} - \psi_i(t - \mathbf{m}_i)\right) \mathbb{1}_{[a_{2i-1}, a_{2i+1}]}$$

La fonction $\vec{T}$ possède au plus ℓ points d'inflexion et $\mathbf{m}_i = \frac{a_{2i-1}+a_{2i+1}}{2}$.

Notre décomposition consiste donc à refaire la même procédure d'extraction sur la fonction résultante $\vec{T}$ jusqu'à ce que on élimine tous les points d'inflexion et par conséquent au final obtenir une tendance. Dans la pratique les fonctions φ_i et ψ_i ne sont pas connues et en même temps les signaux $x(t)$ et $y(t)$ ne sont observés que sur des instants discrets $t_1, \dots, t_N$. L'idée est de donner une approximation de type splines des fonctions φ_i et ψ_i sur l'intervalle $[a_{2i-1}, a_{2i+1}]$, de la manière suivante :

$$\varphi_i(t - \mathbf{m}_i) = \alpha_{i,1}(t - \mathbf{m}_i) + \beta_{i,1}(t - \mathbf{m}_i)^3 + \gamma_{i,1}(t - \mathbf{m}_i)^5$$
$$\psi_i(t - \mathbf{m}_i) = \alpha_{i,2} + \beta_{i,2}(t - \mathbf{m}_i)^2 + \gamma_{i,2}(t - \mathbf{m}_i)^4$$

Les coefficients α, β et γ sont déterminés, comme dans le cas des splines, en utilisant la continuité et la continuité des dérivées du premier et du second ordre des IMF extraites ainsi que les conditions aux bords. Un exemple de décomposition est donné dans les figures (5.7) et (5.8).

5.3 Utilisation de L'EMD pour vérifier l'authenticité de la voix

Les systèmes d'authentification biométrique sont devenus un standard dans les différents protocoles d'accès aux données. Les technologies ASV (*Automatic Speaker Verification*) sont de plus en plus utilisées dans les centres d'appels, les interfaces *homme-machine*, les contrôles d'accès sécurisé pour les services bancaires et commerciaux, voir [23, 20]. Cela a entraîné le développement de nouvelles stratégies basées sur l'apprentissage automatique pour faire face à d'éventuelles attaques par voix de synthèse comme les *deep fake* [11, 9]. En analyse des signaux vocaux, les fréquences des *formants* acoustiques sont considérés comme une signature caractéristique

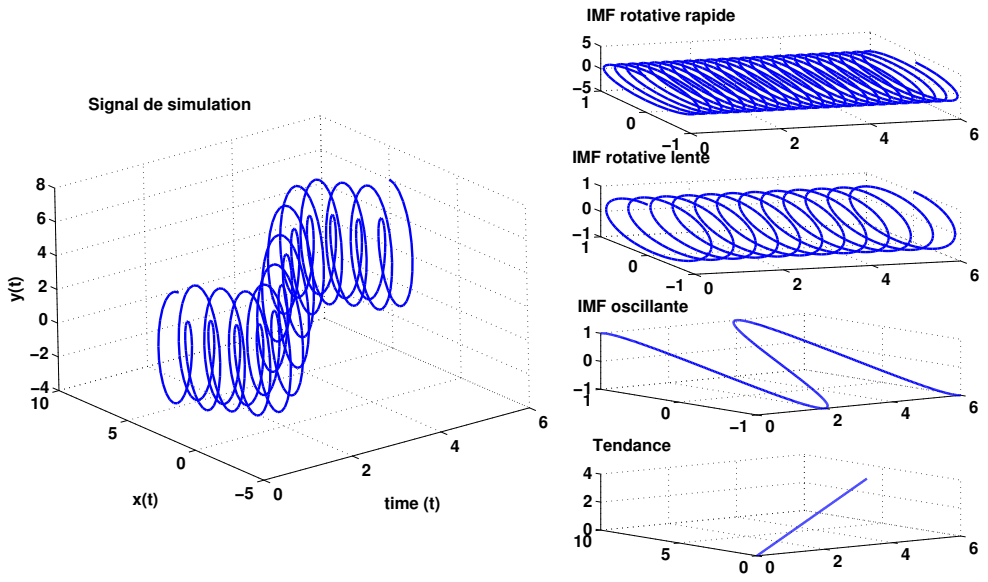


FIGURE 5.7 – A gauche le signal de simulation et à droite ses vraies composantes

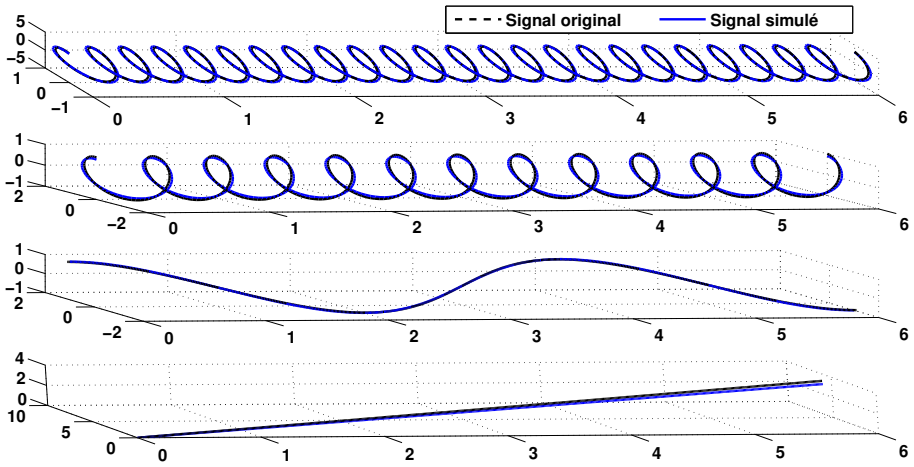


FIGURE 5.8 – Les IMFs extraites sont similaires aux composantes originales du signal 5.7.

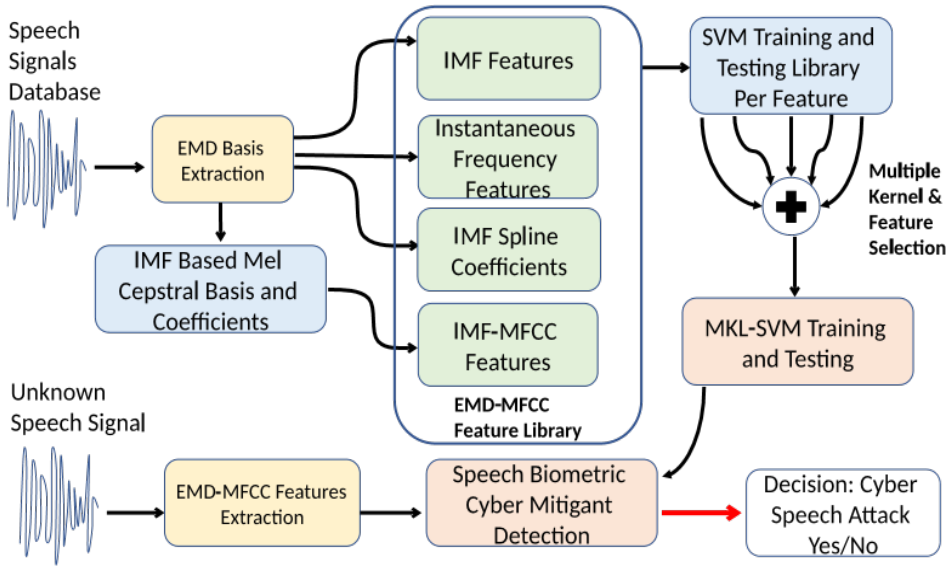


FIGURE 5.9 – Architecture de la méthodologie EMD-MFCC proposée pour la détection des formants.

des cordes vocales donnant une empreinte unique à un orateur [122, 28]. Les formants reflètent l'énergie acoustique de la parole se produisant généralement à chaque bande de fréquence de longueur 1000 Hz. On les note généralement F_1, F_2, F_3... la première F_0 est appelée la fréquence fondamentale et représente la vitesse de vibration des cordes vocales. L'extraction de ces caractéristiques à partir d'un signal vocal pourrait être utilisée pour distinguer une voix humaine d'une voix de synthèse, car elles représentent la façon dont les conduits vocaux créent les sources sonores uniques à un individu. Les MFCC ou *Mel-Frequency Cepstral Coefficients* sont souvent utilisés pour mesurer la concentration d'énergie autour des fréquences des formants. Ces coefficients cepstraux sont calculés à partir de la transformée de Fourier inverse appliquée au logarithme du spectre de puissance. Ils sont habituellement utilisés en traitement de la parole comme des *features* discriminantes pour classifier ces signaux. Nous avons proposé, dans [5], une nouvelle solution d'apprentissage automatique pour l'ASV basée sur l'extraction des caractéristiques fréquentielles des signaux vocaux. Notre proposition combine à la fois les techniques spectrales classiques comme les MFCC et la décomposition modale empirique EMD décrite plus haut. L'idée est d'utiliser les IMFs ainsi que leurs fonctions de fréquences instantanées comme une signature permettant de distinguer une voix humaine authentique d'une voix artificielle répliquée. Plus particulièrement, nous utilisons les coefficients cepstraux sur les IMFs extraites à partir du signal analysé. Notre contribution contient de nombreux éléments clés ; Le fait d'adapter les algorithmes d'analyse temps-fréquence aux signaux non stationnaires a permis d'améliorer la robustesse des features qui ont servi à caractériser la signature ou empreinte vocale. Ensuite, ces nouvelles approches d'extraction de caractéristiques ont été utilisées pour développer un classifieur multi-noyaux basé sur des algorithmes SVM *Support Vector Machine*. L'architecture de la solution proposée est détaillée dans la figure 5.9.

