HAL
open science

Investigations on Stochastic Calculus, Statistics of Processes Applied Statistics for Biology and Medical Research

Nicolas Savy

▶ To cite this version:

Nicolas Savy. Investigations on Stochastic Calculus, Statistics of Processes Applied Statistics for Biology and Medical Research . Applications [stat.AP]. Université de Toulouse 3 Paul Sabatier, 2014. tel-01645597

HAL Id: tel-01645597
https://theses.hal.science/tel-01645597v1

Submitted on 23 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Document de synthèse pour l'obtention d'une

Habilitation à diriger des recherches

Discipline : **Mathématiques**
Spécialité : **Probabilités et Statistiques**

Investigations on Stochastic Calculus, Statistics of Processes

Applied Statistics for Biology and Medical Research

présenté par

Nicolas SAVY

au vu des rapports de

Pr Michael Kosorok	University of North Carolina
Pr Marc Lavielle	INRIA Saclay
Pr David Nualart	University of Kansas

Soutenue publiquement le 18 juin 2014 devant le jury composé de

Pr Serge Cohen	Université Paul Sabatier	Examinateur
Pr Fabrice Gamboa	Université Paul Sabatier	Coordinateur
Pr Michael Kosorok	University of North Carolina	Rapporteur
Pr Thierry Lang	Université Paul Sabatier	Examinateur
Pr Marc Lavielle	INRIA Saclay	Rapporteur
Pr Ivan Nourdin	Université du Luxembourg	Examinateur

Université Paul Sabatier - Toulouse III
Institut de Mathématiques de Toulouse

Document de synthèse pour l'obtention d'une

Habilitation à diriger des recherches

Discipline : **Mathématiques**
Spécialité : **Probabilités et Statistiques**

Investigations on Stochastic Calculus, Statistics of Processes

Applied Statistics for Biology and Medical Research

présenté par

Nicolas SAVY

au vu des rapports de

Pr Michael Kosorok	University of North Carolina
Pr Marc Lavielle	INRIA Saclay
Pr David Nualart	University of Kansas

Soutenue publiquement le 18 juin 2014 devant le jury composé de

Pr Serge Cohen	Université Paul Sabatier	Examinateur
Pr Fabrice Gamboa	Université Paul Sabatier	Coordinateur
Pr Michael Kosorok	University of North Carolina	Rapporteur
Pr Thierry Lang	Université Paul Sabatier	Examinateur
Pr Marc Lavielle	INRIA Saclay	Rapporteur
Pr Ivan Nourdin	Université du Luxembourg	Examinateur

Université Paul Sabatier - Toulouse III
Institut de Mathématiques de Toulouse

Remerciements

First of all, I would like to thank David Nualart, Michael Kosorok and Marc Lavielle for accepting to report on my manuscript and for their interest in my work. I am really honoured to see such eminent scientists for my examination board and I warmly thank each of them for their participation. Je suis très heureux que Serge Cohen ait accepté de prendre place dans ce jury. Il a fait beaucoup lors de mon arrivée à Toulouse et s'est toujours montré bienveillant et de bon conseil. Je l'en remercie très sincèrement. Sans Thierry Lang, il est certain que la seconde partie de ce mémoire n'existerait pas. Je suis vraiment comblé qu'il ai accepté de participer à ce jury. Ivan Nourdin a accepté mon invitation à participer à ce jury. J'en suis très heureux et lui en suis très reconnaissant. Enfin, un énorme merci à Fabrice Gamboa qui a coordonné mon habilitation avec une grande efficacité. Son enthousiasme et son professionalisme m'ont été d'un grand soutien, Merci.

Ensuite, mes remerciements vont à l'ensemble des co-auteurs avec qui j'ai collaboré durant ces années. Je voulais tout particulièrement saluer Vladimir Anisimov, Bernard Bercu, Jean-François Dupuy et Josep Vives. Le temps passe, les co-auteurs commencent à être nombreux, je ne les citerai donc pas tous mais je les remercie très sincèrement.

Laurent Decreusefond a fait beaucoup pour moi durant ma thèse mais également à la période charnière des premières années en tant que Maître de conférences, je tiens à le remercier vivement pour tout ce qu'il m'a appris.

Merci également à tous les membres de l'IMT. Ceux avec qui j'ai eu le plaisir de travailler, Fabien Panloup, Aldéric Joulin, Monique Pontier, Cécile Chouquet... Ceux avec qui il est toujours plaisant de discuter, Patrick Cattiaux, Laurent Miclo, Jean-Michel Loubès, Thierry Klein, Jean-Marc Azaïs, Jean-Yves Dauxois, Didier Concordet... Celui avec qui je partage - avec un grand plaisir - mon bureau depuis mon arrivée à Toulouse, Sébastien Gadat. Enfin je remercie tout spécialement Laure Coutin pour toute l'aide qu'elle m'a apportée, pour sa gentillesse et sa bienveillance. Je n'oublie pas le personnel administratif et technique de l'IMT qui fait un travail remarquable. Merci Françoise, Marie-Laure, Karima, Tamara, Marie-Line, Agnès et les informaticiens.

Travailler à l'interaction Mathématiques / Sciences du vivant ne peut se faire sans une relation étroite entre des équipes de chacune des disciplines. Ce n'est pas forcément facile, les codes, le langage ne sont pas toujours les mêmes. Pour ma part, j'estime avoir beaucoup de chance de travailler avec les équipes que je tiens à remercier maintenant.

- Au premier chef, l'unité INSERM 1027 et notamment l'équipe 1 de Sandrine Andrieu et l'équipe 5 de Thierry Lang. C'est un grand plaisir de travailler avec ces deux équipes. J'en profite pour saluer la remarquable ouverture d'esprit des membres de ces équipes et pour remercier comme il se doit Sandrine Andrieu.

- Le cercle des collaborations s'est aggrandi jusqu'à l'équipe 6 de l'unité INSERM 1027. Autour de la thèse de Caroline Delarue, avec Christine Damase, Isabelle Lacroix et Cécile Chouquet nous navigons entre statistiques et pharmacologie... c'est un vrai plaisir.

- Ensuite, un grand merci à Antoine Blancher, chef du service d'Immunologie du CHU de Toulouse, qui me fait découvrir les méandres de la génétique.

- Merci également a Felipe Guerrero du laboratoire d'Hématologie du CHU de Toulouse.

- Enfin, le cancérople GSO par le biais d'Olivier Claverie a souvent été là pour financer des événements que j'ai organisé et pour encourager des activités de recherches, je l'en remercie.

Je tiens à faire un paragraphe spécial pour les étudiants que j'ai eu le plaisir d'encadrer. Je leur dois beaucoup. Merci à Caroline, Merci à Nathan et Billy actuellement en thèse et surtout un énorme merci à Guillaume, Benoît et Valérie qui ont soutenu leur thèse. Je suis fier d'eux et je leur souhaite beaucoup de bonheur.

Je remercie également mes collègues du département GEA-R de l'IUT de Toulouse. Merci notamment à Fabien, toujours disponible pour arranger nos emplois du temps, Marc mon collègue de bureau, Nadège et Marie-Thérèse les chefs de département, Yann, André, Marc, Nathalie, Ann et tous les enseignants du département. Bien sûre merci également à Sylvie, Marjorie, Jocelyne et Franck de nous faciliter la vie au quotidien.

Je n'oublie pas mes parents, mes frères et plus généralement toute ma famille et ma belle-famille. Une pensée chaleureuse à tous.

Mes plus profonds remerciements vont à Stéphanie. Sans son enthousiasme communicatif pour les sciences du vivant et sa faculté à soulever des questions passionnantes, la seconde partie de ce mémoire n'aurait jamais vu le jour. Merci également pour le soutien au quotidien. Merci pour les encouragements. Merci pour les conseils... Bref, Merci pour tout et bien plus encore.

Je finirai le difficile exercice des remerciements par le plus simple : Gros bisous à Léonel et Achille.

Contents

Introduction **5**
Bibliography . . . 8

A- Investigations on Stochastic Calculus and Statistics of Processes 11

1 Malliavin Calculus and Anticipative Integrals **13**
I Filtered Poisson Process . . . 13
II Skohorod Integral with respect to filtered Poisson process . . . 16
III Anticipative integral for filtered Lévy process . . . 22
Bibliography . . . 28

2 A limit theorem for filtered Poisson processes **31**
I Hilbertian martingale and Radonification . . . 31
II Radonification of martingales associated to filtered processes . . . 32
III Convergence of Hilbertian martingales . . . 33
IV Back to the initial problem . . . 33
Bibliography . . . 34

3 Transportation Inequality and Malliavin Calculus **35**
I Ingredients . . . 35
II Upper bounds on Rubinstein distances . . . 38
III Applications . . . 38
Bibliography . . . 41

4 Properties of Estimators for some diffusion processes. **43**
I Estimator of instantaneous volatility . . . 44
II Large deviation Principle for drift parameter . . . 48
III Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes . . . 61
Bibliography . . . 63

B- Statistic applied to Biology and to Medical Research 65

5 Models for patients' recruitment in clinical trials **67**
I General considerations. . . . 67
II The Bayesian-Poisson models and their performance . . . 71
III Integration of screening-failures . . . 73
IV An additive model for clinical trials' cost modelling [13] . . . 78
Bibliography . . . 80

6 Survival data analysis for prevention in RCT **81**
I Weighted logrank tests . . . 82
II Fleming-Harrington's test . . . 85
III An "omnibus" test: maximum weighted Logrank statistic . . . 89
IV Application: the GuidAge study . . . 94
Bibliography . . . 95

7 Elements for mediation and evolution in Epidemiology **97**
I Approach by Markov Models . . . 98
II Approach by Causal Structural Equations . . . 101
Bibliography . . . 106

8 Various results in interaction with Biology **109**
I Predictive microbiology . . . 109
II Population Genetics . . . 112
Bibliography . . . 113

C- Conclusions and Perspectives **115**

9 Conclusions and Perspectives **117**
I On stochastic calculus . . . 117
II On survival data analysis . . . 118
III On patients' recruitment modelling . . . 118
IV On Clinical Trials Simulation . . . 119
V On mediation analysis . . . 120
VI On Prediction in Social Health . . . 121
Bibliography . . . 121

Curriculum Vitae **122**

Introduction

My research activity divides into two really distinct areas: probability and mathematical statistics of stochastic processes and applied statistics for Medicine and Biology. The manuscript naturally splits into two parts corresponding to these scientific axes. Part A is thus devoted to theoretical problems and Part B devoted to applied problems. I have enjoyed to share my time between those two aspects of mathematics since the very beginning of my Ph.D. thesis. In this first introductory section, I briefly present, in the chronological order, the different questions I have carried on during those last years*. It mixes the two aspects of my research but allows to highlight the links between the different activities.

At the very beginning, during my Phi, I have worked on a family of processes usually called Volterra (or filtered) processes which are defined through another process called the underlying process $\{X_t : t \in [0,T]\}$ (a Brownian motion, a Poisson process or a Lévy process) and a deterministic kernel K by:

$$\left\{ X_t^K = \int_0^t K(t,s)\,\mathrm{d}X_s \,:\, t \in [0,T] \right\}. \tag{1}$$

Details and especially the different assumptions are listed in Section I.1 of Chapter 1. The main motivation for introducing such a family of processes is that it generalizes fractional Brownian motion which appears to be a particular case where $X = B$ a Brownian motion, and $K = K^{(H)}$ a specific kernel defined in Section I.3 of Chapter 1. In order to study models based on such processes, it is crucial to first define a notion of integral with respect to it. That is done, first for fractional Brownian motion in [3, 19], second for Volterra Brownian process in [15] and finally for filtered Poisson processes in my paper with Laurent Decreusefond [18] which is summarized in Section II of Chapter 1. The link between Brownian Volterra process and filtered Poisson process has been clarified by means of a weak convergence theorem established by Laurent Decreusefond and myself in [17]. This result is obtained by the use of radonification techniques and is presented in Chapter 2.

My Ph.D. has been done while I worked as teacher in the "Technological University Institution" (IUT) of Quimper in the Biological Engineering department. It is in this context that I have learnt to teach applied statistics for biology. Some of my colleagues were chemists or biologists and members of a laboratory devoted to microbiology. A team of this laboratory worked on the so-called Predictive microbiology and aimed to model the behaviour of micro-organisms in different culture media (mainly industrial). I have begun to work in collaboration with this team. These investigations have led to publications [12, 13, 23] and a short summary of these contributions are presented in Chapter 8 Section I.

In 2007, I have been recruited by the University of Toulouse III on an assistant professor's position. I continued, in collaboration with Laurent Decreusefond, to study stochastic gradient associated with Poisson processes. This is the main ingredient in the construction of an anticipative integral with respect to filtered Poisson processes. We have extended the field of ours investigations to configurations space equipped with a Poisson measure on which two different gradients co-exist: the differential one [2] and the discrete one [34]. Feyel and Üstünel established in [22] a link between stochastic gradient and transportation inequalities. The question we carried on is what about this link in the configurations space setting since there are two different gradients. That leads us to the paper [16] in collaboration with Aldéric Joulin and described in Chapter 3.

During my Ph.D., Laurent Decreusefond and I have shown a Girsanov's theorem for filtered Poisson process and I was wondering how to apply this theorem to a relevant problem of estimation. Some discussions with Monique Pontier on this topic lead us to the really more attractive problem of instantaneous

*A more complete bibliography is given at the end of each chapter.

volatility estimation in models of the form

$$\mathrm{d}X_t = a_t \, \mathrm{d}t + \sigma_t \, \mathrm{d}B_t$$

where σ is a positive càdlàg semi-martingale. In collaboration with Fabien Panloup, we make use of an approach by power variations. We construct an estimator of σ_t and show its properties. The results [4] are presented in Section I of Chapter 4.

In collaboration with Bernard Bercu (University of Bordeaux) and Laure Coutin (IMT), we have investigated the sharp large deviations properties for functionals associated to fractional Ornstein-Uhlenbeck process defined by

$$\mathrm{d}X_t = \theta \, X_t \; \mathrm{d}t + \; \mathrm{d}B_t^{(H)}.$$

The functional is essentially the Maximum Likelihood Estimator of θ. The proofs follows essentially the same lines as in [10] but fractional Brownian motion induces huge technicalities. Despite its 36 pages, the paper [6] gives the proofs for only one functional. In the same spirit, we have enriched the results of [10] by including in [7] the unstable case ($\theta > 0$). These results are compiled in Section II of Chapter 4.

Chapter 4 ends by a section devoted to the estimation problem for the parameters of an Ornstein-Uhlenbeck process driven itself by another Ornstein-Uhlenbeck processes:

$$\begin{cases} \mathrm{d}X_t &= \theta X_t \, \mathrm{d}t &+ \mathrm{d}V_t, \qquad \text{for } t \in [0, T], \\ \mathrm{d}V_t &= \rho V_t \, \mathrm{d}t &+ \mathrm{d}B_t, \end{cases}$$

where $\theta < 0$, $\rho \leq 0$ and (B_t) is a standard Brownian motion. Studying such a model comes essentially from the fact that it is the continuous-time version of the first-order stable autoregressive process driven by a first-order autoregressive process recently investigated in [8]. The almost sure convergence as well as the asymptotic normality of the maximum likelihood estimators for θ and ρ are established in [9].

Although highly motivated by this research about stochastic processes, I missed a field of application of statistics. A discussion with Stéphanie Savy, specialized in clinical trials management allowed me to find that field of research: Medical research and especially statistics methods for clinical trials. Indeed, there were, at that date, no model for patients recruitment in clinical trials. The recruitment monitoring was done by basic deterministic techniques while it is nothing but a queue. An application of queuing models permitted to develop models and to estimate, not only punctually but also by means of confidence intervals, the date of the end of a clinical trial. This is an information of major importance in clinical trial monitoring. I co-supervise Guillaume Mijoule in the preparation of a Ph.D. thesis on this topic. He started in September 2009 and has developed several models essentially based on empirical Bayesian techniques. In [33] one has shown that such models are really relevant in applications and has investigated sensitivity analysis in order to highlight the role of the parameters in the models. In [5] (in preparation) the problem of drop-out (patients who are included but do not succeed in the inclusion's tests) is plugged in the models. Finally, an economic model is proposed in [32]. This model naturally yields to filtered Cox processes which are an extension of filtered Poisson processes of Chapter 1. Those results are compiled in Chapter 5.

In order to validate those models, I really need real-life datasets. I have contacted the INSERM (French Institute for Medical Research) Unit 1027 of Epidemiology of Toulouse. I explained my problem, they lent me some datasets and it was the very starting point of exiting and fruitful collaborations with several teams of that INSERM Unit: team "Cancer and chronic diseases: social inequalities in health, access to primary and secondary care" managed by Thierry Lang and team "Ageing and Alzheimer disease" managed by Sandrine Andrieu. Moreover, I received of an exemption from teaching during the whole first semester of 2012 in order to develop those relationships with medical researchers.

Team 5 of INSERM Unit 1027 works on social inequalities in health. The causal mechanisms by which exposures at different times of life to certain adverse social, environmental and psychosocial factors may be associated with occurrence of a pathology is a central issue in social epidemiology. The approach used by the team of Thierry Lang is called "life-course" [28, 27]. It is a conceptual model merging methods of human sciences and epidemiology. Under this approach, the susceptibility to a given disease is the result of an inevitable interaction between biological and social phenomena. The health status of an individual is the result of adaptation of the individual to his environment, this adaptation is dependent on the characteristics of the individual (biological, psychological, social), on the environment, themselves influenced

by factors such as socio-economic status (of the country and / or individual). Using a life-course approach therefore incorporates objective measures of health status, but also subjective ones on the onset of the disease or the presence of adverse social circumstances. To deal with such studies, the ideal database is a birth cohort. The team of Thierry Lang is working on the National Child Development Study cohort (NCDS) which follows around 17,000 members since 1958 resulting in a large number of variables (about 23,000). Although extremely rich in informations, the analysis of such a dataset is faced with statistical methodology issues [14]. In order to overpass these difficulties, two approaches have been investigated in the setting of two Ph.D. theses I have co-supervised with Thierry Lang. The one of Dominique Dedieu who has developed an analysis by means of Mixed Markov Hidden Models [20] synthesised in Section I of Chapter 7. Unfortunately, for personal convenience he stopped its Ph.D. at the end of the first year. The one of Benoît Lepage who dealt with the so-called structural causal model introduced in Section II of Chapter 7. These techniques were applied to mediation analysis in [30] and to evolution analysis in [29]

Team 1 of INSERM Unit 1027 deals with ageing and Alzheimer disease. Its manager, Sandrine Andrieu, was in charge of GuidAge's study (Section IV of Chapter 6 for details). That study led to a non-significant effect of the tested drug on Alzheimer disease. The test used for evaluation is logrank's one since the measurement is time to events (dementia). A post-hoc study was made using the well known Fleming-Harrington's test which concluded to to a positive result. Two explanations are valuable: the drug has effectively no efficiency or the methodology was not optimal. In virtue to what is observed in true life, we made the assumption of an ill-posed methodology and began to deal with Fleming-Harrington's test. With Sandrine Andrieu, we have proposed to precise the methodology for dealing with prevention trials in response to a France-Alzheimer call for tenders. Indeed GuidAge, as most of clinical trials on Alzheimer's disease is a prevention trial. In fact, there exists currently no effective treatment for this pathology, making its prevention a priority. Prevention is feasible due to the long asymptomatic latent period of the disease. To date, the rare published articles reporting the results of clinical trials having the appearance of the event "develop a dementia" as a criterion of judgement are negative [21, 31, 35, 36]. The treatment effect occurs late. Thus, the logrank test (which assumes that the hazard rates are proportional) is not appropriated in this setting and weighted logrank tests as Fleming Harrington might be more efficient. France Alzheimer gave us a grant which has funded Valérie Garès's Ph.D.. Sandrine Andrieu and I have supervised her Ph.D. thesis preparation. The aim is to improve the methodology for dealing with survival data in prevention clinical trials. For this we have to keep in mind two important points:

- The parameters of the tests have to be fixed in the research protocol.
- The necessary sample size has to be computed.

The test of interest (Fleming-Harrington) depends of a parameter. The very first question is how to fix that parameter. For this, we make use of stochastic calculus for jump processes in order to reach the distributions under null and alternative hypotheses. Then asymptotic relative efficiency allows to compare the different tests leading to a generating data process in such a way that we are sure that Fleming-Harrington's test is optimal. Hence, it permits to study the performance of the test and overall the sensitivity to the parameter [24]. Another test of interest has been studied essentially because its parameter has a medical reality (time from which the treatment acts). This test is in fact a weighted log-rank test and is compared in [25] to Fleming Harrington's test. Finally, Fleming Harrington's test is a good test for late effect and logrank is a good test for constant effect. Nevertheless, in real life, the trade off between late effect and constant effect is not easy to make when designing a clinical trial. We construct and study in [26] a composite test which does not avoid those assumptions but assume that the proportion of each assumption is known. All those results are summarized in Chapter 6.

Research on anticipative integration was still of actuality. Indeed, I had in mind to clarify the links between the integrals we are able to construct not only with respect to Poisson process but also with respect to Brownian motion and, by means of Lévy Itô decomposition, with respect to Lévy process (with Brownian component). I wondered also how those integrals behave when the underlying process is filtered by a deterministic kernel. I have recently solicited Josep Vives (University of Barcelona) for working together on this topics, the paper [37] is not completely ended but the preliminary results are presented in Section III of Chapter 1.

After this chronological description of my works, what about the structure of the manuscript ? It splits in two parts.
The first part is devoted to the investigations on stochastic calculus and to statistics for stochastic

processes. It contains four chapters. The first one explains the construction of stochastic integrals with respect to filtered processes. The second one shows briefly how a sequence of filtered Poisson processes converges weakly to a Brownian Volterra process. The third chapter is devoted to the link between Transportation Inequality and Malliavin Calculus in configurations space. Finally, this first part ends by a fourth chapter devoted to properties of estimators: instantaneous volatility and drift parameter for various Ornstein Uhlenbeck 's type processes.
The second part presents various works on applied statistics for Biology and Medical research. It contains four chapter. The first chapter presents models for patients recruitment in clinical trials. The second one presents the research on survival data analysis for prevention clinical trials. The third one summarizes a series of papers on epidemiology especially on mediation analysis. Finally the fourth chapter splits itself in two section. The first section presents various results obtained in collaboration with biologists during my stay in Quimper. Section II illustrates recent collaborative works [1, 11] in genetic of populations with Antoine Blancher (Laboratory of Immunogenetics - University of Toulouse III).
The manuscript ends with a description of questions I like to carry out in future.

Bibliography

[1] Alice Aarnink, **Nicolas Savy**, Nicolas Congy, Nicola Rosa, Edward Mee, and Antoine Blancher. Demonstration of the deleterious impact of foeto-maternal mhc compatibility on the success of pregnancy in a macaque model. *Immunogenetics*, 66:105–113, 2014.

[2] Sergio Albeverio, Yuri Kondratiev, and Michael Röckner. Analysis and geometry on configuration spaces. *J. Funct. Anal.*, 154(2):444–500, 1998.

[3] Elisa Alòs, Olivier Mazet, and David Nualart. Stochastic calculus with respect to Gaussian processes. *Ann. Probab.*, 29(2):766–801, 2001.

[4] Alexander Alvarez, Fabien Panloup, Monique Pontier, and **Nicolas Savy**. Estimation of the instantaneous volatility. *Statistic inference for Stochastic processes*, 15:27–59, 2012.

[5] Vladimir Anisimov, Guillaume Mijoule, and **Nicolas Savy**. Statistical modelling of recruitment in multicentre clinical trials with patients' dropout. *Statistics in Medicine*, 2014. In preparation.

[6] Bernard Bercu, Laure Coutin, and **Nicolas Savy**. Sharp large deviations for the fractional Ornstein-Uhlenbeck process. *Theory of Probability and its Applications*, 55(4):575–610, 2011.

[7] Bernard Bercu, Laure Coutin, and **Nicolas Savy**. Sharp large deviation for the non-stationary ornstein-uhlenbeck process. *Stochastic Processes and their Applications*, 122(10):3393–3424, 2012.

[8] Bernard Bercu and Frédéric Proïa. A sharp analysis on the asymptotic behavior of the Durbin-Watson statistic for the first-order autoregressive process. *ESAIM Probab. Stat.*, 17:500–530, 2013.

[9] Bernard Bercu, Frédéric Proïa, and **Nicolas Savy**. On Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes. *Statistics and Probability Letters*, 85:36–44, 2014.

[10] Bernard Bercu and Alain Rouault. Sharp large deviations for the Ornstein-Uhlenbeck process. *Theory Probab. Appl.*, 46(1):1–19, 2002.

[11] Antoine Blancher, Alice Aarnink, **Nicolas Savy**, and Nagushi Takahata. Use of cumulative poisson probability distribution as an estimator of the recombination rate from population-genetic data: example of the Macaca fascicularis major histocompatibility complex. *Genes, Genomes, Genetics*, 2:123–130, 2012.

[12] Louis Coroller, Ivan Leguerinel, Eric Mettler, **Nicolas Savy**, and Pierre Mafart. General model, based on two mixed weibull distributions of bacterial resistance, for describing various shapes of inactivation curves. *Applied Environmental Microbiology*, 72(10):6493–6502, 2006.

[13] Olivier Couvert, Stéphane Gaillard, **Nicolas Savy**, Pierre Mafart, and Ivan Leguérinel. Survival curves of heated bacterial spores: effect of environmental factors on weibull parameters. *International Journal of Food Microbiology*, 101(1):73–81, 2005.

[14] Bianca L. De Stavola, Dorothea Nitsch, Isabel dos Santos Silva, Valerie McCormack, Rebecca Hardy, Vera Mann, Tim J. Cole, Susan Morton, and David A. Leon. Statistical issues in life course epidemiology. *Am J Epidemiol.*, 163(1):84–96, 2006.

[15] Laurent Decreusefond. Stochastic integration with respect to Volterra processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 41(2):123–149, 2005.

[16] Laurent Decreusefond, Aldéric Joulin, and **Nicolas Savy**. Upper bounds on Rubinstein distances on configuration spaces and applications. *Communication on Stochastic Analysis and Applications*, 4(3):377–399, 2010.

[17] Laurent Decreusefond and **Nicolas Savy**. Filtered Brownian motions as weak limit of filtered Poisson processes. *Bernoulli*, 11(2):283–292, 2005.

[18] Laurent Decreusefond and **Nicolas Savy**. Anticipative calculus with respect to filtered Poisson processes. *Annales Institut Henri Poincaré Probabilités Statistiques*, 42(3):343–372, 2006.

[19] Laurent Decreusefond and Ali S. Üstünel. Stochastic analysis of the fractional Brownian motion. *Potential Anal.*, 10(2):177–214, 1999.

[20] Dominique Dedieu, Cyrille Delpierre, Sébastien Gadat, Benoît Lepage, Thierry Lang, and **Nicolas Savy**. Mixed hidden markov model for heterogeneous longitudinal data with missingness and errors in the outcome variable. *Journal de la Société Française de Statistiques*, 155(1):73–98, 2013.

[21] Steven T. DeKosky. Ginkgo Biloba for prevention of dementia: a randomized controlled trial. *Journal of the American Medical Association*, 300(19):2253–2262, 2008.

[22] Denis Feyel and Ali S. Üstünel. Monge-Kantorovitch measure transportation and Monge-Ampère equation on Wiener space. *Probab. Theory Related Fields*, 128(3):347–385, 2004.

[23] Stéphane Gaillard, Ivan Leguérinel, **Nicolas Savy**, and Pierre Mafart. Quantifying the combined effects of the heating time, the temperature and the recovery medium ph on the regrowth lag time of bacillus cereus spores after a heat treatment. *International Journal of Food Microbiology*, 105(1):53–58, 2005.

[24] Valérie Garès, Sandrine Andrieu, Jean-Francois Dupuy, and **Nicolas Savy**. Choosing the parameter of fleming-harringtons test in prevention randomized controlled trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2013. Submitted.

[25] Valérie Garès, Sandrine Andrieu, Jean-Francois Dupuy, and **Nicolas Savy**. Comparison of constant piecewise weighted test and fleming harrington's test - application in clinical trials. *Electronic Journal of Statistics*, 2013. In revision.

[26] Valérie Garès, Sandrine Andrieu, Jean-Francois Dupuy, and **Nicolas Savy**. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. *Statistics in Medicine*, 2013. In revision.

[27] Johan Hallqvist, John Lynch, Mel Bartley, Thierry Lang, and David Blane. Can we disentangle life course processes of accumulation, critical period and social mobility? an analysis of disadvantaged socio-economic positions and myocardial infarction in the stockholm heart epidemiology program. *Social Science and Medicine*, 58(8):1555–1562, 2004.

[28] Diana Kuh and Yoav Ben Shlomo. *A Life Course Approach To Chronic Diseases Epidemiology*. Oxford University Press, Usa, 2004.

[29] Benoît Lepage, Dominique Dedieu, Sébastien Lamy, **Nicolas Savy**, and Thierry Lang. Using directed acyclic graphs for change score analysis. *Epidemiology*, 2012. In revision.

[30] Benoît Lepage, Dominique Dedieu, **Nicolas Savy**, and Thierry Lang. Estimation of a controlled direct effect when an effect of the exposure confounds the mediator-outcome relationship: a comparison of different methods. *Statistic Methods for Medical Research*, 2012. Forthcoming.

[31] Constantine G. Lyketsos. Naproxen and celecoxib do not prevent ad in early results from a randomized controlled trial. *Neurology*, 68(21):1800–1808, 2007.

[32] Guillaume Mijoule, Nathan Minois, Vladimir Anisimov, and **Nicolas Savy**. Additive model for cost modelling in clinical trial. In *Proceedings of the 7th International Workshop on Simulation*, 2013. Forthcoming.

[33] Guillaume Mijoule, Stéphanie Savy, and **Nicolas Savy**. Models for patients' recruitment in clinical trials and sensitivity analysis. *Statistics in Medicine*, 31(16):1655–1674, 2012.

[34] David Nualart and Josep Vives. Anticipative calculus for the Poisson process based on the Fock space. In *Séminaire de Probabilités, XXIV, 1988/89*, volume 1426 of *Lecture Notes in Math.*, pages 154–165. Springer, Berlin, 1990.

[35] Sally A. Shumaker. Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: the women's health initiative memory study: a randomized controlled trial. *Journal of the American Medical Association*, 289(20):2651–2662, 2003.

[36] Sally A. Shumaker. Conjugated equine estrogens and incidence of probable dementia and mild cognitive impairment in postmenopausal women: Women's health initiative memory study. *Journal of the American Medical Association*, 291(24):2947–2958, 2004.

[37] **Nicolas Savy** and Josep Vives. Anticipative integrals with respect to a filtered Lévy process and Lévy-Itô decomposition. *Journal of Theoretical Probability*, 2014. Submitted.

Part A-

Investigations on Stochastic Calculus and Statistics of Processes

Chapter 1

Malliavin Calculus and Anticipative Integral.*

This first chapter is a summary of the results obtained on the construction of stochastic integral with respect to filtered processes. Let us recall that this kind of process is constructed by means of a so-called underlying process X and of a deterministic kernel K as:

$$\left\{ X_t^K = \int_0^t K(t,s)\,\mathrm{d}X_s \,:\, t \in [0,T] \right\},$$

I focus my attention to the special cases of filtered Poisson process during my Ph.D. and to filtered Lévy process during the past years in collaboration with Josep Vives (University of Barcelona).

The chapter organizes as follow: a first section is devoted to some generalities on filtered Poisson processes including the assumptions on the process X and the kernel K. A relevant way to construct an integral with respect to X^K is made in two steps. First one constructs an integral δ with respect to the underlying process X. Second one constructs an operator $\mathcal{K}^*$ closely linked to K in such a way that the integral on X^K is nothing but $\delta(\mathcal{K}^*(u))$. Unfortunately, even for predictable u, $\mathcal{K}^*(u)$ may be anticipative. The integral δ has to be defined in an anticipative way. Section 2 summarizes this construction I have published with Laurent Decreusefond [8]. Finally, the construction of integrals with respect to Lévy processes is presented in a third section extracted from a paper in preparation in collaboration with Josep Vives [16].

I Filtered Poisson Process

I.1 Definitions

Consider $(E,\mathcal{E})$ a measurable space (for notational simplicity, $E=\mathbb{R}$) and ν a positive σ-finite measure on $(E,\mathcal{E})$. Consider Ω the space of simple, locally bounded integer-valued measures on $[0,T]\times E$. One defines the probability $\mathbb{P}$ as the unique measure on Ω such that ω the canonic measure is a Poisson random measure whose intensity is ν. One defines the canonic filtration $(\mathcal{F}_t)$ by:

$$\mathcal{F}_0 \;=\; \{\emptyset,\omega\} \quad \text{and} \quad \mathcal{F}_t \;=\; \sigma\left\{ \int_0^s \int_B \omega(\,\mathrm{d}s,\,\mathrm{d}z)\;,\; s\le t, B\in\mathcal{E} \right\},\; t\in]0,T].$$

Finally, ones denotes $\mathcal{P}$ the predictable σ-algebra on $\Omega\times[0,T]\times E$.

Definition 1.1 *The space $(\Omega,\mathcal{F},(\mathcal{F}_t),\mathbb{P})$ is usually called the Poisson space.*

---

* Publications related to this chapter:

[8] Laurent Decreusefond and **Nicolas Savy**. Anticipative calculus with respect to filtered Poisson processes. *Annales Institut Henri Poincaré Probabilités Statistiques*, 42(3):343–372, 2006.

[16] **Nicolas Savy** and Josep Vivès. Anticipative integrals with respect to filtered Lévy processes and Lévy Itô decomposition. Submitted to *Journal of Theoretical Probability*, 2014.

Consider the sequence $\{(T_n, Z_n) : n \in \mathbb{N}^*\}$ associated with the marked point process (T_n denotes the instant of the n-th jump and Z_n its amplitude) this means

$$\omega(\,\mathrm{d}s,\,\mathrm{d}z) = \sum_{n\in\mathbb{N}^*} \epsilon_{(T_n,Z_n)}(s,z).$$

where ϵ_a denotes the Dirac measure concentrated in $\{a\}$.

Definition 1.2 *The marked point process defined as:*

$$N_t \;=\; \int_0^t \int_E z\,\omega(\,\mathrm{d}s,\,\mathrm{d}z) \;=\; \sum_{n\in\mathbb{N}^*} Z_n\;\mathbb{I}_{\{T_n\le t\}},$$

is called the underlying process and given a kernel $K : [0,T]^2 \to \mathbb{R}$, the process

$$N_t^K \;=\; \int_0^t K(t,s)\,\mathrm{d}N_s \;=\; \int_0^t \int_E zK(t,s)\,\omega(\,\mathrm{d}s,\,\mathrm{d}z) \;=\; \sum_{n\in\mathbb{N}^*} K(t,T_n)\,Z_n\;\mathbb{I}_{\{T_n\le t\}}, \tag{1.1}$$

is called the filtered process. The compensated underlying (resp. compensated filtered) process is:

$$\tilde{N}_t \;=\; \int_0^t \int_E z\,(\omega-\nu)(\,\mathrm{d}s,\,\mathrm{d}z) \quad \left(\textit{resp.}\quad \tilde{N}_t^K \;=\; \int_0^t \int_E zK(t,s)\,(\omega-\nu)(\,\mathrm{d}s,\,\mathrm{d}z)\right).$$

I.2 Assumptions on the kernel

A key assumption for the construction of a filtered process is the following one:

Hypothesis 1.1 *K is triangular in the sense that $K(t,s) = 0$ for all $s > t > 0$.*

Hypothesis 1.1 insures that X^K is an adapted process. In the whole manuscript we assume this assumption fulfilled.

First assumptions set.

Hypothesis 1.2

- *The application $(s,z) \to z\;K(t,s)$ belongs to $\mathcal{L}^2(\nu)$ for any $t > 0$.*
- *K does not explode on the diagonal in the sense that $K(t,t) < \infty$ for all $t \in [0,T]$.*
- *For any $t \ge 0$, the function*

$$K(t,.)\;:\;\begin{array}{ccc} [0,t[& \longrightarrow & \mathbb{R} \\ s & \longrightarrow & K(t,s) \end{array} \quad \textit{is càdlàg.}$$

- *For any $s \ge 0$, the function*

$$K(.,s):\;\begin{array}{ccc} [s,T[& \longrightarrow & \mathbb{R} \\ t & \longrightarrow & K(t,s) \end{array} \quad \textit{has bounded variations.}$$

Second assumptions set. For all $f \in \mathcal{L}^1([0,T])$, the left and right fractional integrals of f are defined as:

$$(I_{0+}^\alpha f)(x) = \frac{1}{\Gamma(\alpha)}\int_0^x f(t)(x-t)^{\alpha-1}\,\mathrm{d}t,\;\; x \ge 0,$$
$$(I_{b-}^\alpha f)(x) = \frac{1}{\Gamma(\alpha)}\int_x^b f(t)(t-x)^{\alpha-1}\,\mathrm{d}t,\;\; x \le b,$$

where $\alpha > 0$ and $I^0 = Id$. For any $\alpha \ge 0$, any $f \in \mathcal{L}^p([0,T])$ and any $g \in \mathcal{L}^q([0,T])$ where $p^{-1} + q^{-1} \le \alpha$, we have:

$$\int_0^T f(s)(I_{0+}^\alpha g)(s)\,\mathrm{d}s = \int_0^T (I_{T-}^\alpha f)(s)g(s)\,\mathrm{d}s.$$

Definition 1.3 *The Besov space* $I_{0+}^{\alpha}(\mathcal{L}^p) \overset{not}{=} \mathcal{I}_{\alpha,p}$ *is usually equipped with the norm:*

$$\|f\|_{\mathcal{I}_{\alpha,p}} = \|I_{0+}^{-\alpha} f\|_{\mathcal{L}^p}.$$

In particular $\mathcal{I}_{\alpha,2}$ *is a separable Hilbert space.*

Consider an Hilbert-Schmidt operator from $\mathcal{L}^2$ into $\mathcal{L}^2$ denoted by K.

Hypothesis 1.3 *There exists* $\alpha > 0$ *such that* K *is a continuous one-to-one linear application from* $\mathcal{L}^2$ *to* $\mathcal{I}_{\alpha+1/2,2}$.

Since the embedding from $\mathcal{I}_{\alpha+1/2,2}$ into $\mathcal{L}^2$ is Hilbert-Schmidt, it guarantees that K is a Hilbert-Schmidt map from $\mathcal{L}^2$ into itself. Thus there is a kernel, still denoted by K, such that the operator K takes the form

$$(Kf)(t) \;=\; \int_0^T K(t,s)f(s)\,\mathrm{d}s \quad \text{with} \quad \int_0^T\int_0^T K^2(t,s)dt\;\mathrm{d}s \;<\; \infty.$$

I.3 Examples

- Considering the kernel $K^{(H)}(t,s) = L_H(t,s)(t-s)^{H-1/2}s^{-|H-1/2|}$, with L_H a bi-continuous function [5], ones deals with the classical fractional Brownian motion. In [7] it is shown that $K^{(H)}$ satisfies assumptions 1.1 and 1.3 for any $H \in (0,1)$.
- Assumptions 1.1 and 1.3 are satisfied for the kernel defined for any $H \in (0,1)$ by:

  $$K(t,s) = \frac{1}{\Gamma(H+1/2)}(t-s)^{H-1/2}\,\mathbb{I}_{[0,t)}(s)$$

  which corresponds to the so-called Lévy fractional Brownian motion since, in terms of function, it coincides with $I_{0+}^{H+1/2}$.
- For any $\alpha > 0$, the kernel $(s,t) \;\to\; e^{\alpha(t-s)}$ allows us to consider Ornstein-Uhlenbeck processes.

I.4 Main properties

Property 1.1 (Properties of sample paths)

- *N^K is not a marked point process.*
- *If K is continuous, N^K has the same jumps times as N.*
- *If K is null on the diagonal, and if K is continuous, N^K is continuous.*
- *Under Assumption 1.2, the process N^K has finite variation on each compact set of $\mathbb{R}$.*
- *Assumption 1.2 insure that N^K is limited by the left. If moreover K is right continuous then N^K is a càdlàg process.*

Property 1.2 (Moments and covariance functions.) *For all $(t,t') \in [0,T]^2$, we have:*

$$\begin{aligned}
\mathbb{E}\left[\tilde{N}_t^K\right] \;&=\; 0,\\
\mathrm{Cov}(\tilde{N}_t^K;\tilde{N}_{t'}^K) \;&=\; \int_0^{t\wedge t'}\int_E z^2 K(t,s)K(t',s)\,\nu(\,\mathrm{d}s,\,\mathrm{d}z) \;=\; \int_0^{t\wedge t'} K(t,s)K(t',s)\,\tilde{\lambda}(\,\mathrm{d}s),
\end{aligned}$$

where $\tilde{\lambda}$ denotes the measure $\tilde{\lambda}(\,\mathrm{d}s) = \int_E z^2\,\nu(\,\mathrm{d}s,\,\mathrm{d}z)$ defined on $[0,T]$.

If the map $(s,t) \to K(t,s)$ is continuous then the filtered Poisson process

$$\left\{\tilde{N}_t^K \;=\; \int_0^t K(t,s)\,\mathrm{d}\tilde{N}_s \;:\; t \geq 0\right\}$$

is not a martingale. However, we have

Property 1.3 (Martingale's property.) *The process defined, for any $r \in [0,T]$ by:*

$$\left\{\tilde{N}_t^{K,r} \;=\; \int_0^t K(r,s)\,\mathrm{d}\tilde{N}_s \;:\; t \geq 0\right\}$$

is a functional-valued martingale.

II Skohorod Integral with respect to filtered Poisson process

Although it is possible to define integral by usual means for marked point process, in order to define an integral for filtered Poisson process, we have to focus on anticipative definition of integrals (see remark 1.2 page 20). In this section we browse the definition given in [8]. It is possible to deal with other approaches (chaotic expansion, $\mathcal{S}$-transform), some elements will be given in Section III.1.

II.1 Skohorod Integral with respect to the underlying process

Integral in the sense of Stieltjès. One can make use of the bounded variations property to defined an integral in the sense of Stieltjès. For any function f measurable or locally bounded or non-negative, one defines the process:

$$\left\{ (f \ * \ \omega)_t \ \overset{not}{=} \ \int_0^t \int_E f(s,z) \ \omega(\,\mathrm{d}s,\,\mathrm{d}z) \ = \ \sum_{n\geq 1} f(T_n, Z_n) \ \mathbb{I}_{\{T_n \leq t\}} \ : \ t \in [0,T] \right\}.$$

Moreover, the process N is bounded variation on each compact set thus for all measurable or locally bounded or non-negative process X one can define the process:

$$\left\{ (X \ \overset{(SL)}{*} \ N)_t \ \overset{not}{=} \ \int_0^t X_s \,\mathrm{d}N_s \ = \ \sum_{n\in\mathbb{N}^*} Z_n \ X_{T_n} \ \mathbb{I}_{\{T_n \leq t\}} \ : \ t \in [0,T] \right\}. \tag{1.2}$$

Stochastic Integral. It is possible to exploit the martingale's property to define a stochastic integral with respect to marked point process.

Hypothesis 1.4 *1. The filtration $\{\mathcal{F}_t : t \in [0,T]\}$ satisfies the "conditions habituelles de la théorie générale des processus".*

2. The process N is adapted.

3. $\mathbb{E}[N_t] \ < \ \infty$ for all t.

Property 1.4 *Assume assumptions 1.4 fulfilled. The measure ν on $[0,T]\times E$ (predictable compensator) is such that:*

- *For any $B \in \mathcal{B}(E)$, the process $\{\nu([0,t]\times B) \ : \ t \in [0,T]\}$ is predictable.*
- *For any non-negative predictable process $f \in \mathcal{L}^1(\nu)$ the process:*

$$\left\{ (f \ \overset{(SI)}{*} \ (\omega - \nu))_t \ \overset{not}{=} \ (f \ * \ \omega)_t \ - \ (f \ \overset{(L)}{*} \ \nu)_t \ : \ t \in [0,T] \right\}$$

is a $(\mathcal{F}_t)_{t\in[0,T]}$-martingale ($- \ \overset{(L)}{*} \ \nu$ stands for the Lebesgue integral with respect to the measure ν).

Theorem 1.1 *For any predictable processes $f \in \mathcal{L}^2(\nu)$, we have, for any $t \in [0,T]$ and $s \in [0,T]$:*

$$\begin{aligned}
\mathbb{E}\left[(f \ \overset{(SI)}{*} \ \tilde{\omega})_t \right] \ &= \ 0, \\
[f \ \overset{(SI)}{*} \ \tilde{\omega} \ ; \ f \ \overset{(SI)}{*} \ \tilde{\omega}]_t \ &= \ (f^2 * \mu)_t, \\
< f \ \overset{(SI)}{*} \ \tilde{\omega} \ ; \ f \ \overset{(SI)}{*} \ \tilde{\omega} >_t \ &= \ (f^2 \ \overset{(L)}{*} \ \nu)_t, \\
\mathbb{E}\left[(f \ \overset{(SI)}{*} \ \tilde{\omega})_t^2 \right] \ &= \ \mathbb{E}\left[(f^2 \ \overset{(L)}{*} \ \nu)_t \right],
\end{aligned}$$

where $\tilde{\omega} \ = \ \omega - \nu$ denotes the compensated measure.

Property 1.5 *The jump process N has for predictable compensator the process A:*

$$\left\{ A_t \ = \ \int_0^t \int_E \ z \ \nu(\,\mathrm{d}s,\,\mathrm{d}z) \ , \ t \in [0,T] \right\}.$$

The compensated jump process is $\tilde{N} \ = \ N \ - \ A$ and a statement equivalent to Theorem 1.1 can be written with $\tilde{N}$, N and A.

Integral in Skohorod sense. To define such an integral Assumption 1.4 page 16 has to be completed by the following ones:

Hypothesis 1.5 *The compensator ν writes $\nu(\mathrm{d}s, \mathrm{d}z) = \lambda(s)\mathrm{d}s\ \eta(\mathrm{d}z)$ where η is a measure of probability (distributions of the Z_k's).*

Hypothesis 1.6 *There exists $m > 0$ such that, for any $s \in [0,T]$, $\lambda(s) > m$ and $\int_0^T \lambda(s)\,\mathrm{d}s < +\infty$.*

Definition 1.4 ([6]) *A functional is said to be cylindric whenever it is of the form*

$$F = f\left(\int_0^T\!\!\int_E f_1(s)g_1(z)\omega(\,\mathrm{d}s,\ \mathrm{d}z), \ldots, \int_0^T\!\!\int_E f_n(s)g_n(z)\omega(\,\mathrm{d}s,\ \mathrm{d}z)\right),$$

where f is a bounded twice differentiable function with bounded derivatives. $f_i g_i \in \mathcal{L}^2(\nu)$ f_i is continuously differentiable with bounded derivatives for all $i = 1, \ldots, n$. We denote by $\mathcal{S}$ the set of cylindric functionals. For any functionals $F \in \mathcal{S}$ and any $h \in \mathcal{L}^2(\nu)$, the directional derivative $DF(h)$ is defined as:

$$\begin{aligned} DF(h) = \quad & -\sum_{i=1}^n \frac{\partial f}{\partial x_i}\left(\int_0^T\!\!\int_E f_1(s)g_1(z)\omega(\,\mathrm{d}s,\,\mathrm{d}z), \ldots, \int_0^T\!\!\int_E f_n(s)g_n(z)\omega(\,\mathrm{d}s,\,\mathrm{d}z)\right) \\ & \cdot \int_0^T\!\!\int_E f_i'(s)g_i(z)\left(\frac{1}{\lambda(s)}\int_0^s h(r,z)\lambda(r)\,dr\right)\omega(\,\mathrm{d}s,\,\mathrm{d}z). \end{aligned}$$

Theorem 1.2 ([6, 8])

- *$\mathcal{S}$ is dense in $\mathbb{L}^2$,*
- *For any $F, G \in \mathcal{S}$, and any $h \in \mathcal{L}^2(\nu)$, $FG \in \mathcal{S}$ and $DFG(h) = FDG(h) + GDF(h)$,*
- *For any $F \in \mathcal{S}$, and any $h \in \mathcal{L}^2(\nu)$, $\mathbb{E}\left[DF(h)\right] = \mathbb{E}\left[F\ .\ (h \overset{(SI)}{*} (\omega - \nu))_T\right]$,*
- *For any $F \in \mathcal{S}$, there exists $\nabla F \in \mathbb{L}^2([0,T] \times E \times \Omega, \nu \otimes \mathrm{d}\mathbb{P})$ measurable with respect to the three variables such that*

$$DF(h) = < \nabla F\ ,\ h >_{\mathcal{L}^2(\nu)} \qquad \forall h \in \mathcal{L}^2(\nu).$$

The results of Theorem 1.2 remains true for the stochastic gradient

$$\begin{aligned} \nabla\ :\ \mathcal{S} &\longrightarrow \mathbb{L}^2([0,T] \times E \times \Omega, \nu \otimes \mathrm{d}\mathbb{P}) \\ F &\longrightarrow \nabla F \end{aligned}$$

We show in [8] that the application $F \to \nabla F$ is closable in $\mathbb{L}^2(\mathbb{P})$. This allows us to extend the domain of the operator ∇. In fact, one introduces on $\mathcal{S}$ the norm

$$\forall F \in \mathcal{S}, \qquad ||F||_{2,1}^2 = ||F||_2^2 + \mathbb{E}\left[||\nabla F||_{\mathcal{L}^2(\nu)}^2\right]$$

and introduces the space $\mathbb{D}_{2,1}$, closure of $\mathcal{S}$ with respect to the norm $||.||_{2,1}$. ∇ is thus defined on $\mathbb{D}_{2,1}$. Theorem 1.2 remains true for $(F,G) \in (\mathbb{D}_{2,1})^2$.

Definition 1.5 *Let ζ be a $\mathcal{L}^2(\nu)$-valued random variable. It is in the domain of δ if and only if, there exists $c(\zeta)$ a constant such that for any $F \in \mathcal{S}$ we have:*

$$|\mathbb{E}\left[DF(\zeta)\right]| \leq c(\zeta)\ ||F||_2,$$

and $\delta(\zeta)$ is defined as:

$$\mathbb{E}\left[F\delta(\zeta)\right] = \mathbb{E}\left[< \nabla G, h >_{\mathcal{L}^2(\nu)}\right] = \mathbb{E}\left[DF(\zeta)\right].$$

Property 1.6 *$\mathcal{L}^2(\nu) \subset \mathrm{Dom}\,(\delta)$ and for any $h \in \mathcal{L}^2(\nu)$ we have, $\delta(h) = (h \overset{(SI)}{*} (\omega - \nu))_T$.*

Consider the space $\mathcal{S}(\mathcal{L}^2(\nu))$ of cylindrical processes of the form:

$$\zeta = \sum F_i v_i \qquad F_i \in \mathcal{S}, v_i \in \mathcal{L}^2(\nu),$$

and the derivative operator

$$D\zeta(h) = \sum DF_i(h) \otimes v_i \qquad F_i \in \mathcal{S}, v_i \in \mathcal{L}^2(\nu).$$

In order to extend the domain of δ to the closure of $\mathcal{S}(\mathcal{L}^2(\nu))$ with respect to a relevant norm, we have to control the moment of order 2 of $\delta(h)$. In fact, being in the domain yield to

$$|\mathbb{E}\left[DF(h)\right]| \;=\; |\mathbb{E}\left[< \nabla F, h >_{\mathcal{L}^2(\nu)}\right]| = \; |\mathbb{E}\left[F\delta(h)\right]| \leq \; ||\delta(h)||_2 \; ||F||_2.$$

In the Brownian setting, it is known that

$$\mathbb{E}\left[\delta(h)^2\right] \;=\; \mathbb{E}\left[||h||^2_{\mathcal{L}^2(\nu)}\right] \;+\; \mathbb{E}\left[\mathrm{trace}(\nabla h \circ \nabla h)\right] \tag{1.3}$$

but it is no more true in the Poisson setting. Instead we have shown:

Theorem 1.3 ([8]) *We have:*

$$\mathbb{E}\left[\delta(\zeta)^2\right] \;=\; \mathbb{E}\left[< \zeta \;;\; \Gamma\zeta >_{\mathcal{L}^2(\nu)}\right] \;+\; \mathbb{E}\left[\mathrm{trace}(\nabla\zeta \circ \nabla\zeta)\right],$$

where

$$\begin{array}{rccc} \Gamma & : & S(\mathcal{H}^\nu) & \to & \mathbb{L}^2(\Omega \times [0,T] \times E,\, \mathrm{d}\mathbb{P} \otimes \nu) \\ & & \zeta & \to & \sum_{i=1}^{\infty} < \zeta, \varepsilon_i >_{\mathcal{H}^\nu} \Gamma(\varepsilon_i) \end{array}$$

with $\Gamma(\varepsilon_i) = \nabla(\delta(\varepsilon_i))$ *and* $\mathcal{H}^\nu$ *the Hilbert space*

$$\mathcal{H}^\nu \;=\; \left\{ h \in \mathcal{L}^2(\nu) \;:\; \frac{\partial h}{\partial s} \in \mathcal{L}^2(\nu) \right\},$$

equipped with the inner product:

$$< g \;,\; h >_{\mathcal{H}^\nu} \;=\; < g \;,\; h >_{\mathcal{L}^2(\nu)} \;+\; < \frac{\partial g}{\partial s} \;,\; \frac{\partial h}{\partial s} >_{\mathcal{L}^2(\nu)},$$

and $\{\varepsilon_i \;:\; i \in \mathbb{N}^*\}$ *is a complete orthonormal basis on* $\mathcal{H}^\nu$.

The operator $\nabla\zeta$ is Hilbert-Schmidt thus $\mathrm{trace}(\nabla\zeta \circ \nabla\zeta)$ exists and we have:

$$\mathbb{E}\left[\mathrm{trace}(\nabla\zeta \circ \nabla\zeta)\right] \;\leq\; \mathbb{E}\left[||\nabla\zeta||^2_{\mathcal{L}^2(\nu)\otimes\mathcal{L}^2(\nu)}\right]$$

and we can consider for $\zeta \in \mathcal{S}(\mathcal{H}^\nu)$ the following norm:

$$\left[||\zeta||^{\Gamma}_{2,1}\right]^2 \;=\; \frac{1}{2}\left(\mathbb{E}\left[||\zeta||^2_{\mathcal{L}^2(\nu)}\right] + \mathbb{E}\left[||\Gamma\zeta||^2_{\mathcal{L}^2(\nu)}\right]\right) \;+\; \mathbb{E}\left[||\nabla\zeta||^2_{\mathcal{L}^2(\nu)\otimes\mathcal{L}^2(\nu)}\right]$$

and consider $\mathbb{D}_{2,1}(\mathcal{H}^\nu)$ the closure of $\mathcal{S}(\mathcal{H}^\nu)$ with respect to this norm.

Property 1.7 *We have* $\mathbb{D}_{2,1}(\mathcal{H}^\nu) \subset \mathrm{Dom}\,(\delta)$ *and the results of Theorem 1.3 remains true for* ζ *a process of* $\mathbb{D}_{2,1}(\mathcal{H}^\nu)$.

By means of this integral for marked point measure, we are able to define an integral for marked Poisson processes. In fact, given a process u (of time variable s) we denote $\tilde{u} \;:\; (s,z) \to z.u(s)$ and if $\tilde{u} \in \mathrm{Dom}\,(\delta)$, we define $\delta^{\tilde{N}}(u) \;=\; \delta(\tilde{u})$. It is easy to check that

Proposition 1.1 $\mathcal{L}^2(\tilde{\lambda}) \subset \mathrm{Dom}\left(\delta^{\tilde{N}}\right)$ *and for any* $h \in \mathcal{L}^2(\tilde{\lambda})$ *we have* $\delta^{\tilde{N}}(h) = (h \overset{(SL)}{*} \tilde{N})_T$.

It is thus natural to define the derivative operator $D^{\tilde{N}}F(u) = DF(\tilde{u})$ for all $F \in \mathcal{S}$ and the associated gradient is thus equal to:

$$\nabla^{\tilde{N}}F \;=\; \frac{1}{\int_E z^2\eta(\,\mathrm{d}z)} \int_E \nabla_{s,z}F \; z \; \eta(\,\mathrm{d}z).$$

We are then able to define for a random process the same concepts as those defined for a random measure (the Hilbert space $\mathcal{H}^{\tilde{\lambda}}$, the space of cylindrical functionals $\mathcal{S}(\mathcal{H}^{\tilde{\lambda}})$, the operator $\Gamma^{\tilde{N}}$ which allows us to define the norm $||.||^{\Gamma^{\tilde{N}}}_{2,1}$ and finally the space $\mathbb{D}^{\tilde{N}}_{2,1}(\mathcal{H}^{\tilde{\lambda}})$ by closure with respect to the norm). This space is included in the domain of $\delta^{\tilde{N}}$ and one shows that:

Theorem 1.4 ([8]) *For all* $\zeta \in \mathbb{D}^{\tilde{N}}_{2,1}(\mathcal{H}^{\tilde{\lambda}})$ *we have:*

$$\mathbb{E}\left[\delta^{\tilde{N}}(\zeta)^2\right] \;=\; \mathbb{E}\left[< \zeta \;;\; \Gamma^{\tilde{N}}\zeta >_{\mathcal{L}^2(\tilde{\lambda})}\right] \;+\; \mathbb{E}\left[\mathrm{trace}(\nabla^{\tilde{N}}\zeta \circ \nabla^{\tilde{N}}\zeta)\right].$$

It is now possible to make bridges between these integrals.

