On martingales, causality, identifiability and model selection

Ph. D. Thesis

Alexander Sokol
Department of Mathematical Sciences
University of Copenhagen
November 2013

This thesis has been submitted to the Ph. D. school of The Faculty of Science, University of Copenhagen.

Alexander Sokol
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen

Email: alexander@math.ku.dk

ISBN: 978-87-7078-386-6

Thesis submission date: 29/11/2013

Thesis supervisor:	Prof. Niels Richard Hansen, University of Copenhagen
Thesis committee:	Prof. Michael Sørensen, University of Copenhagen Assoc. Prof. Jan Pedersen, University of Aarhus Prof. Håkon K. Gjessing, University of Bergen

Preface

This thesis constitutes my Ph. D. dissertation at the University of Copenhagen. The work included here was carried out between December 2010 and November 2013, under the supervision of Prof. Niels Richard Hansen. It was funded through the Programme of Excellence at the University of Copenhagen, as a part of the research project "Statistical Methods for Complex and High Dimensional Models", with Prof. Michael Sørensen as the principal investigator.

The thesis is broadly concerned with probability and statistics, with contributions to the theory of stochastic calculus, exponential martingales, causality, independent component analysis and model selection. Motivating applications are primarily to be found in biological fields, although the results obtained admittedly are more of a theoretical than practical nature.

I am personally not liable to be easily impressed with myself, and in my own opinion, the results presented here are neither particularly original, particularly insightful or particularly substantial. My argument for the raison d'être of this thesis, therefore, is that scientific progress is a slow, gradual and collaborative process, where minute progression also counts for something.

Many people have played parts in the making of this thesis. Most of all, I would like to thank my supervisor, Prof. Niels Richard Hansen, for introducing me to many of the problems discussed in this thesis such as causality and sparse estimation, for many fruitful discussions, for being an active and constructive co-author on several manuscripts and for complaining so comparatively little about my long and verbose emails.

I would also like to thank my other two co-authors, Prof. Marloes Maathuis and Benjamin Falkeborg, for our collaboration. I also owe Marloes further thanks for functioning as a very friendly and accessible contact point during my stay at the Seminar für Statistik at ETH Zürich and for discussions particularly on the topic of causality and stochastic differential equations. Further thanks go to my office-mates

at the seminar, namely Jonas Peters, Michaël Chichignoud and Jan Ernest, who made the stay extra enjoyable, and to the other Ph. D. students and the staff of the seminar in general, for their hospitality.

Closer to home, I would like to thank in particular Prof. Martin Jacobsen and Assoc. Prof. Ernst Hansen at the University of Copenhagen for being ready to discuss my various trivial questions. I also owe thanks to the many co-inhabitants of the Ph. D. ghetto at the university, in particular for the enjoyable conversations with, among others, Anders Jensen, Martin Vincent and Massimiliano Tamborrino. In general, I would like to thank everybody at the Department of Mathematical Sciences at the University of Copenhagen for the various ways in which they have contributed to this thesis by some means.

Even closer to home, I would like to thank my collective, Poco Loco, for being a warm little refuge from the sometimes cold and unforgiving world of mathematics.

Finally, I suppose that I also have to reluctantly admit that there probably are plenty of errors and misprints left in this thesis, and those are of course my own unfortunate responsibility.

Alexander Sokol
November 2013

Summary

In spite of the principal starting point for this thesis being relatively concise, it so happened that during the course of the thesis work, the subjects for our research branched out considerably. Broadly speaking, this thesis is concerned with five main topics: Exponential martingales, the general theory of processes, causality, independent component analysis (ICA) and model selection for nonlinear regression.

In Chapter 1, we give an overview of the research project and its results, and the main subjects of study are introduced. The remaining Chapters 2-10 are self-sufficient manuscripts containing the main results of the research project.

Chapters 2-4 are concerned with the martingale property of exponential martingales. Chapter 2 considers exponential martingales based on counting processes in particular, with the purpose of constructing statistical models involving nonexplosive counting processes with stochastic intensity. Of particular note is the construction of counting processes interacting with diffusions. Chapters 3 and 4 are concerned with the uniformly integrable martingale property from a more abstract point of view, the results of the former being concerned with optimal constants in Novikov-type criteria, and the results of the latter being concerned with a Novikov-type criterion applying both the optional and predictable quadratic variation.

The results in Chapter 5 and Chapter 6 consider the general theory of processes. Essentially no new results are proven here, instead simplified proofs are given of known results, in particular the existence of the dual predictable projection and the quadratic variation, and the applications of these proofs to obtain a simplified theory of stochastic integration are discussed.

Chapter 7 and Chapter 8 center on causality for stochastic differential equations (SDEs). Chapter 7 introduces a notion of causality for SDEs, and we prove that for SDEs driven by Lévy processes, in constrast to results from Pearl's intervention calculus, it holds that postintervention distributions are identifiable from the observational distribution. Chapter 8 considers causality for the particular case of

Ornstein-Uhlenbeck SDEs, where explicit calculations may be made for the postintervention distributions.

Chapter 9 concerns identifiability of the mixing matrix in ICA. It is a well-known result that identifiability of the mixing matrix depends crucially on whether the error distributions are Gaussian or not. We attempt to elucidate what happens in the case where the error distributions are close to but not exactly Gaussian.

Finally, Chapter 10 discusses degrees of freedom in nonlinear regression. Our motivating problem is that of L^1-constrained and L^1-penalized estimation in nonlinear regression. Our objective is to obtain results leading to the calculation of the degrees of freedom of an estimator in order to enable sparse model selection by optimal choice of the penalization parameter. We prove two results related to the degrees of freedom, one theoretical result for constrained estimation, and one more practically applicable for L^1-penalized estimation.

Resumé

I løbet af den tid hvor nærværende afhandling er blevet udarbejdet, har dens forskningsemner bredt sig ud. Generelt omhandler afhandlingen fem hovedemner: Eksponentielle martingaler, generel proces teori, kausalitet, independent component analysis (ICA) og modelselektion for ikke-lineær regression.

I kapitel 1 giver vi et overblik over forskningsprojektet og dets resultater, og vi introducerer de hovedfelter, projektet omhandler. De resterende kapitler kan læses uafhængigt af hinanden og indeholder projektets hovedresultater.

Kapitlerne 2-4 omhandler martingalegenskaben for eksponentielle martingaler. Kapitel 2 betragter eksponentielle martingaler baseret på tælleprocesser. Vores mål er at konstruere statistiske modeller baseret på tælleprocesser med stokastisk intensitet og uden eksplosion. I særdeleshed konstruerer vi fordelinger af interagerende tælleprocesser og diffusioner. Kapitlerne 3 og 4 betragter egenskaben at være en uniformt integrabel martingal fra et mere abstrakt synspunkt. Resultaterne i det første af disse kapitler omhandler optimale konstanter i Novikov-type kriterier, og resultaterne i det andet kapitel omhandler et Novikov-type kriterium som benytter sig af både den optionelle og forudsigelige kvadratiske variation.

Resultaterne i kapitlerne 5 og 6 er relateret til generel proces teori. Grundlæggende set bevises ingen nye resultater. I stedet gives simplificerede beviser af kendte resultater. I særdeleshed bevises eksistensen af den duale forudsigelige projektion samt den kvadratiske variation. Vi diskuterer også hvordan disse resultater kan anvendes til at give en simplificeret udlægning af teorien for stokastiske integraler.

Kapitel 7 og kapitel 8 omhandler et kausalitetsbegreb for stokastiske differentialligninger (SDEer). Kapitel 7 introducerer kausalitetsbegrebet. Endvidere viser vi her at for SDEer drevet af Lévy processer er postinterventionsfordelinger identificerbare fra den observationelle fordeling. Dette står i kontrast til resultaterne fra Pearl's interventionskalkyle. I kapitel 8 anvender vi vores kausalitetsbegreb på Ornstein-Uhlenbeck SDEer, hvor det er muligt at foretage eksplicitte udregninger.

I kapitel 9 undersøger vi identifikation af blandingsmatricen i ICA. Det er et velkendt resultat at identifikation af blandingsmatricen afhæger af hvorvidt fejlfordelingerne er Gaussiske eller ej. Vi forsøger at belyse hvad der sker i tilfældet hvor fejlfordelingerne er tæt på at være Gaussiske uden faktisk at være det.

Til sidst diskuterer vi i kapitel 10 frihedsgrader for estimatorer i ikke-lineær regression. Vores motivation er L^1-begrænset og L^1-penaliseret estimation i en ikke-lineær regressionsmodel. Vores mål er at opnå resultater, der kan lede til udregning af frihedsgraderne for en estimator og derefter til modelselektion ved optimalt valg af penaliseringsparameteren. Vi viser to resultater relateret til frihedsgraderne: Et teoretisk resultat for begrænset estimation, og et mere praktisk anvendeligt resultat for L^1-penaliseret estimation.

Contents

1

Overview of results

1.1 Introduction

In this chapter, we will give an overview of the results obtained in this thesis. The first step towards this is a discussion of the objectives of the thesis, given in Section 1.2, where a description of the research project of the thesis is laid out in the context of previously known results. The results of the thesis can in a natural manner be divided into five categories:

- Exponential martingales
- The general theory of processes
- Causality and interventions
- Identifiability and ICA
- Model selection for nonlinear regression

Each of these subjects are covered separately in Sections 1.3-1.7, where we both outline the results obtained under each headline, as well as discuss our results and relate them to other literature. In Section 1.8, we review perspectives for future research based on the results obtained here.

As mentioned in the summary, the results obtained are given in nine chapters, each chapter corresponding to a manuscript for a paper. Some of these manuscripts are published at the time of this writing. Table 1.1 describes the publication status of each manuscript. The manuscripts "Exponential martingales and changes of measure for counting processes", "Causal interpretation of SDEs" and "On degrees of freedom in nonlinear regression" are co-authored with my supervisor, Prof. Niels Richard

Title	**Status**
Exponential martingales and changes of measure for counting processes	Under review
Optimal Novikov-type criteria for local martingales with jumps	*Elec. Comm. Prob.* **18**
An extended Novikov-type criterion for local martingales with jumps	Under review
An elementary proof that the first hitting time of an F_σ set by a jump process is a stopping time	*Sem. Prob.* **45**
Proving existence results in martingale theory using a subsequence principle	*Comm. Stoch. Anal.* **7**
Causal interpretation of SDEs	Under review
Intervention in Ornstein-Uhlenbeck SDEs	*Proc. 18th EYSM*
Quantifying identifiability in ICA	In preparation
On degrees of freedom in nonlinear regression	In preparation

Table 1.1: Publication status for the manuscripts included in this dissertation.

Hansen, while the manuscript "Quantifying identifiability in ICA" is co-authored with Prof. Marloes Maathuis and Benjamin Falkeborg.

Each of the manuscripts in Chapters 2-10 are designed to be read independently of the others. As some of the chapters concern themselves with similar subjects, some overlap in terms of introductory material and reviews of known results occur. We hope that the benefit of self-sufficiency for each chapter exceeds the possible inconvenience of repetition.

Finally, it should be noted that the vast majority of research results builds intimately on the works of others, and this dissertation is no exception. Many of the ideas employed here are variations or extensions of previously applied techniques. We have endeavored to give credit where credit is due, but our human infirmities will no doubt occassionally have lead us to fail in this respect, though not due to any ill will. Apart from that, we have also at some points intentionally repeated proofs of known results when we felt that the proofs we were able to find in the literature were terse enough to merit expansion for the sake of readability. Examples of this occurs in places such as Lemmas 3.2.2, 6.1.1, 6.5.1, 7.8.1 and 7.9.1, and presumably in other locations as well. In general, this thesis is written under the assumption that it is preferable to include trivial results rather than to exclude them, as this ensures that the reader will not manually have to think through the trivialities on

his or her own.

1.2 Objectives and description of the research project

Before beginning work on this thesis, we set up a series of objectives for the research to be done. However, as time passed, we were led both to change our initial objectives and to add new ones. In the following, we will outline the context of the initial research objectives, describe how these objectives and our corresponding research efforts changed with time, and describe how new objectives from other fields came to be added.

The initial research project outlined for the thesis was concerned with the results about causal inference outlined in [144, 37, 57]. Our objectives were to develop estimation methods which would lead to a practical methodology for applying the results developed by Røysland in [144]. In particular, we were interested in using L^1-penalized estimation, see [64], to estimate graphs describing the causality of the system under observation, and applying our methods to neuronal data similar to those analyzed in for example [132].

The interest of [144] is to estimate causal effects from observational studies in continuous time. In particular, the author of [144] considers the following model of a clinical trial. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ be a filtered probability space satisfying the usual conditions, see [66]. Let N^A, N^C, N^D and N^L be counting processes, where the former three are univariate, and the latter may be multivariate. For convenience, we always assume that these counting processes have intensities, meaning that their compensators are of the form $\int_0^t \lambda_s \, \mathrm{d}s$ for some nonnegative, predictable and locally bounded process λ, see [66, 17] for the definition of compensators, predictability and intensities. We then let

$$A_t = \int_0^t 1_{(s \leq T_A)} \, \mathrm{d}N_s^A, \tag{1.1}$$

$$C_t = \int_0^t 1_{(s \leq T_C)} \, \mathrm{d}N_s^C, \tag{1.2}$$

$$D_t = \int_0^t 1_{(s \leq T_D)} \, \mathrm{d}N_s^D, \tag{1.3}$$

$$L_t = L_0 + \int_0^t H_s^L \, \mathrm{d}N_s^L, \tag{1.4}$$

where T_A, T_C and T_D are the first jump times of the processes N^A, N^C and N^D. Also, L_0 is bounded and $\mathcal{F}_0$ measurable and H^L is a bounded and predictable matrix-valued process. The processes A, C, D and L are then what [144] refers to as observable processes, meaning stochastic integral processes with respect to counting processes. The processes A, C, D and L are meant to measure information about a clinical trial of a patient. In particular,

- A is the counting process for the event of initiating treatment
- C is the counting process for the event of censoring
- D is the counting process for a clinical event such as death
- L is a process measuring the patient's multivariate health condition

As these processes are observable processes, we may apply the theory of local independence developed by Didelez in [37]. In [37], it is described how to define the local independence graph for a collection of observable processes. [144] assumes that under the probability measure P, the local independence graph is as given in Figure 1.2.1. In this graph, the presence of a directed edge from X to Y is interpreted as the possibility of a causal influence of X on Y, while the absence of an edge is interpreted as no causal influence from X to Y.

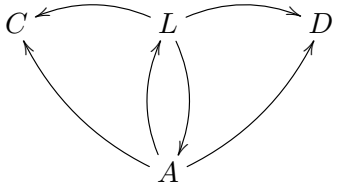


Figure 1.2.1: The local independence graph of (A, C, D, L).

In words, the assumptions made in Figure 1.2.1 is that treatment and patient health status influence each other, such that the health status may be influenced by the treatment given, but the treatment may also depend on the health status. This corresponds to that for example very sick patients may be more likely to receive treatment. Furthermore, clinical events are influenced by the treatment, but is also influenced by natural variation of the health status, and likewise for censoring. The important point here is that the two-way influence between A and L essentially confounds our ability to estimate the effect of the treatment A on the clinical event D. Also, [144] notes that it would be preferable for the censoring not to be influenced by the health status, since this by Theorem 1 of [144] and its comments would lead to the possibility of unbiased estimation of treatment effect when the health status L is unobserved. However, [144] also argues that this lack of influence cannot be assumed in practical observational studies. These considerations leads to the definition of what [144] calls a randomized trial measure $\tilde{P}$ under which the local independence graph is as given in Figure 1.2.2.

The randomized trial measure is a probability measure which is meant to describe a counterfactual experiment, that is, an experiment which in fact did not take place, in which the results of the study were distributed as if the study was randomized such that A is not influenced by L. It is furthermore assumed that L does not influence C. The methodology proposed is now the following: If a randomized trial measure $\tilde{P}$

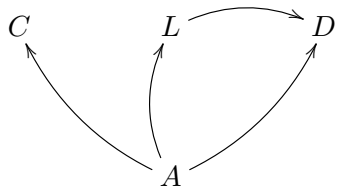


Figure 1.2.2: The local independence graph of (A, C, D, L) under the counterfactual randomized trial measure.

can be constructed which is absolutely continuous with respect to the observational measure P, then reweighed data can be constructed which allows estimation to be carried out under the randomized trial measure $\tilde{P}$ and which allows for unbiased estimation of the treatment effect.

One of the first goals we set ourselves in this context was to try to obtain sufficient conditions for the existence of the randomized trial measure. In Theorem 2 of [144], a sufficient criterion is given. We sought to investigate whether such criteria could be extended. As the construction of the randomized trial measure is based upon a change of measure using a particular class of exponential martingales as the likelihood ratio, this leads directly to the question of when such classes of exponential martingales are true martingales. Answering this question thus became one objective of the thesis. Our efforts in this direction are given in Chapter 2.

The literature regarding exponential martingales is rich, see for example the papers [123, 95, 94, 24, 109, 79, 89, 135]. During our work with this problem, the results of [135] came to our attention. In [135], it is recalled that a classical criterion, see [123], for the exponential martingale $\mathcal{E}(M)$ of a continuous local martingale M with initial value zero to be a uniformly integrable martingale is that

$$E \exp(\tfrac{1}{2}[M]_\infty) < \infty. \tag{1.5}$$

The question is then asked in [135] whether this can be extended to the case where M is not continuous, and it is shown that in general, when $\Delta M \geq -1$, such that $\mathcal{E}(M)$ is nonnegative, $\mathcal{E}(M)$ is a uniformly integrable martingale if only

$$E \exp(\tfrac{1}{2}\langle M^c \rangle_\infty + \langle M^d \rangle_\infty) < \infty, \tag{1.6}$$

and argues that the constants $\frac{1}{2}$ and 1 in front of $\langle M^c \rangle$ and $\langle M^d \rangle$ are optimal, although the proof contains a flaw. Inspired by problems encountered in the case where M is a stochatic integral with respect to a compensated counting process, we asked what the optimal constants were in the case of M with $\Delta M 1_{(\Delta M \neq 0)} \geq a$ for fixed $a \geq -1$. Based on the results obtained there, we also derived a particular sufficient criterion for the case $\Delta M \geq 0$, where a certain symmetry between the cases of using predictable and optional quadratic variations appears. Details of our results for the abstract setting are given in Chapters 3 and 4. A general overview of our results on exponential martingales is given in Section 1.3.

After our investigation of exponential martingales based on counting processes, we did not further pursue work on the results of [144]. However, we did begin considering another problem. The notion of influence or causality applied in [144] is based on the notion of local independence of [37]. Building on this notion, Gégout-Petit and Commenges in [27, 57] introduces a notion of influence between two processes in a class they refer to as $\mathcal{D}'$. This class is a subset of the space of semimartingales. The authors refer to this notion of influence as weak conditional local independence (WCLI), and proceed to propose that this notion is a good starting point for defining the causal influence between two processes. Other notions of influence or causality for continuous-time processes are discussed in [59, 54, 30, 131, 130].

We began considering whether it would be possible, instead of considering two processes, to define a notion of causality for stochastic differential equations. Our hope was to define a notion of causality which would fit well with the classical notion of causality based on directed acyclic graphs (DAGs) as developed in for example the books of Pearl and of Spirtes et al., [126, 161]. We also hoped for a notion of causality which would generalize the results of [57] to the case of non-orthogonal local martingales excluded there, and which at the same time would be easy to make operational in the sense of being applicable to the practical estimation of causal effects and the effects of interventions. Our results in this direction are outlined in Section 1.5, and details of our results are given in Chapter 7 and Chapter 8.

At the same time as this, we also began to pursue a different set of research problems, related to the general theory of processes and the theory of stochastic integration. Since the introduction of the stochastic integral with respect to Brownian motion by Itô in [78], the theory of stochastic integration for semimartingales has become an extensive and well-developed theory, see [85] for a short history. Several monographs are devoted to the introduction of the topic, see for example [36, 66, 143, 87, 83, 134], each contributing with simplified and improved proofs. As the theory is technically demanding, yet essential to several applied fields, most notably mathematical finance and actuarial science, it is of considerable interest to obtain as many simplifications of the proofs of the main results as possible. Our results in this direction include a simple result on the stopping time property of a particular hitting time, and a simplified proof of the existence of the dual predictable projection and the quadratic variation, using methods similar to those pioneered in [13]. Our efforts in this regard are discussed in Section 1.4, and our results are given in detail in Chapter 5 and Chapter 6.

Based on our work on causality for stochastic differential equations, we were confronted with the literature for causal inference in the DAG-based setting. The hospitality of the Seminar für Statistik at ETH Zürich allowed for a gentle introduction to these topics, in particular the results [156, 157] by Shimizu and colleagues on the LiNGAM method for causal discovery in linear structural equation models. The LiNGAM method is based on results related to independent component analysis (ICA), a concept introduced by Comon in [28] and since then developed into a

practical statistical method by the collaborative efforts of many, see for example [23, 28, 74, 75, 76, 122]. Essentially, in its simplest form, ICA is concerned with the following problem. Assume given a p-dimensional vector ε of independent mean zero error variables and a real $p \times p$ matrix A, known as the mixing matrix. Define

$$X = A\varepsilon. \tag{1.7}$$

If we observe samples from X, will we be able to identify A and the distribution of ε, and if so, how? Answering this question is of use in several disciplines, see for example, [29, 10, 12, 86, 172]. One notable result in this direction is that the amount of information embedded in the distribution of X about A depends crucially upon the number of Gaussian coordinates of ε. For example:

(1). If ε contains at most one Gaussian component, A can be identified up to scaling and permutation of columns.

(2). If ε contains only Gaussian components, only $A\Sigma A^t$ can be identified, where Σ is the covariance matrix of ε.

The latter claim (2) is clear since ε, having independent coordinates, is multivariate normal in this case, with mean zero and covariance matrix Σ, and so X is multivariate normal with mean zero and covariance matrix $A\Sigma A^t$. Note that since we have assumed that ε has independent coordinates, Σ is always diagonal. The former claim (1) is non-trivial, see [28] for a proof.

These observations lead naturally to the following question: In what sense does identification of A become harder as the coordinates of ε become closer to Gaussian without ever becoming Gaussian? As none of the coordinates are Gaussian, we are in scenario (1) given above. Nonetheless, as we move closer to scenario (2), we would expect a quantitative shift towards it being more difficult to identify A from samples of the distribution of X. The results we have obtained regarding the elucidation of this question are discussed in Section 1.6, with details given in Chapter 9.

Finally, we also gave thought to the problem of model selection for nonlinear models. This problem came to our attention in the following manner: During our work with causality for SDEs, we often had in mind an application related to gene expression networks. We pictured a network of p genes, each with an expression level X^i, such that X followed an SDE of the form

$$\mathrm{d}X_t = BX_t\,\mathrm{d}t + \mathrm{d}W_t. \tag{1.8}$$

Here, we have put the mean reversion level to zero and the diffusion matrix to the identity for tractability. According to our results on causality for SDEs, the zeroes of the mean reversion speed matrix B controls the causal structure for the gene expression network X. Observing X, we would therefore be particularly interested in sparse estimation of B, as this would give us an understanding of the causal

structure of the network. Now, given observations of X over equidistant time periods $t_k = k\Delta$ for $k = 0 \ldots, n$, a natural loss function for the estimation of B is

$$R(B) = \sum_{k=1}^{n} \|X_{t_k} - \exp(\Delta B) X_{t_{k-1}}\|_2^2. \tag{1.9}$$

We may then consider $\lambda \geq 0$ and obtain a sparse estimator $\hat{B}_\lambda$ of B by letting

$$\hat{B}_\lambda \in \operatorname*{argmin}_{B \in \mathbb{M}(p,p)} R(B) + \lambda \|B\|_1, \tag{1.10}$$

where $\|\cdot\|_1$ denotes the entrywise L^1-norm, and $\mathbb{M}(p,p)$ denotes the space of real $p \times p$ matrices. For any $\lambda \geq 0$, this is a L^1-penalized estimation problem, and can be solved numerically. However, this raises the question of how to choose $\lambda \geq 0$. Inspired by the methodology outlined for example in the book [64] by Hastie et al., we posed the problem of calculating the degrees of freedom for an estimator of the type (1.10). This would allow us to minimize an estimate of the generalization error and thus perform sparse model selection, see Section 1.7 for details.

This problem, however, turned out to be more difficult than expected. Therefore, as a first step, we considered the simpler problem of calculating the degrees of freedom for estimators in nonlinear regression models with independent Gaussian errors. Our results on this problem are given in Chapter 10. A discussion of our results can be found in Section 1.7.

Summing up, the final research objectives which this thesis is organized around are:

- The investigation of sufficient criteria for the exponential martingale of a stochastic integral with respect to a compensated counting process to be a true martingale, with application to the existence of randomized trial measures as given in [144].
- Proof of optimal Novikov-type criteria for the exponential martingale of particular classes of local martingales to be uniformly integrable martingales.
- The design of a notion of causality for stochastic differential equations, and investigation of its relationship to DAG-based causality, its relationship to weak conditional local independence and its prospects for practical application.
- The development of simplified proofs in the general theory of processes and its use in the development of a simplified account of the theory of stochastic integration with respect to general semimartingales.
- The exploration of the asymptotic behaviour of the ICA model as the distribution of the error variables converge to Gaussian distributions in an appropriate sense.

- The elucidation of the concept and calculation of degrees of freedom for nonlinear regression models.

In the following sections, overviews and discussions of our progress in each of these research objectives are given. Finally, Section 1.8 outlines opportunities for further work.

1.3 Exponential martingales

As noted in the previous section, the initial motivation for our work with exponential martingales was the construction of randomized trial measures as defined in [144]. Such randomized trial measures essentially correspond to distributions of counting processes with particular intensities, and can be constructed using a change of measure. To see how this can be carried out, consider the following setup. Assume given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ satisfying the usual conditions. For any semimartingale X with initial value zero, we define

$$\mathcal{E}(X)_t = \exp\left(X_t - \frac{1}{2}[X^c]_t\right) \prod_{0<s\leq t} (1 + \Delta X_s) \exp(-\Delta X_s), \tag{1.11}$$

the stochastic exponential of X. This process is the unique solution in Z to the stochastic integral equation

$$Z_t = 1 + \int_0^t Z_{s-} \, \mathrm{d}X_s. \tag{1.12}$$

In the above, X^c denotes the continuous martingale part of X, see Proposition I.4.27 of [83], and $[X^c]$ denotes the quadratic variation of the process X^c. If X is a local martingale, $\mathcal{E}(X)$ is a local martingale as well, and if $\Delta X \geq -1$, $\mathcal{E}(X)$ is nonnegative. Now assume given a d-dimensional counting process N with d-dimensional intensity λ. Here, λ is generally a nonnegative, predictable and locally bounded process. For convenience, we assume in this introductory section that λ is positive. This assumption can be weakened, see Chapter 2. Also assume given another process μ of this type, and define

$$M_t^i = N_t^i - \int_0^t \lambda_s^i \, \mathrm{d}s \tag{1.13}$$

$$\gamma_t^i = \mu_t^i / \lambda_t^i \tag{1.14}$$

$$H_t^i = \gamma_t^i - 1. \tag{1.15}$$

for $t \geq 0$. The process M is then a local martingale, in the sense that each of its coordinates M^i are local martingales. We also put

$$(H \cdot M)_t = \sum_{i=1}^d \int_0^t H_s^i \, \mathrm{d}M_s^i. \tag{1.16}$$

Note that

$$\Delta(H \cdot M) = \sum_{i=1}^{d} H^i \Delta M^i = \sum_{i=1}^{d} H^i \Delta N^i. \tag{1.17}$$

Making the further assumption that none of the N^i jump at the same time, we obtain $\Delta(H \cdot M) \geq -1$, since $H \geq -1$. Therefore, $\mathcal{E}(H \cdot M)$ is nonnegative. Next, let T be a stopping time, and assume that $\mathcal{E}(H \cdot M)^T$ is a uniformly integrable martingale. Then $E\mathcal{E}(H \cdot M)_T = 1$, so we may define a probability measure Q with Radon-Nikodym derivative $\mathcal{E}(H \cdot M)_T$ with respect to P. By Lemma 2.2.2, which essentially is an application of Girsanov's theorem, it holds that under Q, N is a d-dimensional counting process with intensity $1_{[0,T]}\mu + 1_{(T,\infty)}\lambda$.

This leads to the following conclusion: Using a change of measure with the exponential martingale $\mathcal{E}(H \cdot M)$, it is possible to construct a counting process with intensity μ on $[0, T]$ from a counting process with intensity λ on $[0, T]$. In general, we cannot expect to change the intensity on the whole of $\mathbb{R}_+$, as most counting processes with intensities differing on all of $\mathbb{R}_+$ have singular distributions. For example, the distributions of two homogeneous Poisson processes with different intensities are singular, see Proposition 3.24 of [92]. As many experiments are carried out over a finite and deterministic time period, it is natural to restrict our attention to changing intensities over a deterministic time interval $[0, t]$. This corresponds to having $E\mathcal{E}(H \cdot M)_t = 1$, and this is in particular the case if $\mathcal{E}(H \cdot M)$ is a martingale.

Therefore, summing up, a plan for achieving our objective of constructing randomized trial measures would be to derive sufficient criteria for $\mathcal{E}(H \cdot M)$ to be a martingale. This is the project we carry out in Chapter 2.

As noted earlier, there exist many abstract criteria for an exponential local martingale to be a uniformly integrable martingale. We chose to work with the very general results of Lépingle and Mémin given in [109]. Their main results are restated below. We use the notation $\Pi_p^* A$ for the dual predictable projection, or compensator, of A, see Definition 5.21 of [66].

Theorem ([109], Theorem III.1). *Let M be a local martingale with initial value zero and $\Delta M \geq -1$. Let $R = \inf\{t \geq 0 \mid \Delta M_t = -1\}$. Define a process B by putting $B_t = \frac{1}{2}[M^c]_{t\wedge R} + \sum_{0<s\leq t\wedge R}(1+\Delta M_s)\log(1+\Delta M_s) - \Delta M_s$. If B is locally integrable and $\exp(\Pi_p^* B_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale.*

Theorem ([109], Theorem III.7). *Let M be a local martingale with initial value zero and $\Delta M > -1$. Define $A_t = \frac{1}{2}[M^c]_t + \sum_{0<s\leq t}\log(1+\Delta M_s) - \Delta M_s(1+\Delta M_s)^{-1}$. If $\exp(A_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale.*

The criterion in Theorem III.1 of [109] is known as a predictable type of criterion, involving an exponential moment of a predictable process, while the criterion in Theorem III.7 of [109] is known as an optional type of criterion. These criteria can immediately be translated into the following result.

Theorem 1.3.1. *It holds that $\mathcal{E}(H \cdot M)$ is a uniformly integrable martingale if one of the following two conditions hold:*

$$E \exp\left(\sum_{i=1}^{d} \int_0^\infty (\gamma_s^i \log \gamma_s^i - (\gamma_s^i - 1))\lambda_s^i \, \mathrm{d}s\right) < \infty \quad \textit{or} \tag{1.18}$$

$$E \exp\left(\sum_{i=1}^{d} \int_0^\infty \log \gamma_s^i - \frac{\gamma_s^i - 1}{\gamma_s^i} \, \mathrm{d}N_s^i\right) < \infty. \tag{1.19}$$

Barring the minor detail that we should focus on the martingale property and not the uniformly martingale property, these conditions illustrate well how the change of measure depends on the initial and final intensities λ and μ. To understand the result, consider for convenience the case $\lambda = 1$. The functions being integrated in the conditions (1.18) and (1.19) are then

$$\varphi_p(x) = x \log x - (x - 1) \tag{1.20}$$

$$\varphi_o(x) = \log x - \frac{x - 1}{x}, \tag{1.21}$$

in particular, $\varphi_o(x) = \varphi_p(x)/x$. These functions are plotted in Figure 1.3.1.

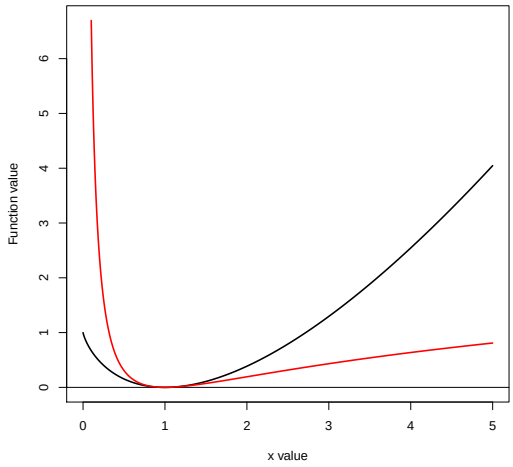


Figure 1.3.1: Black: The function φ_p (1.20). Red: The function φ_o (1.21).

Now, both of the functions φ_p and φ_o tend to infinity as their arguments tend to infinity. This indicates that the change of measure to some new intensity becomes increasingly difficult as the intensity heightens. The behaviour of the two functions are markedly different as their arguments tend to zero, however. Here, φ_p tends to one, while φ_o tends to infinity. This indicates that φ_o might be unsuitable for a change of measure to a small intensity.

To overcome this problem, we considered the heuristic idea of dividing the change of measure up into two parts: A part corresponding to a change of measure to smaller intensity, and a part corresponding to a change of measure to larger intensity. For the change of measure to smaller intensity, Theorem III.1 of [109] can be applied. Furthermore, we used a method of cutting up time intervals into small pieces to further strengthen our sufficient criterion. We later found that a similar methodology had been applied in the proof of Lemma 13 of [135]. The resulting criteria are given as Theorem 2.2.4, restated here for convenience in a simplified version.

Theorem 2.2.4. *It holds that $\mathcal{E}(H \cdot M)$ is a martingale if there is an $\varepsilon > 0$ such that whenever $0 \leq u \leq t$ with $t - u \leq \varepsilon$, one of the following two conditions are satisfied:*

$$E \exp\left(\sum_{i=1}^{d} \int_u^t (\gamma_s^i \log \gamma_s^i - (\gamma_s^i - 1))\lambda_s^i \, \mathrm{d}s\right) < \infty \quad \text{or} \tag{2.4}$$

$$E \exp\left(\sum_{i=1}^{d} \int_u^t \lambda_s^i \, \mathrm{d}s + \int_u^t \log_+ \gamma_s^i \, \mathrm{d}N_s^i\right) < \infty. \tag{2.5}$$

In (2.5), the requirements for applying Theorem III.1 of [109] enter through the presence of $\int_u^t \lambda_s^i \, \mathrm{d}s$. A major benefit of this, however, is that in the case $\lambda = 1$, this term becomes deterministic and as such vanishes from the criterion. This leads to a simple sufficient optional criterion for $\mathcal{E}(H \cdot M)$ to be a martingale in the case where the initial measure is a standard homogeneous Poisson process.

We now discuss the consequences of Theorem 2.2.4. First note that Theorem 2.2.4 can be used to extend the criteria of Theorem 2 in [144] for the existence of randomized trial measures. This is demonstrated in Example 2.3.1. As such, the results provides a contribution towards our initial research objective.

Furthermore, note that by the results already discussed, the martingale property of $\mathcal{E}(H \cdot M)$ implies for any $t \geq 0$ the existence of a probability measure Q such that under Q, N has intensity μ on $[0, t]$. As N is a càdlàg and piecewise constant process with jumps of size one, this yields the construction of a nonexplosive counting process on $[0, t]$ with intensity μ. Therefore, Theorem 2.2.4 can be seen as giving criteria for non-explosion of counting processes. As shown in Section 2.3, both of the criteria of Theorem 2.2.4 are strong enough to replicate the classical result that there exists counting processes with intensities of the form $\mu_t \leq \alpha + \beta N_{t-}$ without explosion, see Section 4.4 of [81]. As the calculations corresponding to the examples of Section 2.3 show, the reduction to arbitrarily small time intervals in Theorem 2.2.4 is essential to achieve this result. Summing up, we have through purely martingale-based methods achieved a replicate of one of the classical criteria for non-explosion. Several other types of counting processes can also be constructed in this way, for example Hawkes processes, again see Section 2.3 for details.

For intensities which are predictable with respect to the filtration generated by the

counting process itself, we can also use methods such as given in [81] to obtain non-explosion. However, Theorem 2.2.4 is also applicable in the case where the intensity is predictable only with respect to other, larger filtrations. In particular, we may consider intensities depending both on the counting process itself and other processes such as diffusion processes. At first, we considered that this might be useful for joint modeling of neuronal spike trains (as a counting process, see [169, 133, 115]) and the membrane potential between spikes (as a diffusion process, see [18, 19, 105, 39]). We later discovered that there in other contexts already had been interest in similar models containing such interacting counting process and diffusion components, see [58, 9].

In order to investigate the existence of such processes, we considered two classes of intensities μ. First, we considered a specification given by:

$$\begin{aligned} \mu_t &= \phi(X_t) \\ \mathrm{d}X_t &= (A(N_t, Z_t) + B(N_t, Z_t)X_t)\,\mathrm{d}t + \sigma(N_t, Z_t)\,\mathrm{d}W_t, \end{aligned} \tag{1.22}$$

where T_n^i denotes the n'th jump time for N^i, $Z_t^i = t - T^i_{N_t^i}$, A is vector valued and B and σ are matrix valued functions, ϕ is a Lipschitz mapping with nonnegative coordinates and W is an $(\mathcal{F}_t)$ Brownian motion independent of N. Second, we considered a specfication given by:

$$\begin{aligned} \mu_t &= |X_{t-}| \\ \mathrm{d}X_t &= (a_{N_t} + b_{N_t}X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t + (\xi_{N_t} - X_{t-})\,\mathrm{d}N_t, \end{aligned} \tag{1.23}$$

where (a_n), (b_n) and (ξ_n), are sequences in $\mathbb{R}$. In Example 2.3.4 and Example 2.3.5, we use (2.4) and (2.5), respectively, to prove that under certain regularity conditions on the parameters in the intensity specification, a change of measure could be applied to construct a counting process with intensity of the types given above. Interpretations of the two types of intensities are given in Section 2.3. Note that because of the diffusion component present in these intensity specifications, the periods of time where the intensity is very small is not easy to characterize. Therefore, in order to develop in particular the latter example, the the fact that the criterion (2.5) is insensitive to small intensities in contrast to (1.19) is essential.

Inspired by the work outlined above, we next considered another, more abstract question. In [135], it was argued by Protter and Shimbo that when M is a local martingale with initial value zero and $\Delta M \geq -1$, then $\mathcal{E}(M)$ is a uniformly integrable martingale if only $E\exp(\frac{1}{2}\langle M^c\rangle_\infty + \langle M^d\rangle_\infty)$ is finite, and the coefficients in front of $\langle M^c\rangle$ and $\langle M^d\rangle$ are optimal. This is a predictable type of sufficient criterion. Here, M^c and M^d denote the continuous and purely discontinuous martingale parts of M, see Theorem 7.25 of [66], and $\langle M\rangle$ denotes the predictable quadratic variation of M. Having noted that there seems to be a distinct difference between predictable and optional sufficient requirements for $\mathcal{E}(M)$ to be a uniformly integrable martingale, depending on the jump sizes of M, we proceeded to ask the following question:

For $a \geq -1$ and a local martingale M with initial value zero and jumps satisfying $\Delta M 1_{(\Delta M \neq 0)} \geq a$, what are the optimal constants $\alpha(a)$ and $\beta(a)$ such that

$$E \exp(\tfrac{1}{2}\langle M^c \rangle_\infty + \alpha(a)\langle M^d \rangle_\infty) < \infty \tag{1.24}$$

and

$$E \exp(\tfrac{1}{2}[M^c]_\infty + \beta(a)[M^d]_\infty) < \infty \tag{1.25}$$

suffice to make $\mathcal{E}(M)$ a uniformly integrable martingale? As the results of [109] are capable of giving sufficient criteria of this type as corollaries, the difficult part of this question is to identify optimality of the constants. Using counterexamples similar to those employed in [135, 109], the optimal constants are identified in Chapter 3, and are given as

$$\alpha(a) = \frac{(1+a)\log(1+a) - a}{a^2} \quad \text{and} \tag{1.26}$$

$$\beta(a) = \frac{(1+a)\log(1+a) - a}{a^2(1+a)}, \tag{1.27}$$

for $a > -1$ with $a \neq 0$, and $\alpha(0) = \beta(0) = \frac{1}{2}$, $\alpha(-1) = 1$. For $a = -1$, there exists no optimal constant for the optional case. These functions are shown in Figure 1.3.2.