Cette technique a été testée de manière extensive pour des signaux vocaux collectés dans différents contextes ; des voix humaines et synthétiques en différentes langues et contextes. Une discussion plus détaillée peut être trouvée dans notre article [5].

Conclusion et perspectives

Pour conclure ce document, je donnerai dans les lignes qui suivent quelques pistes de recherche que j'ai déjà entamées ou que je voudrais développer à moyen et à long terme. Je continuerai à concilier à la fois des recherches purement théoriques en mathématiques appliquées notamment en statistiques des processus et en analyse statistiques des données. Parallèlement à cela, je continuerai à développer des algorithmes et méthodes permettant la mise en application concrète de ces résultats théoriques. Je continuerai également à renforcer des collaborations en interaction avec le monde socioprofessionnel pour la recherche et développement ainsi que des activités de transfert de technologies. Je souhaiterai également continuer mes collaborations fructueuses avec la communauté de volcanologie et apporter le concours des statistiques à la modélisation et l'amélioration des méthodes de surveillance et prédictions des aléas volcaniques.

Processus α-stables et modélisation de phénomènes impulsifs : Il est connu que tout processus gaussien centré est complètement identifié à partir de sa fonction de covariance. Cependant, la caractérisation de la classe plus générale des processus α-stables ($S\alpha S$) ($1 < \alpha < 2$) nécessite la connaissance des mesures spectrales des distributions fini-dimensionnelles de tout ordre. Pour simplifier, un certain nombre de sous-classes ont été étudiées dans la littérature ; Une des plus importantes est obtenue via une intégrale stochastique :

$$X = \left(X_t = \int_T f_t(\lambda) d\xi(\lambda),\ t \in T \right) \tag{6.1}$$

Où $(f_t \in \Lambda_\alpha(\xi),\ t \in T)$ est une famille fonctions et $d\xi$ est une mesure aléatoire symétrique α-stable $S\alpha S$. Le défi majeur dans l'utilisation de la représentation intégrale stochastique (6.1) réside dans la caractérisation de ses distributions fini-dimensionnelles. Deux cas particuliers de la mesure

aléatoire $d\xi$ ont été largement étudiés dans la littérature : ce sont les mesures aléatoires sous-gaussiennes et le cas où $d\xi$ est à accroissements indépendants (dit aussi *indépendamment dispersés*). Ces deux dernières sous familles permettent de donner explicitement la loi du processus sous-jacent.

Plus généralement, la classe de processus satisfaisant une condition de séparabilité (condition $\mathcal{S}$), peut être représentée comme dans (6.1) par des mesures aléatoires à accroissements indépendants et en faisant varier le choix des fonctions $(f_t \in \Lambda_\alpha(\xi),\ t \in T)$. Cependant, dans les applications statistiques, ce sont souvent les fonctions (f_t) qui sont fixes et la mesure aléatoire qui est variée. Cela apparaît, par exemple, dans l'étude de processus harmonisables dont l'intégrale stochastique s'écrit sous la forme :

$$X = \left(X_t = \int_T e^{it\lambda} d\xi(\lambda),\ t \in T \right). \tag{6.2}$$

Dans ce contexte, l'exigence d'une mesure aléatoire indépendamment dispersée $d\xi$ restreint la structure de dépendance du processus X. En tant que tel, il est souhaitable d'introduire une dépendance entre les accroissements de la mesure aléatoire $d\xi$ pour enrichir cette classe de processus. Nous développons plutôt une nouvelle représentation des fonctions caractéristiques des distributions fini-dimensionnelles en terme d'un paramètre qui joue un rôle similaire à la fonction de covariance dans les processus gaussiens. Même lorsque la mesure aléatoire est fixe et que la famille des fonctions (f_t) est issue d'un espace fonctionnel $\Lambda_\alpha(\xi)$, une question naturelle est de savoir quelle est la structure $\Lambda_\alpha(\xi)$ de façon à ce que le processus (6.2) satisfasse la condition de séparabilité $\mathcal{S}$ et avoir ainsi une représentation (6.1).

Statistiques spatiales et applications : L'objectif est de développer de nouvelles techniques de modélisation statistique liées à des données de type Big-Data provenant de réseaux de capteurs spatiaux ou de l'internet des objets(IoT). Nous nous plaçons dans le cadre où chacun des capteurs ou source d'information présente un comportement probabiliste non linéaire provoquant une censure (coupure, détérioration, incertitude etc.) des données. L'enjeu majeur dans ce genre de situation est de mesurer le degré de confiance qu'un utilisateur accordera à une donnée notamment quand les décisions ou prédictions qui en découlent ont un impact critique (sécurité, surveillance, monitoring en transport ...). L'idée est de développer (ou d'adapter) des algorithmes permettant d'établir des règles de préférence ou une hiérarchie dans le choix des sources de façon à en extraire le maximum d'information.

- L'une des pistes que nous voulons explorer est la reconstruction spatio-temporelle à partir de capteurs localisés à des endroits stratégiques. Nous améliorons les techniques proposées en associant à chaque capteurs un poids reflétant le degré de confiance que nous devons lui accorder. Un premier résultat permettra d'obtenir d'une part une reconstruction spatiale optimale du phénomène étudié quantifiant ainsi la qualité de ces mesures ou données. Cet estimateur doit être réalisé sous deux configurations différentes : nœuds hétérogènes, nœuds homogènes.
- La deuxième piste que nous voulons explorer concerne le développement d'algorithmes permettant d'intégrer de multiples sources d'information issues de processus spatiaux connexes afin d'améliorer le déploiement des capteurs et hiérarchiser leurs importances dans le processus de collecte des données. Ce travail revêt un intérêt théorique dans la mesure où il nous permettra de mettre en place de nouvelles technique : à la fois des plans d'expérience,

les plans numériques ainsi que des méthodes d'interpolation spatiale que ce soit dans un contexte Gaussien ou impulsif (α-stable par exemple). Pour une utilisation dans le contexte du Big Data, les algorithmes doivent être optimisés de façon à avoir une faible complexité de calcul. Nous comparons leurs performances sous différentes configurations du système (par exemple, nombre de capteurs, taille, rapport signal à bruit, qualité de détection, etc.).

- Une application directe pourrait être mise en place dans le cadre de projet que nous avons en cours avec le LMV/IRD pour l'amélioration et la surveillance des aléas volcaniques en temps réel.
- Une autre application possible pourrait être mise en place dans le cadre de la thèse CIFRE, que nous venons d'obtenir, et qui vise à générer des données météorologiques à haute résolution géographique/spatiale et temporelle. Ces dernières doivent être cohérentes avec les scénarii futurs du changement climatique pour pouvoir développer un ensemble d'indicateurs agro-climatiques spécifiques aux cultures d'intérêt. L'évolution de ces indicateurs et leur sensibilité aux paramètres météorologiques permettront de quantifier l'impact du changement climatique sur les activités agricoles.

Annexes

Nourddine Azzaoui

Curriculum Vitae

Université Clermont Auvergne
LMBP UMR 6620 - CNRS
(+33) (0)609223238
(+33) (0)473407081
nourddine.azzaoui@uca.fr
math.univ-bpclermont.fr/∼azzaoui

Postes et fonctions occupés

2010–..... **Maître de conférences**,
Université Clermont Auvergne, Clermont Ferrand,
Equipe : Probabilités Analyse et Statistiques.

2008–2010 **Post-doctorat**,
Université De Technologie de Troyes,
Institut charles Delaunay, Laboratoire LM2S.

2006–2008 **Attaché Temporaire d'Enseignement et de Recherches**,
Premier contrat: Université de Bourgogne Dijon ,
Deuxième contrat: Université de Lille 1 , Laboratoire Paul Painlevé.

2005–2006 **Chargé d'études CNRS**,
Institut d'Electronique de Micro-électronique et de Nano-technologie, Lille, 6 mois.