Theorem 1.5 ([8]) *The integrals* $\delta^{\tilde{N}}$, $\overset{(SL)}{*}$ *and* $\overset{(SI)}{*}$ *coincides for predictable processes.*

Theorem 1.6 ([8])

- *Let* $u \in \mathbb{D}_{2,1}(\mathcal{L}^2(\nu))$*, then we have:*

$$\delta(u) \;=\; (u \overset{(SI)}{*} (\omega - \nu))_T - \int_0^T\!\!\int_E \nabla_{s,z} u_{s,z}\; \nu(\,\mathrm{d}s,\,\mathrm{d}z)$$

- *Let* $u \in \mathbb{D}_{2,1}^{\tilde{N}}(\mathcal{L}^2(\tilde{\lambda}))$*, then we have:*

$$\delta^{\tilde{N}}(u) \;=\; (u \overset{(SL)}{*} \tilde{N})_T - \int_0^T \nabla_s^{\tilde{N}} u_s\; \tilde{\lambda}(\,\mathrm{d}s)$$

Remark 1.1 *In [8] reader can find development around covariant derivative and Weitzenböck formula.*

The aim is now to define integrals with respect to filtered Poisson processes. This process is not a martingale thus no stochastic integral can be defined. However, it remains the approach in the sense of Stieltjès since the trajectories have bounded variations. Moreover, by means of the Skohorod integral for the underlying process and of a linear operator, we can define an integral in the sense of Skohorod.

II.2 Fundamental theorem of the operator

Theorem 1.7 (Existence of the operator,[8]) *There exists a linear and continuous map* $\mathcal{K}^* : \mathcal{L}^2(\tilde{\lambda}) \to \mathcal{L}^2(\tilde{\lambda})$ *such that, for any* $t \in [0,T]$*,*

$$\mathcal{K}^*(\,\mathbb{I}_{[0,t]}) \;=\; K(t,.), \tag{1.4}$$

Moreover, denoting $\hat{\mathcal{I}}$ *the closure of the vector space embedded by* $\{\,\mathbb{I}_{[0,t]} \;:\; t \in [0,T]\}$ *with respect to the inner product:*

$$<\,\mathbb{I}_{[0,t]};\,\mathbb{I}_{[0,s]}\,>_{\hat{\mathcal{I}}} \;=\; <K(t,.);K(s,.)>_{\mathcal{L}^2(\tilde{\lambda})},$$

the application $\mathcal{K}^* \;:\; \hat{\mathcal{I}} \;\to\; \mathcal{L}^2(\tilde{\lambda})$ *is an isometry.*

Notation 1.1 ℓ *denotes the Lebesgue measure.*

Sketch of the proof.

- The operator K defined below is continuous and linear from $\mathcal{L}^2(\tilde{\lambda})$ to $\mathcal{L}^2(\ell)$.

$$\begin{aligned} K \;:\; & \mathcal{L}^2(\tilde{\lambda}) \to \mathcal{L}^2(\ell) \\ & f \to \int_0^T K(t,s) f(s) \tilde{\lambda}(\,\mathrm{d}s) \end{aligned}$$

- The operator I_{T-}^T defined below is continuous and linear from $\mathcal{L}^2(\tilde{\lambda})$ to $\mathcal{L}^2(\ell)$:

$$\begin{aligned} I_{T-}^T \;:\; & \mathcal{L}^2(\tilde{\lambda}) \to \mathcal{L}^2(\ell) \\ & f \to \int_{\cdot}^T f(s)\,\mathrm{d}s \end{aligned}$$

- For any $g \in \mathcal{L}^2(\tilde{\lambda})$, the linear form:

$$\begin{aligned} \theta_g \;:\; & \mathcal{L}^2(\tilde{\lambda}) \to \mathbb{R} \\ & f \to \int_0^T g(s) K(f)(s)\,\mathrm{d}s \end{aligned}$$

  is continuous and thus there exists an operator K^* continuous from $\mathcal{L}^2(\ell)$ to $\mathcal{L}^2(\tilde{\lambda})$ such that:

$$\theta_g \;=\; <K^*(g), f>_{\mathcal{L}^2(\tilde{\lambda})}.$$

- Finally one defines the operator:

$$\begin{aligned}\mathcal{K}^* \ : \ & \mathcal{L}^2(\tilde{\lambda}) \to \mathcal{L}^2(\tilde{\lambda}) \\ & f \to K^* \circ \left[I^T_{T-}\right]^{-1}(f)\end{aligned}$$

  which is continuous and we have, formally, for any $f \in \mathcal{L}^2(\tilde{\lambda})$:

$$\int_0^T K^*(\,\epsilon_t) f(s)\tilde{\lambda}(\,\mathrm{d}s) \ = \ K(f)(t) \ = \ \int_0^T K(t,s) f(s)\tilde{\lambda}(\,\mathrm{d}s),$$

  which shows that $K^*(\,\epsilon_t) \ = \ K(t,.)$. Moreover, we have $I^T_{T-}(\,\epsilon_t) \ = \ \mathbb{I}_{[0,t]}$. One deduces that $\mathcal{K}^*(\,\mathbb{I}_{[0,t]}) \ = \ K(t,.)$.

- The isometry is easily shown for $\mathbb{I}_{[0,t]} \in \mathcal{I}$:

$$||\,\mathbb{I}_{[0,t]}||_{\hat{\mathcal{I}}} \ = \ ||K(t,.)||_{\mathcal{L}^2(\tilde{\lambda})} \ = \ ||\mathcal{K}^*(\mathbb{I}_{[0,t]})||_{\mathcal{L}^2(\tilde{\lambda})},$$

  the result follows from a limit procedure, $\mathcal{K}^*$ being continuous.

□

Remark 1.2 *Even for predictable process u, the process $\mathcal{K}^*(u)$ is not necessary predictable.*

It is the reason why we are forced to make use of anticipative integrals.

II.3 Skohorod's Integral with respect to filtered Poisson process

Stieltjès-Lebesgue integrals with respect to pPf. In the sense of Stieltjès-Lebesgue there are two ways for defining integrals. First, we exploit Property 1.1 of pPf trajectories:

Definition 1.6 *For all locally bounded function f, one can define the following process:*

$$\left\{(f \overset{(SL)}{*} N^K)_t \ : \ t \in [0,T]\right\}.$$

Second, we exploit the Lebesgue-Stieltjès integral defined by (1.2) for the underlying Poisson process and the operator $\mathcal{K}^*$:

Definition 1.7 *For all function f such that $\mathcal{K}^*(f)$ is locally bounded, one can define the following process:*

$$\left\{(f \ * \ N^K)_t \ = \ (\mathcal{K}^*(f) \overset{(SL)}{*} N)_t \ : \ t \in [0,T]\right\}.$$

Theses integrals coincides.

Theorem 1.8 ([8]) *For any $f \in \mathcal{L}^2(\tilde{\lambda})$ locally bounded, we have:*

$$(f \overset{(SL)}{*} N^K)_t \ = \ (f \ * \ N^K)_t \qquad \forall t \in [0,T].$$

And the isometry result holds:

Theorem 1.9 ([8]) *For any $f \in \hat{\mathcal{I}}$, $g \in \hat{\mathcal{I}}$ we have:*

$$\mathbb{E}\left[(f \overset{(SL)}{*} \tilde{N}^K) \,.\, (g \overset{(SL)}{*} \tilde{N}^K)\right] \ = \ < f, g >_{\hat{\mathcal{I}}}.$$

Skohorod's Integral with respect to pPf.

Definition 1.8 *For any u such that $\mathcal{K}^*(u) \in \mathrm{Dom}\left(\delta^{\tilde{N}}\right)$ we define the integral with respect to $\tilde{N}^K$ by:*

$$\delta^{\tilde{N}^K}(u) \;=\; \delta^{\tilde{N}}(\mathcal{K}^*(u)) \;=\; \delta(z.\mathcal{K}^*(u)).$$

Definition 1.9 *For any $F \in \mathcal{S}$, for any $h \in \mathcal{L}^2(\tilde{\lambda})$ we define the directional derivative by:*

$$D^{\tilde{N}^K}F(h) \;=\; D^{\tilde{N}}F(\mathcal{K}^*(h)) \;=\; DF(z.\mathcal{K}^*(h)),$$

and thus

$$\begin{aligned} D^{\tilde{N}^K}F(h) \;=\; & -\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}\left(\int_0^T\!\!\int_E f_1(s)g_1(z)\omega(\,\mathrm{d}s,\,\mathrm{d}z),\ldots,\int_0^T\!\!\int_E f_n(s)g_n(z)\omega(\,\mathrm{d}s,\,\mathrm{d}z)\right) \\ & .\int_0^T\!\!\int_E f_i'(s)g_i(z)\left(\frac{z}{\lambda(s)}\int_0^s \mathcal{K}^*(h)(r)\lambda(r)\,dr\right)\omega(\,\mathrm{d}s,\,\mathrm{d}z). \end{aligned}$$

Property 1.8 *For any $F, G \in \mathcal{S}$ and any $h \in \mathcal{L}^2(\tilde{\lambda})$ we have:*

$$\begin{aligned} D^{\tilde{N}^K}(FG)(h) \;&=\; F.D^{\tilde{N}^K}G(h) \;+\; G.D^{\tilde{N}^K}F(h), \\ \mathbb{E}\left[D^{\tilde{N}^K}F(h)\right] \;&=\; \mathbb{E}\left[F.\delta^{\tilde{N}^K}F(h)\right]. \end{aligned}$$

Property 1.9 *There exists a $\mathcal{L}^2(\tilde{\lambda})$-valued random variable $\nabla^{\tilde{N}^K}$ called gradient map associated to $D^{\tilde{N}^K}$. Moreover, we have:*

$$\nabla^{\tilde{N}^K} \;=\; \mathcal{K}\circ\nabla^{\tilde{N}}.$$

It is easily seen that this gradient map satisfies integration by parts property and that it is the adjoint operator of $\delta^{\tilde{N}^K}$. The space which intervenes here is:

$$\mathcal{H}^K \;=\; \left\{\zeta \in \mathcal{L}^2(\tilde{\lambda}) \;:\; \mathcal{K}^*(\zeta) \in \mathcal{H}^{\tilde{\lambda}}\right\}$$

With notations now classical, one defines $\mathcal{S}(\mathcal{H}^K)$ and the operator Γ^K by:

$$\begin{array}{rcccl} \Gamma^K & : & S(\mathcal{H}^K) & \to & \mathbb{L}^2(\Omega\times[0,T]\times E,\,\mathrm{d}\mathbb{P}\otimes\nu) \\ & & \zeta & \to & \mathcal{K}\circ\Gamma^{\tilde{N}}\circ\mathcal{K}^*(\zeta) \end{array}$$

It remains to extent $\mathcal{S}(\mathcal{H}^K)$ by means of an adequate norm:

$$||\zeta||_{2,1}^{\Gamma^K} \;=\; \frac{1}{2}\left[\mathbb{E}\left[||\zeta||_{\mathcal{L}^2(\tilde{\lambda})}\right] \;+\; \mathbb{E}\left[||\Gamma^K\zeta||_{\mathcal{L}^2(\tilde{\lambda})}\right]\right] \;+\; \mathbb{E}\left[||\nabla^{\tilde{N}}(\mathcal{K}^*\zeta)||^2_{\mathcal{L}^2(\tilde{\lambda})\otimes\mathcal{L}^2(\tilde{\lambda})}\right]$$

Consider $\mathbb{D}_{2,1}(\mathcal{H}^K)$ the closure of $\mathcal{S}(\mathcal{H}^K)$ in the sense of the norm. One can then show the following property:

Theorem 1.10 ([8]) *We have $\mathbb{D}_{2,1}(\mathcal{H}^K) \subset \mathrm{Dom}\left(\delta^{\tilde{N}^K}\right)$. Moreover for all $\zeta \in \mathbb{D}_{2,1}^K(\mathcal{H}^K)$ we have:*

$$\mathbb{E}\left[\delta^{\tilde{N}^K}(\zeta)^2\right] \;=\; \mathbb{E}\left[<\zeta\;;\;\Gamma^K\zeta>_{\mathcal{L}^2(\tilde{\lambda})}\right] \;+\; \mathbb{E}\left[\mathrm{trace}(\nabla^{\tilde{N}}(\mathcal{K}^*\zeta)\circ\nabla^{\tilde{N}}(\mathcal{K}^*\zeta))\right].$$

It remains to relate both definitions of integrals.

Theorem 1.11 ([8]) *Let $u \in \mathbb{D}_{2,1}^K(\mathcal{L}^2([0,T])$, then, if the different terms converge we have the following equality:*

$$\delta^{\tilde{N}^K}(u) \;=\; (u \overset{(SL)}{*} \tilde{N}^K)_T \;-\; \int_0^T \nabla_s^{\tilde{N}^K}u_s\tilde{\lambda}(\,\mathrm{d}s)$$

We now state some properties of the integral defined in the sense of Skohorod for filtered Poisson processes.

Theorem 1.12 (Chasles's Relationship [8]) *Let $t \in [0,T]$ fixed. Consider $\mathcal{K}_t^*$ the adjoint operator of $\mathcal{K}$ in $\mathcal{L}_t^2(\tilde{\lambda}) \overset{not}{=} \mathcal{L}^2([0,t],\tilde{\lambda})$. For all $u \in \mathcal{L}_t^2(\tilde{\lambda})$, we have:*

$$\mathcal{K}_t^*(u) \;=\; \mathcal{K}_T^*(u\ \mathbb{I}_{[0,t]})\ \mathbb{I}_{[0,t]},$$

and consequently, we have:

$$\delta^{\tilde{N}}(\mathcal{K}_t^*(u)) \;-\; \delta^{\tilde{N}}(\mathcal{K}_s^*(u)) \;=\; \delta^{\tilde{N}}(\mathcal{K}^*(u\ \mathbb{I}_{]s,t]}))$$

Hypothesis 1.7 *The operator Γ is continuous from $\mathbb{L}^2(\Omega \times [0,T],\ \mathrm{d}\mathbb{P} \otimes \tilde{\lambda})$ to $\mathbb{L}^2(\Omega \times [0,T],\ \mathrm{d}\mathbb{P} \otimes \tilde{\lambda})$.*

Theorem 1.13 (Hölder's continuity of the trajectories [8]) *For any $\alpha \in [1/2, 1[$, under assumptions 1.1, 1.3 and 1.7, considering $u \in \mathbb{D}_{p,1}(\mathcal{H}) \cap \mathrm{Dom}\left(\delta^{\tilde{N}^K}\right)$ with $\alpha p > 1$, the process*

$$\{\delta^K(\mathcal{K}_t^*(u)) \;=\; \delta^{\tilde{N}^K}(u\,\mathbb{I}_{[0,t]}) \;:\; t \in [0,T]\},$$

admits a modification with $(\alpha - 1/p)$-Hölder continuous trajectories. Moreover there is a constant $c > 0$ such that:

$$||\delta^K(u)||_{\mathbb{L}^2(\Omega;\mathrm{Hol}(\alpha-1/p))} \;\leq\; c\,|||\mathcal{K}_T^*|||\,.\,||u||_{\mathbb{D}_{p,1}}$$

Theorem 1.14 (Itô's formula for cylindrical functionals [8]) *Let F a function $\mathcal{C}_b^2$ and assume that $u \in \mathcal{S}(\mathcal{L}^2(\nu))$ this means $u = F.v$ with $F \in \mathcal{S}$ and $v \in \mathcal{L}^2(\nu)$. Let $Z_t = z + (u \overset{(SL)}{*} \tilde{N}^K)_t$. Then $u.F' \circ Z$ is in $\mathrm{Dom}\left(\delta^{\tilde{N}^K}\right)$ and we have, $\mathbb{P}$-almost surely:*

$$F(Z_t) \;=\; F(z) \;+\; (u.F' \circ Z \overset{(SL)}{*} \tilde{N}^K)_t.$$

III Anticipative integral for filtered Lévy process

In this section, we consider a filtered processes defined by (1) page 5 with a deterministic kernel satisfying Assumption 1.2 and for underlying process, a Lévy process $\{L_t, t \in [0,T]\}$. We refer the reader to the book of Sato [14] for a general theory of Lévy processes.

One of the main properties of Lévy processes is the Lévy-Itô decomposition:

Theorem 1.15 (Lévy-Itô decomposition) *There exists a triplet (γ, σ^2, η) with $\gamma \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$ and η a measure on $\mathbb{R}$ called Lévy measure satisfying $\int_{\mathbb{R}} z^2\,\mathrm{d}\eta(z) < \infty$, and such that L can be represented as,*

$$L_t = \gamma t + \sigma B_t + J_t, \tag{1.5}$$

with

- *B a standard Brownian motion,*
- *for any $t \in [0,T]$,*

$$J_t := \lim_{\epsilon \to 0} \int_0^t \int_{|z|>\epsilon} z\,\mathrm{d}\tilde{N}(s,z),$$

where $\tilde{N}$ is the compensated jump measure associated to L:

$$\mathrm{d}\tilde{N}(s,z) = \mathrm{d}N(s,z) - \mathrm{d}s\ \mathrm{d}\eta(z),$$

and N is the jump measure associated to L:

$$N(E) = \mathrm{card}\,\{t \;:\; (t, \Delta L_t) \in E\} \quad \textit{for any } E \in \mathcal{B}(]0,T[\times\mathbb{R}_0),$$

where $\mathbb{R}_0 = \mathbb{R} - \{0\}$, $\Delta L_t = L_t - L_{t-}$, and $\mathrm{card}\,\{A\}$ denotes the cardinal of the set A.

The limit in (1.5) is to be understood in a.s. uniform on every bounded interval's sense.

Remark 1.3 *In the case $\eta = 0$, process L is a Brownian motion with drift γt and volatility σ. In the case $\sigma = 0$ we have a pure jump Lévy process. If, moreover, η is a finite measure we can write $\eta = \eta(\mathbb{R})Q$ with Q a probability distribution on $\mathbb{R}$. In this case the process is a compound Poisson process. Moreover, in virtue of Girsanov's Theorem, we will assume, without lost of generality, that $\gamma = 0$.*

The aim of this section is, first, to highlight the links between the different anticipative integrals we are able to defined with respect to Brownian motion, pure jump processes and Lévy processes and second, to deal with what we will call the Lévy-Itô problem. Lévy-Itô decomposition (1.5) tells us that a Lévy process can be decomposed in two components, a Brownian one and a pure jump one. Thus it is natural to wonder if this decomposition is still true for the integrals considered. One says that the Lévy-Itô problem is true if for any u in the suitable domain, we have, roughly speaking,

$$\delta^L(u) = \delta^B(u) + \delta^J(u).$$

In the Brownian setting, B^K remains a Gaussian process, thus we can define a stochastic integral that we denote $\delta_C^{B,K}$ by the use of the chaos decomposition [1]. This construction is not possible for pure jump and Lévy filtered processes because these processes are no more Lévy processes.

The $\mathcal{S}$-transform allows us to define directly an integral for filtered processes. These integrals are denoted by $\delta_S^{B,K}$ for the Brownian case and $\delta_S^{J,K}$ for the pure jump one, and have been studied in [2] and [3] respectively. For the Lévy case, we will introduce $\delta_S^{L,K}$.

Finally, the more versatile idea is to construct from K a linear operator denoted by $\mathcal{K}^*$ and to define a stochastic integral with respect to X^K, that we will denote by $\delta^{X,\mathcal{K}^*}$, from the one with respect to X denoted δ^X by:

$$\delta^{X,\mathcal{K}^*}(u) = \delta^X(\mathcal{K}^*(u))$$

As noticed in Remark 1.2 page 20, δ^X has to be defined in a anticipative way. So we have to browse the definitions of anticipative integrals with respect to X. Three main constructions can be investigated. First, we consider δ_C^X defined by the use of chaos decomposition (for Brownian motion [12], for the standard Poisson process [13] and for Lévy processes [15]). Second, we consider δ_S^X defined by the use of the $\mathcal{S}$-transform (for Brownian motion [2], for pure jump Lévy processes [3] and for general Lévy processes we have given the definition in [16]).

Remark 1.4 *Notice that we have not investigated δ_G^X the integral defined in Section II.3 as the adjoint of a stochastic gradient. As far as we know, no version of δ_G^L for a general Lévy process has been constructed. However, a direct definition as a dual operator could be introduced from the gradient operator defined in [11]. It is well known that in the Brownian case $\delta_G^B = \delta_C^B$, meanwhile, even in the more simple case developed in [4] and [8], we have $\delta_G^J(u) \neq \delta_C^J(u)$.*

Theorem 1.16 (Lévy-Itô decomposition... continuation.) *The processes B and J that appears in the Lévy Itô Decomposition 1.5 are independent.*

Is it still true for integrals ? One says that the Lévy-Itô problem is complete if moreover $\delta^B(u)$ and $\delta^J(u)$ are independent.

The results shown in [16] are briefly explained here and summarized in the Tables 1.1 and 1.2 below. The questions we deal with are: Are the integrals defined in each column equal? And is, for each line, the Lévy-Itô Problem (LIP) true and complete?

Defined by the use of	$X = L$	$X = B$	$X = J$	LIP true	LIP complete
Chaos	δ_C^L	δ_C^B	δ_C^J	Yes	No
	$\shortparallel$	$\shortparallel$	$\shortparallel$		
$\mathcal{S}$-transform	δ_S^L	δ_S^B	δ_S^J	Yes	No

Table 1.1: Different integrals for the underlying processes.

III.1 Anticipative integral for Lévy process

Integrals based on the chaos decomposition. It is well known, since Itô [10], that Lévy processes enjoy the so-called chaotic representation property in a slightly generalized form. Let us recall here the

Defined by the use of	$X = L$	$X = B$	$X = J$	LIP true	LIP complete
Intrinsic Chaos		$\delta_C^{B,K}$		-	-
		$\shortparallel$			
Intrinsic $\mathcal{S}$-transform	$\delta_S^{L,K}$	$\delta_S^{B,K}$	$\delta_S^{J,K}$	Yes	No
	$\shortparallel$	$\shortparallel$	$\shortparallel$		
$\mathcal{K}^*$ and Chaos	$\delta_C^{L,\mathcal{K}^*}$	$\delta_C^{B,\mathcal{K}^*}$	$\delta_C^{J,\mathcal{K}^*}$	Yes	No
	$\shortparallel$	$\shortparallel$	$\shortparallel$		
$\mathcal{K}^*$ and $\mathcal{S}$-transform	$\delta_S^{L,\mathcal{K}^*}$	$\delta_S^{B,\mathcal{K}^*}$	$\delta_S^{J,\mathcal{K}^*}$	Yes	No

Table 1.2: Different integrals for the filtered processes.

main ideas of this approach. For any Borel set E on $[0,T]\times\mathbb{R}$ we can define the sets $E^* = \{t\in[0,T] : (t,0)\in E\}$ and $E_0 = E - (E\cap([0,T]\times\{0\}))$ and the measure

$$\mu(E) := \sigma^2\int_{E^*} \mathrm{d}t + \int_{E_0} z^2\,\mathrm{d}(\eta\otimes\ell)(z,t).$$

Then for any set E such that $\mu(E)<\infty$ we can introduce the independent random measure

$$M(E) := \sigma\int_{E^*}\mathrm{d}B_t + \lim_{m\to\infty}\int_{E_m} z\,\mathrm{d}\widetilde{N}(t,z),$$

where $E_m = \{(t,z)\in E : \frac{1}{m} < |z| < m\}$ and where the limit is in the $\mathbb{L}^2(\Omega)$ sense. For short we can write

$$\mathrm{d}M(t,z) = 1\!\!1_{[0,T]\times\{0\}}\sigma\,\mathrm{d}B_t\,\epsilon_0(z) + 1\!\!1_{[0,T]\times\mathbb{R}_0} z\,\mathrm{d}\widetilde{N}(t,z),$$

with $\mathbb{R}_0 = \{z\in\mathbb{R},\ z\neq 0\}$.

For any collection of disjoint sets $(E_i)_{i=1,\dots,n}$ such that $\mu(E_i)<\infty$ for all $i=1,\dots,n$, we define the multiple stochastic integral $I_n^M(1\!\!1_{E_1\times\cdots\times E_n})$ of order n with respect to M by

$$I_n^M(1\!\!1_{E_1\times\cdots\times E_n}) = M(E_1)\cdots M(E_n).$$

By linearity, we extend the definition to any elementary function f of the form

$$f(\cdot) = \sum_{i_1,\dots,i_n=1}^{N} a_{i_1,\dots,i_n} 1\!\!1_{A_{i_1}\times\dots\times A_{i_n}}(\cdot),$$

where $A_1,\dots,A_N$ are pairwise disjoint Borel subsets of $[0,T]\times\mathbb{R}$ and $a_{i_1,\dots,i_n}=0$ if two of the indices $i_1,\dots,i_n$ are equal. Then, the multiple integral I_n is extended to

$$\mathcal{L}_n^2 := \mathcal{L}^2(([0,T]\times\mathbb{R})^n;\mathcal{B}(([0,T]\times\mathbb{R})^n);\mu^{\otimes n})$$

due to the fact that the space of all elementary functions is dense in $\mathcal{L}_n^2$ and the property

$$\mathbb{E}\left[I_n^M(\,\mathbb{I}_{E_1\times\cdots\times E_n})I_m^M(\,\mathbb{I}_{F_1\times\cdots\times F_m})\right] = \epsilon_n(m)\,n!\int_{([0,T]\times\mathbb{R})^n}\widetilde{\mathbb{I}_{E_1\times\cdots\times E_n}}\,\widetilde{\mathbb{I}_{F_1\times\cdots\times F_m}}\,\mathrm{d}\mu^{\otimes n},$$

where $\widetilde{f}$ is the symmetrization of the function f.

Remark 1.5 *Notice that if $E=[0,t]\times\mathbb{R}$ for $t\le T$, then*

$$M([0,t]\times\mathbb{R}) = \sigma B_t + J_t = L_t.$$

Theorem 1.17 (Chaotic representation property for Lévy processes.) *If F is a square–integrable random variable, measurable with respect to the filtration generated by L, then F has the unique representation*

$$F = \sum_{n=0}^{\infty} I_n^M(f_n),$$

where $I_0^M(f_0) = f_0 = \mathbb{E}(F)$ and f_n is a symmetric function in $\mathcal{L}_n^2$.

Given this result we can introduce the so-called annihilation and creation operators. We follow here the abstract point of view presented in [13].
On the one hand, we say that a square-integrable random variable F, given by (1.17), belongs to the domain of the annihilation operator D^M, denoted by $\mathbb{D}^M_{2,1}$, if and only if

$$\sum_{n=1}^{\infty} nn!||f_n||^2_{L^2_n} < \infty.$$

In this case we define the random field $D^M F = \{D^M_x F : x \in [0,T] \times \mathbb{R}\}$ as

$$D^M_x F = \sum_{n=1}^{\infty} n I^M_{n-1}(f_n(x,\cdot)).$$

It is well known that D^M defines a linear and closed operator from $\mathbb{L}^2(\Omega,\mathbb{P})$ into $\mathbb{L}^2(\Omega\times[0,T]\times\mathbb{R};\mathbb{P}\otimes\mu)$, with dense domain $\mathbb{D}^M_{2,1}$. Similarly we can define the iterated derivative $D^{M,n}_{x_1,\dots,x_n} = D^M_{x_1}\cdots D^M_{x_n}$ and its domain $\mathbb{D}^{n,2}_M$.
On other hand we define the creation operator δ^M_C. If u has the chaos decomposition

$$u(x) = \sum_{n=0}^{\infty} I^M_n(u_n(x,\cdot)), \quad x \in [0,T]\times\mathbb{R},$$

where $u_n \in \mathcal{L}^2_{n+1}$ is a symmetric function in the last n variables, then $\delta^M_C(u)$ is defined as

$$\delta^M_C(u) = \sum_{n=0}^{\infty} I^M_{n+1}(\tilde{u}_n),$$

provided u belongs to $\mathrm{Dom}\left(\delta^M_C\right)$, that is,

$$\sum_{n=0}^{\infty}(n+1)!||\tilde{u}_n||^2_{\mathbb{L}^2_{n+1}} < \infty.$$

It is easy to see and very well known that there is a duality relation between operators D^M and δ^M_C in the sense that if $F \in \mathbb{D}^M_{2,1}$ and $u \in \mathrm{Dom}\left(\delta^M_C\right)$ we have

$$\mathbb{E}\left[\int_{[0,T]\times\mathbb{R}} u(x) D^M_x F \,\mathrm{d}\mu(x)\right] = \mathbb{E}[\delta^M_C(u)F]. \tag{1.6}$$

So, it can be deduced that δ^M_C is also a linear and closed operator from $\mathbb{L}^2(\Omega\times[0,T]\times\mathbb{R};\mathbb{P}\otimes\mu)$ into $\mathbb{L}^2(\Omega,\mathbb{P})$, with dense domain $\mathrm{Dom}\left(\delta^M_C\right)$.

As particular cases we can consider the multiple stochastic integrals I^W_n and I^J_n defined as I^M when $\eta = 0$ and $\sigma = 0$ respectively. That is, we are denoting $\mathrm{d}W := \sigma\,\mathrm{d}B \otimes \epsilon_0$ and $\mathrm{d}J := z\,\mathrm{d}\tilde{N}$ the random independent measures on $[0,T]\times\mathbb{R}$. Moreover we can consider the corresponding operators D^W, D^J, δ^W_C and δ^J_C.
On the one hand, in the particular case that $\eta = 0$, that is $M = W$, the duality relation becomes

$$\mathbb{E}\left[\int_0^T u(t,0)\sigma^2 D^W_{t,0} F \,\mathrm{d}t\right] = \mathbb{E}[\delta^W_C(u)F].$$

where

$$u(t,z) = \sum_{n=0}^{\infty} I^W_n(u_n(t,z,\cdot)) = \sum_{n=0}^{\infty} I^{\sigma B}_n(u_n(t,z,t_1,0,t_2,0\dots,t_n,0)),$$

and

$$\delta^W_C(u) = \sum_{n=0}^{\infty} I^W_{n+1}(\tilde{u}_n) = \sum_{n=0}^{\infty} I^{\sigma B}_{n+1}(\tilde{u}_n(t,0,t_1,0,\dots,t_n,0)),$$

where $I^{\sigma B}_n$ denotes the multiple stochastic integral defined analogously on $[0,T]$ with respect to the independent random measure $\sigma\,\mathrm{d}B$. Observe that in this case $D^W_{t,0}F = D^{\sigma B}_t F = \frac{1}{\sigma} D^B_t F$.
On other hand, in the particular case $\sigma = 0$, we have

$$\mathbb{E}\left[\int_0^T \int_{\mathbb{R}_0} u(t,z) D^J_{t,z} F z^2 \,\mathrm{d}\eta(z)\,\mathrm{d}t\right] = \mathbb{E}[\delta^J_C(u)F].$$

where

$$u(t,z) = \sum_{n=0}^{\infty} I_n^J(u_n(t,z,\cdot)) = \sum_{n=0}^{\infty} I_n^{J_0}(u_n(t,z,\cdot)\,\mathbb{I}_{\mathbb{R}_0}(\cdot)),$$

and

$$\delta^J(u) = \sum_{n=0}^{\infty} I_{n+1}^J(\tilde{u}_n) = \sum_{n=0}^{\infty} I_{n+1}^{J_0}(\tilde{u}_n(t,z,t_1,z_1,\ldots,t_n,z_n)\,\mathbb{I}_{\mathbb{R}_0}(z)\prod_{i=1}^{n}\mathbb{I}_{\mathbb{R}_0}(z_i)),$$

where $I_n^{J_0}$ denotes the multiple stochastic integral defined analogously on $[0,T]\times\mathbb{R}_0$ with respect to the independent random measure $\mathrm{d}J_0 = z\,\mathrm{d}\tilde{N}$. Observe that in this case that $D_{t,z}^{J_0}F = \mathbb{I}_{\mathbb{R}_0}D_{t,z}^J F$.

The tools are now well defined and we can state the following result:

Theorem 1.18 ([16]) *If $u \in \mathrm{Dom}\left(\delta_C^M\right) \subseteq \mathbb{L}^2(\Omega\times[0,T]\times\mathbb{R})$, we have*

$$\delta_C^W(u) = \sigma\delta_C^B(u(t,0)), \qquad \text{and} \qquad \delta_C^J(u) = \delta_C^{J_0}(u(t,z)\,\mathbb{I}_{\mathbb{R}_0}).$$

Moreover, the Lévy problem is solved and true, in fact

$$\delta_C^M(u) = \delta_C^W(u) + \delta_C^J(u).$$

Remark 1.6 *In order to establish properly this result we have to introduce the canonical Lévy space. This is quite long and technique and is not useful to understand the idea of this section. For details, see [16] and references therein.*

Integrals based on the $\mathcal{S}$-transform. When X is a Brownian motion we refer to [2], when X is a pure jumps process, we refer to [3] and when $X = L$ is a Lévy process, we refer to [16]. Let us briefly explain the construction of such an integral.

Definition 1.10 *For $Y \in \mathbb{L}^2(\Omega,\mathbb{P})$, the $\mathcal{S}$-transform associated to X of Y denoted by $\mathcal{S}(Y)$ is an integral transform defined for any $\phi \in \Xi \subset \mathcal{L}^2(\mathbb{R})$ by*

$$\mathcal{S}(Y)(\phi) = \mathbb{E}_{\mathbb{Q}_\phi}[Y],$$

where

$$\mathrm{d}\mathbb{Q}_\phi = \exp^{\diamondsuit,X}\left(I_1^X(\phi)\right)\ \mathrm{d}\mathbb{P},$$

and $\exp^{\diamondsuit,X}\left(I_1^X(\phi)\right)$ denotes the Wick exponential of $I_1^X(\phi)$ associated to X defined below.

Remark 1.7 *The Wick exponential of $I_1^X(\phi)$ coincides with the Doléans-Dade exponential of $I_1^X(\phi)$.*

Definition 1.11 (Wick exponential in the Brownian setting [2]) *In the Brownian setting, we denote $\Xi = \mathcal{S}(\mathbb{R})$, the Schwartz space of smooth rapidly decreasing functions on $\mathbb{R}$ and,*

$$\exp^{\diamondsuit,B}\left(I_1^B(\phi)\right) = \exp\left(I_1^B(\phi) - \frac{||\phi||^2_{\mathcal{L}^2(\mathbb{R})}}{2}\right).$$

Definition 1.12 (Wick exponential in the pure jumps setting [3]) *In the pure jumps setting, denote $\mathcal{S}(\mathbb{R}^2)$ the Schwartz space of smooth, rapidly decreasing functions on $\mathbb{R}^2$, and consider*

$$\Xi = \left\{\phi \in \mathcal{S}(\mathbb{R}^2):\ \forall(t,z)\in\mathbb{R}\times\mathbb{R}_0,\ \phi(t,z) > -1,\ \frac{\partial\phi}{\partial z}(t,z)\bigg|_{z=0} = 0\right\},$$

and

$$\exp^{\diamondsuit,J_0}\left(I_1^{J_0}(\phi)\right) = \exp\left(\int_0^T\int_{\mathbb{R}_0}\log(1+\phi(t,z))\,\mathrm{d}N(t,z) - \int_0^T\int_{\mathbb{R}_0}\phi(t,z)\,\mathrm{d}\mu(t,z)\right).$$

Definition 1.13 (Wick exponential in the Lévy setting [16].) *In the Lévy setting we consider:*

$$\Xi = \left\{\phi \in \mathcal{S}(\mathbb{R}^2):\ \forall(t,z)\in\mathbb{R}\times\mathbb{R}_0,\ \phi(t,z) > -1,\ \frac{\partial\phi}{\partial z}(t,z)\bigg|_{z=0} = 0\right\},$$

and

$$\exp^{\diamondsuit,M}\left(I_1^M(\phi)\right) = \exp^{\diamondsuit,B}\left(\sigma I_1^B(\phi(\cdot,0)\right)\exp^{\diamondsuit,J_0}\left(I_1^{J_0}(\phi)\,\mathbb{I}_{\mathbb{R}_0}\right).$$

Theorem 1.19 ([2, 3]) *If $\mathcal{S}(Y_1)(\phi) = \mathcal{S}(Y_2)(\phi)$ for all $\phi \in \Xi$, then $Y_1 = Y_2$.*

Theorem 1.19 makes the machinery relevant to define an integral by stating that a process is perfectly described by its $\mathcal{S}$-transform. By this way, the definition of an integral writes:

Definition 1.14 ([2, 3, 16]) *Let Y be a random process. The so-called Hitsuda-Skorokhod integral of Y with respect to the process X denoted $\delta_S^X(Y)$ exists in $\mathbb{L}^2(\Omega)$ if there is a random variable $\Phi \in \mathbb{L}^2(\Omega)$ such that*

$$\mathcal{S}(\Phi)(\phi) = \int_0^T \mathcal{S}(Y(t))(\phi)\ \frac{\partial}{\partial t}\mathcal{S}(X(t))(\phi)\ \mathrm{d}t \qquad \text{for all } \phi \in \Xi. \tag{1.7}$$

Thus in the Brownian setting, it immediately yields to an integral denoted δ_S^B, in a pure jump setting, it yields to an integral denoted $\delta_S^{J_0}$ and in the Lévy setting to an integral denoted δ_S^M.

Links between these two approaches.

Theorem 1.20 ([16])

- *If u belongs to* $\mathrm{Dom}\left(\delta_C^B\right)$, *then $u \in \mathrm{Dom}\left(\delta_S^B\right)$, and $\delta_C^B(u) = \delta_S^B(u)$.*
- *If u belongs to* $\mathrm{Dom}\left(\delta_C^{J_0}\right)$, *then $u \in \mathrm{Dom}\left(\delta_S^{J_0}\right)$, and $\delta_C^{J_0}(u) = \delta_S^{J_0}(u)$.*
- *If u belongs to* $\mathrm{Dom}\left(\delta_C^M\right)$, *then $u \in \mathrm{Dom}\left(\delta_S^M\right)$, and $\delta_C^M(u) = \delta_S^M(u)$.*

Theorem 1.21 ([16]) *The Lévy problem is solved and true. Indeed, for any $u \in \mathrm{Dom}\left(\delta_C^M\right)$, we have:*

$$\delta_S^M(u) = \delta_S^{\sigma B}(u(t,0)) + \delta_S^{J_0}(u\, \mathbb{I}_{\mathbb{R}_0}). \tag{1.8}$$

The Complete Lévy-Itô problem.

Remark 1.8 *It is easily seen that, whatever the setting (Brownian, pure jump and Lévy), the complete Lévy-Itô problem is true if and only if u is deterministic.*

III.2 Anticipative integral for filtered Lévy process

Intrinsic definitions for filtered Brownian motion. The filtered Brownian motion is an isonormal Gaussian process. In fact, let $\mathcal{E}$ be the set of step functions on $[0,T]$. We define the Hilbert space $\mathcal{H}$ as the closure of $\mathcal{E}$ with respect to the inner product

$$\left\langle \mathbb{I}_{[0,s]},\, \mathbb{I}_{[0,t]} \right\rangle_{\mathcal{H}} := \int_0^{t\wedge s} K(s,u)\ K(t,u)\ \mathrm{d}u, \quad \forall (s,t) \in [0,T[^2.$$

We let B^K be the map $B^K(\mathbb{I}_{[0,t]}) := B_t^K$ for every $t \in [0,T]$. The map B^K is a linear isometry from $\mathcal{H}$ to the set $\{B^K(\phi),\ \phi \in \mathcal{H}\}$. This family of random variables is an isonormal Gaussian process and it is possible to define an intrinsic integral by means of chaos decomposition denoted by $\delta_C^{B,K}$. We refer to [12] for details.

We consider in what follows that $X = B, J, L$.

By the use of the $\mathcal{S}$-transform.

Hypothesis 1.8 *Suppose that the mapping $t \to \mathcal{S}\left(X^K(t)\right)(\phi)$ is differentiable for every $\phi \in \Xi$.*

Remark 1.9 *[3, Lemma 5.1] gives assumptions on the kernel for which this assumption is fulfilled.*

Definition 1.15 ([2, 3, 16]) *Suppose $H \subset \mathbb{R}$ is a Borel set and $Y : H \times \Omega \to \mathbb{R}$ is a measurable stochastic process such that $Y(t)$ is square-integrable for each $t \in H$. Y is said to have a Hitsuda-Skorokhod integral with respect to X^K if*

- *for any $\phi \in \Xi$:*
$$\mathcal{S}\left(Y(\cdot)\right)(\phi)\frac{\partial}{\partial t}\mathcal{S}\left(X^K(\cdot)\right)(\phi) \in \mathbb{L}^1(H),$$
- *there is a $\Phi \in \mathbb{L}^2(\Omega)$ such that for any $\phi \in \Xi$:*
$$\mathcal{S}\left(\Phi\right)(\phi) = \int_H \mathcal{S}\left(Y(t)\right)(\phi)\frac{\partial}{\partial t}\mathcal{S}\left(X^K(t)\right)(\phi)\,\mathrm{d}t. \tag{1.9}$$

By Theorem 1.19, Φ is unique and will be denoted $\delta_S^{X,K}(Y)$.

By the use of an operator. Consider now the approach of Section II.3. Here the situation is easier since the measure is $\nu(\mathrm{d}s,\,\mathrm{d}z) = \mathrm{d}s\ \eta(\mathrm{d}z)$ and thus the linear operator for the pure jumps component $\mathcal{K}^*$ is defined from $\mathcal{L}^2([0,T])$ to $\mathcal{L}^2([0,T])$ and is exactly the same as the one for the Brownian component. This allows us to construct an operator for the Lévy setting, still denoted $\mathcal{K}^*$. It is now easy to define an integral with respect the filtered process by means of an integral with respect to the underlying process: for any u such that $\mathcal{K}^*(u) \in \mathrm{Dom}\left(\delta_I^X\right)$, $I = C, S$, $X = B, J, L$,

$$\delta_I^{X,\mathcal{K}^*}(u) = \delta_I^X(\mathcal{K}^*(u)).$$

Relationship between these integrals.

Theorem 1.22 ([12]) *For u such that $u \in \mathrm{Dom}\ \delta_C^{B,K} \cap \mathrm{Dom}\ \delta_C^{B,\mathcal{K}^*}$, $\delta_C^{B,K}(u) = \delta_C^{B,\mathcal{K}^*}(u)$.*

Theorem 1.23 ([16]) *For u such that $\mathcal{K}^*(u) \in \mathrm{Dom}\left(\delta_C^X\right)$, $\delta_C^{X,\mathcal{K}^*}(u) = \delta_S^{X,\mathcal{K}^*}(u)$, $X = B, J, L$.*

Theorem 1.24 ([16]) *For u such that $u \in \mathrm{Dom}\ \delta_C^{X,\mathcal{K}^*} \cap \mathrm{Dom}\ \delta_S^{X,K}$, $\delta_S^{X,\mathcal{K}^*}(u) = \delta_S^{X,K}(u)$, $X = B, J, L$.*

Finally it is now easy to show the Lévy-Itô Problem for filtered processes.

Theorem 1.25

1. *For any $u \in \mathrm{Dom}\left(\delta_C^{L,\mathcal{K}^*}\right)$, $\delta_C^{L,\mathcal{K}^*}(u) = \delta_C^{B,\mathcal{K}^*}(u) + \delta_C^{J,\mathcal{K}^*}(u)$.*
2. *For any $u \in \mathrm{Dom}\left(\delta_S^{L,\mathcal{K}^*}\right)$, $\delta_S^{L,\mathcal{K}^*}(u) = \delta_S^{B,\mathcal{K}^*}(u) + \delta_S^{J,\mathcal{K}^*}(u)$.*
3. *For any $u \in \mathrm{Dom}\left(\delta_S^{L,K}\right)$, $\delta_S^{L,K}(u) = \delta_S^{B,K}(u) + \delta_S^{J,K}(u)$.*

The filtered process and its underlying process has the same filtration this is shown in [9, Theorem 4.8] for Brownian motion and in [8, Theorem 17] for Poisson process and the proof can be easily extended to Lévy process. Thus the Remark 1.8 extends to filtered processes.

Bibliography

[1] Elisa Alòs, Olivier Mazet, and David Nualart. Stochastic calculus with respect to Gaussian processes. *Ann. Probab.*, 29(2):766–801, 2001.

[2] Christian Bender. An S-transform approach to integration with respect to a fractional Brownian motion. *Bernoulli*, 9(6):955–983, 2003.

[3] Christian Bender and Tina Marquardt. Stochastic calculus for convoluted Lévy processes. *Bernoulli*, 14(2), 2008.

[4] Eric A. Carlen and Étienne Pardoux. Differential calculus and integration by parts on Poisson space. In *Stochastics, algebra and analysis in classical and quantum dynamics (Marseille, 1988)*, volume 59 of *Math. Appl.*, pages 63–73. Kluwer Acad. Publ., Dordrecht, 1990.

[5] Laure Coutin and Laurent Decreusefond. Abstract nonlinear filtering theory in the presence of fractional Brownian motion. *Ann. Appl. Probab.*, 9(4):1058–1090, 1999.

[6] Laurent Decreusefond. Perturbation analysis and Malliavin calculus. *Annals of Applied Probability*, 8(2):496–523, 1998.

[7] Laurent Decreusefond. Stochastic calculus with respect to fractional Brownian motion. In *Long Range Dependence : theory and applications.* Murad S. Taqqu, 2000.

[8] Laurent Decreusefond and **Nicolas Savy**. Anticipative calculus with respect to filtered Poisson processes. *Annales Institut Henri Poincaré Probabilités Statistiques*, 42(3):343–372, 2006.

[9] Laurent Decreusefond and Ali S. Üstünel. The Beneš equation and stochastic calculus of variations. *Stochastic Process. Appl.*, 57(2):273–284, 1995.

[10] Kiyosi Itô. Spectral type of the shift transformation of differential processes with stationary increments. *Trans. Amer. Math. Soc.*, 81:253–263, 1956.

[11] Jorge Léon, Josep L. Solé, Frederic Utzet, and Josep Vives. Local Malliavin calculus for Lévy processes and applications. Preprint, 2012.

[12] David Nualart. *The Malliavin calculus and related topics*. Probability and its Applications (New York). Springer-Verlag, New York, 1995.

[13] David Nualart and Josep Vives. Anticipative calculus for the Poisson process based on the Fock space. In *Séminaire de Probabilités, XXIV, 1988/89*, volume 1426 of *Lecture Notes in Math.*, pages 154–165. Springer, Berlin, 1990.

[14] Ken-iti Sato. *Lévy processes and infinitely divisible distributions*, volume 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. Translated from the 1990 Japanese original, Revised by the author.

[15] Josep L. Solé, Frederic Utzet, and Josep Vives. Canonical Lévy process and Malliavin calculus. *Stochastic Process. Appl.*, 117(2):165–187, 2007.

[16] **Nicolas Savy** and Josep Vives. Anticipative integrals with respect to a filtered Lévy process and Lévy-Itô decomposition. *Journal of Theoretical Probability*, 2014. Submitted.

Chapter 2

A limit theorem for filtered Poisson processes.*

In this section, we relate both processes in heart of my works: Brownian Volterra processes and Filtered Poisson processes. We show the weak convergence in a Hölder space, of a sequence of filtered Poisson processes to a Brownian Volterra process when the intensity of the underlying Poisson process tends to infinity.

In a first section, we introduce the ingredients, namely the notion of Hilbertian martingale and the so-called technique of radonification. In a second section, we present the result of convergence of Hilbertian martingales. Finally, we use a projective version of the last convergence result to answer to the question.

I Hilbertian martingale and Radonification

Definition 2.1 ([7]) *Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ a filtered probability space. Let V a separable Hilbert space. A V-valued process X is a $\mathcal{F}$-Hilbertian martingale if and only if:*

$$\begin{aligned} \mathbb{E}\left[\|X_t\|_V\right] &< \infty && \text{for any } t, \\ \mathbb{E}\left[X_t \,|\, \mathcal{F}_s\right] &= X_s, \quad \mathbb{P} \text{ p.s.} && \text{for any } s \geq t. \end{aligned}$$

The analogue of the square bracket is here denoted $\lessdot X \gtrdot$, and defined as the unique predictable process with finite variation with values in the space of positive symmetric nuclear operators from V into V, such that, for u, $v \in V$,

$$\{< X_t, u >_V < X_t, v >_V - < \lessdot X \gtrdot_t u, v >_V, \ t \geq 0\}$$

is a martingale. Since $\lessdot X \gtrdot$, is also a Hilbert-Schmidt operator, we can consider its square root, denoted by $\lessdot X \gtrdot_t^{1/2}$. It is a Hilbert-Schmidt operator because we are dealing with a non-negative definite operator of trace class. We denote by $\mathcal{L}_2(V;V)$, the space of Hilbert-Schmidt maps from V into V. A key result for proving weak convergence of a sequence of Hilbertian martingales is the following one:

Proposition 2.1 ([10]) *Let $\{X^n : n \in \mathbb{N}^*\}$ be a sequence of càdlàg V-valued processes. Then the distributions of the processes $\{X^n, n \geq 1\}$ form a tight sequence of probabilities on $\mathcal{D}(\mathbb{R}^+, V)$ - the Skohorod space of càdlàg functions defined on $\mathbb{R}^+$ with values in V - if the following assumptions are fulfilled:*

1. *For each rational $q \in [0,1]$, the family of random variables $\{X^n(q), n \geq 1\}$ is tight.*
2. *There exists $p > 0$ and a sequence of processes $\{A^n(\delta) : n \in \mathbb{N}^*, \ \delta \in]0,1[\}$ such that:*

$$\mathbb{E}\left[||X^n(t+\delta) - X^n(t)||_V^p \mid \mathcal{F}_t\right] \leq \mathbb{E}\left[A^n(\delta) \mid \mathcal{F}_t\right], \qquad \text{and} \qquad \lim_{\delta\to 0} \limsup_{n\to\infty} \mathbb{E}\left[A^n(\delta)\right] = 0.$$

* Publication related to this chapter:

[2] Laurent Decreusefond and **Nicolas Savy**. Filtered Brownian motions as weak limit of filtered Poisson processes. *Bernoulli*, 11(2):283–292, 2005.

On the one hand, beyond the trivial examples of V-valued Brownian motion or diffusions, it is rather hard to determine whether a V-valued process is a V-valued martingale. On the other hand, it is very easy to see if it is a cylindrical martingale, that is, if $\{< X_t,\, u >_V,\, t \geq 0\}$ is a real-valued martingale for any $u \in V$. The following "radonification" result is thus of paramount interest:

Theorem 2.1 ([1, 9, 4]) *Let E and F two Hilbert spaces and consider $u : E \to F$ a Hilbert-Schmidt operator. Denote $\mathcal{M}_c([0,1],\mathbb{R})$ the space of continuous real-valued martingales space equipped with the norm*

$$\|M\|^2_{\mathcal{M}_c([0,1],\mathbb{R})} = \mathbb{E}\left[\sup_{t\in[0,1]} |M_s|^2\right].$$

If L belongs to $\mathcal{L}(E^*;\mathcal{M}_c([0,1],\mathbb{R}))$, the space of continuous linear applications from the dual of E, denoted E^*, to $\mathcal{M}_c([0,1],\mathbb{R}))$, then $u \circ L$ is a continuous F-valued martingale.

II Radonification of martingales associated to filtered processes

Theorem 2.2 ([2]) *Let M a continuous martingale such that $< M >_t = ct$ for any $t \in [0,1]$. Let K satisfying Assumption 1.3 page 15. Then, for all $\Phi \in (\mathcal{I}_{\alpha+1/2,2})^*$,*

$$\left\{\mathfrak{Z}^M_t(\Phi) := \int_0^t K^*\Phi(s)\,\mathrm{d}M_s, t \in [0,1]\right\}$$

in a continuous martingale. Moreover, for any $\varepsilon \in (0,\alpha]$, there is a $\mathcal{I}_{\alpha-\varepsilon,2}$-valued martingale denoted $\mathfrak{X}^M$ such that, for any $\Phi \in (\mathcal{I}_{\alpha-\varepsilon,2})^*$ we have:

$$\mathfrak{Z}^M_t(\Phi) =< \Phi, \mathfrak{X}^M_t >_{(\mathcal{I}_{\alpha-\varepsilon,2})^*,\mathcal{I}_{\alpha-\varepsilon,2}}$$

Sketch of the proof. Fix $\varepsilon \in (0,\alpha]$. Consider the linear application

$$\begin{array}{rccc} L: & (\mathcal{I}_{\alpha+1/2,2})^* & \longrightarrow & \mathcal{M}_c([0,1],\mathbb{R}) \\ & \Phi & \longrightarrow & \left\{\mathfrak{Z}^M_t(\Phi), t \in [0,1]\right\}. \end{array}$$

One shows that L is in $\mathcal{L}((\mathcal{I}_{\alpha+1/2,2})^*, \mathcal{M}_c([0,1],\mathbb{R}))$. Since the embedding from $\mathcal{I}_{\alpha+1/2,2}$ into $\mathcal{I}_{\beta+1/2,2}$ is Hilbert-Schmidt for $\beta < \alpha - 1/2$, the result is a consequence of Theorem 2.1. □

Remark 2.1 *In the published version, [2] results are shown in the context of semi-martingales but for the sake of simplicity, I prefer to restrain to the martingales setting.*

The key point is the following result which establishes a link with our filtered processes.

Lemma 2.1 *Consider $\alpha > \frac{1}{2}$. For any $t \in [0,1]$, $\mathfrak{Z}^M_t(\epsilon_t)$ is well defined and writes:*

$$\mathfrak{Z}^M_t(\epsilon_t) \;=\; < \epsilon_t,\, \mathfrak{X}^M_t >_{(\mathcal{I}_{\alpha-\varepsilon,2})^*,\mathcal{I}_{\alpha-\varepsilon,2}} \;=\; \int_0^t K(t,s)\,\mathrm{d}M_s.$$

Sketch of the proof. The results comes from some arguments of stochastic calculus and noticing that $\epsilon_t \in (\mathcal{I}_{\alpha+1/2,2})^*$. In fact, it is known ([3, 8]) that:

$$\mathcal{I}_{\alpha+1/2,2} \;\subset\; \mathcal{I}_{\alpha-\varepsilon,2} \;\subset\; \mathrm{Hol}(\alpha-\varepsilon-1/2),$$

where $\mathrm{Hol}(\nu)$ denotes the space of Hölder–continuous, null in zero functions equipped with the norm:

$$||f||_{\mathrm{Hol}(\nu)} = \sup_{t\neq s} \frac{|f(t)-f(s)|}{|t-s|^\nu}.$$

Thus

$$(\mathrm{Hol}(\alpha-\varepsilon-1/2))^* \;\subset\; (\mathcal{I}_{\alpha-\varepsilon,2})^* \;\subset\; (\mathcal{I}_{\alpha+1/2,2})^*.$$

As $\alpha - \varepsilon - 1/2 > 0$ for ε small enough, $\epsilon_t \in (\mathrm{Hol}(\alpha-\varepsilon-1/2))^*$ and thus $\epsilon_t \in (\mathcal{I}_{\alpha+1/2,2})^*$. □

Remark 2.2 *On the one hand, for $\alpha \leq 1/2$, ϵ_t is no more in $(\mathcal{I}_{\alpha-\varepsilon,2})^*$ and $\mathfrak{Z}^M_t(\epsilon_t)$ is meaningless. On the other hand, we know that, for $K(t,s) = (t-s)^{\alpha-1/2}$ and for a Poisson process M, when $\alpha < 1/2$, $\int_0^\cdot K(.,s)\,\mathrm{d}M_s$ is a process which jumps to infinity as soon as the Poisson process jumps. However, $\varepsilon^{-1}\int_{t-\varepsilon}^{t+\varepsilon} \mathfrak{Z}^M_t(s)\,\mathrm{d}s$ is perfectly defined and can substitute $\int_0^t K(t,s)\,\mathrm{d}M_s$ for ε small enough.*

III Convergence of Hilbertian martingales

Here $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$ is the space of Poisson defined by 1.1 page 13 which compensator is λ ds. For any $\lambda \in \mathbb{N}^*$, $N_s^\lambda = \omega([0,s])$. One denotes $\tilde{N}^\lambda$ the process N^λ compensated. Finally X^λ, is defined as:

$$X_t^\lambda = \frac{1}{\sqrt{\lambda}} \int_0^t K(t,s)(\,\mathrm{d}N_s^\lambda - \lambda\,\mathrm{d}s).$$

Theorem 2.3 ([2]) *Consider K satisfying Assumptions 1.1 and 1.3. Consider $\mathfrak{X}^{\hat{N}^\lambda}$ and $\mathfrak{X}^B$ the $\mathcal{I}_{\alpha-\varepsilon,2}$-valued martingales defined by Theorem 2.2 from the martingales $\hat{N}^\lambda$ and B. When λ tends to infinity, the distribution of $\mathfrak{X}^{\hat{N}^\lambda}$ in $\mathcal{D}([0,1], \mathcal{I}_{\alpha-\varepsilon,2})$ converges to the distribution of $\mathfrak{X}^B$.*

Sketch of the proof.