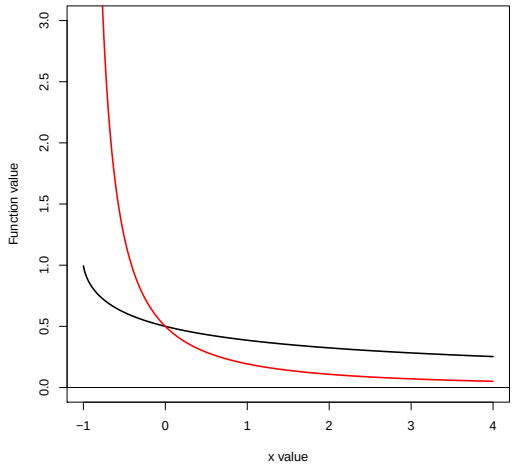


Figure 1.3.2: Black: The function α (1.26). Red: The function β (1.27).

Noting that zero is the unique argument where α and β take the same value, we proceeded next to wonder whether a particularly elegant sufficient criterion could be obtained for this case, perhaps using simplified methods of proof. Inspired by

the methods of Krylov in [101], we prove in Chapter 4 that for any $0 \leq \gamma \leq 1$, the criterion

$$\liminf_{\varepsilon \to 0} \varepsilon \log E \exp\left((1-\varepsilon)\frac{1}{2}(\gamma[M]_\infty + (1-\gamma)\langle M\rangle_\infty)\right) < \infty \tag{1.28}$$

suffices to ensure that $\mathcal{E}(M)$ is a uniformly integrable martingale. This is a combined predictable and optional type of criterion.

1.4 The general theory of processes

Inspiring our work on the general theory of processes was the gradual development of simpler and simpler proofs of many of the main results of the theory of stochastic integration with respect to semimartingales. Much work has been done on this by many authors. For example, in the book [35] by Dellacherie and Meyer, arguably one of the first relatively complete accounts of the general theory of processes and its application to semimartingale theory, much time and effort was spent on proving several of the most difficult theorems of the theory, for example the début theorem, the section theorems, the Doob-Meyer theorem and the existence of the quadratic variation. In later expositions such as [134, 87] by Protter and Kallenberg, respectively, the dependence on the début theorem and the section theorems is removed, using innovations by, among others, [136, 11]. This in particular removes the need for the development of the Choquet theory of capacity ordinarily used for the proof of the début theorem and the section theorems.

Our contributions consist of elementary proofs of two results. The first regards stopping times and the début theorem, details can be found in Chapter 5. The début theorem is the following landmark result, see [35], Section III.44 for a full proof.

Theorem (début theorem). *Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ satisfying the usual conditions. Let X be a progressively measurable stochastic process and let B be a Borel subset of the real numbers. Define*

$$T = \inf\{t \geq 0 \mid X_t \in B\}, \tag{1.29}$$

the first hitting time of B by X. Then T is a stopping time.

This theorem ensures that practically all of the "time-like" variables of use in the general theory of processes in fact qualify as stopping times. That the assumption of the usual conditions in fact is necessary for the result to hold is not obvious, but can be seen from an example given in Section III of [34].

However, for many purposes, such as the development of the stochastic integral with respect to a semimartingale, the full generality of the début theorem is not needed. For example, a classical necessity for the development of this theory is the result that the jumps of a càdlàg adapted process X can be covered by the graphs of a

countable family of stopping times, meaning that there exists a sequence of stopping times (T_n) such that

$$\{(t,\omega) \mid \Delta X_t \neq 0\} \subseteq \cup_{n=1}^{\infty} [\![T_n]\!], \tag{1.30}$$

where $[\![T]\!] = \{(t,\omega) \mid t = T(\omega)\}$. In Chapter III of [66], this result is obtained by putting $T_0^k = 0$ and recursively defining

$$T_{n+1}^k = \inf\{t > T_n^k \mid |X_{T_n^k} - X_t| \geq 1/k\}. \tag{1.31}$$

It is then shown that each T_n^k is a stopping time, and that the family $(T_n^k)_{n\geq 0, k\geq 1}$ covers the jumps of X. In Chapter 5, we prove, using only elementary methods, that for any càdlàg adapted process X and any F_σ set U, meaning a countable union of closed sets, the variable

$$T = \inf\{t \geq 0 \mid \Delta X_t \in U\} \tag{1.32}$$

is a stopping time. A family of stopping times satisfying (1.30) can then be obtained by for example defining $T_0^k = 0$ and recursively

$$T_{n+1}^k = \inf\{t > T_n^k \mid |\Delta X_t| > 1/k\}. \tag{1.33}$$

The difficulty of showing that the variable T of (1.32) is a stopping time arises from the fact that ΔX neither has right nor left limits. In order to circumvent this problem, we consider the cases $0 \in U$ and $0 \notin U$ separately, and in the latter case employ a particular approximation procedure where the exclusion of 0 in the set U essentially allows us to distinguish between whether the limit of particular sequences X_{p_n} and X_{q_n} are equal to X_{s-} or X_s for $\lim p_n = \lim q_n = s$.

Our other contribution is related to the existence of the dual predictable projection and the quadratic variation. To outline our results, we first recall the Doob-Meyer theorem, see Chapter VI of [143] for a proof.

Theorem (Doob-Meyer decomposition). *Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ satisfying the usual conditions. Let X be a submartingale with initial value zero such that $\{X_T \mid T \text{ is a finite stopping time}\}$ is uniformly integrable. Then, there exists a predictable, integrable increasing process A with initial value zero, such that with*

$$M_t = X_t - A_t, \tag{1.34}$$

it holds that M is a uniformly integrable martingale.

In the statement of the above theorem, all processes are càdlàg as well. The theorem has a localized version covering all submartingales, see Section III.3 of [134] for details. The predictable process A is known as the compensator of X. In the case where X is increasing, it is also known as the dual predictable projection of X.

In [13], Beiglböck et al. gives a simplified proof of the Doob-Meyer decomposition theorem using a subsequence principle. While Komlós is given as the source of this subsequence principle, see [98], the result applied is also very similar to general functional analytic results for reflexive Banach spaces: In such spaces, bounded sequences have the property that there exists a sequence of convex combinations of the elements of the tail of the sequence which converge strongly. We sought to further simplify the methods of [13] by restricting ourselves to the case in the Doob-Meyer decomposition where the submartingale is pathwisely increasing. The interest in this lies in the observation that the construction of both the quadratic variation and the stochastic integral in fact can be carried out using only the existence of the compensator for increasing and locally integrable processes, or more generally, processes of locally integrable variation. Our efforts at this are outlined in Chapter 6.

Before describing our proof, we first state the theorem to be proven.

Theorem (existence of the compensator for locally integrable variation processes). *Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ satisfying the usual conditions. Let A be an adapted process of locally integrable variation and initial value zero. Then, there exists a predictable process $\Pi_p^* A$ of locally integrable variation with initial value zero, such that with*

$$M_t = A_t - \Pi_p^* A_t, \tag{1.35}$$

it holds that M is a local martingale.

In order to prove the existence of the compensator $\Pi_p^* A$ of an adapted process A of locally integrable variation with initial value zero, we first use monotone convergence arguments similar to those employed by [11] to reduce to the case where $A_t = \xi 1_{(t \geq T)}$ for ξ bounded and $\mathcal{F}_T$ measurable. In this case, we can apply a simple $\mathcal{L}^2$ version of the subsequence principle also used in [13], and the existence of the compensator follows from the existence of discrete-time compensators and the limes superior arguments of [84].

A different type of argument is needed for the existence of the quadratic variation for a local martingale M with initial value zero. Again, we first state the result we wish to obtain, a full proof can also be found in Chapter VI of [143].

Theorem (existence of the quadratic variation). *Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ satisfying the usual conditions. Let M be a local martingale with initial value zero. Then, there exists an increasing adapted process $[M]$ with initial value zero, such that*

1. $M^2 - [M]$ is a local martingale.
2. $\Delta[M] = (\Delta M)^2$ up to indistinguishability.

To give a simplified proof of this theorem, we first note that by the fundamental theorem of local martingales, see for example Theorem III.29 of [134], we can reduce

the problem to the cases of M bounded or M of integrable variation. We note that the proof of the fundamental theorem of local martingales only requires the existence of the compensator for locally integrable variation processes, which we already have obtained through elementary methods. The existence of the quadratic variation for local martingales of integrable variation and the quadratic covariation with a local martingale of integrable variation is not difficult, this can be obtained through the ordinary integration-by-parts formula and the martingale properties of integrals of predictable processes with respect to local martingales with paths of integrable variation. The remaining challenge is then to obtain the existence of the quadratic variation for a bounded martingale, and it is here that we again employ the subsequence principle.

The fundamental idea of the proof builds on the same method as used in the construction of the quadratic variation for continuous local martingales in for example Section IV.30 of [143]. We consider a bounded martingale M, put $t_k^n = k2^{-n}$ and note that for all n, it holds that

$$M_t^2 = N_t^n + Q_t^n, \tag{1.36}$$

where

$$N_t^n = 2 \sum_{k: t_{k-1}^n < t} M_{t_{k-1}^n}^t (M_{t_k^n}^t - M_{t_{k-1}^n}^t), \tag{1.37}$$

$$Q_t^n = \sum_{k: t_{k-1}^n < t} (M_{t_k^n}^t - M_{t_{k-1}^n}^t)^2. \tag{1.38}$$

Here, N_t^n approximates $2\int_0^t M_{s-}\,\mathrm{d}M_s$, while Q_t^n approximates $[M]_t$. We then use the subsequence principle to choose a sequence of convex combinations of the tail of (N^n) converging to a square-integrable martingale N. We may then prove that there exists a modification $[M]$ of $M^2 - N$ satisfying the requirements to be the quadratic variation of M.

In the case where M is continuous, this yields an even simpler proof, in fact almost a one-page proof, of the existence of the quadratic variation.

1.5 Causality and interventions

Our work with causality and interventions concerns the development of these notions for stochastic differential equations (SDEs). In order to properly understand the context of our work, we begin by giving an exposition of some results from the theory of causal inference in a non-time-dependent context. After doing so, we outline our work on interventions in SDEs and discuss its ramifications.

1.5.1 DAG-based causality

The notion of causality has long been under debate in the field of statistics, see for example [62, 146, 147, 70, 139, 127, 33]. In the theory of causality based on directed acyclic graphs (DAGs) as outlined by Pearl and Spirtes et al. respectively in [126, 161], it is recognized that the distribution of a set of random variables in general cannot be used to uniquely identify causal relationships between variables. Therefore, in order to analyze causality in a formal framework, it is necessary to extend the traditional framework of statistical inference, concerned with inference of distributions, to a framework including a carrier of information about causality. That carrier is the DAG.

In order to introduce the concept of a DAG, recall that a directed graph G on a set of vertices V is a pair (V, E) with $E \subseteq V \times V$. The elements of E are referred to as edges, and if $(\alpha, \beta) \in E$, we say that G contains the directed edge from α to β and write that $\alpha \to \beta$ in G. A path is defined to be an unbroken series of vertices and edges such that no vertices are repeated except possibly the initial and terminal vertices. We say that a path is in G if all its edges are in G. A directed cycle is a path with the same initial and terminal vertices and all arrows pointing in the same direction. For the purposes of this introductory subsection, we assume that the vertex set V is finite.

Definition 1.5.1. We say that a directed graph G is a directed acyclic graph (DAG) if G contains no directed cycles.

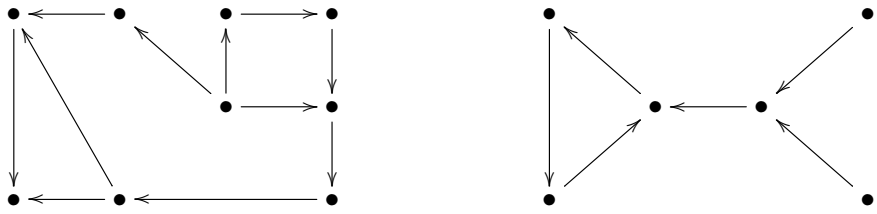


Figure 1.5.1: Left: A graph which is a DAG. Right: A graph which is not a DAG.

Directed graphs lend themselves naturally to graphical representation. In Figure 1.5.1, two graphs are drawn, one which is a DAG and one which is not a DAG. Directed arrows correspond to edges in the graph.

In a DAG G, there is a natural notion of parents, ancestors, descendants and non-descendants. To describe this, we will refer to a path as a forward path if all the edges in the path point in the forward direction, and refer to a path as a backward path if all the edges in the path point in the backward direction, see [126, 107]. For

any DAG G and α a vertex of G, we then use the following notation:

$$\begin{aligned}
\mathrm{pa}_G(\alpha) &= \{\beta \in V \mid (\beta, \alpha) \in E\}, \\
\mathrm{an}_G(\alpha) &= \{\beta \in V \mid \text{there is a forward path from } \beta \text{ to } \alpha \text{ in } G\}, \\
\mathrm{de}_G(\alpha) &= \{\beta \in V \mid \text{there is a forward path from } \alpha \text{ to } \beta \text{ in } G\}, \\
\mathrm{nd}_G(\alpha) &= V \setminus (\mathrm{de}_G(\alpha) \cup \{\alpha\}).
\end{aligned}$$

We refer to $\mathrm{pa}_G(\alpha)$ as the parents of α, to $\mathrm{an}_G(\alpha)$ as the ancestors of α, to $\mathrm{de}_G(\alpha)$ as the descendants of α and to $\mathrm{nd}_G(\alpha)$ as the non-descendants of α. Note that none of these sets contain α. These definitions make sense for any directed graph, but our assumption that G is a DAG ensures that for example ancestors and descendants do not overlap. If the DAG G is clear from the context, we write $\mathrm{pa}(\alpha)$ instead of $\mathrm{pa}_G(\alpha)$ and so forth.

Another description of a DAG can be obtained using the notion of a topological ordering.

Definition 1.5.2. Let G be a directed graph on V and let $\sigma : \{1, \ldots, |V|\} \to V$ be a bijective mapping. We say that σ is an topological ordering for G if it holds that the existence of a forward path from $\sigma(i)$ to $\sigma(j)$ implies $i < j$.

It can be shown that a directed graph G is a DAG if and only if there exists a topological ordering for G. This shows how DAGs essentially are graphs which allow for an ordering of the vertices in the direction of the edges.

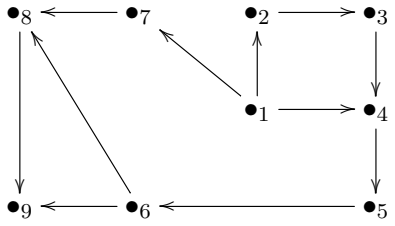


Figure 1.5.2: A topological ordering for the DAG from Figure 1.5.1.

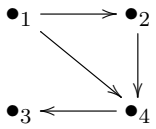
To see how a DAG can be used as a carrier of causal information, we introduce the notion of a structural equation model (SEM). We will define this in a slightly different manner than in the literature in order to make our definition mathematically precise. See Section 1.4.1 of [126] for an alternative formulation. In the following, fix a background probability triple $(\Omega, \mathcal{F}, P)$. For convenience, we only consider SEMs of variables taking values in $\mathbb{R}$.

Definition 1.5.3. A structural equation model (SEM) is a triple of objects of the type $(G, (U_\alpha)_{\alpha \in V}, (f_\alpha)_{\alpha \in V})$, where G is a DAG on V, $(U_\alpha)_{\alpha \in V}$ is a set of real random variables defined on $(\Omega, \mathcal{F}, P)$ and f_α is a function from $\mathbb{R}^{\mathrm{pa}(\alpha)} \times \mathbb{R}$ to $\mathbb{R}$ for $\alpha \in V$. We say that a set of random variables $(X_\alpha)_{\alpha \in V}$ defined on $(\Omega, \mathcal{F}, P)$ is a solution to the SEM if it holds that $X_\alpha = f_\alpha(X_{\mathrm{pa}(\alpha)}, U_\alpha)$ for all $\alpha \in V$.

In Definition 1.5.3, $X_{\mathrm{pa}(\alpha)}$ denotes the subset $(X_\beta)_{\beta\in\mathrm{pa}(\alpha)}$ of the family $(X_\beta)_{\beta\in V}$. The idea behind Definition 1.5.3 is that when $(X_\alpha)_{\alpha\in V}$ is a solution to the SEM, then $(G,(U_\alpha)_{\alpha\in V},(f_\alpha)_{\alpha\in V})$ describes the mechanism yielding $(X_\alpha)_{\alpha\in V}$ from the error, or noise, variables $(U_\alpha)_{\alpha\in V}$. The inclusion in our modeling framework of a mechanism and not just a set of variables is what allows us to define notions related to causality. In particular, a directed edge from α to β in the graph G is interpreted as a possible causal effect of X_α on X_β, while the absence of such an edge is interpreted as the absence of a causal of X_α on X_β. In this way, we may think of the entire SEM as a data-generating mechanism and as the DAG as describing the causal structure of this mechanism.

It is not hard to prove that there always exists a unique solution to a SEM, obtained by algorithmically determining the values X_α according to a topological ordering implied by the DAG, beginning with nodes having no parents, moving on to eligible children of these nodes and so forth. In the literature, it is regularly the case that a SEM simply is defined as a set of variables satisfying a certain set of equations, as this is less cumbersome than working with Definition 1.5.3.

Example 1.5.4. Consider a probability space endowed with four real noise variables $(U_1,\dots,U_4)$. Also consider a DAG G given by



as well as functional relationships

$$\begin{aligned} f_1(u_1) &= u_1 \\ f_2(x_1,u_2) &= 3x_1+u_2 \\ f_3(x_4,u_3) &= 8x_4+u_3 \\ f_4(x_1,x_2,u_4) &= x_1-x_2+u_4. \end{aligned}$$

By evaluating the functions $f_1,\dots,f_4$ in the order of the first, second, fourth and third functions, we find that there exists a unique set of variables $X_1,\dots,X_4$ such that

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= 3X_1+U_2 \\ X_3 &= 8X_4+U_3 \\ X_4 &= X_1-X_2+U_4, \end{aligned}$$

and this is then the solution to the SEM corresponding to the DAG G, the noise variables $U_1,\dots,U_4$ and the functional relationships $f_1,\dots,f_4$. These variables are

given by

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= 3U_1 + U_2 \\ X_3 &= 8U_4 - 16U_1 - 8U_2 + U_3 \\ X_4 &= U_4 - 2U_1 - U_2, \end{aligned}$$

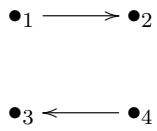
as can be obtained by a simple recursive calculation. ∘

In order to show how to endow this construction with a causal meaning, we define the notion of an intervention in a SEM.

Definition 1.5.5. Consider a SEM $(G, (U_\alpha)_{\alpha\in V}, (f_\alpha)_{\alpha\in V})$. Assume given some element $\beta \in V$ and some $x_\beta \in \mathbb{R}$. Let G' be the graph obtained by removing all edges of the form (α, β) from G. The postintervention SEM obtained by making the intervention $X_\beta := x_\beta$ is the SEM with DAG G', error variables $(U_\alpha)_{\alpha\in V}$ and functional relationships f_α for $\alpha \neq \beta$ and $f_\beta(u) = x_\beta$.

Note that the graph G' in Definition 1.5.5 in fact is a DAG, as removing edges from a DAG always preserves the DAG property. We also refer to the postintervention SEM of Definition 1.5.5 as the SEM obtained by doing $X_\beta := x_\beta$. Example 1.5.6 shows how this notion of intervention for SEMs works in a concrete case.

Example 1.5.6. Proceeding with the same SEM as described in Example 1.5.4, let us consider making the intervention $X_4 := \zeta$. This results in the DAG G' given by



and functional relationships

$$\begin{aligned} f_1(u_1) &= u_1 \\ f_2(x_1, u_2) &= 3x_1 + u_2 \\ f_3(x_4, u_3) &= 8x_4 + u_3 \\ f_4(u_4) &= \zeta. \end{aligned}$$

This SEM has a solution given by

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= 3U_1 + U_2 \\ X_3 &= 8\zeta + U_3 \\ X_4 &= \zeta, \end{aligned}$$

such that these variables satisfy

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= 3X_1 + U_2 \\ X_3 &= 8X_4 + U_3 \\ X_4 &= \zeta. \end{aligned}$$

This is then the result of making the intervention $X_4 := \zeta$. ∘

Let us outline what has been achieved so far. We have defined the notion of a SEM and the solution to a SEM, and we have defined a notion of intervention in a SEM, resulting in a postintervention SEM and corresponding postintervention solution variables. The idea behind these concepts is the following. Consider a set of variables $(X_\alpha)_{\alpha \in V}$ solving a SEM with DAG G. We interpret this as meaning that the data-generating mechanism behind the variables is an algorithm which essentially evaluates the values of the variables recursively in the order of an topological ordering for G. In a more suggestive language, parent variables are the causes of their children, and so forth. This suggests that if we were to imagine making an exogeneous intervention in the system by, say, fixing X_β at the value x_β, the causal links from the parents, and more generally, the ancestors of X_β, would be broken, while the causal links from X_β to its children and descendants would remain intact. This is what the notion of intervention given in Definition 1.5.5 allows us to formalize, and Example 1.5.6 shows this in action.

A classical concrete example which illustrates how the framework of SEMs and SEM interventions parallel the functioning of causality in the real world is the sprinkler example, see p. 15 of [126]. The idea of the example is outlined in Figure 1.5.3.

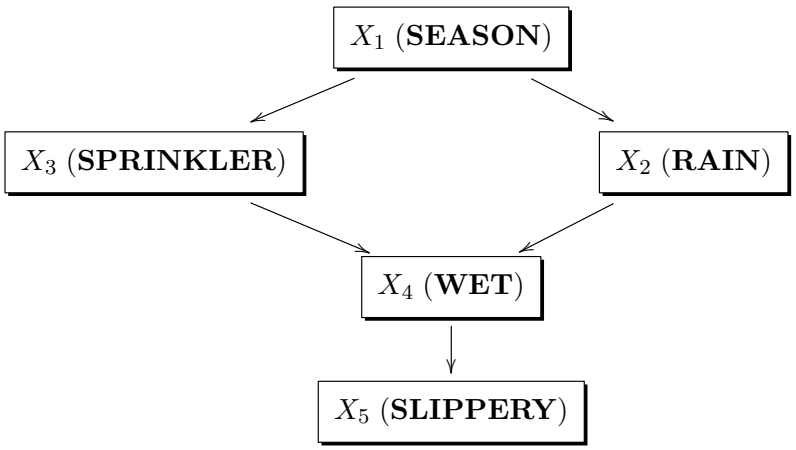


Figure 1.5.3: The DAG for the sprinkler example.

The figure uses a DAG G to depict the causal relationship between five variables:

- X_1 (**SEASON**): The season of the year
- X_2 (**RAIN**): Whether it is raining
- X_3 (**SPRINKLER**): Whether the sprinkler in the garden is turned on
- X_4 (**WET**): Whether the pavement is wet
- X_5 (**SLIPPERY**): Whether the pavement is slippery

Now imagine that the actual relationship between the five variables can be represented as a SEM with DAG G as described in Figure 1.5.3. This would mean, for example, that whether it is raining is a function of the season and noise, and that whether the pavement is wet is a function of whether the sprinkler is on, whether it is raining and noise. For the sake of the example, let us explicitly assume that

$$X_4 = X_2 \vee X_3 \vee U_4, \tag{1.39}$$

such that X_4 is equal to `true` when X_2 or X_3 is `true`, but X_4 may also be `true` due to other factors, represented by the noise variable U_4. We now imagine that we manually turn on the sprinkler, represented by X_3, and ask how we would expect the resulting system to behave. The natural answer is that the causal link from X_1 to X_3 would be broken and that X_4 would always be `true`. This precisely corresponds to what would be the result of an intervention in the SEM according to Definition 1.5.5.

The conclusion is that the notion of a SEM and interventions in a SEM based on its DAG is in accordance with how we ordinarily percieve the functioning of causality in real-world situations. Thus, we appear to have taken the first step towards a useful probabilistic model of causality.

While this is interesting in itself, we have not yet arrived at a statistically operational definition. Next, we outline some important results which are essential to turning the above notion of causality into a practically useful concept. The following three definitions are essential in this regard. Recall that a path in a directed graph is an unbroken series of vertices and edges such that no vertices are repeated except possibly the initial and terminal vertices.

Definition 1.5.7. Let G be a DAG on a vertex set V. Let p be a path in G. Let $C \subseteq V$. We say that p is blocked by C in G when it holds that one of the following is true:

(1). p contains a chain of the form $\alpha \to \gamma \to \beta$ where $\gamma \in C$.

(2). p contains a chain of the form $\alpha \leftarrow \gamma \leftarrow \beta$ where $\gamma \in C$.

(3). p contains a fork of the form $\alpha \leftarrow \gamma \to \beta$ where $\gamma \in C$.

(4). p contains a collider of the form $\alpha \to \gamma \leftarrow \beta$ with $(\mathrm{de}(\gamma) \cup \{\gamma\}) \cap C = \emptyset$.

Definition 1.5.8. Let A, B and C be three disjoint sets in V. We say that C d-separates A and B in G if C blocks every path between A and B in G.

Figure 1.5.4: Left: $\{1, 4\}$ and $\{3\}$ are d-separated by $\{2, 5\}$, but not by $\emptyset$, $\{2\}$ or $\{5\}$. Right: $\{1\}$ and $\{3\}$ are d-separated by $\emptyset$, but not by any nonempty subset of $\{2, 4, 5\}$.

Definition 1.5.9. Let $(X_\alpha)_{\alpha \in V}$ be a family of variables, and let G be a DAG. Let A, B and C be disjoint subsets of V.

1. If $X_A \perp\!\!\!\perp X_B \mid X_C$ whenever C d-separates A and B, then we say that $(X_\alpha)_{\alpha \in V}$ is global Markov with respect to G.
2. If C d-separates A and B whenever $X_A \perp\!\!\!\perp X_B \mid X_C$, we say that $(X_\alpha)_{\alpha \in V}$ is faithful with respect to G.

The intuitive meaning of Definition 1.5.9 is not clear at first sight. The following result makes it easier to understand what the global Markov property is. Similar to [126, 107], we use the following shorthand for densities: With p being the density of $(X_\alpha)_{\alpha \in V}$ and $A, B \subseteq V$, we let $p(x_A)$ denote the density of X_α in x_α, and let $p(x_A \mid x_B)$ denote a conditional density of X_A given $X_B = x_b$, evaluated in x_A.

Theorem 1.5.10 ([107], Section 3.2.2)**.** *Consider a family $(X_\alpha)_{\alpha \in V}$ of variables having a joint density with respect to a product measure. Assume given a DAG G on V. The following three properties are equivalent:*

(1). The joint density p satisfies $p(x) = \prod_{\alpha \in V} p(x_\alpha \mid \mathrm{pa}_G(\alpha))$ almost surely.

(2). With A, B and C disjoint, it holds that if C d-separates A from B, then $X_A \perp\!\!\!\perp X_B \mid X_C$.

(3). For all $\alpha \in V$, it holds that $X_\alpha \perp\!\!\!\perp X_{\mathrm{nd}(\alpha)} \mid X_{\mathrm{pa}(\alpha)}$.

The three properties of Theorem 1.5.10 are referred to as the Markov factorization property, the global Markov property and the local Markov property, respectively. In the case where the family of variables $(X_\alpha)_{\alpha \in V}$ under consideration has a joint density with respect to a product measure and satisfies one of these in this case

equivalent properties with respect to a DAG G, we simply say that $(X_\alpha)_{\alpha\in V}$ is Markov with respect to G. Theorem 1.5.10 shows that under the necessary regularity conditions, the global Markov property is equivalent to X_α being conditionally independent of its non-descendants given its parents. Thinking of such family relationships as corresponding to a notion of time, this is analogous to the classical Markov property for, say, discrete-time Markov chains.

Next, we recall two results which will help make causal inference a practical possibility.

Theorem 1.5.11 ([126], Theorem 1.4.1)**.** *Assume that $(X_\alpha)_{\alpha\in V}$ satisfies a set of structural equations with noise variables $(U_\alpha)_{\alpha\in V}$, DAG G and structural relationships $(f_\alpha)_{\alpha\in V}$, and assume that $(X_\alpha)_{\alpha\in V}$ has a density with respect to a product measure. If the noise variables are independent, $(X_\alpha)_{\alpha\in V}$ is Markov with respect to G.*

Theorem 1.5.12 ([126], Section 3.2.3)**.** *Assume that $(X_\alpha)_{\alpha\in V}$ satisfies a set of structural equations with noise variables $(U_\alpha)_{\alpha\in V}$, DAG G and structural relationships $(f_\alpha)_{\alpha\in V}$. Let $(Y_\alpha)_{\alpha\in V}$ be the solution to the postintervention SEM from doing $X_\beta := \zeta$. Assume that the noise variables are independent and that $(X_\alpha)_{\alpha\in V}$ has a density p with respect to a product measure $\lambda_V = \otimes_{\alpha\in V}\lambda_\alpha$. Then, for λ_β almost all ζ, $(Y_\alpha)_{\alpha\in V}$ has density with respect to the product measure $\tilde{\lambda}_V$ obtained by exchanging the β factor of λ_V with the Dirac measure in ζ, and the density is*

$$q(x) = 1_{(x_\beta=\zeta)} \prod_{\alpha\neq\beta} p(x_\alpha \mid x_{\mathrm{pa}_G(\alpha)}). \tag{1.40}$$

The following example shows what Theorem 1.5.11 and Theorem 1.5.12 state in a practical situation.

Example 1.5.13. Consider again the SEM of Example 1.5.4. Assume that the four noise variables $U_1, \ldots, U_4$ are independent and follow standard normal distributions. As we by the calculations in Example 1.5.4 have

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ -16 & -8 & 1 & 8 \\ -2 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{bmatrix}, \tag{1.41}$$

we find that $(X_1, \ldots, X_4)$ follows a multivariate normal distribution. In particular, $(X_1, \ldots, X_4)$ has a density p with respect to the Lebesgue measure. Theorem 1.5.11 and Theorem 1.5.10 then show that $(X_1, \ldots, X_4)$ satisfies the local Markov property with respect to the DAG of Example 1.5.4, leading to the decomposition

$$p(x) = p(x_1)p(x_2 \mid x_1)p(x_4 \mid x_1, x_2)p(x_3 \mid x_4). \tag{1.42}$$

Consider next the solution $(Y_1, \ldots, Y_4)$ to the postintervention SEM obtained by doing $X_4 := \zeta$, as in Example 1.5.6. By the results obtained in that example, it holds that

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 8\zeta \\ \zeta \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{bmatrix}. \tag{1.43}$$

In this case, the linear transformation of $(U_1, \ldots, U_4)$ is clearly singular, so the vector of variables $(Y_1, \ldots, Y_4)$ does not have a density with respect to the Lebesgue measure. However, $(Y_1, \ldots, Y_4)$ does have a density q with respect to $\lambda \otimes \lambda \otimes \lambda \otimes \delta_\zeta$, where λ denotes the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ and δ_ζ denotes the Dirac measure in ζ. Theorem 1.5.11 and Theorem 1.5.10 then show that $(X_1, \ldots, X_4)$ satisfies the local Markov property with respect to the DAG of Example 1.5.6, leading to

$$q(x) = q(x_1)q(x_4)q(x_2 \mid x_1)q(x_3 \mid x_4). \tag{1.44}$$

In order to relate this to Theorem 1.5.12, note that as $X_1 = Y_1$, we must have $q(x_1) = p(x_1)$. And as both $X_2 = 3X_1 + U_2$ and $Y_2 = 3Y_1 + U_2$ we also obtain $q(x_2 \mid x_1) = p(x_2 \mid x_1)$. Likewise, as $X_3 = 8X_4 + U_3$ and $Y_3 = 8Y_4 + U_3$, we find $q(x_3 \mid x_4) = p(x_3 \mid x_4)$. And as $Y_4 = \zeta$, we obtain $q(x_4) = 1_{(x_4=\zeta)}$, all in all leading to

$$q(x) = 1_{(x_4=\zeta)}p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_4), \tag{1.45}$$

in accordance with Theorem 1.5.12. ∘

Theorem 1.5.11 is a statement about what the distribution of the solution to a SEM is, given certain regularity conditions. This results opens up a pathway to causal inference. To see how this is the case, assume that $(X_\alpha)_{\alpha \in V}$ is a set of observed variables solving some SEM. This SEM represents the causal relationships between the variables. Observing the distribution of $(X_\alpha)_{\alpha \in V}$, we in particular observe all conditional independence relationships between these variables. By Theorem 1.5.11, $(X_\alpha)_{\alpha \in V}$ is Markov with respect to the true DAG G. This implies that we can reason, based solely on the distribution of the observed variables, that the true DAG G for the SEM of $(X_\alpha)_{\alpha \in V}$ must be in the set

$$\{G \mid G \text{ is a DAG such that } (X_\alpha)_{\alpha \in V} \text{ is Markov with respect to } G\}. \tag{1.46}$$

If this set is small, it is realistic to obtain nontrivial information about the causal relationships between the observed variables. In fact, making one more assumption allows us to go even further. Assume now that not only is $(X_\alpha)_{\alpha \in V}$ Markov with respect to the true DAG G, but $(X_\alpha)_{\alpha \in V}$ is also faithful to the true DAG G. The assumption of faithfulness is not an innocent one. In spite of faithfulness having probability one in certain contexts, see Theorem 3.2 of [161], the reasonability of

assuming faithfulnes is a topic under active debate, see [174, 22, 162]. Nevertheless, if we do assume faithfulness, we find that the true DAG must be in the set

$$\{G \mid X_A \perp\!\!\!\perp X_B \mid X_C \text{ if and only if } C \text{ } d\text{-separates } A \text{ from } B \text{ in } G\}. \tag{1.47}$$

In this case, all candidate DAGs G share the same d-separation properties. The equivalence class of DAGs all sharing the same d-separation properties can be characterized and the set of equivalence classes can be shown to be injectively embedded within a certain other class of graphs, called completed patterns by Verma and Pearl in [171], where a characterization of the equivalence classes was first proven. In newer literature, these completed patterns are often referred to as completed partially directed acyclic graphs (CPDAGs). Under the assumption of the Markov property and faithfulness, the true CPDAG is thus uniquely determined by the distribution of the observed variables, and is thus amenable to be inferred from the distribution. Causal inference for the CPDAG based on these ideas is carried out by for example Maathuis et al. in [113].

Several questions, however, remain unanswered. A first question is how we, if we actually know the true DAG G, can reason about the effect of possible interventions. In the case of independent error variables, this is clarified by Theorem 1.5.12. This theorem shows how the postintervention distribution depends on the observational distribution, that is, the distribution of the $(X_\alpha)_{\alpha\in V}$ before any intervention, and the DAG G. What is in fact more important is the objects absent in Theorem 1.5.12: Given the observational density p and the DAG G, the functional relationships (f_α) of the SEM are not necessary to obtain the postintervention density. In other words, if we know the DAG and the density, we can calculate postintervention distributions. The back-door and front-door adjustment criteria developed in Section 3.3 of [126] depend crucially on these results, as they essentially are results about conditions under which (1.40) simplifies. In the context of estimating intervention effects, this means that under the condition of the Markov property and faithfulness, the possible postintervention distributions are limited to distributions calculated using (1.40) for DAGs in the equivalence class of the CPDAG corresponding to the observational distribution. This is also dicsussed in [113].

One important conclusion from all this, however, is that even under the conditions of the Markov property and faithfulness, postintervention effects are not uniquely determined from the observational distribution. And if we do not assume the Markov property, in general corresponding to having error variables in the underlying SEM which are not independent, even the formula (1.40) breaks down. See, however, [166] for results on postintervention distributions in a semi-Markovian framework.

We have now discussed how Theorem 1.5.11 and Theorem 1.5.12 make causal inference possible. One final important observation is the following: In the above, we always assumed that there was some true underlying SEM for our observational variables $(X_\alpha)_{\alpha\in V}$, and we sought to infer the DAG G of this SEM and postintervention distributions for this SEM. However, several of our key concepts and results turned

out only to depend on the DAG G and the observational distribution. In particular, concepts depending only on the DAG G and the distribution are:

- The Markov property
- The faithfulness property
- The postintervention distribution

This means that these can be abstracted to a level relating only to the DAG and the distribution. In our formulation, this is already the case for the Markov and faithfulness properties of Definition 1.5.9. However, by these observations, we may also make the following definition.

Definition 1.5.14. Assume that $(X_\alpha)_{\alpha \in V}$ is a set of variables with density p with respect to the product measure $\lambda_V = \otimes_{\alpha \in V} \lambda$. Assume that $(X_\alpha)_{\alpha \in V}$ is Markov with respect to G. We then define the postintervention distribution of $(X_\alpha)_{\alpha \in V}$ for doing $X_\beta := \zeta$ with respect to G as the distribution with density with respect to the product measure $\tilde{\lambda}_V$ obtained by exchanging the β factor of λ_V with the Dirac measure in ζ, where the density is

$$q(x) = 1_{(x_\beta = \zeta)} \prod_{\alpha \neq \beta} p(x_\alpha \mid x_{\mathrm{pa}_G(\alpha)}). \tag{1.48}$$

See also Lauritzen's exposition [108] for more on this. It is natural in Definition 1.5.14 to restrict ourselves to distributions satisfying the Markov property, since this property is used when deriving Theorem 1.5.12 in the structural equation model context. With Definition 1.5.14, all the important concepts for the reasoning on causal inference carried through above, such as the estimation of the CPDAG and possible postintervention distributions, can be abstracted from the context of SEMs to the context of distributions and DAGs. This captures the fundamentals of an operational statistical theory of causal inference.

1.5.2 A framework for causality and interventions for SDEs

We are now ready to describe our efforts at capturing notions such as causality and interventions in the framework of stochastic differential equations (SDEs). Our work is detailed in Chapter 7 and Chapter 8. Our starting point is the idea proposed by Aalen et al. in Section 4.1 of [1], where it is suggested that in the SDE

$$\mathrm{d}X_t = \mathrm{d}B_t + AX_t \,\mathrm{d}t + \sigma \,\mathrm{d}W_t, \tag{1.49}$$

where B has finite variation and W is a multidimensional Brownian motion, interventions may be defined by putting certain elements of the matrix A equal to zero. Inspired by this, we begin by defining the notion of an intervention formally in a

semimartingale-based SDE framework. Let Z be a d-dimensional semimartingale, and let $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ be a continuous mapping, where $\mathbb{M}(p, d)$ denotes the space of real $p \times d$ matrices. Consider the stochastic differential equation

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t, \tag{1.50}$$

with initial condition X_0. This is formally a stochastic integral equation, see Chapter 7 for details. We refer to Z as the driving semimartingale and to a as the coefficient function. The following definition, replicated from Chapter 7, yields a notion of intervention in such an SDE. We use X^m to denote the m'th coordinate of the p-dimensional process X.

Definition 7.2.2. Consider some $m \leq p$ and $\zeta \in \mathbb{R}$. The stochastic differential equation arising from (1.50) under the intervention $X^m := \zeta$ is

$$\mathrm{d}X_t = b(X_{t-})\,\mathrm{d}Z_t, \tag{1.51}$$

where the initial condition is X_0^i for coordinates $i \neq m$ and ζ for the m'th coordinate, and $b : \mathbb{R}^p \to \mathbb{M}(p, d)$ is given by letting $b_{ij}(x) = a_{ij}(x)$ for $i \neq m$ and $b_{mj}(x) = 0$ for all $x \in \mathbb{R}^p$ and $j \leq d$.