2004–2005 **Chargé d'étude statistiques**,
ENESAD, actuellement AgroSup Dijon , 9 mois,
CESAER UMR 1041: Centre d'Economie et de Sociologie Rurales Appliquées à l'Agriculture et aux Espaces Ruraux.

Responsabilités administratives et nationales

2020–..... **Membre élu au conseil de l'UFR mathématiques** ,
Université Clermont Auvergne, UFR de Mathématiques.

. . . 2019 . . . **Jury de recrutement du Poste de MCF 1086**, *Président du comité de seléction*,
Poste à l'IUT de Clermont-Ferrand recherche et au LMBP .

2018–..... **Responsable du Master de Mathématiques Appliquées, Statistiques**, *Responsable du diplôme*,
UFR de Mathématiques.

2017–2020 **Membre élu au Conseil national des universités**,
Section CNU 26: Mathématiques Appliquées et Applications des Mathématiques.

2012–2017 **Membre élu au conseil du laboratoire de mathématiques UMR CNRS 6620** ,
Université Blaise Pascal, UFR sciences et Technologie.

2015–..... **Membre élu à la commission informatique**,
Laboratoire de Mathématiques Blaise Pascal UMR CNRS 6620,
Département de Mathématiques Informatiques.

Parcours académique et qualifications

Jan. 2010 **Qualification au fonction de Maître de conférences**,
section CNU 61: Génie informatique, automatique et traitement du signal et de l'image.

Jan. 2007 **Qualification au fonctions de Maître de conférences**,
section CNU 26*: Mathématiques Appliquées et Applications des Mathématiques*.

Dec. 2006 **Doctorat de Mathématiques Appliquées et Statistiques**,
Université de Bourgogne Dijon,
Jury: , Christian Robert(Président) Dominique Dehay, Paul Doukhan, Bernard Garel (Rapporteurs), Véronique Maume-Deschamps (Examinatrice), Rachid Sabre, Bernard Schmitt Encadrants..

Encadrement de la recherche

Thèses soutenues

2012–2015 **Thèse de Papa A. Faye**, *soutenue en dec. 2015, actuellement Ingénieur Altran*,
Co-encadrement (30%) avec P. Druilhet (40%) et AF-Yao (30%) ,
Thèse financée par la région Auvergne et BRF-FEDER.

Thèses en cours

2017–2022 **Thèse de Adbillahi Doualeh Ali**, *soutenance mai 2022*,
Co-encadrement (50%) avec A. Guillin (50%),
Thèse financée par l'Agence Universitaire de la Francophonie AUF.

2017–2022 **Thèse de Marta Campi**, *Thèse à University college of London (UCL)* Thèse soumise, defence 2022,
Co-encadrement (50%) avec G. W. Peters (50%).

2019–2022 **Thèse de Julien Hautot**, *Thèse à Université Clermont Auvergne en collaboration avec l'EUPI, soutenance prévue 2022*,
Co-encadrement (50%) avec Céline Teulière (50%).

Post-doctorat

2015–2016 **Malcolm Egan**, *Co-encadrement, il est actuellement chargé de recherche à l'INRIA*,
Financement Université Blaise Pascal, Financement de la région AURA, projet AUDACE.

2018–..... **Alessandro Tadini**, *Co-encadrement, en collaboration avec le LMV*,
Finacement I-site Clermont Cap 20-25.

Master de statistiques M2

2010–..... **Encadrement de plsieurs mémoires de Master M2**,
Master de Mathématiques appliquées, statistiques (MAS). ,
En moyenne deux par an.

Projets de recherches et Industriels

2018–..... **Projet EAUGURE 2**, *Responsable, projet financé par*,
la région Auvergne Rhône Alpes,
La banque publique d'Investissement BPI.

2017–..... **Projet Météo Marketing**, *Co-responsable*, Projet financé par,
Le fond européen FEDER.

2017–..... **Projet PEPS MéteoMarketing**, *Poste de 6 mois d'Ingénieur, financé par*,
Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société AMIES.

2016–..... **Projet TelluS INSU/INSMI**, *Membre*.
Modélisation statistique pour la surveillance des éruptions volcaniques

2012–2015 **Projet Do Well Be** , *Membre*,
Projet financé par l'Agence Nationale de a recherche ANR.

2008–2009 **Projet VIGIRES'EAU** , *Membre*,
Projet financé par l'Agence Nationale de a recherche ANR,
Concepts Systèmes et Outils pour la Sécurité Globale (CSOSG).

2007–2008 **Projet Mv-EMD**, *Membre*,
Projet financé par l'Agence Nationale de a recherche ANR.

Visites et séjours scientifiques

Juin. 2019 **The Institute of Statistical Mathematics**, *1 mois*,
Department of Statistical modelling, Tokyo Japan,.

Mar. 2018 **The Institute of Statistical Mathematics**, *10 jours*,
Department of Statistical modelling, Tokyo Japan,.

Juil. 2016 **The Institute of Statistical Mathematics**, *10 jours*,
Department of Statistical modelling, Tokyo Japan,.

Juil. 2015 **The Institute of Statistical Mathematics**, *10 jours*,
Department of Statistical modelling, Tokyo Japan,.

Mars 2015 **University College of London UCL**, *10 jours*,
Department of Statistical Sciences.

Juil. 2014 **The Institute of Statistical Mathematics**, *10 jours*,
Department of Statistical modelling, Tokyo Japan,.

Mai 2014 **University College of London UCL**, *une semaine*,
Department of Statistical Sciences.

Activité d'animation de la recherche

Dec. 2018 **Computational and Financial Econometrics (CFE 2018)**,
Organisation de la session CO410: Spatio-temporal models for best prediction of climate impacts on societies ,
University of Pisa , Italy.

Juil. 2017 **61th World Statistics Congress ISI2017** ,
Organisation de session STS066: Statistical models and applications of heavy tailed variables and processes ,
Marrakech Maroc.

Dec. 2016 **Computational and Financial Econometrics (CFE 2016)**,
Organisation de la session CO473: α-stable processes with applications in financial econometrics ,
Universitiy of Sevilla , Spain.

Juin 2014 **ANR Do Well Be, Mont-Dore**, *Co-organisateur*,
Workshop de lancement du projet ANR blanc Do Well Be.

Juin 2013 **ANR Do Well Be, Saint Nectaire**, *Co-organisateur*,
Workshop de lancement du projet ANR blanc Do Well Be.

Août 2012 **Journées MAS 2012**, *Co-organisateur*,
Université Blaise Pascal: Laboratoire de Mathématiques.

Séminaires conférences invités et communications orales

Conférences invités et Tutorials

juin. 2018 **IRCICA Lille**,
On sensors selection problems in IoT context and impulsive environement.

Fév. 2018 **The Institute of Statistical Mathematics, Tokyo Japan**, *A theoretical framework for extracting some temporal and frequency features in non-stationnary fractional signals.*,
Department of Statistical modelling.

juil. 2016 **The Institute of Statistical Mathematics Tokyo Japan**, *Construction of some special classes of stable processes that generalizes spatial or temporal Gaussian processes.*

Juil. 2015 **The Institute of Statistical Mathematics, Tokyo Japan**, *Identification and calibration problems in some non-stationary alpha-stables processes.*

Juil. 2014 **The Institute of Statistical Mathematics, Tokyo Japan**, *Bi-measures, Spectral-measures and Other Characterizations for Heavy Tail Processes.*

Mai 2013 **Congrès de la SMAI**,
Analyse du rythme cardiaque de médecins urgentistes du CHU Clermont-Ferrand.

Fév. 2013 **Congrès International du LEM2I**,
Sélection des variables dans un modèle de mélange de régressions normales.

déc. 2012 **IRCICA Lille**,
Frequency representation of stable process and channel modeling .

Présentations en Séminaires

Mar. 2018 **INSA de Lyon CITI Lab** ,
Heavy tailed distributions characterisations and examples of applications in channel modeling.

Janv. 2016 **Séminaire de L'IRMAR Rennes**,
On Construction and classification of some non stationary α-stable processes .

Avril. 2009 **Laboratoire LM2S à l'Université de Troyes**,
Sur une formulation théorique des décompositions modales empiriques et leur extension aux signaux multivariés.

Fév. 2008 **Séminaire invité à l'ENSAI Rennes** ,
Modélisation par des processus alpha-stables harmonisables non stationnaires : exemples d'application en communications .

Mars. 2008 **Université Claude Bernard ISFA Lyon** ,
Estimations spectrales de certains processus α-stables non-stationnaires .

Nov. 2007 **Séminaire au Laboratoire Paul Painlevé** ,
Représentation de type Lévy-Khinchine des processus symétriques alpha-stables en utilisant des bimesures .