- First, one shows that the sequence $\{\mathfrak{X}^{\hat{N}^\lambda} : \lambda \geq 1\}$ is tight in $\mathcal{D}([0,1], \mathcal{I}_{\alpha-\varepsilon,2})$. It is an application of Proposition 2.1 noticing that

$$< \langle \mathfrak{X}^{\hat{N}^\lambda} \rangle_t u,\, v >= \int_0^t K^*(u)(s) K^*(v)(s)\,\mathrm{d}s.$$

- Second, let $\{\mathfrak{X}^{\lambda_k} : k \geq 1\}$ a sub-sequence which converges to a limit say L. We have to show that for any $u \in (\mathcal{I}_{\alpha-\varepsilon,2})^*$, $< u,\, L >=< u,\, \mathfrak{X}^B >$. For this, it is enough to apply stochastic integrals convergence Theorem [5, 6] since in virtue of Lemma 2.1,

$$< u, \mathfrak{X}^{\lambda_k} >_{(\mathcal{I}_{\alpha-\varepsilon,2})^*, \mathcal{I}_{\alpha-\varepsilon,2}} = \int_0^\cdot K^* u(s)\,\mathrm{d}\hat{N}_s^{\lambda_k}.$$

- Finally, it is shown that all convergent sub-sequences have the same limit. Thus distributions of $\mathfrak{X}^{\hat{N}^\lambda}$ in $\mathcal{D}([0,1], \mathcal{I}_{\alpha-\varepsilon,2})$ converge to the distributions of $\mathfrak{X}^B$.

□

IV Back to the initial problem

Corollary 2.1 *Under Assumption 1.3 with $\alpha > \frac{1}{2}$ and 1.1, we have:*

$$\left\{ X_t^\lambda = \int_0^t K(t,s)\,\mathrm{d}\hat{N}_s^\lambda \;:\; t \in [0,1] \right\} \xrightarrow[\lambda \to \infty]{\mathcal{L}(\mathrm{Hol}(\alpha - 1/2 - \varepsilon))} \left\{ X_t = \int_0^t K(t,s)\,\mathrm{d}B_s \;:\; t \in [0,1] \right\}.$$

Sketch of the proof. The application

$$\begin{array}{rccl} B: & \mathcal{I}_{\alpha-\varepsilon,2} & \longrightarrow & \mathrm{Hol}(\alpha - 1/2 - \varepsilon) \\ & f & \longrightarrow & (s \mapsto f(s) =< \epsilon_s\,,\, f >_{(\mathcal{I}_{\alpha-\varepsilon,2})^*, \mathcal{I}_{\alpha-\varepsilon,2}}), \end{array}$$

is well defined ($\alpha > 1/2$, $\mathcal{I}_{\alpha-\varepsilon,2}$ is a sub-space of continuous functions and its dual contains the Dirac measures) and is continuous. Thus, for any F bounded and continuous from $\mathrm{Hol}(\alpha - 1/2 - \varepsilon)$ to $\mathbb{R}$, $F \circ B$ is a continuous from $\mathcal{I}_{\alpha-\varepsilon,2}$ to $\mathbb{R}$. An application of Theorem 2.3 yield to:

$$\mathbb{E}\left[F \circ B(\mathfrak{X}^{\hat{N}^\lambda})\right] \underset{n \to \infty}{\longrightarrow} \mathbb{E}\left[F \circ B(\mathfrak{X}^B)\right],$$

and thus

$$\mathbb{E}\left[F(X^n)\right] \underset{n \to \infty}{\longrightarrow} \mathbb{E}\left[F(X)\right].$$

□

When $\alpha < 1/2$, the function $s \mapsto f(s) =< \epsilon_s \, , \, f >_{(\mathcal{I}_{\alpha-\varepsilon,2})^*, \mathcal{I}_{\alpha-\varepsilon,2}}$ is no more defined for $f \in \mathcal{I}_{\alpha-\varepsilon,2}$.

Corollary 2.2 *Under Assumptions 1.1 and 1.3 with* $0 < \alpha < \frac{1}{2}$*, considering* η *continuous from* $[0,1]$ *to* $(\mathcal{I}_{\alpha-\varepsilon,2})^*$*, we have:*

$$\left\{< \eta_t, \mathfrak{X}_t^n >_{(\mathcal{I}_{\alpha-\varepsilon,2})^*, \mathcal{I}_{\alpha-\varepsilon,2}} \ : \ t \in [0,1]\right\} \xrightarrow[\lambda \to \infty]{\mathcal{L}(\mathcal{C}([0,1];\mathbb{R}))} \left\{< \eta_t, \mathfrak{X}_t >_{(\mathcal{I}_{\alpha-\varepsilon,2})^*, \mathcal{I}_{\alpha-\varepsilon,2}} \ : \ t \in [0,1]\right\}$$

Remark 2.3 *We can choose* η *in such a way that:*

$$< \eta_t, f >_{(\mathcal{I}_{\alpha-\varepsilon,2})^*, \mathcal{I}_{\alpha-\varepsilon,2}} = \varepsilon^{-1} \int_{(t-\varepsilon)\vee 0}^{(t+\varepsilon)\wedge 1} f(s) \, \mathrm{d}s = \varepsilon^{-1}(I_{0+}^1 f((t+\varepsilon)\wedge 1) - I_{0+}^1 f((t-\varepsilon)\vee 0)).$$

Since $f \in \mathcal{I}_{\alpha-\varepsilon,2}$*,* $I_{0+}^1 f$ *is in* $\mathcal{I}_{1+\alpha-\epsilon}$ *sub-space of* $\mathrm{Hol}(1/2+\alpha-\epsilon)$*. It is thus clear that* η *is continuous from* $[0,1]$ *to* $\mathcal{I}^*_{\alpha-\varepsilon,2}$*. The weak convergence result follows.*

Bibliography

[1] Albert Badrikian and Ali S. Üstünel. Radonification of cylindrical semimartingales on Hilbert spaces. *Ann. Math. Blaise Pascal*, 3(1):13–21, 1996.

[2] Laurent Decreusefond and **Nicolas Savy**. Filtered Brownian motions as weak limit of filtered Poisson processes. *Bernoulli*, 11(2):283–292, 2005.

[3] Denis Feyel and Arnaud de La Pradelle. On fractional Brownian processes. *Potential Anal.*, 10(3):273–288, 1999.

[4] Adam Jakubowski, Stanislaw. Kwapień, Paul Raynaud de Fitte, and Jan Rosiński. Radonification of cylindrical semimartingales by a single Hilbert-Schmidt operator. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.*, 5(3):429–440, 2002.

[5] Adam Jakubowski, Jean Mémin, and Gilles Pagès. Convergence en loi des suites d'intégrales stochastiques sur l'espace $\mathbf{d}^1$ de Skorokhod. *Probab. Theory Related Fields*, 81(1):111–137, 1989.

[6] Thomas G. Kurtz and Philip E. Protter. Weak convergence of stochastic integrals and differential equations. In *Probabilistic models for nonlinear partial differential equations (Montecatini Terme, 1995)*, pages 1–41. Springer, Berlin, 1996.

[7] Michel Métivier. *Stochastic partial differential equations in infinite-dimensional spaces.* Scuola Normale Superiore, Pisa, 1988.

[8] Stefan G. Samko, Anatoly A. Kilbas, and Oleg I. Marichev. *Fractional integrals and derivatives.* Gordon and Breach Science Publishers, Yverdon, 1993. Theory and applications, Edited and with a foreword by S. M. Nikolskii, Translated from the 1987 Russian original, Revised by the authors.

[9] Laurent Schwartz. Semi-martingales banachiques: le théorème des trois opérateurs. In *Séminaire de Probabilités, XXVIII*, pages 1–20. Springer, Berlin, 1994.

[10] John B. Walsh. An introduction to stochastic partial differential equations. In *École d'été de probabilités de Saint-Flour, XIV—1984*, volume 1180 of *Lecture Notes in Math.*, pages 265–439. Springer, Berlin, 1986.

Chapter 3

Transportation inequality on the configurations space and Malliavin Calculus*

Let X be a Polish space and ρ a lower semi-continuous distance on $X \times X$, which does not necessarily generate the topology on X. Given two probability measures μ and ν on X, the optimal transportation problem associated to ρ consists in evaluating the distance

$$\mathcal{T}_\rho(\mu,\nu) = \inf_{\gamma\in\Sigma(\mu,\nu)} \int_X \int_X \rho(x,y)\,\mathrm{d}\gamma(x,y), \tag{3.1}$$

where $\Sigma(\mu,\nu)$ is the set of probability measures on $X \times X$ with first (respectively second) marginal μ (respectively ν). The aim of this section is to deal with the following formula due to Feyel and Üstünel [6, Theorem 3.2] which states that

$$\mathcal{T}_{|\cdot|_H}(L.\mu,\mu) \leq \mathbb{E}\left[\left|(\mathcal{I}+\mathcal{L})^{-1}\nabla L\right|_H\right] \tag{3.2}$$

where L is the density of an absolutely continuous probability measure $\nu = L.\mu$ with respect to the Wiener measure μ and $|.|_H$ stands for the norm on the Cameron-Martin space. This result relate transport inequalities (by means of Rubinstein's distance $\mathcal{T}_{|\cdot|_H}$) and stochastic analysis in Wiener space (by means of gradient ∇ and operator $\mathcal{L}$).

Here we focus on a different setting, the space of configurations on a σ-compact metric space. The main difference between this space equipped with a Poisson measure and the Wiener space is that there exists at least two way to define stochastic gradients and the definitions do not coincide. First of all, we introduce the main ingredients to deal with formula (3.2) in the configurations space setting. The points of interest are to define distances which are lower semi-continuous and to define couples Gradient / Distance such that the so-called Rademacher's property is true. Then we have the ingredients to establish the counterpart of (3.2) in configurations space setting and to relate gradients and distances. Finally we give some examples: distance between processes and isoperimetric tails.

I Ingredients

I.1 Transport inequalities

There exists at least one probability measure γ for which the infimum in (3.1) is attained [14, Theorem 4.1]. According to the celebrated Kantorovitch-Rubinstein duality theorem, [14, Theorem 5.10], this minimum is equal to

$$\mathcal{T}_\rho(\mu,\nu) = \sup_{\substack{F\in\rho-\mathrm{Lip}_1\\ F\in\mathcal{L}^1(\mu+\nu)}} \int_X F\,\mathrm{d}(\mu-\nu), \tag{3.3}$$

* Publication related to this chapter:

[5] Laurent Decreusefond, Aldéric Joulin and **Nicolas Savy**. Upper bounds on Rubinstein distances on configuration spaces and applications, *Communication on Stochastic Analysis and Applications*, 2010, 4, 377-399.

where $\rho - \mathrm{Lip}_m$ is the set of bounded Lipschitz continuous functions F from X to $\mathbb{R}$ with Lipschitz constant m:

$$|F(x) - F(y)| \leq m\rho(x, y), \quad x, y \in X.$$

In the context of optimal transportation, $\mathcal{T}_\rho$ is considered as a Rubinstein distance since the cost function is already a distance (see for instance the bibliographical notes in [14, end of Chapter 6]).

I.2 Configurations space

In [5], we have considered the situation where $X = \Gamma_\Lambda$ is the configuration space on a σ-compact metric space Λ with Borel σ-algebra $\mathcal{B}(\Lambda)$, i.e.,

$$\Gamma_\Lambda = \{\omega \subset \Lambda : \omega \cap K \text{ is a finite set for every compact } K \in \mathcal{B}(\Lambda)\}.$$

Here the σ-compactness means that Λ can be partitioned into the union of countably many compact subspaces. We identify $\omega \in \Gamma_\Lambda$ and the positive Radon measure $\sum_{x\in\omega} \varepsilon_x$. Γ_Λ is endowed with the vague topology, i.e., the weakest topology such that for all $f \in \mathcal{C}_0(\Lambda)$ (continuous with compact support on Λ), the following maps

$$\omega \mapsto \int_\Lambda f \, \mathrm{d}\omega = \sum_{x\in\omega} f(x),$$

are continuous. When f is the indicator function of a subset B, we will use the shorter notation $\omega(B)$ for the integral of $\mathbb{I}_B$ with respect to ω. We denote by $\mathcal{B}(\Gamma_\Lambda)$ the corresponding Borel σ-algebra. Let $\mathfrak{M}(\Lambda)$ be the space of positive and diffuse Radon measures on $\mathcal{B}(\Lambda)$ endowed with the corresponding Borel σ-field and equipped with the topology of vague convergence. Given a measure $\sigma \in \mathfrak{M}(\Lambda)$, the probability space under consideration in the remainder of this paper will be the Poisson space $(\Gamma_\Lambda, \mathcal{B}(\Gamma_\Lambda), \mu_\sigma)$, where μ_σ is the Poisson measure of intensity σ, i.e., the probability measure on Γ_Λ fully characterized by

$$\mathbb{E}_{\mu_\sigma}\left[\exp\left(\int_\Lambda f \, \mathrm{d}\omega\right)\right] = \exp\left\{\int_\Lambda (e^f - 1) \, \mathrm{d}\sigma\right\},$$

for all $f \in \mathcal{C}_0(\Lambda)$. Here $\mathbb{E}_{\mu_\sigma}$ stands for the expectation under the measure μ_σ.
Actually, several distance concepts are available between elements of the configuration space Γ_Λ, see for instance [13] for a thorough discussion about this topic. We introduce only three of them which will be useful in the sequel. Let ω and η be two configurations in Γ_Λ.

Trivial distance The trivial distance is simply given by

$$\rho_0(\omega, \eta) \quad = \quad \mathbb{I}_{\{\omega\neq\eta\}}.$$

Total variation distance The total variation distance is defined as

$$\rho_1(\omega, \eta) \quad = \quad \sum_{x\in\Lambda} |\omega(\{x\}) - \eta(\{x\})|.$$

Wasserstein distance If $\Lambda = \mathbb{R}^k$ and κ is the Euclidean distance, the Wasserstein distance is given by

$$\rho_2(\omega, \eta) = \inf_{\beta\in\Sigma(\omega,\eta)} \sqrt{\int_\Lambda \int_\Lambda \kappa(x, y)^2 \, \mathrm{d}\beta(x, y)},$$

where $\Sigma(\omega, \eta)$ denotes the set of configurations $\beta \in \Gamma_{\Lambda\times\Lambda}$ having marginals ω and η, see [4, 11].

In order to use the Kantorovich-Rubinstein duality Theorem, the main property is the following:

Proposition 3.1 ([5, 11]) *For any $i \in \{0, 1, 2\}$, the distance ρ_i is lower semi-continuous on the product space $\Gamma_\Lambda \times \Gamma_\Lambda$ equipped with the product topology.*

Remark 3.1 *As the total variation distance ρ_1, the Wasserstein distance ρ_2 also shares the property that it might takes infinite values. Indeed, if the total masses of two configurations ω and η are finite but differ, then there exists no coupling configuration β in $\Sigma(\omega, \eta)$, hence the distance should be infinite.*

Remark 3.2 *Let us mention that ours definitions of the distance ρ_i is very often infinite itself, as in the Wiener space situation of [6]. These definitions do not coincide with some of the usual definitions of (bounded) distances between point processes, see for instance [11, 3, 13] especially there are not lower semi-continuous, the Kantorovich-Rubinstein duality Theorem is no longer satisfied.*

I.3 Malliavin derivatives and the Rademacher property

Hypothesis 3.1 *Assume now that we have:*

- *A kernel Q on $\Gamma_\Lambda \times \Lambda$, i.e. $Q(\cdot, A)$ is measurable as a function on Γ_Λ for any $A \in \mathcal{B}(\Lambda)$ and $Q(\omega, \cdot)$ is a positive Radon measure on $\mathcal{B}(\Lambda)$ for any $\omega \in \Gamma_\Lambda$. We set $\mathrm{d}\alpha(\omega, x) = Q(\omega, \mathrm{d}x)\, \mathrm{d}\mu_\sigma(\omega)$.*
- *A gradient/Malliavin derivative ∇, defined on the dense subset of $\mathbb{L}^2(\mu_\sigma)$: $\mathrm{Dom}\,\nabla = \{F \in \mathbb{L}^2(\mu_\sigma) : \nabla F \in \mathbb{L}^2(\alpha)\}$.*

Definition 3.1 *Given a distance ρ and a gradient ∇ on Γ_Λ, we say that the couple (∇, ρ) has the Rademacher property whenever*

$$\rho - \mathrm{Lip}_1 \subset \mathrm{Dom}\,\nabla \qquad \textit{and} \qquad |\nabla_x F(\omega)| \leq 1, \quad \alpha\textit{-a.e.} \tag{3.4}$$

Discrete gradient on configuration space. Given a functional $F \in \mathbb{L}^2(\mu_\sigma)$, the discrete gradient of F, denoted by $\nabla^\sharp F$, is defined as

$$\nabla^\sharp_x F(\omega) = F(\omega + \varepsilon_x) - F(\omega), \quad (\omega, x) \in \Gamma_\Lambda \times \Lambda.$$

In particular, $\mathrm{Dom}\,\nabla^\sharp$ is the subspace of $\mathbb{L}^2(\mu_\sigma)$ random variables such that

$$\mathbb{E}_{\mu_\sigma}\left[\int_\Lambda |\nabla^\sharp_x F|^2 \partial\sigma(x)\right] < +\infty.$$

We set $Q^\sharp(\omega, \mathrm{d}x) = \mathrm{d}\sigma(x)$ so that $\alpha^\sharp = \mu_\sigma \otimes \sigma$. According to [9, 12], the Chaotic Representation Property holds on the configuration space. Thus we can define (see Section III.1) the notions of n-th multiple stochastic integral of a real-valued square-integrable symmetric function $f_n \in \mathcal{L}^2(\sigma^{\otimes n})$, the chaotic expression of the gradient, of $\delta^\sharp$ the adjoint operator of $\nabla^\sharp$, of the self-adjoint number operator $\mathcal{L}^\sharp = \delta^\sharp \nabla^\sharp$ and finally of the associated Ornstein-Uhlenbeck semi-group $(P^\sharp_t)_{t\geq 0}$.

Remark 3.3

- *$(P^\sharp_t)_{t\geq 0}$ is exponentially ergodic in $\mathbb{L}^2(\mu_\sigma)$ with respect to the Poisson measure μ_σ.*
- *Moreover, we have the commutation property between gradient and semi-group: if $F \in \mathrm{Dom}\,\nabla^\sharp$,*

$$\nabla^\sharp_x P^\sharp_t F = e^{-t} P^\sharp_t \nabla^\sharp_x F, \quad x \in \Lambda, \quad t \geq 0. \tag{3.5}$$

Proposition 3.2 *Assume that the intensity measure σ is finite on Λ. Then the couples $(\nabla^\sharp, \rho_0)$ and $(\nabla^\sharp, \rho_1)$ satisfy the Rademacher property* (3.4).

Differential gradient on configuration space. Let us introduce another stochastic gradient on the configuration space Γ_Λ which is a derivation, see [1, 11]. Given the Euclidean space $\Lambda = \mathbb{R}^k$, let $V(\Lambda)$ be the space of $\mathcal{C}^\infty$ vector fields on Λ and $V_0(\Lambda) \subset V(\Lambda)$, the subspace consisting of all vector fields with compact support. For $v \in V_0(\Lambda)$, for any $x \in \Lambda$, the curve $t \mapsto \mathcal{V}^v_t(x) \in \Lambda$ is defined as the solution of the following Cauchy problem

$$\begin{cases} \frac{\partial}{\mathrm{d}t}\mathcal{V}^v_t(x) &= v(\mathcal{V}^v_t(x)), \\ \mathcal{V}^v_0(x) &= x. \end{cases} \tag{3.6}$$

The associated flow $(\mathcal{V}^v_t, t \in \mathbb{R})$ induces a curve $(\mathcal{V}^v_t)^*\omega = \omega \circ (\mathcal{V}^v_t)^{-1}$, $t \in \mathbb{R}$, on Γ_Λ: if $\omega = \sum_{x\in\omega} \varepsilon_x$ then $(\mathcal{V}^v_t)^*\omega = \sum_{x\in\omega} \varepsilon_{\mathcal{V}^v_t(x)}$. We are then in position to define a notion of differentiability on Γ_Λ. We take $Q^c(\omega, \mathrm{d}x) = \mathrm{d}\omega(x) = \sum_{y\in\omega} \mathrm{d}\varepsilon_y(x)$ and $\mathrm{d}\alpha^c(\omega, x) = \mathrm{d}\omega(x)\, \mathrm{d}\mu_\sigma(\omega)$. A measurable function $F : \Gamma_\Lambda \to \mathbb{R}$ is said to be differentiable if for any $v \in V_0(\Lambda)$, the following limit exists:

$$\lim_{t\to 0} \frac{F(\mathcal{V}^v_t(\omega)) - F(\omega)}{t}.$$

We denote $\nabla^c_v F(\omega)$ the preceding quantity. The domain of ∇^c is then the set of integrable and differentiable functions such that there exists a process $(\omega, x) \mapsto \nabla^c_x F(\omega)$ which belongs to $\mathbb{L}^2(\alpha^c)$ and satisfies

$$\nabla^c_v F(\omega) = \int_\Lambda \nabla^c_x F(\omega) v(x)\, \mathrm{d}\omega(x).$$

We denote by δ^c the adjoint operator of ∇^c. Given the self-adjoint operator $\mathcal{L}^c = \delta^c \nabla^c$, denote the associated Ornstein-Uhlenbeck semi-group by $(P^c_t)_{t\geq 0}$.

Remark 3.4

- $(P_t^c)_{t\geq 0}$ *is ergodic in* $\mathbb{L}^2(\mu_\sigma)$ *with respect to the Poisson measure* μ_σ *[1].*
- *There is no known commutation relationship between the gradient* ∇^c *and the semi-group* P_t^c.

Proposition 3.3 ([5]) *The couple* (∇^c, ρ_2) *satisfies the Rademacher property* (3.4).

II Upper bounds on Rubinstein distances

Proposition 3.4 ([5]) *Denote* ρ *a lower semi-continuous distance on the configuration space* Γ_Λ *and assume that Hypothesis 3.1 is fulfilled. Assume that the couple* (∇, ρ) *satisfies the Rademacher property* (3.4). *Let* L *be the density of an absolutely continuous probability measure* ν *with respect to* μ_σ. *Then provided the inequality makes sense, the following upper bound on the Rubinstein distance holds:*

$$\mathcal{T}_\rho(\mu_\sigma, \nu) \leq \int_{\Gamma_\Lambda} \int_\Lambda \left| \int_0^{+\infty} \nabla_x P_t L(\omega)\, \mathrm{d}t \right| \mathrm{d}\alpha(\omega, x). \tag{3.7}$$

This first abstract upper bound on the Rubinstein distance is proved using a semi-group method, following the approach emphasized in [7]. Note that the upper bound in the inequality (3.7) is interesting in its own right, but seems to be somewhat difficult to compute in full generality. Hence we turn in the sequel to more concrete situations, i.e., when the gradient of interest is the discrete gradient $\nabla^\sharp$ or the differential one ∇^c and is associated to the convenient distance ρ_i, $i \in \{0, 1, 2\}$, in the sense of the Rademacher property (3.4). Once the abstract estimate (3.7) has been obtained, one notices that it might be simplified whenever a commutation relation between gradient and semi-group holds. Such a property is only verified in the case of the discrete gradient, so that we focus in this part on the couple $(\nabla^\sharp, \rho_1)$.

Theorem 3.1 ([5]) *Let* L *be the density of an absolutely continuous probability measure* ν *with respect to* μ_σ, *and assume that* $L \in \mathrm{Dom}\, \nabla^\sharp$ *and* $\nabla^\sharp L \in L^1(\mu_\sigma \otimes \sigma)$. *Then we get the following estimate:*

$$\mathcal{T}_{\rho_1}(\mu_\sigma, \nu) \leq \mathbb{E}_{\mu_\sigma} \left[\int_\Lambda |\nabla_x^\sharp L|\, \mathrm{d}\sigma(x) \right], \tag{3.8}$$

$$\mathcal{T}_{\rho_1}(\mu_\sigma, \nu) \leq \mathbb{E}_{\mu_\sigma} \left[\int_\Lambda |(\mathcal{I} + \mathcal{L}^\sharp)^{-1} \nabla_x^\sharp L|\, \mathrm{d}\sigma(x) \right]. \tag{3.9}$$

The same inequality also holds under the distance ρ_0.

Remark 3.5

- (3.9) *seems theoretically slightly better than* (3.8) *but often yields to intractable computations, except when the chaos representation of* L *is given.*
- (3.2) *is the very analogue of* (3.9) *on Wiener space.*

Proposition 3.4 for the couple (∇^c, ρ_2) is of theoretical interest, but not really tractable in practise, since no commutation relation has been established yet between the differential gradient ∇^c and the semi-group P_t^c. It is possible to provide another estimate on $\mathcal{T}_{\rho_2}$ through a different approach relying on a time-change argument together with the Girsanov Theorem.

III Applications

III.1 Distance estimates between processes

The purpose of the present part is to apply our main results Theorems to provide distance estimates between a Poisson process and several other more sophisticated processes, such as Cox or Gibbs processes. See for instance the pioneer monograph [3] or also [2, 13] for similar results with respect to another (bounded) distances on the configuration space Γ_Λ.

Proposition 3.5 (Distance between Poisson processes, [5]) *Let μ_τ be a Poisson measure on Γ_Λ of intensity τ. We assume that τ admits a density p with respect to σ such that $p-1 \in \mathcal{L}^1(\sigma)$. Then we have*

$$\mathcal{T}_{\rho_1}(\mu_\sigma, \mu_\tau) \leq \int_\Lambda |p(x)-1| \, \mathrm{d}\sigma(x).$$

Definition 3.2 *A Cox process is a Poisson process with a random intensity. To construct a Cox process, we need to enlarge our probability space. Given an arbitrary probability measure $\mathbb{P}_M$ on $\mathfrak{M}(\Lambda)$, we denote by M the canonical random variable on $(\mathfrak{M}(\Lambda), \mathbb{P}_M)$, i.e. M given by $M(m) = m$ has distribution $\mathbb{P}_M$. On the space $\Gamma_\Lambda \times \mathfrak{M}(\Lambda)$, we consider the probability measures*

$$\mathrm{d}\mu'_M(\omega, m) := \mathrm{d}\mu_m(\omega) \, \mathrm{d}\mathbb{P}_M(m) \quad \textit{and} \quad \mathrm{d}\mu'_\sigma(\omega, m) := \mathrm{d}\mu_\sigma(\omega) \, \mathrm{d}\mathbb{P}_M(m).$$

Note that the second one is the distribution of the independent couple (N, M), where N is the canonical random variable on Γ_Λ with distribution μ_σ. The distribution μ'_M on Γ_Λ is said to be Cox whenever for any function $f \in \mathcal{C}_0(\Lambda)$,

$$\mathbb{E}_{\mu'_M}\left[\exp\left(\int_\Lambda f \, \mathrm{d}\omega\right) \middle| M\right] = \exp\left\{\int_\Lambda (e^f - 1) \, \mathrm{d}M\right\}.$$

In the definition of the distance between μ'_M and μ'_σ, we do not include any information on M, so that the distance ρ_1 remains the same and we have:

$$\mathcal{T}_{\rho_1}(\mu'_\sigma, \mu'_M) \quad = \quad \sup_{F \in \rho_1 - \mathrm{Lip}_1} \int_{\mathfrak{M}(\Lambda)} \left(\int_{\Gamma_\Lambda} F(\omega) \, \mathrm{d}(\mu_\sigma - \mu_m)(\omega)\right) \mathrm{d}\mathbb{P}_M(m).$$

Proposition 3.6 (Distance between a Cox and a Poisson process, [5]) *Assume that μ'_σ-a.s., the measure M is absolutely continuous with respect to σ and that there exists a measurable version of $\mathrm{d}M/\mathrm{d}\sigma$ and such that $\mathrm{d}M/\mathrm{d}\sigma - 1 \in \mathbb{L}^1(\mu'_\sigma \otimes \sigma)$. Then we have*

$$\mathcal{T}_{\rho_1}(\mu'_\sigma, \mu'_M) \leq \mathbb{E}_{\mu'_\sigma}\left[\int_\Lambda \left|\frac{\mathrm{d}M}{\mathrm{d}\sigma}(x) - 1\right| \, \mathrm{d}\sigma(x)\right].$$

Definition 3.3 *Let $\Lambda = \mathbb{R}^k$. The measure ν is a Gibbs measure on Γ_Λ with respect to the reference measure μ_σ, if the density of ν with respect to μ_σ is of the form $L = e^{-V}$, where*

$$V(\omega) := \int_\Lambda \int_\Lambda \phi(x-y) \, \mathrm{d}\omega(x) \, \mathrm{d}\omega(y) < +\infty, \quad \mu_\sigma - a.s.,$$

and where the potential $\phi : \Lambda \to (0, +\infty)$ is such that $\phi(x) = \phi(-x)$ and

$$\int_\Lambda \int_\Lambda \phi(x-y) \, \mathrm{d}\sigma(x) \, \mathrm{d}\sigma(y) < +\infty.$$

Proposition 3.7 (Distance between a Gibbs and a Poisson process, [5]) *The Rubinstein distance $\mathcal{T}_{\rho_1}$ between the Poisson measure μ_σ and the Gibbs measure ν is bounded as follows:*

$$\mathcal{T}_{\rho_1}(\mu_\sigma, \nu) \leq 2 \int_\Lambda \int_\Lambda \phi(x-y) \, \mathrm{d}\sigma(x) \, \mathrm{d}\sigma(y).$$

III.2 Tail and isoperimetric estimates

Tail estimates. Our main result Theorem 3.1 allows us to obtain a first tail estimate.

Proposition 3.8 ([5]) *Let $F \in \rho_1 - \mathrm{Lip}_1$ be centered and let $\lambda > 0$. Consider ν^λ the absolutely continuous probability measure with density $e^{\lambda F}/Z_\lambda$ with respect to μ_σ where $Z_\lambda = \mathbb{E}_{\mu_\sigma}\left[e^{\lambda F}\right]$. Thus we get the deviation inequality available for any $r \geq 0$:*

$$\mu_\sigma\left(F \geq r\right) \leq \exp\left\{r - (r + \|\nabla^\sharp F\|_{1,\infty}) \log\left(1 + \frac{r}{\|\nabla^\sharp F\|_{1,\infty}}\right)\right\}, \tag{3.10}$$

where

$$\|\nabla^\sharp F\|_{1,\infty} := \mu_\sigma - \operatorname{esssup} \int_\Lambda |\nabla^\sharp_x F| \, \mathrm{d}\sigma(x).$$

Remark 3.6 *Such a tail estimate is somewhat similar to that established for instance in [7, 15]. However, in contrast to their results, we do not exhibit at the denominator the sharp variance term*

$$\|\nabla^\sharp F\|_{2,\infty}^2 := \mu_\sigma - \operatorname{esssup} \int_\Lambda |\nabla_x^\sharp F|^2 \, \mathrm{d}\sigma(x),$$

since our method relies on the $\mathbb{L}^1$*-inequality* (3.8).

By the use of Theorem 3.1 with ν the absolutely continuous probability measure with density with respect to μ_σ:

$$L := \frac{1}{\mu_\sigma\left(\omega(K) \geq [\sigma(K)+r]\right)} \, \mathbb{I}_{\{\omega(K)\geq[\sigma(K)+r]\}}, \quad r>0,$$

we are able to recover the multiplicative polynomial factor. The result is:

Proposition 3.9 ([5, 10]) *Given any compact set* $K \subset \Lambda$ *and any* $r>0$*, we have the tail estimate:*

$$\mu_\sigma\left(\omega(K) \geq \sigma(K)+r\right) \quad \leq \quad \frac{[\sigma(K)+r]}{r} \, \frac{e^{[\sigma(K)+r]-\sigma(K)-[\sigma(K)+r]\,\log\left(\frac{[\sigma(K)+r]}{\sigma(K)}\right)}}{\sqrt{2\pi[\sigma(K)+r]}}.$$

Proposition 3.10 ([5]) *Given any fixed configuration* $\eta \in \Gamma_\Lambda$ *and provided the intensity measure* σ *is finite. Then, for any* $r>0$*, we have:*

$$\mu_\sigma\left(\rho_\eta \geq \mathbb{E}_{\mu_\sigma}\left[\rho_\eta\right]+r\right) \leq \frac{\sqrt{2\pi[\sigma(\Lambda)]}[\sigma(\Lambda)]^{[\sigma(\Lambda)]}e^{\frac{1}{12[\sigma(\Lambda)]}}}{\sigma(\Lambda)^{\sigma(\Lambda)}} \, \frac{e^{[\sigma(\Lambda)+r]-[\sigma(\Lambda)]-[\sigma(\Lambda)+r]\,\log\left(\frac{[\sigma(\Lambda)+r]}{[\sigma(\Lambda)+r]-r}\right)}}{\sqrt{2\pi[\sigma(\Lambda)+r]}},$$

where ρ_η *denotes the total variation distance* $\rho_1(\cdot,\eta)$*.*

Hence one deduces that the tail behaviour of the total variation distance is comparable to the previous ones, up to constant multiplicative factors depending on the total mass $\sigma(\Lambda)$.

Isoperimetric inequality. The distance of interest is the trivial distance ρ_0. In the sequel, we assume that the intensity measure σ is finite, so that the domain $\operatorname{Dom}\nabla^\sharp$ contains the indicator functions $\mathbb{I}_A$, $A \in \mathcal{B}(\Gamma_\Lambda)$. Given a Borel set $A \in \mathcal{B}(\Gamma_\Lambda)$, we define its surface measure as

$$\mu_\sigma(\partial A) := \mathbb{E}_{\mu_\sigma}\left[\int_\Lambda |\nabla_x^\sharp \, \mathbb{I}_A| \, \mathrm{d}\sigma(x)\right].$$

Denote h_{μ_σ} the classical isoperimetric constant:

$$h_{\mu_\sigma} = 2 \inf_{0<\mu_\sigma(A)<1} \frac{\mu_\sigma(\partial A)}{\mu_\sigma(A)(1-\mu_\sigma(A))}.$$

We have the following estimate, which is convenient for small total mass $\sigma(\Lambda)$.

Proposition 3.11 ([5, 8]) *Assume that the measure* σ *is finite. Then we have*

$$1 \leq h_{\mu_\sigma} \leq \frac{\sigma(\Lambda)}{1-e^{-\sigma(\Lambda)}}. \tag{3.11}$$

In particular, we have the asymptotic for small total mass:

$$\lim_{\sigma(\Lambda)\to 0} h_{\mu_\sigma} = 1.$$

Remark 3.7 *Houdré and Privault established first the inequality* $h_{\mu_\sigma} \geq 1$ *by using Poincaré inequality in [8]. Our estimate in the right-hand-side of (3.11) is sharp for small values of* $\sigma(\Lambda)$*, but is worse than their estimate for large* $\sigma(\Lambda)$ *since their upper bound is* $8+8\sqrt{\sigma(\Lambda)}$*.*

Bibliography

[1] Sergio Albeverio, Yuri Kondratiev, and Michael Röckner. Analysis and geometry on configuration spaces. *J. Funct. Anal.*, 154(2):444–500, 1998.

[2] Andrew D. Barbour, Timothy C. Brown, and Aihua Xia. Point processes in time and Stein's method. *Stochastics Stochastics Rep.*, 65(1-2):127–151, 1998.

[3] Andrew D. Barbour, Lars Holst, and Svante Janson. *Poisson approximation*, volume 2 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1992. Oxford Science Publications.

[4] Laurent Decreusefond. Wasserstein distance on configuration space. *Potential Anal.*, 28(3):283–300, 2008.

[5] Laurent Decreusefond, Aldéric Joulin, and **Nicolas Savy**. Upper bounds on Rubinstein distances on configuration spaces and applications. *Communication on Stochastic Analysis and Applications*, 4(3):377–399, 2010.

[6] Denis Feyel and Ali S. Üstünel. Monge-Kantorovitch measure transportation and Monge-Ampère equation on Wiener space. *Probab. Theory Related Fields*, 128(3):347–385, 2004.

[7] Christian Houdré and Nicolas Privault. Concentration and deviation inequalities in infinite dimensions via covariance representations. *Bernoulli*, 8(6):697–720, 2002.

[8] Christian Houdré and Nicolas Privault. Isoperimetric and related bounds on configuration spaces. *Statist. Probab. Lett.*, 78(14):2154–2164, 2008.

[9] David Nualart and Josep Vives. Anticipative calculus for the Poisson process based on the Fock space. In *Séminaire de Probabilités, XXIV, 1988/89*, volume 1426 of *Lecture Notes in Math.*, pages 154–165. Springer, Berlin, 1990.

[10] Vygantas Paulauskas. Some comments on inequalities for deviations for infinitely divisible random vectors. *Liet. Mat. Rink.*, 42(4):494–517, 2002.

[11] Michael Röckner and Alexander Schied. Rademacher's theorem on configuration spaces and applications. *J. Funct. Anal.*, 169(2):325–356, 1999.

[12] Juan Ruiz de Chávez. Espaces de Fock pour les processus de Wiener et de Poisson. In *Séminaire de probabilités, XIX, 1983/84*, volume 1123 of *Lecture Notes in Math.*, pages 230–241. Springer, Berlin, 1985.

[13] Dominic Schuhmacher. *Estimation of distances between point process distributions*. PhD thesis, Universität Zürich, 2005.

[14] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[15] Liming Wu. A new modified logarithmic Sobolev inequality for Poisson point processes and several applications. *Probab. Theory Related Fields*, 118(3):427–438, 2000.

Chapter 4

Properties of Estimators for some diffusion processes.*

The research evoked in this chapter deals with statistics of processes especially questions of estimation of parameters. The chapter organizes in three parts. The first part summarizes the article [3] on estimation of instantaneous volatility. This is a question of paramount interest in finance. The technique used are mainly martingale's one. The second part presents two extension of the paper of Bercu and Rouault [11] which deals with large deviations principles for the maximum likelihood estimator of the drift parameter associated to Ornstein-Uhlenbeck processes directed by a Brownian motion B observed on a time interval $[0, T]$:

$$\mathrm{d}X_t = \theta X_t \,\mathrm{d}t + \,\mathrm{d}B_t, \tag{4.1}$$

with initial state $X_0 = 0$ and drift parameter $\theta < 0$. The first extension [7] consists in replacing in (4.1), the Brownian motion by a fractional Brownian motion B^H of Hurst parameter $0 < H < 1$. The second extension [8] consists in considering (4.1) with a non-negative drift parameter. Finally the Chapter ends by a section devoted to Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes defined, over the time interval $[0, T]$, by

$$\left\{ \begin{array}{lclcl} \mathrm{d}X_t & = & \theta X_t \,\mathrm{d}t & + & \mathrm{d}V_t \\ \mathrm{d}V_t & = & \rho V_t \,\mathrm{d}t & + & \mathrm{d}B_t \end{array} \right. \tag{4.2}$$

where $\theta < 0$, $\rho \leq 0$ and (B_t) is a standard Brownian motion. Our motivation for studying (4.2) comes from two observations. On the one hand, the increments of Ornstein-Uhlenbeck processes are not independent which means that the weighted maximum likelihood estimation approach of [23] does not apply directly to our situation. On the other hand, Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes are clearly related with stochastic volatility models in financial mathematics [34]. Furthermore, (4.2) is the continuous-time version of the first-order stable autoregressive process driven by a first-order autoregressive process recently investigated in [9]. The investigations are devoted to the maximum likelihood estimation for θ and ρ. We also introduce the continuous-time Durbin-Watson statistic which will allow us to propose a serial correlation test for Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes. The almost sure convergence as well as the asymptotic normality of our estimates are established. One shall realize that there is a radically different behaviour of the estimator of ρ in the two situations where $\rho < 0$ and $\rho = 0$.

---

* Publications related to this chapter:

[3] Alexander Alvarez, Fabien Panloup, Monique Pontier, and **Nicolas Savy**. Estimation of the instantaneous volatility. *Statistic inference for Stochastic processes*, 15: 27–59, 2012.

[7] Bernard Bercu, Laure Coutin, and **Nicolas Savy**. Sharp large deviations for the fractional Ornstein-Uhlenbeck process. *Theory of Probability and its Applications*, 55(4): 575–610, 2011.

[8] Bernard Bercu, Laure Coutin, and **Nicolas Savy**. Sharp large deviation for the non-stationary Ornstein-Uhlenbeck process. *Stochastic Processes and their Applications*, 122(10): 3393–3424, 2012.

[10] Bernard Bercu, Frédéric Proïa, and **Nicolas Savy**. On Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes. *Statistics and Probability Letters*, 85:36-44, 2014.

I Estimator of instantaneous volatility

In [3], we deal with $(X_t)_{t\geq 0}$ defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ by:

$$\mathrm{d}X_t = a_t \,\mathrm{d}t + \sigma_t \,\mathrm{d}B_t, \quad t \geq 0, \tag{4.3}$$

where B is an $(\mathcal{F}_t)$-adapted Wiener process which satisfies the usual conditions, $a : \mathbb{R}_+ \to \mathbb{R}$ and σ are some càdlàg $(\mathcal{F}_t)$-adapted processes. Furthermore, σ is assumed to be a positive process.

In order to reach an estimator of the volatility it is an usual way to use a discretely observed process and to deal with the power variation of order p. Consider T a positive number and assume that X is observed at times $i\Delta_n$ for all $i = 0, 1, \ldots, [\frac{T}{\Delta_n}]$. In the sequel, we will assume that $\Delta_n \xrightarrow[n\to+\infty]{} 0$. For $p > 0$, we denote by $\hat{B}(p, \Delta_n)$, the process of *power variations of order* p, *i.e.* the stochastic process defined as

$$\widehat{B}(p, \Delta_n)_t := \sum_{i=1}^{[t/\Delta_n]} |\Delta_i^n X|^p, \quad t \in [0, T]$$

where $\Delta_i^n X := X_{i\Delta_n} - X_{(i-1)\Delta_n}$.

The sequence $(\widehat{B}(p, \Delta_n)_t)_n$ is classically known as an estimator of $\int_0^t \sigma_s^p \,\mathrm{d}s$. The study of such estimators of the integrated volatility and its use for the detection of jumps have been deeply studied in the last years (see for instance [6, 36] for the continuous setting, [1, 24, 36] for the discontinuous setting). We are going to recall some existing results about the convergence of this sequence but before, we want to precise the assumptions on σ that will be necessary throughout the section.

Two assumptions are sufficient to establish the properties of this estimator. First an assumption depending on parameter $q \in [1, 2]$ which is related to the behaviour of the small jumps of (σ_t):

Hypothesis 4.1 ($\mathbf{H_q^1}$) *σ is a positive càdlàg semi-martingale such that $\sigma_t = |Y_t|$ where (Y_t) satisfies:*

$$dY_s = b_s \,\mathrm{d}s + \eta_1(s) \,\mathrm{d}W_s + \eta_2(s) \,\mathrm{d}W_s^2 + \int_{\mathbb{R}} y \,\mathbb{I}_{\{|y|\leq 1\}}(\mu(\,\mathrm{d}s,\,\mathrm{d}y) - \nu(\,\mathrm{d}s,\,\mathrm{d}y)) + \int_{\mathbb{R}} y \,\mathbb{I}_{\{|y|>1\}}\mu(\,\mathrm{d}s,\,\mathrm{d}y),$$

where b, η_1, η_2 are adapted càdlàg processes, μ denotes a random measure on $\mathbb{R}_+ \times \mathbb{R}$ with predictable compensator ν satisfying: $\nu(\mathrm{d}t, \mathrm{d}y) = \mathrm{d}t\, F_t(\mathrm{d}y)$ and $(\int (1\wedge|y|^q) F_t(\mathrm{d}y))_{t\geq 0}$ is a locally bounded predictable process.

Second, an assumption depending on parameter $q \in [1, 2]$ too which is a little more constraining control of the jump component:

Hypothesis 4.2 ($\mathbf{H_q^2}$) *For every $T > 0$,*

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,T]} \int_{\{|y|\leq\varepsilon\}} |y|^q F_t(\,\mathrm{d}y) = 0 \qquad a.s.$$

Remark 4.1 *Assumption ($\mathbf{H_q^1}$) implies that (σ_t) is quasi-left continuous and that the jump component has locally-finite q-variation.*

Now, we can recall two results (adapted to our context) about the asymptotic properties of $(\widehat{B}(p, \Delta_n)_t)_n$. On the same topic, we can also quote [5, 6].

Proposition 4.1 ([29, 24]) *Assume ($\mathbf{H_2^2}$). Let p be a positive number and set $m_p := \mathbb{E}\left[|U|^p\right]$ where $U \sim \mathcal{N}(0, 1)$. Then, locally uniformly in t,*

$$\Delta_n^{1-\frac{p}{2}} \hat{B}(p, \Delta_n)_t \xrightarrow[n\to+\infty]{\mathbb{P}} m_p A(p)_t \quad \textit{with} \quad A(p)_t = \int_0^t \sigma_s^p \,\mathrm{d}s.$$

Proposition 4.2 ([2]) *Let $p \geq 2$ and assume Assumption 1 of [2]. Then, the sequence of continuous processes $(Y(n, p))_{n\in\mathbb{N}^*}$ defined for any $n \in \mathbb{N}^*$ by*

$$Y(n, p)_t := \frac{1}{\sqrt{\Delta_n}} \Big(\Delta_n^{1-\frac{p}{2}} \hat{B}(p, \Delta_n)_t - m_p A(p)_t\Big), \quad t \geq 0,$$

converges stably to a random variable $Y(p)$ on an extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t), \tilde{\mathbb{P}})$ of the original filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ such that, for any $t \geq 0$, conditionally on $\mathcal{F}$, $Y(p)_t$ is a centered Gaussian variable with variance $\tilde{\mathbb{E}}[Y(p)_t^2 \mid \mathcal{F}] = (m_{2p} - m_p^2) A(2p)_t$.

It is important to quote that the asymptotic normality is expressed in terms of stable convergence denoted by $\mathcal{L}-s$. This convergence is defined as:

Definition 4.1 *We say that a sequence of random variables (Y_n) converges stably to Y or $Y_n \overset{\mathcal{L}-s}{\Rightarrow} Y$, if there exists an extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ of $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable Y defined on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that for every bounded measurable random variable H, for every bounded continuous function f, $\mathbb{E}\left[Hf(Y_n)\right] \to \tilde{\mathbb{E}}[Hf(Y)]$ when $n \to +\infty$ where $\tilde{\mathbb{E}}$ denotes the expectation on the extension.*

The following lemma is classical and is the reason why we introduce the stable convergence.

Lemma 4.1 *Let (X_n) and (Y_n) be some sequences of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a Polish space E. Assume that (X_n) converges $\mathcal{L}-s$ to X and that (Y_n) converges in probability to Y. Then, the sequence of random variables $(Z_n = X_n + Y_n)$ converges $\mathcal{L}-s$ to $X+Y$.*

Unlike these works, the aim of [3] is to estimate rather the instantaneous volatility. Then, the natural idea is to study estimators of σ_t^p which are built as "derivatives" of the power variations. More precisely, the proposed estimator $(\Sigma(p, \Delta_n, h_n)_t)$ of the instantaneous volatility is a normalized relative increment of cumulative volatility estimator, this relative increment being taken on a smaller and smaller interval.

$$\Sigma(p, \Delta_n, h_n)_t := \frac{\Delta_n^{1-\frac{p}{2}}\left(\hat{B}(p, \Delta_n)_{t+h_n} - \hat{B}(p, \Delta_n)_t\right)}{m_p h_n}, \quad t \leq \bar{T} \text{ with } \bar{T} = T - h_1.$$

Actually, this estimator is the mean of p-variations in a window of length h_n where (h_n) is assumed to be a non-increasing sequence of positive numbers such that h_n tends to 0.

Theorem 4.1 (With Brownian component. [3]) *Let $p = 2$ or $p \geq 3$ and let (X_t) be a stochastic process solution to (4.3). Assume $(\mathbf{H_2^1})$ and $(\mathbf{H_2^2})$. Assume that $\Delta_n = o(h_n)$. Then,*
(i) If $h_n/\sqrt{\Delta_n} \to 0$, $\forall t \in [0, \bar{T}]$,

$$\sqrt{\frac{h_n}{\Delta_n}}(\Sigma(p, \Delta_n, h_n)_t - \sigma_t^p) \xrightarrow[n\to+\infty]{\mathcal{L}-s} \sqrt{\varphi_1(p, t, \sigma)}\ \ U, \tag{4.4}$$

where, conditionally on $\mathcal{F}$, U is a standard Gaussian random variable and $\varphi_1(p, t, \sigma) = \frac{m_{2p} - m_p^2}{m_p^2}\sigma_t^{2p}$.
(ii) If $\sqrt{\Delta_n}/h_n \to \beta \in \mathbb{R}_+$, $\forall t \in [0, \bar{T}]$,

$$\frac{1}{\sqrt{h_n}}(\Sigma(p, \Delta_n, h_n)_t - \sigma_t^p) \xrightarrow[n\to+\infty]{\mathcal{L}-s} \sqrt{\beta^2\varphi_1(p, t, \sigma) + \varphi_2(p, t)}\ \ U,$$

where $\varphi_1(p, t, \sigma)$ and U are defined as before and, $\varphi_2(p, t) = \frac{p^2}{3}(\sigma_t)^{2p-2}\|\eta\|^2(t))$ with $\|\eta\|^2(t) = \eta_1^2(t) + \eta_2^2(t)$.

Remark 4.2 *When the drift term a is null, the result is valid even if $2 < p < 3$. Otherwise, the drift contributes in a bias for the estimator that is not negligible in case $2 < p < 3$.*

Theorem 4.2 (Without Brownian component. [3]) *Let $p = 2$ or $p \geq 3$. and let (X_t) be a stochastic process solution to (4.3). Assume $(\mathbf{H_q^1})$ and $(\mathbf{H_q^2})$ with $q \in [1, 2]$ and suppose that $\eta_1 = \eta_2 = 0$. Assume that $\Delta_n = o(h_n)$. Then,*
(i) If $q \in (1, 2]$, if $\limsup_{n\to+\infty} h_n^{1/2+1/q}/\sqrt{\Delta_n} < +\infty$, $\forall t \in [0, \bar{T}]$,

$$\sqrt{\frac{h_n}{\Delta_n}}(\Sigma(p, \Delta_n, h_n)_t - \sigma_t^p) \xrightarrow[n\to+\infty]{\mathcal{L}-s} \sqrt{\varphi_1(p, t, \sigma)}U,$$

where $\varphi_1(p, t, \sigma)$ and U are defined as in Theorem 4.1.
(ii) Assume that $q = 1$. If $\lim_{n\to+\infty} h_n^3/\Delta_n = 0$, (4.2) holds. If $\lim_{n\to+\infty} h_n^3/\Delta_n = \beta \in \mathbb{R}_+^*$ and if $(\int_{\{0<|y|\leq 1\}} yF_t(\,\mathrm{d}y))_{t\geq 0}$ is càglàd (left-continuous with right limits), then, $\forall t \in [0, \bar{T}]$,

$$\sqrt{\frac{h_n}{\Delta_n}}(\Sigma(p, \Delta_n, h_n)_t - \sigma_t^p) \xrightarrow[n\to+\infty]{\mathcal{L}-s} \sqrt{\varphi_1(p, t, \sigma)}U + \frac{\beta}{2}p\sigma_t^{p-1}\left(b_t - \lim_{s\searrow t}\int_{\{0<|y|\leq 1\}} yF_s(\,\mathrm{d}y)\right). \tag{4.5}$$

Remark 4.3 *When there is no Brownian component in the volatility, i.e. $\eta_1 = \eta_2 = 0$ and when the jump component has locally q-finite variation, we can alleviate the constraint on the sequence (h_n).*

It must be stressed here that the convergence rate depends on the balance between the frequency of observations and the length h_n of the window. Hence, the following proposition justify the choice of a "good pair" (h_n, Δ_n).

Corollary 4.1 *Let $p = 2$ or $p \geq 3$ and assume $\Delta_n = o(h_n)$. Considering the window width h_n, $r_n := \frac{h_n}{\Delta_n}$ corresponds to the number of observations on the interval $[t, t + h_n]$. Suppose $\Delta_n = \frac{1}{n}$ and $r_n := n^\rho$, $0 < \rho < 1$, then $h_n = n^{\rho-1}$.*

- *If there is Brownian component, under Hypotheses $(\mathbf{H_2^1})$, $(\mathbf{H_2^2})$, Theorem 4.1 yields the following convergence rates:*
  - *(i) $\rho < \frac{1}{2}$ yields a convergence rate of order $n^{\rho/2}$,*
  - *(ii) $\rho \geq \frac{1}{2}$ yields a convergence rate of order $n^{(1-\rho)/2}$.*
- *If there is no Brownian component, under Hypotheses $(\mathbf{H_q^1})$ and $(\mathbf{H_q^2})$ with $1 \leq q \leq 2$, Theorem 4.2 yields the following convergence rates:*
  - *(i) if $1 < q \leq 2$, $\rho \leq \frac{2}{2+q}$, yields a convergence rate of order $n^{\rho/2}$,*
  - *(ii) if $q = 1$, $\rho \leq \frac{2}{3}$; the best convergence rate is of order $n^{1/3}$, obtained for $\rho = 2/3$.*

Remark 4.4 *The main restriction of our model is that jumps only occur in the volatility but not in the price. When jumps occur in the log-price X, it seems that we could extend some of the previous announced results by exploiting the fact that convergence properties for the power variations to the cumulated volatility still hold when $p < 2$. However, this extension generates some technicalities which are out of ours objectives.*

Sketch of the proof. First consider the following assumption:

Hypothesis 4.3 $(\mathbf{SH_q})$ *Functions a,b, η_1, η_2, and $\int_0^\cdot \int (|y|^q \wedge 1) F_s(\,\mathrm{d}y)\,\mathrm{d}s$ are bounded and there exists $M > 0$ such that $F_s([-M, M]^c) = 0$ a.s. $\forall s \geq 0$.*

By means of a classical localization procedure we show that it is enough to prove the main theorem under assumption $(\mathbf{SH_q})$:

Lemma 4.2 *Assume that the conclusions of Theorem 4.1 and 4.2 hold for every (X, σ) satisfying $(\mathbf{SH_q})$ and $(\mathbf{H_q^2})$ (with $q \in [1, 2]$ depending on the statement). Then, the conclusions hold for every (X, σ) satisfying $(\mathbf{H_q^1})$ and $(\mathbf{H_q^2})$ with $q \in [1, 2]$.*

Now, following [24], we first decompose $\Sigma(p, \Delta_n, h_n)_t - \sigma_t^p$ as follows:

$$\Sigma(p, \Delta_n, h_n)_t - \sigma_t^p = \frac{Z_{t+h_n}^{(n,p)} - Z_t^{(n,p)}}{m_p h_n} + \Big(\frac{1}{r_n} \sum_{\mathcal{D}_t^n} \sigma_{i\Delta_n}^p - \sigma_t^p\Big), \tag{4.6}$$

where $r_n = h_n/\Delta_n$, $\mathcal{D}_t^n = \Big\{i \in \mathbb{N}^*,\ \Big[\frac{t}{\Delta_n}\Big] + 1 \leq i \leq \Big[\frac{t+h_n}{\Delta_n}\Big]\Big\}$, and

$$Z_t^{(n,p)} := \Delta_n^{1-\frac{p}{2}} \hat{B}(p, \Delta_n)_t - m_p \sum_{i=1}^{[t/\Delta_n]} \Delta_n \sigma_{i\Delta_n}^p.$$

On the one hand, we split

$$\frac{Z_{t+h_n}^{(n,p)} - Z_t^{(n,p)}}{h_n} = \Lambda_1^n(t) + \Lambda_2^n(t) + \Lambda_3^n(t).$$

On the other hand, we decompose the second part of (4.6).

$$\frac{1}{r_n} \sum_{i \in \mathcal{D}_t^n} \sigma_{i\Delta_n}^p - \sigma_t^p = \Lambda_4^n(t) + \Lambda_5^n(t),$$

where Λ_2^n and Λ_4^n are increments of Brownian martingales while Λ_1^n, Λ_3^n and Λ_5^n are remainder terms.