By Definition 7.2.2, intervening takes an SDE as its argument and yields another SDE. It is important to note that intervening does not take an SDE solution and yield another SDE solution. This is similar to how interventions in SEMs as given in Definition 1.5.5 takes a SEM as its argument and yields another SEM, instead of taking a set of variables and yielding another set of variables. Intuitively, in this way, Definition 7.2.2 manages to include the causal mechanism and not only the resulting variables. It is not obvious that Definition 7.2.2 will yield results corresponding to real-world interventions. See, however, Example 7.2.1 for setup where this in fact is the case. See also Subsection 1.7.1 for another idea of a possible application of Definition 7.2.2. Example 1.5.15 shows how Definition 7.2.2 works in a concrete case.

Example 1.5.15. Consider the two-dimensional SDE given by

$$\mathrm{d}X_t^1 = (\theta_1 - X_t^1)\,\mathrm{d}Z_t^1 + \mathrm{d}Z_t^2, \tag{1.52}$$

$$\mathrm{d}X_t^2 = X_t^1\,\mathrm{d}Z_t^2, \tag{1.53}$$

with initial condition (X_0^1, X_0^2). This is an SDE of the type (1.50), with driving semimartingale Z and

$$a(x) = \begin{bmatrix} \theta_1 - x_1 & 1 \\ 0 & x_1 \end{bmatrix}. \tag{1.54}$$

The result of making the intervention $X^1 := \zeta$ is therefore an SDE of the type (1.51), with

$$b(x) = \begin{bmatrix} 0 & 0 \\ 0 & x_1 \end{bmatrix}, \tag{1.55}$$

and initial conditions (ζ, X_0^2). This yields the SDE

$$\mathrm{d}X_t^1 = 0, \tag{1.56}$$

$$\mathrm{d}X_t^2 = X_t^1 \,\mathrm{d}Z_t^2. \tag{1.57}$$

Note that any solution of this SDE will satisfy that $X_t^1 = \zeta$ for all $t \geq 0$ and $\mathrm{d}X_t^2 = \zeta \,\mathrm{d}Z_t^2$. If we instead make the intervention $X^2 := \zeta$, we obtain the SDE

$$\mathrm{d}X_t^1 = (\theta_1 - X_t^1)\,\mathrm{d}Z_t^1 + \mathrm{d}Z_t^2, \tag{1.58}$$

$$\mathrm{d}X_t^2 = 0. \tag{1.59}$$

with initial conditions (X_0^1, ζ). Here, any solution of this SDE satisfies $X_t^2 = \zeta$ for all $t \geq 0$ and $\mathrm{d}X_t^1 = (\theta_1 - X_t^1)\,\mathrm{d}Z_t^1 + \mathrm{d}Z_t^2$. We see that due to the presence of x_1 in the second row of (1.54) and the absence of x_2 in the first row of (1.54), there is an assymmetry in effect of interventions: Interventions $X^1 := \zeta$ influences X^2, but not vice versa. ∘

Note that Definition 7.2.2 yields a formalization of a particular type of intervention, namely where the value of a coordinate of a process is set to constant at all times. It is also possible to define interventions for example at a single timepoint, or interventions where the resulting intervened coordinate is set to another stochastic process, similar to the way various types of interventions are defined in [126].

As we saw in Example 1.5.15, the dependence structure of $a(x)$ on its argument x is important for the effect of interventions. This leads to the following definitions.

Definition 7.4.1. The signature of the SDE (1.50) is the graph S with vertex set $\{1, \ldots, p\}$ and an edge from i to j if it holds that there is k such that the mapping a_{jk} is not independent of the i'th coordinate.

Definition 7.4.2. We say that X^j is locally unaffected by X^i in the SDE (1.50) if there is no edge from i to j in the signature of (1.50).

The lack of independence for a_{jk} of the i'th coordinate referred to in Definition 7.4.1 is a shorthand for the following condition: That there exists $x \in \mathbb{R}^p$ such that $x_i \mapsto a_{jk}(x)$ is not constant. The signature for the SDE of Example 1.5.15 is depicted in Figure 1.5.5. Recalling the conclusions of Example 1.5.15, we also see why it is sensible to refer to the absence of edges in the signature as corresponding to local unaffectedness: The absence of an edge from 2 to 1 in Figure 1.5.5 corresponds to the intervention $X^2 := \zeta$ not affecting X^1.

Before proceeding, we fix some nomenclature. Assume that (1.50) and (1.51) have unique solutions for all interventions, by Theorem V.7 of [134], this is the case whenever the mapping a is Lipschitz. We refer to (1.50) as the observational SDE, to the solution of (1.50) as the observational process, and to the distribution of the solution of (1.50) as the observational distribution. We refer to (1.51) as the postintervention SDE, to the solution of (1.51) as the postintervention process and to the distribution of the solution to (1.51) as the postintervention distribution.

Figure 1.5.5: The signature for the two-dimensional SDE (1.52-1.53).

Our first objective is to understand Definition 7.2.2. We will do this by identifying a SEM such that interventions in this SEM in a limiting sense corresponds to interventions made according to Definition 7.2.2. To do so, consider the Euler scheme for (1.50) corresponding to a time endpoint $T > 0$ and step size Δ. Assume that $N = T/\Delta$ is an integer. The Euler scheme for (1.50) is the set of variables $(X^\Delta)^i_{t_k}$ for $i \leq p$ and $k \geq N$ where $t_k = k\Delta$ given by $X^\Delta_0 = X_0$ and

$$(X^\Delta_{t_k})^i = (X^\Delta_{t_{k-1}})^i + \sum_{j=1}^{d} a_{ij}(X^\Delta_{t_{k-1}})(Z^j_{t_k} - Z^j_{t_{k-1}}). \tag{1.60}$$

In other words, the Euler scheme is given by a discretization of (1.50). Under Lipschitz conditions on a, the Euler scheme always converges to the unique solution of (1.50) in an appropriate sense, see Theorem V.16 of [134]. We now endow this set of variables with noise variables, a DAG and a set of functional relationships corresponding to the natural order in which the Euler scheme is calculated. The formal definition is given in Definition 7.4.3. Here, we explain the definition by an example.

Example 1.5.16. Consider again the two-dimensional SDE (1.52)-(1.53) from Example (1.5.15). Its Euler scheme is given by the recursion

$$(X^\Delta)^1_{t_k} = (X^\Delta)^1_{t_{k-1}} + (\theta_1 - (X^\Delta)^1_{t_{k-1}})(Z^1_{t_k} - Z^1_{t_{k-1}}) + Z^2_{t_k} - Z^2_{t_{k-1}}, \tag{1.61}$$

$$(X^\Delta)^2_{t_k} = (X^\Delta)^2_{t_{k-1}} + (X^\Delta)^1_{t_{k-1}}(Z^2_{t_k} - Z^2_{t_{k-1}}), \tag{1.62}$$

and the DAG corresponding to its endowed SEM is shown in Figure 1.5.6. Note the relationship between this DAG and the signature of Figure 1.5.5. In Figure 1.5.6, we have also included the error variables of the SEM, which are the variables $Z_{t_k} - Z_{t_{k-1}}$ for $k = 1, \ldots, N$. The dotted directed edges are meant to indicate that for each k, $Z_{t_k} - Z_{t_{k-1}}$ is the error variable in the SEM corresponding to all of the variables $(X^\Delta)^i_{t_k}$, with functional relationships given by (1.61)-(1.62). In particular, several of the primary variables of the SEM have the same error variables. This implies that in general, the error variables are not independent, so the distribution of the family of variables $(X^\Delta)^i_{t_k}$ for $i = 1, \ldots, p$ and $k = 0, \ldots, N$ will not be Markov with respect to the DAG of the SEM.

The DAG in Figure 1.5.6 is constructed such that the directed edges correspond to the natural recursive calculation of the variables in the SEM when given the initial values X_0. In detail, the DAG is constructed as follows: As $(X^\Delta)^i_{t_k}$ generally depends

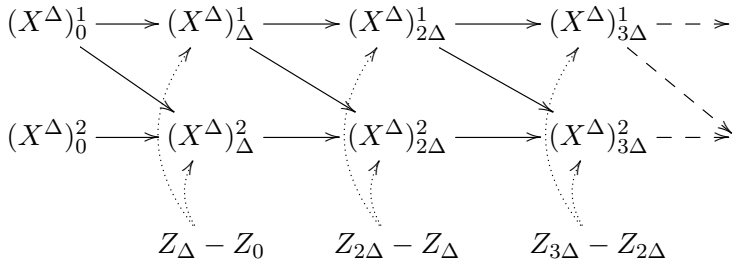


Figure 1.5.6: The DAG for the Euler SEM of (1.61-1.62).

on $(X^\Delta)^i_{t_{k-1}}$, there is always an arrow from $(X^\Delta)^i_{t_{k-1}}$ to $(X^\Delta)^i_{t_k}$. Furthermore, as $(X^\Delta)^2_{t_k}$ depends on $(X^\Delta)^1_{t_{k-1}}$, but $(X^\Delta)^1_{t_k}$ does not depend on $(X^\Delta)^2_{t_{k-1}}$, there is a directed edge from $(X^\Delta)^1_{t_{k-1}}$ to $(X^\Delta)^2_{t_k}$, but no directed edge from $(X^\Delta)^2_{t_{k-1}}$ to $(X^\Delta)^1_{t_k}$. ∘

In Lemma 7.4.5, we argue that the Euler SEM for postintervention SDEs of the type (1.51) obtained from applications of Definition 7.2.2 correspond to postintervention SEMs, recall Definition 1.5.5, obtained from intervening in all variables in the same row, meaning $(X^\Delta)^i_{t_k}$ for all k and some i, of the Euler SEM of (1.50). Essentially, this means that the notion of intervention obtained from Definition 7.2.2 corresponds to the limit of ordinary interventions in the Euler SEM. Now, the DAG of the Euler SEM, as depicted in Figure 1.5.6, has the following properties:

- All directed edges point strictly forward in time
- The driving semimartingale Z occurs only as an error variable

Heuristically speaking, this means that the notion of intervention given in Definition 7.2.2 is justified when:

- Causality propagates forward in time
- There are no instantaneous dependencies
- The driving semimartingale is not directly affected by interventions

This, then, constitutes a rough guide to when the notion of intervention given in Definition 7.2.2 is applicable. Some further points are worthy of notice. Since Definition 7.2.2 corresponds to intervention in the Euler SEM, assuming that Definition 7.2.2 is applicable corresponds to making an assumption on the true DAG of the system. However, this does not actually correspond to assuming that the true DAG is known. Rather, this corresponds to assuming that there is a fixed relationship between the true coefficient a in the SDE and the true DAG. This reduces estimation of the DAG to estimation of a.

Next, we state a theorem on identifiability of postintervention distributions of SDEs. The theorem will use some notions from the theory of Lévy processes and Markov processes, see Section 7.5 for an overview of this. Let D be a bounded neighborhood of zero in $\mathbb{R}^d$. Consider an SDE of the type

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t \tag{1.63}$$

where Z is a d-dimensional Lévy process with D-characteristic triplet (α, C, ν), $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ is Lipschitz and bounded and X_0 is some variable. Then, there exists a unique Feller semigroup (P_t) with the property that any solution of (1.63), independent of the initial distribution and the probability space on which the solution exists, is a Feller process with semigroup (P_t). This result is folklore, and is discussed formally in Lemma 7.5.1 and Section 7.8. Based on this result, it is possible to speak of the semigroup of an SDE such as (1.63).

Theorem 7.5.3. *Consider the SDEs*

$$\mathrm{d}X_t = a(X_{t-})\,\mathrm{d}Z_t, \tag{7.25}$$

$$\mathrm{d}Y_t = \tilde{a}(Y_{t-})\,\mathrm{d}\tilde{Z}_t, \tag{7.26}$$

where Z is a d-dimensional Lévy process, $\tilde{Z}$ is a $\tilde{d}$-dimensional Lévy process and the mappings $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ and $\tilde{a} : \mathbb{R}^p \to \mathbb{M}(p, \tilde{d})$ are Lipschitz and bounded. Assume that (7.25) and (7.26) have the same semigroup, and that the initial values have the same distribution. Then, the postintervention distributions of doing $X^m := \zeta$ in (7.25) and doing $Y^m := \zeta$ in (7.26) are equal for all m and $\zeta \in \mathbb{R}$.

Theorem 7.5.3 is proven in Section 7.8. The theorem shows that for two SDEs driven by Lévy processes having the same initial distribution and semigroup, the postintervention distributions are the same. This theorem is one of our main results on this subject.

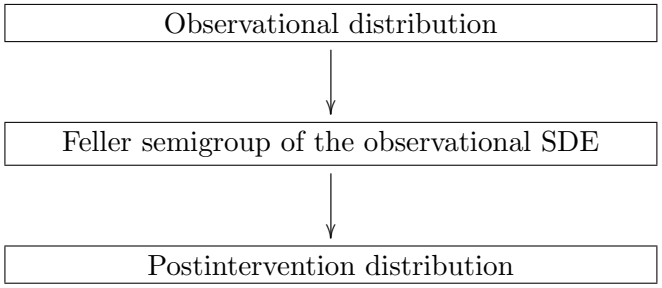


Figure 1.5.7: Line of inference for causality in SDEs.

The theorem points to the possibility of the line of inference depicted in Figure 1.5.7: Under sufficient regularity conditions such as appropriate notions of irreducibility,

the Feller semigroup of a Markov process is identifiable from the observational distribution, and Theorem 7.5.3 allows for deducing postintervention distributions from the Feller semigroup.

In the next subsections, we discuss the ramifications of Theorem 7.5.3. In particular, we discuss the line of inference proposed in Figure 1.5.7, wediscuss how intervention in SDEs work in practical cases, and we relate our theory to other theories of causality.

1.5.3 Discussion of the identifiability theorem

In this subsection, we discuss Theorem 7.5.3 and its consequences. An important first point in connection with this theorem is the following. Recall from our earlier discussion of DAG-based causal inference that postintervention distributions can be obtained from the DAG and the distribution as in Definition 1.5.14 when the underlying variables satisfy the global Markov property with respect to the DAG. As discussed in Example 1.5.16, the DAG of the Euler SEM is not generally Markov with respect to its DAG because of dependency between the error variables. Thus, Theorem 7.5.3 is a result on identifiability of postintervention distributions in the absence of the Markov property.

Another thing to note about Theorem 7.5.3 is the following. Recall that the directed edges of the Euler SEM for the SDE is determined by the signature of the SDE. Furthermore, by Definition 7.4.1, the signature S of the SDE is determined by the structure of how $a(x)$ depends on its arguments, in the sense that S contains a directed edge from i to j if it holds that there is k such that the mapping a_{jk} is not independent of the i'th coordinate. Now consider the particular case of an SDE of the type

$$\mathrm{d}X_t = a(X_t)\,\mathrm{d}W_t, \tag{1.64}$$

where $a : \mathbb{R}^p \to \mathbb{M}(p,p)$ and W is a p-dimensional standard Brownian motion. The distribution of the solution X then depends only on $a(x)$ through $a(x)a(x)^t$. In particular, if $\tilde{a}(x) = a(x)p(x)$ for some function p with $p(x)p(x)^t$ being the identity matrix for all x, then $\tilde{a}(x)\tilde{a}(x)^t = a(x)p(x)p(x)^t a(x)^t = a(x)a(x)^t$, and so the distribution of the solution to the SDE

$$\mathrm{d}X_t = \tilde{a}(X_t)\,\mathrm{d}W_t, \tag{1.65}$$

is the same as that of the solution to (1.64), even though the signature corresponding to $\tilde{a}$ may be different from the one corresponding to a. This is illustrated in Example 1.5.17.

Example 1.5.17. Consider the mapping $a : \mathbb{R}^2 \to \mathbb{M}(2,2)$ defined by putting

$$a(x) = \begin{bmatrix} x_1 & 0 \\ x_2^2/\sqrt{x_1^2+x_2^2} & -x_1x_2/\sqrt{x_1^2+x_2^2} \end{bmatrix} \tag{1.66}$$

whenever x is not zero, and $a(0) = 0$. This mapping satisfies

$$\begin{aligned} a(x)a(x)^t &= \begin{bmatrix} x_1 & 0 \\ x_2^2/\sqrt{x_1^2+x_2^2} & -x_1x_2/\sqrt{x_1^2+x_2^2} \end{bmatrix} \begin{bmatrix} x_1 & x_2^2/\sqrt{x_1^2+x_2^2} \\ 0 & -x_1x_2/\sqrt{x_1^2+x_2^2} \end{bmatrix} \\ &= \begin{bmatrix} x_1^2 & x_1x_2^2/\sqrt{x_1^2+x_2^2} \\ x_1x_2^2/\sqrt{x_1^2+x_2^2} & x_2^2 \end{bmatrix} \end{aligned} \tag{1.67}$$

whenever $x \neq 0$. We will construct another mapping $\tilde{a}$ which has a different signature from a, but which has the same transpose product as a, in the sense of having $\tilde{a}(x)\tilde{a}(x)^t = a(x)a(x)^t$. To do so, define $p : \mathbb{R}^2 \to \mathbb{M}(2,2)$ by

$$p(x) = \frac{1}{\sqrt{x_1^2+x_2^2}} \begin{bmatrix} x_1 & x_2 \\ x_2 & -x_1 \end{bmatrix}, \tag{1.68}$$

for $x \neq 0$ and let $p(0)$ be the identity matrix. Put $\tilde{a}(x) = a(x)p(x)$. We then obtain $\tilde{a}(0) = a(0) = 0$ and

$$\begin{aligned} \tilde{a}(x) &= \frac{1}{\sqrt{x_1^2+x_2^2}} \begin{bmatrix} x_1 & 0 \\ x_2^2/\sqrt{x_1^2+x_2^2} & -x_1x_2/\sqrt{x_1^2+x_2^2} \end{bmatrix} \begin{bmatrix} x_1 & x_2 \\ x_2 & -x_1 \end{bmatrix} \\ &= \frac{1}{\sqrt{x_1^2+x_2^2}} \begin{bmatrix} x_1^2 & x_1x_2 \\ 0 & (x_2^3+x_1^2x_2)/\sqrt{x_1^2+x_2^2} \end{bmatrix} \\ &= \begin{bmatrix} x_1^2/\sqrt{x_1^2+x_2^2} & x_1x_2/\sqrt{x_1^2+x_2^2} \\ 0 & x_2 \end{bmatrix}. \end{aligned} \tag{1.69}$$

Note that the first row of a depends only on the first coordinate, while the second row depends on both coordinates. On the other hand, the first row of $\tilde{a}$ depends on both coordinates, while the second row of $\tilde{a}$ depends only on the second coordinate. This translates into a and $\tilde{a}$ corresponding to different signatures, shown in Figure 1.5.8.

Figure 1.5.8: Left: The signature corresponding to the coefficient function a. Right: The signature corresponding to the coefficient function $\tilde{a}$.

As $p(x)$ is orthonormal for all x, it holds that $\tilde{a}(x)\tilde{a}(x)^t = a(x)a(x)^t$ and so the solutions to the two SDEs

$$\mathrm{d}X_t = a(X_t)\,\mathrm{d}W_t \tag{1.70}$$

$$\mathrm{d}X_t = \tilde{a}(X_t)\,\mathrm{d}W_t \tag{1.71}$$

have the same distribution. Thus, we have explicitly constructed two SDEs with the same solution distributions but with different signatures. Note now that the intervention $X^2 := \zeta$ in (1.70) yields an SDE where the first coordinate satisfies

$$\mathrm{d}X_t^1 = X_t^1 \, \mathrm{d}W_t^1 \tag{1.72}$$

while the intervention $X^2 := \zeta$ in (1.71) yields an SDE where the first coordinate satisfies

$$\mathrm{d}X_t^1 = \frac{(X_t^1)^2}{\sqrt{(X_t^1)^2 + \zeta^2}} \, \mathrm{d}W_t^1 + \frac{X_t^1 \zeta}{\sqrt{(X_t^1)^2 + \zeta^2}} \, \mathrm{d}W_t^2 \tag{1.73}$$

The distribution of the solution to (1.73) is a Markov process with generator

$$Af(x) = \frac{x^4 + (x\zeta)^2}{x^2 + \zeta^2} \frac{\partial^2 f}{\partial x^2}(x) = x^2 \frac{\partial^2 f}{\partial x^2}(x),$$

which is the generator of a geometric Brownian motion with zero drift. This is the same as the generator of the solution to (1.72). Thus, as required in Theorem 7.5.3, the postintervention distributions are the same, even in this case where the signatures are different. ∘

Example 1.5.17 illustrates a rather remarkable fact: For SDE models, the postintervention distributions are identifiable from the observational distribution, even when the signature and thus the resulting DAGs of the Euler SEMs are not identifiable from the observational distribution. One interpretation of this is that for SDEs, the postintervention distributions will be the same for all signatures and thus all resulting DAGs which are compatible with the observational distribution. From this perspective, and in concordance with Theorem 7.5.3, the agreement of the two postintervention distributions in Example 1.5.17 is not so much related to the dependence structure of $a(x)$, but rather on the dependence structure of $a(x)a(x)^t$, or equivalently, $\tilde{a}(x)\tilde{a}(x)^t$.

Some observations which might give a hint as to what is happening in Example 1.5.17 are the following. Considering the first two columns of the Euler SEMs corresponding to the two SDEs (1.70) and (1.71), we have constructed SEMs with DAGs as in Figure 1.5.9.

The SEMs are constructed such that both the distributions and the postintervention distributions of the variables are the same, in spite of the DAGs being different. In particular, recalling the functional relationships of these SEMs, see Definition 7.4.3 for details, the conditional distribution of X_Δ^Δ given $X_0^\Delta = x$ with $\Delta = 1$ will be a normal distribution with mean x and variance

$$\Sigma = \begin{bmatrix} x_1^2 & x_1 x_2^2 / \sqrt{x_1^2 + x_2^2} \\ x_1 x_2^2 / \sqrt{x_1^2 + x_2^2} & x_2^2 \end{bmatrix}. \tag{1.74}$$

Figure 1.5.9: Left: First two columns of the Euler SEM corresponding to the cofficient function a of Example 1.5.17. Right: First two columns of the Euler SEM corresponding to the cofficient function $\tilde{a}$ of Example 1.5.17.

Therefore, the conditional distribution of $(X^\Delta)_\Delta^2$ given $(X^\Delta)_0 = x$ is a normal distribution with mean x_2 and variance x_2^2, and likewise, the conditional distribution of $(X^\Delta)_\Delta^1$ given $(X^\Delta)_0 = x$ is a normal distribution with mean x_1 and variance x_1^2. This implies that $(X^\Delta)_\Delta^2 \perp\!\!\!\perp (X^\Delta)_0^1 \mid (X^\Delta)_0^2$ and $(X^\Delta)_\Delta^1 \perp\!\!\!\perp (X^\Delta)_0^2 \mid (X^\Delta)_0^1$. However, in the left graph of Figure 1.5.9, $(X^\Delta)_\Delta^2$ and $(X^\Delta)_0^1$ are not d-separated by $(X^\Delta)_0^2$, and in the right graph of Figure 1.5.9, $(X^\Delta)_\Delta^1$ and $(X^\Delta)_0^2$ are not d-separated by $(X^\Delta)_0^1$. Thus, neither of the two DAGs of Figure 1.5.9 are faithful to the underlying distribution. The non-uniqueness of the signatures in Example 1.5.17 appears to be related to this lack of faithfulness.

Theorem 7.5.3 has the following important consequence: As postintervention distributions in the context of Lévy driven SDEs only depend on the initial condition and the semigroup, we can abstract the notion of interventions from SDEs to Lévy diffusion distributions, in the same way as we earlier in the context of distributions having the Markov property abstracted the notion of intervention in a SEM to the notion of intervention in a distribution and a DAG, see Definition 1.5.14. We now outline how this can be done. In the following, let $C_0(\mathbb{R}^p)$ denote the space of continuous functions from $\mathbb{R}^p$ to $\mathbb{R}$ vanishing at infinity, and let $C_0^2(\mathbb{R}^p)$ denote the subset of twice continuously differentiable elements of $C_0(\mathbb{R}^p)$.

Definition 1.5.18. Let $A : C_0^2(\mathbb{R}^p) \to C_0(\mathbb{R}^p)$ be linear. We say that A is a Lévy diffusion operator if it can be written as

$$\begin{aligned} Af(x) = &\sum_{i=1}^{p}\sum_{j=1}^{d} a_{ij}(x)\alpha_j \frac{\partial f}{\partial x_i}(x) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(a(x)Ca(x)^t)_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &+ \int f(x + a(x)y) - f(x) - 1_D(y)\sum_{i=1}^{p}\frac{\partial f}{\partial x_i}(x)\sum_{j=1}^{d} a_{ij}(x)y_j \,\mathrm{d}\nu(y), \end{aligned} \tag{1.75}$$

for $f \in C_0^2(\mathbb{R}^p)$ and $x \in \mathbb{R}^p$, where D is a bounded neighborhood of zero in $\mathbb{R}^d$, (α, C, ν) by a Lévy triplet and let $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ is Lipschitz and bounded.

Definition 1.5.19. Let $(P_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p}$ be a family of transition probabilities on the measurable space $(\mathbb{R}^p, \mathcal{B}_p)$. We say that $(P_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p}$ is a Lévy diffusion family if it is a Feller semigroup whose generator is of the type (1.75). If μ is a probability distribution on $(\mathbb{R}^p, \mathcal{B}_p)$ and $(P_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p}$ is a Lévy diffusion family, we say that $(\mu, (P_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p})$ is a Lévy diffusion pair.

In the above, $\mathcal{B}_p$ denotes the Borel-σ-algebra on $\mathbb{R}^p$. In general, the distributions of the solutions to SDEs such as (1.63) will be Feller processes with transition probabilities of the type given in Definition 1.5.19. For notational convenience, we introduce the sample space $D([0,\infty), \mathbb{R}^p)$ of càdlàg paths from $[0,\infty)$ to $\mathbb{R}^p$ and let X° denote the identity on this space. Also, we write $A(D, \alpha, C, \nu, a)$ for the operator defined by (1.75).

Definition 1.5.20. Let $(\mu, (P_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p})$ be a Lévy diffusion pair, where the family of transition probabilities has generator whose restriction to $C_0^2(\mathbb{R}^p)$ is given as $A(D, \alpha, C, \nu, a)$, see (1.75). Let $m \leq p$ and $\zeta \in \mathbb{R}^p$. The postintervention Lévy diffusion pair $(\nu, (Q_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p})$ resulting from the intervention $(X^\circ)^m := \zeta$ is given by letting $\nu = \pi(\mu)$, where the mapping $\pi : \mathbb{R}^p \to \mathbb{R}^p$ inserts ζ on the m'th coordinate, and letting $(Q_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p}$ be the unique family of Feller transition probabilities with generator B whose restriction to $C_0^2(\mathbb{R}^p)$ is equal to $A(D, \alpha, C, \nu, b)$, where $b(x)$ is obtained by exchanging the m'th row of $a(x)$ with zeroes.

The uniqueness required for Definition 1.5.20 follows from the results in Section 7.8. Definition 1.5.20 yields a notion of intervention for Lévy diffusions. Given a Lévy diffusion pair $(\mu, (P_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p})$, it holds due to Theorem 7.5.3 that the postintervention Lévy diffusion pair can be interpreted as the distributional result of considering an arbitrary Lévy driven SDE with initial distribution μ and semigroup $(P_t(x,\cdot))_{t\geq 0, x\in\mathbb{R}^p}$ and making an intervention according to Definition 7.2.2. In this framework, we may thus observe a Lévy diffusion, estimate the initial distribution and semigroup, and consider questions such as what the distributional effects would be of making interventions of the type $(X^\circ)^m := \zeta$, and so forth, corresponding to the latter half of the line of reasoning of Figure 1.5.7. This concludes our discussion of Theorem 7.5.3.

1.5.4 An example based on Ornstein-Uhlenbeck processes

Next, we outline an example discussed in detail in Chapter 8. Consider a three-dimensional Ornstein-Uhlenbeck process of the form

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \mathrm{d}W_t, \tag{1.76}$$

where W is a three-dimensional Brownian motion and B is upper diagonal, meaning that we have

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ 0 & b_{22} & b_{23} \\ 0 & 0 & b_{33} \end{bmatrix}. \tag{1.77}$$

We assume that the diagonal elements of B are all negative. As B is triangular, the set of diagonal elements of B is equal to the set of eigenvalues of B. Therefore, B is stable, meaning that all eigenvalues have strictly negative real parts. Furthermore, all principal submatrices are stable as well. The interpretation of having B upper diagonal is that the levels of all of X^1, X^2 and X^3 influence the average change in X^1, while only the levels of X^2 and X^3 influence the average change in X^2 and only X^3 influences the average change in X^3. This is illustrated in Figure 1.5.10, where the signature of (1.76) is depicted.

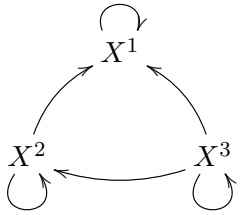


Figure 1.5.10: The signature of (1.76).

As both B and its principal submatrices are stable, there exist stationary distributions both for the observational distribution of X and for the postintervention distributions. All these stationary distributions are normal distributions. The equations below show the stationary means for the case of no intervention, as well as interventions in the second and third coordinates.

$$\text{No intervention:} \quad \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} \tag{1.78}$$

$$X^2 := \zeta \text{ intervention:} \quad \begin{bmatrix} A_1 - \frac{B_{12}}{B_{11}}(\zeta - A_2) \\ \zeta \\ A_3 \end{bmatrix} \tag{1.79}$$

$$X^3 := \zeta \text{ intervention:} \quad \begin{bmatrix} A_1 - \left(\frac{B_{13}}{B_{11}} - \frac{B_{12}B_{23}}{B_{11}B_{22}}\right)(\zeta - A_3) \\ A_2 - \frac{B_{23}}{B_{22}}(\zeta - A_3) \\ \zeta \end{bmatrix}. \tag{1.80}$$

This shows not only that the stationary means in this case lend themselves to closed-form calculation, but also that the results agree with intuition: For example, in the case of the intervention $X^2 := \zeta$, the stationary mean of the first coordinate changes based on the influence of X^2 on X^1 as illustrated in Figure 1.5.10, but there is no

change in the stationary mean of the third coordinate, since there is no influence of X^2 on X^3.

1.5.5 Relationship with weak conditional local independence

Recall that a notion of influence, referred to as weak conditional local independence (WCLI) between two processes in a class of semimartingales is introduced by Gégout-Petit and Commenges in [27, 57]. In this section, we briefly outline a connection between our results and WCLI.

The notion of WCLI builds on the notion of semimartingale characteristics. For an outline of semimartingale characteristics, see Chapter II of [83] or Chapter 7. In [57], the following definition of weak conditional local independence is made.

Definition 1.5.21 ([57], Section 2)**.** We define $\mathcal{D}'$ to be the class of p-dimensional special semimartingales X having decomposition $X = X_0 + M + A$, where M is a square integrable martingale and A is a predictable process with paths of bounded variation, having characteristics (B, C, ν) satisfying that C is a deterministic diagonal matrix.

Definition 1.5.22 ([57], Definition 2)**.** Let X be in $\mathcal{D}'$. We say that X^k is WCLI of X^j if and only if the characteristics B^k and ν^k of X^k are $(\mathcal{F}_t^{-j})_{t\geq 0}$ predictable, where $(\mathcal{F}_t^{-j})_{t\geq 0}$ is the filtration generated by the processes $X^1, \ldots, X^p$ excepting X^j.

The definition is formally well-posed for all special semimartingales. However, [57] chooses to restrict themselves to the class $\mathcal{D}'$ to ensure the interpretability of WCLI. This is explained in Remark 1 of [27], where the following example is considered.

Example 1.5.23. Assume that (X^1, X^2) solves

$$\mathrm{d}X_t^1 = a\,\mathrm{d}t + b\,\mathrm{d}W_t^1 \tag{1.81}$$

$$\mathrm{d}X_t^2 = X_t^1\,\mathrm{d}t + \exp(X_t^1)\,\mathrm{d}W_t^2, \tag{1.82}$$

where (W^1, W^2) is a two-dimensional Brownian motion. For this process, we have that the quadratic covariations of the continuous martingale parts are

$$C = \begin{bmatrix} b^2 t & b\int_0^t \exp(X_s^1)\,\mathrm{d}s \\ b\int_0^t \exp(X_s^1)\,\mathrm{d}s & \int_0^t \exp(2X_t^1)\,\mathrm{d}s \end{bmatrix}, \tag{1.83}$$

which is neither diagonal nor deterministic, so (X^1, X^2) is not in $\mathcal{D}'$. However, we may still ask whether for example B^2 and ν^2 are $(\mathcal{F}_t^{-1})_{t\geq 0}$ predictable. We will argue that the answer is yes. To see this, first note that as X^2 is continuous, the compensator ν^2 of the jump measure of X^2 is zero. Therefore, it is $(\mathcal{F}_t^{-1})_{t\geq 0}$ predictable. The predictable finite variation part B^2 of X^2 is $B_t^2 = \int_0^t X_s^1\,\mathrm{d}s$. Now

note that $(\mathcal{F}_t^{-1})$ is the filtration generated by X^2, so $[X^2]$ is $(\mathcal{F}_t^{-1})_{t\geq 0}$ adapted. Thus, the process $\int_0^t \exp(2X_t^1)\,\mathrm{d}s$ is $(\mathcal{F}_t^{-1})$ adapted, and so the process $\exp(2X_t^1)$ is $(\mathcal{F}_t^{-1})$ adapted, finally yielding that X^1 is $(\mathcal{F}_t^{-1})$ adapted. As B^2 only depends on X^1, this yields that B^2 is $(\mathcal{F}_t^{-1})_{t\geq 0}$ adapted. All in all, we find that if we extend Definition 1.5.22 to (1.81-1.82), X^2 is WCLI of X^1. ∘

In Example 1.5.23, it appears that Definition 1.5.22 does not provide an accurate view of the dependence structure of (X^1, X^2), since we would not want to say that X^2 is WCLI of X^1. The problem is basically that X^1 can be reconstructed through the quadratic variation of X^2, and as a consequence, the measurability properties of B^2 and ν^2 do not provide an accurate image of the structure of the dynamics of (X^1, X^2).

Nonetheless, for our purposes, we will use Definition 1.5.22 for all special semimartingales. This allows us to obtain the following result. Recall that the property of being locally unaffected was given in Definition 7.4.2, stated above.

Theorem 7.6.1. *Let X be the solution to (1.50). Assume that X is a special semimartingale and that Z is a Lévy process. If X^i is locally unaffected by X^m in (1.50), then X^i is WCLI of X^m.*

Theorem 7.6.1 is proven in Section 7.9. Intuitively, Theorem 7.6.1 states that having X^i locally unaffected by X^m yields that X^i in a sense is locally independent of X^m, a notion made precise by having the characteristics of X^i $(\mathcal{F}_t^{-m})$ predictable.

The above discussion, in particular Example 1.5.23, also highlights the benefits of our notion of causality and intervention for SDEs. By working in an SDE framework, we gain access to a candidate SEM in which interventions can be defined and a notion of causality abstracted: The SDE is more than the distribution, and so we have more powerful tools than the semimartingale characteristics at our disposal. This means that we are able to work with semimartingales without orthogonal martingale parts, as Definition 1.5.21 otherwise restricts us to. In Example 7.2.1, we see the importance of this: Natural situations occur where we observe processes without orthogonal martingale parts.

1.5.6 Relationship with DAG-based causal inference

Finally, we relate our notions of causality and interventions for stochastic differential equations to the DAG-based theory. In the DAG-based theory of causal inference, the DAG is the carrier of information about causality. For SDEs, the causal relationships between the coordinates of the SDE are summarized in the signature of the SDE. We are interested in investigating whether there is a connection between the signature and some appropriately chosen corresponding DAG.

To set the scene, we note that in several applications of DAG-based causal inference, the data consist of observations from a continuous-time system assumed to be in

equilibrium, see for example [112, 164, 120, 137]. This can in a natural sense be translated into assuming that the underlying continuous-time system follows an SDE, and that we observe samples from the stationary distribution. In the DAG-based theory, we seek to infer the true DAG. This leads to the following question.

Question 1.5.24. *When the true SDE has signature S, with respect to what DAGs is the stationary distribution global Markov?*

We are not able to give a full answer to this question, but we will make a few considerations which elucidate the problem. We first make the simplifying assumption that the underlying SDE is a p-dimensional Ornstein-Uhlenbeck process as in (1.76). In this case, the signature S has an edge from i to j if and only if B_{ji} is nonzero.

Lemma 1.5.25. *If the Ornstein-Uhlenbeck process has a stationary distribution, then the signature S is not a DAG.*

Proof. Assume contrarily that S is a DAG. Then S has a node i with no parents. This implies that the i'th row of B only contains zeroes. As a consequence, B is not stable, and so there is no stationary distribution. □

According to Lemma 1.5.25, we cannot expect that the true signature is a DAG. This rules out the proposition that when S is a DAG, the stationary distribution is global Markov with respect to S. We may instead consider a related question.

Question 1.5.26. *Consider the case where we may obtain a DAG G from the signature S by removing all loops (that is, edges with the same initial and terminal vertex). Is the stationary distribution Markov with respect to S?*

We provide an example to show that the answer is negative. Consider a three-dimensional Ornstein-Uhlenbeck process with identity diffusion matrix, mean reversion level A and mean reversion speed matrix given by

$$B = \begin{bmatrix} -1 & -1 & 0 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{bmatrix}. \tag{1.84}$$

The signature of this process is illustrated in Figure 1.5.11. We see that the result G of removing all loops from the signature in fact is a DAG.

As B is stable, there exists a unique stationary distribution, which is a normal distribution with mean A and variance Σ given as the solution to $B\Sigma + \Sigma B^t = -I$, see [80]. This is a Lyapounov equation with the solution and inverse of the solution given by

$$\Sigma = \begin{bmatrix} \frac{15}{16} & -\frac{7}{16} & \frac{1}{8} \\ -\frac{7}{16} & \frac{3}{4} & -\frac{1}{4} \\ \frac{1}{8} & -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \quad \text{and} \quad \Sigma^{-1} = \begin{bmatrix} \frac{160}{109} & \frac{96}{109} & \frac{8}{109} \\ \frac{96}{109} & \frac{232}{109} & \frac{92}{109} \\ \frac{8}{109} & \frac{92}{109} & \frac{262}{109} \end{bmatrix}. \tag{1.85}$$

Figure 1.5.11: Left: The signature S of the Ornstein-Uhlenbeck process under consideration. Right: The DAG G obtained by removing loops from the signature.

Now let Z be a variable following a normal distribution with mean A and variance Σ. Let $i, j \in \{1, 2, 3\}$ with $i \neq j$ and let k be the unique element of $\{1, 2, 3\} \setminus \{i, j\}$. By Proposition 5.2 of [107], it holds that Z_i and Z_j are conditionally independent given Z_k if and only if $(\Sigma^{-1})_{ij} = 0$. If Z were Markov with respect to the DAG G, it should hold that X_1 and X_3 are conditionally independent given X_2. However, $(\Sigma^{-1})_{13}$ is nonzero. Thus, the stationary distribution is not global Markov with respect to G, and the answer to Question 1.5.26 is negative.

1.6 Identifiability and ICA

In this section, we outline our work on identifiability in ICA models. We begin by outlining the LiNGAM method and ICA. After this, we describe our identifiability result and consider the numerical evaluation of our result.