Oct. 2007 **Séminaire invité au Laboratoire LSP à Toulouse** ,
Sur l'estimation spectrale de certains processus alpha-stables non stationnaires .

Sep. 2007 **Rencontres des jeunes statisticiens** ,
Estimations spectrales de certains processus α-stables non-stationnaires .

Mai. 2007 **Séminaire invité à l'université Pierre et Marie Curie - Paris 6** ,
Application des processus α-stables dans la modélisation du canal 60 Ghz .

Mars. 2007 **Séminaire invité à l'université de Rennes 1** ,
Sur une nouvelle représentation spectrale de certains processus alpha-stables non stationnaires .

Vulgarisation et diffusion auprès des entreprises:

Nov. 2017 **Journées Maintenance Montluçon** ,
Exemple de modèles de prédiction de pannes : Modèle de Cox avec covariables .

Avr. 2017 **Printemps 2017 Maths Entreprises en Auvergne** ,
Quelques outils de traitement de données de grandes dimensions dans le cotexte du Bigdata .

Avr. 2017 **Rencontre avec des mathématiciens** ,
Des mathématiques pour planifier, Exposé à l' hôtel de région d'Auvergne .

Synthèse des enseignements par établissement

2010–..... **Université Clermont Auvergne**, *total de (245h).*
- Master de statistiques: Analyse des données (50h), Séries chronoogiques (25h), Logiciels de statistiques (50h)
- Licence L3: Probabilités et statistiques (32h),
- Licence L2: Probabilités et statistiques (36h),
- Licence L1: Mathématiques Appliquées aux autres sciences (26h), Staistiques pour Psycho. (26h)

2010–2018 **Polytech Clermont-Ferrand**, *un total de (66h)* .
- 5ème Année: Modèles de Régression (16 h)
- 3ème Année: Probabilités et statistiques (34h), Méthodes statistiques (16h)

2008–2010 **Université de Technologie de Troyes**, *un total de (80h).*
- 4ème Année: Transmission de l'information (34h)
- 4ème Année: Traitement du signal (34h), Matlab (12h)

2007–2008 **Université de Lille 1**, *un total de (93h)* .
- Licence L2 SPI: Algèbre Linéaire (30h), Calcul differentiel et Intégration (30h)
- Licences L1 SVT: Mathématiques (33h)

2007–2008 **Telecom Lille** , *un total de (99h).*

- 1ère année: Mathématiques pour l'ingénieur (77h)
- 1ère année: Probabilités et statistiques (22h)

2006–2007 **Université de Bourgogne Dijon**, *Un total de (70h)* .

- Licence L2 Psycho: Statistiques et Probabilités (50h)
- Licence L1: Mathématiques pour SVTE (8h), Logiciel Maple (12h).

2006–2007 **Ensbana Supagro Dijon**, *un total de (36h).*

- 5ème Année: Analyses des données (28h), Logiciel de statistiques SAS (8h)

Bibliographie

[1] **N. AZZAOUI**, G. W. Peters, A. Guillin et M. Egan. *Spectral Characterization of the Non-Independent Increment Family of α-Stable Processes that Generalize Gaussian Process Models.* doi : 10.2139/ssrn.2892547.

[2] **N. AZZAOUI**, G. W. Peters, A. Guillin et M. Egan. *Supplement to : 'Spectral Characterization of the Family α-Stable Processes that Generalize Gaussian Process Models.'* doi : 10.2139/ssrn.2892570.

[3] A. Tadini, **N. AZZAOUI**, O. Roche, P. Samaniego, B. Bernard, A. Bevilacqua, S. Hidalgo, A. Guillin et M. Gouhier. « Tephra fallout probabilistic hazard maps for Cotopaxi and Guagua Pichincha volcanoes (Ecuador) with uncertainty quantification ». In : *Journal of Geophysical Research : Solid Earth* 127.2 (2022). doi : 10.1029/2021JB022780.

[4] D. Abdillahi-Ali, **N. AZZAOUI**, A. Guillin, G. Le Mailloux et T. Matsui. « Penalized Least Square in Sparse Setting with Convex Penalty and Non Gaussian Errors ». In : *Acta Mathematica Scientia* 41.6 (2021), p. 2198-2216. doi : 10.1007/s10473-021-0624-0.

[5] M. Campi, G. W. Peters, **N. AZZAOUI** et T. Matsui. « Machine Learning Mitigants for Speech Based Cyber Risk ». In : *IEEE Access* 9 (2021), p. 136831-136860. doi : 10.1109/ACCESS.2021.3117080.

[6] O. Roche, **N. AZZAOUI** et A. Guillin. « Discharge rate of explosive volcanic eruption controls runout distance of pyroclastic density currents ». In : *Earth and Planetary Science Letters* 568 (2021), p. 117017. doi : 10.1016/j.epsl.2021.117017.

[7] A. Tadini, O. Roche, P. Samaniego, **N. AZZAOUI**, A. Bevilacqua, A. Guillin, M. Gouhier, B. Bernard, W. Aspinall, S. Hidalgo et al. « Eruption type probability and eruption source parameters at Cotopaxi and Guagua Pichincha volcanoes (Ecuador) with uncertainty quantification ». In : *Bulletin of Volcanology* 83.5 (2021), p. 1-25. doi : 10.1007/s00445-021-01458-z.

[8] P. Cattiaux et A. Guillin. « On the Poincaré constant of log-concave measures ». In : *Geometric Aspects of Functional Analysis*. Springer, 2020, p. 171-217.

[9] K. Salah et S. Halim. « Kernel function and dimensionality reduction effects on speaker verification system ». In : *2020 International Conference on Electrical Engineering (ICEE)*. IEEE. 2020, p. 1-4.

[10] A. Tadini, O. Roche, P. Samaniego, A. Guillin, **N. AZZAOUI**, M. Gouhier, M. de'Michieli Vitturi, F. Pardini, J. Eychenne, B. Bernard et al. « Quantifying the uncertainty of a coupled plume and tephra dispersal model : PLUME-MOM/HYSPLIT simulations applied to Andean volcanoes ». In : *Journal of Geophysical Research : Solid Earth* 125.2 (2020). doi : 10.1029/2019JB018390.

[11] M. Y. Faisal et S. Suyanto. « SpecAugment impact on automatic speaker verification system ». In : *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE. 2019, p. 305-308.

[12] M. Gouhier, J. Eychenne, **N. AZZAOUI**, A. Guillin, M. Deslandes, M. Poret, A. Costa et P. Husson. « Low efficiency of large volcanic eruptions in transporting very fine ash into the atmosphere ». In : *Nature : Scientific reports* 9.1 (2019), p. 1-12. doi : 10.1038/s41598-019-38595-7.

[13] A. Tadini, O. Roche, P. Samaniego, A. Guillin, **N. AZZAOUI**, M. Gouhier, A. Bevilacqua, M. De'Michieli Vitturi, F. Pardini, B. Bernard et al. « Developing probabilistic tephra fallout hazard maps for Cotopaxi and Guagua Pichincha volcanoes, Ecuador, with uncertainty quantification ». In : t. 2019. 2019, p. V23I-0313. url : lien.

[14] S. Coly, P. Druilhet, **N. AZZAOUI** et P. A. Faye. *Plans d'échantillonnage spatiaux dans un contexte de surveillance*. Sous la dir. de 50èmes Journées de Statistique. 2018. url : lien.

[15] A. A.Bellec G.Lecué. « Towards the study of least squares estimators with convex penalty ». In : *Arxiv* (2017).

[16] M. Egan, S. M. Perlaza et V. Kungurtsev. *Capacity Sensitivity of Continuous Channels*. Rapp. tech. 9012. Lyon, France : INRIA, jan. 2017.

[17] P.-A. Faye, P. Druilhet, **N. AZZAOUI** et A.-F. Yao. *Bayesian spatio-temporal kriging with misspecified black-box*. 2017. url : https://hal.archives-ouvertes.fr/hal-01226169.

[18] M. L. de Freitas, M. Egan, L. Clavier, A. Goupil, G. W. Peters et **N. AZZAOUI**. « Capacity Bounds for Additive Symmetric α-Stable Noise Channels ». In : *IEEE Transactions on Information Theory* 63.8 (août 2017), p. 5115-5123. doi : 10.1109/TIT.2017.2676104.

[19] I. Keita, G. Takato, T. Matsui, G. W. Peters et **N. AZZAOUI**. « Feature Extraction Using Empirical Mode Decomposition for Tire Sensing ». In : *Infomation-Based Induction Sciences and Machine Learning (IBIS-ML2017)*. 2017, p. 315-320. url : https://www.ieice.org/ken/paper/20171110nbZh/eng/.

[20] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco et H. Delgado. « ASVspoof : the automatic speaker verification spoofing and countermeasures challenge ». In : *IEEE Journal of Selected Topics in Signal Processing* 11.4 (2017), p. 588-604.