In order to deal with Λ_2^n and Λ_4^n we make use of the following lemma which is a corollary of a result [19] on the stable-CLT for martingale increments adapted to our specific framework.

Lemma 4.3 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space. For $n \geq 1$, let $\zeta_2^n, \zeta_3^n, \ldots, \zeta_{k_n}^n$ denote some martingale increments with respect to the sub-σ-fields of $\mathcal{F}$: $\bar{\mathcal{F}}_{n,1} \subset \bar{\mathcal{F}}_{n,2} \subset \ldots \subset \bar{\mathcal{F}}_{n,k_n}$. Set $S_n = \sum_{i=2}^{k_n} \zeta_i^n$ and $\mathcal{G} = \cap_{n\geq 1} \bar{\mathcal{F}}_{n,1}$. Assume that $n \to \bar{\mathcal{F}}_{n,k_n}$ is a non-increasing sequence of σ-fields such that $\cap_{n\geq 1} \bar{\mathcal{F}}_{n,k_n} = \mathcal{G}$. Then, if the following conditions hold:*

(i) There exists a $\mathcal{G}$-measurable random variable η such that

$$\sum_{i=2}^{k_n} \mathbb{E}\left[(\zeta_i^n)^2 | \bar{\mathcal{F}}_{n,i-1}\right] \xrightarrow[n\to+\infty]{\mathbb{P}} \eta,$$

(ii) For every $\varepsilon > 0$,

$$\sum_{i=2}^{k_n} \mathbb{E}\left[(\zeta_i^n)^2 \, \mathbb{I}_{\left\{|\zeta_i^n|^2 \geq \varepsilon\right\}} | \bar{\mathcal{F}}_{n,i-1}\right] \xrightarrow[n\to+\infty]{\mathbb{P}} 0,$$

then, (S_n) converges stably to S where S is defined on an extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ and such that conditionally on $\mathcal{F}$, the distribution of S is a centered Gaussian law with variance η.

Proposition 4.3 *Assume that $\Delta_n = o(h_n)$ and $(\mathbf{SH_2})$.*
(i). Then,

$$\rho_n \left(\Lambda_2^n(t) + \Lambda_4^n(t)\right) \xrightarrow[n\to+\infty]{\mathcal{L}-s} f(t,p)U,$$

where $U \sim \mathcal{N}(0,1)$, U is independent of $\mathcal{F}_t$ and

$$(f^2(t,p),\ \rho_n) = \begin{cases} \left(\varphi_1(p,t,\sigma),\ \sqrt{r_n}\right) & \text{if } h_n = o(\sqrt{\Delta_n}), \\ \left(\beta^2\varphi_1(p,t,\sigma) + \varphi_2(p,t),\ \frac{1}{\sqrt{h_n}}\right) & \text{if } \frac{\sqrt{\Delta_n}}{h_n} \to \beta \in \mathbb{R}_+^*, \\ \left(\frac{1}{3}p^2(\sigma_t)^{2p-2}\|\eta\|^2(t),\ \frac{1}{\sqrt{h_n}}\right) & \text{if } \frac{\sqrt{\Delta_n}}{h_n} \to 0. \end{cases}$$

(ii). In case of pure jump process, meaning we assume that $\eta_1 = \eta_2 = 0$, then, $\Lambda_4 = 0$ and, for every $t \in [0,T]$,

$$\sqrt{\frac{h_n}{\Delta_n}} \Lambda_2^n(t) \xrightarrow[n\to+\infty]{\mathcal{L}-s} f(t,p)U,$$

with $f^2(t,p) = \varphi_1(p,t,\sigma)$.

Finally for the remainder terms, we show that:

Proposition 4.4

- *Assume $(\mathbf{SH_q})$ and $(\mathbf{H_q^2})$ with $q \in]1,2]$. Then, for every $t \in [0,T]$:*

$$\frac{1}{h_n^{1/q}} \Lambda_5^n(t) \xrightarrow[n\to+\infty]{\mathbb{P}} 0. \tag{4.7}$$

  Assume that the previous assumptions hold with $q = 1$ and that $\left(\int_{\{0<|y|\leq 1\}} yF_t(\,\mathrm{d}y)\right)_{t\geq 0}$ is càglàd, then, for every $t \in [0,T]$:

$$\frac{1}{h_n}\Lambda_5^n(t) \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{\theta_t^0}{2} \quad \textit{with} \quad \theta_t^0 := p\sigma_t^{p-1}\left(b_t - \lim_{s\searrow t}\int_{\{0<|y|\leq 1\}} yF_s(\,\mathrm{d}y)\right) + \frac{p(p-1)}{2}\sigma_t^{p-2}\|\eta\|^2(t).$$

- *Assume $(\mathbf{SH_2})$. Then, for every $t \in [0,\bar{T}]$,*

$$\sqrt{\frac{h_n}{\Delta_n}} \Lambda_1^n(t) \xrightarrow[n\to+\infty]{\mathbb{L}^2} 0. \tag{4.8}$$

- *Assume $(\mathbf{SH_2})$. Then, if $p = 2$ or $p \geq 3$, for every $t \in [0,\bar{T}]$,*

$$\max\left(\sqrt{\frac{h_n}{\Delta_n}}, \sqrt{\frac{1}{h_n}}\right) \Lambda_3^n(t) \xrightarrow[n\to+\infty]{\mathbb{L}^1} 0. \tag{4.9}$$

Focus for instance on statements (4.4) and (4.2), the other ones follow exactly the same lines. On the one hand, using Proposition 4.4 together with the fact that, under the assumptions,

$$\sup_{n\geq 1}(\sqrt{h_n/\Delta_n})h_n^{1/q} < +\infty$$

we deduce from (4.7), (4.8) and (4.9) that, for $p \in \{1/2\} \cup [3, +\infty[$

$$\sqrt{\frac{h_n}{\Delta_n}}\Lambda_5^n(t) \xrightarrow[n\to+\infty]{\mathbb{P}} 0, \quad \sqrt{\frac{h_n}{\Delta_n}}\Lambda_1^n(t) \xrightarrow[n\to+\infty]{\mathbb{P}} 0, \sqrt{\frac{h_n}{\Delta_n}}\Lambda_3^n(t) \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \quad \text{since } \Delta_n^{\frac{p-2}{2}\wedge\frac{1}{2}} = \Delta_n^{\frac{1}{2}} \text{ if } p \geq 3.$$

On the other hand, under the assumptions of Theorems 4.1(i) and 4.2(i), one deduces from Proposition 4.3(i) and (ii) respectively that,

$$\sqrt{\frac{h_n}{\Delta_n}}\left(\Lambda_2^n(t) + \Lambda_4^n(t)\right) \xrightarrow[n\to+\infty]{\mathcal{L}-s} f(t,p)U,$$

Therefore, (4.4) and (4.2) follow from Lemma 4.1 applied with $Y = 0$ and from the decomposition of the error stated statement (4.6). □

Theorems 4.1 and 4.2 allow us to build a confidence region to estimate, for all t, parameter σ_t. The main remark is that the length of the asymptotic confidence intervals of σ_t, is about $r_n^{-\frac{1}{2}}\frac{\sqrt{m_{2p}-m_p^2}}{pm_p}$, and this length order is unhappily increasing with p, so it could be not so good to use $p > 2$.

II Large deviation Principle for drift parameter

II.1 Maximum likelihood estimator of drift parameter

In the Brownian setting, the definition of the maximum likelihood estimator of the drift parameter can be written:

$$\widehat{\theta}_T = \frac{\int_0^T X_t \,\mathrm{d}X_t}{\int_0^T X_t^2 \,\mathrm{d}t}. \tag{4.10}$$

In fact, in this setting, these quantities are perfectly defined in the sense of Itô's calculus. This is no more true in the fractional Brownian motion setting where the numerator of (4.10) is *a priori* not well-defined. To deal with fractional Brownian motion, we make use of the Gaussian martingale [31]:

$$M_t = \int_0^t w(t,s)\,\mathrm{d}B_s^H, \qquad t > 0,$$

where w is a weighting function defined for any $0 < s < t$, by $w(t,s) = w_H^{-1}s^{-H+1/2}(t-s)^{-H+1/2}$ with w_H a positive normalization constant. Its quadratic variation is

$$<M>_t = \lambda_H^{-1}t^{2-2H}.$$

The problem rewrites by means of

$$Y_t = \int_0^t w(t,s)\,\mathrm{d}X_s = \theta\int_0^t w(t,s)X_s\,\mathrm{d}s + M_t. \tag{4.11}$$

In [26] authors have shown that (4.11) can be rewritten

$$Y_t = \theta\int_0^t Q_s\,\mathrm{d}<M>_s + M_t$$

where (Q_t) satisfies, for any $t > 0$,

$$Q_t = \frac{l_H}{2}\left(t^{2H-1}Y_t + \int_0^t s^{2H-1}\,\mathrm{d}Y_s\right).$$

The score function (derivative of the log-likelihood from the observations on the interval $[0,T]$), is given by:

$$\Sigma_T(\theta) = \int_0^T Q_t \, \mathrm{d}Y_t - \theta \int_0^T Q_t^2 \, \mathrm{d}{<}M{>}_t \, .$$

One deduces the expression of the maximum likelihood estimator of θ, solution of $\Sigma_T(\theta) = 0$:

$$\widehat{\theta}_T = \frac{\int_0^T Q_t \, \mathrm{d}Y_t}{\int_0^T Q_t^2 \, \mathrm{d}{<}M{>}_t}.$$

Remark 4.5 *In the sequel, one supposes $1/2 < H < 1$. It is not a restriction of the case $0 < H < 1$ in virtue of Jost's formula [25] of transformation of process (B_t^H) in (B_t^{1-H}):*

$$B_t^H = \left(\frac{2H}{\Gamma(2H)\Gamma(3-2H)} \right)^{1/2} \int_0^t (t-s)^{2H-1} \, \mathrm{d}B_s^{1-H}.$$

Remark 4.6 *In [11] and in [7] a study of the Large Deviation principle for the energy defined as:*

$$S_T = \int_0^T X_t^2 \, \mathrm{d}t, \qquad \textit{in the Brownian setting,}$$

$$S_T = \int_0^T Q_t^2 \, \mathrm{d}{<}M{>}_t \qquad \textit{in the fractional Brownian setting}$$

is also proposed. The results are stated in paragraph II.5.

II.2 Two questions, one tool

Definition 4.2 ([15]) *We say that a family of real random variables (Z_T) satisfies a Large Deviation Principle (LDP) with rate function I, if I is a lower semi-continuous function from $\mathbb{R}$ to $[0,+\infty]$ such that, for any closed set $F \subset \mathbb{R}$,*

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}(Z_T \in F) \leq - \inf_{x\in F} I(x),$$

while for any open set $G \subset \mathbb{R}$,

$$- \inf_{x\in G} I(x) \leq \liminf_{T\to\infty} \frac{1}{T} \log \mathbb{P}(Z_T \in G).$$

Moreover, I is a good rate function if its level sets are compact subsets of $\mathbb{R}$.

A classical tool for proving an LDP is the normalized cumulant generating function. In our setting, for an LDP of S_T and $\widehat{\theta}_T$ we make use of

$$\mathcal{L}_T(a,b) = \frac{1}{T} \log \mathbb{E}[\exp(\mathcal{Z}_T(a,b))]$$

where, for any $(a,b) \in \mathbb{R}$,

$$\mathcal{Z}_T(a,b) = a \int_0^T X_t \, \mathrm{d}X_t + b \int_0^T X_t^2 \, \mathrm{d}t, \qquad \text{in Brownian setting,}$$

$$\mathcal{Z}_T(a,b) = a \int_0^T Q_t \, \mathrm{d}Y_t + b \int_0^T Q_t^2 \, \mathrm{d}{<}M{>}_t, \qquad \text{in fractional Brownian setting.}$$

The random variable $\mathcal{Z}_T(a,b)$ allows us an unified presentation of our results. In order to establish a LDP for S_T and $\widehat{\theta}_T$, it is enough to prove an LDP for $\mathcal{Z}_T(0,a)$ and $\mathcal{Z}_T(a,-ca)$, respectively. As a matter of fact, we have for all $a, c \geq 0$,

$$\mathbb{P}(S_T \geq cT) = \mathbb{P}(Z_T(0,a) \geq acT) \qquad \text{and} \qquad \mathbb{P}(\widehat{\theta}_T \geq c) = \mathbb{P}(Z_T(a,-ac) \geq 0).$$

The following lemmas provide an asymptotic expansion for $\mathcal{L}_T$ which enlightens the role of the limit $\mathcal{L}$ of $\mathcal{L}_T$ for the LDP, as well as the first order terms $\mathcal{H}$ and $\mathcal{K}_T$ for the sharp LDP.

Lemma 4.4 (Stable Brownian case [11]) *Let Δ be the effective domain of the limit $\mathcal{L}$ of $\mathcal{L}_T$*

$$\Delta = \Big\{(a,b) \in \mathbb{R}^2 \ / \ \theta^2 - 2b > 0 \text{ and } \sqrt{\theta^2 - 2b} > a + \theta\Big\}$$

Then, for any (a,b) in the interior of Δ, denote

$$\varphi(b) = \sqrt{\theta^2 - 2b} \qquad \text{and} \qquad \tau(a,b) = \varphi(b) - (a+\theta),$$

we have the decomposition

$$\mathcal{L}_T(a,b) = \mathcal{L}(a,b) + \frac{1}{T}\mathcal{H}(a,b) + \frac{1}{T}\mathcal{R}_T(a,b),$$

where

$$\mathcal{L}(a,b) = -\frac{1}{2}(a+\theta+\varphi(b)), \tag{4.12}$$

$$\mathcal{H}(a,b) = -\frac{1}{2}\log\left(\frac{\tau(a,b)}{2\varphi(b)}\right), \tag{4.13}$$

$$\mathcal{R}_T(a,b) = -\frac{1}{2}\log\left(1 + \frac{2\varphi(b) - \tau(a,b)}{\tau(a,b)}e^{-2T\varphi(b)}\right). \tag{4.14}$$

Finally, the remainder $\mathcal{R}_T(a,b)$ goes exponentially fast to 0:

$$\mathcal{R}_T(a,b) = \mathcal{O}(e^{-2T\varphi(b)}).$$

Lemma 4.5 (Stable fractional Brownian case [7, 26]) *Let Δ_H be the effective domain of the limit $\mathcal{L}$ of $\mathcal{L}_T$*

$$\Delta_H = \Big\{(a,b) \in \mathbb{R}^2 \ / \ \theta^2 - 2b > 0 \text{ and } \sqrt{\theta^2 - 2b} > \max(a+\theta; -\delta_H(a+\theta))\Big\}$$

where $\delta_H = (1-\sin(\pi H))/(1+sin(\pi H))$. Then, for any (a,b) in the interior of Δ_H, denote $r_T(b) = r_H(\varphi(b)T/2)\exp(-T\varphi(b)) - 1$, we have the decomposition

$$\mathcal{L}_T(a,b) = \mathcal{L}(a,b) + \frac{1}{T}\mathcal{H}(a,b) + \frac{1}{T}\mathcal{K}_T(a,b) + \frac{1}{T}\mathcal{R}_T(a,b), \tag{4.15}$$

where

$$\mathcal{K}_T(a,b) = -\frac{1}{2}\log\left(1 + \frac{(2\varphi(b) - \tau(a,b))}{2\varphi(b)}r_T(b)\right),$$

with φ and τ defined by (4.4) and $\mathcal{L}(a,b)$ and $\mathcal{H}(a,b)$ defined by (4.12) and (4.13) respectively. Finally, the remainder is

$$\mathcal{R}_T(a,b) = -\frac{1}{2}\log\left(1 + \frac{(2\varphi(b) - \tau(a,b))^2}{\tau(a,b)(2\varphi(b) + r_T(b)(2\varphi(b) - \tau(a,b)))}e^{-2T\varphi(b)}\right),$$

with r_H defined for all $z \in \mathbb{C}$ with $|\arg z| < \pi$ by

$$r_H(z) = \frac{\pi z}{\sin(\pi H)}\Big(I_H(z)I_{1-H}(z) + I_{-H}(z)I_{H-1}(z)\Big),$$

here, I_H is the modified Bessel function of the first kind [28].

Sketch of the proof. By Girsanov's theorem, $\mathcal{L}_T(a,b)$ can be rewritten as

$$\begin{aligned}\mathcal{L}_T(a,b) &= \frac{1}{T}\log\mathbb{E}\Big[\exp\Big(a\int_0^T Q_t\,\mathrm{d}Y_t + bS_T\Big)\Big],\\ &= \frac{1}{T}\log\mathbb{E}_\varphi\Big[\exp\Big((a+\theta-\varphi)\int_0^T Q_t\,\mathrm{d}Y_t + \frac{1}{2}(2b-\theta^2+\varphi^2)S_T\Big)\Big],\end{aligned}$$

for all $\varphi \in \mathbb{R}$, where $\mathbb{E}_\varphi$ stands for the expectation after the usual change of probability

$$\frac{d\mathbb{P}_\varphi}{d\mathbb{P}} = \exp\Big((\varphi-\theta)\int_0^T Q_t\,\mathrm{d}Y_t - \frac{1}{2}(\varphi^2-\theta^2)S_T\Big).$$

If $\underline{\theta^2 - 2b > 0}$, we can choose $\varphi = \sqrt{\theta^2 - 2b}$ and $\tau = \varphi - (a + \theta)$ which leads to

$$\mathcal{L}_T(a, b) = \frac{1}{T} \log \mathbb{E}_\varphi \Big[\exp\Big(-\tau \int_0^T Q_t \, \mathrm{d}Y_t \Big) \Big].$$

By Itô's formula, we also have

$$\int_0^T Q_t \, \mathrm{d}Y_t = \frac{1}{2} \Big(l_H Y_T \int_0^T t^{2H-1} \, \mathrm{d}Y_t - T \Big).$$

Consequently, we obtain that

$$\mathcal{L}_T(a, b) = \frac{\tau}{2} + \frac{1}{T} \log \mathbb{E}_\varphi \Big[\exp\Big(-\frac{\tau l_H}{2} Y_T \int_0^T t^{2H-1} \, \mathrm{d}Y_t \Big) \Big].$$

Under the new probability $\mathbb{P}_\varphi$, the pair $(Y_T, \int_0^T t^{2H-1} \, \mathrm{d}Y_t)$ is Gaussian with mean zero and covariance matrix $\Gamma_T(\varphi)$. If $\underline{\tau > 0}$, relation (5.12) of [26] gives an explicit expression of the expectation. The rest of the proof are mainly computations on Gaussian distribution. These computations are possible only if

$$1 + \frac{(2\varphi - \tau)}{2\varphi} r_T > 0$$

leading to $\underline{\sqrt{\theta^2 - 2b} > -\delta_H(a + \theta)}$. The underlined expressions leads to the expression of the domain. □

Remark 4.7. *By use of the duplication formula for the gamma function [28], one can realize that if $H = 1/2$, $r_H(z) = e^{2z} + e^{-2z}$ which immediately leads to $r_T(b) = e^{-2T\varphi(b)}$. Consequently, in that particular case, $\mathcal{K}_T(a, b)$ as well as $\mathcal{R}_T(a, b)$ go exponentially fast to zero and we find again Lemma 2.1 of [11] which is the keystone for all results of [11].*

Lemma 4.6 (Unstable Brownian case [8]) *Let Δ be the effective domain of the limit $\mathcal{L}$ of $\mathcal{L}_T$*

$$\Delta = \Big\{ (a, b) \in \mathbb{R}^2 \ / \ \theta^2 - 2b > 0 \text{ and } \sqrt{\theta^2 - 2b} > a + \theta \Big\}$$

Then, for any (a, b) in the interior of Δ, denote

$$\varphi(b) = -\sqrt{\theta^2 - 2b} \qquad \text{and} \qquad \tau(a, b) = a + \theta - \varphi(b),$$

we have the decomposition (4.12) *with*

$$\mathcal{L}(a, b) = -\frac{1}{2}(a + \theta - \varphi(b)), \tag{4.16}$$

$$\mathcal{H}(a, b) = -\frac{1}{2} \log \left(\frac{2\varphi(b) + \tau(a, b)}{2\varphi(b)} \right), \tag{4.17}$$

$$\mathcal{R}_T(a, b) = -\frac{1}{2} \log \left(1 - \frac{\varphi(b)}{2\varphi(b) + \tau(a, b)} e^{2T\varphi(b)} \right).$$

Finally, the remainder $\mathcal{R}_T(a, b)$ goes exponentially fast to 0:

$$\mathcal{R}_T(a, b) = \mathcal{O}(e^{2T\varphi(b)}).$$

Sketch of the proof. We have:

$$\begin{aligned} \mathcal{L}_T(a, b) &= \frac{1}{T} \log \mathbb{E} \left[\exp\Big(a \int_0^T X_t \, \mathrm{d}X_t + b \int_0^T X_t^2 \, \mathrm{d}t \Big) \right], \\ &= \frac{1}{T} \log \mathbb{E}_\varphi \left[\exp\Big((a + \theta - \varphi) \int_0^T X_t \, \mathrm{d}X_t + \frac{1}{2}(2b - \theta^2 + \varphi^2) \int_0^T X_t^2 \, \mathrm{d}t \Big) \right], \end{aligned}$$

for all $\varphi \in \mathbb{R}$, where $\mathbb{E}_\varphi$ stands for the expectation after the change of measures

$$\frac{d\mathbb{P}_\varphi}{d\mathbb{P}} = \exp\Big((\varphi - \theta) \int_0^T X_t \, \mathrm{d}X_t - \frac{1}{2}(\varphi^2 - \theta^2) \int_0^T X_t^2 \, \mathrm{d}t \Big).$$

Choosing $\varphi = \varphi(b)$ such that $\tau(a,b) = a + \theta - \varphi(b) < 0$, we obtain that

$$\mathcal{L}_T(a,b) = \frac{1}{T} \log \mathbb{E}_\varphi \left[\exp\Big(\tau(a,b) \int_0^T X_t \, \mathrm{d}X_t \Big) \right].$$

For this, consider $(a,b) \in \Delta = \{(a,b) \in \mathbb{R}, \quad \theta^2 - 2b > 0, \quad a + \theta < \sqrt{\theta^2 - 2b}\}$ and choose $\varphi = \varphi(b)$ with $\varphi^2(b) = \theta^2 - 2b$. Under the measure $\mathbb{P}_\varphi$, X_T is a Gaussian random variable with zero mean and variance $\sigma_T^2(b)$ given by

$$\sigma_T^2(b) = -\frac{1 - \exp(2\varphi(b)T)}{2\varphi(b)}.$$

Which converges by taking $\varphi(b) = -\sqrt{\theta^2 - 2b}$. □

II.3 Maximum likelihood estimator

Consistency and asymptotic normality

Proposition 4.5 (See [30] for Brownian case and [26, 32, 33] for fractional Brownian case.) *In the stable, unstable, and explosive cases of the Brownian setting and in the stable case of fractional Brownian setting, we have:*

$$\widehat{\theta}_T \xrightarrow[T\to\infty]{a.s.} \theta$$

In the Brownian setting, the asymptotic normality is totally different in the three situations.

Proposition 4.6

- *If $\theta < 0$, the process (X_t) is positive recurrent and [30],*

$$\sqrt{T}(\widehat{\theta}_T - \theta) \xrightarrow[T\to\infty]{\mathcal{L}} \mathcal{N}(0, -2\theta).$$

- *If $\theta = 0$, the process (X_t) is null recurrent and [21],*

$$T(\widehat{\theta}_T - \theta) \xrightarrow[T\to\infty]{\mathcal{L}} \frac{\int_0^1 B_t \, \mathrm{d}B_t}{\int_0^1 B_t^2 \, \mathrm{d}t} = \frac{B_1^2 - 1}{2\int_0^1 B_t^2 \, \mathrm{d}t}$$

- *If $\theta > 0$, the process (X_t) is transient and [20, 27],*

$$\exp(\theta T)(\widehat{\theta}_T - \theta) \xrightarrow[T\to\infty]{\mathcal{L}} 2\theta \left(\frac{Y}{Z}\right)$$

  where Y, Z are two independent Gaussian $\mathcal{N}(0,1)$ random variables which implies that the limiting ratio Y/Z has a Cauchy distribution.

Proposition 4.7 ([7, 12]) *In the stable case of the fractional Brownian setting, we also have the CLT*

$$\sqrt{T}\Big(\widehat{\theta}_T - \theta\Big) \xrightarrow[T\to\infty]{\mathcal{L}} \mathcal{N}\Big(0, -2\theta\Big).$$

Large deviations principle

Theorem 4.3 (See [22] for Brownian case and [7] for fractional Brownian case.). *The maximum likelihood estimator $(\widehat{\theta}_T)$ satisfies an LDP with good rate function*

$$I(c) = \begin{cases} -\dfrac{(c-\theta)^2}{4c} & \text{if } c < \dfrac{\theta}{3}, \\ 2c - \theta & \text{if } c \geq \dfrac{\theta}{3}. \end{cases}$$

Remark 4.8 *One can observe that the rate function I is totally free of the parameter H. For the energy, the situation is quite different (see Theorem 4.9).*

Sketch of the proof. In order to establish the large deviation properties of $(\widehat{\theta}_T)$, we shall make use of the auxiliary random variable defined for all $c \in \mathbb{R}$ by

$$Z_T(c) = \int_0^T Q_t \, \mathrm{d}Y_t - c \int_0^T Q_t^2 \, \mathrm{d}{<}M{>}_t,$$

where we recall that $\mathbb{P}(\widehat{\theta}_T \geq c) = \mathbb{P}(Z_T(c) \geq 0)$. Let

$$D_H = \Big\{ a \in \mathbb{R} \ / \ \theta^2 + 2ac > 0 \text{ and } \sqrt{\theta^2 + 2ac} > \max(a + \theta; -\delta_H(a + \theta)) \Big\}.$$

After some straightforward calculations, it is not hard to see that

$$D_H = \begin{cases}]a_1^H, a_2^H[& \text{if } \ c \leq \dfrac{\theta}{2}, \\ ]a_1^H, a^c[& \text{if } \ c > \dfrac{\theta}{2}, \end{cases}$$

where $a^c = 2(c - \theta)$. In addition, for all $a \in D_H$, let

$$L(a) = \mathcal{L}(a, -ca) = -\frac{1}{2}\left(a + \theta + \sqrt{\theta^2 + 2ac}\right).$$

The function L is not steep as the derivative of L is finite at the boundary of D_H. In this setting, $I(c) = -\inf_{a \in \overline{D_H}} L(a)$. Moreover, $L^{'}(a) = 0$ if and only if $a = a_c$ with $a_c = (c^2 - \theta^2)/(2c)$ and $a_c \in D_H$ whenever $c < \theta/3$. Then, when $c < \theta/3$, $I(c) = -L(a_c)$ and when $c > \theta/3$, $I(c) = -L(a^c)$ because L is decreasing. □

We shall now focus our attention on the explosive case $\theta > 0$. It immediately follows from (4.1) that

$$X_T = \exp(\theta T) \int_0^T \exp(-\theta t) \, \mathrm{d}B_t.$$

The Gaussian process $(Y_T = \exp(-\theta T) X_T)$ converges almost surely and in mean square to the Gaussian non-degenerate random variable

$$Y = \int_0^\infty \exp(-\theta t) \, \mathrm{d}B_t.$$

Hence, via Toeplitz's lemma

$$\frac{1}{\exp(2\theta T)} \int_0^T X_t^2 \, \mathrm{d}t \underset{T \to \infty}{\overset{a.s.}{\longrightarrow}} \frac{Y^2}{2\theta}.$$

Consequently, one can expect for $(\widehat{\theta}_T)$ an LDP with speed $\exp(2\theta T)$. However, $(\widehat{\theta}_T)$ is a sequence of self-normalized random variables and we shall show that $(\widehat{\theta}_T)$ satisfies an LDP similar to that of Theorem 4.3 with speed T.

Theorem 4.4 ([8]). *If $\theta > 0$, then $(\widehat{\theta}_T)$ satisfies an LDP with speed T and good rate function*

$$I(c) = \begin{cases} -\dfrac{(c-\theta)^2}{4c} & \textit{if} \quad c \leq -\theta, \\ \theta & \textit{if} \quad |c| < \theta, \\ 0 & \textit{if} \quad c = \theta, \\ 2c - \theta & \textit{if} \quad c > \theta. \end{cases} \tag{4.18}$$

Remark 4.9 *The unstable case $\theta = 0$ can be handled exactly as the explosive case $\theta > 0$ since Lemma 4.6 is also true in the unstable situation. Consequently, we directly obtain the LDP and SLDP for $(\widehat{\theta}_T)$ in the unstable case by replacing θ by 0 in the previous results.*

Sharp large deviations results

By the use of Lemmas 4.4, 4.5 and 4.6 we are able to perform sharp large deviations principles inspired by the well-known Bahadur-Rao Theorem [4] on the sample mean.

Remark 4.10 *In papers [7, 8, 11] all the constraints are explicit. For enlighten the writing of the manuscript, only the constants of interest are explicit.*

Theorem 4.5 ([11]) *Consider the Ornstein-Uhlenbeck process given by* (4.1) *where the drift parameter* $\theta < 0$. *The maximum likelihood estimator* $(\hat{\theta}_T)$ *satisfies an SLDP.*
a) For all $c < \theta$, *there exists a sequence* $(b_{c,k})$ *such that, for any* $p > 0$ *and* T *large enough*

$$\mathbb{P}(\widehat{\theta}_T \leq c) = \frac{-\exp(-TI(c) + H^s(c))}{a_c \sigma_c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{b_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where

$$a_c = \frac{c^2 - \theta^2}{2c} \qquad \text{and} \qquad \sigma_c^2 = -\frac{1}{2c}$$

while, for all $\theta < c < \theta/3$,

$$\mathbb{P}(\widehat{\theta}_T \geq c) = \frac{\exp(-TI(c) + H^s(c))}{a_c \sigma_c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{b_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right].$$

b) For all $c > \theta/3$ *with* $c \neq 0$, *there exists a sequence* $(d_{c,k})$ *such that, for any* $p > 0$ *and* T *large enough*

$$\mathbb{P}(\widehat{\theta}_T \geq c) = \frac{\exp(-TI(c) + K^s(c))}{a_c \sigma_c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{d_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where

$$a_c = 2(c - \theta) \qquad \text{and} \qquad \sigma_c^2 = \frac{c^2}{2(2c - \theta)^3}$$

c) For $c = \theta/3$, *there exists a sequence* (e_k) *such that, for any* $p > 0$ *and* T *large enough*

$$\mathbb{P}(\widehat{\theta}_T \geq c) = \frac{\exp(-TI(c))}{2\pi T^{1/4}} \frac{\Gamma(1/4)}{a_\theta^{3/4} \sigma_\theta} \left[1 + \sum_{k=1}^{2p} \frac{e_k}{(\sqrt{T})^k} + \mathcal{O}\Big(\frac{1}{T^p \sqrt{T}}\Big) \right]$$

where

$$a_\theta = -\frac{4\theta}{3} \qquad \text{and} \qquad \sigma_\theta^2 = -\frac{3}{2\theta}.$$

d) Finally, for $c = 0$, *for any* $p > 0$ *and for* T *large enough*

$$\mathbb{P}(\widehat{\theta}_T \geq 0) = 2\frac{\exp(-TI(c))}{\sqrt{2\pi T}\sqrt{-2\theta}} \left[1 + \sum_{k=1}^{p} \frac{(2k)!}{2^{2k}\theta^k T^k k!} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right].$$

Theorem 4.6 ([8]) *Consider the Ornstein-Uhlenbeck process given by* (4.1) *where the drift parameter* $\theta > 0$. *The maximum likelihood estimator* $(\hat{\theta}_T)$ *satisfies an SLDP.*
a) For all $c < -\theta$, *there exists a sequence* $(b_{c,k})$ *such that, for any* $p > 0$ *and* T *large enough*

$$\mathbb{P}(\widehat{\theta}_T \leq c) = \frac{-\exp(-TI(c) + H^e(a))}{a_c \sigma_c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{b_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where

$$a_c = \frac{c^2 - \theta^2}{2c} \qquad \text{and} \qquad \sigma_c^2 = -\frac{1}{2c}$$

b) For all $c > \theta$, *there exists a sequence* $(d_{c,k})$ *such that, for any* $p > 0$ *and* T *large enough*

$$\mathbb{P}(\widehat{\theta}_T \geq c) = \frac{\exp(-TI(c) + K^e(c))}{a_c \sigma_c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{d_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where

$$a_c = 2(c-\theta) \qquad \text{and} \qquad \sigma_c^2 = \frac{c^2}{2(2c-\theta)^3}$$

c) For all $|c| < \theta$ with $c \neq 0$, there exists a sequence $(e_{c,k})$ such that, for any $p > 0$ and T large enough

$$\mathbb{P}(\widehat{\theta}_T \leq c) = \frac{\exp(-TI(c) + J^e(c))}{a_c \sigma_c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{e_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where

$$a_c = \frac{\theta}{c+\theta} \qquad \text{and} \qquad \sigma_c^2 = \frac{c^2}{2\theta^3}$$

d) For $c = -\theta$, there exists a sequence (f_k) such that, for any $p > 0$ and T large enough

$$\mathbb{P}(\widehat{\theta}_T \leq c) = \frac{\exp(-TI(c))}{2\pi T^{1/4}} \frac{\Gamma(1/4)}{a_\theta^{3/4} \sigma_\theta} \left[1 + \sum_{k=1}^{2p} \frac{f_k}{(\sqrt{T})^k} + \mathcal{O}\Big(\frac{1}{T^p \sqrt{T}}\Big) \right]$$

where

$$a_\theta = \sqrt{\theta} \qquad \text{and} \qquad \sigma_\theta^2 = \frac{1}{2\theta}.$$

e) Finally, for $c = 0$, for any $p > 0$ and for T large enough

$$\mathbb{P}(\widehat{\theta}_T \leq 0) = 2\frac{\exp(-TI(c))\sqrt{2\theta T}}{\sqrt{2\pi}} \left[1 + \sum_{k=1}^{p} \frac{(-1)^k (\theta T e^{-2\theta T})^k}{(2k+1)k!} + \mathcal{O}\Big((Te^{-2\theta T})^{p+1}\Big) \right].$$

Theorem 4.7 ([8]) *Consider the Ornstein-Uhlenbeck process given by* (4.1) *where the drift parameter $\theta = 0$. The maximum likelihood estimator $(\hat{\theta}_T)$ satisfies an SLDP.*
a) For all $c < 0$, there exists a sequence $(b_{c,k})$ such that, for any $p > 0$ and T large enough

$$\mathbb{P}(\widehat{\theta}_T \leq c) = \frac{-2\exp(-TI(c))}{a_c \sigma_c \sqrt{6\pi T}} \left[1 + \sum_{k=1}^{p} \frac{b_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where $a_c = c/2$ and $\sigma_c^2 = -1/(2c)$.
b) For all $c > 0$, there exists a sequence $(d_{c,k})$ such that, for any $p > 0$ and T large enough

$$\mathbb{P}(\widehat{\theta}_T \geq c) = \frac{2\exp(-TI(c))}{a_c \sigma_c \sqrt{6\pi T}} \left[1 + \sum_{k=1}^{p} \frac{d_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where $a_c = 2c$ and $\sigma_c^2 = 1/(16c)$.

Theorem 4.8 ([7]) *Consider the Ornstein-Uhlenbeck process given by* (4.1) *driven by a fractional Brownian motion where the drift parameter $\theta < 0$. The maximum likelihood estimator $(\hat{\theta}_T)$ satisfies an SLDP.*
a) For all $\theta < c < \theta/3$, there exists a sequence $(b_{c,k}^H)$ such that, for any $p > 0$ and T large enough,

$$\mathbb{P}(\hat{\theta}_T \geq c) = \frac{\exp(-TI(c) + J^f(c) + K_H^f(c))}{\sigma_c a_c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{b_{c,k}^H}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

while, for $c < \theta$,

$$\mathbb{P}(\hat{\theta}_T \leq c) = -\frac{\exp(-TI(c) + J^f(c) + K_H^f(c))}{\sigma_c a_c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{b_{c,k}^H}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where

$$a_c = \frac{c^2 - \theta^2}{2c} \qquad \text{and} \qquad \sigma_c^2 = -\frac{1}{2c}.$$

b) For all $c > \theta/3$ with $c \neq 0$, there exists a sequence $(d_{c,k}^H)$ such that, for any $p > 0$ and T large enough,

$$\mathbb{P}(\hat{\theta}_T \geq c) = \frac{\exp(-TI(c) + P^f(c))\sqrt{\sin(\pi H)}}{\sigma^c a^c \sqrt{2\pi T}} \left[1 + \sum_{k=1}^{p} \frac{d_{c,k}^H}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big) \right]$$

where

$$a^c = 2(c-\theta) \qquad \text{and} \qquad (\sigma^c)^2 = \frac{c^2}{2(2c-\theta)^3}.$$

c) For $c = 0$, for any $p > 0$ and for T large enough,

$$\mathbb{P}(\hat{\theta}_T \geq 0) = 2\,\frac{\exp(-TI(c))\sqrt{\sin(\pi H)}}{\sqrt{2\pi T}\sqrt{-2\theta}}\left[1 + \sum_{k=1}^{p}\frac{d_k^H}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big)\right]$$

d) For $c = \theta/3$, there exists a sequence (d_k^H) such that, for any $p > 0$ and T large enough

$$\mathbb{P}(\hat{\theta}_T \geq \frac{\theta}{3}) = \frac{\exp(-TI(c))\,\Gamma\left(\frac{1}{4}\right)}{4\pi T^{1/4}\,a_\theta^{3/4}\,\sigma_\theta}\sqrt{\sin(\pi H)}\left[1 + \sum_{k=1}^{p}\frac{e_k^H}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big)\right]$$

where a_θ and σ_θ are given by

$$a_\theta = -\frac{4\theta}{3} \qquad \text{and} \qquad \sigma_\theta^2 = -\frac{3}{2\theta}.$$

II.4 Idea of the proofs of the SLDP

The proofs of those theorems follow the same lines. In order to give an idea of these lines, we will focus on the explosive case in the Brownian setting of Theorem 4.6. The case $c = 0$ is really different of the other ones and is treated by hand in [8]. In the sequel, we avoid this case and assume $c \neq 0$.

The key-point

As we have already seen when we dealt with Large Deviation Principle, $I(c) = -\mathcal{L}(a_c)$ where a_c, belongs to the effective domain D_c. In order to reach a Sharp Large Deviation Principle, we have to distinguish two situations: the "easy" one ($c < -\theta$) for which $a_c \in D_c$ and the "hard" for which a_c is on the border. In this "hard" case it is necessary to make use of a slight modification of the strategy of time varying change of probability proposed by Bryc and Dembo [13]. For this the key point is the following Lemma:

Lemma 4.7

- *In each "hard case" there exists a unique family (a_T), which belongs to the interior of D_c and converges to its border a_c as T goes to infinity where $a_c = 2(c-\theta)$ for all $c > \theta$ or 0 elsewhere.*
- *Moreover a_T is solution of the implicit equation*

$$\mathcal{L}'(a) + \frac{1}{T}\mathcal{H}'(a) = 0$$

where $\mathcal{L}(a) = \mathcal{L}(a, -ac)$ and $\mathcal{H}(a) = \mathcal{H}(a, -ac)$ are given by (4.16) and (4.17).

- *Finally, one can find sequences (a_k) (different for each case) such that, for any $p > 0$ and T large enough,*

$$a_T = \sum_{k=0}^{p}\frac{a_k}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big), \qquad \text{for all } c > \theta,$$

$$a_T = \sum_{k=1}^{p}\frac{a_k}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big), \qquad \text{for all } |c| < \theta \text{ and } c \neq 0,$$

$$a_T = \sum_{k=1}^{2p}\frac{a_k}{(\sqrt{T})^k} + \mathcal{O}\Big(\frac{1}{T^p\sqrt{T}}\Big) \qquad \text{for all } c = -\theta.$$

Splitting of the problem

Let $\alpha_T = a_c$ if $c < -\theta$ and $\alpha_T = a_T$ otherwise. Consider the change of probability

$$\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{P}} = \exp\Big(\alpha_T Z_T(c) - T\mathcal{L}_T(\alpha_T)\Big)$$

and denote by $\mathbb{E}_T$ the expectation associated with $\mathbb{P}_T$. We clearly have

$$\begin{aligned}\mathbb{P}(\widehat{\theta}_T \leq c) &= \mathbb{P}(Z_T(c) \leq 0) = \mathbb{E}[\,\mathbb{I}_{\{Z_T(c)\leq 0\}}],\\ &= \mathbb{E}_T\Big[\exp(-\alpha_T Z_T(c) + T\mathcal{L}_T(\alpha_T))\,\mathbb{I}_{\{Z_T(c)\leq 0\}}\Big],\\ &= \exp\Big(T\mathcal{L}_T(\alpha_T)\Big)\mathbb{E}_T\Big[\exp(-\alpha_T Z_T(c))\,\mathbb{I}_{\{Z_T(c)\leq 0\}}\Big].\end{aligned}$$

Consequently, we can split $\mathbb{P}(\widehat{\theta}_T \leq c)$ into two terms, $\mathbb{P}(\widehat{\theta}_T \leq c) = A_T B_T$ with

$$\begin{aligned}A_T &= \exp(T\mathcal{L}_T(\alpha_T)),\\ B_T &= \mathbb{E}_T[\exp(-\alpha_T Z_T(c))\,\mathbb{I}_{\{Z_T(c)\leq 0\}}].\end{aligned}$$

Expansion of A_T

The proof of the expansion of A_T is nothing but computations by means of Lemma 4.6 together with the expansions of a_T given by Lemma 4.7 and the definition (4.18) of I. The results are the following ones:

Lemma 4.8 *There exists sequences (γ_k) (different from a line to another) such that, for any $p > 0$ and T large enough,*

- *For all $c < -\theta$,*
$$A_T = \exp\left(-TI(c) + H^e(c)\right)\Big(1 + \mathcal{O}\Big(e^{2Tc}\Big)\Big),$$
- *For all $c > \theta$*
$$A_T = \exp\left(-TI(c) + P(c)\right)\sqrt{eT}\left[1 + \sum_{k=1}^{p}\frac{\gamma_k}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big)\right],$$
- *For all $|c| < \theta$ and $c \neq 0$*
$$A_T = \exp\left(-TI(c) + P(c)\right)\sqrt{eT}\left[1 + \sum_{k=1}^{p}\frac{\gamma_k}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big)\right],$$
- *For $c = -\theta$*
$$A_T = \exp\left(-TI(c)\right)(e\theta T)^{1/4}\left[1 + \sum_{k=1}^{2p}\frac{\gamma_k}{(\sqrt{T})^k} + \mathcal{O}\Big(\frac{1}{T^p\sqrt{T}}\Big)\right].$$

In the Brownian setting, the remainder $\mathcal{R}_T(a_T)$ goes to zero exponentially fast and thus does not contribute to the limit. This is no more true in the fractional Brownian setting. It is the reason why the decomposition (4.15) is different.

Expansion of the B_T

The proof of the expansion of B_T is more technical. The results are the following ones:

Lemma 4.9 *There exists a sequence (β_k) (different in the different cases) such that, for any $p > 0$ and T large enough,*

- *For all $c < -\theta$,*
$$B_T = \frac{\beta_0}{\sqrt{T}}\left[1 + \sum_{k=1}^{p}\frac{\beta_k}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big)\right],$$
- *For all $c > \theta$ and $|c| < \theta$ and $c \neq 0$*
$$B_T = \sum_{k=1}^{p}\frac{\delta_k}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big),$$
- *For $c = -\theta$*
$$B_T = \sum_{k=1}^{2p}\frac{\delta_k}{(\sqrt{T})^k} + \mathcal{O}\Big(\frac{1}{T^p\sqrt{T}}\Big).$$

Sketch of the proof. First, denote

$$\beta_T = \begin{cases} \sigma_c\sqrt{T} & \text{if} \quad c < -\theta, \\ \sqrt{T} & \text{if} \quad c = -\theta, \\ T & \text{if} \quad |c| < \theta, \\ -T & \text{if} \quad c > \theta. \end{cases}$$

One can observe that we always have $\alpha_T\beta_T < 0$. Then whatever the case, we can rewrite

$$B_T = \mathbb{E}_T\Big[\exp(-\alpha_T\beta_T U_T)\,\mathbb{I}_{\{U_T\leq 0\}}\Big] \qquad \text{where} \qquad U_T = \frac{Z_T(c)}{\beta_T}.$$

Denote by Φ_T the characteristic function of U_T under $\mathbb{P}_T$. It is easily seen that for all $u \in \mathbb{R}$,

$$\Phi_T(u) = \exp\left(T\mathcal{L}_T\left(\alpha_T + \frac{iu}{\beta_T}\right) - T\mathcal{L}_T(\alpha_T)\right).$$

Computations lead us to the following Lemmas:

Lemma 4.10 *For $c < -\theta$, the distribution of U_T under $\mathbb{P}_T$ converges, as T goes to infinity, to an $\mathcal{N}(0,1)$ distribution. Moreover, for any $p > 0$, there exist integers $q(p)$, $r(p)$ and a sequence $(\varphi_{k,l})$ independent of p, such that, for T large enough*

$$\Phi_T(u) = \exp\Big(-\frac{u^2}{2}\Big)\left[1 + \frac{1}{\sqrt{T}}\sum_{k=0}^{2p}\sum_{l=k+1}^{q(p)}\frac{\varphi_{k,l}u^l}{(\sqrt{T})^k} + \mathcal{O}\Big(\frac{\max(1,|u|^{r(p)})}{T^{p+1}}\Big)\right]$$

and the remainder $\mathcal{O}$ is uniform as soon as $|u| \leq sT^{1/6}$ with $s > 0$.

Lemma 4.11 *For $c > \theta$, the distribution of U_T under $\mathbb{P}_T$ converges, as T goes to infinity, to the distribution of $\gamma(N^2-1)$, where N is an $\mathcal{N}(0,1)$ random variable and $\gamma > 0$. Moreover, for any $p > 0$, there exist integers $q(p)$, $r(p)$, $s(p)$ and a sequence $(\varphi_{k,l,m})$ independent of p, such that, for T large enough*

$$\Phi_T(u) = \frac{\exp(-i\gamma u)}{\sqrt{1-2i\gamma u}}\exp\left(-\frac{\sigma_c^2u^2}{2T}\right)\left[1 + \sum_{k=1}^{p}\sum_{l=k+1}^{q(p)}\sum_{m=0}^{r(p)}\frac{\varphi_{k,l,m}u^l}{T^k(1-2i\gamma u)^m} + \mathcal{O}\Big(\frac{\max(1,|u|^{s(p)})}{T^{p+1}}\Big)\right]$$

the remainder $\mathcal{O}$ is uniform as soon as $|u| \leq sT^{2/3}$ with $s > 0$.

Lemma 4.12 *For $|c| < \theta$ with $c \neq 0$, the distribution of U_T under $\mathbb{P}_T$ converges, as T goes to infinity, to the distribution of $\gamma(N^2-1)$, where N is an $\mathcal{N}(0,1)$ random variable and $\gamma > 0$. Moreover, for any $p > 0$, there exist integers $q(p)$, $r(p)$, $s(p)$ and a sequence $(\varphi_{k,l,m})$ independent of p, such that, for T large enough*

$$\Phi_T(u) = \frac{\exp(-i\gamma u)}{\sqrt{1-2i\gamma u}}\exp\left(-\frac{\sigma_c^2u^2}{2T}\right)\left[1 + \sum_{k=1}^{p}\sum_{l=k+1}^{q(p)}\sum_{m=0}^{r(p)}\frac{\varphi_{k,l,m}u^l}{T^k(1-2i\gamma u)^m} + \mathcal{O}\Big(\frac{\max(1,|u|^{s(p)})}{T^{p+1}}\Big)\right]$$

the remainder $\mathcal{O}$ is uniform as soon as $|u| \leq sT^{2/3}$ with $s > 0$.

Lemma 4.13 *For $c = -\theta$, the distribution of U_T under $\mathbb{P}_T$ converges, as T goes to infinity, to the distribution of $\sigma_\theta N + \gamma_\theta(M^2-1)$, where N and M are two independent $\mathcal{N}(0,1)$ random variables, $\sigma^2 > 0$ and $\gamma_\theta > 0$. Moreover, for any $p > 0$, there exist integers $q(p)$, $r(p)$, $s(p)$ and a sequence $(\varphi_{k,l,m})$ independent of p, such that, for T large enough*

$$\Phi_T(u) = \frac{\exp\left(-i\gamma_\theta u\right)}{\sqrt{1-2i\gamma_\theta u}}\exp\left(-\frac{\sigma_\theta^2u^2}{2}\right)\left[1 + \frac{1}{\sqrt{T}}\sum_{k=0}^{2p}\sum_{l=k+1}^{q(p)}\sum_{m=0}^{r(p)}\frac{\varphi_{k,l,m}u^l}{(\sqrt{T})^k(1-2i\gamma_\theta u)^m} + \mathcal{O}\Big(\frac{\max(1,|u|^{s(p)})}{T^{p+1}}\Big)\right]$$

the remainder $\mathcal{O}$ is uniform as soon as $|u| \leq sT^{1/6}$ with $s > 0$.

The proof ends by an application of Parseval's formula which allows us to rewrite B_T

$$B_T = -\frac{1}{2\pi\alpha_T\beta_T}\int_{-\infty}^{\infty}\left(1+\frac{iu}{\alpha_T\beta_T}\right)^{-1}\Phi_T(u)\,\mathrm{d}u.$$

This step is possible if and only if for T large enough, Φ_T belongs to $\mathbb{L}^2(\mathbb{R})$. This is a very technical and difficult (especially in the fractional setting) part of the proof. Finally we split B_T into two terms, $B_T = C_T + D_T$ where

$$\begin{aligned} C_T &= -\frac{1}{2\pi\alpha_T\beta_T}\int_{|u|\le s_T}\left(1+\frac{iu}{\alpha_T\beta_T}\right)^{-1}\Phi_T(u)\,\mathrm{d}u, \\ D_T &= -\frac{1}{2\pi\alpha_T\beta_T}\int_{|u|> s_T}\left(1+\frac{iu}{\alpha_T\beta_T}\right)^{-1}\Phi_T(u)\,\mathrm{d}u. \end{aligned} \tag{4.19}$$

where s_T is chosen in such a way that there are positive constants C and $0<\nu<1$ satisfying

$$\min\left(\frac{Ts_T^2}{\beta_T^2}, \frac{T\sqrt{s_T}}{\sqrt{|\beta_T|}}\right) \ge CT^{\nu} \tag{4.20}$$

and there exist two positive constants d and D such that

$$|D_T| \le d\ T\ \exp(-DT^{\nu}). \tag{4.21}$$

We choose s_T large enough to satisfy (4.20) and small enough to enable us to intervene integral and summation into (4.19). Fortunately this is possible whatever the case. In fact,

- In the case $c<-\theta$, it works with $s_T = sT^{1/6}$ with $s>0$ and $\nu = 1/3$.
- In the other cases, it works with $s_T = sT^{2/3}$ with $s>0$ and $\nu = 1/3$.

The expansion of C_T thus follows from that of Φ_T and some tedious but standard calculations on the $\mathcal{N}(0,1)$ distribution if the case $c<-\theta$ or via a careful use of the contour integral lemma for the Gamma function given in Lemma 7.3 of [11] for the other cases. Finally, (4.21) tells us that the expansion of B_T is nothing but that of C_T which ends the proof. □

II.5 The energy

Results on the energy were investigated only in the stable case in the Brownian setting in [11] and in the fractional Brownian setting in [7]. In order to stress the proof of the large deviation principle, we focus on the fractional Brownian setting. For this, we shall make use of Lemma 4.5 with $a=0$ and $b=a$. On the one hand, let

$$D_H = \left\{a\in\mathbb{R}\ /\ \theta^2-2a>0 \text{ and } \sqrt{\theta^2-2a} > -\delta_H\theta\right\}.$$

It is not hard to see that $D_H =]-\infty, a_H[$ where

$$a_H = \frac{\theta^2}{2}(1-\delta_H^2).$$

Consequently, as $|\delta_H|<1$, one can observe that the origin always belongs to the interior of D_H. On the other hand, for all $a\in D_H$,

$$\mathcal{L}(a) = \mathcal{L}(0,a) = -\frac{1}{2}(\theta+\sqrt{\theta^2-2a}),$$

The main difficulty comparing to [11] is that the function L is not steep. Indeed, $\mathcal{L}'(a_H) = -1/(2\theta\delta_H)$. Moreover, for all $c>0$, $\mathcal{L}'(a)=c$ if and only if $a=a_c$ with $a_c = (4\theta^2c^2-1)/(8c^2)$. Hence, $a_c<a_H$ whenever $0<c<-1/(2\theta\delta_H)$. Denote by I the Fenchel-Legendre transform of the function $\mathcal{L}$, the large deviation properties of (S_T/T) states as follows.

Theorem 4.9. *The sequence (S_T/T) satisfies a LDP with good rate function*

$$I(c) = \begin{cases} \dfrac{(2\theta c+1)^2}{8c} & \text{if } 0 < c \leq -\frac{1}{2\theta\delta_H}, \\ \dfrac{c\theta^2}{2}(1-\delta_H^2) + \dfrac{\theta}{2}(1-\delta_H) & \text{if } c \geq -\frac{1}{2\theta\delta_H}, \\ +\infty & \text{otherwise.} \end{cases}$$

Remark 4.11 *In the particular case $H = 1/2$, then $\delta_H = 0$ and the LDP for (S_T/T) is exactly the one established by Bryc and Dembo [13] for general centered Gaussian processes.*

Theorem 4.10 *The sequence (S_T/T) satisfies a SLDP.*
a) For all $-1/(2\theta) < c < -1/(2\theta\delta_H)$, there exists a sequence $(b^H_{c,k})$ such that, for any $p > 0$ and T large enough,

$$\mathbb{P}(S_T \geq cT) = \frac{\exp(-TI(c) + J^f(c) + K^f_H(c))}{a_c\sigma_c\sqrt{2\pi T}}\left[1 + \sum_{k=1}^{p}\frac{b^H_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big)\right]$$

while, for $0 < c < -1/(2\theta)$,

$$\mathbb{P}(S_T \leq cT) = -\frac{\exp(-TI(c) + J^f(c) + K^f_H(c))}{a_c\sigma_c\sqrt{2\pi T}}\left[1 + \sum_{k=1}^{p}\frac{b^H_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big)\right]$$

where

$$a_c = \frac{4\theta^2c^2-1}{8c^2} \qquad \text{and} \qquad \sigma_c^2 = 4c^3.$$

b) For all $c > -1/(2\theta\delta_H)$, there exists a sequence $(d^H_{c,k})$ such that, for any $p > 0$ and T large enough

$$\mathbb{P}(S_T \geq cT) = \frac{\exp(-TI(c) + P^f_H(c))}{a_H\sigma_H\sqrt{2\pi T}}\left[1 + \sum_{k=1}^{p}\frac{d^H_{c,k}}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big)\right]$$

where

$$a_H = \frac{\theta^2(1-\delta_H^2)}{2} \qquad \text{and} \qquad \sigma_H^2 = -\frac{1}{2\theta^3\delta_H^3}. \tag{4.22}$$

c) For $c = -1/(2\theta\delta_H)$, there exists a sequence (d^H_k) such that, for any $p > 0$ and T large enough

$$\mathbb{P}(S_T \geq cT) = \frac{\exp(-TI(c) + K^f_H)\Gamma(1/4)}{2\pi a_H\sigma_H T^{1/4}}\left[1 + \sum_{k=1}^{2p}\frac{d^H_k}{(\sqrt{T})^k} + \mathcal{O}\Big(\frac{1}{T^p\sqrt{T}}\Big)\right]$$

where a_H and σ_H^2 are given by (4.22).