1.6.1 LiNGAM, ICA and problem statement

As mentioned in Section 1.2, our work on ICA was inspired by open problems related to the LiNGAM method for causal discovery mentioned at a course given at the Seminar für Statistik at ETH Zürich. The LiNGAM method was introduced by Shimizu et al. in [156]. The fundamental idea of that paper is as follows. As in Section 1.5, consider a SEM $(G, (U_i), (f_i))$, with $i = 1, \ldots, p$, corresponding to the vertex set $V = \{1, \ldots, p\}$, and consider a solution $X = (X_1, \ldots, X_p)$ to the SEM. In general, the DAG G is not identifiable from the distribution of X. In order to overcome this, [156] considers what is known as a restricted structural equation model. Working in the context of restricted structural equation models essentially corresponds to restricting the functions f_i allowed in the SEM, see for example [128, 73, 180] for more on this. In [156], the restriction made is that all the functions f_i are assumed to be linear and that the coefficient of f_i corresponding to the error term is assumed to be equal to one. The resulting solution to the SEM then satisfies

a set of equations of the form

$$X = BX + U, \tag{1.86}$$

where B is a $p \times p$ matrix. The parental dependency structure of the f_i forces a certain acyclic structure on B, namely that $B_{ij} = 0$ whenever j is not a parent of i in the DAG G. This corresponds to the existence of a permutation of the rows and columns of B resulting in a lower triangular matrix, see again [156]. In Appendix A of [156], it is argued that when the components of U are independent and non-Gaussian, the acyclic structure of B ensures that the DAG can be uniquely identified from the distribution of X. This is in contrast to the general case discussed in Section 1.5, and is the essential benefit of considering SEMs with linear functions f_i. The results in [156] relies on rewriting (1.86) to

$$X = AU \tag{1.87}$$

with $A = (I - B)^{-1}$. In the case where the coordinates of U are independent and non-Gaussian, a result in [28] states that A is identifiable from the distribution of X up to scaling and permutation of columns. This result, however, relies essentially on the assumption of non-Gaussianity. A natural question is therefore: How difficult is it to estimate A in (1.87) from the distribution of X when the coordinates of U are non-Gaussian but close to Gaussian? In its abstract form, this question is related neither to LiNGAM nor to causal inference. Rather, it is related to the problem of estimation of A in the model

$$X = A\varepsilon \tag{1.88}$$

where ε has independent nondegenerate coordinates. The model (1.88) is known as the independent component analysis (ICA) model, and in this context, A is known as the mixing matrix. As a special case of the results given in [50], we obtain the following for invertible A and B:

- If all coordinates of ε are standard Gaussian, $A\varepsilon$ and $B\varepsilon$ have the same distribution precisely when $AA^t = BB^t$.
- If no coordinates of ε are Gaussian, $A\varepsilon$ and $B\varepsilon$ have the same distribution whenever $A = B\Lambda P$ for some invertible diagonal matrix Λ and a permutation matrix P.

We express this by saying that in the case of Gaussian errors, the mixing matrix is identifiable from the distribution of X up to transpose products, while in the case of non-Gaussian errors, the mixing matrix is identifiable from the distribution of X up to scaling and column permutation.

Now consider the case where we only have finitely many samples $X_1, \ldots, X_n$ from the distribution of X, and the error distribution is non-Gaussian but close to Gaussian.

Having only finitely many samples, our data might as well appear to be from a distribution with or without Gaussian errors. It is therefore a priori unclear which of the two above scenarios are reflected in our ability to identify the mixing matrix. This leads to the following question.

Question 1.6.1. *Consider the statistical model of distributions of the form $A\varepsilon$, where ε has independent nondegenerate coordinates, and the true distribution of ε is non-Gaussian but close to Gaussian. Assume given a true mixing matrix A and samples $X_1, \ldots, X_n$ from the true distribution. Will our ability to estimate the mixing matrix based on $X_1, \ldots, X_n$ primarily be up to transpose products or up to scaling and permutation of columns?*

Formal analysis of the ICA model can be done in several ways. Define, for $A \in \mathbb{M}(p,p)$, a mapping $L_A : \mathbb{R}^p \to \mathbb{R}^p$ by $L_A(x) = Ax$. We may then consider the statistical model

$$\{L_A(R) \mid A \in \mathbb{M}(p,p), R \in \mathcal{P}(p)\}, \tag{1.89}$$

where $\mathcal{P}(p)$ is the set of probability measures on $(\mathbb{R}^p, \mathcal{B}_p)$ with independent nondegenerate coordinates. This model is semiparametric. Alternatively, we may fix $R \in \mathcal{P}(p)$ and consider the statistical model

$$\{L_A(R) \mid A \in \mathbb{M}(p,p)\}, \tag{1.90}$$

which is parametric. Work on quantifying the difficulty of identifying A from finitely many samples of the distribution of X has been done by several authors, see for example [5, 23, 77, 125]. In particular, in [23], Chen and Bickel presents an algorithm for estimation of the mixing matrix in (1.89), and the asymptotic variance is calculated. In [125], a Cramér-Rao lower bound is calculated by Ollila et al. for the model (1.90) under certain restrictions on the error distribution.

We follow [125] by restricting our attention to (1.90). Furthermore, for simplicity, we consider the case where all the coordinates of the error distribution have the same distribution. In other words, we fix a nondegenerate distribution ζ on $(\mathbb{R}, \mathcal{B})$, define $R = \zeta^{\otimes p}$ and consider the statistical model (1.90). In order to guide our thinking, we first remark on the case where we do not have finitely many samples, but where instead the full distribution of $L_A(\zeta^{\otimes p})$ is known. Because we have fixed a probability measure ζ, some of the scaling indeterminacy for identifiability of the mixing matrix vanishes, and the remaining indeterminacy can be split into two cases. In the following, F^A denotes the cumulative distribution function of $L_A(\zeta^{\otimes p})$. From the results in for example [50], we obtain the following.

Lemma 9.3.5. *Assume that ζ is a non-degenerate mean zero probability measure on $(\mathbb{R}, \mathcal{B})$. Let $A, B \in \mathbb{M}(p,p)$ be invertible. Then the following hold:*

1. *If ζ is Gaussian, then $F^A = F^B$ if and only if $AA^t = BB^t$.*

2. *If ζ is non-Gaussian and symmetric, then $F^A = F^B$ if and only if $A = B\Lambda P$ for some permutation matrix P and a diagonal matrix Λ satisfying $\Lambda^2 = I$.*

3. *If ζ is non-symmetric, then $F^A = F^B$ if and only if $A = BP$ for some permutation matrix P.*

We conclude that concerning identifiability in (1.90), three different scenarios are possible, leading to the following question.

Question 1.6.2. *Fix a nondegenerate distribution ζ on $(\mathbb{R}, \mathcal{B})$, put $R = \zeta^{\otimes p}$ and consider the statistical model (1.90). Assume given a true mixing matrix A and samples $X_1, \ldots, X_n$ from $L_A(R)$. Which of the three scenarios outlined in Lemma 9.3.5 reflect our ability to estimate the mixing matrix based on $X_1, \ldots, X_n$?*

We expect that the scenario guiding identifiability of the mixing matrix in Question 1.6.2 depends on whether the number of observations or the closeness of ζ to Gaussian or symmetric variables is dominant. That is, if ζ is non-symmetric but close to Gaussian, while n is exceedingly large, the number of samples may be large enough to nullify the effects of having errors close to Gaussian.

Based on this line of thinking, in order to elucidate Question 1.6.2, we choose to restrict our attention to asymptotic scenarios, meaning that we consider a sequence of error distributions $P_e(\beta_n)$ and investigate identifiability of the mixing matrix as n tends to infinity and $P_e(\beta_n)$ correspondingly converges to some limiting distribution ζ. The limiting behaviour of this sequence of models would then reflect the relative dominance of the increase in the number of samples and the convergence of the error distribution to, say, a Gaussian distribution. This is reflected in the following question.

Question 1.6.3. *Consider a family of nondegenerate distributions $P_e(\beta)$ on $(\mathbb{R}, \mathcal{B})$ for $0 \leq \beta \leq 1$. Let (β_n) be a sequence converging to zero. Put $R(\beta_n) = P_e(\beta_n)^{\otimes p}$ and consider the statistical model (1.90) based on $R(\beta_n)$. Assume given a true mixing matrix A and samples $X_{n1}, \ldots, X_{nn}$ from $L_A(R(\beta_n))$. In view of Lemma 9.3.5, as n tends to infinity, is our ability to estimate the mixing matrix based on $X_{n1}, \ldots, X_{nn}$ determined by the properties of $P_e(\beta_n)$ or by the properties of the limit ζ of $P_e(\beta_n)$?*

In the hope of obtaining estimator-independent results related to Question 1.6.3, we choose to consider the empirical cumulative distribution function $\mathbb{F}^A_{\beta_n}$ of $X_{n1}, \ldots, X_{nn}$. Let $I_x = (-\infty, x_1] \times \cdots \times (-\infty, x_p]$, $\mathbb{F}^A_{\beta_n} : \mathbb{R}^p \to [0, 1]$ is defined by

$$\mathbb{F}^A_{\beta_n}(x) = \frac{1}{n} \sum_{k=1}^{n} 1_{I_x}(X_{nk}). \tag{1.91}$$

For a cumulative distribution function F on $\mathbb{R}^p$, we say that a random field W on $\mathbb{R}^p$ is an F-Gaussian field if it is a p-dimensional mean zero Gaussian field with

covariance function $R : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ given by $R(x,y) = F(x \wedge y) - F(x)F(y)$. Here, $x \wedge y$ denotes the coordinate wise minimum of x and y. By results from empirical process theory, see for example [170], it holds in the case where the error distribution does not depend on n that $\sqrt{n}(\mathbb{F}^A_{\beta_n} - F^A)$ converges weakly in $\mathcal{L}_\infty(\mathbb{R}^p)$ to an F^A-Gaussian field, so that $\mathbb{F}^A_{\beta_n}$ converges to F^A at rate $1/\sqrt{n}$. Here, $\mathcal{L}_\infty(\mathbb{R}^p)$ denotes the space of bounded and Borel measurable functions from $\mathbb{R}^p$ to $\mathbb{R}$. In the hope that closeness of $\mathbb{F}^A_{\beta_n}$ to F^B reflects the inability to distinguish A and B based on $X_{n1}, \ldots, X_{nn}$, we may make a final reduction of our initial question to the following.

Question 1.6.4. *Consider a family of nondegenerate distributions $P_e(\beta)$ on $(\mathbb{R}, \mathcal{B})$ for $0 \leq \beta \leq 1$. Let (β_n) be a sequence converging to zero. Put $R(\beta_n) = P_e(\beta_n)^{\otimes p}$ and consider the statistical model (1.90) based on $R(\beta_n)$. Assume given a true mixing matrix A and samples $X_{n1}, \ldots, X_{nn}$ from $L_A(R(\beta_n))$. In view of Lemma 9.3.5, as n tends to infinity, is the closeness of $\mathbb{F}^A_{\beta_n}$ to F^B determined by the properties of $P_e(\beta_n)$ or by the properties of the limit ζ of $P_e(\beta_n)$?*

Our contribution consists of results related to Question 1.6.4 for particular asymptotic scenarios.

1.6.2 An identifiability result

We now outline the results obtained. For details and proofs, see Chapter 9. We first make precise the asymptotic scenario we consider. Consider two distinct fixed mean zero probability measures ξ and ζ on $(\mathbb{R}, \mathcal{B})$. Define $P_e(\beta) = \beta\xi + (1-\beta)\zeta$. We refer to $P_e(\beta)$ as a contaminated ζ distribution. Fix a matrix $A \in \mathbb{M}(p,p)$. As before, we let F^A be the cumulative distribution function of $L_A(\zeta^{\otimes p})$. Furthermore, we let F^A_β be the cumulative distribution function of $L_A(P_e(\beta)^{\otimes p})$. Note in particular that $F^A = F^A_0$. Consider a probability space $(\Omega, \mathcal{F}, P)$ endowed with a triangular array $(X_{nk})_{1 \leq k \leq n}$ such that for each n, the variables $X_{n1}, \ldots, X_{nn}$ are independent variables with cumulative distribution function $F^A_{\beta_n}$. Let $\mathbb{F}^A_{\beta_n}$ be the empirical cumulative distribution function of $X_{n1}, \ldots, X_{nn}$. Also assume that we are given an F^A-Gaussian field W on $(\Omega, \mathcal{F}, P)$. Let $\beta_n = n^{-\rho}$ for some $\rho > 0$. Note that in this case, $P_e(\beta_n)$ converges to ζ. Finally, define for all $A \in \mathbb{M}(p,p)$,

$$\Gamma_1(A) = \sum_{k=1}^{p} L_A\left(\zeta^{\otimes(k-1)} \otimes \frac{\xi - \zeta}{\|\xi - \zeta\|_\infty} \otimes \zeta^{\otimes(p-k)}\right). \tag{1.92}$$

Here, $\|\cdot\|_\infty$ denotes the Kolmogorov norm on the space of signed measures, given by

$$\|\mu\|_\infty = \sup_{x \in \mathbb{R}^p} |\mu((-\infty, x_1] \times \cdots \times (-\infty, x_p])| \tag{1.93}$$

for any signed measure μ on $(\mathbb{R}^p, \mathcal{B}_p)$. The following is a simplified combination of two of our main results, Theorem 9.4.3 and Theorem 9.4.5.

Theorem 1.6.5. *Assume that c is a continuity point of the distribution of $\|W\|_\infty$. Let $A, B \in \mathbb{M}(p,p)$. The following holds:*

1. *If $\rho > 1/2$ and $F^A = F^B$, then*

$$\lim_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) = P(\|W\|_\infty > c). \tag{1.94}$$

2. *If $0 < \rho < 1/2$ and either $F^A \neq F^B$ or $F^A = F^B$ and $\Gamma_1(A) \neq \Gamma_1(B)$, then*

$$\lim_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) = 1. \tag{1.95}$$

Loosely speaking, Theorem 1.6.5 states that for the asymptotic scenarios considered, the answer to Question 1.6.4 is: For $\rho > 1/2$, closeness of the empirical cumulative distribution function $\mathbb{F}^A_{\beta_n}$ to F^B is solely determined by ζ, while for $0 < \rho < 1/2$, closeness of the empirical cumulative distribution function to F^B is not solely determined by ζ.

To understand the results of Theorem 1.6.5 in detail, consider the case where $A, B \in \mathbb{M}(p,p)$ are invertible. Assume that $AA^t = BB^t$ while $A \neq B\Lambda P$ for all diagonal Λ with $\Lambda^2 = I$ and all permutation matrices P. Let ζ be a nondegenerate Gaussian distribution and let ξ be such that $P_e(\beta)$ is non-Gaussian for all $\beta \in (0,1)$. These assumptions imply the following:

1. For all n, $F^A_{\beta_n} \neq F^B_{\beta_n}$. From this, we would not expect $\mathbb{F}^A_{\beta_n}$ and F_{β_n} to be close.
2. $F^A = F^B$. From this, if $F^A_{\beta_n}$ and $F^B_{\beta_n}$ are close enough to F^A and F^B, respectively, we would expect $\mathbb{F}^A_{\beta_n}$ to be close to $F^B_{\beta_n}$.

See also Corollary 9.4.4. Now assume that $\beta_n = n^{-\rho}$ for $\rho > 1/2$. The first part of Theorem 1.6.5 shows that (1.94) holds, indicating that in this asymptotic scenario, the probability of having $\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c$ does not tend to one. Therefore, $\mathbb{F}^A_{\beta_n}$ and $F^B_{\beta_n}$ cannot be distinguished at rate $1/\sqrt{n}$, and so it is the latter of the two above scenarios which dominates. This indicates that for $\beta_n = n^{-\rho}$ with $\rho > 1/2$, the effect of having an error distribution close to Gaussian is more important than the increase in sample size.

On the other hand, if we consider $\beta_n = n^{-\rho}$ and assume that $F^A = F^B$ while also having $\Gamma_1(A) \neq \Gamma_1(B)$, the second part of Theorem 1.6.5 shows that (1.95) holds, indicating that in this case, the probability of having $\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c$ does in fact tend to one, and so the first of the two above scenarios dominate, corresponding to the effect of having an error distribution close to Gaussian being less important than the increase in sample size. In this case, $\mathbb{F}^A_{\beta_n}$ and $F^B_{\beta_n}$ are asymptotically distinguishable at rate $1/\sqrt{n}$.

1.6.3 Numerical experiments

In this section, we carry out some numerical experiments related to the results outlined in the previous subsection. We will make two different numerical experiments: One evaluating the results of Theorem 1.6.5, and one evaulating the applicability of Theorem 1.6.5 to the practical application of ICA.

We begin by considering numerical confirmation of the results of Theorem 1.6.5. In particular, we are interested in the numerical evaluation of

$$p(\rho) = \lim_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \tag{1.96}$$

for $\beta_n = n^{-\rho}$ and varying $\rho > 0$, with appropriate choices of c, A, B, ξ and ζ. According to Theorem 1.6.5, under suitable regularity conditions, we should see that for $\rho > 1/2$, yielding a fast decrease in the level of contamination of the error distribution, that (1.96) is constant and less than one, while for $0 < \rho < 1/2$, (1.96) is constant and equal to one. A simple strategy for the evaulation of (1.96) is Monte Carlo simulation, using the approximation

$$p(\rho) \approx \frac{1}{N}\sum_{k=1}^{N} 1_{(Z_k > c)} \tag{1.97}$$

for some large N and n, where $Z_1, \ldots, Z_N$ are independent variables with distribution $\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty$. The simulation and evaluation of $\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty$ are not straightforward, as we have to evaluate both a supremum taken over an unbounded set and have to evaluate $F^B_{\beta_n}$, which is not generally known in closed form. To make our experiments feasible, we consider the scenario where $p = 2$, ζ is the standard normal distribution and ξ is the standard exponential distribution. We furthermore fix $\alpha \in (0, 1)$ and consider the two matrices

$$A = \begin{bmatrix} 1 & 0 \\ \alpha & \sqrt{1-\alpha^2} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} \sqrt{1-\alpha^2} & \alpha \\ 0 & 1 \end{bmatrix}.$$

Note that

$$AA^t = BB^t = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}, \tag{1.98}$$

while $A \neq B\Lambda P$ for all diagonal Λ with $\Lambda^2 = I$ and permutation matrices P. Thus, we are in the scenario where $F^A_\beta \neq F^B_\beta$ for $0 < \beta \leq 1$, while $F^A = F^B$. For definiteness, we put $\alpha = 0.4$ and $c = 1.5$. The key benefit of this setup is that the cumulative distribution function F^B_β can be calculated in semi-analytical form. As for the supremum $\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty$, the fact that $\mathbb{F}^A_{\beta_n}$ is an empirical cumulative distribution function and $F^B_{\beta_n}$ is a cumulative distribution function implies that the supremum can be reduced to a finite maximum. These observations make it feasible

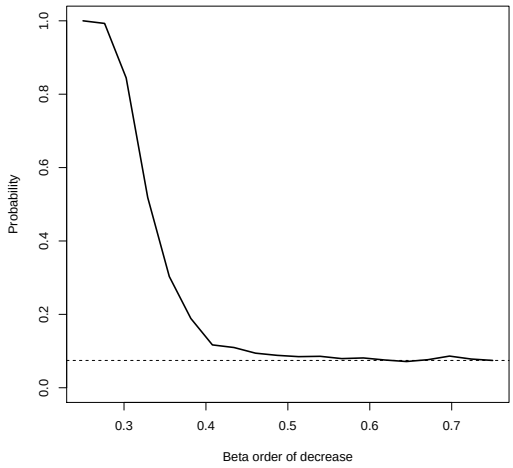


Figure 1.6.1: Plot of Monte Carlo estimates of $p(\rho)$ for $\beta_n = n^{-\rho}$ with ρ varying from 0.25 to 0.75 and $n = 50000$.

to evaluate (1.97), even though this evaluation remains computationally intensive. Figure 1.6.1 shows the results of the Monte Carlo simulations for evaluating $p(\rho)$.

Figure 1.6.1 indicates that in accordance with Theorem 1.6.5, $p(\rho)$ is the same for $\rho > 1/2$. The figure, however, does not indicate that $p(\rho)$ is one for $0 < \rho < 1/2$, as Theorem 1.6.5 states. To explain this, we note the following, also mentioned in Section 9.5. Based on the results of Section 9.3, $\|F^A_\beta - F^B_\beta\|_\infty$ is asymptotically linear in β as β tends to zero. Therefore, we would expect that for large n and some constant $k > 0$, we have

$$P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \approx P(\sqrt{n}\|F^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \approx 1_{(\sqrt{n}k\beta_n > c)}. \tag{1.99}$$

Thus, $P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \approx 1$ when $n^{1/2-\rho} = \sqrt{n}\beta_n > c/k$, corresponding to $n > \exp((\log c/k)/(1/2-\rho))$. For $\rho < 1/2$ close to $1/2$, this latter number is extremely large, making detection of the limiting value of 1 difficult. We therefore expect that with larger values of n, our numerical results would have been in accordance with Theorem 1.6.5. However, computational limitations make this infeasible.

Next, we investigate how our results translate into identifiability for ICA models in practice. To this end, fix α, β and n. We continue to consider the scenario where $p = 2$, ζ is the standard normal distribution and ξ is the standard exponential distribution. Assume that $X_1, \ldots, X_n$ are samples from the bivariate distribution $L_A(P_e(\beta) \otimes P_e(\beta))$. Thus, A is the true mixing matrix. Further assume that we have some estimate $\hat{A}$ of the mixing matrix. We will use $\hat{A}$ to devise a simple test of the hypothesis that A is the true mixing matrix against the alternative that B is the true mixing matrix. Recalling (1.98), we find that the distributions $L_A(\zeta \otimes \zeta)$ and

$L_B(\zeta \otimes \zeta)$ are the the same, and it will thus be impossible to devise a test with good properties when ζ is the error distribution. As a corollary, for the error distribution $P_e(\beta) = \beta\xi + (1-\beta)\zeta$, we expect that when β is small, any test will perform badly.

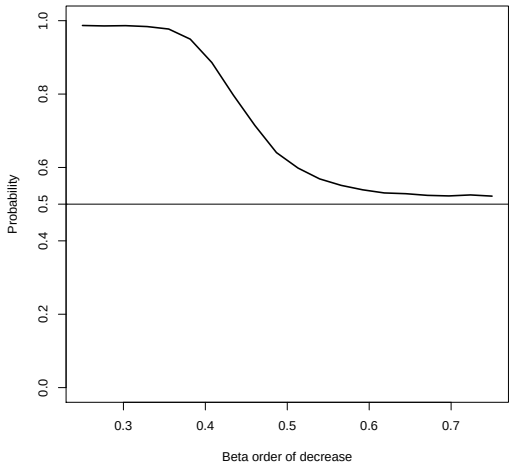


Figure 1.6.2: Plot of Monte Carlo estimates of $Q(n, \beta_n)$ for $\beta_n = n^{-\rho}$ with ρ varying from 0.25 to 0.75 and $n = 50000$.

In order to be able to apply mainstream ICA estimators, we will assume that the true error distribution is unknown. We can then only expect any estimator $\hat{A}$ to estimate the mixing matrix up to scaling and permutation of columns. To remove part of the scaling indeterminacy, we assume that $\hat{A}$ has rows of unit Euclidean norm, since this is the case for both the candidates A and B of the mixing matrix. We then calculate

$$d_A = \min \|A - \hat{A}\Lambda P\| \tag{1.100}$$

$$d_B = \min \|B - \hat{A}\Lambda P\|. \tag{1.101}$$

where the minimum is taken over all 2×2 diagonal Λ with $|\Lambda_{ii}| = 1$ and all 2×2 permutation matrices P, meaning that the minimum is over at most eight different values. Here, the norm is the Frobenius norm, meaning the entrywise Euclidean norm. If $d_A < d_B$, we accept the hypothesis that A is the true mixing matrix, while if $d_B \leq d_A$, we would accept the alternative that B is the true mixing matrix. Define

$$Q(n, \beta) = P(d_A < d_B), \tag{1.102}$$

the probability of identifying the correct mixing matrix. We are interested in the behaviour of $Q(n, \beta)$ for various asymptotic scenarios in n and β. To make the experiment concrete, we let the estimator $\hat{A}$ be the result of the fastICA algorithm

as implemented in the R package `fastICA`, applied to the data $X_1, \ldots, X_n$ and with normalized rows.

In Figure 1.6.2, we plot estimates of $\lim_n Q(n, \beta_n)$ for the scenario $\beta_n = n^{-\rho}$ with ρ varying between 0.25 and 0.75. Note that even for $\rho > 0.5$, the probability of identifying the correct mixing matrix is strictly above $1/2$. At first glance, this might seem to be at odds with the results of Theorem 1.6.5. However, Theorem 1.6.5 is not a result about identification in ICA, but rather an indication of the behaviour of identification in ICA. In particular, Theorem 1.6.5 indicates that for $\rho > 1/2$, we should not be able to identify the correct mixing matrix at rate $1/\sqrt{n}$, but the theorem does not disallow identification entirely.

Ultimately, however, Figure 1.6.2 does not appear to have much of a relationship with neither Figure 1.6.1 nor Theorem 1.6.5. We conclude that in order to obtain information about practical identifiability in ICA from Theorem 1.6.5, a more subtle approach is necessary.

1.6.4 Concluding comments

Compared to our most general research question, Question 1.6.1, our results on identification in ICA models, centered around Question 1.6.4, are inconclusive, mainly because it is unclear how answers to Question 1.6.4 ultimately help to answer Question 1.6.1. In particular, finite-sample results such as those required by Question 1.6.1 would be much more informative than the asymptotic results obtained when answering Question 1.6.4.

Nonetheless, our efforts have aided in setting up a framework of analysis and some ideas and techniques which may be helpful for future research efforts. It is our hope that our current work will be a stepping stone towards more precise results on identifiability in ICA.

1.7 Model selection in nonlinear regression

In this section, we outline our work on degrees of freedom in nonlinear regression. Details can be found in Chapter 10. We commence by explaining our motivating problem, an estimation problem for Ornstein-Uhlenbeck processes. After this, we show how this type of estimation problem can be handled in the simple linear regression case. Finally, we consider the extension of the methods from linear regression to nonlinear regression. Ultimately, this does not solve our initial problem, but is a step along the road to a solution.

1.7.1 An estimation problem for Ornstein-Uhlenbeck processes

We first recall the motivating problem outlined in Section 1.2. We consider an Ornstein-Uhlenbeck diffusion model on the form

$$\mathrm{d}X_t = BX_t\,\mathrm{d}t + \mathrm{d}W_t, \tag{1.103}$$

where W is a p-dimensional Brownian motion and $B \in \mathbb{M}(p,p)$. We are interested in estimation of the mean reversion speed matrix B. As an application, we think of the coordinates of X as expression levels of a set of p genes. Applying our notion of causality for SDEs, the zeroes of B determine the signature of the SDE (1.103).

We hope that the underlying true parameter is sparse, corresponding to a sparse causal structure. We would therefore like to obtain an estimate of B which is sparse as well. We refer to this type of problem as a variable selection problem, or more generally, as a model selection problem. The rationale behind this nomenclature is that a sparse estimate of B would, through its zeroes, "select" a submodel of the full model with reduced dimensionality.

Now assume that we observe X at equidistant timepoints $t_k = \Delta k$ for $k = 0, \ldots, n$. By a loss function, we mean a function penalizing errors in prediction, see Chapter 3 of [96]. As the conditional mean of X_{t_k} given $X_{t_{k-1}}$ is $\exp(tB)X_{t_{k-1}}$, a reasonable loss function $R : \mathbb{M}(p,p) \to [0,\infty)$ for estimation of B based on the observations $X_{t_0}, \ldots, X_{t_n}$ is

$$R(B) = \sum_{k=1}^{n} \|X_{t_k} - \exp(\Delta B)X_{t_{k-1}}\|_2^2, \tag{1.104}$$

since this loss function compares conditional means with actual values. Note that Δ in (1.104) does not refer to a jump, as previously in this chapter, but instead refers to the distance between the timepoints where X is observed.

A simple estimator of B is then given by

$$\hat{B} \in \operatorname*{argmin}_{B \in \mathbb{M}(p,p)} R(B). \tag{1.105}$$

In Figure 1.7.1, an example of a B matrix as well as an estimate based on (1.105) for simulated data is shown. We see that while the example B matrix is sparse, the estimate is not sparse at all, as no entries of the estimate is zero. In order to obtain sparse estimates, we instead consider L^1-penalized estimators of the form

$$\hat{B}_\lambda \in \operatorname*{argmin}_{B \in \mathbb{M}(p,p)} R(B) + \lambda\|B\|_1, \tag{1.106}$$

where $\|\cdot\|_1$ denotes the entrywise L^1-norm and $\lambda \geq 0$. This is a nonsmooth regression problem. A numerical algorithm for solving this problem may be obtained by iteratively considering linear approximations of $B \mapsto X_{t_k} - \exp(\Delta B)X_{t_{k-1}}$ and

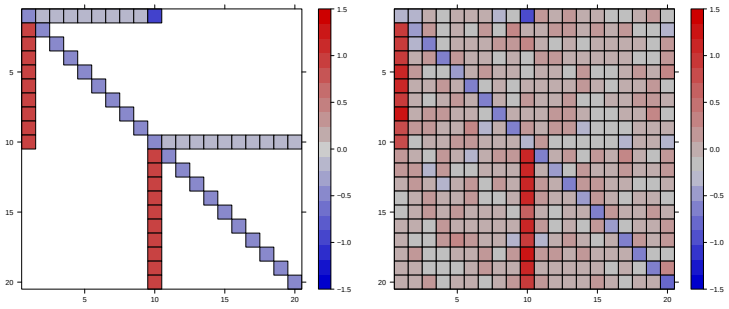


Figure 1.7.1: Left: An example B matrix for the model (1.103). Right: The estimate obtained by minization of the loss function (1.104) for simulated data.

applying a coordinate-wise minimization algorithm as outlined by Friedman et al. in [56]. In Figure 1.7.2, we show examples of estimates obtained from solving (1.106) for the same simulated data as applied in Figure 1.7.1, for varying levels of λ. We see that as expected, increasing levels of λ yield ever sparser estimates of B. Thus, based on (1.106), we obtain a family $(\hat{B}_\lambda)$ of estimates of B with increasing levels of sparsity.

In order to use this in practice, we need to be able to select a good level of sparsity, or equivalently, we need to choose λ in some sensible way. In order to obtain a plan for this, we consider the linear regression case and review a methodology which is applicable there.

1.7.2 An example based on the LASSO

For the purposes of our example, we consider a linear regression model of the form

$$Y = X\beta + \varepsilon \tag{1.107}$$

where $X \in \mathbb{M}(n, p)$ and ε follows an n-dimensional Gaussian distribution with mean zero and variance σI_n, where I_n denotes the identity matrix. The LASSO estimator, see for example the book [64] by Hastie et al. or the paper [167] by Tibshirani, is given by

$$\hat{\beta}_\lambda \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1. \tag{1.108}$$

Similarly to (1.106), $\hat{\beta}_\lambda$ tends to increase in sparsity as λ increases. In order to see this, we consider $n = 100$, $p = 5000$ and an artificially generated design matrix X and true sparse parameter β, containing only 50 nonzero values. Efficient algorithms exist for the calculation of the LASSO estimates for varying values of λ, see for example

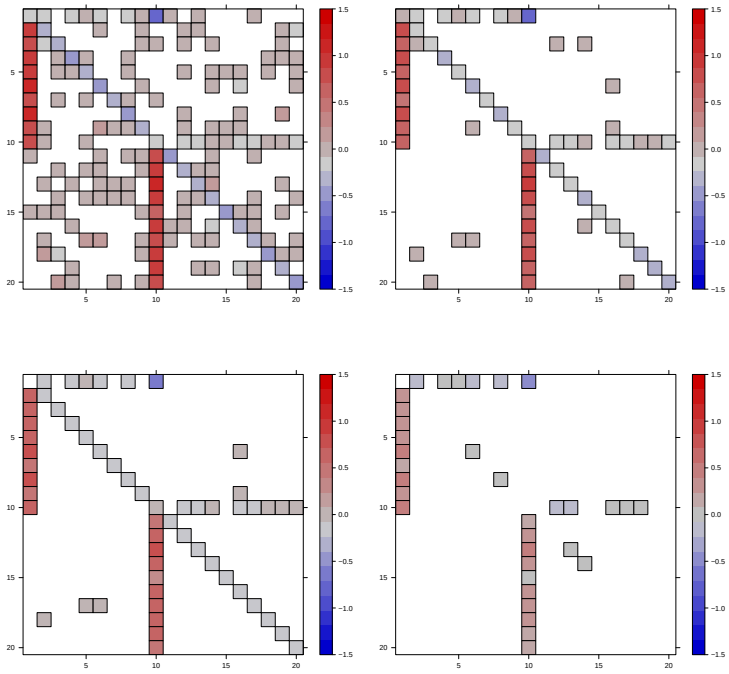


Figure 1.7.2: From left to right: Estimates of B obtained from (1.106) for λ equal to 0.05, 0.20, 0.30 and 0.50.

[56, 45, 64]. In Figure 1.7.3, we show selected coordinates of $\hat{\beta}_\lambda$ for varying levels of λ. The figure confirms that as expected, as λ increases, more and more coordinates of $\hat{\beta}_\lambda$ become zero.

As in the previous subsection, we now face the problem of making a good choice of λ. As λ measures the sparsity of our parameter and thus heuristically speaking the complexity of our model, we may think of the choice of λ as a problem of model selection. In order to choose λ, we consider a function which measures the quality of our estimator. One natural such measure of quality is the mean squared prediction error, given by

$$\mathrm{MSPE}_\beta(\hat{\beta}_\lambda) = E_\beta \|Y - X\hat{\beta}_\lambda(Y)\|_2^2, \tag{1.109}$$

where $\|\cdot\|_2$ denotes the Euclidean norm and E_β denotes expectation given that β is the true parameter. We write $\hat{\beta}_\lambda(Y)$ instead of just $\hat{\beta}_\lambda$ to emphasize that $\hat{\beta}_\lambda$ is a function of Y. The mean squared prediction error can be estimated by the training error, given by

$$\mathrm{err}(\hat{\beta}_\lambda) = \|Y - X\hat{\beta}_\lambda(Y)\|_2^2. \tag{1.110}$$

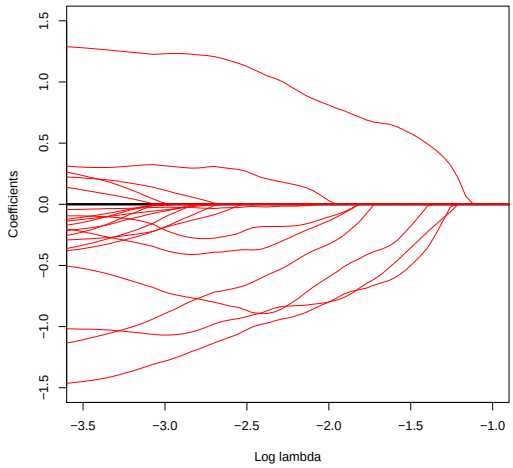


Figure 1.7.3: LASSO estimates for selected coordinates of $\hat{\beta}_\lambda$ as λ varies. For convenience, the abscissa denotes $\log \lambda$ and not λ. The figure illustrates that as λ increases, $\hat{\beta}_\lambda$ becomes sparse.

We may then choose λ as the argument minimum of (1.110). However, it is immediate that this will always favor $\lambda = 0$, as $\hat{\beta}_\lambda(Y)$ for $\lambda = 0$ in fact is the argument minimum of $\|Y - X\beta\|_2^2$. Therefore, we need another quality measure. To obtain this, we make the observation that in many cases, we would like our estimator to be able to correctly predict future responses Y^*, and not only responses Y already observed. This leads to the idea of the generalization error, see for example the paper [44] by Efron, given by

$$\mathrm{Err}_\beta(\hat{\beta}_\lambda) = E_\beta \|Y^* - X\hat{\beta}_\lambda(Y)\|_2^2, \tag{1.111}$$

where Y^* is independent of Y and has the same distribution as Y. The generalization error measures the expected performance of $\hat{\beta}_\lambda$ when used for prediction in a new sample. We expect that if we choose λ as the argument minimum of (1.111), we obtain a choice of λ which balances good prediction on our current sample, since Y^* and Y after all follow the same distribution, with flexibility of the estimator, as (1.111) favors not fitting too strongly to our observed response, Y. Of course, as (1.111) is a mean with respect to β, it is not known. We can therefore not minimize it over λ, but we may try to minimize estimates of it. A natural estimate to make is

$$\widehat{\mathrm{Err}}_\beta(\hat{\beta}_\lambda) = \|Y^* - X\hat{\beta}_\lambda(Y)\|_2^2, \tag{1.112}$$

but this latter is not observable, as we by definition do not observe the hypothetical

future response Y^*. However, as shown in [44], it holds that

$$\mathrm{Err}_\beta(\hat{\beta}_\lambda) = E_\beta\|Y - X\hat{\beta}_\lambda(Y)\|_2^2 + 2\sum_{i=1}^{n}\mathrm{Cov}_\beta(Y_i, X\hat{\beta}_\lambda(Y)_i). \tag{1.113}$$

Heuristically, (1.113) shows how the generalization error and the mean squared prediction error differ: As we fit our estimate harder to our observed data Y, the covariance between our reponse and our predicted values in (1.113) increases, and so the discrepancy between the mean squared prediction error and the generalization error increases. In the expression (1.113) for the generalization error, the hypothetical future response Y^* is no longer present. However, the covariance term does not lend itself to simple estimation. To circumvent this, we may apply a result of Stein's paper [163], which shows that under suitable regularity conditions, it holds that

$$\sum_{i=1}^{n}\mathrm{Cov}_\beta(Y_i, X\hat{\beta}_\lambda(Y)_i) = \sigma^2 E_\beta(\mathrm{div}\, X\hat{\beta}_\lambda)(Y), \tag{1.114}$$

where div denotes the divergence, that is, the sum of the partial derivatives. It should be noted that the regularity conditions necessary for (1.114) to hold are not innocent. In particular, to prove (1.114), it is in [163] explicitly used that the error variables ε are independent and follow identical Gaussian distributions. When (1.114) is substituted in (1.113), we obtain

$$\mathrm{Err}_\beta(\hat{\beta}_\lambda) = E_\beta\|Y - X\hat{\beta}_\lambda(Y)\|_2^2 + 2\sigma^2 E_\beta(\mathrm{div}\, X\hat{\beta}_\lambda)(Y). \tag{1.115}$$

This implies that the generalization error may be estimated as

$$\widehat{\mathrm{Err}}_\beta(\hat{\beta}_\lambda) = \mathrm{err}(\hat{\beta}_\lambda) + 2\hat{\sigma}^2(\mathrm{div}\, X\hat{\beta}_\lambda)(Y), \tag{1.116}$$

where $\hat{\sigma}$ is some estimator of σ. In general, it is not obvious how this estimator of σ should be chosen. For the purposes of this example, we will somewhat unfairly pretend that we know the true σ. To complete our program, then, it remains to calculate the divergence. In [168], it is shown by Tibshirani and Taylor that

$$E_\beta(\mathrm{div}\, X\hat{\beta}_\lambda)(Y) = E_\beta\ \mathrm{rank}\ X_{(\cdot,\mathcal{A})}, \tag{1.117}$$

where $X_{(\cdot,\mathcal{A})}$ denotes the submatrix of the design matrix X corresponding to the columns $\mathcal{A}$, where $\mathcal{A}$ is the stochastic set of nonzero elements of $\hat{\beta}_\lambda$. Combining these results, we come to the conclusion that we for any $\lambda \geq 0$ may obtain an unbiased estimate of the generalization error corresponding to the LASSO estimate by putting

$$\widehat{\mathrm{Err}}_\beta(\hat{\beta}_\lambda) = \mathrm{err}(\hat{\beta}_\lambda) + 2\sigma^2\mathrm{rank}\ X_{(\cdot,\mathcal{A})}. \tag{1.118}$$

In Figure 1.7.4, we plot this estimate of the generalization error for our data for varying $\lambda \geq 0$. Note that the function plotted has a jagged, discontinuous shape.