[21] M. Egan, M. de Freitas, L. Clavier, A. Goupil, G. Peters et N. Azzaoui. « Achievable rates for additive isotropic alpha-stable noise channels ». In : *IEEE International Symposium on Information Theory*. 2016.

[22] S. van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

[23] K. Sriskandaraja, V. Sethu, E. Ambikairajah et H. Li. « Front-end for antispoofing countermeasures in speaker verification : Scattering spectral decomposition ». In : *IEEE Journal of Selected Topics in Signal Processing* 11.4 (2016), p. 632-643.

[24] N. Farsad, W. Guo, C.-B. Chae et A. Eckford. « Stable distributions as noise models for molecular communication ». In : *Proc. IEEE Global Communications Conference*. 2015.

[25] M. Gouhier, A. Guillin, **N. AZZAOUI**, J. Eychenne et S. Valade. « Source mass eruption rate retrieved from satellite-based data using statistical modelling ». In : *EGU General Assembly Conference Abstracts*. T. 17. 2015. url : https://ui.adsabs.harvard.edu/abs/2015EGUGA..1710222G/abstract.

[26] **N. AZZAOUI**, L. Clavier, A. Guillin et G. W. Peters. « Spectral Measures of α-Stable Distributions : An Overview and Natural Applications in Wireless Communications ». In : *Theoretical Aspects of Spatial-Temporal Modeling* (1er jan. 2015). doi : 10.1007/978-4-431-55336-6_3.

[27] X. Yan, L. Clavier, G. W. Peters, **N. AZZAOUI**, F. Septier et I. Nevat. « Skew-t copula for dependence modelling of impulsive (α-stable) interference ». In : *Proc. IEEE Int. Conf. Communications (ICC)*. Juin 2015, p. 4816-4821. doi : 10.1109/ICC.2015.7249085.

[28] M. Bashar, M. T. Ahmed, M. Syduzzaman, P. J. Ray et A. Islam. « Text-independent speaker identification system using average pitch and formant analysis ». In : *International Journal on Information Theory (IJIT)* 3.3 (2014), p. 23-30.

[29] **N. AZZAOUI**, A. Guillin, M. Gouhier, J. Eychenne et S. Valade. *Modélisation statistique pour la surveillance des éruptions volcaniques*. Sous la dir. de R. d'Auvergne. 2014, p. 153-170. url : lien.

[30] **N. AZZAOUI**, A. Guillin, F. Dutheil, G. Boudet, A. Chamoux, C. Perrier, J. Schmidt et P. R. Bertrand. « Classifying heartrate by change detection and wavelet methods for emergency physicians ». In : *Essaim* 45 (2014), p. 48-57. doi : 10.1051/proc/201445005.

[31] B. Nikfar, T. Akbudak et A. Han Vinck. « MIMO capacity of class A impulsive noise channel for different levels of information availability at transmitter ». In : *IEEE International Symposium on Power Line Communications and Its Applications*. 2014.

[32] C. Bonadonna et A. Costa. « Estimating the volume of tephra deposits : a new simple strategy ». In : *Geology* 40.5 (2012), p. 415-418.

[33] B. Hafidi et **N. AZZAOUI**. « Criteria for longitudinal data model selection based on Kullback's symmetric divergence ». In : *Revue ARIMA* 15 (2012), p. 83-99. url : https://arima.episciences.org/1959.

[34] **N. AZZAOUI** et L. Clavier. « UWB channel modeling for objects evolving in impulsive environnements ». In : *Proc. IEEE Wireless Communications and Networking Conf. Workshops (WCNCW)*. Avr. 2012, p. 191-195. doi : 10.1109/WCNCW.2012.6215488.

[35] N. Städler et P. Bühlmann. « Missing values : sparse inverse covariance estimation and an extension to sparse regression ». In : *Statistics and Computing* 22.1 (2012), p. 219-235.

[36] Y. Wu et S. Verdú. « Functional properties of minimum mean-square error and mutual information ». In : *IEEE Transactions on Information Theory* 58.3 (2012), p. 1289-1301.

[37] A. Ayache et P. R. Bertrand. « Discretization error of wavelet coefficient for fractal like processes ». In : (2011).

[38] P. R. Bertrand, M. Fhima et A. Guillin. « Off-line detection of multiple change points by the filtered derivative with p-value method ». In : *Sequential Analysis* 30.2 (2011), p. 172-207.

[39] C. Bonadonna, R. Genco, M. Gouhier, M. Pistolesi, R. Cioni, F. Alfano, A. Hoskuldsson et M. Ripepe. « Tephra sedimentation during the 2010 Eyjafjallajökull eruption (Iceland) from deposit, radar, and satellite observations ». In : *Journal of Geophysical Research : Solid Earth* 116.B12 (2011).

[40] L. Fillatre, P. Honeine, I. Nikiforov, C. Richard, H. Snoussi, **N. AZZAOUI**, B. Guépié, Z. Noumir, S. Deveughèle et H. Yin. « Vigires'eau : Surveillance en temps réel de la qualité de l'eau potable d'un réseau de distribution en vue de la détection d'intrusions ». In : *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'11),(ANR-CSOSG)*. 2011. url : lien.

[41] S. Petry, C. Flexeder et G. Tutz. « Pairwise fused lasso ». In : (2011).

[42] A. Rabbachin, T. Quek, H. Shin et M. Win. « Cognitive Network Interference ». In : 29.2 (fév. 2011), p. 480-493.

[43] J. Wang, E. Kuruoglu et T. Zhou. « Alpha-stable channel capacity ». In : *IEEE Communications Letters* 15.10 (2011), p. 1107-1109.

[44] N. Azzaoui et L. Clavier. « Statistical channel model based on α-stable random processes and application to the 60 GHz ultra wide band channel ». In : *IEEE Transactions on Communications* 58.5 (2010), p. 1457-1467.

[45] P. Cardieri. « Modeling interference in wireless ad hoc networks ». In : *IEEE Communications Surveys & Tutorials* 12.4 (2010), p. 551-572.

[46] L. Clavier et **N. AZZAOUI**. *Processus alpha-stables et communications numériques*. Sous la dir. de J. M. et Journée en l'honneur de Jacques Neveu. 2010. url : pdf.

[47] J. Friedman, T. Hastie et R. Tibshirani. « Regularization paths for generalized linear models via coordinate descent ». In : *Journal of statistical software* 33.1 (2010), p. 1.

[48] H. E. Ghannudi, L. Clavier, N. Azzaoui, F. Septier et P.-A. Rolland. « α-stable interference modeling and Cauchy receiver for an IR-UWB *ad hoc* network ». In : *IEEE Trans. Commun.* 58 (juin 2010), p. 1748-1757.

[49] H. E. Ghannudi, L. Clavier, N. Azzaoui, F. Septier et P. Rolland. « α-stable interference modeling and Cauchy receiver for an IR-UWB *ad hoc* network ». In : 58.6 (juin 2010), p. 1748-1757.

[50] A. Miraoui, H. Snoussi, J. Duchêne et **N. AZZAOUI**. « On the detection of elderly equilibrium degradation using multivariate-EMD ». In : *Proc. IEEE Globecom Workshops*. Déc. 2010, p. 2049-2053. doi : 10.1109/GLOCOMW.2010.5700305.

[51] P. Pinto et M. Win. « Communication in a Poisson Field of Interferers–Part I : Interference Distribution and Error Probability ». In : 9.7 (juill. 2010), p. 2176-2186. doi : 10.1109/TWC.2010.07.060438.

[52] P. Pinto et M. Win. « Communication in a Poisson Field of Interferers-Part II : Channel Capacity and Interference Spectrum ». In : 9.7 (juill. 2010), p. 2187-2195. doi : 10.1109/TWC.2010.07.071283.

[53] P. Pinto et M. Win. « Communication in a poisson field of interferers-part II : channel capacity and interference spectrum ». In : *IEEE Transactions on Wireless Communications* 9.7 (2010), p. 2187-2195.

[54] Y. Chen et N. Beaulieu. « Interference Analysis of UWB Systems for IEEE Channel Models Using First- and Second-Order Moments ». In : 57.3 (mars 2009), p. 622-625.

[55] A. Miraoui, **N. AZZAOUI**, H. Snoussi et J. Duchêne. *Détection de dégradation de l'équilibre chez les personnes âgées*. Sous la dir. de C. rendu des Journées SFTAG. 2009. url : lien.

[56] A. Molisch. « Ultra-Wide-Band Propagation Channels ». In : 97.2 (fév. 2009), p. 353-371.