Sketch of the proof. The proof follows the same lines as these for the maximum likelihood estimator of θ. Let L_T be the normalized cumulant generating function of S_T. We can split $\mathbb{P}(S_T \geq cT)$ into two terms, $\mathbb{P}(S_T \geq cT) = A_T B_T$ with,

- when $a_c = (4\theta^2c^2-1)/(8c^2)$, belongs to the domain D_H, i.e. $-1/(2\theta) < c < -1/(2\theta\delta_H)$,

$$A_T = \exp(T(L_T(a_c) - ca_c)), \quad \text{and} \quad B_T = \mathbb{E}_T\Big[\exp(-a_c(S_T - cT))\,\mathbb{I}_{\{S_T \geq cT\}}\Big],$$

  where $\mathbb{E}_T$ stands for the expectation after the usual change of probability

$$\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{P}} = \exp\Big(a_cS_T - TL_T(a_c)\Big),$$

- when $a_c = (4\theta^2c^2-1)/(8c^2)$, is on the border of the domain D_H,

$$A_T = \exp\left(TL_T(a_T) - cTa_T\right), \quad \text{and} \quad B_T = \mathbb{E}_T\Big[\exp(-a_T(S_T - cT))\,\mathbb{I}_{\{S_T \geq cT\}}\Big],$$

where $\mathbb{E}_T$ stands for the expectation after the usual change of probability

$$\frac{\mathrm{d}\mathbb{P}_T}{\mathrm{d}\mathbb{P}} = \exp\Big(a_T S_T - T L_T(a_T)\Big),$$

and a_T is a unique element which belongs to the interior of D_H and converges to its border. After some tedious but straightforward calculations, we show that there exists a sequence (a_k) such that, for any $p > 0$ and T large enough,

$$a_T = \sum_{k=0}^{p} \frac{a_k}{T^k} + \mathcal{O}\Big(\frac{1}{T^{p+1}}\Big), \quad \text{or} \quad a_T = \sum_{k=0}^{2p} \frac{a_k}{(\sqrt{T})^k} + \mathcal{O}\Big(\frac{1}{T^p\sqrt{T}}\Big) \quad \text{in the case of equality.}$$

The proof now splits into two parts, the first one is devoted to the expansion of A_T which is nothing but computations while the second one gives the expansion of B_T which can be rewritten as

$$B_T = \mathbb{E}_T\Big[\exp(-a_c\sigma_c\sqrt{T}U_T)\,\mathbb{I}_{\{U_T\geq 0\}}\Big], \quad \text{where} \quad U_T = \frac{S_T - cT}{\sigma_c\sqrt{T}}, \quad \text{for all } -1/(2\theta) < c < -1/(2\theta\delta_H),$$

$$B_T = \mathbb{E}_T\Big[\exp(-a_T T U_T)\,\mathbb{I}_{\{U_T\geq 0\}}\Big], \quad \text{where} \quad U_T = \frac{S_T - cT}{T}, \quad \text{for all } c > -1/(2\theta\delta_H),$$

$$B_T = \mathbb{E}_T\Big[\exp(-a_T\sqrt{T}U_T)\,\mathbb{I}_{\{U_T\geq 0\}}\Big], \quad \text{where} \quad U_T = \frac{S_T - cT}{\sqrt{T}}, \quad \text{for } c = -1/(2\theta\delta_H).$$

Denoting Φ_T the characteristic function of U_T under $\mathbb{P}_T$, the hard question in the fractional setting is to show that for T large enough, Φ_T belongs to $L^2(\mathbb{R})$. In fact, in contrast with [11], it is impossible here to make use of the Karhunen-Loève expansion of the process (X_t) and we have to make computations by hand. The expansion of B_T follows from that of Φ_T, an application of Parseval's formula, standard calculus on the $\mathcal{N}(0,1)$ distribution and a careful use of the contour integral Lemma for the Gamma functions. □

III Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes

The maximum likelihood estimator of the parameter θ of the model (4.2) page 43 is given by

$$\widehat{\theta}_T = \frac{\int_0^T X_t\,\mathrm{d}X_t}{\int_0^T X_t^2\,\mathrm{d}t} = \frac{X_T^2 - T}{2\int_0^T X_t^2\,\mathrm{d}t}. \tag{4.23}$$

The estimation of the parameter ρ of the model (4.2) requires the evaluation of the residuals generated by the estimation of θ at stage T. For all $0 \leq t \leq T$, denote

$$\widehat{V}_t = X_t - \widehat{\theta}_T\Sigma_t \tag{4.24}$$

where

$$\Sigma_t = \int_0^t X_s\,\mathrm{d}s.$$

By analogy with (4.23) and on the basis of the residuals (4.24), we estimate ρ by

$$\widehat{\rho}_T = \frac{\widehat{V}_T^2 - T}{2\int_0^T \widehat{V}_t^2\,\mathrm{d}t}.$$

Therefore, we are in the position to define the continuous-time version of the discrete-time Durbin-Watson statistic [9, 16, 17, 18],

$$\widehat{D}_T = \frac{2\int_0^T \widehat{V}_t^2\,\mathrm{d}t - \widehat{V}_T^2 + T}{\int_0^T \widehat{V}_t^2\,\mathrm{d}t},$$

which clearly means that $\widehat{D}_T = 2(1-\widehat{\rho}_T)$. We shall make use of $\widehat{D}_T$ to build a serial correlation statistical test for the Ornstein-Uhlenbeck driven noise, that is to test whether or not $\rho = 0$ (not presented here).

Theorem 4.11 ([10]) *We have the almost sure convergences*

$$\widehat{\theta}_T \xrightarrow[T\to\infty]{a.s.} \theta^*, \qquad \widehat{\rho}_T \xrightarrow[T\to\infty]{a.s.} \rho^*,$$

where

$$\theta^* = \theta + \rho \qquad \text{and} \qquad \rho^* = \frac{\theta\rho(\theta+\rho)}{(\theta+\rho)^2 + \theta\rho}.$$

Sketch of the proof. The arguments are essentially strong law of large number for martingales, the fact that most of the processes introduced are positive recurrent and computations by means of Itô's formula. □

Theorem 4.12 ([10]) *If $\rho < 0$, we have the joint asymptotic normality*

$$\sqrt{T}\begin{pmatrix}\widehat{\theta}_T - \theta^* \\ \widehat{\rho}_T - \rho^*\end{pmatrix} \xrightarrow[T\to\infty]{\mathcal{L}} \mathcal{N}(0,\Gamma)$$

where the asymptotic covariance matrix

$$\Gamma = \begin{pmatrix}\sigma_\theta^2 & \ell \\ \ell & \sigma_\rho^2\end{pmatrix}$$

with

$$\sigma_\theta^2 = -2\theta^*, \qquad \ell = \frac{2\rho^*\left((\theta^*)^2 - \theta\rho\right)}{(\theta^*)^2 + \theta\rho} \quad \text{and} \quad \sigma_\rho^2 = -\frac{2\rho^*\left((\theta^*)^6 + \theta\rho\left((\theta^*)^4 - \theta\rho\left(2(\theta^*)^2 - \theta\rho\right)\right)\right)}{\left((\theta^*)^2 + \theta\rho\right)^3}.$$

In particular, we have

$$\sqrt{T}\left(\widehat{\theta}_T - \theta^*\right) \xrightarrow[T\to\infty]{\mathcal{L}} \mathcal{N}(0,\sigma_\theta^2),$$

and

$$\sqrt{T}\left(\widehat{\rho}_T - \rho^*\right) \xrightarrow[T\to\infty]{\mathcal{L}} \mathcal{N}(0,\sigma_\rho^2).$$

Sketch of the proof. The keypoint of the proof is an application of Central limit Theorem for continuous-time vector martingales. □

Theorem 4.13 *If $\rho = 0$, we have the convergence in distribution*

$$T\,\widehat{\rho}_T \xrightarrow[T\to\infty]{\mathcal{L}} \mathcal{W}$$

where the limiting distribution $\mathcal{W}$ is given by

$$\mathcal{W} = \frac{\int_0^1 B_s \,\mathrm{d}B_s}{\int_0^1 B_s^2 \,\mathrm{d}s} = \frac{B_1^2 - 1}{2\int_0^1 B_s^2 \,\mathrm{d}s}$$

and (B_t) is a standard Brownian motion.

Sketch of the proof. The result comes from the self-similarity property of the Brownian motion (W_t), Theorem 4.11, and the continuous mapping theorem. □

Remark 4.12 *The asymptotic behaviour of $\widehat{\rho}_T$ when $\rho < 0$ and $\rho = 0$ is closely related to the results previously established for the unstable discrete-time autoregressive process, see [14], [21], [35]. According to Corollary 3.1.3 of [14], we can express*

$$\mathcal{W} = \frac{\mathcal{T}^2 - 1}{2\mathcal{S}}$$

where $\mathcal{T}$ and $\mathcal{S}$ are given by the Karhunen-Loeve expansions

$$\mathcal{T} = \sqrt{2}\sum_{n=1}^{\infty}\gamma_n Z_n \qquad \text{and} \qquad \mathcal{S} = \sum_{n=1}^{\infty}\gamma_n^2 Z_n^2$$

with $\gamma_n = 2(-1)^n/((2n-1)\pi)$ and (Z_n) is a sequence of independent random variables with $\mathcal{N}(0,1)$ distribution.

Remark 4.13 *For all $0 \le t \le T$, the residuals $\widehat{V}_t$ given by (4.24) depend on $\widehat{\theta}_T$. It would have been more natural to make use of the estimator of θ at stage t instead of stage T, in order to produce a recursive estimate. In this situation, Theorem 4.11 still holds but we have been unable to prove Theorem 4.12.*

Bibliography

[1] Yacine Aït-Sahalia and Jean Jacod. Volatility estimators for discretely sampled Lévy processes. *Ann. Statist.*, 35(1):355–392, 2007.

[2] Yacine Aït-Sahalia and Jean Jacod. Testing for jumps in a discretely observed process. *Ann. Statist.*, 37(1):184–222, 2009.

[3] Alexander Alvarez, Fabien Panloup, Monique Pontier, and **Nicolas Savy**. Estimation of the instantaneous volatility. *Statistic inference for Stochastic processes*, 15:27–59, 2012.

[4] Raghu R. Bahadur and R. Ranga Rao. On deviations of the sample mean. *Ann. Math. Statist.*, 31:1015–1027, 1960.

[5] Ole E. Barndorff-Nielsen, Svend Erik Graversen, Jean Jacod, Mark Podolskij, and Neil Shephard. A central limit theorem for realised power and bipower variations of continuous semimartingales. In *From stochastic calculus to mathematical finance*, pages 33–68. Springer, Berlin, 2006.

[6] Ole E. Barndorff-Nielsen and Neil Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(2):253–280, 2002.

[7] Bernard Bercu, Laure Coutin, and **Nicolas Savy**. Sharp large deviations for the fractional Ornstein-Uhlenbeck process. *Theory of Probability and its Applications*, 55(4):575–610, 2011.

[8] Bernard Bercu, Laure Coutin, and **Nicolas Savy**. Sharp large deviation for the non-stationary ornstein-uhlenbeck process. *Stochastic Processes and their Applications*, 122(10):3393–3424, 2012.

[9] Bernard Bercu and Frédéric Proïa. A sharp analysis on the asymptotic behavior of the Durbin-Watson statistic for the first-order autoregressive process. *ESAIM Probab. Stat.*, 16, 2012.

[10] Bernard Bercu, Frédéric Proïa, and **Nicolas Savy**. On Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes. *Statistics and Probability Letters*, 85:36–44, 2014.

[11] Bernard Bercu and Alain Rouault. Sharp large deviations for the Ornstein-Uhlenbeck process. *Theory Probab. Appl.*, 46(1):1–19, 2002.

[12] Alexandre Brouste and Marina Kleptsyna. Asymptotic properties of MLE for partially observed fractional diffusion system. *Stat. Inference Stoch. Process.*, 13(1):1–13, 2010.

[13] Włodzimierz Bryc and Amir Dembo. Large deviations for quadratic functionals of Gaussian processes. *J. Theoret. Probab.*, 10(2):307–332, 1997. Dedicated to Murray Rosenblatt.

[14] Ngai H. Chan and Ching Z. Wei. Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.*, 16 (1):367–401, 1988.

[15] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.

[16] James Durbin and Geoffrey S. Watson. Testing for serial correlation in least squares regression. I. *Biometrika*, 37:409–428, 1950.

[17] James Durbin and Geoffrey S. Watson. Testing for serial correlation in least squares regression. II. *Biometrika*, 38:159–178, 1951.

[18] James Durbin and Geoffrey S. Watson. Testing for serial correlation in least squares regession. III. *Biometrika*, 58:1–19, 1971.

[19] Geoffrey K. Eagleson. Martingale convergence to mixtures of infinitely divisible laws. *Ann. Probability*, 3(3):557–562, 1975.

[20] Paul D. Feigin. Maximum likelihood estimation for continuous-time stochastic processes. *Advances in Appl. Probability*, 8(4):712–736, 1976.

[21] Paul D. Feigin. Some comments concerning a curious singularity. *J. Appl. Probab.*, 16(2):440–444, 1979.

[22] Danielle Florens-Landais and Huyên Pham. Large deviations in estimation of an Ornstein-Uhlenbeck model. *J. Appl. Probab.*, 36(1):60–77, 1999.

[23] Yaozhong Hu and Hongwei Long. Parameter estimation for Ornstein-Uhlenbeck processes driven by α-stable Lévy motions. *Commun. Stoch. Anal.*, 1(2):175–192, 2007.

[24] Jean Jacod. Asymptotic properties of realized power variations and related functionals of semi-martingales. *Stochastic Process. Appl.*, 118(4):517–559, 2008.

[25] Céline Jost. Transformation formulas for fractional Brownian motion. *Stochastic Process. Appl.*, 116(10):1341–1357, 2006.

[26] Marina L. Kleptsyna and Alain Le Breton. Statistical analysis of the fractional Ornstein-Uhlenbeck type process. *Stat. Inference Stoch. Process.*, 5(3):229–248, 2002.

[27] Yuri A. Kutoyants. *Statistical inference for ergodic diffusion processes.* Springer Series in Statistics. Springer-Verlag London Ltd., London, 2004.

[28] Nikolai N. Lebedev. *Special functions and their applications.* Revised English edition. Translated and edited by Richard A. Silverman. Prentice-Hall Inc., Englewood Cliffs, N.J., 1965.

[29] Dominique Lépingle. La variation d'ordre p des semi-martingales. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 36(4):295–316, 1976.

[30] Robert S. Liptser and Albert N. Shiryaev. *Statistics of random processes. II*, volume 6 of *Applications of Mathematics (New York).* Springer-Verlag, Berlin, 2001.

[31] Ilkka Norros, Esko Valkeila, and Jorma Virtamo. An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions. *Bernoulli*, 5(4):571–587, 1999.

[32] B. L. S. Prakasa Rao. Sequential estimation for fractional ornstein-uhlenbeck type process. *Sequential Anal.*, 23(1):33–44, 2004.

[33] B. L. S. Prakasa Rao. Estimation for translation of a process driven by fractional brownian motion. *Stoch. Anal. Appl.*, 23(6):1199–1212, 2005.

[34] Wim Schoutens. *Stochastic processes and orthogonal polynomials*, volume 146 of *Lecture Notes in Statistics.* Springer-Verlag, New York, 2000.

[35] John S. White. The limiting distribution of the serial correlation coefficient in the explosive case. *Ann. Math. Statist.*, 29:1188–1197, 1958.

[36] Jeannette H. C. Woerner. Power and multipower variation: inference for high frequency data. In *Stochastic finance*, pages 343–364. Springer, New York, 2006.

Part B-

Statistic applied to Biology and to Medical Research

Chapter 5

Models for patients' recruitment in clinical trials*

This chapter is a summary of the results obtained during the Ph.D. thesis preparation of Guillaume Mijoule. I have co-supervised this thesis with Prof. Laure Coutin (IMT). Guillaume Mijoule defended his Ph.D. June 3rd, 2013 and the examination board was Prof. Stephen Senn (CMML - referee) - Prof. Adeline Sansom (University of Paris V - referee) - Prof. Vladimir Anisimov (Quintiles - University of Glasgow) - Prof. Antoine Chambaz (University of Paris X) - Prof. Laure Coutin and myself (University of Toulouse III).

After a section devoted to general considerations on patient' recruitment, a second section deals with the models of paramount interest introduced by Anisimov [8] and by myself in [14]. The description of these models, the estimation procedure and the predictive method are briefly described. A third section explained how to integrate drop-out in these models [1]. Finally a fourth section presents the first step for a clinical cost modelling [13].

I General considerations.

In order to get marketing authorization, a new product has to succeed in clinical trials. A clinical trial is based on statistical considerations in order to show the product efficiency, taking into account the variability of the environment. It is a well known fact that the power of this test is linked to the number of patients we deal with. If an inadequate number of enrolled patients is used, then the study may fail to reject the null hypothesis due to lack of power. So the number of patients to include is a fixed parameter of the trial. There has been much effort in computing the sample size for clinical trials. Its computation is now standard and mandatory in trial protocol (see Consort Group works [17]). On the other side relatively little attention is focused on improving the predictions of the recruitment process. Indeed, till now the most of techniques used by pharma companies are based on deterministic models and various *ad hoc* techniques. Rojavin [16] says "Patient recruitment and retention remains until now more of an art rather than a science".

The problem of predicting patients recruitment and evaluating the recruitment time in clinical trials has been given much attention during the past years. Using a Poisson process to describe the recruitment process is now an accepted approach [18, 19, 10, 11]. Meanwhile, a huge variability of the recruitment process makes the question quite hard to investigate, thus, stochastic modelling has to be developed. Up to now, the easier to handle and more relevant models are so-called Poisson-Gamma model introduced in [8] and Poisson-Pareto model introduced in [14]. Those models assume that patients arrive at different

* Publications related to this chapter:

[1] Vladimir Anisimov, Guillaume Mijoule, and **Nicolas Savy**. Statistical modelling of recruitment in multicentre clinical trials with patients' drop-out. *Statistics in Medicine*, In progress, 2014.

[13] Guillaume Mijoule, Nathan Minois, Vladimir Anisimov and **Nicolas Savy**. Additive Model for Cost Modelling in Clinical Trial, *Proceedings of the 7th International Workshop on Simulation*, Rimini, May 2013. Forthcoming.

[14] Guillaume Mijoule, Stéphanie Savy and **Nicolas Savy**. Models for patients' recruitment in clinical trials and sensitivity analysis. *Statistics in Medicine*, 31(16): 1655–1674, 2012.

centres according to Poisson processes where the rates are Gamma-distributed or Pareto-distributed.

Keeping in mind these techniques, the models of patients' recruitment can be widely enriched. It is used as a basis for developing techniques for the analysis of the effects of unstratified and centre-stratified randomization [5], predictive event modelling [6] and predicting randomization process [7]. Drop-out modelling question is plugged in the Poisson-Gamma model in [1]. Indeed, patient drop-out is a critical point of most medical studies. In the framework of clinical trial, the management of drop-out is really a keystone point to deal with because first, it can yield to a lack of power in statistical analysis and second, it can be informative, especially in cancer research where drop out are mainly due to harmful side-effects and/or lack of efficacy of the treatment being studied. Finally a model for clinical trials' cost in introduced in [13].

I.1 Why model patients' recruitment ?

The computation of the Necessary Sample Size (NSS for short), denoted by N_R, is mandatory in every clinical trial protocol. Thus, in the framework of a clinical trial or more generally in medical research, an important question is how long it takes to recruit a given number of patients N_R. Indeed, this is of paramount interest for planning trials because of scientific concern, economic and ethical reasons.

- **Ethical concern**, because it is not satisfactory to continue a study in vain.
- **Economical concern**, a clinical trial is an expensive study in itself and, as the duration of the trials is included in the duration of the exclusive right to exploit the drug (20 years) [20], a delay generates an enormous loss of income (Fig 5.1). Moreover, an improvement of the planning and monitoring of a trial reduces costs and save money.

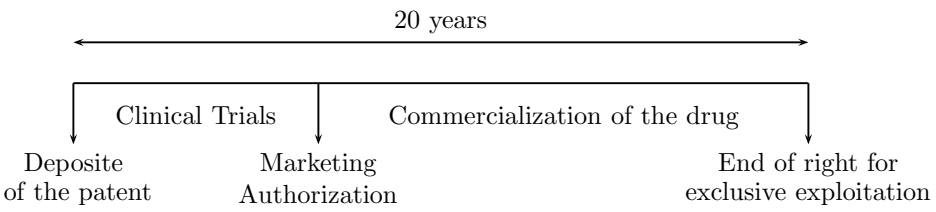

Figure 5.1: Exploitation period of a drug.

- **Scientific concern**, because new drugs are increasingly developed and approved by regulatory agencies and when accrual rates are too low, there may be new informations available during the enrolment period such as the results of other trials or a change in the understanding of the underlying biology.

For these reasons, stopping or continuing a trial is thus a decision with huge consequences and it will be useful to have some objective tools based on scientific criteria to take it.

I.2 How to model patients' recruitment ?

History of the model

Few authors have considered the problem of patients' recruitment. The reader can refer to [9] for a systematic review of the existing models for recruitment. As far as we know, the pioneer work of Morgan [15] where an estimation of the total study duration is proposed as a function of inclusion duration and based on data from previous clinical trials. Let us cite Lee [12] for a model of the recruitment by Poisson processes. Inspired by the queueing theory, this point of view has been widely developed. Poisson process appears as a natural assumption in literature [19]. In [19] a model of a multicentric trial based on Poisson process is introduced. Poisson processes depend on only one parameter, which is the rate of enrolment in our case. In [11] Carter and co-authors have developed models based on Poisson Processes and have noticed that the use of the historic mean is a too simple model and the necessity to take into consideration the variability of the rate.

Anisimov and Fedorov [8] proposed to use a doubly stochastic Poisson process to take into consideration the variation in recruitment rates between different centres. This model, called as a Poisson-gamma

model, assumes that the patients arrive at different centres according to Poisson processes with the rates viewed as independent gamma distributed random variables. In [8] the procedure of parameters estimation at interim stage using empirical Bayesian technique have been suggested. The model has been validated using data from a large number of real trials [4]. The Poisson-gamma model was developed further for predicting recruitment process at initial and interim stages [2], to account for the situations when the centres opening dates may not be known and assumed to be uniformly distributed in some intervals [3, 14], some centres can be closed or open in the future [7], for sensitivity analysis to parameter errors [14].

A Poisson-gamma model is used in [1] as a starting point for the patient arrival process and develop technique further assuming that each patient can be lost during a screening process following patient arrival. In fact we do not know if an included patient will complete the study. In order to overpass this problem, one uses to overvalue the NSS of 10 to 20% (arbitrary but classical value). This arbitrariness obviously damages the care system performance for ethical and economic reasons. The aim of a modelling of drop-out in patients' recruitment is to substitute the arbitrariness of the drop out estimation by an 'on-going' estimation. Then we would be able to quantify the drop-out and provide predictions of the optimal number of patients accounting for sample size needed and costs of the trial. Suppose that screening interval, which is the time that a patient has to stay initially in clinical centre to complete some preliminary tests for inclusion-exclusion criteria and to be randomized into the study, is assumed to be a fixed positive number R which is the same for all patients (Fig. 5.3 page 73). We assume that a patient can be lost either at the start of the screening process with some probability or during the screening interval at some random time. We consider a few models for drop-out.

The model.

Consider now a multicentric clinical trial which parameters are:

- N_R the necessary sample size (number of patients we have to recruit) is fixed and related to the statistical analysis,
- T_R the expected time for the inclusion of these N_R patients,
- C the number of centres,

We model the enrolment in centre c ($c = 1, \ldots, C$) by a Cox process starting at u_c denoted by $\{N^c(t),\, t \geq u_c\}$. Assume that $\min\{u_c\,;\, c = 1, \ldots, C\} = 0$. The distribution of the rate λ_c will be denoted by $\mathcal{L}(\theta, c)$ and its density denoted by $p_{\theta,c}$. Finally, for any $t \geq 0$,

$$\mathcal{N}^C(t) = \sum_{c=1}^{C} N^c(t)$$

Remark 5.1 *By considering time-dependent λ:*

$$\lambda(t) = \lambda\, \mathbb{I}_{\{t \geq u_c\}}$$

One can deal with models which start at $t = 0$. For notational simplicity, one assumes, excepted when specified, that $u_c \equiv 0$.

The actual end time of the study is denoted by T and is the random variable (stopping time) defined as:

$$T = \left\{ \inf_{t \geq 0} \,:\, \mathcal{N}^C(t) = N_R \right\}.$$

I.3 What model patients' recruitment for ?

θ known : Feasibility of the trial.

The model is completely defined and we are able to calculate the probability of finishing on time and the expectation of the duration of the trial.

Theorem 5.1 ([14]) *For any θ, $C \geq 1$, $t \geq 0$ and $N \geq 0$, we have*

$$\begin{aligned}
\mathbb{P}\left[\mathcal{N}^C(t) \geq N\right] &= 1 - \mathbb{E}_{\lambda_1,\cdots,\lambda_C}\left[\sum_{k=0}^{N-1} e^{-(\lambda_1+\cdots+\lambda_C)t}\frac{[(\lambda_1+\cdots+\lambda_C)t]^k}{k!}\right] \\
&= 1 - \sum_{k=0}^{N-1}\frac{t^k}{k!}\int_{\mathbb{R}^C}(x_1+\ldots+x_C)^k e^{-t(x_1+\ldots+x_C)}\prod_{c=1}^{C} p_{\theta,c}(x_c)\ dx_c, \qquad (5.1) \\
\mathbb{E}\left[T_n\right] &= \mathbb{E}_{\lambda_1,\cdots,\lambda_C}\left[\frac{N}{\lambda_1+\cdots+\lambda_C}\right] \\
&= N\int_{\mathbb{R}^C}\frac{\prod_{c=1}^{C} p_{\theta,c}(x_c)}{x_1+\ldots+x_C}dx_1\ldots dx_C.
\end{aligned}$$

Remark 5.2 *If all the u_c are not null, such quantities can still be evaluated by means of a Monte Carlo method.*

Now, as already pointed out in [2, 8], we are able to consider some tools to monitor a clinical trial:

- We can investigate the **feasibility of the trial** which is given by

$$\mathbb{P}\left[\mathcal{N}^C(T_R) \geq N_R\right].$$

- Given a fixed probability (say 80% for instance) we can calculate an **estimation of the duration** of the trial up to this probability, that is the time:

$$T \quad s.t. \quad \mathbb{P}\left[\mathcal{N}^C(T) \geq N_R\right] = 0.80.$$

- Given a fixed probability (say 80% for instance) we can calculate an **estimation of the number of centres** necessary for ending the trial on time up to this probability by:

$$C \quad s.t. \quad \mathbb{P}\left[\mathcal{N}^C(T_R) \geq N_R\right] \geq 0.80.$$

θ known : On going study.

Consider now an ongoing study at time t_1. During the period $[0, t_1]$ N_1 patients are assumed to be included. We will denote by $\mathcal{F}_{t_1}$ the history (filtration) of the enrolment process until time t_1. The key point that makes this approach of paramount interest is that we can reach the probability of including the $N - N_1$ remaining patients before the deadline and to estimate the duration of the trial. In fact one can apply Theorem 5.1, where the expectations are taken with respect to the forward distributions of the λ_c's, that is the predictive distributions based on using interim data.

Remark 5.3 *Data used for this machinery are not linked to the response of the patient to the treatment but to the inclusion in the trial data. Consequently, this technique does break the blindness.*

Remark 5.4 *We have two choices to evaluate the integral* (5.1)*: calculate it explicitly when a closed form of the integral is available or use Monte Carlo simulations.*

Now, the same kind of tools as those introduced in the previous section can be used by replacing

$$\mathbb{P}\left[\mathcal{N}^C(T) \geq N_R\right] \qquad \text{by} \qquad \mathbb{P}\left[\mathcal{N}^C(T) \geq N_R \mid \mathcal{F}_{t_1}\right].$$

We are also able to introduce corrective actions on the trial:

- We are able to **estimate of the value of the recruitment rate** to reach the deadline. When the rate is constant or when the expected rate is easily linked to the parameters θ, we can calculate an estimation of the rate necessary to reach the deadline, that is the value $\tilde{\theta}$ such that:

$$\mathbb{P}\left[\mathcal{N}^C(T_R) \geq N_R \mid \mathcal{F}_{t_1}\right] = 0,80.$$

  This action is quite artificial because in practice it is quite hard to change the rate of recruitment. Meanwhile it is a useful tool for taking a decision on the continuation of a clinical trial.

- We are able to **estimate of the number of centres to open** in order to reach the deadline. The overall rate of inclusion is a sum of the two random variables

$$\Lambda = \Lambda_A + \Lambda_B$$

where Λ_A is the contribution of the already opened centres (their distributions are thus the forward ones), and Λ_B is the contribution of the new centres, the distribution does not depend of the history $\mathcal{F}_{t_1}$. Replace C by $C+1$ increases the probability to end the trial on time, and this probability tends to 1 as C grows to infinity. In order to calculate the smallest number C of centres to open, a simple procedure consists in incrementing C until reaching the desired probability (here 0.80).

θ unknown : On going study.

In most cases, θ is unknown or given by the investigator (and often overestimated). The classic idea is to replace the real parameter θ by an estimation $\hat{\theta}$ in each relationship. For this, we use the data collected on $[0, t_1]$ to estimate θ. The error made on predictions (that is, on $\mathbb{P}_\theta\left[\mathcal{N}^C(T_R) \geq N\right]$) when replacing the true parameters θ by the estimated parameters $\hat{\theta}$ is discussed in [14]).

II The Bayesian-Poisson models and their performance

In this section, we develop the three main models of interest. Given a interim recruitment analysis at t_1, we denote by:

- n_c the number of patients recruited by centre c,
- $\tau_c = t_1 - u_c$ the duration of activity of this centre.

Given these informations, we derive the forward distribution for each model and, if the parameters are unknown, we give their Maximum Likelihood Estimator. Finally we compare their performances on a real data set. In [14] reader can find more deeper properties on the estimators, methods for comparing the models and sensitivity analyses.

II.1 The Γ-Poisson model

For this model, the distribution of the rate is Gamma whose distribution is given by:

$$f_{(\alpha,\beta)}(\lambda) = \lambda^{\alpha-1}\frac{\beta^\alpha\, e^{-\beta\,\lambda}}{\Gamma(\alpha)}\,\mathbb{I}_{\{\lambda\geq 0\}} \quad \text{with} \quad \Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1}\, e^{-t}\,\mathrm{d}t.$$

Proposition 5.1 ([8]) *Assume θ is known, given $\{(n_c,\tau_c), 1\leq c\leq C\}$, the forward distribution has for density $p^{t_1}_{\theta,c}(x)$ and is the density of a $\Gamma(\alpha+n_c, \beta+\tau_c)$ distribution.*

Proposition 5.2 ([8]) *Assume θ is unknown, given $\{(n_c,\tau_c), 1\leq c\leq C\}$, the maximum likelihood estimation of (α,β) is obtained by the maximisation of:*

$$\ln L = C\alpha\ln\beta - C\ln\Gamma(\alpha) + \sum_{c=1}^{C}\left[\ln\Gamma(\alpha+n_c) - (\alpha+n_c)\ln(\beta+\tau_c) + \tau_c\right].$$

II.2 The Pareto-Poisson model

For this model, the distribution of the rate is Pareto whose distribution is given by:

$$f_{(\gamma,\delta)}(\lambda) = \gamma\frac{\delta^\gamma}{\lambda^{\gamma+1}}\,\mathbf{1}_{\{\lambda\geq\delta\}}.$$

Proposition 5.3 ([14]) *Assume θ is known, given data $\{(n_c,\tau_c), 1\leq c\leq C\}$, the forward distribution has for density $p^{t_1}_{\theta,c}$ and is*

$$p^{t_1}_{\theta,c}(x) = \frac{\tau_c^{n_c-\gamma}}{\Gamma_{inc}(n_c-\gamma,\delta\tau_c)}e^{-x\tau_c}x^{n_c-\gamma-1}\;\mathbb{I}_{\{x\geq\delta\}} \quad \textit{where} \quad \Gamma_{inc}(y,x) = \int_x^\infty e^{-t}t^{y-1}dt,$$

is usually called (upper) incomplete-Gamma function and exists in many mathematical software packages.

Proposition 5.4 ([14]) *Assume θ is unknown, given $\{(k_c, \tau_c), 1 \leq c \leq C\}$, the MLE of (γ, δ) is obtained by the maximisation of:*

$$\ln L = C\gamma \ln \delta + C \ln \gamma + \sum_{c=1}^{C} \left[\ln \Gamma_{inc}(n_c - \gamma, \delta\tau_c) - (n_c - \gamma) \ln \tau_c + \tau_c \right].$$

II.3 The $\mathcal{U}\Gamma$-Poisson model

In some particular studies we do not observe the opening time of the centre (u_c is unknown). In case, we can assume that the opening time is a random variable uniformly distributed in $[s'_c, s_c]$.
Instead of observing u_c, we often observe the time of first inclusion of each centre, ρ_c. In case, from we use data from an ongoing study at time t_1, and set $s'_c = t_1 - \rho_c$ and $s_c = t_1$. If at t_1 the first inclusion has not occurred, we put $s'_c = 0$. Any any case, s'_c and s_c do depend on t_1.

Theorem 5.2 ([14]) *Assume θ is known, given data $\{(n_c, s'_c, s_c), 1 \leq c \leq C\}$, if we put*

$$m = \sum_{c=1}^{C} \frac{\alpha + n_c}{s_c - s'_c} \ln \left(\frac{\beta + s_c}{\beta + s'_c} \right) \qquad \text{and} \qquad v = \sum_{c=1}^{C} (\alpha + n_c) \frac{1}{(\beta + s'_c)(\beta + s_c)},$$

we can approximate the overall forward rate by a Gamma distribution by matching the first two moments $\Lambda \overset{d}{\approx} \Gamma(A, B)$ with

$$A = \frac{m^2}{v} \qquad \text{and} \qquad B = \frac{m}{v},$$

and we have

$$\mathbb{E}\left[T_N\right] \approx N \frac{m}{m^2 - v}.$$

Proposition 5.5 ([14]) *Assume θ is unknown, given $\{(n_c, s'_c, s_c), 1 \leq c \leq C\}$, the MLE of (α, β) is obtained by the maximisation of:*

$$\ln L = C\alpha \ln \beta + C \ln \Gamma(\alpha) + \sum_{c=1}^{C} \left[\ln \Gamma(\alpha + n_c) - \ln(s_c - s'_c) + \ln(J(\alpha, \beta, n_c, s'_c, s_c) \right],$$

by putting $J(\alpha, \beta, n, s', s) = \int_{s'+\beta}^{s+\beta} t^{-n-\alpha}(t - \beta)^n dt$.

II.4 Comparison on real data

In the setting of the case studied, we plan to include 610 patients in 3 years. 77 centres are devoted to this trial. In the protocol of this trial, we would plan an ongoing study at the end of the first year, at 1.5 years and at 2 years. The opening dates of the centres are not known, so we use can make use of $\mathcal{U}\Gamma$-Poisson model. Results are collected in Table 5.1.

The model	Time $t_1 = 1$	Time $t_1 = 1.5$	Time $t_1 = 2$
Γ-Poisson	3.31 $[3.29, 3.33]$	2.63 $[2.61, 2.65]$	2.44 $[2.43, 2.45]$
Π-Poisson	2.63 $[2.61, 2.65]$	2.39 $[2.37, 2.41]$	2.36 $[2.35, 2.37]$
$\mathcal{U}\Gamma$-Poisson	2.60 $[2.58, 2.62]$	2.34 $[2.33, 2.35]$	2.36 $[2.35, 2.36]$

Table 5.1: Estimation of the trial duration: the expectation on the first line, the 95% confidence interval on the second one.

Figure 5.2 represents the recruitments process. The dots represents the cumulative number of included patients. The solid line represents the expectation of the recruitment and the doted lines are the 80% confidence interval for that expectation. If you are over the solid line, all is right, if you are under the line, you have to verify the behaviour of the trial by another ongoing study. This is an useful figure for

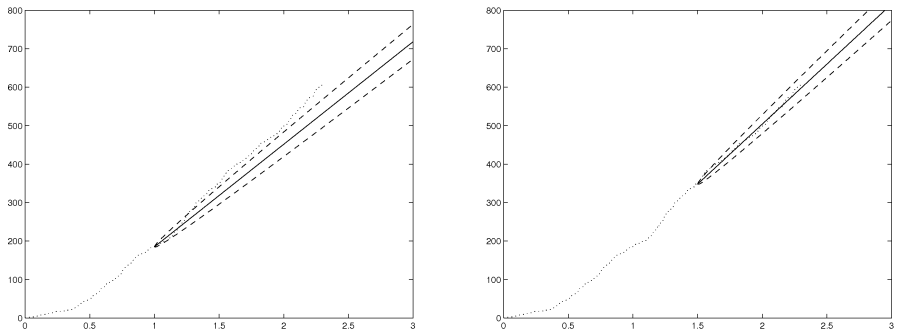

Figure 5.2: Predicted Enrolment behaviour at time 1 (left) and at time 1.5 (right)).

trialists to monitor the clinical trial.

Finally, the study ended in 2.31 year. Compare with the results in Table 5.1, the model has predicted the end of the trial with an error of 15 days, 10 months before the end.

III Integration of screening-failures

III.1 Models for recruitment with patients' drop-out.

Two kinds of drop-out can be considered in the model:

- drop-out at the inclusion (drop-out (1) on Figure 5.3).
- drop-out at any time during the screening period (drop-out (2) on Figure 5.3). Since R is small compared with the duration of the follow up, this part of the model can be neglected.

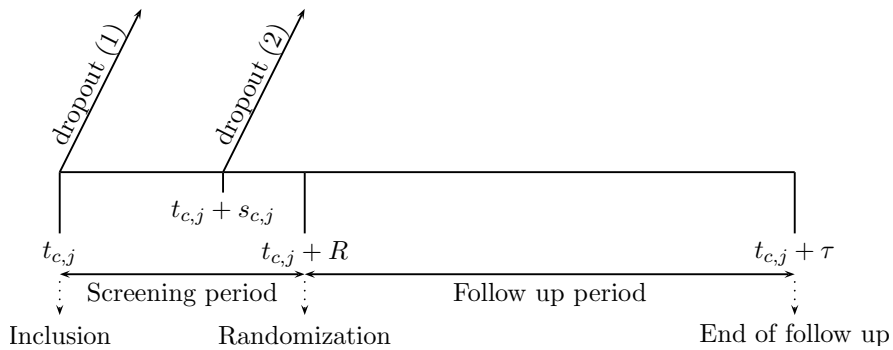

Figure 5.3: Step of patients inclusion in Clinical trials.

Consider a patient j arriving at centre c at time $t_{c,j}$. In order to consider drop-out at time $t_{c,j}$, it is enough to consider a certain probability of drop-out $1 - r_c$. We introduce two models:

Model 1. The variation in randomization probability between different centres is not taken into account. For all $1 \leq c \leq C$, $r_c = r$ where r is a fixed constant in $[0, 1]$.

Model 2. The variation in randomization probability between different centres is described using a beta distribution. The variables $\{r_c \ ; \ 1 \leq c \leq C\}$ are i.i.d.r.v. having a beta distribution with parameters (ψ_1, ψ_2).

Otherwise, the patient drops the study at some time $t_{c,j} + s_{c,j}$ during the screening interval if $s_{c,j} \leq R$. If neither one of these events happen, the patient is successfully randomized at time $t_{c,j} + R$ and registered to participate in the trial. Three models can be considered:

Model 3. For all $1 \leq c \leq C$, $r_c = r$ and the values $\{Z_{c,j}(\theta), j \geq 1 \ ; \ 1 \leq c \leq C\}$ are i.i.d.r.v. having an exponential distribution with parameter θ (the same for all centres).

Model 4. For all $1 \leq c \leq C$, $r_c = r$ and the values $\{Z_{c,j}(\theta_c), j \geq 1 \ ; \ 1 \leq c \leq C\}$ given θ_c are i.i.d.r.v. having an exponential distribution with parameter θ_c, where the values $\{\theta_c \ ; \ 1 \leq c \leq C\}$ are i.i.d.r.v. having a gamma distribution with parameters (α_2, β_2).

Model 5. The variables $\{r_c \ ; \ 1 \leq c \leq C\}$ are i.i.d.r.v. having a beta distribution with some parameters (ψ_1, ψ_2). The values $\{Z_{c,j}(\theta_c), j \geq 1 \ ; \ 1 \leq c \leq C\}$ given θ_c are i.i.d.r.v. having an exponential distribution with parameter θ_c, where the values $\{\theta_c \ ; \ 1 \leq c \leq C\}$ are i.i.d.r.v. having a gamma distribution with parameters (α_2, β_2).

In order to formalize the problem, let us define the independent families of indicators $\{\chi_{c,j}(r_c), j \geq 1 \ ; \ 1 \leq c \leq C\}$, where for a given r_c the variables $\{\chi_{c,j}(r_c), j \geq 1\}$ are conditionally independent and for any $1 \leq c \leq C$ and any $j \geq 1$,

$$\mathbb{P}(\chi_{c,j}(r_c) = 0) = 1 - \mathbb{P}(\chi_{c,j}(r_c) = 1) = r_c.$$

Now, for each centre c, at any time $t \geq 0$, recall that $N^c(t)$ denotes the number of patients included at time t, we define three processes:

- **randomized patients:**

$$N^{c,R}(t) = \text{card}\left\{j \ : \ u_c \leq t_{c,j} \leq t - R \text{ and } \chi_{c,j}(r_c) = 0, \ Z_{c,j}(\theta_c) \geq R\right\},$$

- **lost patients:**

$$N^{c,L}(t) = \text{card}\left\{j \ : \ u_c \leq t_{c,j} \leq t \text{ and } \{\chi_{c,j}(r_c) = 1\} \cup \{Z_{c,j}(\theta_c) \leq \min(R, t - t_{c,j})\}\right\},$$

- **patients in screening process:**

$$N^{c,S}(t) = N_t^c - N_t^{c,R} - N_t^{c,L}.$$

Finally, denote $\mathcal{N}^X = \sum_{c=1}^{C} N^{c,X}$ for $X := R, L, S$. The trial stops as soon as the desired number of randomized patients N_R is reached, that is when $\mathcal{N}^R(t) = N_R$.

As we see, model 5 is the most advanced model that accounts for the variation in the probability of drop-out upon patient arrival and in the distribution of drop-out time during screening process across clinical centres. For each model we consider the procedure of estimating unknown parameters and predicting in time the future process of randomized patients and the total recruitment time.

If the drop-out time were unknown, it would be impossible to distinguish between a patient lost upon arrival or during screening process, and the distinction within the model would be irrelevant. Thus, for models 3-5, we have, for each patient, to know the arrival and drop-out (or randomization) time.

III.2 Estimation

Let t_1 be some interim time and assume for simplicity that $\tau_c \geq R$ for any $c = 1, \ldots, C$. Consider that center c has recruited n_c patients and that k_c has been randomized. The recruitment process is assumed to be Poisson-Gamma modelled with unknown parameters (α, β). Thus, for any $c = 1, \ldots, C$, the random variable $N^c(t_1)$ follows a negative binomial distribution:

$$N^c(t_1) \sim \text{NegBin}\left(\alpha, \frac{\mu\tau_c}{\alpha + \mu\tau_c}\right).$$

where $\mu = \mathbb{E}[\lambda] = \alpha/\beta$. Assume for simplicity that there is no screening delay.

Model 1.

Theorem 5.3 ([1]) *Given data $\{(n_c, k_c, \tau_c), 1 \leq c \leq C\}$, the log-likelihood function writes:*

$$\mathcal{L}_1(\alpha, \mu, r) = \mathcal{L}_{1,1}(\alpha, \mu) + \mathcal{L}_{1,2}(r),$$

with

$$\mathcal{L}_{1,1}(\alpha,\mu) = \sum_{c=1}^{C} \ln\Gamma(n_c+\alpha) - C\ln\Gamma(\alpha) + N_1(\ln\mu - \ln\alpha) - \sum_{c=1}^{C}(n_c+\alpha)\ln(1+\mu\tau_c/\alpha) + B, \tag{5.2}$$

$$\mathcal{L}_{1,2}(r) = \sum_{c=1}^{C}\Big[k_c\ln r + (n_c-k_c)\ln(1-r)\Big] + B. \tag{5.3}$$

where B is some generic constant independent of the parameters and $N_1 = \sum_{c=1}^{C} n_i$ is the total number of recruited patients up to time t_1. The maximum likelihood estimator is given by

$$\widehat{r} = \Big(\sum_{c=1}^{C} n_c\Big)^{-1}\sum_{c=1}^{C} k_c. \tag{5.4}$$

Sketch of the proof. In centre c the number of randomized patients $N^{c,R}(t_1)$ has a binomial distribution with parameters (n_c, r_c):

$$N^{c,R}(t_1)\,|\,\{N^c(t_1) = n_c\} \sim \mathrm{Bin}(n_c, r_c).$$

Then the log-likelihood function can be written in the form:

$$\mathcal{L}_1(\alpha,\mu,r) = \sum_{c=1}^{C}\ln\left[\mathrm{NegBin}\left(n_c;\alpha,\frac{\mu\tau_c}{\alpha+\mu\tau_c}\right)\right] + \sum_{c=1}^{C}\ln[\mathrm{Bin}(k_c;n_c,r)].$$

The parameter r is separated from (α,μ) and $\mathcal{L}_1(\alpha,\mu,r)$ can be re-written in the form: $\mathcal{L}_1(\alpha,\mu,r) = \mathcal{L}_{1,1}(\alpha,\mu) + \mathcal{L}_{1,2}(r)$, with$\mathcal{L}_{1,1}(\alpha,\mu)$ given by (5.2) and $\mathcal{L}_{1,2}(r)$ by (5.3). Taking derivative in r yields to (5.4). □

Remark 5.5 *Whatever the model, parameters (α,μ) can be estimated using log-likelihood function $\mathcal{L}_{1,1}$ given by* (5.2) *and a two-dimensional optimization procedure.*

Remark 5.6 *Note that if there is a screening delay, then such patients that entered screening process but the results of their screening procedure are unknown yet should be excluded in the calculations of probability of randomization, otherwise this probability will be underestimated. Therefore, instead of n_i we should count $\widetilde{n_i}$, the number of patients with known screening results.*

Model 2.

Theorem 5.4 ([1]) *Given data $\{(n_c,k_c,\tau_c),\ 1\le c\le C\}$, the log-likelihood function writes:*

$$\mathcal{L}_2(\alpha,\mu,\psi_1,\psi_2) = \mathcal{L}_{2,1}(\alpha,\mu) + \mathcal{L}_{2,2}(\psi_1,\psi_2),$$

where $\mathcal{L}_{2,1} = \mathcal{L}_{1,1}$ given by (5.2)*, and*

$$\mathcal{L}_{2,2}(\psi_1,\psi_2) = \sum_{c=1}^{C}\ln\mathcal{B}(k_c+\psi_1, n_c-k_c+\psi_2) - M\ln\mathcal{B}(\psi_1,\psi_2) + B, \tag{5.5}$$

where $\mathcal{B}(\psi_1,\psi_2) = \int_0^1 x^{\psi_1-1}(1-x)^{\psi_2-1}dx$ is a beta function. Parameters (ψ_1,ψ_2) can be estimated using log-likelihood function $\mathcal{L}_{2,2}$and a two-dimensional optimization procedure.

Sketch of the proof. In centre c the number of randomized patients $N^{c,R}(t_1)$ has a Beta-binomial distribution (n_c,ψ_1,ψ_2):

$$N^{c,R}(t_1)\,|\,\{N^c(t_1) = n_c\} \sim \mathrm{Bin}(n_c, \mathrm{Beta}(\psi_1,\psi_2)).$$

□

Models 3, 4, 5.

For calculation of the likelihood function for models 3, 4, 5, we need to account more information. Assume that at time t_1, we observe patients arrival times $\{t_{c,j} \leq t_1 \ ; \ j \geq 1 \ , \ 1 \leq c \leq C\}$ and the last time $\{s_{c,j}, j \geq 1 \ ; \ 1 \leq c \leq C\}$ they are in the screening process (i.e $t_{c,j} \leq s_{c,j} \leq t_1$) . Let

$$\begin{aligned}
\mathcal{D}_1^c &= \{j \geq 1, \text{ s.t. } \quad s_{c,j} = t_{c,j}\}, \\
\mathcal{D}_2^c &= \{j \geq 1, \text{ s.t. } \quad s_{c,j} = (t_{c,j} + R) \wedge t_1\}, \\
\mathcal{D}_3^c &= \{j \geq 1, \text{ s.t. } \quad t_{c,j} < s_{c,j} < (t_{c,j} + R) \wedge t_1\}.
\end{aligned}$$

and $\mathcal{D}_1 = \bigcup_{c=1}^C \mathcal{D}_1^c$, $\mathcal{D}_2 = \bigcup_{c=1}^C \mathcal{D}_2^i$, $\mathcal{D}_3 = \bigcup_{c=1}^C \mathcal{D}_3^i$ and for any $1 \leq c \leq C$. Denote by:

- $m_c = \text{card}\,\{\mathcal{D}_3^c\}$ the number of patients lost in the middle of screening process in centre c,
- $\tilde{k}_c = \text{card}\,\{\mathcal{D}_2^c\} + \text{card}\,\{\mathcal{D}_3^c\}$ the number of patients that are not lost immediately upon arrival in centre c,
- $l_c = \sum_j (s_{c,j} - t_{c,j})$ the sum of screening durations in centre c.

Model 3.

Theorem 5.5 ([1]) *Given data $\{(t_{c,j}, s_{c,j}, \tau_c), j \geq 1 \ ; \ 1 \leq c \leq C\}$ the log-likelihood function writes:*

$$\mathcal{L}_3(\alpha, \mu, r, \theta) = \mathcal{L}_{3,1}(\alpha, \mu) + \sum_{c=1}^C \Big[(n_c - \tilde{k}_c)\ln(1-r) + \tilde{k}_c \ln r + m_c \ln\theta - \theta l_c\Big],$$

where $\mathcal{L}_{3,1} = \mathcal{L}_{1,1}$. The maximum likelihood estimators of r and θ are thus given by

$$\widehat{r} = \Big(\sum_{c=1}^C n_c\Big)^{-1} \sum_{c=1}^C \tilde{k}_c \quad \textit{and} \quad \widehat{\theta} = \Big(\sum_{c=1}^C l_c\Big)^{-1} \sum_{c=1}^C m_c, \tag{5.6}$$

Sketch of the proof. Conditioning on parameters $\{(\theta_c, r_c)\,;\, 1 \leq c \leq C\}$, we can write a general expression for the likelihood

$$\mathbf{L}\left[(t_{c,j}); (s_{c,j})\right] = \exp\left[\mathcal{L}_{1,1}(\alpha, \mu)\right] \times \prod_{c=1}^C \mathbb{E}\left[\mathbf{L}_2\left(\theta_c, r_c; (t_{c,j}), (s_{c,j})\right)\right], \tag{5.7}$$

where $\mathcal{L}_{1,1}$ is given in (5.2), and the expectation in (5.7) is taken when θ_c and r_c vary according to their respective distributions defined by models 3-5 gives:

$$\begin{aligned}
\mathbf{L}_2\left(\theta_c, r_c; (t_{c,j}), (s_{c,j})\right) &= \prod_{\mathcal{D}_1}(1 - r_c) \prod_{\mathcal{D}_2} r_c \exp(-\theta_c(s_{c,j} - t_{c,j})) \times \prod_{\mathcal{D}_3} r_c\theta_c \exp\left(-\theta_c(s_{c,j} - t_{c,j})\right), \\
&= (1 - r_c)^{n_c - \tilde{k}_c} r_c^{\tilde{k}_c} \theta_c^{m_c} \exp\left(-\theta_c l_c\right).
\end{aligned}$$

Taking derivatives in r and θ we get the MLE (5.6). □

Model 4.

Theorem 5.6 ([1]) *Given data $\{(t_{c,j}, s_{c,j}, \tau_c), j \geq 1 \ ; \ 1 \leq c \leq C\}$ the log-likelihood function writes:*

$$\mathcal{L}_4(\alpha, \mu, r, \alpha_2, \beta_2) = \mathcal{L}_{4,1}(\alpha, \mu) + \sum_{c=1}^C \Big[(n_c - \tilde{k}_c)\ln(1-r) + \tilde{k}_c \ln r\Big] + \mathcal{L}_{4,2}(\alpha_2, \beta_2), \tag{5.8}$$

where $\mathcal{L}_{4,1} = \mathcal{L}_{1,1}$ and

$$\mathcal{L}_{4,2}(\alpha_2, \beta_2) = \sum_{c=1}^C \Big[\ln\Gamma(m_c + \alpha_2) - \ln\Gamma(\alpha_2) + \alpha_2 \ln\beta_2 - (m_c + \alpha_2)\ln(\beta_2 + l_c)\Big]. \tag{5.9}$$

Thus

$$\widehat{r} = \Big(\sum_{c=1}^C n_c\Big)^{-1} \sum_{c=1}^C \tilde{k}_c. \tag{5.10}$$

and (α_2, β_2) can be estimated using log-likelihood function $\mathcal{L}_{4,2}$ and a two-dimensional optimization procedure.

Sketch of the proof. The result comes from the posterior distributions of the rates of inclusion, probabilities of instantaneous drop-out and rate of drop-out:

$$\begin{aligned}\widehat{\lambda}_c &= \mathrm{Ga}(\widehat{\alpha} + n_c, \widehat{\beta} + \tau_c),\\ \widehat{r} &= \Big(\sum_{c=1}^{C} n_c\Big)^{-1} \sum_{c=1}^{C} \tilde{k}_c,\\ \widehat{\theta}_c &= \mathrm{Ga}(\widehat{\alpha}_2 + m_c, \widehat{\beta}_2 + l_c), \quad c = 1, \dots, C.\end{aligned} \tag{5.11}$$

□

Model 5.

Theorem 5.7 ([1]) *Given data $\{(t_{c,j}, s_{c,j}, \tau_c), j \geq 1\ ;\ 1 \leq c \leq C\}$ the log-likelihood function writes:*

$$\mathcal{L}_5(\alpha, \mu, \psi_1, \psi_2, \alpha_2, \beta_2) = \mathcal{L}_{5,1}(\alpha, \mu) + \mathcal{L}_{5,2}(\alpha_2, \beta_2) + \mathcal{L}_{5,3}(\psi_1, \psi_2),$$

where $\mathcal{L}_{5,1} = \mathcal{L}_{1,1}$ and $\mathcal{L}_{5,2} = \mathcal{L}_{4,2}$ and

$$\mathcal{L}_{5,3}(\psi_1, \psi_2) = \sum_{c=1}^{C} \Big[\ln \mathcal{B}(\tilde{k}_c + \psi_1, n_c - \tilde{k}_c + \psi_2) - \ln \mathcal{B}(\psi_1, \psi_2)\Big], \tag{5.12}$$

Parameters (α, μ), (ψ_1, ψ_2) and (α_2, β_2) can be estimated using two-dimensional optimization procedures for corresponding functions $\mathcal{L}(\cdot)$.

Sketch of the proof. The result comes from the posterior distributions of the rates of inclusion, probabilities of instantaneous drop-out and rate of drop-out:

$$\widehat{\lambda_c} = \mathrm{Ga}(\widehat{\alpha} + n_c, \widehat{\beta} + \tau_c), \tag{5.13}$$

$$\widehat{r_c} = \mathrm{Beta}(\widehat{\psi_1} + \tilde{k}_c, \widehat{\psi_2} + n_c - \tilde{k}_c), \tag{5.14}$$

$$\widehat{\theta_c} = \mathrm{Ga}(\widehat{\alpha}_2 + m_c, \widehat{\beta}_2 + l_c), \quad c = 1, \dots, C. \tag{5.15}$$

□

III.3 Prediction

predictive process.

In the sequel, Π_a stand for a Poisson process with rate a. Given data $\{(n_c, \tau_c)\ ;\ 1 \leq c \leq C\}$ at interim time t_1, let ν_c be the number of patients entered screening stage at centre c in the interval $[t_1 - R, t_1]$ and k_c be the total number of randomized patients up to time t_1.

For models 1 and 2, given ν_c, the number of patients in centre c that will be randomized in the interval $[t_1, t_1 + R]$ is a binomial random variable $\mathrm{Bin}(\nu_c, r)$ and the times when these patients are randomized are uniformly distributed in $[t_1, t_1 + R]$. The predicted number of randomized patients in $[t_1, t_1 + R]$ is $\mathrm{Bin}(\nu_c, \widehat{r})$, where $\widehat{r}$ is defined in (5.4). The number of patients randomized after time $t_1 + R$ can be considered as thinning of the process $\mathcal{N}^c$ with probability r.

Theorem 5.8 ([1]) *For any $t > t_1 + R$, the predicted process of the number of randomized patients in centre c, $\{\widehat{k}_c(t),\ t \geq t_1 + R\}$, is developing as*

- ***Model 1.***
$$\widehat{k}_c(t) = k_c + \mathrm{Bin}(\nu_c, \widehat{r}) + \Pi_{\widehat{r}\,\widehat{\lambda}_c}(t - t_1 - R).$$
where $\widehat{r}$ is given by (5.4).

- ***Model 2.***
$$\widehat{k}_c i(t) = k_c + \mathrm{Bin}(\nu_c, \widehat{r}_c) + \Pi_{\widehat{r}_c\,\widehat{\lambda}_c}(t - t_1 - R).$$
where $\widehat{r}_c$ is given by (5.5).