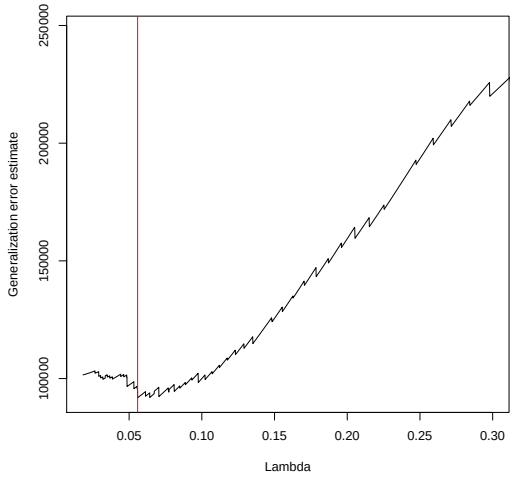


Figure 1.7.4: The estimated generalization error (1.118) for varying levels of λ. The vertical red line identifies the λ corresponding to the minimal generalization error estimate.

The points of discontinuity occur when the number of nonzero variables changes, leading to a change in the rank of $X_{(\cdot,\mathcal{A})}$.

Picking the λ corresponding to the minimum estimated generalization error in this case leads to choosing $\hat{\lambda} = 0.0558$. This yields a sparse estimate $\hat{\beta}_{\hat{\lambda}}$ with 78 nonzero entries, which should be compared with the number of nonzero entries in the true β parameter, 50. What this example shows is that using the notion of generalization error and rewriting the generalization error in terms of the training error plus a divergence term, we are able to obtain a method for selecting λ through minimization of the estimated generalization error. This allows us to select an estimate $\hat{\beta}_{\hat{\lambda}}$ to use for model selection and provides an alternative to for example the cross-validation methods as described in [64].

1.7.3 Model selection in nonlinear regression

The next question to be posed is whether the program which helped us select λ in the linear regression case can be carried through in a more general setting. For the Ornstein-Uhlenbeck model (1.103) and the mean reversion speed matrix estimator (1.106), a natural definition of the generalization error is

$$\mathrm{Err}_B(\hat{B}_\lambda) = E_B \sum_{k=1}^{n} \|X^*_{t_k} - \exp(\Delta \hat{B}_\lambda(X_{t_0,\ldots,t_n}))X^*_{t_{k-1}}\|_2^2, \tag{1.119}$$

where X^* is independent of X and has the same distribution as X. This measures the predictive performance of the estimator $\hat{B}_\lambda$, based on $X_{t_0}, \ldots, X_{t_n}$, when applied to a new, independent set of data. Ideally, we would like to derive expressions for (1.119) in terms of the training error and a type of bias correction term. We have not been able to achieve this.

Instead of solving our main problem, we have considered a simpler problem, which nonetheless covers a more general class of models than the linear regression model. We chose to consider the nonlinear regression model

$$Y = \varphi(\beta) + \varepsilon \tag{1.120}$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ is continuous and ε follows an n-dimensional Gaussian distribution with mean zero and variance σI_n. This model will take the place of the more complicated Ornstein-Uhlenbeck model. Our goal is to consider estimators in the model (1.120) and prove results which pave the way for estimation of the generalization error for these estimators. For an estimator $\hat{\beta}$ in the nonlinear regression model, it is straightforward to define the training and generalization error as

$$\mathrm{err}_\beta(\hat{\beta}) = \|Y - \varphi(\hat{\beta}(Y))\|_2^2, \tag{1.121}$$

$$\mathrm{Err}_\beta(\hat{\beta}) = E_\beta\|Y^* - \varphi(\hat{\beta}(Y))\|_2^2. \tag{1.122}$$

The results of [44] also apply to this case, and yield

$$\mathrm{Err}_\beta(\hat{\beta}) = \mathrm{err}_\beta(\hat{\beta}) + 2\sum_{i=1}^{n} \mathrm{Cov}_\beta(Y_i, \varphi(\hat{\beta})(Y)_i). \tag{1.123}$$

At this juncture, it is natural to introduce

$$\mathrm{df}_\beta(\hat{\beta}) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}_\beta(Y_i, \varphi(\hat{\beta})(Y)_i). \tag{1.124}$$

We refer to $\mathrm{df}_\beta(\hat{\beta})$ as the degrees of freedom of the estimator $\hat{\beta}$. We then obtain the central identity

$$\mathrm{Err}_\beta(\hat{\beta}) = \mathrm{err}_\beta(\hat{\beta}) + 2\sigma^2\mathrm{df}_\beta(\hat{\beta}). \tag{1.125}$$

Furthermore, if the regularity conditions required by [163] hold, we also obtain

$$\mathrm{df}_\beta(\hat{\beta}) = E_\beta(\mathrm{div}\,\varphi \circ \hat{\beta})(Y), \tag{1.126}$$

which is known as the divergence form of the degrees of freedom, or simply as Stein's unbiased risk estimate (SURE). The term "degrees of freedom" is apt, because in several cases, $\mathrm{df}_\beta(\hat{\beta})$ reduces to something that sensibly can be interpreted as the "dimension" of the model. The identity (1.125) shows that if we are to estimate the generalization error, it suffices to calculate a workable expression for the degrees of freedom.

We will take an interest in calculating the degrees of freedom for select classes of estimators in this model. We consider two types of estimators, namely

$$\hat{\beta}(y) \in \operatorname*{argmin}_{\beta \in K} \|y - \varphi(\beta)\|_2^2 \tag{1.127}$$

for compact $K \subseteq \mathbb{R}^p$, and

$$\hat{\beta}_\lambda \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - \varphi(\beta)\|_2^2 + \lambda \|\beta\|_1, \tag{1.128}$$

for $\lambda \geq 0$. These two classes are related, in the sense that they are both liable to provide sparse estimates. Here, (1.127) yields sparse estimates when K is chosen as for example a centered ball in the L^1-norm. Obtaining expressions for the degrees of freedom opens up the door to estimating the generalization error (1.125) and choosing optimal λ for estimators of the types (1.127) and (1.128).

Before proceeding to the presentation of our results, we remark that the discussion of degrees of freedom, both for various types of estimators in the linear regression model, and for estimators in more complicated models, is not a novel undertaking. Results related to this can be found for example in [38, 40, 42, 93, 99, 118, 155, 176], which covers both general considerations on the nature of the degrees of freedom as well as results in specific cases such as L^1-penalized linear regression, support vector regression, partial least squares and shape-restricted regression.

Our main contribution is the following two theorems related to the degrees of freedom for (1.127) and (1.128). In the statement of the theorems, we neglect to detail the regularity conditions required for the theorems to hold. For more precise statements, see Section 10.3.

Theorem 10.3.1. *Let $\hat{\beta}$ be as in (1.127). Subject to regularity criteria, it holds that $\mathrm{df}(\hat{\beta}) = \sum_{i=1}^n \int_{\mathbb{R}^n} \psi(y)\, \mathrm{d}\mu^i(y)$, where ψ is the density of Y and $(\mu^i)_{i \leq n}$ is a family of nonnegative Radon measures corresponding to the partial derivatives of $\varphi \circ \hat{\beta}$ in a generalized function sense.*

Theorem 10.3.2. *Let $\hat{\beta}$ be as in (1.128). Subject to regularity criteria, it holds that $\mathrm{df}(\hat{\beta}) = E_\beta \mathrm{tr}\, A(Y)_{(\cdot,\mathcal{A})} B(Y, \hat{\beta})^{-1}_{(\mathcal{A},\mathcal{A})} A(Y)^t_{(\cdot,\mathcal{A})}$ where $B \in \mathbb{M}(p,p)$ and $A \in \mathbb{M}(n,p)$ are given by*

$$B_{ij}(y,\beta) = \sum_{k=1}^n \frac{\partial \varphi_k}{\partial \beta_i}(\beta) \frac{\partial \varphi_k}{\partial \beta_j}(\beta) - (y_k - \varphi(\beta)_k) \frac{\partial^2 \varphi_k}{\partial \beta_i \partial \beta_j}(\beta), \tag{1.129}$$

$$A_{ki}(\beta) = \frac{\partial \varphi_k}{\partial \beta_i}(\beta), \tag{1.130}$$

and $\mathcal{A}$ is the active set of the estimator, $\mathcal{A} = \{i \leq p \mid \hat{\beta}_i \neq 0\}$.

It is essential to point out that while the regularity required for the conclusions of Theorem 10.3.1 to hold are quite innocent, the same cannot be said for Theorem

10.3.2. Therefore, at present, Theorem 10.3.2 should be seen as a "moral theorem" giving an expression for the natural candidate of the degrees of freedom of an L^1-penalized estimator in the nonlinear case.

Before concluding this section, we comment on the content of the two theorems, and outline further work to be done. Consider first the case of the constrained estimator (1.127). We will argue that heuristically, Theorem 10.3.1 leads to an extended divergence formula for the degrees of freedom. To see this, we introduce, for any nonempty compact set $K \subseteq \mathbb{R}^p$, the metric projection onto K as the multifunction

$$\mathrm{pr}_K(y) = \operatorname*{argmin}_{x \in K} \|x - y\|_2^2. \tag{1.131}$$

We then obtain that $\varphi(\hat{\beta}(y)) \in \mathrm{pr}_{\varphi(K)}(y)$ for all $y \in \mathbb{R}^n$. It is this property which ensures the existence of the Radon measures occurring in Theorem 10.3.1. These nonnegative Radon measures in fact corresponds to distributional second-order partial derivatives of the convex mapping $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = \|x\|_2/2 - d_{\varphi(K)}(x)^2/2, \tag{1.132}$$

where $d_{\varphi(K)} = \inf_{y \in K} \|x - y\|_2$. Now let $\mu^i = \mu^i_{ac} + \mu^i_s$ be the Lebesgue decomposition of μ^i with respect to the Lebesgue measure on $\mathbb{R}^n$. By Alexandrov's theorem, see Theorem 6.4.1 of the book [52] by Evans and Gariepy, it holds for Lebesgue almost all $x \in \mathbb{R}^n$ that

$$\lim_{y \to x} \frac{f(y) - f(x) - (y - x)^t Df(x) - \frac{1}{2}(y - x)^t D^2 f(x)(y - x)}{\|y - x\|_2^2} = 0, \tag{1.133}$$

where $Df(x)$ denotes the gradient of f, which exists Lebesgue almost everywhere by Theorem 6.3.1 and Theorem 3.1.2 of [52], and $D^2 f_{ii}(x)$ is the Lebesgue density of μ^i_{ac}. As it can be shown that $\mathrm{pr}_{\varphi(K)}$ is single-valued and equal to Df Lebesgue almost everywhere, this indicates that $D^2 f$ in fact Lebesgue almost everywhere is the derivative of $\mathrm{pr}_{\varphi(K)}$, or equivalently, $\varphi \circ \hat{\beta}$. Inserting this into the degrees of freedom formula of Theorem 10.3.1, we obtain

$$\mathrm{df}(\hat{\beta}) = E \operatorname{div}(\varphi \circ \hat{\beta})(Y) + \sum_{i=1}^{n} \int_{\mathbb{R}^n} \psi(y) \, \mathrm{d}\mu^i_s(y). \tag{1.134}$$

As the first term of (1.134) exactly corresponds to the divergence form of the degrees of freedom, this yields an extended divergence-type formla for constrained estimators. In the case where the singular measures μ^i_s vanish, we reclaim the ordinary divergence form of the degrees of freedom.

In order to rigorously prove (1.134), it would be necessary to argue that $D^2 f$ in fact can be identified with the derivative of $\mathrm{pr}_{\varphi(K)}$. This multifunction is single-valued almost everywhere. However, it may still be the case that the set where the multifunction is not single-valued is dense in $\mathbb{R}^n$, rendering the classical notion of

differentiability invalid for $\mathrm{pr}_{\varphi(K)}$. One plan for rigorously proving (1.134) would be to apply extended differentiation theories for set-valued functions such as developed in [141, 16].

As for Theorem 10.3.2, the main challenge is to obtain verifiable sufficient criteria for the degrees of freedom formula to hold. It is at present unclear how to do this.

Further work in this direction involves degrees of freedom results for the case where the variance of ε in (1.120) is not diagonal, or when there are more complicated dependencies in the minimization problem corresponding to the estimator, as in (1.104).

1.8 Directions for future research

In this section, we outline some possible topics for further research.

Nonexplosion criteria for counting processes. The results outlined in Section 1.3 give sufficient criteria ensuring the construction of nonexplosive counting processes with particular stochastic intensities. It is of general interest to expand further on the situations where these criteria can be applied. In particular, it would be of interest to apply such criteria or related martingale methods to obtain criteria for nonexplosion in the case where the candidate intensity exhibits high degrees of non-monotonicity, as is for example the case for self-exciting processes such as Hawkes processes.

Monotonicity properties of exponential martingales. As mentioned in Chapter 2, the martingale property for exponential martingales $\mathcal{E}(M)$ is generally not "monotone" in M in any natural sense. This is argued in [94], Example 1.13. Now assume that N is a homogeneous Poisson process, with the same setup as in Chapter 2. In spite of the results of [94], as the martingale property of $\mathcal{E}(H \cdot M)$ with $H = \mu - 1$ heuristically corresponds to non-explosion of counting processes with intensity μ, it is natural to conjecture the following. Consider nonnegative, predictable and locally bounded processes μ and μ^*. Does it hold for this particular type of exponential martingales that $\mu^* \leq \mu$ implies that if $\mathcal{E}(H \cdot M)$ is a martingale, then $\mathcal{E}(H^* \cdot M)$ is a martingale as well? Here, $H^* = \mu^* - 1$.

The relationship between the signature and distributions of SDEs. Example 1.5.17 shows that for carefully chosen examples, two SDEs may have different signatures, yet have the same distributions and postintervention distributions. This is in some sense an irregular phenomenon. It would be of interest to understand when this type of behaviour can occur, and whether it is unlikely to happen in practice, in the sense of for example having the parameters allowing this to occur being of Lebesgue measure zero.

Optimization problems for postintervention distributions. The notion of causality for SDEs outlined in Section 1.5 indicates optimization problems of interest.

For example, assume given an Ornstein-Uhlenbeck SDE of the type

$$\mathrm{d}X_t = B(X_t - A)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t \tag{1.135}$$

where the parameters are assumed to have been estimated. If X corresponds to a gene expression network, we might have an interest in understanding which genes to knock out, similar to the problems considered in the data example of [113], in order to optimize some particular feature of the postintervention distribution.

SDEs and DAG-based inference. The results of Subsection 1.5.6 indicate a general lack of correspondence between the causal structure of an SDE and the DAG-structure of its stationary distribution. Nonetheless, practical results show that applying DAG-based causal inference to continuous-time systems assumed to be in equilibrium can nonetheless yield good results, see for example [112, 120]. It is of interest to understand under which conditions estimation of the DAG for the stationary distribution can yield good information about the signature of an underlying SDE for the continuous-time system.

SDEs and more general types of interventions. Definition 7.2.2 yields a notion of intervention in SDEs where a coordinate of the process is set to a constant value at all timepoints. It would be of interest to consider extensions of this to the case where the process is only intervened on at particular times, or where the resulting intervened process is not set to a constant. Also, it would be of interest to extend the notion of interventions to cases covering situations such as the one described in Example 7.4.7.

Identifiability of the mixing matrix in ICA. Our results on identifiability in ICA, outlined in Section 1.6, are hardly conclusive. It would be of interest to expand on the results obtained there, or alternatively, consider the same problem from another perspective. We have obtained theoretical results for one type of scenario where behaviour can be distinguished according to the asymptotic closeness of the error distributions to a Gaussian distribution. It would be of interest both to obtain more theoretically interesting scenarios of this type, as well as to obtain designs for numerical experiments indicating how practical identifiability in ICA is determined by the way in which the error distribution is close to but not quite Gaussian.

Degrees of freedom in nonlinear regression with independent Gaussian errors. In Section 1.7, we obtained candidate expressions for the degrees of freedom for constrained and L^1-penalized estimators in the context of nonlinear regression with independent Gaussian errors. Several extensions of these results would be beneficial. Regarding constrained estimation and Theorem 10.3.1, rigorous proof is needed that the density of the absolutely continuous part of μ^i in fact corresponds to the partial derivative of $\varphi \circ \hat{\beta}$. Furthermore, it is of interest to obtain sufficient criteria on K to ensure that the singular part of μ^i vanishes such that the ordinary divergence form of the degrees of freedom can be reclaimed. As for L^1-penalized estimation and Theorem 10.3.2, it would be of interest to obtain verifiable sufficient criteria for the degrees of freedom formula to hold. Also, it would be beneficial to investigate what

results from the literature can be obtained as special cases of the degrees of freedom formula of Theorem 10.3.2.

Degrees of freedom in extended models. Our initial motivation for considering degrees of freedom was model selection for discretely observed Ornstein-Uhlenbeck processes of the form (1.103). The training error estimators in this type of model could naturally be taken to be of the form

$$\|X_{t_k} - \exp(\Delta \hat{B}(X_{t_0}, \ldots, X_{t_n}))X_{t_{k-1}}\|_2^2, \tag{1.136}$$

while the generalization error is

$$E\|X^*_{t_k} - \exp(\Delta \hat{B}(X_{t_0}, \ldots, X_{t_n}))X^*_{t_{k-1}}\|_2^2, \tag{1.137}$$

with X^* having the same distribution as X while being independent of X. The covariance and divergence calculations applied in Chapter 10 to relate the training and generalization errors do not apply here. It would be of interest to understand the relationship between the training and generalization errors in this more complicated model. A first step in this direction could be to consider degrees of freedom in nonlinear regression models $Y = \varphi(\beta) + \varepsilon$ in the case where, instead of having ε be multivariate normal with mean zero and variance σI_n, we allow a general covariance matrix Σ.

2

Exponential martingales and changes of measure for counting processes

Alexander Sokol and Niels Richard Hansen

2010 Mathematics Subject Classification. Primary 60G44; Secondary 60G55.

Key words and phrases. Counting Process, Martingale, Exponential martingale, Girsanov, Intensity, Uniform integrability, Absolute continuity.

Abstract. We give sufficient criteria for the Doléans-Dade exponential of a stochastic integral with respect to a counting process local martingale to be a true martingale. The criteria are sufficiently weak to be useful and verifiable, as illustrated by several non-trivial examples, without introducing artificial constraints. In particular, they make it possible to construct nonexplosive point processes with intensities adapted to a general filtration by a change of measure.

2.1 Introduction

The motivation for this paper is the problem of constructing nonexplosive dynamic processes via a change of measure on the background probability space. The objective is to derive verifiable conditions in a context of counting processes for the exponential

martingale to be a true martingale without introducing artificial constraints. As discussed recently by [60], it is of general interest to formulate a statistical model of a dynamic counting process in terms of a family of candidate intensities, and it is then essential to be able to verify that the intensities give well-defined nonexplosive models. To this end, we need conditions on the candidate intensities. If the intensity is adapted to the filtration generated by the counting process itself, precise results are obtainable by transferring the problem to a canonical setup, see [81]. Exercise 4.4.5 in [81] gives the following result. For an intensity process λ such that

$$\lambda_t \leq a(N_{t-}) \tag{2.1}$$

for a sequence $a(n)$ satisfying $\sum_{n=1}^{\infty} \frac{1}{a(n)} = \infty$, it holds that there is a probability space with a nonexplosive counting process N having intensity λ under P. The result is mentioned in [60] as the Jacobsen condition.

Alternative approaches to ensure the existence of a nonexplosive counting process with a given intensity are also surveyed in [60]. One of the most general, explicit conditions mentioned in [60] is (25). This is a growth condition on λ_t^a with $a > 1$. A consequence of our results is that (25) of [60] also is a sufficient criterion for nonexplosion when $a \geq 1$, and not only when $a > 1$.

Our starting point is the paper by Lépingle and Mémin, [109], and their general results, which we adapt to the specific case of Doléans-Dade exponentials of stochastic integrals with respect to counting process local martingales. We also employ a reduction to arbitrarily small time intervals, which considerably improves the usefulness of the criteria obtained. Furthermore, we illustrate how the criteria can be verified. We consider, in particular, examples of interacting diffusion and jump processes for which the general framework is suitable.

2.2 Summary of results

In this section, we state and discuss our main results, postponing proofs to Section 2.4. Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ satisfying the usual conditions, see [134], Section I.1 for the definition of this as well as other standard probabilistic concepts. We say that N is a nonexplosive d-dimensional counting process if N is càdlàg and piecewise constant with jumps of size one, and no coordinates of N jump at the same time. We say that a process X is locally bounded if there is a sequence of stopping times increasing almost surely to infinity such that $X^{T_n} 1_{(T_n>0)}$ is bounded. Let λ be a nonnegative, predictable and locally bounded d-dimensional process. Then λ is almost surely integrable on compacts with respect to the Lebesgue measure. We say that N is a counting process with intensity λ if it holds that $N_t^i - \int_0^t \lambda_s^i \, \mathrm{d}s$ is a local martingale for each i. Note in particular that since the predictable σ-algebra considered is the one generated by the filtration $(\mathcal{F}_t)_{t\geq 0}$, the intensity is allowed to depend on other processes than just N.

We recall the definition of Doléans-Dade exponentials. In the following, all semimartingales X are assumed to have càdlàg paths, that is, $X(\omega)$ is càdlàg for all $\omega \in \Omega$. By X_{t-}, we denote the limit of X_s as s tends to t from below, and we write $\Delta X_t = X_t - X_{t-}$ for the jump of X at t. Assume given a semimartingale X with initial value zero. By Theorem II.37 of [134] and Theorem I.4.61 of [83], the stochastic differential equation $Z_t = 1 + \int_0^t Z_{s-} \, \mathrm{d}X_s$ has a càdlàg adapted solution, unique up to indistinguishability, and the solution is

$$\mathcal{E}(X)_t = \exp\left(X_t - \frac{1}{2}[X^c]_t\right) \prod_{0<s\leq t} (1 + \Delta X_s) \exp(-\Delta X_s), \tag{2.2}$$

where X^c is the continuous martingale part of X, see Proposition I.4.27 of [83], and $[X^c]$ denotes the quadratic variation process. If X is a local martingale, $\mathcal{E}(X)$ is a local martingale as well, and in this case, we refer to $\mathcal{E}(X)$ as an exponential martingale. The case $\Delta X \geq -1$ is be of particular importance to us. In this case, $\mathcal{E}(X)$ is nonnegative, and we may put $R = \inf\{t \geq 0 \mid \Delta X_t = -1\}$ and obtain

$$\mathcal{E}(X)_t = 1_{(t<R)} \exp\left(X_t - \frac{1}{2}[X^c]_t + \sum_{0<s\leq t} \log(1 + \Delta X_s) - \Delta X_s\right). \tag{2.3}$$

Now assume given a d-dimensional nonexplosive counting process N with nonnegative, predictable and locally bounded intensity λ, and assume given another d-dimensional nonnegative, predictable and locally bounded process μ.

Definition 2.2.1. We say that μ is λ-compatible if it holds for all $\omega \in \Omega$ that $\mu_t^i(\omega) = 0$ whenever $\lambda_t^i(\omega) = 0$, and if the process γ defined by $\gamma_t^i = \mu_t^i(\lambda_t^i)^{-1}$ for $i \leq d$ is locally bounded.

In Definition 2.2.1, we use the convention that zero divided by zero is equal to one. Now assume that μ is λ-compatible. Define M to be the d-dimensional local martingale given by $M_t^i = N_t^i - \int_0^t \lambda_s^i \, \mathrm{d}s$. Put $\gamma_t^i = \mu_t^i(\lambda_t^i)^{-1}$ and $H_t^i = \gamma_t^i - 1$ for $t \geq 0$. As we have assumed that μ is λ-compatible, γ and H are both well-defined and locally bounded real-valued processes. We define $(H \cdot M)_t = \sum_{i=1}^d \int_0^t H_s^i \, \mathrm{d}M_s^i$, $H \cdot M$ is then a one-dimensional process.

The following lemma shows that given λ and μ, $\mathcal{E}(H \cdot M)$ is the relevant exponential martingale to consider for changing the distribution of N from a counting process with intensity λ to a counting process with intensity μ.

Lemma 2.2.2. *Let T be a stopping time and assume that $\mathcal{E}(H \cdot M)^T$ is a uniformly integrable martingale. With Q being the probability measure with Radon-Nikodym derivative $\mathcal{E}(H \cdot M)_T$ with respect to P, it holds that N is a counting process under Q with intensity $1_{[0,T]}\mu + 1_{(T,\infty)}\lambda$. In particular, if $\mathcal{E}(H \cdot M)$ is a martingale, it holds for any $t \geq 0$ and with Q_t being the probability measure with Radon-Nikodym derivative $\mathcal{E}(H \cdot M)_t$ with respect to P that N is a counting process under Q_t with intensity $1_{[0,t]}\mu + 1_{(t,\infty)}\lambda$.*

In general, we cannot expect $\mathcal{E}(H \cdot M)$ to be a uniformly integrable martingale, only an ordinary martingale, because the distributions of counting processes with intensities which differ sufficiently in general will be singular. For example, the distributions of two homogeneous Poisson processes with different intensities are singular, see Proposition 3.24 of [92].

As an aside, note that the measure Q_T obtained in Lemma 2.2.2 of course always will be absolutely continuous with respect to P. A natural question to ask is when Q_T and P are equivalent. This is the case when the Radon-Nikodym derivative is almost surely positive. Lemma 2.2.3 gives a condition for this.

Lemma 2.2.3. *If the set of zeroes of μ has Lebesgue measure zero, $\mathcal{E}(H \cdot M)$ is almost surely positive.*

Finally, we state our sufficient criteria for $\mathcal{E}(H \cdot M)$ to be a true martingale. Defining $\log_+ x = \max\{0, \log x\}$ for $x \geq 0$, with the convention that the logarithm of zero is minus infinity, our main results are the following.

Theorem 2.2.4. *Assume that λ and μ are nonnegative, predictable and locally bounded. Assume that μ is λ-compatible. It holds that $\mathcal{E}(H \cdot M)$ is a martingale if there is an $\varepsilon > 0$ such that whenever $0 \leq u \leq t$ with $t - u \leq \varepsilon$, one of the following two conditions are satisfied:*

$$E \exp\left(\sum_{i=1}^{d} \int_u^t (\gamma_s^i \log \gamma_s^i - (\gamma_s^i - 1))\lambda_s^i \, \mathrm{d}s\right) < \infty \quad \textit{or} \tag{2.4}$$

$$E \exp\left(\sum_{i=1}^{d} \int_u^t \lambda_s^i \, \mathrm{d}s + \int_u^t \log_+ \gamma_s^i \, \mathrm{d}N_s^i\right) < \infty. \tag{2.5}$$

Corollary 2.2.5. *Assume that $\lambda = 1$ and assume that μ is nonnegative, predictable and locally bounded. Then μ is λ-compatible. It holds that $\mathcal{E}(H \cdot M)$ is a martingale if there is an $\varepsilon > 0$ such that whenever $0 \leq u \leq t$ with $t - u \leq \varepsilon$, one of the following two conditions are satisfied:*

$$E \exp\left(\sum_{i=1}^{d} \int_u^t \mu_s^i \log_+ \mu_s^i \, \mathrm{d}s\right) < \infty \quad \textit{or} \quad E \exp\left(\sum_{i=1}^{d} \int_u^t \log_+ \mu_s^i \, \mathrm{d}N_s^i\right) < \infty. \tag{2.6}$$

The immediate use of Theorem 2.2.4 and its corollary is as an existence result for nonexplosive counting processes with particular intensities, as the change of measure obtained from the martingale property of $\mathcal{E}(H \cdot M)$ yields the existence of a nonexplosive counting process distribution with given intensity μ on a bounded time interval $[0, t]$. That this is the case may be seen from Lemma 2.2.2, which shows that under the measure Q_t with Radon-Nikodym derivative $\mathcal{E}(H \cdot M)_t$ with respect to P, N is a counting process with intensity $1_{[0,t]}\mu + 1_{(t,\infty)}\lambda$.

Note that it is not necessarily possible to use the family (Q_t) to obtain the existence of a limiting probability measure Q_∞ under which N has intensity μ on all of $\mathbb{R}_+$. Such a limiting probability would require extension results such as discussed in the appendix of [55]. See also the discussion following Example 2.3.3.

As a specific application of our results, let us assume that we are interested in constructing a statistical model for a nonexplosive counting processes. We assume given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ and a d-dimensional counting process N such that under P, N has intensity $\lambda_t = 1$. Fix a timepoint t and let us assume that we are interested in considering a statistical model on the time interval $[0, t]$ based on a family of intensities $(\mu_\theta)_{\theta\in\Theta}$. If μ_θ satisfies the criteria of Corollary 2.2.5, then $\mathcal{E}(H_\theta \cdot M)$ is a martingale, and so $\mathcal{E}(H_\theta \cdot M)_t$ has unit mean. Letting Q_θ be the probability measure with Radon-Nikodym derivative $\mathcal{E}(H_\theta \cdot M)_t$ with respect to P, it holds that under Q_θ, N is a counting process with intensity, and the intensity is μ_θ on $[0, t]$. Furthermore, the family $(Q_\theta)_{\theta\in\Theta}$ is dominated by P, and the likelihood function is known in explicit form. Thus, Corollary 2.2.5 has allowed us to construct the statistical model and prove that explosion does not occur.

As regards checking the criteria in practice, an important property to note is that the criteria only need to be checked locally, in the sense that it is only necessary to find some $\varepsilon > 0$ such that the criteria hold for $0 \leq u \leq t$ with $t - u \leq \varepsilon$. This seemingly innocent property makes it possible to apply the criteria in several interesting situations. In particular, it allows us to extend the criterion (25) of [60] from $a > 1$ to $a \geq 1$, see Example 2.3.3.

Instead of considering Theorem 2.2.4 as a criterion for nonexplosion, we may also think of it simply as a sufficient criterion for the Doléans-Dade exponential $\mathcal{E}(M)$ of a particular type of local martingale M to be a true martingale. For M a local martingale with $\Delta M \geq -1$ and initial value zero, the question of when $\mathcal{E}(M)$ is a uniformly integrable martingale or a true martingale has been treated many times in the literature, see for example [123, 95, 94, 24] for results in the case of continuous M, and [109, 79, 89] for results in the general case.

In particular, a considerable family of criteria related to this problem is discussed in [89]. We remark that the proof of Theorem 2.2.4 applies Theorem III.1 and Theorem III.7 of [109]. In the parlance of [89], Theorem III.1 of [109] corresponds to condition $I(0, 1)$. There exists a slight improvement of condition $I(0, 1)$, namely condition $I(0, 1-)$, also proven in [89]. Applying this condition instead of condition $I(0, 1)$ does not lead to significant improvements of our results. We further remark that Theorem III.7 of [109] does not have an analogue in the hierarchy of [89]. In general, the criteria on which the results of Theorem 2.2.4 are built are among the strongest known, and optimality properties of these criteria are known. Therefore, we expect that no significant improvements of Theorem 2.2.4 are feasible.

The remainder of the paper is organized as follows. Section 2.3 gives some examples of applications of the results. In Section 2.4, we present the proof of the main results.

Section 2.5 contains supplementary results for Section 2.3.

2.3 Examples

In this section, we give examples where the conditions in Theorem 2.2.4 and Corollary 2.2.5 may be verified. Our first example shows how Theorem 2.2.4 under certain circumstances allows for changes of the intensity where the new intensity is an affine function of the old intensity. Such criteria were also discussed in Theorem 2 of [144], where the new intensity μ was assumed to be related to the initial intensity λ by the relationship $|\mu_t - \lambda_t| \leq \theta\sqrt{\lambda_t}$.

Example 2.3.1. Assume that d is equal to one. Assume that $\lambda_s \geq \delta$ for some $\delta > 0$ and that $\mu_t \leq \alpha + \beta\lambda_s$. If there is $\varepsilon > 0$ such that for $0 \leq u \leq t$ with $t - u \leq \varepsilon$, $\int_u^t \lambda_s \, \mathrm{d}s$ has an exponential moment of order $(1 + (\alpha\delta^{-1} + \beta)\log_+(\alpha\delta^{-1} + \beta))$, then $\mathcal{E}(H \cdot M)$ is a martingale.

Proof of Example 2.3.1. By our assumptions, $\gamma_t = \alpha\lambda_t^{-1} + \beta \leq \alpha\delta^{-1} + \beta$. Using that $x \log x - (x-1) \leq 1 + x \log x \leq 1 + x \log_+ x$ for any $x \geq 0$, we obtain

$$\int_u^t (\gamma_s^i \log \gamma_s^i - (\gamma_s^i - 1))\lambda_s \, \mathrm{d}s \leq (1 + (\alpha\delta^{-1} + \beta)\log_+(\alpha\delta^{-1} + \beta)) \int_u^t \lambda_s \, \mathrm{d}s, \quad (2.7)$$

so the first criterion of Theorem 2.2.4 yields the result. □

In the remainder of the examples, we assume that $\lambda = 1$, such that N is a d-dimensional standard Poisson process. We consider particular cases where Corollary 2.2.5 may be applied. For Example 2.3.2 below, we first introduce some notation. Let X be a semimartingale. If the quadratic variation process $[X]$ is locally integrable, the dual predictable projection $\Pi_p^*[X]$ is well-defined, see Definition 5.21 of [66] and Section III.5 of [134]. In this case, we put $\langle X \rangle = \Pi_p^*[X]$ and refer to $\langle X \rangle$ as the predictable quadratic variation process of X.

Example 2.3.2. Assume that μ is a nonnegative, predictable and locally integrable process, and assume that there is $\varepsilon > 0$ such that $\exp(\varepsilon\langle H \cdot M \rangle_t)$ is integrable for all $t \geq 0$. In this case, the first criterion of Corollary 2.2.5 may be applied to show that $\mathcal{E}(H \cdot M)$ is a martingale.

Proof of Example 2.3.2. Let $\varepsilon > 0$ be given such that $\exp(\varepsilon\langle H \cdot M \rangle_t)$ is integrable for all $t \geq 0$. Pick $K > 0$ so large that $x \log_+ x \leq \varepsilon(x-1)^2$ holds for $x \geq K$, then $E\exp(\sum_{i=1}^d \int_u^t \mu_s^i \log_+ \mu_s^i \, \mathrm{d}s) \leq \exp(dtC) E\exp(\varepsilon \sum_{i=1}^d \int_0^t (H_s^i)^2 \, \mathrm{d}s)$, where we define $C = \sup_{-1 \leq x \leq K} x \log_+ x$. As N has no common jumps, however, we have $[H \cdot M]_t = \sum_{i=1}^d \int_0^t (H^i)_s^2 \, \mathrm{d}N_s^i$. Therefore, as H is predictable, we obtain

$$\langle H \cdot M \rangle_t = \Pi_p^* \sum_{i=1}^d \int_0^t (H^i)_s^2 \, \mathrm{d}N_s^i = \sum_{i=1}^d \int_0^t (H_s^i)^2 \, \mathrm{d}\Pi_p^* N_s^i = \sum_{i=1}^d \int_0^t (H_s^i)^2 \, \mathrm{d}s. \quad (2.8)$$

All in all, we conclude

$$E\exp\left(\sum_{i=1}^{d}\int_u^t \mu_s^i \log_+ \mu_s^i \,\mathrm{d}s\right) \leq \exp(dtC) E\exp\left(\varepsilon\langle H\cdot M\rangle_t\right) < \infty, \tag{2.9}$$

and the result follows by Corollary 2.2.5. □

Example 2.3.2 is noteworthy because of the following. In [135], applying the results of [109], the following Novikov-type criterion is demonstrated: If M is a locally square integrable local martingale with $\Delta M \geq -1$ and $\exp(\frac{1}{2}\langle M^c\rangle_\infty + \langle M^d\rangle_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale. Here, M^c and M^d denote the continuous and purely discontinuous parts of the local martingale, respectively, see Theorem 7.25 of [66]. Furthermore, [135] argue by example that the constant 1 in front of $\langle M^d\rangle_\infty$ cannot in general be exchanged with $1-\varepsilon$ for any $\varepsilon > 0$. Example 2.3.2, however, shows that when proving the martingale property instead of the uniformly integrable martingale property, for the particular type of local martingale considered here, the constant 1 may in fact be exchanged with any positive number. This is a consequence of the particular form of $\langle H\cdot M\rangle$ combined with the fact that we are endeavouring to prove the martingale property and not the uniformly integrable martingale property.

Example 2.3.3. Assume that μ is a nonnegative, predictable and locally integrable process satisfying $\mu_t^i \leq \alpha + \beta\sum_{j=1}^d N_{t-}^j$. Then both criteria of Corollary 2.2.5 may be applied to obtain that $\mathcal{E}(H\cdot M)$ is a martingale.

Proof of Example 2.3.3. We begin by considering the use of the first moment condition of Corollary 2.2.5. As $x\log_+ x$ is increasing in x, it suffices to consider the case where $\alpha > 1$ and $\beta > 0$, such that μ is positive. Fix $\varepsilon > 0$, and let $0 \leq u \leq t$ with $t - u \leq \varepsilon$. We then obtain, with $N_t^S = \sum_{j=1}^d N_t^j$,

$$\exp\left(\sum_{i=1}^{d}\int_u^t \mu_s^i \log_+ \mu_s^i \,\mathrm{d}s\right) \leq \exp\left(\varepsilon d(\alpha+\beta N_t^S)\log(\alpha+\beta N_t^S)\right). \tag{2.10}$$

Now, for k large enough, $\varepsilon d(\alpha+\beta k)\log(\alpha+\beta k) \leq 4\varepsilon d\beta k\log k$. Therefore, we find that $\exp(\sum_{i=1}^d\int_u^t \mu_s^i\log\mu_s^i\,\mathrm{d}s)$ is integrable if only $\exp(4\varepsilon d\beta N_t^S\log N_t^S)$ is integrable. Now note that N_t^S is Poisson distributed with parameter dt, so by choosing ε with $4\varepsilon d\beta < 1$, we obtain the desired integrability using Lemma 2.5.1. The first moment condition of Corollary 2.2.5 now yields the result.