[57] **N. AZZAOUI**, A. Miraoui, H. Snoussi et J. Duchene. « Empirical Mode Decomposition for vectorial bi-dimensional signals ». In : *Proc. Int. Workshop Multidimensional (nD) Systems*. Juin 2009, p. 1-4. doi : 10.1109/NDS.2009.5191553.

[58] **N. AZZAOUI**, H. Snoussi et J. Duchene. « An enhanced empirical modal decomposition without sifting ». In : *Proc. IEEE/SP 15th Workshop Statistical Signal Processing*. Août 2009, p. 796-799. doi : 10.1109/SSP.2009.5278474.

[59] W. I. Rose et A. J. Durant. « Fine ash content of explosive eruptions ». In : *Journal of Volcanology and Geothermal Research* 186.1-2 (2009), p. 32-39.

[60] M. Win, P. Pinto et L. Shepp. « A Mathematical Theory of Network Interference and Its Applications ». In : 97.2 (fév. 2009), p. 205-230.

[61] J.-M. Bardet, H. Bibi et A. Jouini. « Adaptive wavelet-based estimator of the memory parameter for stationary Gaussian processes ». In : *Bernoulli* 14.3 (2008), p. 691-724.

[62] P. Besbeas et B. Morgan. « Improved estimation of the stable laws ». In : *Statistics and Computing* 18.2 (2008), p. 219-231.

[63] L. Clavier, **N. AZZAOUI** et W. Sawaya. *UWB and 60 GHz channel model as an α-stable random process*. Rapp. tech. Technical Report TD (08) 634, Lille, France, 2008. url : Lien.

[64] J. Yeh, S. Fan et J. Shieh. « Human heart beat analysis using a modified algorithm of detrended fluctuation analysis based on empirical mode decomposition ». In : *Medical Engineering and Physics* (2008).

[65] D. Yinfeng, L. Yingmin, X. Mingkui et L. Ming. « Analysis of earthquake ground motions using an improved Hilbert–Huang transform ». In : *Soil Dynamics and Earthquake Engineering* 28.1 (2008), p. 7-19.

[66] J.-M. Bardet et P. Bertrand. « Identification of the multiscale fractional Brownian motion with biomechanical applications ». In : *Journal of Time Series Analysis* 28.1 (2007), p. 1-52.

[67] J. Gao. *Nonlinear Time Series Semiparametric and Nonparametric Methods*. Chapman & Hall/CRC, 2007.

[68] H. E. Ghannudi, L. Clavier et P. Rolland. « Modeling Multiple Access Interference in Ad Hoc Networks Based on IR-UWB Signals Up-Converted to 60 GHz ». In : *European Conference on Wireless Technologies*. Munich, Germany, oct. 2007, p. 106-109.

[69] P. Goncalves, P. Abry, G. Rilling et P. Flandrin. « Fractal Dimension Estimation : Empirical Mode Decomposition Versuswavelets ». In : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*. T. 3. 15–20 April 2007, p. III-1153-III-1156. doi : 10.1109/ICASSP.2007.366889.

[70] W. Liu et L. Wang. « BER Analysis in A Generalized UWB Frequency Selective Channel With Randomly Arriving Clusters and Rays ». In : juin 2007, p. 4281-4286.

[71] **N. AZZAOUI** et L. Clavier. « An Impulse Response Model for the 60 Ghz Channel Based on Spectral Techniques of α-stable Processes ». In : *Proc. IEEE Int. Conf. Communications*. Juin 2007, p. 5040-5045. doi : 10.1109/ICC.2007.832.

[72] C. Park et T. Rappaport. « Short-range Wireless Communications for Next-Generation Networks : UWB, 60 GHz Millimeter-Wave WPAN, and ZigBee ». In : 14.4 (août 2007), p. 70-78.

[73] G. Rilling, P. Flandrin, P. Gonalves et J. M. Lilly. « Bivariate Empirical Mode Decomposition ». In : *h* 14.12 (déc. 2007), p. 936-939. doi : 10.1109/LSP.2007.904710.

[74] S. Scollo, P. Del Carlo et M. Coltelli. « Tephra fallout of 2001 Etna flank eruption : Analysis of the deposit and plume dispersion ». In : *Journal of Volcanology and Geothermal Research* 160.1-2 (2007), p. 147-164.

[75] T. Tanaka et D. Mandic. « Complex Empirical Mode Decomposition ». In : *Signal Processing Letters, IEEE* 14.2 (2007), p. 101-104.

[76] W. Wu, C. Lee, C. Wang et C. Chao. « Signal-to-Interference-Plus-Noise ratio Analysis for Direct-Sequence Ultra-Wideband Systems in Generalized Saleh-Valenzuela Channels ». In : *IEEE Journal of selected Topics in signal processing* 1.3 (oct. 2007), p. 483-497.

[77] R. Azari, L. Li et C. Tsai. « Longitudinal data model selection ». In : *Computational Statistics and Data Analysis* 50.11 (2006), p. 3053-3066.

[78] B. Bosco, R. Emrick, S. Franson, J. Holmes et S. Rockwell. « Emerging Commercial Applications using the 60 GHz Unlicensed Band : Opportunities and Challenges ». In : déc. 2006, p. 1-4.

[79] D. Cabric, M. Chen, D. Sobel, S. Wang, J. Yang et R. Brodersen. « Novel Radio Architectures for UWB, 60 GHz, and Cognitive Wireless Systems ». In : *EURASIP Journal on Wireless Communications and Networking* 2006 (2006), Article ID 17957, 18 pages.

[80] J. Fiorina et W. Hachem. « On the asymptotic distribution of the correlation receiver output for time-hopped UWB signals ». In : 54.7 (juill. 2006), p. 2529-2545.

[81] B. Hafidi. « A small-sample criterion based on Kullback's symmetric divergence for vector autoregressive modeling ». In : *Statistics and Probability Letters* 76.15 (2006), p. 1647-1654.

[82] B. Hafidi et A. Mkhadri. « A corrected Akaike criterion based on Kullback's symmetric divergence : applications in time series, multiple and multivariate regression ». In : *Computational Statistics and Data Analysis* 50.6 (2006), p. 1524-1550.

[83] S. Kizhner, K. Blank, T. Flatley, N. E. Huang, D. Petrick et P. Hestnes. « On certain theoretical developments underlying the Hilbert-Huang transform ». In : *Proc. IEEE Aerospace Conference*. Avr. 2006, 14pp. doi : 10.1109/AERO.2006.1656061.

[84] **N. AZZAOUI**. « Analyse et Estimations Spectrales des Processus alpha-Stables non-Stationnaires ». Université de Bourgogne, 2006. url : lien.

[85] P. Pinto, C.-C. Chong, A. Giorgetti, M. Chiani et M. Win. « Narrowband Communication in a Poisson Field of Ultrawideband Iinterferers ». In : *The IEEE 2006 International Conference on Ultra-Wideband (ICUWB)*. Sept. 2006, p. 387-392.

[86] G. Rilling et P. Flandrin. « on the Influence of Sampling on the Empirical Mode Decomposition ». In : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2006*. T. 3. 14–19 May 2006, p. III-III. doi : 10.1109/ICASSP.2006.1660686.

[87] D. Zha et T. Qiu. « Underwater sources location in non-Gaussian impulsive noise environments ». In : *Digital Signal Processing* 16 (2006), p. 149-163.

[88] F. Barthe, P. Cattiaux et C. Roberto. « Concentration for independent random variables with heavy tails ». In : *Applied Mathematics Research eXpress* 2005.2 (2005), p. 39-60.

[89] F. Barthe, P. Cattiaux et C. Roberto. « Concentration for independent random variables with heavy tails ». In : *Applied Mathematics Research eXpress* 2005.2 (2005), p. 39-60.

[90] C. Bonadonna et B. Houghton. « Total grain-size distribution and volume of tephra-fall deposits ». In : *Bulletin of Volcanology* 67.5 (2005), p. 441-456.

[91] N. Demesh et S. Chekhmenok. « Estimation of the spectral density of a homogeneous random stable discrete time field ». In : *Statistics and Operations Research Transactions* 29 (2005), p. 101-118.

[92] D. Hedeker et R. Gibbons. *Applied Longitudinal Data Analysis*. Wiley ; John Wiley distributor, 2005.

[93] A. Molisch. « Ultrawideband propagation channels-theory, measurement, and modeling ». In : 54.5 (sept. 2005), p. 1528-1545.

[94] G. Rilling, P. Flandrin et P. Goncalves. « Empirical mode decomposition, fractional Gaussian noise and Hurst exponent estimation ». In : *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*. T. 4. 18–23 March 2005, p. iv/489-iv/492. doi : 10.1109/ICASSP.2005.1416052.

[95] W. Stute et L. Zhu. « Nonparametric checks for single-index models ». In : *Annals of Statistics* 33.3 (2005), p. 1048-1083.