For models 3 to 5, given ν_c the number of patients with unknown screening outcome in centre c, $\nu_c = \text{card}\{\Omega_c\}$ with

$$\Omega_c = \{j \geq 1 : t_1 - R < t_{c,j} \leq t_1 \text{ and } s_{c,j} > t_1 - t_{c,j}\}.$$

Given data at t_1 and θ_c, the number of randomized patients between t_1 and $t_1 + R$ in centre c is the sum of ν_c independent Bernoulli r.v. with probabilities $e^{-\theta_c(R-t_1+t_{c,j})}$ denoted as $\text{Ber}\left(e^{-\theta_c(t_{c,j}+R-t_1)}\right)$.

Theorem 5.9 ([1]) *For any $t > t_1 + R$, the predicted process of the number of randomized patients in centre c, $\{\widehat{k}_c(t),\ t \geq t_1 + R\}$, is developing as*

- ***Model 3.***

$$\widehat{k}_c(t) = k_c + \Pi_{\widehat{p}\,\widehat{\lambda}_c}(t - t_1 - R) + \sum_{j \in \Omega_c} \text{Ber}\left(e^{-\widehat{\theta}(t_{c,j}+R-t_1)}\right),$$

  where the probability of non-drop-out is $\widehat{p} = \widehat{r}e^{-\widehat{\theta}R}$ and $(\widehat{r}, \widehat{\theta})$ are given in (5.6).

- ***Model 4-5.***

$$\widehat{k}_c(t) = k_c + \Pi_{\widehat{p}_c\,\widehat{\lambda}_c}(t - t_1 - R) + \sum_{j \in \Omega_c} \text{Ber}\left(e^{-\widehat{\theta}_c(t_{c,j}+R-t_1)}\right),$$

  where $\widehat{p}_c = \widehat{r}_c \exp(-\widehat{\theta}_c R)$, and $(\widehat{r}_c, \widehat{\theta}_c)$ are given by (5.11).

Predictive bounds.

The main interest of this technique is to construct predictive bounds. Indeed, for C large enough ($C > 20$) we can use these expressions of the predictive process to create $(1-\delta)$-predictive bounds for $\widehat{k}(t)$ using a normal approximation similar to [7]. For this, we make use of the expressions of

$$\mathbb{E}\left[\widehat{k}(t) \mid \text{data}\right] \qquad \text{and} \qquad \mathbb{V}\left[\widehat{k}(t) \mid \text{data}\right].$$

IV An additive model for clinical trials' cost modelling [13]

This section is devoted to the first step of a model for clinical trials' cost. The aim is to calculate the cost of the trial from predefined data and from the data collected at an interim time.

IV.1 Description of the model

We assume we can categorize the different costs of a clinical trial as follows :

- a fixed cost K_1 for a screened patient,
- a fixed cost K_2 for a randomized patient (on top of the screening cost),
- a time-depending cost g for a randomized patient satisfying technical assumptions,
- a fixed cost F_c for an opened centre c,
- a time-depending cost G_c for an opened centre c.

Recall that $N^{c,R}(t)$ (resp. $N^c(t)$) is the number of randomized (resp. screened) patients at time t in c-th centre. The model we investigate for the cost of centre c at time t is:

$$C^c(t) = K_1 N^{c,R}(t) + K_2 N^c(t) + \sum_{0 \leq T_n^c \leq t} g(t, T_n^c) + F_c + G_c t, \tag{5.16}$$

And the total cost at time t is:

$$\mathcal{C}(t) = \sum_{i=1}^{C} C^c(t). \tag{5.17}$$

Notice that

$$\sum_{0 \leq T_n^c \leq t} g(t, T_n^c) = \int_0^t g(t,s) dN^{c,R}(s),$$

this process is a Filtered Poisson process as studied in Chapter 1. Isn't it marvellous... The computation of the expectation of (5.17) at time t is not a big deal but of no interest. The one of interest is to compute the expectation of $\mathcal{C}(T)$ where T is

$$T = \left\{ \inf_{t \geq 0} \; : \; \mathcal{N}^{c,R}(t) = N_R \right\}.$$

and is a stopping time in the natural filtration of $\mathcal{N}^{c,R}$. As seen in Chapter 1, the standard Itô calculus is not applicable with F.P.P. and thus we cannot apply standard argument of stochastic calculus.

IV.2 Calculation of the mean cost

Non-Bayesian setting

First assume the recruitment rates $(\lambda_c)_{1 \leq c \leq C}$ and probabilities of screening success $(r_c)_{1 \leq c \leq C}$ are known. Then we have the expansion

$$\mathcal{N} = \mathcal{N}^R + \mathcal{N}^L \tag{5.18}$$

where $\mathcal{N}^R$ is the aforementioned Poisson process of randomized patient, with rate $\Lambda_1 = \sum_{c=1}^{C} r_c \lambda_c$, and $\mathcal{N}^L$ is an independent Poisson process with rate $\Lambda_2 = \sum_{c=1}^{C} (1 - r_c)\lambda_c$, representing the number of screening failures over time.

Theorem 5.10 ([13]) *Let $K_1' = K_1 + K_2$, and $p_\tau(dt) := e^{-\Lambda_1 t} t^{N_R - 1} \, \mathbb{I}_{\{t>0\}} dt$ be the pdf of T. Then*

$$\mathbb{E}\left[\mathcal{C}(T)\right] = K_1' N_R + K_2 N_R \frac{\Lambda_2}{\Lambda_1} + \int_0^{+\infty} g(t,t) p_\tau(dt) + (N_R - 1) \int_0^{+\infty} \int_0^t g(t,s)\, ds \, \frac{p_\tau(dt)}{t} + G \frac{N_R}{\Lambda_1} + F. \tag{5.19}$$

Remark 5.7 *All functions in* (5.19) *are positive and measurable, so the integrals are well defined.*

Bayesian setting

Now, we assume the initial rates are distributed according to a Gamma distribution and the probabilities of screening as a Beta distribution. At some interim time t_1, assume c-th centre has screened n_c patients and randomized k_c patients. Recall that, by Bayesian re-estimation, given n_c and k_c, the rate λ_c has a Gamma distribution with parameters $(\alpha + n_c, \beta + t_1)$, and probability of screening r_c has a Beta distribution with parameters $(\psi_1 + k_c, \psi_2 + n_c - k_c)$. A consequence of Theorem 5.10 is the following corollary.

Corollary 5.1 *In the Bayesian setting, the mean cost is*

$$\begin{aligned}\mathbb{E}\left[\mathcal{C}(T)\right] = K_1' N_R + K_2 N_R \mathbb{E}\left[\frac{\Lambda_2}{\Lambda_1}\right] + \int_0^{+\infty} g(t,t) \mathbb{E}\left[e^{-\Lambda_1 t}\right] t^{N_R - 1} dt \\ + (N_R - 1) \int_0^{+\infty} \int_0^t g(t,s)\, ds \, \mathbb{E}\left[e^{-\Lambda_1 t}\right] t^{N_R - 2} dt + G N_R \mathbb{E}\left[\Lambda_1^{-1}\right] + F.\end{aligned}$$

IV.3 Closure of a centre

Assume a linear cost over time for each randomized patient. The function g is defined as $g(t,s) = K_3(t - s)\, \mathbb{I}_{\{t \geq s\}}$, where K_3 is some positive constant. Does the closure of j-th centre implies money savings? Next corollary answers this question.

Corollary 5.2 *The closure of j-th centre implies a variation of the cost of the trial $\Delta\mathcal{C}_j$ which is*

$$\Delta\mathcal{C}_j = N_R \mathbb{E}\left[\frac{\lambda_j}{\Lambda(\Lambda_1 - r_j \lambda_j)} \left(K_2(r_j \Lambda_2 - (1 - r_j)\Lambda_1) + \frac{1}{2} K_3 (N_R - 1) r_j + r_j G - G_j \frac{\Lambda_1}{\lambda_j} \right)\right] - F_j,$$

where $G = \sum_{c=1}^{C} G_c$.

Bibliography

[1] Vladimir Anisimov, Guillaume Mijoule, and **Nicolas Savy**. Statistical modelling of recruitment in multicentre clinical trials with patients' dropout. *Statistics in Medicine*, 2014. In preparation.

[2] Vladimir V. Anisimov. Using mixed Poisson models in patient recruitment in multicentre clinical trials. In *Proceegings of the World Congress on Ingineering*, volume II, pages 1046–1049, London, United Kingdom, 2008.

[3] Vladimir V. Anisimov. Predictive modelling of recruitment and drug supply in multicenter clinical trials. In *Proceedings of the Joint Statistical Meeting, ASA*, pages 1248–1259, Washington, USA, August 2009.

[4] Vladimir V. Anisimov. Recruitment modeling and predicting in clinical trials. *Pharmaceutical Outsourcing*, 10(1):44–48, 2009.

[5] Vladimir V. Anisimov. Effects of unstratified and centre-stratified randomization in multi-centre clinical trials. *Pharmaceutical Statistics*, 10(1):50–59, 2011.

[6] Vladimir V. Anisimov. Predictive event modelling in multicentre clinical trials with waiting time to response. *Pharmaceutical Statistics*, 10(6):517–522, 2011.

[7] Vladimir V. Anisimov. Statistical modeling of clinical trials (recruitment and randomization). *Comm. Statist. Theory Methods*, 40(19-20):3684–3699, 2011.

[8] Vladimir V. Anisimov and Valerii V. Fedorov. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in Medicine*, 26(27):4958–4975, 2007.

[9] Katharine D. Barnard, Louise Dent, and Andrew Cook. A systematic review of models to predict recruitment to multicentre trials. *BMC Medical Research Methodology*, 63(10), 2010.

[10] Rickey E. Carter, Susan C. Sonne, and Kathleen T. Brady. Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC Medical Research Methodology*, 5(11):1–5, 2005.

[11] Rickey Edward Carter. Application of stochastic processes to participant recruitment in clinical trials. *Controlled Clinical Trials*, 25(5):429–436, 2004.

[12] Young Jack Lee. Interim recruitment goals in clinical trials. *Journal of Chronic Diseases*, 36(5):379–389, 1983.

[13] Guillaume Mijoule, Nathan Minois, Vladimir Anisimov, and **Nicolas Savy**. Additive model for cost modelling in clinical trial. In *Proceedings of the 7th International Workshop on Simulation*, 2013. Forthcoming.

[14] Guillaume Mijoule, Stéphanie Savy, and **Nicolas Savy**. Models for patients' recruitment in clinical trials and sensitivity analysis. *Statistics in Medicine*, 31(16):1655–1674, 2012.

[15] Timothy M. Morgan. Nonparametric estimation of duration of accrual and total study length for clinical trials. *Biometrics*, 43(4):903–912, 1987.

[16] Mikhail Rojavin. Patient recruitment and retention: From art to science. *Contemporary Clinical Trials*, 30(5):387–387, 2009.

[17] Kenneth F. Schulz, Douglas G. Altman, and David Moher. Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, In Press, Corrected Proof:–, 2010.

[18] Stephen Senn. *Statistical Issues in Drug Development*. John Wiley & Sons, Chichester, 1997.

[19] Stephen Senn. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine*, 17:1753–1765, 1998.

[20] Anne-Priscille Vlasto. Brevets et médicament en france. pourquoi l'application du droit des brevets au médicament est-elle autant critiquée ? *Médecine & Droit*, 2007(82):25–32, 2007.

Chapter 6

Survival data analysis for prevention Randomized Controlled Trials *

This chapter is a summary of the results obtained during the Ph.D. thesis of Valérie Garès who I have supervised with Prof. Sandrine Andrieu (INSERM Unit 1027 - Toulouse) for epidemiological aspects. Valérie Garès defended her Ph.D. April 15th, 2014 and the examination board was composed of Prof. Aurélien Latouche (CNAM Paris - referee) - Prof. John O'Quigley (UPMC ParisVI - referee) - Prof. Jean-Yves Dauxois (INSA de Toulouse) - Prof. Jean-François Dupuy (INSA de Rennes), Prof. Sandrine Andrieu and myself (University of Toulouse III). It deals with weighted logrank tests which are constructed by plugging a weight function $(W_n(s),\ s \in \mathbb{R}^+)$, depending of the sample size n in the logrank statistic. We focus on the application of these well known tests in the framework of a clinical trial where two problems of paramount importance appears:

- the choice of a particular weight,
- the computation of the necessary sample size.

The choice of a weight is motivated by the kind of deviation to the null hypothesis (of equality of the survival functions) that we are interested in detecting. The Fleming-Harrington weight (see [3]) is defined as

$$W_n^{p,q}(s) = [\hat{S}_n(s)]^p\ [1 - \hat{S}_n(s)]^q \tag{6.1}$$

where $p \geq 0$, $q \geq 0$ and $\hat{S}_n$ is the Kaplan-Meier estimator of the survival function S under the null hypothesis. To detect late effect, one emphasizes what happen at the end of the follow-up period taking $p = 0$ and $q \geq 0$. In what follows, we shall refer the resulting weighted test to as the "Fleming-Harrington test", and denote it by FH(q). We will focus on this test because it depends only of one parameter and is implemented in most software. However how to choose this parameter has not been investigated yet.

In [4], we have studied in details Fleming-Harrington's test. The performances in terms of empirical power has been investigated by means of simulations studies. We focus our attention on the choice of the parameter q which is not directly interpretable in terms of late effects. This question is of paramount importance for clinical trials and is not easy. Hopefully, a sensitivity analysis of the role of q has shown that Fleming-Harrington's test is only few sensitive to the value of q. The choice of q close to 3 is thus a good choice in most situations.

---

* Publications related to this chapter:

[4] Valérie Garès, Sandrine Andrieu, Jean-François Dupuy, and **Nicolas Savy**. Choosing the parameter of Fleming-Harrington's test in prevention randomized controlled trials. Submitted to *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2014.

[5] Valérie Garès, Sandrine Andrieu, Jean-François Dupuy, and **Nicolas Savy**. Comparison of constant piecewise weighted test and Fleming Harrington's test - Application in clinical trials. In revision. *Electronic Journal of Statistics*, 2013.

[6] Valérie Garès, Sandrine Andrieu, Jean-François Dupuy, and **Nicolas Savy**. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. In revision. *Statistics in Medicine*, 2013.

In the family of weighted logrank' tests, a second weight of interest is the constant piecewise weight (CPW for short) defined as

$$W^{t^*}(t) = \begin{cases} 0 & \text{if } t \leq t^*, \\ 1 & \text{if } t > t^*, \end{cases} \tag{6.2}$$

for some t^*. The resulting weighted logrank statistic (subsequently referred to as the "CPWL statistic" and denoted by CPWL(t^*)) has been studied in [12]. One appealing feature of the weight (6.2) is that the parameter t^* is directly interpretable in terms of late effects. In practice, a reasonable value of t^* should therefore be based on the investigator's *a priori* knowledge about the late effects. In [5] we have studied this test and noticed that CPWL's test suffers from being sensitive to the value of t^*. In view of this result, the Fleming-Harrington weight (6.1) appears to be more appealing than the CPW (6.2). However, in practice, it is easier to identify a reasonable range of values for t^* than to choose q. By comparing the Fleming-Harrington and CPWL tests (using arguments from the asymptotic efficiency theory and some numerical comparisons), we are able to elucidate the relationship between q and t^*. From this, we establish some rules for choosing q from a given t^*.

Finally both tests assumes that effects are late. It is a huge decision to take for whom design a clinical trial. In order to overpass this difficulty, we propose in [6] a test which avoids this assumption by taking the maximum between logrank and Fleming-Harrington statistics. We investigate the performances of this test.

After a section devoted to generalities on weighted logrank tests, we focus our attention in a second section to Fleming-Harrington's test especially its performance and a comparison with CPWL's test. A third section introduced the maximum weighted logrank's test. Finally an application to GuidAge's study is discussion.

I Weighted logrank tests

I.1 Notations and definitions

Let T be a non-negative random variable with cumulative distribution function F, survival function $S = 1 - F$, hazard function λ, and cumulative hazard function $\Lambda(t) = \int_0^t \lambda(s)ds$. T denotes the duration from some time origin to the occurrence of some event of interest. In what follows, T is assumed to be right-censored that is, we only observe the events that occur before a certain time C. Letting T^i and C^i be respectively the latent survival and censoring times for the i-th individual, the observations consist of n independent couples $(X^i, \delta^i)_{i=1\ldots n}$, where $X^i = \min(T^i, C^i)$ and $\delta^i = \mathbb{I}_{\{T^i \leq C^i\}}$. We assume that T^i and C^i are independent for every $i = 1, \ldots, n$. Let G be the distribution function of the $(C^i)_{i=1,\ldots,n}$, τ denote the total duration of the study, and $\tau' = \inf_{t\geq 0}\{\pi(t) = 0\}$, where $\pi(t) = (1 - F(t))(1 - G(t))$. We assume that $\tau < \tau'$. For every $t \geq 0$, we also define the random variables

$$N_n(t) = \sum_{i=1}^n \mathbb{I}_{\{X^i \leq t,\, \delta^i = 1\}} \qquad \text{and} \qquad Y_n(t) = \sum_{i=1}^n \mathbb{I}_{\{X^i \geq t\}}.$$

$N_n(t)$ is the number of failures at t and $Y_n(t)$ is the number of at-risk subjects at time t^-.

We consider a clinical trial with two arms, where n_T patients receive a drug (or treatment) and n_P patients receive a placebo (with $n = n_P + n_T$). In what follows, all the random variables and related quantities (cumulative distribution function, survival function...) for the treatment (respectively placebo) group are upper-indexed by T (respectively P). For example, we note $N_n = N^P_{n_P} + N^T_{n_T}$ and $Y_n = Y^P_{n_P} + Y^T_{n_T}$.

I.2 Asymptotic distributions

Consider the following null and alternative hypotheses:

$$\begin{cases} \mathcal{H}_0 & : F^T = F^P = F_{\theta^0}, \\ \mathcal{H}_1 & : F^T = F_{\theta^T} \quad \text{and} \quad F^P = F_{\theta^P}. \end{cases} \tag{6.3}$$

To solve this testing problem, one focuses our attention on Weighted Logrank tests which are defined as

$$\mathrm{LR}_{W_n}(t) = \int_0^t W_n(s) \left(\frac{n_P + n_T}{n_P n_T}\right)^{1/2} \frac{Y^P_{n_P}(s) Y^T_{n_T}(s)}{Y_n(s)} \left[\frac{dN^P_{n_P}(s)}{Y^P_{n_P}(s)} - \frac{dN^T_{n_T}(s)}{Y^T_{n_T}(s)}\right],$$

where (W_n) is a sequence of adapted, bounded, non-negative and predictable weighting processes.

Hypothesis 6.1 *As $n \to \infty$, $n_P/n \to 1/2$ and $n_T/n \to 1/2$.*

Hypothesis 6.2 *There exists a function $w \in \mathbb{D}$ (where $\mathbb{D}$ is the Skohorod space of càdlàg functions) such that $W_n(s) \xrightarrow{a.s.} w(s)$ as $n \to \infty$.*

Theorem 6.1 *Assume that Hypotheses 6.1 and 6.2 are fulfilled.*
Then, under $\mathcal{H}_0$, LR_{W_n} converges weakly to $\mathbb{G}_0$ a zero-mean Gaussian process with covariance function

$$\sigma^2_{\theta^0} : (t_1, t_2) \to \int_0^{t_1 \wedge t_2} w^2(s) \frac{\pi^P(s-)\pi^T(s-)}{\pi(s-)} (1 - \Delta\Lambda_{\theta^0}(s)) d\Lambda_{\theta^0}(s).$$

And under $\mathcal{H}_1$, $\mathrm{LR}_{W_n} - \sqrt{n}\mu^{\mathbb{G}_1}_{(\theta^T,\theta^P)}$ converges weakly to $\mathbb{G}_1$ a zero-mean Gaussian process with covariance function $\sigma^2_{(\theta^T,\theta^P)} = \sigma^2_{\theta^P} + \sigma^2_{\theta^T}$, where for $j = T, P$,

$$\sigma^2_{\theta^j} : (t_1, t_2) \to \frac{1}{2} \int_0^{t_1 \wedge t_2} \frac{k^2(s)}{\pi^j(s-)} (1 - \Delta\Lambda_{\theta^j}(s)) d\Lambda_{\theta^j}(s) \quad \textit{and} \quad \mu^{\mathbb{G}_1}_{(\theta^T,\theta^P)} : t \to \frac{1}{2} \int_0^t k(s)(d\Lambda_{\theta^P}(s) - d\Lambda_{\theta^T}(s)).$$

with

$$k(s) = w(s) \frac{\pi^P(s-)\pi^T(s-)}{\pi(s-)}.$$

Sketch of the proof. The following result is well-known (see [7]) but can be shown in an elegant way by the use of stochastic integrals convergence Theorem [9]. The key-point is the following result which plugs the problem in a martingale environment.

Theorem 6.2 ([7]) *For $i = P, T$ the process $M^i_{n_i} = N^i_{n_i} - A^i_{n_i}$ is a martingale with predictable compensator $A^i_{n_i}$ defined by:*

$$t \to A^i_{n_i}(t) = \int_0^t Y^i_{n_i}(s) d\Lambda^i(s).$$

In fact, it is easily seen that LR_{W_n} can be written

$$\mathrm{LR}_{W_n} = \mathrm{LRM}^P_{W_n} - \mathrm{LRM}^T_{W_n} + \mathrm{LRC}^P_{W_n} - \mathrm{LRC}^T_{W_n}, \tag{6.4}$$

where for $(i, j) \in \{T, P\}$, $i \neq j$, and $t \geq 0$,

$$\mathrm{LRM}^i_{W_n}(t) = \int_0^t \sqrt{\frac{n_j}{n}} W_n(s) \frac{n}{Y_n(s)} \frac{Y^j_{n_j}(s)}{n_j} \frac{dM^i_{n_i}(s)}{\sqrt{n_i}}, \tag{6.5}$$

$$\mathrm{LRC}^i_{W_n}(t) = \sqrt{n} \int_0^t W_n(s) \frac{n}{Y_n(s)} \frac{Y^i_{n_i}(s)}{n_i} \frac{Y^j_{n_j}(s)}{n_j} \sqrt{\frac{n_i n_j}{n\, n}} d\Lambda_{\theta^i}(s). \tag{6.6}$$

The terms $\mathrm{LRM}^i_{W_n}$ express as a stochastic integral with respect to the martingale $\tilde{M}$:

$$\tilde{M}^i_{n_i}(s) = \frac{M^i_{n_i}(s)}{\sqrt{n_i}} \quad \text{and} \quad \mathrm{LRM}^i_{W_n}(t) = \int_0^t H^i_{n_i}(s) d\tilde{M}^i_{n_i}(s).$$

The following convergence results hold:

Lemma 6.1
- *For $i = T, P$, martingales $\tilde{M}^i_{n_i}$ converge weakly to $\mathbb{M}$ a zero-mean Gaussian process with covariance function given by (see [1]):*

$$\mathrm{Cov}(\mathbb{M}(t_1), \mathbb{M}(t_2)) = \int_0^{t_1 \wedge t_2} (1 - \Delta\Lambda^i(s)) \pi^i(s-) d\Lambda^i(s).$$

- *For $(i, j) \in \{T, P\}$, $i \neq j$, we have:*

$$H^{i,j}_n = \sqrt{\frac{n_j}{n}} W_n \frac{n}{Y_n} \frac{Y^j_{n_j}}{n_j} \xrightarrow[n \to \infty]{a.s.} \frac{1}{\sqrt{2}} \frac{k}{\pi^i_-}.$$

In remains to apply the convergence for stochastic integral Theorem:

Theorem 6.3 ([9]) *Let M_n a sequence of real-valued martingales and H_n a sequence of real-valued càdlàg predictable processes such that:*

- *(M_n) satisfies the uniform tightness (UT) condition,*
- $(H_n, M_n) \overset{\mathcal{L}(\mathbb{D}^2)}{\to} (H_\infty, M_\infty)$

then

$$\int_0^t H_n(s)dM_n(s) \xrightarrow{\mathcal{L}(\mathbb{D})} \int_0^t H_\infty(s)dM_\infty(s), \tag{6.7}$$

Notice that the limit of $H_n^{i,j}$ is deterministic thus the limit process is the stochastic integral of a deterministic process with respect to a Gaussian martingale, it is thus a Gaussian process itself.

The terms $\mathrm{LRC}^i_{W_n}$ is 0 under assumption $\mathcal{H}_0$. Under $\mathcal{H}_1$, noticing that,

$$\frac{Y_n}{n} \xrightarrow[n\to\infty]{p.s.} \pi_-, \qquad \frac{Y^i_{n_i}}{n_i} \xrightarrow[n\to\infty]{p.s.} \pi^i_-, \quad \text{and} \quad \frac{Y^j_{n_j}}{n_j} \xrightarrow[n\to\infty]{p.s.} \pi^j_-,$$

we have:

$$\frac{1}{\sqrt{n}}\,\mathrm{LRC}^i_{W_n} \xrightarrow[n\to\infty]{a.s} \frac{1}{2}\int_0^{\cdot} k(s)d\Lambda_{\theta^i}(s).$$

which yields, after some algebra, to the value of $\mu^{\mathbb{G}_1}_{(\theta^T,\theta^P)}$. □

I.3 Application to the computation of the NSS

Before launching a clinical trial, one needs to know how much resource is needed to ensure that the study has enough power to detect the difference of interest. Assuming a type I censoring scheme, we provide a sample size formula for testing the hypotheses (6.3) using a weighted logrank test.

Theorem 6.4 *Fix the alternative hypotheses (this means the values of θ^P and θ^T) and assume that $n_T = n_P = \frac{n}{2}$. The sample size needed to achieve a power $1-\beta$ with a type I error α, when testing the hypotheses (6.3) using a weighted logrank test, is given by :*

$$n = 2.\frac{\sigma_1^2}{\mu^2}.(z_{1-\alpha/2} + z_{1-\beta})^2, \tag{6.8}$$

where z_γ denotes the quantile of order γ of a standard normal distribution and

$$\mu = \int_0^\tau w(s)\frac{\pi^P(s)\pi^T(s)}{\pi(s)}(d\Lambda_{\theta^P}(s) - d\Lambda_{\theta^T}(s)),$$
$$\sigma_1^2 = \int_0^\tau w^2(s)\left(\frac{\pi^P(s)(\pi^T(s))^2}{(\pi(s))^2}d\Lambda_{\theta^P}(s) + \frac{(\pi^P(s))^2\pi^T(s)}{(\pi(s))^2}d\Lambda_{\theta^T}(s)\right).$$

I.4 Asymptotic Relative Efficiency

Theorem 6.1 insures that under $\mathcal{H}_1$, the asymptotic distribution of $LR_{W_n}(t)$ is degenerate (see [4] and references therein). As a consequence, the weighted logrank tests are consistent and their respective powers converge to 1 as n tends to infinity ([3, 7]). In this setting, an appropriate comparison procedure is to investigate the behaviour of the tests under a sequence of alternatives converging to the null hypothesis as n tends to infinity. A relevant choice of the alternatives $(\theta^P_{n_P})$ and $(\theta^T_{n_T})$ in (6.3) is

$$\theta^P_{n_P} = \theta^0 + d\left(\frac{n_T}{n_P(n_P+n_T)}\right)^{1/2} \text{and } \theta^T_{n_T} = \theta^0 - d\left(\frac{n_P}{n_T(n_P+n_T)}\right)^{1/2} \tag{6.9}$$

where $d \in \mathbb{R}$ is a constant (see [3]) and for which, under assumptions developed below, we have a finite constant μ_{θ^0} such that:

$$\sqrt{n}\,\mu^{\mathbb{G}_0}_{(\theta^T,\theta^P)} \xrightarrow[n\to\infty]{a.s.} \mu_{\theta^0}.$$

This is the idea of Asymptotic Relative Efficiency (ARE) in the sense of Pitman (see [11] for a definition and a detailed exposition).

Hypothesis 6.3 *The function $\theta \to \lambda_\theta$ is differentiable at θ^0 and $\frac{\partial \lambda_\theta}{\partial \theta}\big|_{\theta=\theta^0} \neq 0$.*

Theorem 6.5 ([7]) *Assume that Assumption 6.3 holds and let*

$$k(s) = w(s)\frac{\pi^P(s)\pi^T(s)}{\pi(s)}. \tag{6.10}$$

Then

$$\sqrt{n}\,\mu^{\mathbb{G}_0}_{(\theta^T,\theta^P)} \xrightarrow[n\to\infty]{a.s.} \mu_{\theta^0} \qquad \text{with} \qquad \mu_{\theta^0} = \int_0^{\cdot} c\frac{k(s)}{\lambda_{\theta^0}(s)}\,\frac{\partial \lambda_\theta}{\partial \theta}(s)\bigg|_{\theta=\theta^0} d\Lambda_{\theta^0}(s).$$

Then one deduces the following result which expresses the Pitman's ARE of two weighted logrank statistics as the ratio of their respective asymptotic efficiencies (AE for short) and expresses the expression of the limit weight function which makes the AE maximal. This results is of paramount interest for investigating the performance of the test.

Theorem 6.6 ([7]) *Let $LR_{W^1_n}$ and $LR_{W^2_n}$ be two weighted logrank statistics satisfying the Assumptions 6.1, 6.2, 6.3. Consider a sequence of alternatives of the form (6.3), with $\theta^P_{n_P}$ and $\theta^T_{n_T}$ defined by (6.9). Then the Pitman ARE of $LR_{W^1_n}$ with respect to $LR_{W^2_n}$ is given by:*

$$ARE(LR_{W^1_n}, LR_{W^2_n}) = \frac{AE(LR_{W^1_n})}{AE(LR_{W^2_n})} \qquad \text{where} \qquad AE(LR_{W^j_n}) = \frac{\left(\int_0^\tau \frac{k_j(s)}{\lambda_{\theta^0}(s)}\,\frac{\partial \lambda_\theta}{\partial \theta}(s)\big|_{\theta=\theta^0} d\Lambda_{\theta^0}(s)\right)^2}{\int_0^\tau (k_j)^2(s)\frac{\pi(s)}{\pi^P(s)\pi^T(s)}d\Lambda_{\theta^0}(s)}. \tag{6.11}$$

Moreover, the weighted logrank statistic with maximal AE has a limit weight function w such that k in (6.10) is given by:

$$k : s \to \kappa\frac{1}{\lambda_{\theta^0}(s)}\,\frac{\partial \lambda_\theta}{\partial \theta}\bigg|_{\theta=\theta^0}(s)\left(\frac{\pi^P(s)\pi^T(s)}{\pi(s)}\right),$$

where κ is a constant.

II Fleming-Harrington's test

Fleming-Harrington's weight (6.1) with $p = 0$ and $q \geq 0$ emphasizes the differences at the end of the follow-up. It is thus a better tool than logrank's test to detect late effect. However, it depends on a parameter which has to be specified in the protocol. Here we evaluate the performance of this test: its power and its sensitivity to the parameter q. These investigation can be performed by simulation studies. This is possible thank to Theorem 6.7 below. In fact, it yields to a method for generating data optimal for Fleming-Harrington's test in the sense of ARE.

II.1 Optimality of Fleming Harrington's test

In logrank testing, a useful strategy is to consider the particular pattern of "shift assumptions up to a change of time" for the alternative hypothesis (see [7]). This can be defined through the following family of distribution functions:

$$F_\theta(t) = \Psi(g(t) + \theta),\ t \in \mathbb{R}^+, \qquad \theta \in \Theta, \tag{6.12}$$

where $g : [0, \infty[\to] - \infty, u^+[$ (with $u^+ \in \bar{\mathbb{R}}$) is a differentiable non-decreasing function, and Ψ is a continuous cumulative distribution function with positive density Ψ' and an almost everywhere continuous second derivative Ψ'' (see [4, 5] for more details). Under a shift alternative and a relevant choice for g, Theorem 6.6 allows to express the hazards of the treatment and placebo groups up to a shift $\Delta = \theta^P - \theta^T$:

Theorem 6.7 ([4]) *Given a shift Δ, the Fleming-Harrington test with $q > 0$ has maximum AE to test*

$$\begin{cases} \mathcal{H}_0 & : \lambda^T = \lambda^P, \\ \mathcal{H}_1 & : \lambda^T = \lambda^P\ \Gamma^q(., \Delta), \end{cases} \tag{6.13}$$

where for any $t \in \mathbb{R}^+$,

$$\Gamma^q(t, \Delta) = \frac{L^q((\mathcal{L}^q)^{-1}(\mathcal{L}^q(S^P(t)) + \Delta))}{L^q(S^P(t))},$$

and $\mathcal{L}^q :]0,1[\to \mathbb{R}^-$ *is a one-to-one map defined as the primitive of the function defined from* $]0,1[$ *to* $\mathbb{R}^-$ *by:*

$$x \to \frac{1}{xL^q(x)} \qquad \text{with} \quad L^q(x) = -B_{inc}(x-1, q+1, p),$$

and B_{inc} *is the incomplete beta function* $B_{inc}(x,a,b) = \int_0^x s^{a-1}(1-s)^{b-1}ds$.

Remark 6.1 *The same reasoning can be made in the setting of early effect detection* ($p \geq 0$ *and* $q = 0$ *in* (6.1)*). In this case, the function* $\mathcal{L}^p$ *(analogue of* $\mathcal{L}^q$*) is explicit and the computations are easy. Here the function is no more explicit and the computations are made by numerical integration techniques.*

II.2 Performances of Fleming-Harrington's test

The performances of the test have been investigated by means of empirical level and empirical power. We simulate data according to a generating process under which the Fleming-Harrington test with parameter q_S is optimal (in the sequel, q_S will stand for "the q value used for simulating the data") and one performs different tests on this generated data set.

Data generating process (DGP1). Let $q_S > 0$, $c = S^P(\tau)$, a sample size n and a discrepancy rate r defined as

$$r = \frac{S^T(\tau) - S^P(\tau)}{1 - S^P(\tau)}, \tag{6.14}$$

which in preventive clinical trials, is usually fixed by the investigator, the data generating process is:

- The data in the placebo group are simulated from an exponential distribution with parameter $a > 0$, where a is fixed from the desired proportion of censored data:

$$a = -\frac{\ln(S^P(\tau))}{\tau}. \tag{6.15}$$

- The data in the treatment group are simulated from the hazard function

$$\lambda^T(t) = a\ \frac{L^{q_S}((\mathcal{L}^{q_S})^{-1}(\mathcal{L}^{q_S}(e^{-at}) + \Delta(q_S)))}{L^{q_S}(e^{-at})} \tag{6.16}$$

  with $\Delta(q_S)$ given by

$$\Delta(q_S) = \theta^T - \theta^P = \mathcal{L}^{q_S}(r(1 - S^P(\tau)) + S^P(\tau)) - \mathcal{L}^{q_S}(S^P(\tau)).$$

  This data set is optimal for FH(q_S) in virtue of Theorem 6.7.

We consider well-balanced placebo and treatment groups that is, $n_P = n_T = \frac{n}{2}$. Such a sample generated from this data generating process is denoted by $\mathcal{S}_1(q_S, n, r, c)$.

Simulation design. We simulate $N = 2000$ samples $\mathcal{S}_1(q_S, n, r, c)$ for each $q_S \in \{0,1,2,3,4,5\}$. The logrank test and the Fleming-Harrington tests with $q = q_T$ with q_T successively equal to $1,2,3,4$ are applied to each of the N samples, and the empirical powers of all these tests are obtained (in what follows, q_T will stand for "the q value used for testing the data". In [4, 5] we investigate sensitivity analysis of the test with respect to:

- the sample size considering several values for n ($n = 100, 500, 1000, 2000$),
- the censoring considering several values for c ($c = 0.2, 0.5, 0.8$),
- the discrepancy rate considering several values for r ($r = 0.1, 0.2, 0.3$).

Results. The Fleming-Harrington's test appears to respect the nominal level. From the Table 6.1, [4, 5] and their supplementary documents, the power of Fleming-Harrington's test increases with n and r, and decreases when the censoring increases. In each scenario, we note that the Fleming-Harrington test has maximal power when q_T is taken equal to q_S. We also observe that the empirical power of the Fleming-Harrington test only slightly varies when q_T varies, which means that the sensitivity of the Fleming-Harrington test to the value of q_T is very small. Therefore, misspecifying q_T will only have a limited impact on the result of the test. This is a nice feature of the Fleming-Harrington test in view of its application in clinical trials.

q_S	Logrank	$q_T = 1$	$q_T = 2$	$q_T = 3$	$q_T = 4$
0	**0.640**	0.534	0.420	0.349	0.294
1	0.620	**0.743**	0.713	0.670	0.632
2	0.609	0.845	**0.877**	0.871	0.853
3	0.593	0.873	0.912	**0.914**	**0.914**
4	0.587	0.887	0.940	0.957	**0.961**
5	0.588	0.910	0.962	0.974	**0.980**

Table 6.1: Empirical power of FH tests for various q_T when the data are generated under the optimal hypothesis for FH(q_S) with $c = 0.8$, $r = 0.2$, $n = 2000$.

II.3 Comparison with the Constant Piecewise Weighted Logrank's test

In the situation where late effect appears at a certain date t^*, epidemiologists have in mind to use two logrank tests, one on the interval $[0, t^*]$ and one on the interval $]t^*, \tau]$. This can be formalized by the use of a weighted logrank test with a constant piecewise weight function defined by (6.2). This statistic has been studied in [12] in which we can found the analogue of Theorem 6.7. It is thus possible to generate data under which the CPWL(t^*)'s test is optimal. In [5], we investigate this test in the very same way as those of previous section. We notice that the CPWL's has good performances in term of empirical level and empirical power but is more sensitive to the value of t^* than Fleming-Harrington's test to the value of q.

Comparison by means of simulations.

Data generating process (DGP2). Let t_S^*, $c = S^P(\tau)$, a sample size n and a discrepancy rate r (in what follows, t_S^* will stand for "the value of t^* used for simulating the data"). We simulate data according to a generating process under which the CPWL(t_S^*) test is optimal. For this, the data in the placebo group are simulated from an exponential distribution with parameter a given by (6.15), and the data in the treatment group are simulated from the hazard function

$$\lambda^T(t) = a(1 - \Delta(t_S^*)\mathbb{I}_{\{t > t_S^*\}}) \tag{6.17}$$

where $\Delta(t_S^*)$ is given by

$$\Delta(t_S^*) = \frac{1}{a} \ln\left(\frac{S^T(\tau)}{S^P(\tau)}\right) \frac{1}{\tau - t_S^*}.$$

We consider well-balanced placebo and treatment groups. A sample simulated from this data generating process is denoted by $\mathcal{S}_2(t_S^*, n, r, c)$.

Simulation studies. To investigate the behaviour of the CPWL test (respectively Fleming-Harrington test) when the data are simulated under optimal alternatives for Fleming-Harrington test (respectively CPWL test). We consider two sets of scenarios for late differences:

- For each $q_S \in \{0, 1, 2, 3, 4\}$, we simulate $N = 2000$ samples $\mathcal{S}_1(q_S, 2000, 0.2, 0.8)$. We apply to the N samples the logrank test and the CPWL($\tau.\tau_T^*$) with $\tau_T^* = 0.2, 0.4, 0.6, 0.8$.
- For each $\tau_S = 0, 0.2, 0.4, 0.6$, we simulate $N = 2000$ samples $\mathcal{S}_2(\tau.\tau_S^*, 2000, 0.2, 0.8)$. We apply to the N samples the logrank test and Fleming-Harrington tests with $q = q_T = 1, 2, 3, 4$.

In each situation, we are thus able to calculate the empirical power. The Table 6.2 gives the empirical power of the CPWL test for the various combinations of q_S and t_T^* (that is, for data generated under optimal alternatives for Fleming-Harrington test).

Similarly, Table 6.3 gives the empirical power of Fleming-Harrington test when the data are generated under optimal alternatives for the CPWL. We provide results for $r = 0.2$, $c = 0.8$ and $n = 2000$.

As expected, we observe from the Table 6.2 that as q_S increases, the value of τ_T^* which ensures the largest power for a CPWL($\tau.\tau_T^*$) test increases (a similar remark holds from the Table 6.3 when τ_S^* increases). We also note that the power of the Fleming-Harrington test is less sensitive to q_T (for a given τ_S^*) than the power of CPWL($\tau.\tau^*$) is to τ^* for a given q_S. This confirms our previous finding that the Fleming-Harrington test is less sensitive to q than the CPWL test is to τ^*. In this sense, the Fleming-Harrington test should be preferred in practice.

q_S	Logrank	$\tau_T^* = 0.2$	$\tau_T^* = 0.4$	$\tau_T^* = 0.6$	$\tau_T^* = 0.8$
0	**0.644**	0.543	0.420	0.294	0.167
1	0.650	0.715	**0.719**	0.624	0.425
2	0.605	0.723	**0.790**	0.773	0.630
3	0.578	0.707	0.831	**0.873**	0.783
4	0.601	0.715	0.856	**0.918**	0.882

Table 6.2: Empirical power of the CPWL($\tau.\tau_T^*$) test when the data are generated under the optimal hypothesis for FH(q_S). $c = 0.8$, $r = 0.2$, $n = 2000$.

t_S^*	$q_T = 0$	$q_T = 1$	$q_T = 2$	$q_T = 3$	$q_T = 4$
0	**0.635**	0.512	0.402	0.329	0.276
0.2	0.620	**0.694**	0.608	0.515	0.452
0.4	0.623	**0.822**	0.814	0.766	0.707
0.6	0.594	0.896	0.948	**0.957**	0.953

Table 6.3: Empirical power of FH(q_T) when the data are generated under the optimal hypothesis for CPWL($\tau.\tau_S^*$). $c = 0.8$, $r = 0.2$, $n = 2000$.

Comparison by the use of the ARE

We have to be a bit more precise on the definition of ARE. If $(LR_{W_n^1})$ and $(LR_{W_n^2})$ are two sequences of weighted logrank tests, $\widetilde{ARE}(LR_{W_n^1}, LR_{W_n^2})$ will denote the ARE of $LR_{W_n^1}$ with respect to $LR_{W_n^2}$ under a sequence of alternatives such that $AE(LR_{W_n^2})$ is maximal. The following commutativity property holds:

Theorem 6.8 ([5]) *Assume that the Assumptions 6.1, 6.2 and 6.3 hold. Then*

$$\widetilde{ARE}(LR_{W_n^1}, LR_{W_n^2}) = \widetilde{ARE}(LR_{W_n^2}, LR_{W_n^1}).$$

Suppose that T is exponentially distributed under $\mathcal{H}_0$ and that the right-censoring time C is of type I. This means that under $\mathcal{H}_0$, $\pi(t) = S(t) = \exp(-at)$ and $\lambda(t) = a$, for $t \in [0, \tau[$. Then, in virtue of Theorem 6.8, for every $q \in \mathbb{R}^+$ and $t^* \in [0, \tau[$, the function:

$$f(q, t^*) = \widetilde{ARE}(LR_{W_n^q}, LR_{W_n^{t^*}}) = \widetilde{ARE}(LR_{W_n^{t^*}}, LR_{W_n^q}),$$

is well-defined and we have:

Theorem 6.9 *Given $t^* \in [0, \tau[$, there exists a unique $q(t^*) \in \mathbb{R}^+$ such that*

$$\max_{q \in \mathbb{R}^+} f(q, t^*) = f(q(t^*), t^*).$$

Given $q \in \mathbb{R}^+$, there exists a unique $t^*(q) \in [0, \tau[$ such that

$$\max_{t^* \in [0, \tau[} f(q, t^*) = f(q, t^*(q)).$$

Theorem 6.9 proves the existence and unicity of the maximum of the partial applications $q \to f(q, x)$ and $x \to f(q, x)$, but it tells nothing about the shape of the relations $q \to t^*(q)$ and $t^* \to q(t^*)$.

Numerical methods can be used to obtain some informations about these relations. In [5], plots of the graph of f are given and one observes that for every q (respectively t^*), there is a unique t^* (respectively q) such that the ARE is maximal. As an illustration, Table 6.4 provides the correspondence between t^* and q when $c = 0.8$, $n = 2000$ and $r = 0.2$.

Choice of the value of q

For a given trial, our proposal is thus to:

1. choose t^* based on *a priori* knowledge about the expected late effects,

FH(q)	$q =$	1	2	3	4
CPWL($\tau.\tau^*(q)$)	$\tau^*(q) =$	0.3	0.5	0.6	0.7
CPWL($\tau.\tau^*$)	$\tau^* =$	0.2	0.4	0.6	0.8
FH(q)	$q(\tau^*) =$	0.5	1.2	2.4	5.9

Table 6.4: Correspondence between q and t^* to give $f(q, t^*)$ maximal.

2. identify and use the test FH(q) which is the closest from CPWL(t^*) in terms of asymptotic efficiency.

Remark 6.2 *This procedure should be relevant only if the map $t^* \to q(t^*)$ is not too sensitive to the value of t^*. Graphically, this map is not a straight line, thus its sensitivity to a variation of t^* depends on t^*. But one can observe that the range of t^* where $t^* \to q(t^*)$ is sensitive is limited to a relatively extreme domain, which ensures a good stability of the choice of q for most of the t^* values.*

Sample size calculations

In [5], we investigate a comparison of those two tests by means of the necessary sample size. We calculate the sample size needed for testing the hypotheses (6.13) (respectively (6.17)) using Fleming-Harrington test (respectively the CPWL test) by Theorem 6.4 applied to the specific weight. Various setting are considered letting $\alpha = 0.05$, $\beta = 0.2$, the censoring fraction: 0.2, 0.5, 0.8 and the rate value: 0.1, 0.2, 0.3.

For both tests, as expected, the sample size needed to achieve the prescribed power and level increases as the censoring increases, and decreases when the rate r increases. Also, for Fleming-Harrington test, the sample size decreases when q increases from 1 (the sample size is sometimes larger for $q = 1$ than for $q = 0$). For the CPWL test, the sample size decreases when t^* increases from 0. We observe that the sample size needed for the Fleming-Harrington test is generally larger than for the CPWL test. The difference stays moderate however in most of the cases.

We have shown that the choice of q value procedure does not result in an unreasonable increase of the necessary sample size. The strategy proposed in the previous section therefore retains the nice features of both tests (namely the interpretation of t^* in terms of late effects and the robustness of FH(q) to the value of q).

III An "omnibus" test: maximum weighted Logrank statistic

The article [6] is devoted to the construction of a statistic for testing the hypothesis of equality of two survival distributions. This statistic denoted MWL is designed to have good power against both the late effects and proportional hazards alternatives. It is constructed as the maximum of the logrank and Fleming-Harrington statistics.

In what follows, we investigate the asymptotic distribution of the proposed statistic and we assess its performance via simulations studies. We also propose a sample size calculation procedure for this test.

III.1 MWL: definition and asymptotic distribution

We consider the testing problem

$$\begin{cases} \mathcal{H}_0 & : \ F^T = F^P = F, \\ \mathcal{H}_1 & : \ \cup_{i=1}^m \{F^T = \Psi^{q_i}(g + \theta^T(i)) \text{ and } F^P = \Psi^{q_i}(g + \theta^P(i))\}, \end{cases} \tag{6.18}$$

where $\Psi^q := \Psi^{0,q}$ is defined by (6.12) and for a given late effect alternative of the type q_i, the shift $\Delta(q_i)$ is given by $\Delta(q_i) := \theta^T(i) - \theta^P(i)$. For every $i = 1, \ldots, m$, let p_i be a known probability that reflects the investigator's degree of certainty that a late effect of type q_i occurs, with $\sum_{i=1}^m p_i = 1$ (note that if $q_i = 0$, the i-th alternative is proportional hazards). The maximum weighted logrank statistic is defined, for $\vec{q} = (q_1, \ldots, q_m) \in \mathbb{N}^m$ where $q_i \neq q_j$ for $i \neq j$ and for any $t \geq 0$ by

$$\mathrm{MWL}_n^{\vec{q}}(t) = \max_{i=1,\ldots,m} \left(\left| \frac{\mathrm{FH}_n^{q_i}(t)}{\widehat{\sigma}_{q_i}(t)} \right| \right).$$

To construct a decision rule, we need the asymptotic distribution of the process $\mathrm{MWL}_n^{\vec{q}} := \{\mathrm{MWL}_n^{\vec{q}}(t), t \geq 0\}$ under $\mathcal{H}_0$.

Theorem 6.10 ([4]) *Under $\mathcal{H}_0$, $\mathrm{MWL}_n^{\vec{q}}$ converges weakly to $\max_{i=1,\ldots,m}(|\tilde{\mathbb{G}}^{q_i}|)$ where $(\tilde{\mathbb{G}}^{q_1}, \ldots, \tilde{\mathbb{G}}^{q_m})$ is a m-variate zero mean Gaussian process with covariance function defined for any $i, j = 1, \ldots, m$ by*

$$(\tilde{\Sigma}_{i,j}^{\mathcal{H}_0})^2 : (t_1, t_2) \to \mathbb{E}[\tilde{\mathbb{G}}^{q_i}(t_1)\tilde{\mathbb{G}}^{q_j}(t_2)] = \frac{(\Sigma_{i,j}^{\mathcal{H}_0})^2(t_1 \wedge t_2)}{\Sigma_{i,i}^{\mathcal{H}_0}(t_1)\Sigma_{j,j}^{\mathcal{H}_0}(t_2)},$$

with

$$(\Sigma_{i,j}^{\mathcal{H}_0})^2 : t \to \int_0^t w^{q_i}(s)w^{q_j}(s)\frac{\pi^P(s)\pi^T(s)}{\pi(s)}d\Lambda_{\theta^0}(s)$$

and $w^q(s) = (1 - S(s))^q$.

Sketch of the proof. Under $\mathcal{H}_0$, it is easily seen that $(\mathrm{FH}_n^{q_1}/\widehat{\sigma}_{q_1}, \ldots, \mathrm{FH}_n^{q_m}/\widehat{\sigma}_{q_m}) \xrightarrow{\mathcal{L}(\mathbb{D}^m)} (\tilde{\mathbb{G}}^{q_1}, \ldots, \tilde{\mathbb{G}}^{q_m})$ where $(\tilde{\mathbb{G}}^{q_1}, \ldots, \tilde{\mathbb{G}}^{q_m})$ is as above. Moreover, the function $(F_1, \ldots, F_m) \to \max_{i=1,\ldots m}(|F_i|)$ is continuous from $\mathbb{D}^m$ to $\mathbb{D}$. □

Under $\mathcal{H}_1$, we are not able to reach the asymptotic distribution of $\mathrm{MWL}_n^{\vec{q}}$. However, we can establish the following weak convergence result which is sufficient to derive a sample size computation algorithm (see the section III.3 below). For $k = T, P$, let

$$\Lambda_{\theta^k} : t \to -\ln\left(\sum_{i=1}^m p_i(1 - \Psi^{q_i}(g(t) + \theta^k(i)))\right) \tag{6.19}$$

and for $i = 1, \ldots, m$, let

$$\mu^{q_i} : t \to \frac{1}{2}\int_0^t k^{q_i}(s)(d\Lambda_{\theta^P}(s) - d\Lambda_{\theta^T}(s)) \quad \text{with} \quad k^{q_i}(s) = (1 - S(s))^{q_i}\frac{\pi^P(s)\pi^T(s)}{\pi(s)}$$

$$(\Sigma_{i,j}^{\mathcal{H}_1})^2 : t \to \frac{1}{2}\int_0^t w^{q_i}(s)w^{q_j}(s)\left[\pi^P(s)\left(\frac{\pi^T(s)}{\pi(s)}\right)^2 d\Lambda_{\theta^P}(s) + \pi^T(s)\left(\frac{\pi^P(s)}{\pi(s)}\right)^2 d\Lambda_{\theta^T}(s)\right].$$

Theorem 6.11 ([4]) *Under $\mathcal{H}_1$, the m-variate process*

$$(\mathrm{FH}_n^{q_1}/\widehat{\sigma}_{q_1}, \ldots, \mathrm{FH}_n^{q_m}/\widehat{\sigma}_{q_m}) - \sqrt{n}\left(\mu^{q_1}/\Sigma_{1,1}^{\mathcal{H}_1}, \ldots, \mu^{q_m}/\Sigma_{m,m}^{\mathcal{H}_1}\right),$$

converges weakly to $(\tilde{\mathbb{G}}'^{q_1}, \ldots, \tilde{\mathbb{G}}'^{q_m})$ with $(\tilde{\mathbb{G}}'^{q_1}, \ldots, \tilde{\mathbb{G}}'^{q_m})$ a zero mean m-variate Gaussian process with covariance function

$$(\tilde{\Sigma}_{i,j}^{\mathcal{H}_1})^2 : (t_1, t_2) \to \mathbb{E}[\tilde{\mathbb{G}}'^{q_i}(t_1)\tilde{\mathbb{G}}'^{q_j}(t_2)] = \frac{(\Sigma_{i,j}^{\mathcal{H}_1})^2(t_1 \wedge t_2)}{\Sigma_{i,i}^{\mathcal{H}_1}(t_1)\Sigma_{j,j}^{\mathcal{H}_1}(t_2)}, \quad i, j = 1, \ldots, m. \tag{6.20}$$

Sketch of the proof. Under $\mathcal{H}_1$ similar arguments as those of Theorem 6.10 allow to prove that

$$(\mathrm{FH}_n^{q_1}/\widehat{\sigma}_{q_1}, \ldots, \mathrm{FH}_n^{q_m}/\widehat{\sigma}_{q_m}) - \sqrt{n}\left(\mu^{q_1}/\Sigma_{1,1}^{\mathcal{H}_1}, \ldots, \mu^{q_m}/\Sigma_{m,m}^{\mathcal{H}_1}\right) \xrightarrow{\mathcal{L}(\mathbb{D}^m)} (\tilde{\mathbb{G}}'^{q_1}, \ldots, \tilde{\mathbb{G}}'^{q_m})$$

where $(\tilde{\mathbb{G}}'^{q_1}, \ldots, \tilde{\mathbb{G}}'^{q_m})$ is defined above. Moreover under $\mathcal{H}_1$, it follows from the Bayes formula that the distribution of T in the group k $(k = T, P)$ is given by

$$\mathbb{P}_{\mathcal{H}_1}\{T < t\} = \sum_{i=1}^m \mathbb{P}_{\mathcal{H}_1}\{T < t \mid \Omega_i\}\,\mathbb{P}\{\Omega_i\} = \sum_{i=1}^m p_i\Psi^{q_i}(g(t) + \theta^k(i)),$$

where Ω_i is the event "a late effects of type q_i occurs". If follows that $\Lambda_{\theta^k}(t)$ is expressed as (6.19). □

Remark 6.3 *It is important to notice that $\mathrm{MWL}_n^{\vec{q}}$ is not necessarily the optimal statistic for testing the hypothesis* (6.18).

As shown in Section II, Fleming-Harrington's test is quite insensitive to the value of q provided $q > 0$. Thus one can restrict to $\vec{q} = (0, q)$ and in what follows, we will consider the test $\mathrm{MWL}_n^{\vec{q}}(\tau)$ with $\vec{q} = (0, q)$ (for notational simplicity, we will denote this test by MWL^q).

III.2 A simulation study

In this section, we assess via simulations the properties of the test statistic MWL^q. First, we evaluate the level and power (against the alternatives of proportional hazards (case 1) and late effects (case 2)) of MWL^q. Then we investigate the sensitivity of MWL^q to q.

Data Generating Process. The data generating process is the one described for Fleming Harrington's test page 86. It is important to notice that DGP1 is valuable for proportional hazards alternative $q = 0$ and for late effects alternative $q > 0$. In fact, for simulating the data in the treatment group, we consider, for a proportional hazards alternative, the following hazard function in the treatment group

$$\lambda^T(t) = ae^{\Delta(0)}.$$

which is exactly (6.16) with $q_S = 0$.

Performances of the test: Simulation Design. We consider several simulation scenarios obtained by combining various censoring proportions ($c = 0.2, 0.5, 0.8$), discrepancy rates ($r = 0.1, 0.2, 0.3$) and sample sizes ($n = 100, 1000, 2000$ with $n_P = n_T = n/2$). 2000 data sets are simulated for each combination of c, r and n. The logrank test FH^0 (LR thereafter), Fleming-Harrington's test for late effect FH^3 and the proposed test MWL^3 are applied to the resulting data (the nominal level is set to 0.05) and their respective empirical powers over the 2000 data sets are obtained.

Performances of the test: Results. The proposed MWL^q respects the nominal level. From Table 6.6, we observe that the power of the tests $\mathrm{LR}, \mathrm{FH}^3$ and MWL^3 increases with n and r and decreases when censoring proportion increases. We also verify that the logrank test (respectively Fleming-Harrington's test FH^3) has maximum power in the case 1 (respectively case 2) but performs badly when late effects (respectively proportional hazards) are present. In contrast, the proposed maximum weighted logrank test MWL^3 performs well in both cases. Its power is close to the maximum power whatever the true alternative is. To see this, we calculate for each test the relative variation (RV) of its empirical power p with respect to the maximum achieved power p_{max}. This is defined as

$$RV = \frac{|p - p_{max}|}{p_{max}}.$$

One clearly observes that the maximum weighted logrank test minimizes the relative variation in almost every simulation scenario. Moreover, this relative variation is rather stable with respect to c, n and r, which is not the case for Fleming-Harrington's test (case 1) and the logrank test (case 2).