Consider instead using the second moment condition of Corollary 2.2.5. Again, it suffices to consider $\alpha > 1$ and $\beta > 0$. We fix $\varepsilon > 0$ and consider $0 \leq u \leq t$ satisfying $|t-u| \leq \varepsilon$. We put $N_t^S = \sum_{j=1}^d N_t^j$ and define a mapping $\varphi : \mathbb{N}_0 \to \mathbb{R}$ by $\varphi(n) = E\exp(\int_0^{t-u}\log\beta(n+N_s^S)\,\mathrm{d}N_s^S)$. Let $m \in \mathbb{N}$ be such that $\alpha \leq \beta m$, we then obtain $\alpha + \beta x \leq \beta(m+x)$. As N^S is a Poisson process of rate d, we find that

conditionally on N_u^S, the processes $s \mapsto N_{s+u}^S - N_u^S$ and $s \mapsto N_s^S$ have the same distribution. Therefore, we obtain by conditioning on N_u^S that

$$E\exp\left(\sum_{i=1}^{d}\int_u^t \log\mu_s^i \,\mathrm{d}N_s^i\right) \le E\exp\left(\sum_{i=1}^{d}\int_u^t \log\beta\left(m+\sum_{j=1}^{d}N_s^j\right)\,\mathrm{d}N_s^i\right) \tag{2.11}$$
$$= E\exp\left(\int_u^t \log\beta(m+N_s^S)\,\mathrm{d}N_s^S\right) = E\exp\left(\int_u^t \log\beta(m+N_u^S+N_s^S-N_u^S)\,\mathrm{d}N_s^S\right)$$
$$= E\exp\left(\int_0^{t-u} \log\beta(m+N_u^S+N_{s+u}^S-N_u^S)\,\mathrm{d}(N_{s+u}^S-N_u^S)\right) = E\varphi(m+N_u^S).$$

Now note that

$$\begin{aligned}\varphi(n) &= E\exp\left(\int_0^{t-u}\log\beta(n+N_s^S)\,\mathrm{d}N_s^S\right) = E\exp\left(\sum_{k=1}^{N_{t-u}^S}\log\beta(n+k)\right)\\ &= \sum_{p=0}^{\infty}\exp\left(\sum_{k=1}^{p}\log\beta(n+k)\right)\frac{((t-u)d)^p}{p!}\exp(-(t-u)d)\\ &= \exp(-(t-u)d)\sum_{p=0}^{\infty}\beta^p\left(\prod_{k=1}^{p}(n+k)\right)\frac{((t-u)d)^p}{p!}\\ &= \exp(-(t-u)d)\sum_{p=0}^{\infty}(\beta(t-u)d)^p\frac{(n+p)!}{n!p!}.\end{aligned} \tag{2.12}$$

Whenever $|x|<1$, we have $\sum_{p=0}^{\infty}x^p\frac{(n+p)!}{n!p!} = (1-x)^{-(n+1)}$ by formula (15.1.8) of [3], and we therefore conclude, whenever $\beta(t-u)d<1$, that

$$\varphi(n) = \frac{\exp(-(t-u)d)}{(1-\beta(t-u)d)^{n+1}}. \tag{2.13}$$

Therefore, in this case,

$$\begin{aligned}E\exp\left(\sum_{i=1}^{d}\int_u^t\log\mu_s^i\,\mathrm{d}N_s^i\right) &\le E\varphi(m+N_u^S) = E\frac{\exp(-(t-u)d)}{(1-\beta(t-u)d)^{m+N_u^S+1}}\\ &= \exp(-td)\sum_{p=0}^{\infty}(1-\beta(t-u)d)^{-(p+m+1)}\frac{(ud)^p}{p!}\\ &= \frac{1}{(1-\beta(t-u)d)^{m+1}}\exp\left(-td+\frac{ud}{1-\beta(t-u)d}\right).\end{aligned} \tag{2.14}$$

We conclude that the second moment condition of Corollary 2.2.5 yields the result, using ε such that $\beta\varepsilon d<1$. □

The above is the extension of criterion (25) of [60] from $a > 1$ to $a \geq 1$ mentioned earlier. For the case of intensities predictable with respect to the filtration generated by N, the existence of nonexplosive counting processes with intensities affinely bounded by the total number of jumps as in Example 2.3.3 is well known, see Example 4.4.5 of [81]. The abstract construction of Example 2.3.3 covers the general case of intensities predictable with respect to $(\mathcal{F}_t)$ and yields a family $(\Omega, \mathcal{F}_t, Q_t)_{t\geq 0}$ of probability spaces such that $(N_s)_{s\leq t}$ has intensity μ on $[0,t]$ under Q_t, here Q_t is the measure with Radon-Nikodym derivative $\mathcal{E}(H \cdot M)_t$ with respect to P. Additional structure on the probability space is needed to guarantee the existence of the inverse limit $(\Omega, \sigma(\cup_{t\geq 0}\mathcal{F}_t), Q)$ with the restriction of Q to $\mathcal{F}_t$ being equal to Q_t, such that $(N_s)_{s\geq 0}$ has intensity μ under Q, see [25] and [55].

If the process μ is exactly affine in the sense that $\mu_t^i = \alpha + \beta \sum_{j=1}^d N_{t-}^j$, the martingale property of $\mathcal{E}(H \cdot M)$ may be obtained by direct calculation. However, this does not in itself imply that the same result holds when we only have the inequality $\mu_t^i \leq \alpha + \beta \sum_{j=1}^d N_{t-}^j$. In general, such "monotonicity" properties of the martingale property for exponential martingales do not hold, see for example [94], Example 1.13.

Next, we consider two examples involving intensities given as solutions to stochastic differential equations. In both cases, we assume given a Brownian motion relative to the given filtration $(\mathcal{F}_t)$, meaning in the d-dimensional case that $(W^i)_t^2 - t$ is an $(\mathcal{F}_t)$ martingale for $i \leq d$ and $W_t^i W_t^j$ is an $(\mathcal{F}_t)$ martingale for $i, j \leq d$ with $i \neq j$. We call such a process an $(\mathcal{F}_t)$-Brownian motion. By Lévy's characterisation of Brownian motion for general filtered probability spaces, see Theorem IV.33.1 of [143], this requirement ensures that the characteristic properties of the Brownian motion interact well with the filtration $(\mathcal{F}_t)$. By $\mathbb{M}(d,d)$, we denote the set of $d \times d$ matrices with real entries.

Example 2.3.4. Consider mappings $A : \mathbb{N}_0^d \times \mathbb{R}_+^d \to \mathbb{R}^d$, $B : \mathbb{N}_0^d \times \mathbb{R}_+^d \to \mathbb{M}(d,d)$ and $\sigma : \mathbb{N}_0^d \times \mathbb{R}_+^d \to \mathbb{M}(d,d)$ such that for all $\eta \in \mathbb{N}_0^d$, $A(\eta, \cdot)$, $B(\eta, \cdot)$ and $\sigma(\eta, \cdot)$ are continuous and bounded and such that σ always is positive definite. With T_n^i denoting the n'th jump time for N^i and $Z_t^i = t - T_{N_t^i}^i$, let X be a solution to the d-dimensional stochastic differential equation

$$\mathrm{d}X_t = (A(N_t, Z_t) + B(N_t, Z_t)X_t)\,\mathrm{d}t + \sigma(N_t, Z_t)\,\mathrm{d}W_t \tag{2.15}$$

with initial value x_0 in $\mathbb{R}^d$, where W is an $(\mathcal{F}_t)$ Brownian motion independent of N. Let $\phi : \mathbb{R}^d \to \mathbb{R}_+^d$ be Lipschitz and put $\mu_t = \phi(X_t)$. Assume that $\delta > 0$ and $c_A, c_B, c_\sigma > 0$ exist such that

$$\sup_{t\geq 0} \|A(\eta, t)\|_2 \leq c_A \|\eta\|_1^{1-\delta} \tag{2.16}$$

$$\sup_{t\geq 0} \|\sigma(\eta, t)\|_2 \leq c_\sigma \|\eta\|_1^{(1-\delta)/2} \tag{2.17}$$

$$\sup_{t\geq 0} \|B(\eta, t)\|_2 \leq c_B, \tag{2.18}$$

where $\|\cdot\|_2$ in the first case denotes the Euclidean norm and in the two latter cases denote the operator norm induced by the Euclidean norm, and $\|\cdot\|_1$ denotes the $\mathcal{L}^1$ norm in $\mathbb{N}_0^d$. Then, the first criterion of Corollary 2.2.5 may be applied to obtain that $\mathcal{E}(H \cdot M)$ is a martingale.

Proof of Example 2.3.4. We need to show that the first criterion of Corollary 2.2.5 is applicable. We may assume without loss of generality that $0 < \delta < 1$. It suffices to prove that for any $t > 0$, $E\exp(\sum_{i=1}^d \int_0^t \mu_s^i \log_+ \mu_s^i \,\mathrm{d}s)$ is finite. Fix $t > 0$. By Jensen's inequality, we find

$$E\exp\left(\sum_{i=1}^d \int_0^t \mu_s^i \log_+ \mu_s^i \,\mathrm{d}s\right) \leq \frac{1}{t}\int_0^t E\exp\left(t\sum_{i=1}^d \mu_s^i \log_+ \mu_s^i\right) \mathrm{d}s. \tag{2.19}$$

We wish to bound the expectation inside the integral by an expression depending continuously on s. Recall that we have assumed that ϕ is Lipschitz, so there exists $\gamma > 0$ such that $\|\phi(x)\|_\infty \leq \gamma\|x\|_2$, yielding $\phi^i(x) \leq \gamma\|x\|_2$ for all $i \leq d$, and so $E\exp(t\sum_{i=1}^d \mu_s^i \log_+ \mu_s^i) \leq E\exp(td\gamma\|X_s\|_2 \log_+ \gamma\|X_s\|_2)$. Next, let $0 < \zeta < 1$. It holds for all $x \geq 0$ that $\log_+ x \leq \zeta^{-1}x^\zeta$. Therefore, defining $\rho = td\gamma^{1+\zeta}\zeta^{-1}$, we conclude

$$E\exp\left(t\sum_{i=1}^d \mu_s^i \log_+ \mu_s^i\right) \leq E\exp\left(\rho\|X_s\|_2^{\zeta+1}\right). \tag{2.20}$$

We will calculate this expectation by conditioning on N. Let η denote a counting process path, and let (τ_n) denote the event times of η. By the explicit representation in Lemma 2.5.2 as well as the results on pathwise stochastic integration in [90], it holds that conditionally on $N = \eta$, X_s has the same distribution as Y_s^η, where

$$Y_s^\eta = C_s^{-1}\left(x_0 + \int_0^s C_v A(\eta_v, v - \tau_{\eta_v})\,\mathrm{d}v + \int_0^s C_v \sigma(\eta_v, v - \tau_{\eta_v})\,\mathrm{d}W_v\right), \tag{2.21}$$

which is a normal distribution with mean ξ_s^η and variance Σ_s^η, where

$$\xi_s^\eta = C_s^{-1}\left(x_0 + \int_0^s C_v A(\eta_v, v - \tau_{\eta_v})\,\mathrm{d}v\right) \tag{2.22}$$

$$\Sigma_s^\eta = C_s^{-1}\int_0^s (C_v\sigma(\eta_v, v - \tau_{\eta_v}))^t (C_v\sigma(\eta_v, v - \tau_{\eta_v}))\,\mathrm{d}s (C^{-1})_s^t, \tag{2.23}$$

and where $C_s = \exp(-\int_0^s B(\eta_v, v - \tau_{\eta_v})\,\mathrm{d}v)$. With $\|\cdot\|_2$ denoting the matrix operator norm induced by the Euclidean norm, Lemma 2.5.3 yields

$$\begin{aligned} E\exp(\rho\|X_s\|_2^{1+\zeta}) &= \int E\left(\exp(\rho\|X_s\|_2^{1+\zeta})\middle| N = \eta\right) \mathrm{d}N(P)(\eta) \\ &\leq k_d E\exp(a(\rho,\zeta)\|\xi_s^N\|^{1+\zeta}) \exp\left(b(\rho,\zeta)\|\Sigma_s^N\|_2^{\frac{1+\zeta}{1-\zeta}}\right), \end{aligned} \tag{2.24}$$

with a and b as in the statement of the lemma. Next, we consider bounds for $\|\xi_s^\eta\|$ and $\|\Sigma_s^\eta\|_2$. Note that $\|C_s\|_2 \leq \exp(\int_0^s \|B(\eta_v, v - \tau_{\eta_v})\|_2 \,\mathrm{d}v) \leq \exp(sc_B)$, where we have applied standard norm inequalities, see Theorem 10.10 of [68] and Lemma 1.4 of [51], and similarly, $\|C_s^{-1}\|_2 \leq \exp(sc_B)$. Therefore, recalling that $0 < \delta < 1$ so that $x \mapsto x^{1-\delta}$ is increasing,

$$\|\xi_s^\eta\|_2 \leq \exp(sc_B) \left(\|x_0\|_2 + sc_A \exp(sc_B) \|\eta_s\|_1^{1-\delta} \right). \tag{2.25}$$

Similarly, we obtain

$$\|\Sigma_s^\eta\|_2 \leq s \exp(4sc_B) c_\sigma^2 \|\eta_s\|_1^{1-\delta}. \tag{2.26}$$

In particular, for appropriate continuous functions a_ξ, b_ξ and b_Σ from $\mathbb{R}_+$ to $\mathbb{R}$, depending on ζ, we obtain the two bounds

$$\|\xi_s^\eta\|_2^{1+\zeta} \leq a_\xi(s) + b_\xi(s) \|\eta_s\|_1^{(1-\delta)(1+\zeta)} \tag{2.27}$$

$$\|\Sigma_s^\eta\|_2^{\frac{1+\zeta}{1-\zeta}} \leq b_\Sigma(s) \|\eta_s\|_1^{(1-\delta)\frac{1+\zeta}{1-\zeta}}. \tag{2.28}$$

We then conclude

$$\begin{aligned}
&E \exp(\rho \|X_s\|_2^{1+\zeta}) \\
&\leq k_d E \exp\left(a(\rho,\zeta) \left(a_\xi(s) + b_\xi(s) \|N_s\|_1^{(1-\delta)(1+\zeta)} \right) + b(\rho,\zeta) b_\Sigma(s) \|N_s\|_1^{(1-\delta)\frac{1+\zeta}{1-\zeta}} \right) \\
&\leq k_d \exp\left(a(\rho,\zeta) a_\xi(s) \right) E \exp\left((a(\rho,\zeta) b_\xi(s) + b(\rho,\zeta) b_\Sigma(s)) \|N_s\|_1^{(1-\delta)\frac{1+\zeta}{1-\zeta}} \right).
\end{aligned} \tag{2.29}$$

The above depends on given constants δ, c_A, c_B and c_σ, as well as the constant ζ which we may choose arbitrarily in $(0,1)$. We now choose ζ so small in $(0,1)$ that $(1-\delta)(1+\zeta)(1-\zeta)^{-1} \leq 1$. Recalling that for any Poisson distributed variable Z with intensity λ and any $c \in \mathbb{R}$, it holds that $E \exp(cZ) = \exp((\exp(c) - 1)\lambda)$, we may conclude

$$\begin{aligned}
&E \exp(\rho \|X_s\|_2^{1+\zeta}) \\
&\leq k_d \exp\left(a(\rho,\zeta) a_\xi(s) \right) E \exp\left((a(\rho,\zeta) b_\xi(s) + b(\rho,\zeta) b_\Sigma(s)) \|N_s\|_1 \right) \\
&= k_d \exp\left(a(\rho,\zeta) a_\xi(s) \right) \exp((\exp(a(\rho,\zeta) b_\xi(s) + b(\rho,\zeta) b_\Sigma(s)) - 1) ds).
\end{aligned} \tag{2.30}$$

All in all, we may now define, for $0 \leq s \leq t$,

$$\varphi(s) = k_d \exp\left(a(\rho,\zeta) a_\xi(s) \right) \exp((a(\rho,\zeta) \exp(b_\xi(s) + b(\rho,\zeta) b_\Sigma(s)) - 1) ds), \tag{2.31}$$

and obtain $E \exp(t \sum_{i=1}^d \mu_s^i \log_+ \mu_s^i) \leq \varphi(s)$ for all such s. The functions a_ξ, b_ξ and b_Σ depends continuously on s. Therefore, φ is a continuous function of s. In particular, the integral of φ over $[0,t]$ is finite. Recalling our first estimates, this leads us to conclude that for any $t \geq 0$, it holds that $E \exp(\sum_{i=1}^d \int_0^t \mu_s^i \log_+ \mu_s^i \,\mathrm{d}s)$ is finite, and so the first integrability criterion of Corollary 2.2.5 is satisfied. □

Example 2.3.5. Let $(\xi_n)_{n\geq 0}$, $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ be sequences in $\mathbb{R}$. Assume that $b_n \neq 0$ for $n \geq 0$ and assume that X satisfies the one-dimensional stochastic differential equation

$$\mathrm{d}X_t = a_{N_t} + b_{N_t} X_t \,\mathrm{d}t + \sigma \,\mathrm{d}W_t + (\xi_{N_t} - X_{t-}) \,\mathrm{d}N_t, \tag{2.32}$$

with initial value ξ_0 and $\sigma > 0$, where W is an $(\mathcal{F}_t)$ Brownian motion independent of N. Put $\mu_t = |X_{t-}|$. Assume that there are $\alpha, \beta > 0$ such that

$$|\xi_n| \leq \alpha + \beta n \tag{2.33}$$

$$|a_n/b_n| \leq \alpha + \beta n \tag{2.34}$$

$$|b_n| \leq \alpha. \tag{2.35}$$

Then, the second criterion of Corollary 2.2.5 may be applied to obtain that $\mathcal{E}(H \cdot M)$ is a martingale.

Proof of Example 2.3.5. We want to show that the second moment condition of Corollary 2.2.5 is applicable. To this end, we first construct an explicit solution to the stochastic differential equation defining X. With T_n denoting the n'th event time for N, define the process W^n by $W_t^n = W_{T_n+t} - W_{T_n}$ and define $\mathcal{F}_t^n = \mathcal{F}_{T_n+t}$. By Theorem I.12.1 of [142], W^n is independent of $\mathcal{F}_{T_n}$ and has the distribution of a Brownian motion. Again using Theorem I.12.1 of [142] with the stopping time $T_n + s$, we have for $0 \leq s \leq t$ that

$$\begin{aligned} E(W_t^n | \mathcal{F}_s^n) &= E(W_{T_n+t} - W_{T_n} | \mathcal{F}_{T_n+s}) \\ &= E(W_{T_n+t} - W_{T_n+s} | \mathcal{F}_{T_n+s}) + W_{T_n+s} - W_{T_n} \\ &= W_{T_n+s} - W_{T_n} = W_s^n, \end{aligned} \tag{2.36}$$

and Lévy's characterisation Theorem for Brownian motion relative to a filtration, see [143], Theorem IV.33.1, shows that W^n is an $(\mathcal{F}_t^n)$-Brownian motion. We may then use the Itô existence and uniqueness theorem, see Theorem 11.2 of [143], concluding that on the same probability space that carries the Poisson process N, the Brownian motion W and in particular the $(\mathcal{F}_t^n)$-Brownian motion W^n, there exist unique processes X^n satisfying $\mathrm{d}X_t^n = a_n + b_n X_t^n \,\mathrm{d}t + \sigma \,\mathrm{d}W_t^n$ with constant initial values ξ_n. Whenever $T_n \leq t < T_{n+1}$, we then have

$$\begin{aligned} X_{t-T_n}^n &= \xi_n + \int_0^{t-T_n} a_n + b_n X_s^n \,\mathrm{d}s + \int_0^{t-T_n} \sigma \,\mathrm{d}W_s^n \\ &= \xi_n + \int_{T_n}^{t} a_n + b_n X_{s-T_n}^n \,\mathrm{d}s + \int_{T_n}^{t} \sigma \,\mathrm{d}W_s. \end{aligned} \tag{2.37}$$

The process $\sum_{n=0}^{\infty} X_{t-T_n}^n 1_{[T_n, T_{n+1})}(t)$ thus satisfies the same stochastic differential equation as X. By pathwise uniqueness for each X^n, $X_t = \sum_{n=0}^{\infty} X_{t-T_n}^n 1_{[T_n, T_{n+1})}(t)$.

The above deliberations yield an explicit representation for the stochastic differential equation defining the intensity. Next, we check that the second moment condition

of Corollary 2.2.5 is applicable. With $S_k = T_k - T_{k-1}$ denoting the sequence of interarrival times, we then obtain for the moment condition to be investigated that

$$E\exp\left(\int_u^t \log_+ |X_{s-}|\,\mathrm{d}N_s\right) \leq E\exp\left(\int_u^t \log(1+|X_{s-}|)\,\mathrm{d}N_s\right) \tag{2.38}$$
$$= E\prod_{k=N_u+1}^{N_t}(1+|X^{k-1}_{T_k-T_{k-1}}|) = E\prod_{k=N_u+1}^{N_t}(1+|X^{k-1}_{S_k}|).$$

In order to obtain the finiteness of this expression, we wish to condition on N. Given a counting process trajectory η, we refer to the event times of η by (τ_n), $\tau_0 = 0$, and we let (s_n) be the corresponding interarrival times, $s_n = \tau_n - \tau_{n-1}$. We then have

$$E\prod_{k=N_u+1}^{N_t}(1+|X^{k-1}_{S_k}|) = \int E\left(\prod_{k=\eta_u+1}^{\eta_t}(1+|X^{k-1}_{s_k}|)\,\middle|\, N=\eta\right)\mathrm{d}N(P)(\eta). \tag{2.39}$$

Next, we argue that given N, the variables $(X^{k-1}_{s_k})_{k\geq 1}$ are mutually independent, in the sense that it $N(P)$ almost surely holds that the conditional distribution of the variables $(X^{k-1}_{s_k})_{k\geq 1}$ given $N=\eta$ is the product measure of each of the marginal conditional distributions.

Applying Theorem V.10.4 of [143] and the Doob-Dynkin Lemma, see the first lemma of Section A.IV.3 of [41], there is a measurable mapping $G_{k-1} : C[0,s_k] \to \mathbb{R}$ such that $X^{k-1}_{s_k}$ is the transformation under G_{k-1} of the first s_k coordinates of W^{k-1}. We apply this result to obtain the conditional independence of $X^{k-1}_{s_k}$ given $N=\eta$. As $X^{k-1}_{s_k}$ is a transformation of $(W^{k-1})^{s_k}$, it will suffice to show that the processes $(W^{k-1})^{s_k}$ are conditionally independent given $N=\eta$. To this end, we recall that W is independent of N, and note that $(W^{k-1})^{s_k}_t = W_{(\tau_{k-1}+t)\wedge\tau_k} - W_{\tau_{k-1}}$. Therefore, $(W^{k-1})^{s_k}$ is $\mathcal{F}_{\tau_k}$ measurable. By Theorem I.12.1 of [142], W^{k-1} is independent of $\mathcal{F}_{\tau_{k-1}}$. Inductively, it follows that conditionally on $N=\eta$, the sequence of processes $(W^{k-1})^{s_k}$ are mutually independent. Therefore, conditionally on N, the variables $(X^{k-1}_{s_k})_{k\geq 1}$ are mutually independent.

Applying this conditional indepedence, we may now conclude

$$E\prod_{k=N_u+1}^{N_t}(1+|X^{k-1}_{S_k}|) = \int E\left(\prod_{k=\eta_u+1}^{\eta_t}(1+|X^{k-1}_{s_k}|)\,\middle|\, N=\eta\right)\mathrm{d}N(P)(\eta)$$
$$= E\prod_{k=N_u+1}^{N_t} E(1+|X^{k-1}_{S_k}||N). \tag{2.40}$$

Next, we develop a simple bound for $E(|X^{k-1}_{S_k}||N)$. Consider again a counting process path η, we then almost surely have $E(|X^{k-1}_{S_k}||N=\eta) = E|X^{k-1}_{s_k}|$, where $X^{k-1}_{s_k}$ is given by $X^{k-1}_{s_k} = \xi_{k-1} + \int_0^{s_k} a_{k-1} + b_{k-1}X^{k-1}_t\,\mathrm{d}t + \sigma W^{k-1}_{s_k}$. By (3.42) of [61], we then

find that $X_{s_k}^{k-1}$ is normally distributed with mean and variance given by

$$EX_{s_k}^{k-1} = -\frac{a_{k-1}}{b_{k-1}} + \exp(s_k b_{k-1}) \left(\xi_{k-1} + \frac{a_{k-1}}{b_{k-1}} \right). \tag{2.41}$$

$$VX_{s_k}^{k-1} = \sigma^2 \int_0^{s_k} \exp(2b_{k-1}(s_k - u)) \, \mathrm{d}u. \tag{2.42}$$

By our assumptions on a_k, b_k and ξ_k, we then obtain

$$\begin{aligned} E|X_{s_k}^{k-1}| &\leq |EX_{s_k}^{k-1}| + \sqrt{VX_{s_k}^{k-1}} E(X_{s_k}^{k-1} - EX_{s_k}^{k-1}) / \sqrt{VX_{s_k}^{k-1}} \\ &\leq \left| \frac{a_{k-1}}{b_{k-1}} \right| + \exp(s_k b_{k-1}) \left(|\xi_{k-1}| + \left| \frac{a_{k-1}}{b_{k-1}} \right| \right) + \sqrt{2/\pi} \sigma \sqrt{s_k} \exp(2 s_k b_{k-1}) \\ &\leq \alpha + \beta(k-1) + 2\exp(s_k \alpha)(\alpha + \beta(k-1)) + \sqrt{2/\pi} \sigma \sqrt{s_k} \exp(2 s_k \alpha). \end{aligned} \tag{2.43}$$

Therefore, we see that by defining $\alpha^*(v) = \alpha + 2\alpha \exp(v\alpha) + \sqrt{2/\pi}\sigma\sqrt{v}\exp(2v\alpha)$ and $\beta^*(v) = \beta + 2\beta\exp(v\alpha)$, we have $E|X_{s_k}^{k-1}| \leq \alpha^*(s_k) + \beta^*(s_k)(k-1)$. Next, note that for $k \leq N_t$, it holds that $T_k \leq T_{N_t} \leq t$. Therefore, for any k with $N_u + 1 \leq k \leq N_t$, it holds that $S_k \leq t$. As α^* and β^* are increasing, we then find

$$\begin{aligned} E \prod_{k=N_u+1}^{N_t} E(1 + |X_{S_k}^{k-1}||N) &\leq E \prod_{k=N_u+1}^{N_t} (1 + \alpha^*(S_k) + \beta^*(S_k)(k-1)) \\ &\leq E \prod_{k=N_u+1}^{N_t} (1 + \alpha^*(t) + \beta^*(t)(k-1)) \\ &= E \exp\left(\int_u^t \log(1 + \alpha^*(t) + \beta^*(t) N_{s-}) \, \mathrm{d}N_s \right). \end{aligned} \tag{2.44}$$

Proceeding as in the the proof of Example 2.3.3 using the second moment condition of Corollary 2.2.5, it follows that for $\varepsilon > 0$ small enough and $0 \leq u \leq t$ with $t - u \leq \varepsilon$, the above is finite, and so the moment condition is satisfied. □

Examples 2.3.4 and 2.3.5 show how Corollary 2.2.5 may be used to construct counting processes with intensities not adapted to the filtration induced by N itself. Note that by Corollary 11.5.3 of [158], W is always independent of N, so the independence requirements in the above are mentioned only for clarity. Also note that in Example 2.3.4, the required bounds on the coefficients hold independently of the norms on $\mathbb{N}_0^d$, $\mathbb{R}^d$ and $\mathbb{M}(d,d)$ chosen, since all norms on finite-dimensional vector spaces are equivalent.

The interpretation of the two examples are as follows. In Example 2.3.4, the intensity is a transformed diffusion process with mean reversion level, mean reversion speed and diffusion coefficient which are deterministic between jumps. A simple example may be obtained as follows. Let X be a solution to the one-dimensional stochastic differential equation

$$\mathrm{d}X_t = \beta(\alpha \exp(-\gamma(t - T_{N_t})) - X_t) \, \mathrm{d}t + \sigma \, \mathrm{d}W_t, \tag{2.45}$$

where $\alpha, \beta, \gamma \geq 0$ and T_n is the n'th event time of N. Define $\mu_t = |X_t|$. μ is then a process of the type given in Example 2.3.4. Except when X is nonpositive, μ behaves as a diffusion immediately after each jump of N, with a mean reversion level α, reverting to this level at rate β, and furthermore, the mean reversion level decreases exponentially with rate γ in $t - T_{N_t}$, which is the time since the last jump of N.

In Example 2.3.5, the intensity is the absolute value of a linear diffusion process with constant coefficients between jumps. Furthermore, the intensity is reset to the level ξ_n at the n'th jump of N.

Example 2.3.6. Consider mappings $\phi_i : \mathbb{R} \to [0, \infty)$ and $h_{ij} : [0, \infty) \to \mathbb{R}$. Define

$$\mu_t^i = \phi_i \left(\sum_{j=1}^{d} \int_0^{t-} h_{ij}(t-s) \, \mathrm{d}N_s^j \right). \tag{2.46}$$

If ϕ^i is Borel measurable with $\phi_i(x) \leq |x|$ and h_{ij} is bounded, then $\mathcal{E}(H \cdot M)$ is a martingale.

Proof of Example 2.3.6. By Lemma 2.5.4, the process $\sum_{j=1}^{d} \int_0^{t-} h_{ij}(t-s) \, \mathrm{d}N_s^j$ is predictable. As ϕ^i is Borel measurable, it then follows that μ^i is predictable. As ϕ^i is nonnegative, μ is nonnegative. And by stopping at event times, we find that μ is locally bounded. Thus, μ is nonnegative, predictable and locally bounded. Letting $c > 0$ be such that $\|h_{ji}\|_\infty \leq c$ for all $i, j \leq d$, we obtain

$$\mu_t^i \leq \left| \sum_{j=1}^{d} \int_0^{t-} h_{ij}(t-s) \, \mathrm{d}N_s^j \right| \leq \sum_{j=1}^{d} \int_0^{t-} |h_{ij}(t-s)| \, \mathrm{d}N_s^j \leq c \sum_{j=1}^{d} N_{t-}^j, \tag{2.47}$$

and the result follows from Example 2.3.3. □

Example 2.3.6 yields a change of measure to a probability measure where the counting process is a multidimensional Hawkes process. In general, many specifications of ϕ and h will yield exploding counting processes and there will exist no measure change yielding the required intensity change.

The above examples all give various types of sufficient criteria for the martingale property of $\mathcal{E}(H \cdot M)$ using Corollary 2.2.5. As an aside, we may ask whether the classical necessary and sufficient criterion for nonexplosion for piecewise constant intensities, see Theorem 2.3.2 of [121], may be replicated as a criterion for the martingale property of $\mathcal{E}(H \cdot M)$. The following example shows that this is the case.

Example 2.3.7. Let $d = 1$, let (α_n) be a sequence of positive numbers and let $\mu_t = \alpha_{N_{t-}}$. Then $\mathcal{E}(H \cdot M)$ is a martingale if and only if $\sum_{n=0}^{\infty} \frac{1}{\alpha_n}$ is divergent.

Proof of Example 2.3.7. Let T_n be the n'th jump time of N, then (T_n) is a localising

sequence. We have

$$\begin{aligned} E\mathcal{E}(\mu \cdot M - M)_{T_n} &= E \exp\left(T_n - \int_0^{T_n} \mu_s \, \mathrm{d}s + \int_0^{T_n} \log \mu_s \, \mathrm{d}N_s \right) \\ &= E \exp\left(-\sum_{k=1}^{n} (\alpha_{k-1} - 1)(T_n - T_{n-1}) + \sum_{k=1}^{n} \log \alpha_{k-1} \right) \\ &= \prod_{k=1}^{n} \alpha_{k-1}(1 - (1 - \alpha_{k-1}))^{-1} = 1, \end{aligned} \tag{2.48}$$

so $\mathcal{E}(M)^{T_n}$ is a uniformly integrable martingale by Lemma 2.4.2. Therefore, by Lemma 2.5.5, $\mathcal{E}(\mu \cdot M - M)$ is a martingale if and only if $\lim_n E\mathcal{E}(\mu \cdot M - M)_{T_n} 1_{(T_n \leq t)}$ is zero for all $t \geq 0$. Now let $(\Omega', \mathcal{F}', P')$ be an auxiliary probability space endowed with a sequence (U_n) of independent exponentially distributed variables, where U_n has intensity α_n. Let P_n be the measure with Radon-Nikodym derivative $\mathcal{E}(\mu \cdot M - M)_{T_n}$ with respect to P. By Lemma 2.2.2, under P_n, N has intensity $\mu 1_{[0,T_n]} + 1_{(T_n,\infty)}$. In particular, the distribution of T_n under P_n is then the same as the distribution of $\sum_{k=1}^{n} U_k$ under P', and so

$$\begin{aligned} \lim_n E\mathcal{E}(M)_{T_n} 1_{(T_n \leq t)} &= \lim_n P_n(T_n \leq t) \\ &= \lim_n P'\left(\sum_{k=1}^{n} U_k \leq t \right) = P'\left(\sum_{k=1}^{\infty} U_k \leq t \right), \end{aligned} \tag{2.49}$$

since $\cap_{n=1}^{\infty}(\sum_{k=1}^{n} U_k \leq t) = (\sum_{k=1}^{\infty} U_k \leq t)$. Now, as $\sum_{k=1}^{\infty} \frac{1}{\alpha_k}$ diverges, Theorem 2.3.2 of [121] shows that $\sum_{k=1}^{\infty} U_k$ is almost surely infinite, so $P'(\sum_{k=1}^{\infty} U_k \leq t) = 0$. The result now follows from Lemma 2.5.5. □

2.4 Proofs of the main results

In this section, we present the proofs of the results stated in Section 2.2. We begin by recalling some folklore results on supermartingales and exponential martingales. For completeness, we give proofs of these results as well.

Lemma 2.4.1. *Let X be a nonnegative supermartingale. Then X is a uniformly integrable martingale if and only if $EX_\infty = EX_0$, and X is a martingale if and only if it holds for all $t \geq 0$ that $EX_t = EX_0$.*

Proof. First note that for a nonnegative supermartingale X, $0 \leq EX_t \leq EX_0$ for all $t \geq 0$. Therefore, $(X_t)_{t \geq 0}$ is bounded in $\mathcal{L}^1$, and so X_∞, the almost sure limit of X_t, always exists, see Theorem II.69.1 of [142]. Now, if X is a uniformly integrable martingale, it is immediate that $EX_\infty = EX_0$. Conversely, assume $EX_\infty = EX_0$. Fix $t \geq 0$. It then holds that $EX_t = EX_\infty$, yielding $E(X_t - E(X_\infty|\mathcal{F}_t)) = 0$. As $X_t \geq E(X_\infty|\mathcal{F}_t)$, this implies $X_t = E(X_\infty|\mathcal{F}_t)$, so X is a uniformly integrable martingale. The martingale case follows by considering the stopped process X^t. □

Recall that if M is a local martingale with $\Delta M \geq -1$ and initial value zero, $\mathcal{E}(M)$ is a nonnegative local martingale and a supermartingale, $E\mathcal{E}(M)_t \leq 1$ and $\mathcal{E}(M)_\infty$ always exists as an almost sure limit with $E\mathcal{E}(M)_\infty \leq 1$. Applying Lemma 2.4.1 to the case of Doléans-Dade exponentials then yields the following useful result.

Lemma 2.4.2. *Let M be a local martingale with $\Delta M \geq -1$ and initial value zero. $\mathcal{E}(M)$ is a uniformly integrable martingale if and only if $E\mathcal{E}(M)_\infty = 1$, and $\mathcal{E}(M)$ is a martingale if and only if $E\mathcal{E}(M)_t = 1$ for all $t \geq 0$.*

Now consider given a d-dimensional nonexplosive counting process N with nonnegative, predictable and locally bounded intensity λ as well as another nonnegative, predictable and locally bounded process μ which is λ-compatible. As in Section 2.2, M is the d-dimensional local martingale defined by $M_t^i = N_t^i - \int_0^t \lambda_s^i \, \mathrm{d}s$. Furthermore, we also use the notation that $\gamma^i = \mu_t^i(\lambda_t^i)^{-1}$ and $H_t^i = \gamma_t^i - 1$. Recall that the assumption that μ is λ-compatible by convention implies that both γ and H are locally bounded. Integrals are vector integrals in the sense that $H \cdot M$ denotes the one-dimensional process defined by $H \cdot M = \sum_{i=1}^d H^i \cdot M^i$.

We first prove Lemma 2.2.2, the result stated in Section 2.2 as the reason for taking interest in the martingale property of $\mathcal{E}(H \cdot M)$ when considering changing the intensity of a counting process. Recall that Π_p^* denotes the dual predictable projection, see Definition 5.21 of [66].

Lemma 2.4.3. *Let M be a local martingale with $\Delta M \geq -1$ and let T be a stopping time. Assume that $\mathcal{E}(M)^T$ is a uniformly integrable martingale. Let Q be the probability measure having Radon-Nikodym derivative $\mathcal{E}(M)_T$ with respect to P. If L is a local martingale under P such that $[L, M^T]$ is locally integrable under P, then $L - \langle L, M^T \rangle$ is a local martingale under Q, where the angle bracket is calculated under P.*

Proof. First note that as Q has a density with respect to P, Q is absolutely continuous with respect to P. With Z being the likelihood process for Q with respect to P, meaning that $Z_t = E(\frac{\mathrm{d}Q}{\mathrm{d}P}|\mathcal{F}_t)$, we have $Z_t = E(\mathcal{E}(M)_\infty^T|\mathcal{F}_t) = \mathcal{E}(M)_t^T$ up to indistinguishability. In particular, $Z_0 = 1$ almost surely. By an examination of the proof of the predictable Girsanov theorem, Theorem III.41 of [134], we therefore find that the theorem can be applied in spite of our not having assumed that $\mathcal{F}_0$ is a sub-σ-algebra of the P-completion of $\{\emptyset, \Omega\}$, as the theorem in [134] otherwise requires.

Now consider a process L which is a local martingale under P such that $[L, M^T]$ is locally integrable under P. Note that $[L, \mathcal{E}(M^T)] = [L, \mathcal{E}(M^T)_- \cdot M^T] = \mathcal{E}(M)_-^T \cdot [L, M^T]$. As $\mathcal{E}(M)_-$ is left-continuous, it is locally bounded. Therefore, as $[L, M^T]$ is locally integrable, the process $\mathcal{E}(M)_-^T \cdot [L, M^T]$ is locally integrable as well. Thus, $[L, \mathcal{E}(M^T)]$ is locally integrable under P, so the predictable covariation of this process is well-defined under P. Then, Theorem III.41 of [134] applies and yields that the

process given by $L_u - \int_0^u \mathcal{E}(M^T)_{s-}^{-1} \, \mathrm{d}\langle \mathcal{E}(M^T), L\rangle_s$ is a Q local martingale, where the angle bracket is calculated under P. Noting that

$$\begin{aligned} L_u - \int_0^u \frac{1}{\mathcal{E}(M^T)_{s-}} \, \mathrm{d}\langle \mathcal{E}(M^T), L\rangle_s &= L_u - \int_0^u \frac{1}{\mathcal{E}(M^T)_{s-}} \, \mathrm{d}\langle \mathcal{E}(M^T)_- \cdot M^t, L\rangle_s \\ &= L_u - \langle L, M^T\rangle_u, \end{aligned} \tag{2.50}$$

the result follows. □

Proof of Lemma 2.2.2. Fix a stopping time T. By definition, Q has Radon-Nikodym derivative $\mathcal{E}(H \cdot M)_T$ with respect to P. We wish to apply Lemma 2.4.3 in order to prove the result. We first check that $[M^i, (H \cdot M)^T]$ is locally integrable under P. Note that $[M^i, M^j]_t = \sum_{0<s\leq t} \Delta M_s^i \Delta M_s^j = \sum_{0<s\leq t} \Delta N_s^i \Delta N_s^j = [N^i, N^j]$, since M^i has finite variation, in particular $[M^i] = N^i$. As the coordinates of N have no common jumps, we have $[M^i, (H \cdot M)^T] = H^i 1_{[0,T]} \cdot [N^i]$. Because we have assumed that H is locally bounded, this is locally integrable. From Lemma 2.4.3, we then conclude that $M^i - \langle M^i, (H \cdot M)^T\rangle$ is a local martingale under Q. Next, under P, $(\Pi_p^* N^i)_t = \int_0^t \lambda_s^i \, \mathrm{d}s$, and H and $1_{[0,T]}$ are predictable. Therefore, we obtain the equalities $\langle M^i, (H \cdot M)^T\rangle_s = \Pi_p^*(H^i 1_{[0,T]} \cdot [N^i])_s = \int_0^s H_u^i 1_{(u\leq T)} \lambda_u^i \, \mathrm{d}u$, which allows us to conclude that

$$\begin{aligned} M_s^i - \langle M^i, (H \cdot M)^T\rangle_s &= N_s^i - \int_0^s \lambda_s^i \, \mathrm{d}s - \int_0^s H_u^i 1_{(u\leq T)} \lambda_u^i \, \mathrm{d}u \\ &= N_t^i - \int_0^s \mu_u^i 1_{[0,T]}(u) + \lambda_u^i 1_{(T,\infty)}(u) \, \mathrm{d}u. \end{aligned} \tag{2.51}$$

This proves that under Q, N has intensity $1_{[0,T]}\mu + 1_{(T,\infty)}\lambda$. The results for the case where $\mathcal{E}(H \cdot M)$ is a martingale then follows by considering stopping times which are constant. □

Next, we prove Lemma 2.2.3, which yields a sufficient criterion for the probability measure Q constructed using an exponential martingale to be equivalent to our starting probability measure P.