[96] H. Zou et T. Hastie. « Regularization and variable selection via the elastic net ». In : *Journal of the royal statistical society : series B (statistical methodology)* 67.2 (2005), p. 301-320.

[97] J. Cavanaugh. « Criteria for Linear Model Selection Based on Kullback's Symmetric Divergence ». In : *Australian & New Zealand Journal of Statistics* 46.2 (2004), p. 257-274.

[98] G. Fitzmaurice, N. Laird et J. Ware. *Applied Longitudinal Analysis*. Wiley-IEEE, 2004.

[99] P. Flandrin, G. Rilling et P. Goncalves. « Empirical mode decomposition as a filter bank ». In : 11.2 (fév. 2004), p. 112-114. doi : 10.1109/LSP.2003.821662.

[100] S. Ghassemzadeh, R. Jana, C. Rice, W. Turin et V. Tarokh. « Measurement and modeling of an ultra-wide bandwidth indoor channel ». In : *IEEE Transactions on Communications* 52.10 (2004), p. 1786-1796.

[101] S. Kizhner, T. P. Flatley, N. E. Huang, K. Blank et E. Conwell. « On the Hilbert-Huang transform data processing system development ». In : *Proc. IEEE Aerospace Conference*. T. 3. Juin 2004.

[102] Y. Xia, W. Li, H. Tong et D. Zhang. « A goodness-of-fit test for single-index models ». In : *Statistica Sinica* 14.1 (2004), p. 1-27.

[103] L. Yang et G. Giannakis. « Ultra-wideband communications : an idea whose time has come ». In : 21.6 (nov. 2004), p. 26-54.

[104] N. Azzaoui, L. Clavier et R. Sabre. « Path delay model based on α-stable distribution for the 60GHz indoor channel ». In : t. 3. Déc. 2003, p. 1638-1643.

[105] A. Lapidoth et S. Moser. « Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels ». In : *IEEE Transactions on Information Theory* 49.10 (2003), p. 2426-2467.

[106] **N. AZZAOUI**, L. Clavier et R. Sabre. « Path delay model based on α-stable distribution for the 60 GHz indoor channel ». In : *Proc. IEEE Global Telecommunications Conf. GLOBECOM '03*. T. 3. Déc. 2003, 1638-1643 vol.3. doi : 10.1109/GLOCOM.2003.1258515.

[107] L. Clavier, M. Fryziel, C. Garnier, Y. Delignon et D. Boulinguez. « Performance of DS-CDMA on the 60 GHz channel ». In : t. 5. Sept. 2002, p. 2332-2336.

[108] P. Diggle. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

[109] G. Durgin. *Space-Time Wireless Channels*. Prentice Hall PTR, oct. 2002.

[110] M. Fryziel, C. Loyez, L. Clavier et N. Rolland. « Path loss model of the 60 GHz radio channel ». In : *Microwave and optical technology letters* 34.3 (août 2002), p. 158-162.

[111] Y. Lostalen, Y. Corre, Y. Louët, Y. L. Helloco, S. Collonge et G. E. Zein. « Comparison of measurements and simulations in indoor environments for wireless local area networks ». In : t. 1. Mai 2002, p. 389-393.

[112] N. Moraitis et P. Constantinou. « Indoor channel modeling at 60 GHz for wireless LAN applications ». In : t. 3. Sept. 2002, p. 1203-1207.

[113] P. Smulders. « Exploiting the 60 GHz band for local wireless multimedia access : prospects and future direction ». In : 40.1 (jan. 2002), p. 140-147.

[114] H. Xu, V. Hukshya et T. Rappaport. « Spatial and Temporal Characteristics of 60 GHz Indoor Channel ». In : 20.3 (avr. 2002), p. 620-630.

[115] L. Clavier, M. Rachdi, M. Fryziel, Y. Delignon, V. L. Thuc, C. Garnier et P. Rolland. « Wide band 60GHz indoor channel : characterization and statistical modelling ». In : t. 4. Oct. 2001, p. 2098-2102.

[116] A. L. Goldberger. « Heartbeats, hormones, and health : is variability the spice of life ? » In : *American Journal of Respiratory and Critical Care Medicine* 163.6 (2001), p. 1289-1290.

[117] M. Al-Nuaimi et A. Siamarou. « Coherence bandwidth and K-factor measurements for indoor wireless radio channels at 62.4GHz ». In : avr. 2001, p. 275-278.

[118] M. Lavielle et E. Moulines. « Least-squares estimation of an unknown number of shifts in a time series ». In : *Journal of time series analysis* 21.1 (2000), p. 33-59.

[119] P. Billingsley. *Convergence of Probability Measures*. John Wiley et Sons, 1999.

[120] J. Cavanaugh. « A large-sample model selection criterion based on Kullback's symmetric divergence ». In : *Statistics and Probability Letters* 42.4 (1999), p. 333-343.

[121] H. Fofack et J. Nolan. « Tail behavior, modes and other characteristics of stable distributions ». In : *Extremes* 2.1 (1999), p. 39-58.

[122] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash et K. Johnson. « Formants of children, women, and men : The effects of vocal intensity variation ». In : *The Journal of the Acoustical Society of America* 106.3 (1999), p. 1532-1542.

[123] P. C. Ivanov, L. A. N. Amaral, A. L. Goldberger, S. Havlin, M. G. Rosenblum, Z. R. Struzik et H. E. Stanley. « Multifractality in human heartbeat dynamics ». In : *Nature* 399.6735 (1999), p. 461-465.

[124] J. Kunisch, E. Zollinger, J. Pamp et A. Winkelmann. « MEDIAN 60 GHz Wideband Indoor Radio Channel Measurements and Model ». In : t. 4. Sept. 1999, p. 2393-2397.

[125] S. Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[126] D. Middleton. « Non-Gaussian noise models in signal processing for telecommunications : New methods and results for class A and class B noise models ». In : *IEEE Trans. Inf. Theory* 45.4 (mai 1999), p. 1129-1149.

[127] Y. Xia, H. Tong et W. Li. « On extended partially linear single-index models ». In : *Biometrika* 86.4 (1999), p. 831.

[128] C. Bonadonna, G. Ernst et R. Sparks. « Thickness variations and volume estimates of tephra fall deposits : the importance of particle Reynolds number ». In : *Journal of Volcanology and Geothermal Research* 81.3-4 (1998), p. 173-187.

[129] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung et H. Liu. « The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis ». In : *PROCEEDINGS-ROYAL SOCIETY OF LONDON A* 454 (1998), p. 903-995.

[130] R. Carroll, J. Fan, I. Gijbels et M. Wand. « Generalized partially linear single-index models ». In : *Journal of the American Statistical Association* 92.438 (1997), p. 477-489.

[131] J. Cavanaugh. « Unifying the derivations for the Akaike and corrected Akaike information criteria ». In : *Statistics and Probability Letters* 33.2 (1997), p. 201-208.

[132] Y. Fujikoshi et K. Satoh. « Modified AIC and Cp in multivariate linear regression ». In : *Biometrika* 84.3 (1997), p. 707.

[133] J. Gao et H. Liang. « Statistical inference in single-index and partially nonlinear models ». In : *Annals of the Institute of Statistical Mathematics* 49.3 (1997), p. 493-517.

[134] P. Smulders et L. Correia. « Characterisation of propagation in 60GHz radio channels ». In : *Electronic and communication engineering journal* 9.2 (avr. 1997), p. 73-80.

[135] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger et al. « Heart rate variability : standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology ». In : (1996).

[136] Y. Fan et Q. Li. « Consistent model specification tests : omitted variables and semiparametric functional forms ». In : *Econometrica : Journal of the Econometric Society* 64.4 (1996), p. 865-890.

[137] N. Huang, S. Long et Z. Shen. « The mechanism for frequency downshift in nonlinear wave evolution ». In : *Advances in applied mechanics* 32 (1996), p. 59-117.

[138] R. Tibshirani. « Regression shrinkage and selection via the lasso ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 58.1 (1996), p. 267-288.

[139] A. Weron. *Simulation and Chaotic Behavior of Stable Processes*. Birkhauser, 1996.

[140] D. Hand et M. Crowder. *Practical Longitudinal Data Analysis*. Chapman & Hall/CRC, 1995.

[141] C. L. Nikias et M. Shao. *Signal processing with α-stable distributions and applications*. Sous la dir. de W. inter science. J.Wiley, 1995.

[142] R. Sabre. « Discrete estimation of spectral density for symmetric stable process ». In : *Statistica* 60 (1995), p. 494-519.

[143] P. Smulders, J. Fernandes et A. Wagemans. « Frequency domain measurements of millimeter wave indoor radio channels ». In : 44.6 (déc. 1995), p. 1017-1022.