These findings suggest that the proposed maximum weighted logrank test is an appealing compromise between the logrank and Fleming-Harrington tests when one wishes to test the equality of survival distributions without assuming *a priori* whether the true alternative is proportional hazards or late effects.

Sensitivity to q of the test: Simulation Design. Using MWL^q requires choosing q and one may wonder whether the test is sensitive to this value. To elucidate this issue, we generate, by the use of DGP1, 2000 data sets for each $q_S = 0, \ldots, 8$. Then for each q_S, we obtain the empirical power of the logrank test, of Fleming-Harrington's tests $\mathrm{FH}^1, \ldots, \mathrm{FH}^8$, and of $\mathrm{MWL}^1, \ldots, \mathrm{MWL}^8$. This simulation study is run for $c = 0.8$, $n = 2000$ and $r = 0.2$.

Sensitivity to q of the test: Results. The results of the sensitivity study are given in the table 6.5. We observe that using MWL^3 ensures a good power whatever the true alternative is (including the proportional hazards). Moreover, the test MWL^3 is more stable (in terms of power) than FH^3.

q_S	LR	FH^1	FH^2	FH^3	FH^4	FH^5	FH^6	FH^7	FH^8
0	**0.629**	0.526	0.416	0.334	5 0.289	0.256	0.229	0.211	0.192
1	0.625	**0.756**	0.744	0.702	0.655	0.611	0.580	0.545	0.513
2	0.609	0.839	**0.864**	0.863	0.850	0.835	0.812	0.781	0.754
3	0.623	0.869	0.919	**0.925**	0.922	0.910	0.901	0.890	0.87
4	0.626	0.891	0.943	0.959	0.961	**0.963**	0.961	0.960	0.953
5	0.608	0.911	0.963	0.976	0.978	**0.982**	0.980	0.979	0.976
6	0.597	0.916	0.964	0.982	0.987	0.989	**0.991**	**0.991**	0.990
7	0.562	0.910	0.962	0.978	0.986	**0.987**	**0.987**	0.986	0.986
8	0.564	0.909	0.970	0.988	0.993	0.995	**0.996**	**0.996**	**0.996**
q_s		MWL^1	MWL^2	MWL^3	MWL^4	MWL^5	MWL^6	MWL^7	MWL^8
0		**0.620**	0.606	0.589	0.584	0.582	0.579	0.571	0.571
1		0.729	**0.731**	0.720	0.692	0.679	0.671	0.657	0.650
2		0.797	**0.828**	0.826	0.816	0.801	0.784	0.773	0.758
3		0.833	0.881	**0.897**	0.896	0.888	0.880	0.875	0.859
4		0.864	0.923	0.936	**0.946**	0.945	0.943	0.940	0.934
5		0.880	0.947	0.959	0.967	0.968	**0.970**	0.967	0.968
6		0.887	0.950	0.971	0.978	0.983	**0.984**	0.983	0.983
7		0.874	0.946	0.967	0.974	0.977	0.978	**0.979**	0.978
8		0.881	0.954	0.977	0.985	0.989	0.988	0.989	**0.990**

Table 6.5: **Sensitivity to q of FH^q and MWL^q.** On each line, the data are generated using the procedure described in the section III.2, by using the value q_S ($q_S = 1, \ldots, 8$). The empirical power of LR, FH^q and MWL^q ($q = 1, \ldots, 8$) are obtained based on 2000 samples ($n = 2000$, $c = 0.8$ and $r = 0.2$). The values in bold represent the maximum over the lines.

			$q_S = 0$						$q_S = 3$					
n	c	r	LR		FH^3		MWL^3		LR		FH^3		MWL^3	
			EP	RV	EP	RV	EP	RV	EP	RV	EP	RV	EP	RV
100	0.2	0.1	**0.182**	-	0.113	0.38	0.163	**0.10**	0.105	0.44	**0.188**	-	0.152	**0.19**
		0.2	**0.486**	-	0.260	0.47	0.442	**0.09**	0.246	0.51	**0.500**	-	0.444	**0.11**
		0.3	**0.780**	-	0.461	0.41	0.736	**0.06**	0.531	0.38	**0.857**	-	0.831	**0.03**
	0.5	0.1	**0.074**	-	0.048	0.35	0.064	**0.14**	0.067	0.23	**0.085**	0.02	0.087	**-**
		0.2	**0.178**	-	0.107	0.40	0.167	**0.06**	0.145	0.43	**0.255**	-	0.224	**0.12**
		0.3	**0.310**	-	0.175	0.44	0.287	**0.07**	0.239	0.51	**0.489**	-	0.432	**0.12**
	0.8	0.1	**0.069**	-	0.052	0.25	0.060	**0.13**	0.060	0.21	**0.076**	-	0.073	**0.04**
		0.2	0.081	**0.06**	0.070	0.19	**0.086**	-	0.080	0.34	**0.121**	-	0.111	**0.08**
		0.3	**0.115**	-	0.083	0.28	0.105	**0.09**	0.119	0.43	**0.208**	-	0.189	**0.09**
1000	0.2	0.1	**0.905**	-	0.567	0.37	0.865	**0.04**	0.574	0.37	**0.914**	-	0.883	**0.03**
		0.2	**1.000**	-	1.000	-	1.000	-	0.990	**0.01**	**1.000**	-	1.000	-
		0.3	**1.000**	-	1.000	-	1.000	-	1.000	-	**1.000**	-	1.000	-
	0.5	0.1	**0.362**	-	0.199	0.45	0.324	**0.10**	0.259	0.49	**0.505**	-	0.445	**0.12**
		0.2	**0.899**	-	0.578	0.36	0.858	**0.05**	0.790	0.19	**0.979**	-	0.972	**0.01**
		0.3	**0.998**	-	0.899	0.10	0.997	**-**	0.985	0.02	**1.000**	-	1.000	-
	0.8	0.05	**0.070**	-	0.061	0.13	0.070	-	0.076	0.17	0.092	-	0.089	**0.03**
		0.1	**0.124**	-	0.090	0.27	0.117	**0.06**	0.124	0.44	**0.220**	-	0.195	**0.11**
		0.2	**0.381**	-	0.193	0.49	0.337	**0.12**	0.335	0.50	**0.665**	-	0.604	**0.09**
		0.3	**0.680**	-	0.361	0.47	0.620	**0.09**	0.682	0.29	**0.964**	-	0.945	**0.02**
2000	0.2	0.1	**0.997**	-	0.863	0.13	0.995	-	0.861	0.14	**0.999**	-	0.996	**-**
		0.2	**1.000**	-	1.000	-	1.000	-	1.000	-	**1.000**	-	1.000	-
		0.3	**1.000**	-	1.000	-	1.000	-	1.000	-	**1.000**	-	1.000	-
	0.5	0.1	**0.632**	-	0.312	0.51	0.566	**0.10**	0.163	0.46	**0.303**	-	0.272	**0.10**
		0.2	**0.994**	-	0.853	0.14	0.991	-	0.968	0.03	**1.000**	-	1.000	-
		0.3	**1.000**	-	0.995	0.01	1.000	-	1.000	-	**1.000**	-	1.000	-
	0.8	0.1	**0.205**	-	0.118	0.42	0.178	**0.13**	0.212	0.47	**0.397**	-	0.351	**0.12**
		0.2	**0.629**	-	0.334	0.47	0.589	**0.06**	0.623	0.33	**0.925**	-	0.897	**0.03**
		0.3	**0.947**	-	0.665	0.30	0.928	**0.02**	0.926	0.07	**0.999**	-	0.999	-

Table 6.6: **Empirical power (EP) and its relative variation (RV) for the logrank** (LR) **test, Fleming-Harrington's test** FH^3 **and maximum weighted logrank test** MWL^3. The data are simulated according to the procedure described in the section III.2, with $q_S = 0$ and $q_S = 3$. For every simulation scenario, the largest power and smallest RV are indicated in bold (c: censoring proportion, n: sample size, r: discrepancy rate (6.14)).

III.3 Sample size calculation

Several sample size formulas have been proposed for weighted logrank tests (see for example [8, 10, 2]). In [6], we derive a sample size formula for the test MWL^q. This formula being implicit, we describe a numerical algorithm for evaluating the necessary sample size. The proposed methodology is evaluated in a numerical study. First, we obtain the necessary sample size for testing the hypothesis (6.18) with MWL^q under various settings defined by the censoring proportion c ($c = 0.2, 0.5, 0.8$), the discrepancy rate r ($r = 0.1, 0.2, 0.3$) and the probability p_1 ($p_1 = 0.2, 0.5, 0.8$) (p_1 reflects the investigator's degree of certainty that proportional hazards occur). Then, we compare the sample sizes required by the logrank, Fleming-Harrington and maximum weighted logrank tests in a typical setting of a prevention trial that is, we consider $c = 0.8$ and $r = 0.2$ with $\alpha = 0.05$ and $\beta = 0.2$. p_1 is taken equal to 0.5, reflecting the fact that no preference is given to the proportional hazards or late effects alternative. We investigate the sensitivity of the necessary sample size to the probability p_1.

As may be expected, the necessary sample size for MWL^q increases when the censoring proportion increases and when the discrepancy rate r decreases (smaller late effects require more patients to be detected). Note also that the necessary sample size for MWL^q decreases when q increases (as for Fleming-Harrington's test). The necessary sample size also increases when p_1 increases: as the suspicion of proportional hazards increases, one needs more patients to decide whether proportional hazards or late effects occur. Finally, the necessary sample size for MWL^q is larger than for FH^q but the difference stays moderate.

IV Application: the GuidAge study

GuidAge is a randomized parallel-group double-blind trial registered to ClinicalTrials.gov under the number NCT00276510. Its primary outcome is the conversion of elderly subjects to probable Alzheimer's disease. The setting of the trial is as follows:

- it includes french elderly subjects ($\geq$ 70 years) who are free of dementia but have expressed a spontaneous memory complaint to their general practitioner
- the patients are randomized to either a preventive daily 240 mg dose of a standardised ginkgo biloba extract (EGb761) or a placebo
- the patients are followed during 5 years by their physician and in expert memory centres (712 physicians and 25 memory centres participate in the trial)

A former analysis of the trial was based on the logrank test. Assuming that under EGb761, the conversion rate from memory complaint to Alzheimer's disease is 25% less than under the placebo, the Alzheimer's disease-free rate after a 5-years long follow-up is equal to 89.63% under EGb761 and to 86.18% under the placebo. The total sample size ($n = 2800$) was calculated by letting $\alpha = 0.05$, $\beta = 0.2$ and by taking account of the drop out rate over the 5 years of follow-up. The result of the logrank test was negative, yielding the conclusion that there is no significant effect of the EGb 761.

However, EGb761 is a preventive treatment whose efficiency may require some preliminary exposure before an effect occurs. In that case, we rather suggest to use the statistic MWL^3 to test for a treatment effect. We set $p_1 = 0.5$ that is, we do not favor any of the proportional hazards or late effects alternatives (note that under this setting, the necessary sample size is $n = 2351$ and taking care of the 5% drop-out rate, it is $n = 3001$). The results of the analysis are given in the Table 6.7. The proposed MWL^3 has a p-value equal to 0.008 and from this, we conclude to a significant effect of the EGb761. Since the logrank test is not significant, this effect should be a late effect (for conciseness, we also investigated Fleming-Harrington's tests FH^q ($q = 2, 3, 4$) and the maximum weighted logrank tests MWL^q with $q = 2$ and 4. All these tests appear to detect a late effect of the EGb761).

	Logrank	FH^2	FH^3	FH^4	MWL^2	MWL^3	MWL^4
Test statistic	1.027	2.562	2.814	2.882	2.562	2.814	2.882
p-value	0.304	**0.010**	**0.004**	**0.003**	**0.018**	**0.008**	**0.006**

Table 6.7: **GuidAge study**. Various statistics and their p-values (in bold: significant results at the level 5%).

Bibliography

[1] Jean-Yves Dauxois. Convergence des processus de Nelson-Aalen et de Kaplan-Meier par une méthode de martingale. *Comptes Rendus de l'Académie des Sciences. Série I. Mathématique*, 328(11):1081–1084, 1999.

[2] Kevin H. Eng and Michael R. Kosorok. A sample size formula for the supremum log-rank statistic. *Biometrics*, 61(1):86–91, 2005.

[3] Thomas R. Fleming and David P. Harrington. *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1991.

[4] Valérie Garès, Sandrine Andrieu, Jean-Francois Dupuy, and **Nicolas Savy**. Choosing the parameter of fleming-harringtons test in prevention randomized controlled trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2013. Submitted.

[5] Valérie Garès, Sandrine Andrieu, Jean-Francois Dupuy, and **Nicolas Savy**. Comparison of constant piecewise weighted test and fleming harrington's test - application in clinical trials. *Electronic Journal of Statistics*, 2013. In revision.

[6] Valérie Garès, Sandrine Andrieu, Jean-Francois Dupuy, and **Nicolas Savy**. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. *Statistics in Medicine*, 2013. In revision.

[7] Richard Gill. *Censoring and stochastic integrals*. Mathematisch Centrum, 1980.

[8] M. Halperin, E. Rogot, J. Gurian, and F. Ederer. Sample sizes for medical tirials with special reference to long-term therapy. *Biometrics*, 21:13–24, 1967.

[9] Adam Jakubowski, Jean Mémin, and Gilles Pagès. Convergence en loi des suites d'intégrales stochastiques sur l'espace $\mathbf{D}^1$ de Skorokhod. *Probability Theory and Related Fields*, 81(1):111–137, 1989.

[10] Edward Lakatos and K. K. Gordon Lan. A comparison of sample size methods for the logrank statistic. *Statist. Med.*, 11:179–191, 1992.

[11] Aad W. Van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[12] David M. Zucker and Edward Lakatos. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika*, 77(4):853–864, 1990.

Chapter 7

Various elements for analysing mediation and evolution in Epidemiology*

This chapter is a summary of the results obtained by Dominique Dedieu and Benoît Lepage during their Ph.D. theses I have supervised with Pr Thierry Lang (INSERM Unit 1027). Benoît Lepage defended his Ph.D. June 21th, 2013 and the examination board was composed of Prof. Roch Giorgi (University of Marseille - referee) - Prof. Bruno Falissard (University of Paris XI - referee) - Prof. Daniel Commenges (Bordeaux Institute for Public Health) - Prof. Paul Landais (University of Montpellier) - Prof. Thierry Lang and myself (University of Toulouse III). Unfortunately, Dominique Dedieu stopped, for personal reasons, his thesis at the end of the first year.

As explained in the Introduction, the main skill of Thierry Lang's team is the so-called life-course approach of epidemiology. To deal with this approach, birth cohorts are of paramount interest. The one used for our investigations is the National Child Development Study cohort (NCDS) conducted in the UK by the Centre for Longitudinal Studies [19] . This study initially included all live births between 3 and 9 March 1958 in England, Wales and Scotland (just over 17,000 births). Surveys were subsequently conducted on the cohort members at ages 7, 11, 16, 23, 33, 42, 46 and 50 years resulting in a large number of variables (about 23,000). In the introduction one points out that this methods generates statistical methodology issues we aim to overpass.

- A large number of variables, not necessary of the same kind have to be considered. The model must take that complexity into account as well as the longitudinal nature.
- The collection of declarative variables over several decades yields to measurement errors.
- Monitoring data on 50 years of life necessarily induce missing data.
- The model must be able to describe the articulation of causal factors that are linked over time (mediation). That question of the evaluation of a causal effect by an observational study is the subject of an extensive literature [17].

The chapter organizes as follow. Section I is devoted to the approach followed by Dominique Dedieu by means of Mixed Markov Hidden Models. MMHM allows to deal with mediation seen as the transition from a state to another. Notice that these models tell nothing about causality. The "Hidden" aspect permit to consider error in declaration while the use of mixed model for transition include the error in

* Publications related to this chapter:

[7] Dominique Dedieu, Cyrille Delpierre, Sébastien Gadat, Benoît Lepage, Thierry Lang, **Nicolas Savy**. Mixed Hidden Markov Model for Heterogeneous Longitudinal Data with Missingness and Errors in the Outcome Variable. *Journal de la Société Française de Statistiques*, 155(1), pp. 73-98, 2014.

[13] Benoît Lepage, Dominique Dedieu, Sébastien Lamy, and **Nicolas Savy**, Thierry Lang. Using directed acyclic graphs for change score analysis. *Epidemiology*, 2013. In revision.

[14] Benoît Lepage, Dominique Dedieu, **Nicolas Savy**, Thierry Lang. Estimation of a controlled direct effect when an effect of the exposure confounds the mediator-outcome relationship: a comparison of different methods. *Statistic Methods for Medical Research*, 2013. Forthcoming.

measurement. Section II focuses on the works of Benoît Lepage and deals with the so-called structural causal model, described by Pearl [18]. It is the combination of features of the potential outcome framework of Rubin, of path analysis and of structural equation modelling. In the longitudinal context, marginal structural models have been shown to be useful to describe and evaluate causal effects.

I Approach by Markov Models

I.1 Generalities

Langeheine in [11] stresses that latent classes models are a general solution to successfully cope with measurement errors. In those works, a true latent (or hidden) quantity is distinguished from the measured (or declared) quantity. In a longitudinal framework, observations form a stochastic process and these observations depend on a second hidden process. Hidden Markov Models (HMM) belong to such a type of longitudinal models and has been intensively applied for analysing problems with measurement errors. The real health condition is described through a continuous hidden Markov process, and HMM allows to consider misclassification errors. Furthermore, HMM may help to deal with Missing Non At Random (MNAR) data. The corresponding model, fitted to clinical trial data, includes two extended Markovian processes for the outcome (which is partially hidden) and for the missingness indicator (which is completely observed), respectively, the latter being related to the former.

The model introduced in [7], incorporates both data missingness and error measurement. Finally, monitored cohort over the long term (frequently encountered in life-course epidemiology) may raise the problem of time heterogeneity in the response process as well as in the health condition transition, or even in the outcome definition. The outcome variable may concern present-day health conditions as well as certain past health-related events, which is inconsistent with the usual Markov assumption. This is the case in the NCDS 1958 cohort. It seems thus interesting to extend the usual Markov framework to take past events into account.

In the multi-state model, which includes HMM, Commenges notices in [5] that the assumption of homogeneous state transitions was very stringent, while in most cases the studied population is heterogeneous with regard to some relevant characteristics. He defines a model for state transitions involving observed covariates. However Commenges also observes in [4] that there may remain an unexplained heterogeneity following the adjustment for available covariates. This requires the introduction of random effects into the transition models. In this context, [1] introduced MHMM (Mixed Hidden Markov Models) which was applied to multiple sclerosis data.

I.2 The model

Each subject is described by a stochastic process $(S_t)_{t\geq 0}$ which quantifies the health state of the subject. Of course, $(S_t)_{t\geq 0}$ is not directly observed and is only known through the subject's declarations, which is represented using another stochastic process $(Y_t)_{t\geq 0}$. The process $(Y_t)_{t\geq 0}$ is related to the real hidden health state $(S_t)_{t\geq 0}$. Fig. 7.1 gives an example of such a model applied to NCDS' 58 cohort. The results and the interpretations are presented in [7].

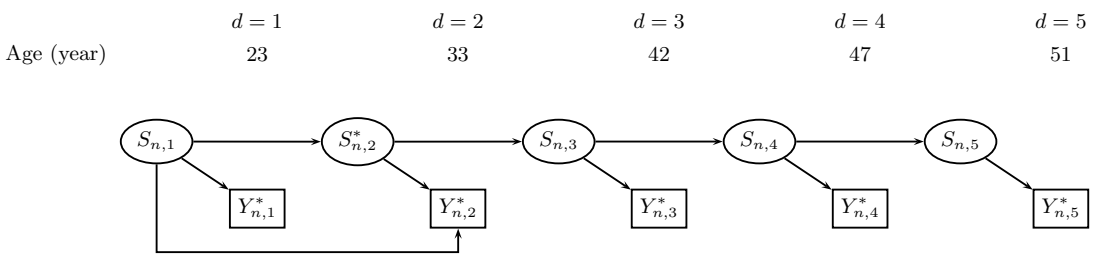

Figure 7.1: Study of cancer with the NCDS 1958 cohort by means of MHMM. $Y^*_{n,d}$: response of the subject n at date d and $S_{n,d}$ or $S^*_{n,d}$: true health state of the subject n on each time interval.

Longitudinal structure

We consider the evolution of N independent subjects. We assume that the time $0 \leq t \leq T$ is discretely sampled into a finite set of intervals $]t_d; t_{d+1}]$, with $1 \leq d \leq D$ such that $t_0 = 0$ and $t_{D+1} = T$. The intervals are assumed to be known at the beginning of the study. Hence, we denote by $(S_{n,d})_{1\leq n\leq N, 1\leq d\leq D}$ (resp. $(Y_{n,d})_{1\leq n\leq N, 1\leq d\leq D}$) the real health state (resp. the observed declarations) of subject n $1 \leq n \leq N$ at "time" d. The real health state $S_{n,d}$ is then described according to three possible states $\{0, 1, 2\}$ which code the situation of subject n at time d:

- if the disease is absent and the subject is alive in $]t_d; t_{d+1}]$, then $S_{n,d} = 0$,
- if the disease is present at any time of $]t_d; t_{d+1}]$, then $S_{n,d} = 1$,
- if the subject dies in $]t_d; t_{d+1}]$, $S_{n,d} = 2$.

The process $(Y_{n,d})_{1\leq n\leq N, 1\leq d\leq D}$ is described according to four possible states $\{0, 1, 2, 3\}$. Each of these states depends of course of the declaration of the subject n:

- $Y_{n,d} = 0$ if no disease is signalled along $]t_d; t_{d+1}]$,
- $Y_{n,d} = 1$ if the disease has been stated during $]t_d; t_{d+1}]$,
- $Y_{n,d} = 2$ if no response is obtained in $]t_d; t_{d+1}]$,
- $Y_{n,d} = 3$ if the subject dies during $]t_d; t_{d+1}]$.

For any subject n and any time d, the response $Y_{n,d}$ randomly depends on covariates. The observed ones are denoted $(\mathbf{X}_{n,d})_{1\leq d\leq D,\ 1\leq n\leq N}$ and the unobserved ones $(\mathbf{W}_n)_{1\leq n\leq N}$ since there are supposed homogeneous (independent on d) for the sake of simplicity. It is then natural to assume the following filtration properties

- $S_{n,d}$ is independent of $\{S_{n,d-k},\ 1 < k \leq d-1\}$ conditionally to $(S_{n,d-1}, \mathbf{X}_{n,d-1}, \mathbf{W}_n)$,
- $Y_{n,k}$ is independent of $\{Y_{n,d},\ d \neq k\}$ conditionally to $(S_{n,k}, \mathbf{X}_{n,k}, \mathbf{W}_n)$.

According to these several assumptions, we then obtain N Markov processes $(S_{n,d})_{d=1,\dots,D}$ which form, along with the $(Y_{n,d})_{d=1,\dots,D}$ processes, a MHMM. We will omit in the sequel the conditioning for covariates $\mathbf{X}_{n,1}, \dots, \mathbf{X}_{n,D}$ for clarity purposes.

Real state transitions model

These transitions concern the evolution of the true health state $(S_{n,d})_{1\leq n\leq N, 1\leq d\leq D}$ and are embedded in a Markov dynamic. Our model implies a time heterogeneity on stochastic behaviours for each of the subject introduced through the use of some covariates $\mathbf{X}_{n,d}$ (observed) and $\mathbf{W}_n$ (unobserved). These covariates directly influence the formal transition:

$$f_d(s, q, \mathbf{X}_{n,d}, \mathbf{W}_n, \theta^{trans}) = \mathbb{P}_{\theta^{trans}}(S_{n,d+1} = q \,|\, S_{n,d} = s, \mathbf{X}_{n,d}, \mathbf{W}_n). \tag{7.1}$$

Covariates $\mathbf{X}_{n,d}$ and $\mathbf{W}_n$ are taken into account by the use of a General Linear Model (GLM). More precisely, let us denote $\boldsymbol{\theta}^{trans}$ a the set of parameters $\theta^{trans} = (\theta^{trans,\mathbf{X}}, \theta^{trans,0})$ which stands for the influence of covariates $\mathbf{X}$ as well as the random effects that do not depend on the covariates. For each admissible transition $s \mapsto q$, $\theta^{trans,\mathbf{X}}_{s,q}$ is an unknown matrix which acts on the observed covariates $\mathbf{X}_{n,d}$ at time d on subject n. Second, the set of parameters $(\theta^{trans,0}_{s,q,d})_{s,q,d}$ stands for the natural transition from state s to state q at time d. At last, the covariates $(\mathbf{W}_n)_{1\leq n\leq N}$ model the individual randomness from one subject to another and each $\mathbf{W}_n$ is also a vector of $\mathbb{R}^7$. A linear predictor $\eta_{s,q,d}$ is defined as

$$\forall (s,q) \in \{0,1\} \times \{0,1,2\}, \forall d \geq 0, \quad \eta_{s,q,d}(\mathbf{X}_{n,d}, \mathbf{W}_n) = \theta^{trans,0}_{s,q,d} + \mathbf{X}'_{n,d}\boldsymbol{\theta}^{trans,\mathbf{X}}_{s,q} + (\mathbf{W}_n)_{s,q}\,, \tag{7.2}$$

and the transitions probabilities are defined by a multinomial logit model with η:

$$\forall (s,q) \in \{0,1\} \times \{0,1,2\}, \forall d \geq 0, \quad f_d(s, q, \mathbf{X}_{n,d}, \mathbf{W}_n, \theta^{trans}) = \frac{\exp(\eta_{s,q,d}(\mathbf{X}_{n,d}, \mathbf{W}_n))}{\sum_i \exp(\eta_{s,i,d}(\mathbf{X}_{i,d}, \mathbf{W}_i))}. \tag{7.3}$$

Response model

We describe here the probability to obtain response $Y_{n,d} = q$ from a real state $S_{n,d} = s$. The transition probabilities mainly rely on an emission parameter $\boldsymbol{\theta}^{em}$. From any state of $\{0, 1\}$, four responses are possible, each of them being related with an emission (or response) probability. Each emission has a specific interpretation :

- Response $Y_{n,d} = 1$ (disease) from a state $S_{n,d} = 0$ (no disease) is considered as an error.
- Response $Y_{n,d} = 0$ from a state $S_{n,d} = 1$ has several interpretations :
  - i) Subject n may not be ill when the question was asked and became sick just after or while the collecting of the data. If the question being asked concerns only the current health state, as it could apply to long-term observational cohorts, the information is lost.
  - ii) The diagnostic has not been told to the subject.
  - iii) The subject may present a denial behaviour.
- Non-response $Y_{n,d} = 2$ is only possible from $S_{n,d} \in \{0, 1\}$
- Of course, $Y_{n,d} = 3$ if and only if $S_{n,d} = 2$.

The parameter $\boldsymbol{\theta}^{em}$ quantify exactly the randomness in the response emission :

$$g_d(s, y, \boldsymbol{\theta}^{em}) := \mathbb{P}(Y_{n,d} = y \,|\, S_{n,d} = s) = \boldsymbol{\theta}^{em}_{s,y,d}. \tag{7.4}$$

Remark 7.1 *It would also be possible to describe a more general emission process which may involve the unobserved covariates $\boldsymbol{W}_n$ through a GLM following the same strategy already used for the definition of the functions f_d and $\eta_{s,q,d}$ introduced above.*

Initial state and unobserved covariates

We end the model statement by the description of the initial values of $(S_{n,1})_{1\leq n\leq N}$. In this view, we assume without loss of generality that for any subject n, $S_{n,1}$ belongs to $\{0, 1\}$ (initially dead subject won't be considered!). We then define $\boldsymbol{\theta}^{ini}$ as

$$\boldsymbol{\theta}^{ini,0} = \mathbb{P}(S_{n,1} = 0) \qquad \boldsymbol{\theta}^{ini,1} = \mathbb{P}(S_{n,1} = 1) = 1 - \boldsymbol{\theta}^{ini,0}.$$

Extension to retrospective data.

In some longitudinal studies, the subjects may be asked questions concerning both their present health state and their past health state. In this section, one presents an adaptation to the MHMM described in the previous section in order to analyse such data. We assume that at a certain (but not necessarily any) date t the subjects are asked two questions : "are you ill *now* ?" and "have you *ever* been ill ?".

Let fix $n \in \{1, \ldots, N\}$ a subject. We denote $Y^*_{n,d}$ the random response variable in the date interval d. We assume that $Y^*_{n,d}$ may only stand for the first question, or only for the second question, or may gather the response to both questions. In this latter case, we assume that the "non-response" level stands for both, and then we obtain six levels ((0,0), (0,1), (1,0), (1,1), the non-response levels (.,2) and (2,.) corresponds, by assumption, to one level and level 3 (death)). $(0, 0)$ tells us that the patient is not ill and have not been ill. The Markov hypothesis does no longer stand as $Y^*_{n,d}$ depends not only on the current state $S_{n,d}$ but on the complete state history $(S_{n,1}, ..., S_{n,d})$. Let us assume that $Y^*_{n,d}$ is independent from $Y^*_{n,k}$ $(k \neq d)$ conditionally to $S^*_{n,d} = (S_{n,d}, S'_{n,d-1})$, with $S'_{n,d-1}$ adopting value 1 if there exists some $k < d$ with $S_{n,k} = 1$ (with in addition $S_{n,l} \neq 2$ for all $l < d$).

We prove in [7] that the processes $(S^*_{n,d})_{d=1,\ldots,D}$ and $(Y^*_{n,d})_{d=1,\ldots,D}$ form a HMM. It is then possible to consider $(S^*_{n,d})_{d=1,\ldots,D}$ as a five-states Markov process taking values in $\{(0, 0), (0, 1), (1, 0), (1, 1), (2, 0)\}$. Notice that, due to the definition of the state memory, some state transitions are deterministic. Indeed, the transitions $(0, 0) \to (\cdot, 1)$ and $(0, 1) \to (\cdot, 0)$ are not possible.

I.3 The estimation procedure

Performing estimations for such mixed effects models is challenging, particularly due to the impossibility of computing the expected likelihood in a closed form expression, which generally implies expensive computational methods. Different approaches to the problem of discrete-time MHMM parametric estimation have been developed recently. In [1] author performs such estimations by the use of MCEM algorithm. In [2] authors develop, on a general framework, a stochastic EM approach (SEM). Using a SEM algorithm instead of performing, for example, a numerical integration. Delattre, in [8], makes use of SAEM algorithm a variant of SEM algorithm developed in [9, 12]. The convergence of SAEM has been studied by many different authors (see [12] and references therein) and is established under some assumptions especially in the context of the exponential family. These assumptions are valid for the MHMM emission model used by Delattre but no more in our multinomial logit setting. In the context of [7], we prefer to perform a punctual SEM estimation by simply averaging on the stochastic estimations. Moreover, Delattre proposes to simulate all "individual" transition parameters (missing covariates and real health states). Here the Metropolis-Hastings sampler may have a heavy cost as regards time computation (as it is an iterative procedure which must be performed for each subject). We prefer to compute and exact integration over the real health state (their number is limited in our model) and to use the simulation step of the SEM for the missing covariates. Details on the estimation procedure are given in the core of [7].

I.4 Results

In [7], we focus our attention on the MHMM described by Figure 7.1. We have made three kinds of investigations on this model.

- We investigate the quality of our estimation procedure by means of simulation studies.
- We investigate the robustness of the model by means of simulation study on perturbed models.
- We applied the model on NCDS' 58 dataset to put in light the mediated effect of an early social class on cancer through a smoking behaviour.

The results on simulated datasets are acceptable both in terms of estimation quality and in terms of model robustness. However, the method for estimating the effects of factors influencing the occurrence of this endpoint are difficult to implement, especially when the number of subjects is important. Moreover, despite the implementation of a stochastic algorithm, the computation time is very long and prevent full exploitation of the data. To overpass this difficulty one considers sub-samples what could yield to an underestimation of the variability of the estimates. The strategy used here could have been improved but, unfortunately, Dominique Dedieu decided to stop these investigations.

II Approach by Causal Structural Equations

In the context of observational studies, various methods can be proposed to estimate causal effects. The usual epidemiological methods such as analysis of a sub-sample of cases - paired control ("nested case-control study") may be used. But in a context of observation, for which exposure factors are not randomized, the recent literature dealing with the assessment of causal effects up most often in the context of the theory of "potential outcomes". Identifiability of the causal effect can be investigated by the use of Directed Acyclic Graphs (DAG).

II.1 An utilization of DAG to identify a causal effect in evolution study

In clinical epidemiology, we sometimes need to estimate the effect of an exposure of interest E (for example an anti-hypertensive treatment) on change from baseline of a time-dependent quantitative outcome (for example blood pressure at time t, denoted $\mathrm{BP}(t)$). The exposure E is observed at the beginning t_1 of the study (although it may have occurred before the beginning of the study), and a change score is defined by the difference $\Delta\,\mathrm{BP}$ in blood pressure between the beginning t_1 and the end t_2 of the study:

$$\Delta\,\mathrm{BP} = \mathrm{BP}(t_2) - \mathrm{BP}(t_1).$$

Two methods of estimating the effect of E on change from baseline have been regularly discussed: computing a linear regression of $\Delta\,\mathrm{BP}$ adjusted for baseline value $\mathrm{BP}(t_1)$ (sometimes called analysis of covariance)

or unadjusted for baseline value (sometimes called "simple analysis of change score"). For the individual i ($i = 1, \ldots, I$), the linear regression of $\Delta\,\mathrm{BP}$ on E adjusted for $\mathrm{BP}(t_1)$ is:

$$\Delta\,\mathrm{BP}_i = \mu + \tau_{\mathrm{BP}_1}\,\mathrm{BP}(t_1)_i + \tau_E E_i + \epsilon_i,$$

It is known that the regression coefficient τ_E can also be estimated using a linear regression of $\mathrm{BP}(t_2)$ on E adjusted for $\mathrm{BP}(t_1)$:

$$\mathrm{BP}(t_2)_i = \mu + (\tau_{\mathrm{BP}_1} + 1)\,\mathrm{BP}(t_1)_i + \tau_E E_i + \epsilon_i,$$

The linear regression of $\Delta\,\mathrm{BP}$ on E unadjusted for $\mathrm{BP}(t_1)$ is:

$$\Delta\,\mathrm{BP}_i = \mu' + \tau'_E E_i + \epsilon'_i,$$

The causal effect of E on $\Delta\,\mathrm{BP}$ is estimated by the regression coefficients τ_E or τ'_E according to the model chosen.

In some situations, the models can lead to very different results. This paradox was pointed out by Lord in the case of a non-randomized study exploring the effect of sex on weight change. In the literature, several factors have been discussed to choose the best model, for instance:

- Design of the study. Randomized - non-randomized [23] - cut-off design [22].
- Regression to the mean [23, 10].

In [13] we propose to guide the choice of a statistical model (linear regression adjusted or unadjusted for baseline outcome level) to estimate the causal effect of an exposure on change, using DAGs to represent each possible combination of study designs, regression to the mean phenomena. We used graphical rules and path analysis principles to interpret DAGs. As a complement, we illustrated each situation by simulated data compatible with the causal structure of the DAG and we estimated the effect of the exposure on change using both linear regressions.

For a sake of simplicity, we will deal with a binary exposure ($E = 1$ for exposure versus $E = 0$ for non-exposure). We also focused on the situation with a complete and equal follow-up for every participant (in the case of variable follow-up, one would have to discuss additional hypotheses about independence of the length of follow-up with other variables in the system). Four situations has been considered (see Figure 7.2 for the DAGs):

- Figure 1 represents four causal structures corresponding to randomized trials. The exposure E is independent of the baseline blood pressure $\mathrm{BP}(t_1)$ because of randomization.
- In Figure 2 the four initial causal structures have an additional baseline confounder (or a set of confounders) C which influences the exposure as well as $\mathrm{BP}(t_1)$ and $\mathrm{BP}(t_2)$.
- In Figure 3, we add a causal influence from the observed blood pressure $\mathrm{BP}^*(t_1)$ (defined below) to the exposure E in the causal structures of Figure 1. For example, an anti-hypertensive treatment may be more frequently given to patients with higher observed blood pressure at the beginning of the study. Such a causal structure also corresponds to the "cut-off design" mentioned by Senn.
- In Figure 4, the causal structures differ from the previous ones by an exposure which starts before the beginning of the study and influences both $\mathrm{BP}(t_1)$ and $\mathrm{BP}(t_2)$ in exposed subjects, as in the examples given for instance in [10].

And for each situation $X = 1, 2, 3$ or 4, we enrich the situation by

- the possibility that $\mathrm{BP}(t_1)$ influences $\Delta\,\mathrm{BP}$, for example through an intermediate mechanism (represented by the variable M in Figures XC and XD).
- adding the age at the beginning of the study ($age(t_1)$) which can be used to model the natural evolution of blood pressure with ageing in Figure XB and XD.

The variable R is a set of pre-existing individual characteristics which continuously influences blood pressure.

We used the notation $\mathrm{BP}^*(t)$ and $\Delta\,\mathrm{BP}^*$ for the observed blood pressure and change values. The observed blood pressure is influenced by the unmeasured (latent) blood pressure $\mathrm{BP}(t_1)$ and $\mathrm{BP}(t_2)$

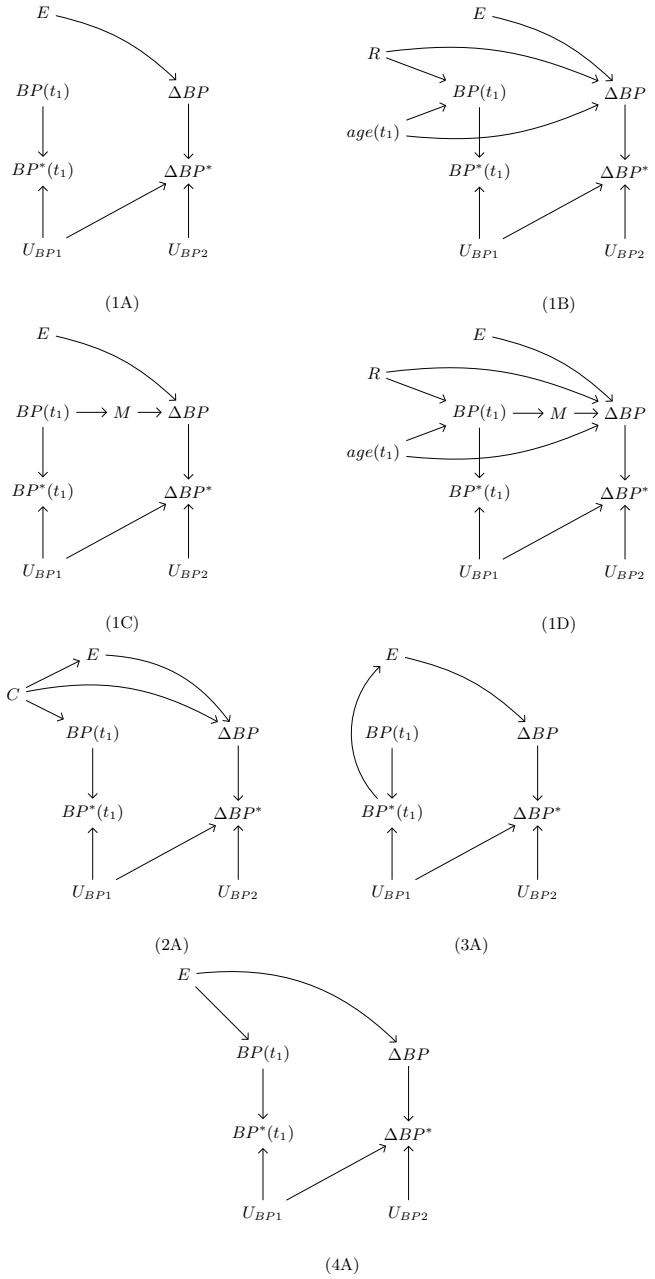

Figure 7.2: DAGs of the causal structures.

and intra-individual terms denoted U_{BP_1} and U_{BP_2} (which can include intra-individual variability and measurement error). We assume that for $j = 1, 2$,

$$\mathrm{BP}_j^*(t)_i = \mathrm{BP}_j(t)_i + (U_{\mathrm{BP}_j})_i$$

where U_{BP_j} are independent exogenous variables from a Gaussian distribution of mean 0 and variance σ_U^2.

The analyses of these DAGs together with paths analyses allow to generate data in accordance with the causal structure of the DAGs. In each scenario, 1050 samples of size 500 have been simulated according to the causal structures represented in Figures 1 to 4. The values are calculable and the bias for each method can be estimated. The results are ordered in Table 7.1.

	Figure 1		Figure 2	
Sub-figure	adjusted for $BP^*(t_1)$	unadjusted for $BP^*(t_1)$	adjusted for $BP^*(t_1)$	unadjusted for $BP^*(t_1)$
A	unbiased	unbiased	biased	unbiased
B	unbiased	unbiased	biased	unbiased
C	unbiased	unbiased	biased	biased
D	unbiased	unbiased	biased	biased

	Figure 3		Figure 4	
Sub-figure	adjusted for $BP^*(t_1)$	unadjusted for $BP^*(t_1)$	adjusted for $BP^*(t_1)$	unadjusted for $BP^*(t_1)$
A	unbiased	biased	biased	unbiased
B	unbiased	biased	biased	unbiased
C	unbiased	biased	biased	unbiased
D	unbiased	biased	biased	unbiased

Table 7.1: Bias in the estimation of the effect of the exposure E on blood pressure change (Δ BP), using linear regressions adjusted or unadjusted for $BP(t_1)$, in simulated datasets compatible with the causal structures represented in Figures 1 to 4.

II.2 An utilization of causal structural models in analysis of mediation

Generalities on mediation analysis

Mediation analyses aim to assess the role of intermediate factors (or mediators, denoted by M) in the causal relationship between an exposure X and an outcome Y (Figure 7.3).

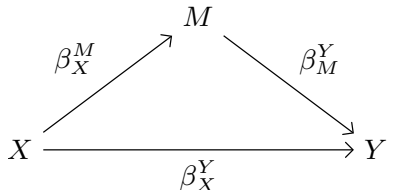

Figure 7.3: DAG for the representation of a classical causal situation in mediation analysis.

Currently, different methods exist for a statistical analysis of those mediation paths.

- In the social, psychological or medical literature, most mediation analyses have been based on two or three linear models: the total effect of the exposure is estimated as the effect of the exposure X on the outcome Y, the direct effect is estimated by the residual association between X and Y adjusted for the mediators and the indirect effect is the difference between total and direct effects. The indirect effect may also be estimated by the product of the path coefficient of X on M and the path coefficient of M on Y. It is easy to handle for models such as Figure 7.3.
- Structural Equations Models (SEM) have been proposed and used to analyse more complex theoretical models (several exposure factors, several mediators,...). SEM define a priori relationships assumed to be causal between variables (observed or not) and allows to estimate three kinds of effects:
  - direct effects which measure the effect of a variable on the events which does not pass through the identified mediation factors,
  - the indirect effects which are represented by the set of paths mediated by at least one intermediate variable between events and variable of interest.

– total effect which is the sum of direct effect and the various indirect effects.

Structural equations Models are related with Directed Acyclic Graphs (DAG) by Pearl [15]. Thanks to this theory, one can underscore a graphical criterion ("d-separation") which allows to detect fitting bias in structural schemes.

Remark 7.2 *This approach identifies, in causal chain, points that explain the largest share of variability and therefore where it is better to act in order to break the chains.*

- In [25], analysis of mediation is considered from counter-factual models of causality. Those methods spotlight several limits to SEM especially when one moves away from the assumptions of normality, of linearity, or in the presence of qualitative variables or interaction between the initial exposure and the mediating factor, or in some situations confusion. In contrast, the concept of "controlled direct effect" [16] defined in the framework of the theory of potential outcomes allows to describe such situations. Several methods are proposed to estimate those controlled direct effects: "Inverse probability of Treatment Weighted" (IPTW) method [21, 24], "g-computation" [20], "sequential g-estimation" [26], "structural nested models" [21]...

Theory of potential outcomes allows to define mediated effects in a more general way than the one of SEM [26]. Specific definitions of direct and indirect effects have been given in the literature on causal inference. Based on the notions of counter-factual and using the potential outcome notation: $Y_x(u)$ is the value Y would take in the statistical unit u under a given exposure $X = x$; $M_x(u)$ the value of M under the exposure $X = x$; and $Y_{xm}(u)$ the value of Y under the exposure $X = x$ when M is controlled at the level $M = m$. The controlled direct effect (CDE) is the effect of the exposure that would be observed if the mediator was controlled at the value m in the whole population. The average CDE when M is fixed at m is defined by:

$$\mathrm{CDE}_m = \mathbb{E}\left[Y_{xm}\right] - \mathbb{E}\left[Y_{x^*m}\right]$$

where $X = x^*$ is the reference value. In order to estimate a CDE without bias in an observational study, some unmeasured confounding assumptions are necessary and estimation based on adjusting for the mediator can be inappropriate. Graphical conditions for the identification of CDEs have been described by Pearl [17].

Analysis of a biased case

When estimating the CDE_m with a regression of X on Y adjusted for M, the bias comes from conditioning on the collider M ($X \to M \leftarrow Z$). In [14], we focus on the causal structure defined by the DAG in Figure 7.4 which is a standard situation in Epidemiology. In this situation, the estimation is known for being biased. One explores the size of the bias when varying the effect of edges of the DAG in simulated data sets as well as in a real data (IHPAF study) example and compares methods of estimations. We mainly focused on the case of no unmeasured confounding, where functional relations between variables are linear with no interactions or effect modifications.

To do this, several configurations of the causal structure has been explored by varying the intensity of each arc of the paths between X to Y. The controlled direct effect of X on Y (for a fixed level of the mediator M) has been estimated for each simulated sample by different methods. One compare five methods of estimating the average CDE_m.

- **Simple adjusting for M.**
- **Marginal Structural Models estimated via IPTW [21].** That is used to estimate the marginal expected value of the potential outcomes Y_{xm} when we fix the exposure and the mediator to $X = x$ and $M = m$.
- **Marginal Structural Models estimated via IPTW with truncated weights [3].** This method has been suggested to decrease the standard error (SE) of the estimated causal effects resulting from extreme weights in MSMs, especially when X and M are continuous. However, weight truncation can also increase the bias of the estimation. In [3] weights were truncated at the 1st and the 99th percentiles of the weight distribution.
- **Sequential g-estimation for structural mean models [26].**
- **g-computation with Monte Carlo simulation [6].**

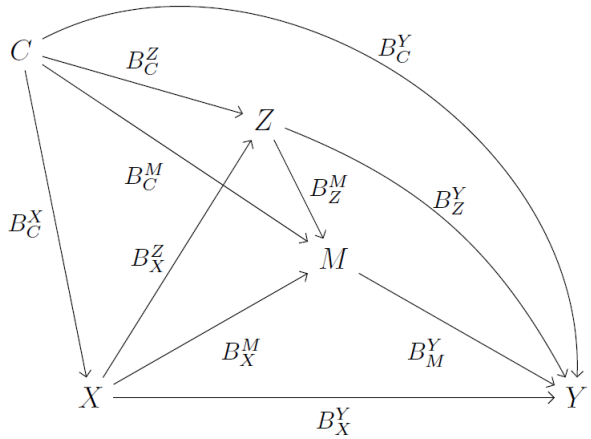

Figure 7.4: DAG of the situation of interest.

We computed the 95% CI of the CDEs using a non-parametric bootstrap method, based on 3000 bootstrap samples. The five above-mentioned estimation methods were applied in every bootstrap sample and the bounds of the 95% CI were calculated from the distribution of the CDE bootstrap estimates.

Results.

The results of this work illustrate that the sequential method or the g-computation show the best guarantees of performance in both case exposure variables and / or variables of mediation continuous and binary. The structural equation models gives similar results but are less well suited to face qualitative judgement criteria or to take into account possible interactions between exposure and mediators. Estimated by inverse weighting, is of comparable quality to the g-computation in case of exposure variables and binary mediation. In a situation with several mediating variables and several intermediate confusion factors measured repeatedly during the follow-up, as is the case in our analysis on the relationship between adversity and future health status data from cohort 58, the computation method of g-computation appears as the most suitable.

Bibliography

[1] Rachel M. Altman. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210, 2007.

[2] Gilles Celeux and Jean Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports*, 41(1-2):119–134, 1992.

[3] Stephen R. Cole and Miguel A. Hernàn. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–64, 2008.

[4] Daniel Commenges. Multi-state models in epidemiology. *Lifetime Data Analysis*, 5(4):315–327, 1999.

[5] Daniel Commenges. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11(2):167–182, 2002.

[6] Rhian M. Daniel, Bianca L. De Stavola, and Simon N. Cousens. g-formula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal*, 11(4):479–517, 2011.

[7] Dominique Dedieu, Cyrille Delpierre, Sébastien Gadat, Benoît Lepage, Thierry Lang, and **Nicolas Savy**. Mixed hidden markov model for heterogeneous longitudinal data with missingness and errors in the outcome variable. *Journal de la Société Française de Statistiques*, 155(1):73–98, 2013.

[8] Maud Delattre. Inference in mixed hidden Markov models and applications to medical studies. *Journal de la Société Française de Statistique*, 151(1):90–105, 2010.

[9] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[10] Maria Glymour, Jennifer Weuve, Lisa F Berkman, Ichiro Kawachi, and James M Robins. When is baseline adjustment useful in analyses of change? an example with education and cognitive change. *American Journal of Epidemiology*, 162(3):267–78, 2005.

[11] Jacques A. Hagenaars and Allan L. McCutcheon, editors. *Applied latent class analysis.* Cambridge University Press, Cambridge, 2002.

[12] Estelle Kuhn and Marc Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

[13] Benoît Lepage, Dominique Dedieu, Sébastien Lamy, **Nicolas Savy**, and Thierry Lang. Using directed acyclic graphs for change score analysis. *Epidemiology*, 2012. In revision.

[14] Benoît Lepage, Dominique Dedieu, **Nicolas Savy**, and Thierry Lang. Estimation of a controlled direct effect when an effect of the exposure confounds the mediator-outcome relationship: a comparison of different methods. *Statistic Methods for Medical Research*, 2012. Forthcoming.

[15] Judea Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2000.

[16] Judea Pearl. Direct and indirect effects. In D. Koller and J. Breese, editors, *Proceedings of the Seventeenth Conference on Uncertainy in Articial Intelligence*, pages 411–420. Morgan Kaufmann, San Francisco, CA, 2001.

[17] Judea Pearl. *Causality: models, reasoning, and inference.* Cambridge University Press, Cambridge, U.K.; New York, second edition, 2009.

[18] Judea Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):Article7, 2010.

[19] Chris Power and Jane Elliott. Cohort profile: 1958 british birth cohort (national child development study). *International journal of epidemiology*, 35(1):34–41, 2006.

[20] James M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–55, 1992.

[21] James M. Robins, Miguel A. Hernàn, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–60, 2000.

[22] Stephen Senn. Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25(24):4334–44, 2006.

[23] Gerard J.P. Van Breukelen. Ancova versus change from baseline: more power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, 59(9):920–925, 2006.

[24] Tyler J. Vanderweele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26, 2009.

[25] Tyler J. Vanderweele and S. Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468, 2009.

[26] Stijn Vansteelandt. Estimating direct effects in cohort and case-control studies. *Epidemiology*, 20(6):851–60, 2009.

Chapter 8

Various results in interaction with Biology

This chapter briefly presents some papers in interaction with biology. The first section is a summary of a series of three articles [8, 9, 12] written during my first position in Quimper. These articles are the result of a collaborative works with the Laboratory of Applied Microbiology of Quimper and deal with predictive microbiology. The main lines of that topic are described below. The second section presents in a few words the results obtained in collaboration with Prof. Antoine Blancher, director of the Molecular Immunology Laboratory of Toulouse. In the two papers [1, 4] we have written, statistics come support the arguments of geneticists.

I Predictive microbiology*

The aim of predictive microbiology is to describe the evolution of bacterial populations in food. The extreme complexity of the behaviour of bacteria associated with worries of risk management make this modelling difficult. It is usually done in two steps:

- Primary modelling: one describes the number of bacteria as a function of time $N(t) = f_\theta(t)$ where θ is a family of parameters.
- Secondary modelling: on expresses θ as a function of environmental parameters.

The constraints of predictive microbiology is to get simple models whose parameters have a biological reality and which fit well to the observed data.

Various primary models can be apprehended. In [8, 9] we worked on heat treatment this means t is to be understood as the heat treatment duration (not time). In [12] we worked on regrowth after thermal treatment t is the time after this treatment.

I.1 Primary models

Primary model for bacteria growing

In [12], we study the impact of heating treatment on bacteria regrowth, focusing on the lag time. It is a well-known fact that the curve $(t, \log N_t)$ can be decomposed in three steps: an initial constant one, an exponential one and a late constant one. The exponential period is characterised by the maximal growth

* Publications related to this paragraph:

[8] Louis Coroller, Ivan Leguerinel, Eric Mettler, **Nicolas Savy**, and Pierre Mafart. General model, based on two mixed Weibull distributions of bacterial resistance, for describing various shapes of inactivation curves. *Applied Environmental Microbiology*, 72(10):6493–6502, 2006.

[9] Olivier Couvert, Stéphane Gaillard, **Nicolas Savy**, Pierre Mafart, and Ivan Leguérinel. Survival curves of heated bacterial spores: effect of environmental factors on Weibull parameters. *International Journal of Food Microbiology*, 101(1):73–81, 2005.

[12] Stéphane Gaillard, Ivan Leguérinel, **Nicolas Savy**, and Pierre Mafart. Quantifying the combined effects of the heating time, the temperature and the recovery medium pH on the regrowth lag time of *bacillus cereus* spores after a heat treatment. *International Journal of Food Microbiology*, 105(1):53–58, 2005.

rate μ_{max}, slope of the tangent of the curve at the inflection point. Lag time λ is usually defined as the intercept of this tangent with the horizontal line passing through $\log N_0$. The main topic of this paper is to deal with the secondary model, see the brief summary in section I.3

Primary model for heating treatment [9]

The first order kinetic model describing inactivation of micro-organisms is generally expresses by:

$$N_t = N_0 \ \exp(-kt), \qquad \text{or} \quad \log N_t = \log N_0 - \frac{t}{D}, \tag{8.1}$$

where N_0 is the initial number of cells, N_t the number of surviving cells after a duration of heat treatment t and k is the first order parameter. The second expression is preferred because the classical D-value presents a simple biological significance: time that leads to a 10-fold reduction of surviving population and is easily estimated from a simple linear regression. At least up to 2005, this concept governed canning process calculation and was the one currently used in Food and Pharmaceutical Industry. However, in many cases, the survival curves of heated bacteria does not present a log linear relation: a concave or upward concavity of curves has frequently been observed [6]. So the bacterial heat resistance cannot be evaluated from the classical D-value. Consequently, many authors proposed models. These models show accuracy but are either over parametrized or have parameters without any physical or biological significance. Moreover, the complexity of these models hinders their application in heat treatment process calculations. Other authors who considered the survival curve as a cumulative form of temporal distribution of lethality event distribution, presented a probabilistic approach (for instance [14]). The Weibull frequency distribution model invoked to describe the time to failure in mechanical system was applied to bacterial death time. Different forms of this model were presented in the literature; however, the decimal logarithm form below, which is close to (8.1), seems more suitable to describe non-log linear survival curves

$$\log N_t = \log N_0 - \left(\frac{t}{\delta}\right)^p. \tag{8.2}$$

where δ is the first reduction time that leads to a 10-fold reduction of the surviving population and p the shape parameter.

Primary model for heating treatment subjected to an acidic stress [8]

Cells of *Listeria monocytogenes* or *Salmonella enterica serovar Typhimurium* taken from six characteristic stages of growth were subjected to an acidic stress (pH 3.3). As expected, the bacterial resistance increased from the end of the exponential phase to the late stationary phase. Moreover, the shapes of the survival curves gradually evolved as the physiological states of the cells changed. In [8] a new primary models, based on two mixed Weibull distributions of cell resistance, is proposed to describe the survival curves and the change in the pattern with the modifications of resistance of two assumed sub-populations:

$$N_t = N_0 \ \left[f\, 10^{\left(-\frac{t}{\delta_1}\right)^{p_1}} + (1-f)\, 10^{\left(-\frac{t}{\delta_2}\right)^{p_2}} \right], \tag{8.3}$$

where t is time, N_0 is the initial bacterial concentration, f is the fraction of the original population in the major group and the subscripts 1 and 2 indicate the two different sub-populations. Sub-population 1 is more sensitive to stress than sub-population 2 is ($\delta_1 < \delta_2$). This model was compared to Whiting's model [16] which was the model of reference.