Lemma 2.4.4. *Let N have intensity λ. If X is a process which is nonnegative, predictable and locally bounded, and it holds almost surely that pathwisely, the set of zeroes of X has Lebesgue measure zero, then it almost surely holds that the zeroes of X are disjoint from the jump times of N^i for all i.*

Proof. As X is predictable, the set of zeroes of X is a predictable set. Thus, the integral process $\int_0^t 1_{(X_s=0)} \, \mathrm{d}M_s^i$ is a local martingale. Let (T_n) be a localising sequence such that $\int_0^t 1_{(X_s=0)} 1_{(t\leq T_n)} \, \mathrm{d}N_s^i$ is bounded and such that $\int_0^t 1_{(X_s=0)} 1_{(t\leq T_n)} \, \mathrm{d}M_s^i$ is a true martingale. Then $E\int_0^t 1_{(X_s=0)} 1_{(t\leq T_n)} \, \mathrm{d}M_s^i = 0$, and so by our assumptions, $E\int_0^t 1_{(X_s=0)} 1_{(t\leq T_n)} \, \mathrm{d}N_s^i = 0$ as well, leading us to conclude that $\int_0^\infty 1_{(X_s=0)} \, \mathrm{d}N_s^i$ is almost surely zero. This implies that almost surely, the set of zeroes of X is disjoint from the jump times of N^i. As the coordinate i was arbitrary, the result follows. □

Proof of Lemma 2.2.3. Note that $\Delta(H \cdot M)_t = \sum_{i=1}^{d} H_t^i \Delta N_t^i$. By Lemma 2.4.4, the set of zeroes of μ^i is disjoint from the jump times of N^i. Therefore, the set of zeroes of γ^i is disjoint from the jump times of N^i as well, and so the set where H^i is -1 is disjoint from the jump times of N^i. We conclude that almost surely, $H \cdot M$ has no jumps of size -1. Theorem I.4.61 of [83] then shows that $\mathcal{E}(H \cdot M)$ is almost surely positive. □

Finally, we prove Theorem 2.2.4 and its corollary. We first state the two main theorems of [109] which we will apply to integrals of compensated counting processes in order to obtain our results. The two main theorems from that article are Theorem III.1 and Theorem III.7, given below.

Theorem 2.4.5. *Let M be a local martingale with initial value zero and jumps satisfying $\Delta M \geq -1$. Let $R = \inf\{t \geq 0 \mid \Delta M_t = -1\}$. Define B by putting $B_t = \frac{1}{2}[M^c]_{t\wedge R} + \sum_{0<s\leq t\wedge R}(1+\Delta M_s)\log(1+\Delta M_s) - \Delta M_s$. If B is locally integrable and $\exp(\Pi_p^* B_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale.*

Theorem 2.4.6. *Let M be a local martingale with initial value zero and $\Delta M > -1$. Define A by putting $A_t = \frac{1}{2}[M^c]_t + \sum_{0<s\leq t} \log(1+\Delta M_s) - \frac{\Delta M_s}{1+\Delta M_s}$. If $\exp(A_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale.*

The following two lemmas are ingredients for the proof of Theorem 2.2.4. The first lemma allows us to restrict our attention to small deterministic time intervals when proving the martingale property of exponential martingales. This technique is well-known, see for example Corollary 3.5.14 of [91]. The second lemma decomposes an exponential martingale into the product of two exponential martingales, corresponding to successive changes of intensity from λ to μ and μ to $\mu + \nu$. This will, colloquially speaking, allow us to consider the large and small parts of μ separately when proving the martingale property.

Lemma 2.4.7. *Let M be a local martingale with $\Delta M \geq -1$, and let $\varepsilon > 0$. If $\mathcal{E}(M^t - M^u)$ is a martingale whenever $0 \leq u \leq t$ with $t - u \leq \varepsilon$, then $\mathcal{E}(M)$ is a martingale.*

Proof. Let $\varepsilon > 0$ be given such that $\mathcal{E}(M^t - M^u)$ is a martingale for $0 \leq u \leq t$ with $t - u \leq \varepsilon$. By Lemma 2.4.2, to show that $\mathcal{E}(M)$ is a martingale, it suffices to show $E\mathcal{E}(M)_t = 1$ for all $t \geq 0$. Fix $t \geq 0$ and note that $[M^{t+\varepsilon} - M^t, M^t] = 0$, so Theorem II.38 of [134] yields

$$\mathcal{E}(M^{t+\varepsilon}) = \mathcal{E}(M^t)\mathcal{E}(M^{t+\varepsilon} - M^t). \tag{2.52}$$

Therefore, we obtain

$$\begin{aligned} E\mathcal{E}(M)_{t+\varepsilon} &= E\mathcal{E}(M^{t+\varepsilon})_{t+\varepsilon} = E\mathcal{E}(M)_t E(\mathcal{E}(M^{t+\varepsilon} - M^t)_{t+\varepsilon}|\mathcal{F}_t) \\ &= E\mathcal{E}(M)_t \mathcal{E}(M^{t+\varepsilon} - M^t)_t = E\mathcal{E}(M)_t. \end{aligned} \tag{2.53}$$

As $E\mathcal{E}(M)_0 = 1$, this implies $E\mathcal{E}(M)_t = 1$ for all $t \geq 0$, and so we conclude that $\mathcal{E}(M)$ is a martingale. □

Lemma 2.4.8. *Let ν be nonnegative, predictable and locally bounded. Assume that μ is λ-compatible and that $\mu+\nu$ is μ-compatible. Then $\mu+\nu$ is also λ-compatible. Define three processes $(H_\lambda^{\mu+\nu})_t^i = (\mu_t^i+\nu_t^i)(\lambda_t^i)^{-1}-1$, $(H_\lambda^{\mu})_t^i = \mu_t^i(\lambda_t^i)^{-1}-1$ and $(H_\mu^{\mu+\nu})_t^i = (\mu_t^i+\nu_t^i)(\mu_t^i)^{-1}-1$. Define d-dimensional processes M^λ and M^μ by puting $(M^\lambda)_t^i = N_t^i - \int_0^t \lambda_s^i \,\mathrm{d}s$ and putting $(M^\mu)^i = N_t^i - \int_0^t \mu_s^i \,\mathrm{d}s$. Then, it holds that $\mathcal{E}(H_\lambda^{\mu+\nu}\cdot M^\lambda) = \mathcal{E}(H_\lambda^{\mu}\cdot M^\lambda)\mathcal{E}(H_\mu^{\mu+\nu}\cdot M^\mu)$.*

Proof. That $\mu+\nu$ is λ-compatible follows as $\mu+\nu$ is μ-compatible and μ is λ-compatible. Furthermore, M^λ and M^μ are processes of finite variation, so we find

$$\begin{aligned}[H_\lambda^\mu\cdot M^\lambda, H_\mu^{\mu+\nu}\cdot M^\mu]_t &= \sum_{0<s\le t}\Delta(H_\lambda^\mu\cdot M^\lambda)_s\Delta(H_\mu^{\mu+\nu}\cdot M^\mu)_s\\ &= \sum_{i=1}^d\sum_{0<s\le t}(H_\lambda^\mu)_s^i(H_\mu^{\mu+\nu})_s^i\Delta N_s^i\\ &= \sum_{i=1}^d\int_0^t (H_\lambda^\mu)_s^i(H_\mu^{\mu+\nu})_s^i\,\mathrm{d}N_s^i. \end{aligned}\tag{2.54}$$

Therefore, $\mathcal{E}(H_\lambda^\mu\cdot M^\lambda)\mathcal{E}(H_\mu^{\mu+\nu}\cdot M^\mu) = \mathcal{E}(H_\lambda^\mu\cdot M^\lambda + H_\mu^{\mu+\nu}\cdot M^\mu + H_\lambda^\mu H_\mu^{\mu+\nu}\cdot N)$ by Theorem II.38 of [134]. We find

$$\begin{aligned}&(H_\lambda^\mu\cdot M^\lambda + H_\mu^{\mu+\nu}\cdot M^\mu + H_\lambda^\mu H_\mu^{\mu+\nu}\cdot N)_t \\ &= \sum_{i=1}^d\int_0^t (H_\lambda^\mu)_s^i + (H_\mu^{\mu+\nu})_s^i + (H_\lambda^\mu)_s^i(H_\mu^{\mu+\nu})_s^i\,\mathrm{d}N_s^i - \int_0^t (H_\lambda^\mu)_s^i\lambda_s^i + (H_\mu^{\mu+\nu})_s^i\mu_s^i\,\mathrm{d}s.\end{aligned}\tag{2.55}$$

Now noting that

$$\begin{aligned}&(H_\lambda^\mu)_t^i + (H_\mu^{\mu+\nu})_t^i + (H_\lambda^\mu)_t^i(H_\mu^{\mu+\nu})_t^i\\ &= \left(\frac{\mu_t^i}{\lambda_t^i}-1\right) + \left(\frac{\mu_t^i+\nu_t^i}{\mu_t^i}-1\right) + \left(\frac{\mu_t^i}{\lambda_t^i}-1\right)\left(\frac{\mu_t^i+\nu_t^i}{\mu_t^i}-1\right)\\ &= \frac{\mu_t^i}{\lambda_t^i}\frac{\mu_t^i+\nu_t^i}{\mu_t^i} - 1 = \frac{\mu_t^i+\nu_t^i}{\lambda_t^i} - 1 = (H_\lambda^{\mu+\nu})_t^i\end{aligned}\tag{2.56}$$

as well as $(H_\lambda^\mu)_t^i\lambda_t^i + (H_\mu^{\mu+\nu})_t^i\mu_t^i = \mu_t^i - \lambda_t^i + \mu_t^i + \nu_t^i - \mu_t^i = \mu_t^i+\nu_t^i-\lambda_t^i$, we may conclude

$$\begin{aligned}&H_\lambda^\mu\cdot M^\lambda + H_\mu^{\mu+\nu}\cdot M^\mu + H_\lambda^\mu H_\mu^{\mu+\nu}\cdot N\\ &= \sum_{i=1}^d\int_0^t (H_\lambda^{\mu+\nu})_s^i\,\mathrm{d}N_s^i - \sum_{i=1}^d\int_0^t (H_\lambda^{\mu+\nu})_s^i\lambda_s^i\,\mathrm{d}s = H_\lambda^{\mu+\nu}\cdot M^\lambda,\end{aligned}\tag{2.57}$$

yielding the desired result. □

Proof of Theorem 2.2.4. By Lemma 2.4.7, it suffices to show the martingale property of $\mathcal{E}((H \cdot M)^t - (H \cdot M)^u)$ when $0 \leq u \leq t$ with $t - u \leq \varepsilon$. Let such a pair of u and t be given and let $L = (H \cdot M)^t - (H \cdot M)^u$. With R and B as in Theorem 2.4.5, we have for $r \geq 0$ that

$$B_r = \sum_{i=1}^{d} \int_0^r 1_{[0,R]}(s) 1_{(u,t]}(s)((1 + H_s^i)\log(1 + H_s^i) - H_s^i)\,\mathrm{d}N_s^i. \tag{2.58}$$

From this, we obtain that B is locally integrable, and as $1_{[0,R]}$ is a predictable process, we have

$$\begin{aligned}(\Pi_p^* B)_\infty &= \sum_{i=1}^{d} \int_u^t 1_{[0,R]}(s)(\gamma_s^i \log \gamma_s^i - (\gamma_s^i - 1))\lambda_s^i \,\mathrm{d}s \\ &\leq \sum_{i=1}^{d} \int_u^t (\gamma_s^i \log \gamma_s^i - (\gamma_s^i - 1))\lambda_s^i \,\mathrm{d}s. \end{aligned} \tag{2.59}$$

Therefore, if the first integrability criterion is satisfied, $\mathcal{E}(L)$ is a uniformly integrable martingale by Theorem 2.4.5, in particular a martingale. This proves the first claim.

Next, we consider the case where the second integrability criterion is satisfied. We will use Lemma 2.4.8 to prove that $\mathcal{E}((H \cdot M)^t - (H \cdot M)^u)$ is a martingale in this case. To this end, we define predictable d-dimensional processes μ^- and μ^+ by

$$(\mu^-)_s^i = \mu_s^i 1_{(\mu_s^i \leq \lambda_s^i)} + \lambda_s^i 1_{(\mu_s^i > \lambda_s^i)} \tag{2.60}$$

$$(\mu^+)_s^i = (\mu_s^i - \lambda_s^i) 1_{(\mu_s^i > \lambda_s^i)}. \tag{2.61}$$

We then have $\mu = \mu^+ + \mu^-$. Also define two processes $(\gamma^*)^i = (\mu^-)^i(\lambda^i)^{-1}$ and $(\gamma^{**})^i = \mu^i((\mu^-)^i)^{-1}$, and $H^* = \gamma^* - 1$ and $H^{**} = \gamma^{**} - 1$. Now, as λ and μ are predictable, μ^- and μ^+ are predictable as well. Furthermore, μ^- and μ^+ are both nonnegative and locally bounded. By inspection, μ^- is λ-compatible and μ is μ^--compatible. Now define $(M^-)_t^i = N_t^i - \int_0^t (1_{(u,t]}(s)(H^*)_s^i + 1)\lambda_s^i \,\mathrm{d}s$. Define $L^* = (H^* \cdot M)^t - (H^* \cdot M)^u$ and $L^{**} = (H^{**} \cdot M^-)^t - (H^{**} \cdot M^-)^u$. Note that $L^* = H^* 1_{(u,t]} \cdot M$, $L^{**} = H^{**} 1_{(u,t]} \cdot M^-$ and $L = H 1_{(u,t]} \cdot M$. Invoking Lemma 2.4.8, we obtain $\mathcal{E}(L) = \mathcal{E}(L^*)\mathcal{E}(L^{**})$. We will apply Theorem 2.4.5 to the local martingale L^*. By the same calculations as earlier, this can be done if we can prove that

$$E \exp\left(\sum_{i=1}^{d} \int_u^t ((\gamma^*)_s^i \log(\gamma^*)_s^i - ((\gamma^*)_s^i - 1))\lambda_s^i \,\mathrm{d}s\right) < \infty. \tag{2.62}$$

However, we always have $0 \leq (\gamma^*)^i \leq 1$. As $x \log x \leq 0$ when $0 \leq x \leq 1$, we obtain $(\gamma^*)_s^i \log(\gamma^*)_s^i - (\gamma^*)_s^i \leq 0$, and so it suffices to note that

$$E \exp\left(\sum_{i=1}^{d} \int_u^t \lambda_s^i \,\mathrm{d}s\right) \leq E \exp\left(\sum_{i=1}^{d} \int_u^t \lambda_s^i \,\mathrm{d}s + \int_u^t \log_+ \gamma_s^i \,\mathrm{d}N_s^i\right) < \infty, \tag{2.63}$$

so Theorem 2.4.5 shows that $\mathcal{E}(L^*)$ is a uniformly integrable martingale. Let Q be the measure with Radon-Nikodym derivative $\mathcal{E}(L^*)_\infty$ with respect to P. We then have $E^P\mathcal{E}(L)_\infty = E^Q\mathcal{E}(L^{**})_\infty$. To show that $\mathcal{E}(L)$ is a uniformly integrable martingale, it suffices to show that this is equal to one. To do so, we will apply Theorem 2.4.6 to show that $\mathcal{E}(L^{**})$ is a uniformly integrable martingale under Q. To this end, first note that by Lemma 2.2.2, N^i has intensity $(1_{(u,t]}(H^*)^i + 1)\lambda^i$ under Q. Therefore, M^- is a local martingale under Q, and so L^{**} is a local martingale under Q as well. Next, $(H^{**})^i_t = (\gamma^{**})^i_t - 1 = 1_{(\mu^i_t \le \lambda^i_t)} + \gamma^i_t 1_{(\mu^i_t > \lambda^i_t)} - 1 \ge 0$, so $\Delta L^{**} \ge 0 > -1$, and therefore Theorem 2.4.6 is applicable. Now, with A as in Theorem 2.4.6, we have

$$\begin{aligned} A_\infty &= \frac{1}{2}[(L^{**})^c]_\infty + \sum_{0<s} \log(1+\Delta L^{**}_s) - \frac{\Delta L^{**}_s}{1+\Delta L^{**}_s} \\ &\le \sum_{i=1}^d \int_u^t \log \frac{\mu^i_s}{(\mu^-)^i_s}\, \mathrm{d}N^i_s = \sum_{i=1}^d \int_u^t \log_+ \gamma^i_s \, \mathrm{d}N^i_s. \end{aligned} \tag{2.64}$$

Also, since $-1 \le H^* \le 0$, we find that $-1 \le \Delta L^* \le 0$ and thus, whenever L^* has no jumps of size -1, we obtain

$$\begin{aligned} \mathcal{E}(L^*)_\infty &= \exp\left(L^*_\infty + \sum_{0<s} \log(1+\Delta L^*_s) - \Delta L^*_s \right) \\ &\le \exp\left(L^*_\infty - \sum_{0<s} \Delta L^*_s \right) = \exp\left(-\sum_{i=1}^d \int_u^t (H^*)^i_s \lambda^i_s \, \mathrm{d}s \right) \\ &\le \exp\left(\sum_{i=1}^d \int_u^t \lambda^i_s \, \mathrm{d}s \right), \end{aligned} \tag{2.65}$$

which leads to

$$E^Q \exp(A_\infty) = E\mathcal{E}(L^*)_\infty \exp(A_\infty) \le E \exp\left(\sum_{i=1}^d \int_u^t \lambda^i_s \, \mathrm{d}s + \int_u^t \log_+ \gamma^i_s \, \mathrm{d}N^i_s \right), \tag{2.66}$$

and this is finite by assumption. Theorem 2.4.6 then shows that L^{**} is a uniformly integrable martingale under Q, so $E^Q\mathcal{E}(L^{**})_\infty = 1$. Therefore, $E^P\mathcal{E}(L)_\infty = 1$. Thus, $\mathcal{E}(L)$ is a uniformly integrable martingale, in particular a martingale. This completes the proof. □

Proof of Corollary 2.2.5. First note that

$$x \log x - (x-1) \le 1 + x \log x \le 1 + x \log_+ x \tag{2.67}$$

for $x \ge 0$. Therefore, as $\lambda = 1$, the first moment condition of Theorem 2.2.4 reduces to the first moment condition in the statement of the corollary. Furthermore, $E\exp(\sum_{i=1}^d \int_u^t \lambda^i_s \, \mathrm{d}s + \int_u^t \log_+ \gamma^i_s \, \mathrm{d}N^i_s) = e^{d(t-u)} E\exp(\sum_{i=1}^d \int_u^t \log_+ \gamma^i_s \, \mathrm{d}N^i_s)$ as $\lambda = 1$ and the result for the second moment condition of the corollary follows. This completes the proof. □

2.5 Supplementary results

Lemma 2.5.1. *Let Z be Poisson distributed with parameter μ. Then $\exp(\varepsilon Z \log Z)$ is integrable whenever $0 \leq \varepsilon < 1$.*

Proof. This follows by an application of Stirling's formula, see (6.11.2) of [182], and comparison with a geometric series. □

Lemma 2.5.2. *Consider mappings $A : \mathbb{N}_0^d \times \mathbb{R}_+^d \to \mathbb{R}^d$, $B : \mathbb{N}_0^d \times \mathbb{R}_+^d \to \mathbb{M}(d,d)$ and $\sigma : \mathbb{N}_0^d \times \mathbb{R}_+^d \to \mathbb{M}(d,d)$ such that $A(\eta,\cdot)$, $B(\eta,\cdot)$ and $\sigma(\eta,\cdot)$ are bounded and continuous for $\eta \in \mathbb{N}_0^d$. Let W be a d-dimensional $(\mathcal{F}_t)$ Brownian motion. Let T_n^i be the n'th event time for N^i and let $Z_t^i = t - T^i_{N_t^i}$. The stochatic differential equation*

$$\mathrm{d}X_t = (A(N_t, Z_t) + B(N_t, Z_t)X_t)\,\mathrm{d}t + \sigma(N_t, Z_t)\,\mathrm{d}W_t \tag{2.68}$$

is exact, in the sense that for any initial value, it has a pathwise unique solution. Defining $C_t = \exp(-\int_0^t B(N_s, Z_s)\,\mathrm{d}s)$, the solution is

$$X_t = C_t^{-1}\left(X_0 + \int_0^t C_s A(N_s, Z_s)\,\mathrm{d}s + \int_0^t C_s \sigma(N_s, Z_s)\,\mathrm{d}W_s\right). \tag{2.69}$$

Proof. Let $\tilde{A}_s = A(N_s, Z_s)$, and define $\tilde{B}$ and $\tilde{\sigma}$ analogously. Note that as N and Z are adapted, $\tilde{A}$ is adapted as well, since $A(\eta,\cdot)$ is continuous and therefore Borel measurable for all $\eta \in \mathbb{N}_0^d$. As the process also is right-continuous and locally bounded, all integrals are well-defined, and similarly for $\tilde{B}$ and $\tilde{\sigma}$. Let X_0 be some initial value. Assume that X is a solution to the stochastic differential equation. Note that each entry of C_t is differentiable as a function of t, and $\frac{\mathrm{d}}{\mathrm{d}t}C_t^{ij} = (-\tilde{B}_t C_t)^{ij}$. The integration-by-parts formula yields

$$(C_t X_t)_i = X_0^i + \sum_{j=1}^d \int_0^t C_s^{ij}\,\mathrm{d}X_s^j - \int_0^t X_s^j (\tilde{B}_s C_s)^{ij}\,\mathrm{d}s. \tag{2.70}$$

This implies $(C_t X_t)_i = X_0^i + \sum_{j=1}^d \int_0^t C_s^{ij}\tilde{A}_s^j\,\mathrm{d}s + \int_0^t C_s^{ij}\sum_{k=1}^d \tilde{\sigma}_s^{jk}\,\mathrm{d}W_s^k$, since X is a solution, leading to

$$X_t = C_t^{-1}\left(X_0 + \int_0^t C_s A(N_s, Z_s)\,\mathrm{d}s + \int_0^t C_s \sigma(N_s, Z_s)\,\mathrm{d}W_s\right). \tag{2.71}$$

This proves pathwise uniqueness. Applying the integration-by-parts formula to the above shows that the proposed solution in fact is a solution, yielding existence. □

Lemma 2.5.3. *Let X be a d-dimensional normally distributed variable with mean ξ and positive definite variance Σ. Let $c > 0$ and $0 < \varepsilon < 1$. Then $\exp(c\|X\|_2^{1+\varepsilon})$ is*

integrable. Furthermore, defining $a(c,\varepsilon) = 2^{1+\varepsilon}c$ *and* $b(c,\varepsilon) = 16^{(1+\varepsilon)/(1-\varepsilon)}c^{2/(1-\varepsilon)}$, *it holds that*

$$E\exp(c\|X\|_2^{1+\varepsilon}) \leq k_d \exp(a(c,\varepsilon)\|\xi\|^{1+\varepsilon}) \exp\left(b(c,\varepsilon)\|\Sigma\|_2^{\frac{1+\varepsilon}{1-\varepsilon}}\right), \tag{2.72}$$

where $k_d = A_d m_{d-1}(\sqrt{2}\sqrt{\pi^{d-1}})^{-1}$, A_d *is the area of the unit sphere in* d *dimensions and* m_d *is the* d*'th absolute moment of the standard normal distribution.*

Proof. By [104], p. 181, Σ has a unique symmetric positive definite square root $\Sigma^{1/2}$ such that $\Sigma = (\Sigma^{1/2})^2$. Furthermore, with $Y = \Sigma^{-1/2}(X-\xi)$, it holds that $X = \xi + \Sigma^{\frac{1}{2}}Y$, where Y is d-dimensionally standard normally distributed. With $\|\cdot\|_2$ denoting the operator norm induced by the Euclidean norm, we get

$$\begin{aligned} E\exp(c\|X\|_2^{1+\varepsilon}) &= E\exp(c\|\xi + \Sigma^{\frac{1}{2}}Y\|_2^{1+\varepsilon}) \leq E\exp(c(\|\xi\|_2 + \|\Sigma^{\frac{1}{2}}Y\|_2)^{1+\varepsilon}) \\ &\leq E\exp(c2^{1+\varepsilon}(\|\xi\|_2^{1+\varepsilon} + \|\Sigma^{\frac{1}{2}}Y\|_2^{1+\varepsilon})) \\ &\leq \exp(c2^{1+\varepsilon}\|\xi\|^{1+\varepsilon}) E\exp\left(c2^{1+\varepsilon}\|\Sigma\|_2^{(1+\varepsilon)/2}\|Y\|_2^{1+\varepsilon}\right). \end{aligned} \tag{2.73}$$

Switching to polar coordinates (see [150], page 149) we obtain, with A_d denoting the area of the unit sphere in d dimensions and $C = c2^{1+\varepsilon}\|\Sigma\|_2^{(1+\varepsilon)/2}$,

$$\begin{aligned} &E\exp\left(c2^{1+\varepsilon}\|\Sigma\|_2^{(1+\varepsilon)/2}\|Y\|_2^{1+\varepsilon}\right) \\ &= \int_{\mathbb{R}^d} \exp\left(C\|x\|_2^{1+\varepsilon}\right) \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}\|x\|_2^2\right) \mathrm{d}x \\ &= \frac{A_d}{\sqrt{(2\pi)^{d-1}}} \int_0^\infty \exp\left(Cr^{1+\varepsilon}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}r^2\right) r^{d-1}\,\mathrm{d}r. \end{aligned} \tag{2.74}$$

Using a change of variables, we obtain the bound

$$\begin{aligned} &\int_0^\infty \exp\left(Cr^{1+\varepsilon}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}r^2\right) r^{d-1}\,\mathrm{d}r \\ &\leq \int_0^\infty r^{d-1}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{4}r^2\right) \mathrm{d}r \sup_{s\geq 0} \exp\left(Cs^{1+\varepsilon} - \frac{1}{4}s^2\right) \\ &= 2^{d/2}\int_0^\infty r^{d-1}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}r^2\right) \mathrm{d}r \sup_{s\geq 0} \exp\left(Cs^{1+\varepsilon} - \frac{1}{4}s^2\right). \end{aligned} \tag{2.75}$$

With m_d denoting the d'th absolute moment of the standard normal distribution, we have $\int_0^\infty r^{d-1}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}r^2\right)\mathrm{d}r = \frac{1}{2}m_{d-1}$. Also, defining $\phi(r) = Cr^{1+\varepsilon} - \frac{1}{4}r^2$ for $r \geq 0$, ϕ has a global maximum at $r^* = (2C(1+\varepsilon))^{1/(1-\varepsilon)}$ with $\phi(r^*) \leq 4^{\frac{1+\varepsilon}{1-\varepsilon}}C^{\frac{2}{1-\varepsilon}}$. As the exponential mapping is increasing, this allows us to conclude

$$\int_0^\infty \exp\left(Cr^{1+\varepsilon}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}r^2\right) r^{d-1}\,\mathrm{d}r \leq 2^{d/2-1}m_{d-1}\exp\left(4^{\frac{1+\varepsilon}{1-\varepsilon}}C^{\frac{2}{1-\varepsilon}}\right). \tag{2.76}$$

Recalling our definition of C, we have $4^{\frac{1+\varepsilon}{1-\varepsilon}} C^{\frac{2}{1-\varepsilon}} = 16^{\frac{1+\varepsilon}{1-\varepsilon}} c^{\frac{2}{1-\varepsilon}} \|\Sigma\|_2^{\frac{1+\varepsilon}{1-\varepsilon}}$. Therefore, defining $a(c,\varepsilon) = 2^{1+\varepsilon}c$ and $b(c,\varepsilon) = 16^{(1+\varepsilon)/(1-\varepsilon)} c^{2/(1-\varepsilon)}$, we finally conclude

$$E\exp(c\|X\|_2^{1+\varepsilon}) \leq \frac{A_d m_{d-1}}{\sqrt{2}\sqrt{\pi^{d-1}}} \exp(a(c,\varepsilon)\|\xi\|^{1+\varepsilon}) \exp\left(b(c,\varepsilon)\|\Sigma\|_2^{\frac{1+\varepsilon}{1-\varepsilon}}\right), \tag{2.77}$$

which proves the lemma. □

Lemma 2.5.4. *Let N be a point process, let $h : \mathbb{R}_+ \to \mathbb{R}$ be Borel measurable and define $\mu_t = \int_0^{t-} h(t-s)\,\mathrm{d}N_s$. Then μ is a predictable process.*

Proof. As h is Borel measurable, there exists a sequence of simple Borel measurable functions $h_n : \mathbb{R}_+ \to \mathbb{R}$ converging pointwise to h. As N pathwisely only jumps finitely many times on compact intervals, we have $\mu_t = \lim_{n\to\infty} \mu_t^n$, where the limit is pointwise and $\mu_t^n = \int_0^{t-} h_n(t-s)\,\mathrm{d}N_s$. Thus, it suffices to show that μ^n is predictable. Assume for definiteness that $h_n = \sum_{k=1}^{m_n} c_{nk} 1_{A_{nk}}$, where $c_{nk} \in \mathbb{R}$ and A_{nk} is a Borel set in $\mathbb{R}_+$. With T_n denoting the n'th event time for N, we have

$$\mu_t^n = \sum_{n=1}^{\infty} h_n(t-T_n) 1_{(T_n<t)} = \sum_{n=1}^{\infty} \sum_{k=1}^{m_n} c_{nk} 1_{(t-T_n\in A_{nk})} 1_{(T_n<t)} \tag{2.78}$$

From this, we conclude that in order to show the result, it suffices to show that for any stopping time T and any Borel set in $\mathbb{R}_+$, the process $X_t^A = 1_{(t-T_n\in A)}$ is predictable. Let T be a stopping time and let $\mathbb{D}$ be the class of Borel sets in $\mathbb{R}_+$ such that this holds. Then $\mathbb{D}$ is a Dynkin class. Furthermore, for $a \geq 0$, we have $X_t^A = 1_{(t-T_n\in(a,\infty))} = 1_{(T_n+a<t)}$. This shows that X^A is left-continuous and adapted, and so predictable. By Dynkin's lemma, X^A is predictable for all Borel sets A in $\mathbb{R}_+$. This proves the lemma. □

Lemma 2.5.5. *Let (T_n) be a localising sequence and assume that $\mathcal{E}(M)^{T_n}$ is a martingale. $\mathcal{E}(M)$ is a martingale if and only if $\lim_n E\mathcal{E}(M)_{T_n} 1_{(T_n\leq t)} = 0$ for each $t \geq 0$.*

Proof. By our assumptions on the martingale property of $\mathcal{E}(M)^{T_n}$, it holds that $E\mathcal{E}(M)_{T_n} 1_{(T_n\leq t)} = 1 - E\mathcal{E}(M)_t 1_{(T_n>t)}$. By the dominated convergence theorem, $\lim_n E\mathcal{E}(M)_t 1_{(T_n>t)} = E\mathcal{E}(M)_t$. Thus, $\lim_n E\mathcal{E}(M)_{T_n} 1_{(T_n\leq t)} = 1 - E\mathcal{E}(M)_t$, and so Lemma 2.4.2 yields the result. □

3

Optimal Novikov-type criteria for local martingales with jumps

Alexander Sokol

2010 Mathematics Subject Classification. Primary 60G44; Secondary 60G40.

Key words and phrases. Martingale, Exponential martingale, Uniform integrability, Novikov, Optimal, Poisson process.

Abstract. We consider local martingales M with initial value zero and jumps larger than a for some a larger than or equal to -1, and prove Novikov-type criteria for the exponential local martingale to be a uniformly integrable martingale. We obtain criteria using both the quadratic variation and the predictable quadratic variation. We prove optimality of the coefficients in the criteria. As a corollary, we obtain a verbatim extension of the classical Novikov criterion for continuous local martingales to the case of local martingales with initial value zero and nonnegative jumps.

3.1 Introduction

The motivation of this paper is the question of when an exponential local martingale is a uniformly integrable martingale. Before introducing this problem, we fix our notation and recall some results from stochastic analysis.

Assume given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ satisfying the usual conditions, see [134] for the definition of this and other probabilistic concepts such as being a local martingale, locally integrable, locally square-integrable, and for the quadratic variation and quadratic covariation et cetera. For any local martingale M, we say that M has initial value zero if $M_0 = 0$. For any local martingale M with initial value zero, we denote by $[M]$ the quadratic variation of M, that is, the unique increasing adapted process with initial value zero such that $\Delta[M] = (\Delta M)^2$ and $M^2 - [M]$ is a local martingale. If M furthermore is locally square integrable, we denote by $\langle M \rangle$ the predictable quadratic variation of M, which is the unique increasing predictable process with initial value zero such that $[M] - \langle M \rangle$ is a local martingale.

For any local martingale with initial value zero, there exists by Theorem 7.25 of [66] a unique decomposition $M = M^c + M^d$, where M^c is a continuous local martingale and M^d is a purely discontinuous local martingale, both with initial value zero. Here, we say that a local martingale with initial value zero is purely discontinuous if it has zero quadratic covariation with any continuous local martingale with initial value zero. We refer to M^c as the continuous martingale part of M, and refer to M^d as the purely discontinuous martingale part of M.

Let M be a local martingale with initial value zero and $\Delta M \geq -1$. The exponential martingale of M, also known as the Doléans-Dade exponential of M, is the unique càdlàg solution in Z to the stochastic differential equation $Z_t = 1 + \int_0^t Z_{s-}\,\mathrm{d}M_s$, given explicitly as

$$\mathcal{E}(M)_t = \exp\left(M_t - \frac{1}{2}[M^c]_t\right) \prod_{0<s\leq t} (1 + \Delta M_s)\exp(-\Delta M_s), \tag{3.1}$$

see Theorem II.37 of [134]. Applying Theorem 9.2 of [66], we find that Z always is a local martingale with initial value one. Also, $\mathcal{E}(M)$ is always nonnegative. We wish to understand when $\mathcal{E}(M)$ is a uniformly integrable martingale.

The question of when $\mathcal{E}(M)$ is a uniformly integrable martingale has been considered many times in the literature, and is not only of theoretical interest, but has several applications in connection with other topics. In particular, exponential martingales are of use in mathematical finance, where checking uniform integrability of a particular exponential martingale can be used to prove absence of arbitrage and obtain equivalent martingale measures for option pricing. For more on this, see [135] or chapters 10 and 11 of [15]. Also, exponential martingales arise naturally in connection with maximum likelihood estimation for stochastic processes, where the likelihood viewed as a stochastic process often is an exponential martingale which is a true martingale, see for example the likelihood for parameter estimation for Poisson processes given in (3.43) of [92] or the likelihood for parameter estimation for diffusion processes given in Theorem 1.12 of [103]. Finally, exponential martingales which are true martingales can be used in the explicit construction of various probabilistic objects, for example solutions to stochastic differential equations, as in

Section 5.3.B of [91].

Several sufficient criteria for $\mathcal{E}(M)$ to be a uniformly integrable martingale are known. First results in this regard were obtained by [123] for the case of continuous local martingales. Here, we are interested in the case where the local martingale M is not necessarily continuous. Sufficient criteria for $\mathcal{E}(M)$ to be a uniformly integrable martingale in this case have been obtained by [109, 79, 124, 175, 89].

We now explain the particular result to be obtained in this paper. In [123], the following result was obtained: If M is a continuous local martingale with initial value zero and $\exp(\frac{1}{2}[M]_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale. This criterion is known as Novikov's criterion. We wish to understand whether this result can be extended to local martingales which are not continuous.

In the case with jumps, another process in addition to the quadratic variation process is relevant: the predictable quadratic variation. As noted earlier, the predictable quadratic variation is defined for any locally square-integrable local martingale M with initial value zero, is denoted $\langle M \rangle$, and is the unique predictable, increasing and locally integrable process with initial value zero such that $[M] - \langle M \rangle$ is a local martingale, see p. 124 of [134]. For a continuous local martingale M with initial value zero, we have that M always is locally square integrable and $\langle M \rangle = [M]$.

Using the predictable quadratic variation, the following result is demonstrated in Theorem 9 of [135]. Let M be a locally square integrable local martingale with initial value zero and $\Delta M \geq -1$. It then holds that if $\exp(\frac{1}{2}\langle M^c \rangle_\infty + \langle M^d \rangle_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale. This is an extension of the classical Novikov criterion of [123] to the case with jumps. [135] also argue in Example 10 that the constants in front of $\langle M^c \rangle$ and $\langle M^d \rangle$ are optimal, although their argument contains a flaw, namely that the formula (28) in that paper does not hold.

In this paper, we specialize our efforts to the case where M has jumps larger than or equal to a for some $a \geq -1$ and prove results of the same type, requiring either that M is a locally square integrable local martingale and that the variable $\exp(\frac{1}{2}\langle M^c \rangle_\infty + \alpha(a)\langle M^d \rangle_\infty)$ is integrable for some $\alpha(a)$, or that the variable $\exp(\frac{1}{2}[M^c]_\infty + \beta(a)[M^d]_\infty)$ is integrable for some $\beta(a)$. For all $a \geq -1$, we identify the optimal values of $\alpha(a)$ and $\beta(a)$, in particular giving an argument circumventing the problems of Example 10 in [135]. Our results are stated as Theorem 3.2.4 and Theorem 3.2.5. In particular, we obtain that for local martingales M with initial value zero and $\Delta M \geq 0$, $\mathcal{E}(M)$ is a uniformly integrable martingale if $\exp(\frac{1}{2}[M]_\infty)$ is integrable or if M is locally square integrable and $\exp(\frac{1}{2}\langle M \rangle_\infty)$ is integrable, and we obtain that both the constants in the exponents and the requirement on the jumps of M are optimal. This result is stated as Corollary 3.2.6 and yields a verbatim extension of the Novikov criterion to local martingales M with initial value zero and $\Delta M \geq 0$.

3.2 Main results and proofs

In this section, we apply the results of [109] to obtain optimal constants in Novikov-type criteria for local martingales with jumps. For $a > -1$ with $a \neq 0$, we define

$$\alpha(a) = \frac{(1+a)\log(1+a) - a}{a^2} \quad \text{and} \tag{3.2}$$

$$\beta(a) = \frac{(1+a)\log(1+a) - a}{a^2(1+a)}, \tag{3.3}$$

and put $\alpha(0) = \beta(0) = \frac{1}{2}$ and $\alpha(-1) = 1$. The functions α and β will yield the optimal constants in the criteria we will be demonstrating. Before proving our main results, Theorem 3.2.4 and Theorem 3.2.5, we state three lemmas.