[144] G. Tsihrintzis, C. Nikias et M. Shao. « Performance of Optimum and Suboptimum Receivers in the Presence of Impulsive Noise Modeled as an Alpha-stable Process ». In : 43.2 (fév. 1995), p. 904-914.

[145] S. Ambike, J. Ilow et D. Hatzinakos. « Detection for binary transmission in a mixture of Gaussian noise and impulsive noise modeled as an alpha-stable process ». In : 1.3 (mars 1994), p. 55-57.

[146] E. Bedrick et C. Tsai. « Model Selection for Multivariate Regression in Small Samples ». In : *Biometrics* 50.1 (1994), p. 226-231.

[147] G. Samorodnitsky et M. Taqqu. « Stable non-Gaussian processes : Stochastic models with infinite variance ». In : (1994).

[148] G. Samorodnitsky et M. Taqqu. *Stable Non-Gaussian Random Processes : Stochastic Models with Infinite Variance*. Chapmann et Hall, 1994.

[149] P. Smulders et J. Fernandes. « Wide-band simulations and measurements of MM-wave indoor radio channels ». In : t. 2. Sept. 1994, p. 501-504.

[150] S. Wen et W. I. Rose. « Retrieval of sizes and total masses of particles in volcanic clouds using AVHRR bands 4 and 5 ». In : *Journal of Geophysical Research : Atmospheres* 99.D3 (1994), p. 5421-5431.

[151] M. Basseville, I. V. Nikiforov et al. *Detection of abrupt changes : theory and application*. T. 104. prentice Hall Englewood Cliffs, 1993.

[152] R. Jones. *Longitudinal Data With Serial Correlation : a state-space approach*. Chapman & Hall/CRC, 1993.

[153] S. Wales et D. Rickard. « Wideband propagation measurements of short range millimetric radio channels ». In : *Race 1043 - Electronics and Communication Engineering Journal* 5.4 (août 1993), p. 249-254.

[154] J. Fierstein et M. Nathenson. « Another look at the calculation of fallout tephra volumes ». In : *Bulletin of volcanology* 54.2 (1992), p. 156-167.

[155] E. Sousa. « Performance of a Spread Spectrum Packet Radio Network in a Poisson Field of Interferers ». In : 38.6 (nov. 1992), p. 1743-1754.

[156] T. Rappaport, S. Seidel et K. Takamizawa. « Statistical Channel Impulse Response Models for Factory and Open Plan Building Radio Communication System Design ». In : 39.5 (mai 1991), p. 794-807.

[157] C. Hurvich, R. Shumway et C. Tsai. « Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples ». In : *Biometrika* 77.4 (1990), p. 709.

[158] A. Verbyla. « A conditional derivation of residual maximum likelihood. » In : *AUST. J. STAT.* 32.2 (1990), p. 227-230.

[159] C. Hurvich et C. Tsai. « Regression and time series model selection in small samples ». In : *Biometrika* 76.2 (1989), p. 297.

[160] A. Prata. « Infrared radiative transfer calculations for volcanic ash clouds ». In : *Geophysical research letters* 16.11 (1989), p. 1293-1296.

[161] A. Prata. « Observation of volcanic ash clouds using AVHRR□2 radiances ». In : *Int. J. Rem. Sens* 10.4 (1989), p. 751-761.

[162] D. M. Pyle. « The thickness, volume and grainsize of tephra fall deposits ». In : *Bulletin of Volcanology* 51.1 (1989), p. 1-15.

[163] S. Seidel, K. Takamizawa et T. Rappaport. « Application of Second-Order Statistics for an Indoor Radio Channel Model ». In : San Francisco, CA, mai 1989, p. 888-892.

[164] A. A. M. Saleh et R. A. Valenzuela. « A statistical model for indoor multipath propagation ». In : SAC-5.2 (fév. 1987), p. 128-137.

[165] J. H. McCulloch. « Simple Consistent Estimators of Stable distribution Parameters ». In : *Communications on Satatistical Simulations* 15.4 (1986), p. 1109-1136.

[166] V. Zolotarev. *One-dimensional stable distributions (translations of mathematical monographs, vol. 65)*. The American Mathematical Society, 1986.

[167] E. Masry et S. Cambanis. « Spectral density estimation for stationary stable processes ». In : *Stochastic Processes And Their Applications* 18 (1984), p. 1-30.

[168] S. Cambanis. « Complex symmetric stable variables and processes ». In : *Contributions to Statistics : Essays in Honour of Norman L. Johnson* (1983). Sous la dir. de P. Sen. P.K. Sen, ed., North Holland, New York, p. 63-79.

[169] I. Koutrouvelis. « Iterative Procedure for the Estimation of the Parameters of Stable Laws ». In : *Communications in Statistics - Simulation and Computation* 10.1 (1981), p. 17-28.

[170] R. Lepage. « Multidimensional infinitely divisible variables and processes ». In : *Lecture Notes in Math.* 860.2 (1980), p. 279-284.

[171] H. Suzuki. « A Statistical Model for Urban Radio Propagation ». In : COM-25 (août 1979), p. 673-680.

[172] D. Middleton. « Statistical-physical models of electromagnetic interference ». In : *IEEE Transactions on Electromagnetic Compatibility* 19.3 (1977), p. 106-127.

[173] W. DuMouchel. « On the Asymptotic Normality of the Maximum-Likelihood Estimate when Sampling from a Stable Distribution ». In : *Annals of Statistics* 3 (1973), p. 948-957.

[174] S. Arimoto. « An algorithm for computing the capacity of arbitrary discrete memoryless channels ». In : *IEEE Transactions on Information Theory* 18 (1972), p. 14-20.

[175] R. Blahut. « Computation of channel capacity and rate-distortion functions ». In : *IEEE Transactions on Information Theory* 18.4 (1972), p. 460-473.

[176] G. Turin, F. Clapp, T. Johnson, S. Fine et D. Lavry. « A statistical model of urban multi-path propagation ». In : VT-21 (fév. 1972), p. 1-9.

[177] H. Patterson et R. Thompson. « Recovery of inter-block information when block sizes are unequal ». In : *Biometrika* 58.3 (1971), p. 545.

[178] E. Fama et R. Roll. « Some Properties of Symmetric Stable Distributions ». In : *Journal of the American Statistical Association* 63.323 (1968), p. 817-836.

[179] S. Kullback. *Information Theory and Statistics*. Dover Publications, 1968.

[180] E. A. Nadaraya. « On estimating regression ». In : *Theory of Probability & Its Applications* 9.1 (1964), p. 141-142.

[181] G. S. Watson. « Smooth regression analysis ». In : *Sankhyā : The Indian Journal of Statistics, Series A* (1964), p. 359-372.

[182] P. Bello. « Characterization of Randomly Time-Variant Linear Channels ». In : 11.4 (déc. 1963), p. 360-393.

[183] C. E. Shannon. « Two-way communication channels ». In : *Proc. 4th Berkeley Symp. Math., Statist. Probabil.* T. 1. Juin 1960, p. 611-644.

[184] V. Zolotarev. « Mellin-Stieltjes transforms in probability theory ». In : *Theory Probab. Appl.* 2.4 (1957), p. 433-460.

[185] C. Shannon. « A Mathematical Theory of Communication ». In : *Bell System Technical Journal* 27 (juill. 1948), 379-423 and 623-656.

Abstract

This document summarises my research work on statistical modelling and data processing issued from different application areas. The general themes of this research are part of the mainstream of statistical data analysis, signal processing and their applications. We insist, in particular, on the multi-disciplinary character of the results obtained. These results combine both theoretical work and application purposes. The main part of this work concerned the spectral analysis of α-stable processes and their applications in communication, data processing and in impulsive signals modelling The other part, which is not too far from the spectral analysis, deals with the time-frequency analysis of specific non-stationary signals and their applications, in particular in the analysis of physiological signals and data. The third part is devoted to regression models in small high dimensional contextes. Many applications have been devloped for variable and model selection for volcanological data among others.

Résumé

Ce document de synthèse résume mes travaux de recherche sur la modélisation et le traitement statistique de données issues de différents domaines d'applications. Les thématiques générales de ces recherches s'inscrivent dans le cadre de l'analyse statistique des données et du traitement du signal ainsi que de leurs applications. Nous insistons notamment sur le caractère multi-disciplinaire des résultats obtenus qui allient à la fois un travail souvent théorique avec des finalités applicatives. La part importante de ces travaux a concerné l'analyse spectrale des processus α-stables ainsi que leurs applications en communication, en traitement de données et des signaux impulsifs. L'autre volet, qui n'est pas très éloigné de l'analyse spectrale a concerné l'analyse temps-fréquences de certains signaux non stationnaires et leurs applications ; en particulier en analyse de signaux physiologiques. La troisième partie est consacrée aux modèles de régressions, de sélection des variables ainsi que leurs applications en volcanologie entre autres.