The parameters of the proposed model (8.3) were stable and showed consistent evolution according to the initial physiological state of the bacterial population. Compared to the Whiting's model, the proposed model allowed a better fit and more accurate estimation of the parameters. Finally, the parameters of the simplified model had biological significance, which facilitated their interpretation.

I.2 Secondary models

For a long time, secondary modelling remained at stage mono-factorial, where only the heating temperature was taken into account. The model is that of Bigelow [3]:

$$\log D = \log D^* - \frac{T - T^*}{Z_T},$$

where D^* is the calculated D-value at temperature T^* and Z_T is a distance of T from T^*, which leads to a 10-fold reduction D-value. By the use of factorial design, it is possible to put in light the environmental factors which affect the heat resistance of bacteria, the pH of the heating medium and the pH of the recovery medium (pHV) are shown to be of prominence. Couvert *et al.* [10] has developed an extended Bigelow model to describe both effects of heating and recovery medium pH on the apparent bacterial spore heat resistance.

$$\log D = \log D^* - \frac{T - T^*}{Z_T} - \left|\frac{pH - pH^*}{Z_{pH}}\right| - \left|\frac{pH' - pH'^*}{Z_{pH'}}\right|^2,$$

where $pH*$ and pH'^* are the reference heat treatment and recovery medium pH fixed to 7, Z_{pH} is a distance of pH from $pH*$, which leads to a 10-fold reduction D-value. z_{pH} quantifies the influence of heat medium pH influence on bacterial heat resistance. $Z_{pH'}$ is a distance of pH' from pH'^*, which leads to a 10-fold reduction in apparent D-value and characterizes the influence of the pH on the recovery of the microorganism after a heat treatment. $D*$ is the calculated D-value corresponding to pH^* and pH'^* conditions.

Like the Bigelow model, Couvert's model was suitable for the calculation of δ-values as well as for those of D-values. However, the influence of heating temperature on the value of p is not clear and variable according to several authors.

I.3 Parameters estimation and confidence intervals

Two techniques were applied to reach confidence intervals of the model parameters. First, the method described in [13] based on works of Beale [2] and a second one by bootstrap.

Summary of the results of [9]

In [14] authors observed that the shape parameter (p) of Weibull model seams to be characteristic of the bacterial seed. In [9] we compare a model where the shape parameter is estimated for each sample of the factorial design and a model with a constant value of p for given microbial population. The bootstrap procedure had to be adapted in order to take into account the difference in the sample size of each sample path.

The strength shape of the projections and the high correlation coefficient indicate a structural correlation between model parameters. Three Weibull model parameters were fitted to each inactivation data set and correlation coefficients were determined from the evaluated confidence region, for the 18 environmental conditions studied presents the estimates of the structural correlation between parameters for all kinetics. Thus, Weibull model parameters ($\log N_0, \delta, p$) are dependent (Figure 8.1): an error on δ will be balanced by an error on p in the same way. Finally, a single value of p estimated from the

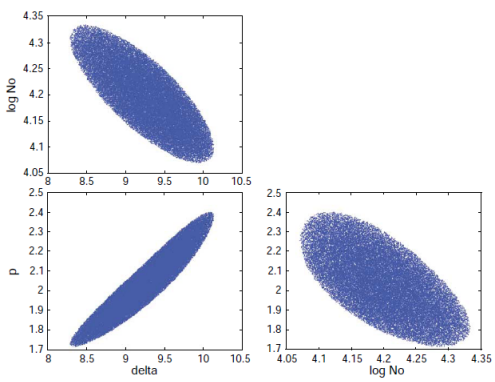

Figure 8.1: Projection of the confident region on three orthogonal planes, from *Bacillus pumilus* A40 data (heating temperature: 95.8 °C, heating and recovery medium pH: 7).

whole set of kinetics eliminates the structural correlation between δ and p parameters as well as $\log N_0$

and p parameters and decreases the structural correlation between $\log N_0$ and δ. The Weibull model parameters become independent.

Couvert's model including the dependence temperature and heating and recovery medium pH, was fitted on the δ-values evaluated with the two calculation methods. Observed and calculated values were compared and one showed a slightly higher accuracy of Couvert's model when the δ-values were evaluated with single p-value. The Weibull model is suitable for describing log linear, or not, heat survival curves. However, a simplification of this model consisting in getting a single overall estimation of p-value per strain, regardless of environmental conditions of heat treatment and recovery, seems to be enough for bacterial food predictive modelling and canning process calculation.

Summary of the results of [12]

The purpose of this study was to quantify the lag time of re-growth of heated spores of *Bacillus cereus* as a function of the conditions of the heat treatment: temperature, duration and pH of the recovery medium. For a given heating temperature, curves plotting lag times versus time of heating show more or less complex patterns. However, under a heating time corresponding to a decrease of 2 decimal logarithms of the surviving populations of spores, a linear relationship between the lag time of growth and the time of the previous heat treatment can be observed: $\lambda = \lambda_0 + mt$, where t is the heating time. The slope m of this linear relationship followed itself a Bigelow type linear relationship,

$$\log m = \log m^* + \frac{T - T^*}{\zeta_T} + \frac{pH - pH_{opt}}{\zeta_{pH}},$$

the slope of which yielded ζ-values very close to the observed conventional values. It was then concluded that the slope of the regrowth lag time versus the heating time followed a linear relationship with the sterilisation value reached in the course of the previous heat treatment. A sharp effect of the pH of the medium which could be described by a simple secondary model was observed. As expected, the observed intercept of the linear relationship between lag time and heating time (lag without previous heating) was dependent on only the pH of the medium and not on the heating temperature.

II Population Genetics†

The Molecular Immunology Laboratory of Toulouse used to work on genetic markers (microsatellite of variable inter-allelic length) localized in the region of the major histocompatibility complex (MHC). Ours investigations are based on a sample of 750 macaques from Mauritius population [5]. One genotyped 750 Mauritian animals for 17 microsatellites markers spreads across the MCH. This allowed us to characterize seven frequent MHC haplotypes (a haplotype corresponds to the combination of alleles present on a given chromosome) and 25 percent of recombinant haplotypes (see therafter). The major advantage of the Mauritian population is strong bottleneck that followed the founding of the macaque population by the release in the wild of a low number of founders captured in Indonesia. Indeed the macaque Mauritian population was founded about 400 years by the introduction of some animals from Indonesia (most likely is Java but some authors also believe in Sumatra, and perhaps also Malaysia). By historical sources we know that these animals were introduced by Dutch sailors. The founder animals quickly adapted to the Mauritian environment and has the population quickly grew. According to historical sources we know that about one century after the arrival of animals, their number was such that they were everywhere on the island and they ravaged crops. Many settlers left the island for this reason. The population of macaques stabilized due to limitation of food resources. Because of the small number of founders, it is easy to understand why the number of frequent (and most probably founding) haplotypes is limited in the nowadays population (we found only seven founding haplotypes). At each generation recombination occurred between parental haplotypes, in theory at random, explaining that 25% of haplotypes are recombinant in the nowadays population.

† Publications related to this paragraph:

[1] Alice Aarnink, **Nicolas Savy**, Nicolas Congy, Nicola Rosa, Edward Mee and Antoine Blancher. Demonstration of the deleterious impact of foeto-maternal MHC compatibility on the success of pregnancy in a macaque model. *Immunogenetics*, 66 pp. 105-113, 2014.

[4] Antoine Blancher, Alice Aarnink, **Nicolas Savy**, and Nagushi Takahata. Use of cumulative Poison probability distribution as an estimator of the recombination rate from population-genetic data: example of the Macaca fascicularis MHC. *Genes, Genomes, Genetics*, 2:123–130, 2012.

The first work [4] is an estimation of recombination rates from the frequencies of non-recombinant haplotypes and single, dual, triple, ..., of order n recombinant haplotypes, obtained from a sample of macaques from Mauritius. In the second work [1], we studied from the MHC genotypes of animals of a British breeding macaques from Mauritius, the impact of the compatibility between the mother and her offspring in the pregnancy outcomes.

II.1 Estimated recombination rates

In a population founded at a given time with a limited number of animals having founding haplotypes, at each generation there is a decrease in the frequency of founding haplotypes due to recombination . This decrease is of course directly related to the frequency of recombination. Theoretically, we can model the haplotype frequencies with $0, 1, 2, 3, 4, \ldots, n$ recombinations after X generations by a Poisson distribution. In [4], we describe a method to estimate the rate of recombination per generation from the genotypes of a large individual sample of an expanding population, for which the founding event is dated. The idea is simply to fit a Poisson distribution to data by maximum likelihood. On the genetic point of view, the paper develops the difficulties of describing the haplotype frequencies. The problems are multiple and complex. First, the recombination of two identical haplotypes of a homozygous animal is mute. The observed frequency has been corrected to actual frequency by estimating the average frequency of homozygotes in the population of Mauritius. Second, a difficulty with the lack of markers whose density does not describe all the recombination events (double recombination between two markers is obviously mute). Finally, by dividing the intensity of the Poisson distribution by the number of generations (50-100) from the date of the founding population, we deduced that the rate of recombination in the MHC is approximately 0.004 to 0.008 in the Mauritian macaque population.

Usual methods for calculating the rate of recombination is based on the theory of coalescence [11]. Our estimate was compared with the results obtained by the first Bayesian algorithm (is even now one of the best): the "PHASE" software [15]. The results are quite comparable. The method proposed in [4] presents an alternative to the usual method of coalescence within genetically isolated populations that experienced a strong bottleneck and when the founding dates of the population is precisely known.

Finally, a model of recombination in an population equivalent to that of Mauritius has been simulated by Kiril Kryukov, a informatician of the laboratory of Professor Saitou (National Institute of genetics, Mishima, Japan), in order to deduce the recombination rate which would be consistent with the observed results (proportion of recombinant haplotypes we observed in the sample of 750 animals). The results are the same as ours.

II.2 Impact of compatibility between the mother and her offspring in pregnancy outcomes

In [1], we studied the frequency of pairs mother/child compatible for the MHC and the pairs mother / child MHC semi-compatible (the child can not be totally incompatible with his mother having inherited the latter half of MHC alleles). The problem then boils down to a comparison between observed and theoretical numbers (the probability of generating offspring compatible with the mother is 0.5 for breeding pairs informative in the sense that we have defined in this article). View the modest sample size, we opted for a direct comparison with a binomial or multinomial distribution by using the results in [7]. The observed repartition was clearly outside the interval of confidence of 99 %, and therefore most probably resulted from a selection of the semi-compatible foetuses during pregnancy. We concluded that MHC fully compatible cynomolgus macaque foetuses have a selective survival disadvantage in comparison with foetuses inheriting a paternal MHC haplotype differing from maternal haplotypes.

Bibliography

[1] Alice Aarnink, **Nicolas Savy**, Nicolas Congy, Nicola Rosa, Edward Mee, and Antoine Blancher. Demonstration of the deleterious impact of foeto-maternal mhc compatibility on the success of pregnancy in a macaque model. *Immunogenetics*, 66:105–113, 2014.

[2] Evelyn M. L. Beale. Confidence regions in non-linear estimation. *J. Roy. Statist. Soc. Ser. B*, 22:41–88, 1960.

[3] Willard D. Bigelow. The logarithmic nature of thermal death time curves. *J. Infect. Diseases*, 29:528–536, 1921.

[4] Antoine Blancher, Alice Aarnink, **Nicolas Savy**, and Nagushi Takahata. Use of cumulative poisson probability distribution as an estimator of the recombination rate from population-genetic data: example of the Macaca fascicularis major histocompatibility complex. *Genes, Genomes, Genetics*, 2:123–130, 2012.

[5] Maxime Bonhomme, Antoine Blancher, Sergi Cuartero, Lounès Chikhi, and Brigitte Crouau-Roy. Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. *Mol Ecol.*, 17(4):1009–19, 2008.

[6] Olivier Cerf. Tailing of survival curves of bacterial spores. *Int. J. Appl. Bacteriol.*, 42:1–19, 1977.

[7] Djalil Chafaï and Didier Concordet. Confidence regions for the multinomial parameter with small sample size. *J. Amer. Statist. Assoc.*, 104(487):1071–1079, 2009.

[8] Louis Coroller, Ivan Leguerinel, Eric Mettler, **Nicolas Savy**, and Pierre Mafart. General model, based on two mixed weibull distributions of bacterial resistance, for describing various shapes of inactivation curves. *Applied Environmental Microbiology*, 72(10):6493–6502, 2006.

[9] Olivier Couvert, Stéphane Gaillard, **Nicolas Savy**, Pierre Mafart, and Ivan Leguérinel. Survival curves of heated bacterial spores: effect of environmental factors on weibull parameters. *International Journal of Food Microbiology*, 101(1):73–81, 2005.

[10] Olivier Couvert, Ivan Leguerinel, and Pierre Mafart. Modelling the overall effect of pH on the apparent heat resistance of Bacillus Cereus spores. *Int J Food Microbiol.*, 49(1-2):57–62, 1999.

[11] Dana C. Crawford, Tushar Bhangale, Na Li, Garrett Hellenthal, and Mark J. Rieder. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.*, 36:700–706, 2004.

[12] Stéphane Gaillard, Ivan Leguérinel, **Nicolas Savy**, and Pierre Mafart. Quantifying the combined effects of the heating time, the temperature and the recovery medium ph on the regrowth lag time of bacillus cereus spores after a heat treatment. *International Journal of Food Microbiology*, 105(1):53–58, 2005.

[13] Jean R. Lobry, Laurent Rosso, and Jean-Pierre Flandrois. A fortran subroutine for the determination of parameter confidence limits in non-linear models. *Binary*, 3:86–93, 1991.

[14] Pierre Mafart, Olivier Couvert, Stéphane Gaillard, and Ivan Leguerinel. On calculating sterility in thermal preservation methods: application of the Weibull frequency distribution model. *Int J Food Microbiol.*, 72(1-2):107–13, 2002.

[15] Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data.. *American Journal of Human Genetics*, 73:1162–1169, 2003.

[16] Richard C. Whiting, Solomon Sackitey, S. Calderone, K. Morely, and J.G. Phillips. Model for the survival of staphylococcus aureus in nongrowth environments. *Int J Food Microbiol*, 31(1-3):231–43, 1996.

Part C-

Conclusions and Perspectives

Chapter 9

Conclusions and Perspectives

In the same vein of what I have presented in this manuscript, the questions I would like to deal with in future are on both fundamental probability and applied statistics for biology and medical research. In contrast with the works I expected to do in theoretical probabilities, applied statistics for medical research or for biology are collaborative works. That is important because it allows to construct interdisciplinary projects. In what follows, the questions I am interested in are described and the main lines of projects in which I am invested are written.

I On stochastic calculus

As pointed out in Chapter 1, fractional Brownian motion or more generally filtered Brownian motion are processes that do not satisfy the two nice properties of the stochastic calculus: martingale property and strong Markov property. This raises two questions of interest but pretty hard to handle.

- First, what are the stopping times T satisfying

$$\mathbb{E}\left[B_T^K\right] = 0 \ \ ? \tag{9.1}$$

  The set of such stopping times is not empty. The deterministic times satisfy (9.1). Further, it is not the set of all bounded stopping times since it is a characterisation of martingales.

- Second, a nice property of Brownian motion is the existence of a local time process denoted $L^{\frac{1}{2}}$. A theorem of interest is the following one:

  Theorem 9.1 ([10]) *Let m a measure of Radon on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and E denotes its support. For $y \in E$ and $\omega \in \Omega$ fixed, we define:*

$$\begin{aligned} A_{y,\omega}(t) &= \frac{1}{2}\int_E L^{\frac{1}{2}}(t, x-y)(\omega)\, m(dx) \qquad \text{for any } t > 0, \\ A_{y,\omega}(0) &= 0. \end{aligned}$$

  Consider, for any $t \geq 0$, the stopping times $\tau_t = \inf\{s \geq 0 \,:\, A_s > t\}$ and the process $Y_t = B_{\tau_t} + y$.

  1. *If E is an interval then $(Y, (\mathcal{F}_{\tau_t})_{t\geq 0}, \mathbb{P}_y, y \in E)$ is a strong Markov process continuous on E and such that $Y_0 = y$.*
  2. *If E is a countable set compounded of isolated points then $(Y, (\mathcal{F}_{\tau_t})_{t\geq 0}, \mathbb{P}_y, y \in E)$ is a birth and death process.*

  In the fractional Brownian setting there are two different definitions of local time. One is by means of Gaussian property [2] and the other one by means of Tanaka's formula [3]. But, private of the strong Markov property, we are unable to prove an analogue of Theorem 9.1 and to describe the behaviour of a fractional Brownian motion in local time scale.

These questions have been stated in my thesis without any results. I let down this question during last years. In 2009, I have met Yimin Xiao (Michigan State University) and discussed on the first point. He suggested me to begin by stating the problem for Lévy fractional Brownian motion whose increments are stationary. We plan to work together on this question but time ran out. It is still an exiting work for

future.

Concerning stochastic integration with respect to Lévy processes described in Section III of Chapter 1, I have pointed out that the things go pretty well because we deal with Lévy process whose pure jumps component has for compensator $\nu(\,\mathrm{d}s,\,\mathrm{d}z) = \,\mathrm{d}s\,\eta(\,\mathrm{d}z)$. The operators $\mathcal{K}^{*,B}$ and $\mathcal{K}^{*,J}$ associated to the Brownian and pure jumps components are exactly the same. It is in fact possible to extend those results to processes defined as the sum of a Brownian motion and a pure jumps process with compensator $\nu(\,\mathrm{d}s,\,\mathrm{d}z) = \lambda(s)\,\mathrm{d}s\,\eta(\,\mathrm{d}z)$ exactly like in Section II of Chapter 1. The operator $\mathcal{K}^{*,B}$ and $\mathcal{K}^{*,J}$ are not the same but the machinery still goes on considering the operator which for any $f \in \mathcal{H}$ associate the function:

$$\mathcal{K}^{*,L}(f)(s,z) = \mathcal{K}^{*,B}(f)(s)\,\mathbb{I}_{\{z=0\}} + z\,\mathcal{K}^{*,J}(f)(s)\,\mathbb{I}_{\{z\neq 0\}}, \quad (s,z) \in [0,T] \times \mathbb{R}.$$

II On survival data analysis

STAFAV's project `http://www.math.u-psud.fr/~stafav/` aims to develop statistics in Sub-Saharan Francophone Africa. It implements doctorates of statistics in joint supervision between French university and Sub-Saharan University. Fabrice Billy Webe follows that program and is preparing in Toulouse a Ph.D. in applied statistics for epidemiology of HIV. INSERM Unit 1027, has precisely a question of interest in that topics. Cohort NADIS (New Aids Data Information System) deals with adult sero-positive patients initiated an HAART (Highly Active Antiretroviral Therapy) between January 2000 and June 2008. Cohort NADIS aims to analyse the incidence of stopping the HAART and its prognostic factors. At first glance it is a question of survival data analysis since the outcome is a time to events but in this case, there are four events of interest (intolerance, treatment failure, therapeutic simplification and finally "other causes"). Modelling competing risks data with covariables is a well-known question which has been investigated by at least two ways:

- an approach favouring the cause-specific risk function proposed by Cox [4],
- an approach favouring the cumulative incidence of the event of interest proposed by Fine & Gray [5].

That leads epidemiologists to a dilemma as to the choice of one method over the other.

Fabrice Billy Webe has just started in November 2013 a Ph.D. I supervised with Jean-Yves Dauxois (INSA Toulouse). The aim is to clarify the interpretation of these tools and to precise in what situation it is more convenient to use the first one rather than the second one. For this, a very first question to investigate or at least to clarify is how to generate data in competing risk framework which do not favour a method rather than another.

III On patients' recruitment modelling

First, the models proposed in Chapter 5 are obviously of interest for pharma companies and for institutional trials. Its practical use has to be popularised. It is one of the major objective of future works on this topics. Second, these models can be widely improved in order to cover a wider panorama of trials especially by including cohort construction whose specificity is the large duration of the follow up period. Third, these models can be used to compare clinical trials recruitment. These questions are embedded in a project called SMPR (Stochastic Modelling of Patients' Recruitment). I am the principal investigator of this project which involves statisticians - epidemiologists - specialists of clinical trials designing and clinicians. This project benefits of a grant from IRESP (Public Health Research Institute) for the period 2013-2015. This grant allows us to fund the Ph.D. of Nathan Minois (sept. 2013 - Sept. 2016) that I supervise with the help of Sandrine Andrieu for epidemiological and methodological aspects.

III.1 Recruitment models for survival data

In the setting of survival data analysis one usually uses the expected number of events to reach the necessary sample size. In this framework it is thus more interesting to deal with the number of events rather then the number of patients to be recruited. By the use of a Poisson-gamma model for recruitment and exponential models for survival data, we are able to reach this quantity. Anisimov [1] has already

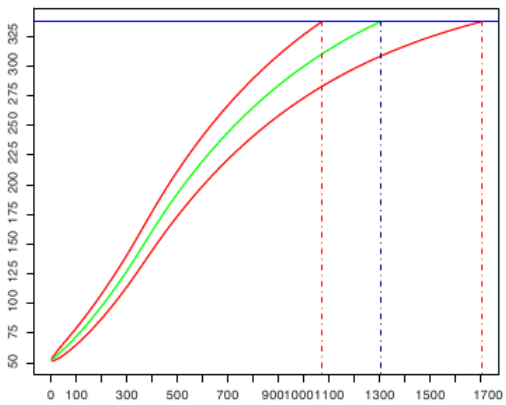

Figure 9.1: The recruitment stop at day 365. The expected number of events is the central curve and the others are the 90%-confidence intervals. We aim to reach 337 events, the expected duration is thus 1304 days and the 90%-confidence interval (1072,1705).

used this machinery in an oncology setting and the model provided very good predictions.

Our aim is to develop this model in order to give a better comprehension of the drop-out process. For this purpose, the recruited patients will be plugged in a continuous time Markov chain where the states correspond to different stages of disease progression of the patient during the trial (On risk, Drop-out, Recurrence, Death). The model will allow to provide an estimation of the number of different events in time at the end of the trial and to predict the time needed to observe a given number of events after a fixed duration of recruitment (in the spirit of Figure 9.1). Estimation is to be understood from data collected at an interim time.

III.2 Secondary recruitment models

The second objective is to develop tests in order to explain the accrual rate in a specific study or to compare the accrual rates between different studies enrolling the same type of population or between different centres of a same trial. In fact, it is possible to fit a regression model on the parameters directing the recruitment process introducing covariates (for instance size of the city where the centre is based, prevalence of the disease on study). The rate of the underlying Poisson process (or its expectation for a Cox process) can be expressed using a linear (for instance) regression model:

$$\lambda = \lambda_0 + \gamma^t X + \epsilon \tag{9.2}$$

where X is the vector of covariates, γ is the associated vector of coefficients and ϵ is a random residual value assumed to be normally distributed. This model allows to set comparison tests. That aspect is of interest in public health because it allows putting in light factors that influence the dynamics of recruitment. The very first application we have in mind is to compare the type of centre ("CHU" vs "CHG") and in each centre, the number of ongoing trials, the number of active investigators, and their participation in others trials simultaneously for the same conditions or not. Theses tests will be used to compare IFM 2005-02 and IFM 2009-02 clinical trial.

IV On Clinical Trials Simulation

The design of a clinical trials is an hard step of medical research. Indeed, the huge variability of phenomenon such as side-effect, safety, tolerability, effectiveness, recruitment... makes difficult the decision process. Approximatively 90% of drugs failed during clinical development. This is a major problem for scientific, ethical and economical reasons. Once we see the efficiency of patients' recruitment models, it is natural to go further with simulations and to model the whole clinical trial (including dose/responds, drop-out, side-effect and so on). Good Practices for Clinical Trials Simulation have been discussed in Holford *et al* [7]. They shed light on three main principles quoted below:

- ***Clarity**: The report of the simulation should be understandable in terms of scope and conclusions by intended users such as those responsible for committing resources to a clinical trial.*

- ***Completeness**: The assumptions, methods and critical results should be described in sufficient detail to be reproduced by an independent team.*

- ***Parsimony**: The complexity of the models and simulation procedures should be no more than necessary to meet the objectives of the simulation project. Program codes sufficient to generate models, simulate trials and perform replication and simulation project level analyses should be retained but there is no need to store simulated trial and analysis results which can be reproduced from these codes.*

The evolution of those techniques are recounted in two "state of the art" papers: one in 2000 [8] and one in 2010 [9]. By means of these papers, one notices that, except in huge pharma companies, the use of such techniques was, to date, not so popular. However, under the pulse of regulatory agencies, tools for CTS are developed. "*CTS is a centrepiece of the critical path initiative and part of the future of drug development*", (FDA, 2007). In July 2013, regulatory agencies in the U.S. and Europe have endorsed a quantitative simulation tool that allows researchers to model clinical trials in mild to moderate Alzheimer's disease.

In order to start those investigations in a good way, essentially with a precise idea of what has been done and how, the Cancéropole (Institution devoted to research on cancer) supports a workshop in Toulouse in April 2015 on this topic. I am in charge of its organisation. Moreover, in the framework of a medico-economic project with Toulouse's Hospital, we aim to use simulation techniques to estimate the necessary sample size of a clinical trials we are designing.

V On mediation analysis

By means of the project IBISS (Biological Incorporation of Social Health Inequalities), granted by the National Research Agency (ANR), some funds are devoted to continue the investigations on mediation analysis. I am the investigator of this research work-package of IBISS's project which aims

- first, to develop a measure of the mediated effects,

- second, to propose a methodology on how to manage missing data in databases.

V.1 Measure of mediated effects

As seen in Chapter 7, in Structural Equation Models, the better way to measure the mediated effect is to calculate the product of the effect along the mediation path. As this approach does not define the mediation in an intrinsic way, we try to define a new measurement of the mediation by means of a coefficient widely inspired of the determination coefficient (R^2) for the linear models. The objective is to find a decomposition of the variance explained by a variable as the sum of the direct part and of the mediated one.

V.2 Missing data

This question is of paramount interest in epidemiology and the whole INSERM unit 1027 wonders how to deal with such data. On the pulse of Benoît Lepage and me, a research group (compound of statisticians and epidemiologists) has been created in November 2013. The aim is, first to clarify the ideas on the use of implementation techniques for missing data, second to browse the methods that not necessitate implementation and third, to lead research on topics of interest, especially missing data in longitudinal setting and missing non at random data.

The arrival of Chloé Dimeglio, on a post-doctoral position on biostatistics at INSERM Unit 1027, will certainly boost these works.

VI On Prediction in Social Health

In January 2014, the Project IMPACTISS "Social and economic determinants of health: feasibility and political as well as social acceptability of health impact assessment and benefits for primary prevention" has been proposed in response to the call for tenders INCA - IRESP (National Cancer Institute - Public Health Research Institute). I am in charge of a work-package.

This project deals with the Health Impact Assessment (HIA). According to its most common definition, the HIA is a set of procedures, methods and tools to assess the potential positive and negative effects of a project, program or policy on health. Further, it gives information on the social distribution of these effects. The objective is to make recommendations to policy-makers, in order to maximize the positive impacts and reduce negative effects. There are two questions of paramount interest we aim to investigate in that project: first, how to measure the impact of a health policy and second, what is the effect of an action on social determinants of health. Both questions are particularly difficult to deal with. Moreover, not so much attention has been given on those problems. An idea is to use mathematical modelling [6]. The literature offers various examples of HIA tools especially in the field of Health-Environment. We focus our attention on DYNAMO-HIA (Dynamic Modelling for Health Impact Assessment) software. It is based on Markov modelling and used partial micro-simulation to calculate disease and mortality from historical data. Our goal is, first, to improve that software or at least to let it work in our social epidemiology's setting. Second, one makes use of g-computation (see Section II.2 of Chapter 7) to simulate scenarios which will be compared to those of DYNAMO-HIA. Finally we aim to validate those results on real life dataset.

In case this project is accepted, the grant allows us to recruit a Ph.D. I will supervise with Thierry Lang.

Bibliography

[1] Vladimir V. Anisimov. Predictive event modelling in multicentre clinical trials with waiting time to response. *Pharmaceutical Statistics*, 10(6):517–522, 2011.

[2] Simeon M. Berman. Local times and sample function properties of stationary Gaussian processes. *Trans. Amer. Math. Soc.*, 137:277–299, 1969.

[3] Laure Coutin, David Nualart, and Ciprian A. Tudor. Tanaka formula for the fractional Brownian motion. *Stochastic Process. Appl.*, 94(2):301–315, 2001.

[4] David R. Cox. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972.

[5] Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, Vol.94(446):496–509, 2009.

[6] Penelope Hawe and Louise Potvin. What is population health intervention research? *Can J Public Health*, 100(1):Suppl I8–14, 2009.

[7] Nicholas H.G. Holford, M. Hale, H.C. Ko, J.L. Steimer, L.B. Sheiner, and C.C. Peck, editors. *Simulation in drug development: Good practices*, volume Proceedings of "Modeling and Simulation of Clinical Trials: Best Practices Workshop", 1999.

[8] Nicholas H.G. Holford, H.C. Kimko, J.P. Monteleone, and Carl C. Peck. Simulation of clinical trials. *Annu Rev Pharmacol Toxicol.*, 40:209–34, 2000.

[9] Nicholas H.G. Holford, S.C. Ma, and B.A. Ploeger. Clinical trial simulation: a review. *Clin Pharmacol Ther.*, 88(2):166–82, 2010.

[10] Eric Thibault. *Files d'attentes, Approximations diffusions, Caractéristiques transitoires.* PhD thesis, University of Rennes 1, 1998.

Curriculum Vitae

Personal Details

	Born November 21, 1973, Le Mans (France).
Current Position:	Assistant professor. University Paul Sabatier - Toulouse 3. IUT A, department GEA. * Toulouse Mathematics Institute, Team Statistic and Probabilities.
Research Adress:	Institut de Mathématiques de Toulouse, 118, route de Narbonne - 31062 Toulouse Cedex 9, France. Tel. : +33(0)5 61 55 62 20 Fax : +33(0)5 61 55 60 89. mail : `nicolas.savy@math.univ-toulouse.fr`
Teaching Adress:	IUT de Toulouse 3A - Département GEA-Rangueil, 133, avenue de Rangueil - 31077 Toulouse Cedex 4, France. Tel. : +33(0)5 61 25 84 06 Fax : +33(0)5 61 25 88 01. mail : `nicolas.savy@iut-tlse3.fr`
Web Page:	`http://math.univ-toulouse.fr/~savy`

Education

1998 – 2003:	**PhD in Fundamental Mathematics and Applications,** University of Rennes 1, *"Mouvement Brownien fractionnaire, applications aux télécommunications. Calcul stochastique relativement à des processus fractionnaires".* Examination board: Prof. Jean Mémin, Prof. Laurent Decreusefond, Prof. Jean-Bernard Gravereaux (Advisor), Prof. Ying Hu, Prof. David Nualart (Reviewer), Prof. Nicolas Privault (Reviewer).
1996 – 1997:	**"Agrégation" and "CAPES" in Mathematics** (state teaching qualifications), University of Rennes 1.
1995 – 1996:	**Master's degree in Fundamental Mathematics and Applications**, University of Rennes 1.

Professional Experience

- **Since November 2006: Assistant Professor**
  University Paul Sabatier - Toulouse 3 - IUT Department GEA.
  Toulouse Mathematics Institute, Team Statistic and Probabilities.
- **September 1999-October 2006: Teacher**
  "Université de Bretagne Occidentale" - IUT of Quimper - Department "Biological Engineering".
  Mathematics Research Institute of Rennes, Team Stochastic Processes.
- **September 1998-August 1999: Trainee teacher**
  "Middle School" Bourgchevreuil at Cesson Sévigné (35).

*IUT : Technological University Institution - GEA : Business and Administration Management

Teaching Activities

University Paul Sabatier - Toulouse 3

- Since 2006: IUT GEA - Toulouse
  - **1st year Bachelor's degree:** Mathematics applied to management
  - **1st year Bachelor's degree:** Probabilities
  - **2nd year Bachelor's degree:** Preparation to competitive examination
- Since 2006: other components
  - **Research Master's degree in Clinical Epidemiology:** Survival data analysis
  - **Research Master's degree in Applied Mathematics:** Stochastic calculus
  - **3rd year Bachelor's degree:** Inferential statistics and linear models

University "Bretagne Occidentale"

- 1999 - 2006: IUT Biological Engineering - Quimper
  - **1st year Bachelor's degree:** Statistics for biologists
  - **1st year Bachelor's degree:** Informatics
  - **1st year Bachelor's degree:** Mathematics
  - **2nd year Bachelor's degree:** Industrial statistics
  - **3rd year Bachelor's degree:** Design of experiments
- 2000 - 2006: FITI2A (Engineering School)
  - **3rd year Bachelor's degree:** Statistics

Guest lecturer

- January 2007: Havana University (Cuba)
  - **Research Master's degree:** Stochastic Calculus and Application to Finance.
- June 2012: Havana University (Cuba)
  - **Research Master's degree:** Methodology for Therapeutic Trials.

Research Activities

Articles in Probability

Submitted for publication

[P8] J. Vives, **N. Savy**, *Anticipative Integrals with respect to a filtered Lévy Process and Lévy-Itô decomposition*, 2014. Submitted to Journal of Theoretical Probability.

Published

[P7] B. Bercu, F. Proïa, **N. Savy**, *On Ornstein-Uhlenbeck driven by Ornstein-Uhlenbeck processes,* Statistics and Probability Letters, 85 pp. 36-44, 2014.

[P6] B. Bercu, L. Coutin, **N. Savy**, *Sharp large deviation for the non-stationary Ornstein-Uhlenbeck process*, Stochastic Processes and their Applications, 122(10) pp. 3393-3424, 2012.

[P5] A. Alvarez, F. Panloup, M. Pontier, **N. Savy**, *Estimation of the instantaneous volatility and detection of volatility jumps,* Statistical Inference for Stochastic Processes, 15 pp. 27-59, 2012.

[P4] L. Decreusefond, A. Joulin, **N. Savy**, *Rubinstein distances on configuration spaces,* Communications on Stochastic Analysis, 4(3) pp. 377–399, 2010.

[P3] B. Bercu, L. Coutin, **N. Savy**, *Sharp large deviations for the fractional Ornstein-Uhlenbeck process,* Theory of Probability and its Applications, 55(4) pp. 575-610, 2011.

[P2] L. Decreusefond, **N. Savy**, *Anticipative calculus with respect to filtered Poisson processes,* Annales de l'Institut Henri Poincaré - Probabilités et Statistiques, 42(3) pp. 343–372, 2006.

[P1] L. Decreusefond, **N. Savy**, *Filtered Brownian motion as weak limit of Filtered Poisson processes,* Bernoulli, 11(2) pp. 283-292, 2005.

Articles in Statistics

In preparation, submitted for publication or in revision

[S8] V. Anisimov, G. Mijoule, **N. Savy**, *Statistical modelling of recruitment in multicentre clinical trials with patients' drop-out*, 2013. In preparation for Statistics in medicine.

[S7] V. Garès, S. Andrieu, J.F. Dupuy, **N. Savy**, *Comparison of constant piecewise weighted test and Fleming Harrington's test - Application in clinical trials,* Electronic Journal of Statistics, 2013. In revision.

[S6] V. Garès, S. Andrieu, J.F. Dupuy, **N. Savy**, *An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials*, Statistics in medicine, 2013. In revision.

[S5] V. Garès, S. Andrieu, J.F. Dupuy, **N. Savy**, *Choosing the parameter of Fleming-Harrington's test in prevention randomized controlled trials*, Journal of the Royal Statistical Society: Series C , 2014. Submitted.

[S4] B. Lepage, S. Lamy, D. Dedieu, **N. Savy**, T. Lang, *Estimating the Causal Effect of an Exposure on Change from Baseline Using Directed Acyclic Graphs and Path Analysis*, Epidemiology, 2012. In revision.

Published or forthcoming

[S3] D. Dedieu, C. Delpierre, S. Gadat, B. Lepage, T. Lang, **N. Savy** *Discrete Time Mixed Hidden Markov Models with State Memory for Longitudinal Data - Application to the epidemiology of cancer in the NCDS 1958 study*, "Journal de la SFDS", 155(1) pp. 73-98, 2014.

[S2] B. Lepage, D. Dedieu, **N. Savy**, T. Lang, *Estimation of a controlled direct effect when an effect of the exposure confounds the mediator-outcome relationship: a comparison of different methods*, Statistical Methods for Medical Research, 2012. Forthcoming.

[S1] G. Mijoule, S. Savy, **N. Savy**, *A methodological approach for a dynamical patients recruitment in clinical trials,* Statistics in Medicine, 31(16) pp. 1655-1674, 2012.

Articles in Statistics for Medical or Biological research

Published or forthcoming

[MB5] A. Aarnink, E. Mee, **N. Savy**, N. Caugy, N. Rose, A. Blancher, *Demonstration of the deleterious impact of foeto-maternal MHC compatibility on the success of pregnancy in a macaque model*, Immunogenetics, 66(2) pp.105-113, 2014.

[MB4] A. Blancher, A. Aarnink, **N. Savy**, N. Takahata, *Use of cumulative Poisson probability distribution as an estimator of the recombination rate from population-genetic data: example of the Macaca fascicularis MHC*, Genes, Genomes, Genetics, 2 pp. 123-130, 2012.

[MB3] L. Coroller, I. Leguérinel, E. Mettler, **N. Savy**, P. Mafart, *General Model, Based on Two Mixed Weibull Distributions of Bacterial Resistance, for Describing Various Shapes of Inactivation Curves,* Applied and Environmental Microbiology, 72(10) pp. 6493-6502, 2006.

[MB2] S. Gaillard, I. Leguérinel, **N. Savy**, P. Mafart, *Quantifying the combinated effects of the heating time, the temperature and the recovery medium pH on the regrowth lag time of Bacillus Cereus spores after a heat treatment,* International Journal of Food Microbiology, 105(1) pp. 53-58, 2005.

[MB1] O. Couvert, S. Gaillard, **N. Savy**, I. Leguérinel, P. Mafart, *Survival curves of heated bacterial spores: effect of environmental factors on Weibull parameters,* International Journal of Food Microbiology, 101(1) pp. 73-81, 2005.

Books chapters

[BC1] V. Anisimov, G. Mijoule, **N. Savy**, *Additive model for cost modelling in clinical trial*, Proceedings of "7th International Workshop on Simulation", Rimini, 2013. Forthcoming.

Conference proceedings

Published Conference proceedings

[PP5] **N. Savy**, *Stochastic modelling of recruitment in clinical trials*, Proceedings of "2nd International Conference on Biometrics and Biostatistics", Chicago, 2013, published in J. Biomet. Biostat., 4(4) pp. 36, 2013.

[PP4] B. Lepage, S. Lamy, **N. Savy**, T. Lang, *Utilisation des graphes acycliques dirigés pour le choix d'un modèle d'analyse du changement*, Proceedings of "3e Colloque thématique de l'Adelf", Toulouse, 2012, published in Revue d'épidémiologie et de Santé Publique, 61(S2) pp. S108, 2013.

[PP3] B. Lepage, D. Dedieu, V. Ehlinger, M. Kelly-Irving, **N. Savy**, T. Lang, *Analyse de la médiation par g-computation : application dans la cohorte britannique NCDS 58*, Proceedings of "5e Congrès International d'Épidémiologie Adelf-Epiter", Bruxelles, 2012, published in Revue d'épidémiologie et de Santé Publique, 60(S2) pp. S71, 2012.

[PP2] B. Lepage, D. Dedieu, **N. Savy**, T. Lang, *Comparaison de quatre méthodes d'estimation d'un effet direct et exploration de l'importance des biais potentiels : étude par simulation*, Proceedings of "EPI-CLIN", Marseilles, 2011, published in Revue d'épidémiologie et de Santé Publique, 59(S1) pp. S17, 2011.

[PP1] **N . Savy**, G. Mijoule, S. Savy, *Approche méthodologique du recrutement de patients*, Proceedings of the "6e Journées du Cancérople GSO", published in Bulletin du Cancer, 97(4) pp. S80, 2010.

Unpublished Conference proceedings

[UP7] V. Garès, J.F. Dupuy, **N. Savy**, *Un nouveau test pour l'analyse de données de prévention en recherche clinique*, Proceedings of "45e Journées de Statistique", Toulouse, 2013.

[UP6] F. Proïa, **N. Savy**, B. Bercu *Le processus de Ornstein-Uhlenbeck engendré par le processus de Ornstein-Uhlenbeck*, Proceedings of "45e Journées de Statistique", Toulouse, 2013.

[UP5] V. Garès, S. Andrieu, J.F. Dupuy, **N. Savy**, *On the use of Fleming and Harrington's test to detect late effects in clinical trials*, Proceedings of "Statistical Models and Methods for Reliability and Survival Analysis and Their Validation", Bordeaux, 2012.

[UP4] G. Mijoule, **N. Savy**, *Modélisation du recrutement de patients lors d'un essai clinique*, Proceedings of "44e Journées de Statistique", Bruxelles, 2012.

[UP3] **N. Savy**, B. Bercu, L. Coutin, *Grandes Déviations Précises pour des processus de Ornstein Uhlenbeck*, Proceedings of "44e Journées de Statistique", Bruxelles, 2012.

[UP2] V. Garès, J.F. Dupuy, **N. Savy**, *Utilisation du test de Fleming et Harrington pour la détection d'effets tardifs en recherche clinique*, Proceedings of "44e Journées de Statistique", Bruxelles, 2012.

[UP1] O. Couvert, **N. Savy**, P. Mafart, *Monte Carlo Simulation of F-value distribution*, Proceedings of "Predictive Modelling in Foods", Quimper, 2003.

Posters

[**Po4**] **N. Savy**, S. Savy, S. Andrieu, *Les essais cliniques simulés. Des avancés récentes ?,* "EpiClin 2014", Bordeaux, 2014.

[**Po3**] G. Mijoule, S. Savy, **N. Savy**, *Modélisation de la phase d'enrlement lors d'essais cliniques,* "6e Journées du Cancérople GSO", Toulouse, 2010.

[**Po2**] L. Decreusefond, J. Ledoux, **N. Savy**, *Simulation Monte Carlo de l'application de transport optimal,* "Journées STAR", Rennes, 2005.

[**Po1**] O. Couvert, **N. Savy**, P. Mafart, *Monte Carlo Simulation of F-value distribution,* 4th International Conference "Predictive Modelling in Foods", Quimper, 2003.

Books

[**B1**] **N. Savy**, *Probabilités et Statistiques : pour modéliser et décider,* Ellipses Editions, 2006.

Oral Presentations

June 10, 2013 *Stochastic modelling of recruitment in clinical trials*, "2nd International Conference on Biometrics and Biostatistics", Chicago (USA).

May 22, 2013 *Additive model for cost modelling in clinical trial*, "7th International Workshop on Simulation", Rimini (Italie).

June 22, 2012 *Models for patients' recruitment in clinical trials*, Seminar of Havana's University, Havana (Cuba).

June 21, 2012 *On the use of Fleming-Harrington's test to detect late effects in Clinical Trials*, Seminar of Havana's University, Havana (Cuba).

May 24, 2012 *Grandes déviations précises pour des processus de Ornstein Uhlenbeck*, "44e Journées de Statistique", Bruxelles (Belgium).

April 24, 2012 *On the use of Fleming-Harrington's test to detect late effects in Clinical Trials*, Seminar of Queen Mary University, London (UK).

June 10, 2011 *Sharp Large Deviation for some Ornstein Uhlenbeck processes*, Seminar SPAAF, Lyon.

May 26, 2011 *Sharp Large Deviation for some Ornstein Uhlenbeck processes*, Seminar of Barcelona University, Barcelona (Spain).

Sept. 28, 2010 *Modélisation de la phase d'enrlement lors d'essais cliniques*, "Journée Biostatistiques et Statistiques Médicales", Toulouse.

Sept. 01, 2010 *Principe de Grandes Déviations précises pour des processus de Ornstein Uhlenbeck fractionnaires*, "Journées M.A.S ", Bordeaux.

Sept. 04, 2009 *Utilisation des techniques de files d'attente en recherche clinique*, Seminar of Epidemiology, Toulouse.

May 13, 2009 *A propos de la méthodologie statistique*, "Journée du groupe Méthodologie du Cancérpole GSO", Toulouse.

Nov. 22, 2006 *Théorème de convergence d'une suite de processus de Poisson filtrés vers son homologue Brownien*, Seminar of Statistics, Toulouse.

Nov. 18, 2004 *Théorème de convergence d'une suite de processus de Poisson filtrés vers son homologue Brownien*, Seminar of probabilities, Brest.

Sept. 07, 2004 *Intégrale anticipative pour des processus de Poisson marqués et extension aux processus de Poisson-Volterra*, "Journées M.A.S.", Nancy.

June 03, 2004 *Quelques résultats sur les processus de Poisson-Volterra*, "Journées d'Analyse Stochastique", Paris.

June 17, 2002 *Intégrale stochastique pour des processus de Poisson filtrés*, Seminar of probabilities, Rennes.

May 2, 2002 *Intégrale stochastique pour des processus de Poisson filtrés*, "Congrès des Jeunes Probabilistes", Aussois (France).

Jan. 10, 2000 *Comportement asymptotique de la capacité de stockage d'une file à entrée Brownienne fractionnaire*, Seminar of probabilities, Rennes.

Organization of conferences

- **Member of the Scientific committee** for the workshop *"Dynamic predictions for repeated markers and repeated events: models and validation in cancer""*, Bordeaux, October 10 and 11, 2013.
- **President of the Organization committee** for the workshop *"Analyse de données longitudinales de cancer, Modèles de Markov cachés"*, Toulouse Mathematics Institute, Toulouse, October 18 and 19, 2012.
- **Member of the Organization committee** for the international conference *"Random differential equations and Gaussian fields"*, Château de Mons, Caussens (32), June 15 to 19, 2009.
- **Member of the Organization committee** for the international workshop *"Selfdecomposability and Fractional Processes"*, Toulouse Mathematics Institute, Toulouse, November 20 and 21, 2008.
- **Member of the Organization committee** for the workshop *"Journée en l'honneur de Monique Pontier"*, Toulouse Mathematics Institute, Toulouse, June 11, 2008.
- **Member of the Organization committee** for the *"International Conference on Probabilities and Statistics"*, Toulouse Mathematics Institute, Toulouse, June 14 and 15, 2007.

Ph-D Supervisions

Defended Ph-D theses

- **Guillaume Mijoule**

  "Modélisation du processus d'inclusion de patients dans un essai clinique multicentrique".

  - Co-Advisor: Pr Laure Coutin (IMT)
  - Started in September 2009, examined June 3rd, 2013
  - Current position: temporary assistant professor at University Paris X.

- **Benoit Lepage**

  "Prise en compte des hypothèses de causalité dans l'analyse d'une évolution et l'analyse de la médiation".

  - Co-Advisor: Pr Thierry Lang (INSERM Unit 1027)
  - Started in January 2010, examined June 21, 2013
  - Current position: assistant professor at University Toulouse 3.

- **Valérie Garès**

  "Améliorer la performance des analyses de survie dans le cadre des essais de prévention et application la maladie d'Alzheimer".

  - Co-Advisor: Pr Sandrine Andrieu (INSERM Unit 1027)
  - Started in September 2010, examined April 15th, 2014

Ph-D theses in progress

- **Nathan Minois**

  "Modèles de Markov Cachés appliqués à l'étude des maladies chroniques".

  - Co-Advisor: Pr Sandrine Andrieu (INSERM Unit 1027)
  - Started in October 2013.

- **Fabrice Billy Webe**

  "Risques compétitifs".

  - Co-Advisor: Pr Jean-Yves Dauxois (IMT)
  - Started in November 2013.

Ph-D theses steering committees member

Defended Ph-D theses

- **Stéphane Gaillard**

  "Modélisation de la thermorésistance, de la viabilité et du comportement à la recroissance de Bacillus cereus, en fonction de la température, du pH et de l'activité aqueuse".

  – Advisor: Pr Pierre Mafart (University of "Bretagne Occidentale")

  – Started in October 1999, examined December 19th, 2003 (member of the examination board).

- **Olivier Couvert**

  "Prise en compte de l'influence du pH dans l'optimisation des traitements thermiques".

  – Advisor: Pr Pierre Mafart (University of "Bretagne Occidentale")

  – Started in October 1998, examined April 21th, 2002

  – Current position: assistant professor at University of "Bretagne Occidentale".

Ph-D thesis in progress

- **Caroline Delarue**

  "Analyse statistique de l'exposition aux psychotropes pendant la grossesse et survenue de malformations congénitales et/ou de pathologies infantile".

  – Advisors: Pr Claire Damase (INSERM Unit 1027) and Cécile Chouquet (IMT)

  – Started in September 2012.

Masters supervision

- **Guillaume Mijoule**

  – *"Le processus de Poisson Markov-modulé"*,

  – Research Master's degree in Applied Mathematics - March to June 2009

- **Valerie Garès**

  – *"Approche processus stochastiques de la survie"*,

  – Research Master's degree in Applied Mathematics - March to June 2010.

- **Anne-Claire Brunet**

  – *"Utilisation de techniques de classification pour l'amélioration du budget du CHU de Toulouse"*,

  – Professional Master's degree in Statistics and Economy - April to September 2011.

- **Isabelle Bouissière**

  – *"Utilisation de techniques GLM en Hématologie"*,

  – Professional Master's degree "IMAT" - April to September 2012.

- **Nathan Minois**

  – *"Modélisation de la phase d'inclusion de patients lors d'essais cliniques"*,

  – Professional Master's degree "IMAT" - April to September 2013.

Industrial Partnerships

ADRIA	Advisor on implementation of "AFNOR" normes for validation of "rapid" methods in microbiology.
AGRAUXINE	Advisor on implementation of experimental designs to optimize culture media.
SAUPIQUET	Application of Monte-Carlo technique to reduce sterilisation schedules in the framework of Olivier Couvert's PhD thesis.
DANONE	Modelling of bacteria growth and destruction in the framework of Stéphane Gaillard's PhD thesis.

Administrative Activities

Research

- From 2009 to 2013, elected member of mathematics scientific commission at University Toulouse 3.
- Since 2009, member of recruitment commission for applied mathematics assistant professors at University of Toulouse 3.
- In 2008, elected member of recruitment commission for applied and fundamental mathematics assistant professors at Toulouse 3A's IUT.
- **Referee** for the journals Computational Statistics and Data Analysis, Applied Mathematical Modelling, ESAIM: Probability and Statistics, Journal of Statistical Planning and Inference, Journal of Cancer Therapy, Stochastics, Statistical Inference for Stochastic Processes, Applied Mathematical Modelling, International Journal of Probability and Statistics, Scientific Research and Essays.
- **Research groups**
  - **2010 –:** "Stats et Santé": Executive committee member.
  - **2010 –:** "Mathématiques et Entreprises": member.
  - **1999 – 2002:** Member of INTAS's project 99016 (Universities of Barcelona, Berlin, Helsinki, Moscou, Kiev, Donets and Rennes).
- **Projects and grants**
  - **2013 – 2016:** Project *"Incorporation Biologique et Inégalités Sociales de Santé"* granted by (work-package leader)
  - **2013 – 2016:** Project *"Improved predictability of sub-chronic GMO toxicity by identification of early biomarkers of toxicity"* granted by French Ministry of Ecology
  - **2013 – 2015:** Project *"Statistic Methods for Patients Recruitment and Clinical Trials Design"* (Principal Investigator) granted by IRESP (Public Health Research Institute)
  - **2011 – 2013:** Project *"Essais de prévention dans la démence de type Alzheimer : Améliorer la performance des outils statistiques"* (Principal Investigator) granted by France-Alzheimer
  - **2010 – 2012:** Project *"Méthodes d'analyses appliqués à l'épidémiologie biographique ; exploration des chemins de causalité entre environnement social, habitudes alimentaires et cancer"* (work-package leader) granted by National Cancer Institute - Public Health Research Institute
  - **2004 – 2006:** Project *"TRansport OptiMal et Applications aux Techniques de l'Information et de la Communication"* granted by CNRS (National Center for Scientific Research)

Teaching

- Since 2012, GEA department commission elected member.
- Since 2010, **joint-coordinator for Research Master's** "Clinical Epidemiology" at University of Toulouse 3. Responsible for Biostatistics and Clinical trials teaching units.
- From June 2002 to Juin 2005, Teaching Director for Department "Biological Engineering" of Quimper's IUT.

Summary

My research activity is mainly devoted to Stochastic processes and to Applied Statistics for Biology and Medical Research. That manuscript naturally splits in two parts.
The first part evokes my investigations in the field of probability and statistics of stochastic processes. It contains four chapters. The first one gives the main lines of the construction of stochastic integrals with respect to filtered processes (filtered Poisson processes and filtered Lévy processes). Those integrals are anticipative and defined by means of Malliavin Calculus. In the second chapter we show briefly how a sequence of filtered Poisson processes converges weakly to a Brownian Volterra process. The proof of that result makes use of radonification techniques. The third chapter clarifies the link between Transportation Inequality and Malliavin Calculus in the space of configurations. This first part ends by a fourth chapter devoted to statistics of processes. Two problems are considered: first the construction and the proofs of mains properties (essentially convergence and asymptotic normality) of an estimator of instantaneous volatility in a diffusion process, second, the proofs of Large and Sharp large deviations principles for functionals associated to various Ornstein Uhlenbeck's type processes.
The second part presents works on applied statistics for Biology and Medical research. It contains four chapters. The first chapter presents empirical Bayesian models which aim to capture the behaviour of the dynamic of patients' recruitment in clinical trials. Those models are of paramount interest to estimate the end of a clinical trial from on-going data. The second chapter presents research on survival data analysis for prevention clinical trials. It contains mainly the study of the so-called Fleming-Harrington's test and its comparison with another weighted logrank's test of interest. Finally a composite test is introduced and studied. The third chapter summarizes a series of papers on epidemiology especially on mediation analysis. Two aspects are evoked, by means of Mixed Hidden Markov Models and by means of causality analysis tools. Finally a fourth chapter allows me to include various results obtained in collaboration with biologists during my stay in Quimper on the topic of Predictive Microbiology and recently in genetic of populations.
Manuscript ends with a description of questions I would like to carry out in future.

Résumé

Mon activité de recherche est principalement consacrée aux processus stochastiques et aux statistiques appliquées à la Biologie et à la Recherche Médicale. Ce manuscrit se divise naturellement en deux parties.
La première partie évoque mes recherches dans le domaine des probabilités et statistiques des processus stochastiques. Il contient quatre chapitres. Le premier donne les grandes lignes de la construction d'intégrales stochastiques par rapport à des processus filtrés (processus de Poisson filtrés et processus de Lévy filtrés). Ces intégrales sont anticipatives et sont définies au moyen du calcul de Malliavin. Dans le deuxième chapitre, nous montrons brièvement comment une suite de processus de Poisson filtrés converge faiblement vers un processus de Volterra brownien. La preuve de ce résultat utilise des techniques de radonification. Le troisième chapitre vise à clarifier le lien entre les inégalités de transport et le calcul de Malliavin sur l'espace des configurations. Cette première partie se termine par un quatrième chapitre consacré à la statistique des processus. Deux problèmes sont considérés : d'abord la construction et les preuves des principales propriétés (essentiellement la convergence forte et la normalité asymptotique) d'un estimateur de la volatilité instantanée dans un processus de diffusion, d'autre part, les preuves de principes de grandes déviation et de grandes déviations précises pour des fonctionnelles associées à divers processus de type Ornstein-Uhlenbeck.
La deuxième partie présente divers travaux de statistiques appliquées à la biologie et de la recherche médicale. Elle contient quatre chapitres. Le premier chapitre présente des modèles Bayésiens empiriques qui visent à capter le comportement de la dynamique du recrutement des patients dans les essais cliniques. Ces modèles sont d'un intérêt primordial pour estimer la fin d'un essai clinique à partir des données en cours. Le deuxième chapitre présente les recherches menées sur l'analyse de données de survie pour les essais cliniques de prévention. Il contient principalement l'étude du test de Fleming - Harrington et sa comparaison avec un autre test de logrank pondéré intéressant. Enfin un test composite est introduit et étudié. Le troisième chapitre résume une série d'articles de statistiques appliquées à l'épidémiologie en particulier sur l'analyse de la médiation. Deux aspects sont abordés, par le biais de modèles de Markov cachés mixtes et au moyen d'outils d'analyse de causalité. Enfin, un quatrième chapitre me permet d'inclure divers résultats obtenus en collaboration avec des biologistes pendant mon séjour à Quimper sur le thème de la microbiologie prévisionnelle et récemment en génétique des populations.
Le manuscrit se termine par une description de questions que j'aimerais traiter à l'avenir.