Lemma 3.2.1. *The functions α and β are continuous, positive and strictly decreasing. Furthermore, $\beta(a)$ tends to infinity as a tends to minus one.*

Proof. We first prove the result on α. Define $h(a) = (1+a)\log(1+a) - a$ for $a > -1$ and $h(-1) = 1$. Note that h is differentiable with $h'(a) = \log(1+a)$. By the l'Hôpital rule, we have

$$\lim_{a\to -1} h(a) = 1 + \lim_{a\to -1} \frac{\log(1+a)}{(1+a)^{-1}} = 1 - \lim_{a\to -1} \frac{(1+a)^{-1}}{(1+a)^{-2}} = 1, \tag{3.4}$$

which yields that h and α are continuous at -1. Similarly,

$$\lim_{a\to 0} \alpha(a) = \lim_{a\to 0} \frac{\log(1+a)}{2a} = \lim_{a\to 0} \frac{1}{2(1+a)} = \frac{1}{2}, \tag{3.5}$$

so α is continuous at 0. As h is zero at zero, $h(a)$ is positive for $a \neq 0$, from which is follows that α is positive. It remains to show that α is strictly decreasing. For $a \geq -1$ with $a \notin \{-1, 0\}$, we have that α is differentiable with

$$\begin{aligned} \alpha'(a) &= \frac{a^2\log(1+a) - 2((1+a)\log(1+a) - a)a}{a^4} \\ &= \frac{2a^2 - a(2+a)\log(1+a)}{a^4}. \end{aligned} \tag{3.6}$$

By the l'Hôpital rule, we obtain

$$\begin{aligned} \lim_{a\to 0} \alpha'(a) &= \lim_{a\to 0} \frac{4a - 2(1+a)\log(1+a) - a(2+a)(1+a)^{-1}}{4a^3} \\ &= \lim_{a\to 0} \frac{a(2+a)(1+a)^{-2} - 2\log(1+a)}{12a^2} \\ &= -\lim_{a\to 0} \frac{2a(2+a)(1+a)^{-3}}{24a} = -\frac{1}{12}\lim_{a\to 0} \frac{2+a}{(1+a)^3} = -\frac{1}{6}, \end{aligned} \tag{3.7}$$

so defining $\alpha'(0) = -\frac{1}{6}$, we obtain that α' is a continuous mapping on $(-1, \infty)$, and as α' is the derivative of α for $a \geq -1$ with $a \notin \{-1, 0\}$, α' is also the derivative of α for $(1, \infty)$. In order to show that α is strictly decreasing, it then suffices to show that that $2a^2 - a(2+a)\log(1+a)$ is negative for $a > -1$ with $a \neq 0$. Now, for $a \neq 0$, note that

$$\frac{\mathrm{d}}{\mathrm{d}a}(2a - (2+a)\log(1+a)) = 2 - \log(1+a) - \frac{2+a}{1+a} \quad \text{and} \tag{3.8}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}a^2}(2a - (2+a)\log(1+a)) = \frac{1}{(1+a)^2} - \frac{1}{1+a} = -\frac{a}{(1+a)^2}. \tag{3.9}$$

From this, we conclude that $a \mapsto 2a - (2+a)\log(1+a)$ is positive for $-1 < a < 0$ and negative for $a > 0$. Therefore, $a \mapsto 2a^2 - a(2+a)\log(1+a)$ is negative for $a > -1$ with $a \neq 0$. As a consequence, α is strictly decreasing. As $\beta(a) = \alpha(a)/(1+a)$, the results on β follow from those on α. □

Lemma 3.2.2. *Let N be a standard Poisson process, let b and λ be in $\mathbb{R}$, and define $f_b(\lambda) = \exp(-\lambda) + \lambda(1+b) - 1$. With $L_t^b = \exp(-\lambda(N_t - (1+b)t) - t f_b(\lambda))$, L^b is a nonnegative martingale with respect to the filtration induced by N.*

Proof. Let $\mathcal{G}_t = \sigma(N_s)_{s \leq t}$. Fix $0 \leq s \leq t$. As $N_t - N_s$ is independent of $\mathcal{G}_s$ and follows a Poisson distribution with parameter $t - s$, we obtain

$$E(\exp(-\lambda(N_t - N_s))|\mathcal{G}_s) = \exp((t-s)(\exp(-\lambda) - 1)), \tag{3.10}$$

which implies

$$\begin{aligned} E(L_t^b|\mathcal{G}_s) &= E(\exp(-\lambda(N_t - N_s))|\mathcal{G}_s)\exp(-\lambda N_s)\exp(\lambda(1+b)t - t f_b(\lambda)) \\ &= \exp((t-s)(\exp(-\lambda) - 1))\exp(-\lambda N_s)\exp(\lambda(1+b)t - t f_b(\lambda)) \\ &= \exp(-\lambda(N_s - (1+b)s) - s f_b(\lambda)) = L_s^b, \end{aligned} \tag{3.11}$$

proving the lemma. □

Lemma 3.2.3. *Let M be a local martingale with initial value zero and $\Delta M \geq -1$. Then $E\mathcal{E}(M)_\infty \leq 1$, and $\mathcal{E}(M)$ is a uniformly integrable martingale if and only if $E\mathcal{E}(M)_\infty = 1$.*

Proof. This follows from the the optional sampling theorem for nonnegative supermartingales. □

In the proof of Theorem 3.2.4, note that for a standard Poisson process N, it holds that with $M_t = N_t - t$, $\langle M \rangle_t = t$, since $[M]_t = N_t$ by Definition VI.37.6 of [143] and since $\langle M \rangle$ is the unique predictable and locally integrable increasing process making $[M] - \langle M \rangle$ a local martingale.

Theorem 3.2.4. *Fix $a \geq -1$. Let M be a locally square integrable local martingale with initial value zero and $\Delta M 1_{(\Delta M \neq 0)} \geq a$. If $\exp(\frac{1}{2}\langle M^c\rangle_\infty + \alpha(a)\langle M^d\rangle_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale. Furthermore, for all $a \geq -1$, the coefficients $\frac{1}{2}$ and $\alpha(a)$ in front of $\langle M^c\rangle_\infty$ and $\langle M^d\rangle_\infty$ are optimal in the sense that the criterion is false if any of the coefficients are reduced.*

Proof. Sufficiency. With $h(x) = (1+x)\log(1+x) - x$, we find by Lemma 3.2.1 that for $-1 \leq a \leq x$, $\alpha(a) \geq \alpha(x)$, which implies $h(x) \leq \alpha(a)x^2$. Letting $a \geq -1$ and letting M be a locally square integrable local martingale with initial value zero, $\Delta M 1_{(\Delta M \neq 0)} \geq a$ and $\exp(\frac{1}{2}\langle M^c\rangle_\infty + \alpha(a)\langle M^d\rangle_\infty)$ integrable, we obtain for all $t \geq 0$ the inequality $(1+\Delta M_t)\log(1+\Delta M_t) - \Delta M_t \leq \alpha(a)(\Delta M_t)^2$, and so Theorem III.1 of [109] shows that $\mathcal{E}(M)$ is a uniformly integrable martingale. Thus, the condition is sufficient.

As regards optimality of the coefficients, optimality of the coefficient $\frac{1}{2}$ in front of $\langle M^c\rangle$ is well-known, see [123]. It therefore suffices to to prove optimality of the coefficient $\alpha(a)$ in front of $\langle M^d\rangle$. To do so, we need to show the following: That for each $\varepsilon > 0$, there exists a locally square integrable local martingale with initial value zero and $\Delta M 1_{(\Delta M \neq 0)} \geq a$ such that $\exp(\frac{1}{2}\langle M^c\rangle_\infty + (1-\varepsilon)\alpha(a)\langle M^d\rangle_\infty)$ is integrable, while $\mathcal{E}(M)$ is not a uniformly integrable martingale.

The case $a > 0$. Let $\varepsilon, b > 0$, put $T_b = \inf\{t \geq 0 \mid N_t - (1+b)t = -1\}$ and define $M_t = a(N_t^{T_b} - t \wedge T_b)$. We claim that we may choose $b > 0$ such that M satisfies the requirements stated above. It holds that M is a locally square integrable local martingale with initial value zero and $\Delta M 1_{(\Delta M \neq 0)} \geq a$, and M is purely discontinuous by Definition 7.21 of [66] since it is of locally integrable variation. In particular, $M^c = 0$, so it suffices to show that $\exp((1-\varepsilon)\alpha(a)\langle M\rangle_\infty)$ is integrable while $\mathcal{E}(M)$ is not a uniformly integrable martingale. To show this, we first argue that T_b is almost surely finite. To this end, note that since $t \mapsto N_t - (1+b)t$ only has nonnegative jumps, has initial value zero and decreases between jumps, the process hits -1 if and only if it is less than or equal to -1 immediately before one of its jumps. Therefore, with U_n denoting the n'th jump time of N, we have

$$\begin{aligned} P(T_b = \infty) &= P(\cap_{n=1}^\infty (N_{U_n-} - (1+b)U_n > -1)) \\ &= P(\cap_{n=1}^\infty (n > U_n(1+b))) \\ &\leq P(\limsup_{n\to\infty} U_n/n \leq (1+b)^{-1}), \end{aligned} \tag{3.12}$$

which is zero, as $\lim_{n\to\infty} U_n/n = 1$ almost surely by the law of large numbers, and $(1+b)^{-1} < 1$ as $b > 0$. Thus, T_b is almost surely finite, and by the path properties of N, $N_{T_b} = (1+b)T_b - 1$ almost surely. We then obtain

$$\begin{aligned} \mathcal{E}(M)_\infty &= \exp(a(N_{T_b} - T_b) + N_{T_b}(\log(1+a) - a)) \\ &= \exp(N_{T_b}\log(1+a) - aT_b) \\ &= \exp(((1+b)T_b - 1)\log(1+a) - aT_b) \\ &= (1+a)^{-1}\exp(T_b((1+b)\log(1+a) - a)). \end{aligned} \tag{3.13}$$

Recalling Lemma 3.2.3, we wish to choose $b > 0$ such that $E\mathcal{E}(M)_\infty < 1$ and $E\exp((1-\varepsilon)\alpha\langle M\rangle_\infty) < \infty$ holds simultaneously. Note that $\langle M\rangle_\infty = a^2 T_b$. Therefore, we need to select a positive b with the properties that

$$E\exp(T_b((1+b)\log(1+a) - a)) < 1 + a \text{ and} \tag{3.14}$$

$$E\exp(T_b a^2(1-\varepsilon)\alpha(a)) < \infty. \tag{3.15}$$

Consider some $b > 0$ and let f_b be as in Lemma 3.2.2. By that same lemma, the process L^b defined by putting $L_t^b = \exp(-\lambda(N_t - (1+b)t) - tf_b(\lambda))$ is a martingale. In particular, it is a nonnegative supermartingale with initial value one, so Theorem II.77.5 of [142] yields $1 \geq EL_{T_b}^b = E\exp(\lambda - T_b f_b(\lambda))$. As a consequence, we obtain $E\exp(-T_b f_b(\lambda)) \leq \exp(-\lambda)$. Note that $f_b'(\lambda) = -\exp(-\lambda) + 1 + b$, such that f_b takes its minimum at $-\log(1+b)$. Therefore, $-f_b$ takes its maximum at $-\log(1+b)$, and we find that the maximum is $h(b)$. In particular, $E\exp(T_b h(b))$ is finite. Next, define λ by putting $\lambda(b) = -\log((1+a)\frac{b}{a})$, we then have $E\exp(-T_b f_b(\lambda(b))) \leq (1+a)\frac{b}{a}$, which is strictly less than $1 + a$ whenever $b < a$. Thus, if we can choose $b \in (0, a)$ such that

$$(1+b)\log(1+a) - a \leq -f_b(\lambda(b)) \text{ and} \tag{3.16}$$

$$a^2(1-\varepsilon)\alpha(a) \leq h(b), \tag{3.17}$$

we will have achieved our end, since (3.16) implies (3.14) and (3.17) implies (3.15). To this end, note that

$$\begin{aligned} -f_b(\lambda(b)) &= -\exp(\log((1+a)\tfrac{b}{a})) + \log((1+a)\tfrac{b}{a})(1+b) + 1 \\ &= 1 - (1+a)\tfrac{b}{a} + (1+b)\log((1+a)\tfrac{b}{a}) \\ &= 1 - (1+a)\tfrac{b}{a} + (1+b)\log(1+a) + (1+b)\log\tfrac{b}{a}, \end{aligned} \tag{3.18}$$

such that, by rearrangement, (3.16) is equivalent to

$$0 \leq 1 + a - (1+a)\tfrac{b}{a} + (1+b)\log\tfrac{b}{a}, \tag{3.19}$$

and therefore, as $1 - \frac{b}{a} > 0$ for $0 < b < a$, equivalent to

$$(1+b)\frac{\log\frac{b}{a}}{\frac{b}{a} - 1} \leq 1 + a, \tag{3.20}$$

which, as $\log x \leq x - 1$ for $x > 0$, is satisfied for all $0 < b < a$. Thus, it suffices to choose $b \in (0, a)$ such that (3.17) is satisfied, corresponding to choosing $b \in (0, a)$ such that $(1-\varepsilon)h(a) \leq h(b)$. As h is positive and continuous on $(0, \infty)$, this is possible by choosing b close enough to a. With this choice of b, we now obtain M yielding an example proving that the coefficient $\alpha(a)$ is optimal. This concludes the proof of optimality in the case $a > 0$.

The case $a = 0$. Let $\varepsilon > 0$. To prove optimality, we wish to identify a locally square integrable local martingale M with initial value zero and $\Delta M 1_{(\Delta M \neq 0)} \geq 0$ such

that $\exp((1-\varepsilon)\alpha(0)\langle M\rangle_\infty)$ is integrable while $\mathcal{E}(M)$ is not a uniformly integrable martingale. Recalling that α is positive and continuous, pick $a > 0$ so close to zero that $(1-\varepsilon)\alpha(0) \leq (1-\frac{1}{2}\varepsilon)\alpha(a)$. By what was already shown, there is a locally square integrable local martingale M with initial value zero and $\Delta M 1_{(\Delta M \neq 0)} \geq a$ such that $\exp((1-\frac{1}{2}\varepsilon)\alpha(a)\langle M\rangle_\infty)$ is integrable while $\mathcal{E}(M)$ is not a uniformly integrable martingale. As $\exp((1-\varepsilon)\alpha(0)\langle M\rangle_\infty)$ is integrable in this case, this shows that $\alpha(0)$ is optimal.

The case $-1 < a < 0$. Let $\varepsilon > 0$, let $-1 < b < 0$, let $c > 0$ and define a stopping time T_{bc} by putting $T_{bc} = \inf\{t \geq 0 \mid N_t - (1+b)t \geq c\}$. Also define a local martingale M by $M_t = a(N_t^{T_{bc}} - t \wedge T_{bc})$. We claim that we can choose $b \in (-1, 0)$ and $c > 0$ such that M satisfies the requirements to show optimality. Similarly to the case $a > 0$, M is a purely discontinuous locally square integrable local martingale with initial value zero and $\Delta M 1_{(\Delta M \neq 0)} \geq a$, so it suffices to show that $\exp((1-\varepsilon)\alpha(a)\langle M\rangle_\infty)$ is integrable while $\mathcal{E}(M)$ is not a uniformly integrable martingale. We first investigate some properties of T_{bc}. As $t \mapsto N_t - (1+b)t$ only has nonnegative jumps, has initial value zero and decreases between jumps, the process advances beyond c at some point if and only it advances beyond c at one of its jump times. Therefore, with U_n denoting the n'th jump time of N,

$$\begin{aligned} P(T_{bc} = \infty) &= P(\cap_{n=1}^\infty (N_{U_n} - (1+b)U_n < c)) \\ &= P(\cap_{n=1}^\infty (n - c < U_n(1+b))) \\ &\leq P(\liminf_{n\to\infty} U_n/n \geq (1+b)^{-1}), \end{aligned} \tag{3.21}$$

which is zero, as U_n/n tends to one almost surely and $(1+b)^{-1} > 1$. Thus, T_{bc} is almost surely finite. Furthermore, by the path properties of N, $N_{T_{bc}} \geq (1+b)T_{bc} + c$ and $N_{T_{bc}} \leq (1+b)T_{bc} + c + 1$ almost surely. Since $\log(1+a) \leq 0$, we in particular obtain $N_{T_{bc}} \log(1+a) \leq ((1+b)T_{bc} + c)\log(1+a)$ almost surely. From this, we conclude that

$$\begin{aligned} \mathcal{E}(M)_\infty &= \exp(a(N_{T_{bc}} - T_{bc}) + N_{T_{bc}}(\log(1+a) - a)) \\ &= \exp(N_{T_{bc}} \log(1+a) - aT_{bc}) \\ &\leq \exp(((1+b)T_{bc} + c)\log(1+a) - aT_{bc}) \\ &= (1+a)^c \exp(T_{bc}((1+b)\log(1+a) - a)). \end{aligned} \tag{3.22}$$

We wish to choose $-1 < b < 0$ and $c > 0$ such that $E\exp((1-\varepsilon)\alpha(a)\langle M\rangle_\infty) < \infty$ and $E\mathcal{E}(M)_\infty < 1$ holds simultaneously. As $\langle M\rangle_\infty = a^2 T_{bc}$, this is equivalent to choosing $-1 < b < 0$ and $c > 0$ such that

$$E\exp(T_{bc}((1+b)\log(1+a) - a)) < (1+a)^{-c} \text{ and} \tag{3.23}$$

$$E\exp(T_{bc}a^2(1-\varepsilon)\alpha(a)) < \infty. \tag{3.24}$$

Let f_b and L^b be as in Lemma 3.2.2. The process L^b is then a nonnegative supermartingale. As $N_{T_{bc}} \leq (1+b)T_{bc} + c + 1$, the optional stopping theorem allows us

to conclude that for $\lambda \geq 0$,

$$\begin{aligned} 1 \geq EL^b_{T_{bc}} &= E\exp(-\lambda(N_{T_{bc}} - (1+b)T_{bc}) - T_{bc}f_b(\lambda)) \\ &\geq E\exp(-(c+1)\lambda - T_{bc}f_b(\lambda)), \end{aligned} \tag{3.25}$$

so that $E\exp(-T_{bc}f_b(\lambda)) \leq \exp((c+1)\lambda)$. As in the case $a > 0$, $-f_b$ takes its maximum at $-\log(1+b)$, and the maximum is $h(b)$, leading us to conclude that $E\exp(T_{bc}h(b))$ is finite. Put $\lambda(b,c) = (c+1)^{-1}\log((1+a)^{-c}\frac{b}{a})$. For all $b \in (a,0)$, $\frac{b}{a} < 1$, leading to $E\exp(-T_{bc}f_b(\lambda(b,c))) \leq (1+a)^{-c}\frac{b}{a} < (1+a)^{-c}$. Therefore, if we can choose $b \in (a,0)$ and $c > 0$ such that

$$(1+b)\log(1+a) - a \leq -f_b(\lambda(b,c)) \text{ and} \tag{3.26}$$

$$a^2(1-\varepsilon)\alpha(a) \leq h(b), \tag{3.27}$$

we will have obtained existence of a local maringale yielding the desired optimality of $\alpha(a)$. We first note that $a^2(1-\varepsilon)\alpha(a) \leq h(b)$ is equivalent to $(1-\varepsilon)h(a) \leq h(b)$. As h is continuous and positive on $(-1,0)$, we find that (3.27) is satisfied for $a < b < 0$ with b close enough to a. Next, we turn our attention to (3.26). We have

$$\begin{aligned} -f_b(\lambda(b,c)) &= -\exp\left(-\frac{1}{c+1}\log\left((1+a)^{-c}\frac{b}{a}\right)\right) - \frac{1+b}{c+1}\log\left((1+a)^{-c}\frac{b}{a}\right) + 1 \\ &= 1 - (1+a)^{\frac{c}{c+1}}\left(\frac{b}{a}\right)^{-\frac{1}{(c+1)}} - \frac{1+b}{c+1}\log\left((1+a)^{-c}\frac{b}{a}\right) \\ &= 1 - (1+a)^{\frac{c}{c+1}}\left(\frac{a}{b}\right)^{\frac{1}{c+1}} + \frac{c(1+b)}{c+1}\log(1+a) + \frac{1+b}{c+1}\log\frac{a}{b}, \end{aligned} \tag{3.28}$$

such that (3.26) is equivalent to

$$0 \leq 1 + a - (1+a)^{\frac{c}{c+1}}\left(\frac{a}{b}\right)^{\frac{1}{c+1}} + \frac{1+b}{c+1}\left(\log\frac{a}{b} - \log(1+a)\right). \tag{3.29}$$

Fixing $a < b < 0$, we wish to argue that for b close enough to a, (3.29) holds for c large enough. To this end, let $\rho_b(c)$ denote the right-hand side of (3.29). Then $\lim_{c\to\infty}\rho_b(c) = 0$. We also note that $\frac{\mathrm{d}}{\mathrm{d}c}\frac{1}{c+1} = -\frac{1}{(c+1)^2}$ and $\frac{\mathrm{d}}{\mathrm{d}c}\frac{c}{c+1} = \frac{1}{(c+1)^2}$, yielding

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}c}(1+a)^{\frac{c}{c+1}}\left(\frac{a}{b}\right)^{\frac{1}{c+1}} \\ &= \frac{\mathrm{d}}{\mathrm{d}c}\exp\left(\frac{c}{c+1}\log(1+a) + \frac{1}{c+1}\log\frac{a}{b}\right) \\ &= \left(\frac{\log(1+a)}{(c+1)^2} - \frac{\log\frac{a}{b}}{(c+1)^2}\right)\exp\left(\frac{c}{c+1}\log(1+a) + \frac{1}{c+1}\log\frac{a}{b}\right) \\ &= \frac{\log(1+a) - \log\frac{a}{b}}{(c+1)^2}\exp\left(\frac{c}{c+1}\log(1+a) + \frac{1}{c+1}\log\frac{a}{b}\right) \end{aligned} \tag{3.30}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}c}\frac{1+b}{c+1}\left(\log\frac{a}{b}-\log(1+a)\right)=-\frac{1+b}{(c+1)^2}\left(\log\frac{a}{b}-\log(1+a)\right)$$
$$=(1+b)\frac{\log(1+a)-\log\frac{a}{b}}{(c+1)^2}, \tag{3.31}$$

which leads to

$$\rho_b'(c)=\frac{\log(1+a)-\log\frac{a}{b}}{(c+1)^2}\left(1+b-\exp\left(\frac{c}{c+1}\log(1+a)+\frac{1}{c+1}\log\frac{a}{b}\right)\right). \tag{3.32}$$

Now note that for $a<b$, we obtain

$$\lim_{c\to\infty}1+b-\exp\left(\frac{c}{c+1}\log(1+a)+\frac{1}{c+1}\log\frac{a}{b}\right)=1+b-(1+a)>0, \tag{3.33}$$

and for b close enough to a, $\log(1+a)-\log\frac{a}{b}<0$, since $a<0$. Therefore, for all c large enough, $\rho_b'(c)<0$. Consider such a c, we then obtain

$$\rho_b(c)=\lim_{y\to\infty}\rho_b(c)-\rho_b(y)=-\lim_{y\to\infty}\int_c^y\rho_b'(z)\,\mathrm{d}z>0. \tag{3.34}$$

Thus, we conclude that for b close enough to a, it holds that $\rho_b(c)>0$ for c large enough.

We now collect our conclusions in order to obtain $b\in(a,0)$ and $c>0$ satisfying (3.26) and (3.27). First choose $b\in(a,0)$ so close to a that $(1-\varepsilon)h(a)\leq h(b)$ and $\log(1+a)-\log\frac{a}{b}<0$. Pick c so large that $\rho_b(c)>0$. By our deliberations, (3.26) and (3.27) then both hold, demonstrating the existence of a locally square integrable local martingale M with initial value zero and $\Delta M1_{(\Delta M\neq 0)}\geq a$ such that $\exp((1-\varepsilon)\alpha(a)\langle M\rangle_\infty)$ is integrable while $\mathcal{E}(M)$ is not a uniformly integrable martingale.

The case $a=-1$. Let $\varepsilon>0$. We wish to identify a purely discontinous locally square integrable local martingale M with $\Delta M1_{(\Delta M\neq 0)}\geq -1$ such that integrability of $\exp((1-\varepsilon)\alpha(-1)\langle M\rangle_\infty)$ holds while $\mathcal{E}(M)$ is not a uniformly integrable martingale. We proceed as in the case $a=0$. By positivity and continuity of α, take $a>0$ so close to -1 that $(1-\varepsilon)\alpha(-1)\leq(1-\frac{1}{2}\varepsilon)\alpha(a)$. By what was shown in the previous case, there exists a purely discontinuous locally square integrable local martingale M with initial value zero and $\Delta M1_{(\Delta M\neq 0)}\geq a$ such that $\exp((1-\frac{1}{2}\varepsilon)\alpha(a)\langle M\rangle_\infty)$ is integrable while $\mathcal{E}(M)$ is not a uniformly integrable martingale. As it then also holds that $\exp((1-\varepsilon)\alpha(-1)\langle M\rangle_\infty)$ is integrable, this shows that $\alpha(-1)$ is optimal. □

Theorem 3.2.5. *Fix $a>-1$. Let M be a local martingale with initial vaule zero and $\Delta M1_{(\Delta M\neq 0)}\geq a$. If $\exp(\frac{1}{2}[M^c]_\infty+\beta(a)[M^d]_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale. Furthermore, for all $a>-1$, the coefficients $\frac{1}{2}$ and $\beta(a)$ in front of $[M^c]_\infty$ and $[M^d]_\infty$ are optimal in the sense that the criterion is false if any of the coefficients are reduced.*

Furthermore, there exists no $\beta(-1)$ such that for M with $\Delta M 1_{(\Delta M \neq 0)} \geq -1$, integrability of $\exp(\frac{1}{2}[M^c]_\infty + \beta(-1)[M^d]_\infty)$ suffices to ensure that $\mathcal{E}(M)$ is a uniformly integrable martingale.

Proof. Sufficiency. We proceed in a manner closely related to the proof of Theorem 3.2.4. Defining g by putting $g(x) = \log(1+x) - x/(1+x)$, we find by Lemma 3.2.1 that for $-1 < a \leq x$, $\beta(a) \geq \beta(x)$, yielding $g(x) \leq \beta(a)x^2$. Letting $a > -1$ and letting M be a locally square integrable local martingale with initial value zero, $\Delta M 1_{(\Delta M \neq 0)} \geq a$ and $\exp(\frac{1}{2}[M^c]_\infty + \beta(a)[M^d]_\infty)$ integrable, we obtain for all $t \geq 0$ that $\log(1+\Delta M_t) - \Delta M_t/(1+\Delta M_t) \leq \beta(a)(\Delta M_t)^2$, and so Theorem III.7 of [109] shows that $\mathcal{E}(M)$ is a uniformly integrable martingale. Thus, the condition is sufficient.

As in Theorem 3.2.4, optimality of the $\frac{1}{2}$ in front of $[M^c]$ follows from [123], so it suffices to consider the coefficient $\beta(a)$ in front of $[M^d]$. Thus, for $a > -1$, we need for each $\varepsilon > 0$, to find a locally square integrable local martingale with initial value zero and $\Delta M 1_{(\Delta M \neq 0)} \geq a$ such that $\exp(\frac{1}{2}[M^c]_\infty + (1-\varepsilon)\beta(a)[M^d]_\infty)$ is integrable, while $\mathcal{E}(M)$ is not a uniformly integrable martingale.

The case $a > 0$. Let $\varepsilon, b > 0$, put $T_b = \inf\{t \geq 0 \mid N_t - (1+b)t = -1\}$ and define $M_t = a(N_t^{T_b} - t \wedge T_b)$. Noting that $[M]_\infty = a^2 N_{T_b}$, we may argue as in the proof of Theorem 3.2.4 and obtain that it suffices to identify $b > 0$ such that

$$E\exp(T_b((1+b)\log(1+a) - a)) < 1 + a \text{ and} \tag{3.35}$$

$$E\exp(N_{T_b}a^2(1-\varepsilon)\beta(a)) < \infty. \tag{3.36}$$

Let f_b be as in Lemma 3.2.2. As in the proof of Theorem 3.2.4, we obtain that $E\exp(T_b h(b))$ is finite, where $h(x) = (1+x)\log(1+x) - x$, and furthermore obtain that with $\lambda(b) = -\log((1+a)\frac{b}{a})$, $E\exp(-T_b f_b(\lambda(b))) < 1 + a$ for $b < a$. As $N_{T_b} = (1+b)T_b - 1$ almost surely and $g(b) = h(b)/(1+b)$, we then also obtain that $E\exp(N_{T_b}g(b))$ is finite. Thus, if we can choose $b \in (0, a)$ such that

$$(1+b)\log(1+a) - a \leq -f_b(\lambda(b)) \text{ and} \tag{3.37}$$

$$a^2(1-\varepsilon)\beta(a) \leq g(b), \tag{3.38}$$

we will obtain the desired result, as (3.37) implies (3.35) and (3.38) implies (3.36). As earlier noted, (3.37) always holds for $0 < b < a$. As for (3.38), this requirement is equivalent to having that $(1-\varepsilon)g(a) \leq g(b)$ for some $b \in (0, a)$, which by continuity of g can be obtained by choosing b close enough to a. Choosing b in this manner, we obtain M yielding an example proving that the coefficient $\beta(a)$ is optimal. This concludes the proof of optimality in the case $a > 0$.

The case $a = 0$. This follows similarly to the corresponding case in the proof of Theorem 3.2.4.

The case $-1 < a < 0$. Let $\varepsilon > 0$, let $-1 < b < 0$, let $c > 0$ and define a stopping time T_{bc} by putting $T_{bc} = \inf\{t \geq 0 \mid N_t - (1+b)t \geq c\}$. Also define a local martingale

M by $M_t = a(N_t^{T_{bc}} - t \wedge T_{bc})$. As in the proof of Theorem 3.2.4, in order to obtain the desired counterexample, it suffices to choose $-1 < b < 0$ and $c > 0$ such that

$$E \exp(T_{bc}((1+b)\log(1+a) - a)) < (1+a)^{-c} \text{ and} \tag{3.39}$$

$$E \exp(T_{bc} a^2 (1-\varepsilon)\beta(a)) < \infty. \tag{3.40}$$

With f_b as in Lemma 3.2.2, we find as in the proof of Theorem 3.2.4 that $\exp(T_{bc}h(b))$ has finite mean. Furthermore, defining $\lambda(b,c) = (c+1)^{-1}\log((1+a)^{-c}\frac{b}{a})$, it holds for b with $a < b \leq (1+a)^c a$ that $\lambda(b,c) \geq 0$ and $E\exp(-T_{bc}f_b(\lambda(b,c))) < (1+a)^{-c}$. Also, as $N_{T_{bc}} \leq (1+b)T_{bc} + c + 1$, $E\exp(N_{T_{bc}}(1+b)^{-1}h(b))$ and thus $E\exp(N_{T_{bc}}g(b))$ is finite. Therefore, if we can choose $b \in (a,0)$ and $c > 0$ such that

$$(1+b)\log(1+a) - a \leq -f_b(\lambda(b,c)) \text{ and} \tag{3.41}$$

$$a^2(1-\varepsilon)\beta(a) \leq g(b), \tag{3.42}$$

we obtain the desired result. By arguments as in the proof of the corresponding case of Theorem 3.2.4, we find that by first picking b close enough to a and then c large enough, we can ensure that both (3.41) and (3.42) hold, yielding optimality for this case.

The case $a = -1$. For this case, we need to show that for any $\gamma \geq 0$, it does not hold that finiteness of $E\exp(\gamma[M^d]_\infty)$ implies that $\mathcal{E}(M)$ is a uniformly integrable martingale. Let $\gamma \geq 0$. By Lemma 3.2.1, $\beta(a)$ tends to infinity as a tends to -1. Therefore, we may pick $a > -1$ so small that $\beta(a) \geq \gamma$. By what we already have shown, there exists M with initial value zero and $\Delta M 1_{(\Delta M \neq 0)} \geq -1$ such that $E\exp(\beta(a)[M^d]_\infty)$ and thus $E\exp(\gamma[M^d]_\infty)$ is finite, while $\mathcal{E}(M)$ is not a uniformly integrable martingale. □

Corollary 3.2.6. *Let M be a local martingale with initial value zero and $\Delta M \geq 0$. If $\exp(\frac{1}{2}[M]_\infty)$ is integrable or if M is locally square integrable and $\exp(\frac{1}{2}\langle M\rangle_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale. Furthermore, this criterion is optimal in the sense that if either the constant $\frac{1}{2}$ is reduced, or the requirement on the jumps is weakened to $\Delta M \geq -\varepsilon$ for some $\varepsilon > 0$, the criterion ceases to be sufficient.*

Proof. That the constant $\frac{1}{2}$ cannot be reduced follows from Theorem 3.2.4 and Theorem 3.2.5. That the requirement on the jumps cannot be reduced follows by combining Theorem 3.2.4 and Theorem 3.2.5 with the fact that α and β both are strictly decreasing by Lemma 3.2.1. □

4

An extended Novikov-type criterion for local martingales with jumps

Alexander Sokol

2010 Mathematics Subject Classification. Primary 60G44; Secondary 60G40.

Key words and phrases. Martingale, Exponential martingale, Uniform integrability, Novikov.

Abstract. For local martingales with nonnegative jumps, we prove a sufficient criterion for the corresponding exponential martingale to be a uniformly integrable martingale. The criterion is in terms of exponential moments of a convex combination of the optional and predictable quadratic variation. The result extends earlier known criteria.

4.1 Introduction

In [123], Novikov introduced a sufficient criterion for the exponential martingale of a continuous local martingale to be a uniformly integrable martingale. In this paper, we prove a similar result in the case where the local martingale is not continuous, but is assumed to have nonnegative jumps. The novelty of our criterion rests in that our result is stronger than previously known results, in that it combines optional

and predictable components and in that our proof of the criterion demonstrates a straightforward two-step structure. We begin by fixing our notation and recalling some results from stochastic analysis.

Assume given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ satisfying the usual conditions, see [134] for the definition of this and other probabilistic concepts such as localising sequences, local martingales, the quadratic covariation et cetera. For any local martingale M, we say that M has initial value zero if $M_0 = 0$. For any local martingale M with initial value zero, we denote by $[M]$ the quadratic variation of M, that is, the unique increasing adapted process with initial value zero such that $M^2 - [M]$ is a local martingale.

If A is an adapted increasing process with initial value zero, we say that A is integrable if EA_∞ is finite, and we say that A is locally integrable if A^{T_n} is integrable for some localising sequence (T_n), that is, a sequence of stopping times increasing to infinity. If A is an adapted process with initial value zero and paths of finite variation, we say that A is locally integrable if the variation process is locally integrable. Whenever A is adapted, has initial value zero, is of finite variation and is locally integrable, there exists a predictable process $\Pi_p^* A$ with those same properties such that $A - \Pi_p^* A$ is a local martingale, see Definition VI.21.3 of [143]. We refer to $\Pi_p^* A$ as the dual predictable projection of A, or simply as the compensator of A.

If M is locally square integrable, it holds that $[M]$ is locally integrable, and we denote by $\langle M \rangle$ the compensator of $[M]$. We refer to $\langle M \rangle$ as the predictable quadratic variation of M. It then holds that $M^2 - \langle M \rangle$ is a local martingale.

For any local martingale with initial value zero, there exists by Theorem 7.25 of [66] a unique decomposition $M = M^c + M^d$, where M^c is a continuous local martingale and M^d is a purely discontinuous local martingale, both with initial value zero. Here, we say that a local martingale with initial value zero is purely discontinuous if it has zero quadratic covariation with any continuous local martingale with initial value zero. We refer to M^c as the continuous martingale part of M, and refer to M^d as the purely discontinuous martingale part of M.

With M a local martingale with initial value zero and $\Delta M \geq 0$, the exponential martingale of M, also known as the Doléans-Dade exponential of M, is given by

$$\mathcal{E}(M)_t = \exp\left(M_t - \frac{1}{2}[M^c]_t\right) \prod_{0<s\leq t} (1 + \Delta M_s) \exp(-\Delta M_s). \tag{4.1}$$

The process $\mathcal{E}(M)$ is the unique càdlàg solution in Z to the stochastic differential equation $Z_t = 1 + \int_0^t Z_{s-} \, \mathrm{d}M_s$, see Theorem II.37 of [134]. By Theorem 9.2 of [66], $\mathcal{E}(M)$ is always a local martingale with initial value one. We are interested in sufficient criteria to ensure that $\mathcal{E}(M)$ is a uniformly integrable martingale. This is a classical question in probability theory, with applications in finance, stochatic differential equations and statistical inference for continuously observed stochastic processes, see for example [135, 15, 91, 92, 103]. For the case when M is continuous,

sufficient criteria ensuring that $\mathcal{E}(M)$ is a uniformly integrable martingale have been obtained in [123, 24, 94, 95, 119]. For the case when M has jumps, see [109, 79, 124, 175, 89].

We now explain the particular result to be obtained in this paper. In [123], the following result was obtained: If M is a continuous local martingale with initial value zero and $\exp(\frac{1}{2}[M]_\infty)$ is integrable, then $\mathcal{E}(M)$ is a uniformly integrable martingale. This criterion is known as Novikov's criterion. In [101], it was shown that for a continuous local martingale M with initial value zero, the condition

$$\liminf_{\varepsilon\to 0} \varepsilon \log E \exp\left((1-\varepsilon)\frac{1}{2}[M]_\infty\right) < \infty \tag{4.2}$$

suffices to ensure that $\mathcal{E}(M)$ is a uniformly integrable martingale. This is an extension of the result in [123]. And in [160], optimal constants $\alpha(a)$ and $\beta(a)$ for $a > -1$ were identified such that when $\Delta M 1_{(\Delta M \neq 0)} \geq a$, integrability of $\exp(\alpha(a)[M]_\infty)$ and $\exp(\beta(a)[M]_\infty)$ suffices to ensure that $\mathcal{E}(M)$ is a uniformly integrable martingale, and it was noted that for the case $a = 0$, $\alpha(a) = \beta(a) = \frac{1}{2}$. Thus, the case where $\Delta M \geq 0$ presents a higher level of regularity than the general case. In this note, we prove that when $\Delta M \geq 0$, the condition

$$\liminf_{\varepsilon\to 0} \varepsilon \log E \exp\left((1-\varepsilon)\frac{1}{2}(\alpha[M]_\infty + (1-\alpha)\langle M\rangle_\infty)\right) < \infty \tag{4.3}$$

suffices to ensure that $\mathcal{E}(M)$ is a uniformly integrable martingale, thus extending the results of [123] and [101]. Note that while sufficiency of simple Novikov-type criteria such as those given in [160] follow from the results of [109], the condition (4.3) does not. Also, to the best of the knowledge of the author, the condition (4.3) is the first one obtained applying both the quadratic variation and the predictable quadratic variation at the same time.

4.2 Main results and proofs

In this section, we will prove the following theorem.

Theorem 4.2.1. *Let M be a locally square integrable local martingale with initial value zero and $\Delta M \geq 0$. Fix $0 \leq \alpha \leq 1$ and assume that*

$$\liminf_{\varepsilon\to 0} \varepsilon \log E \exp\left((1-\varepsilon)\frac{1}{2}(\alpha[M]_\infty + (1-\alpha)\langle M\rangle_\infty)\right) < \infty. \tag{4.4}$$

Then $\mathcal{E}(M)$ is a uniformly integrable martingale. If $\alpha = 1$, the result also holds without the assumption that M is locally square integrable. Furthermore, for all $0 \leq \alpha \leq 1$, the constant $1/2$ in (4.4) is optimal.

Optimality of the constant $1/2$ is well known, see [123]. We begin by considering the proof for the case $\alpha = 1$, where local square integrability is not required. Our proof in this case rests on the following two elementary martingale lemmas and the following real analysis lemma.

Lemma 4.2.2. *Let M be a local martingale with initial value zero and $\Delta M \geq 0$. Then $E\mathcal{E}(M)_\infty \leq 1$, and $\mathcal{E}(M)$ is a uniformly integrable martingale if and only if $E\mathcal{E}(M)_\infty = 1$.*

Proof. This follows from the the optional sampling theorem for nonnegative supermartingales. □

Lemma 4.2.3. *Let M be a local martingale with initial value zero. Let $\mathcal{C}$ denote the set of all bounded stopping times. If there exists $a > 1$ such that $(M_T)_{T\in\mathcal{C}}$ is bounded in $\mathcal{L}^a$, then M is a uniformly integrable martingale.*

Proof. As $(M_T)_{T\in\mathcal{C}}$ is bounded in $\mathcal{L}^a$, $(M_T)_{T\in\mathcal{C}}$ is uniformly integrable. Let (T_n) be a localising sequence such that M^{T_n} is a uniformly integrable martingale for each $n \geq 1$. Let S be a bounded stopping time. Then $(M_{T_n\wedge S})_{n\geq 1}$ is uniformly integrable as well. As $M_{T_n\wedge S}$ converges almost surely to M_S, we conclude that M_S is integrable and that $M_{T_n\wedge S}$ converges in $\mathcal{L}^1$ to M_S. As M^{T_n} is a uniformly integrable martingale, $EM_S^{T_n} = 0$ by the optional stopping theorem, and thus $EM_S = 0$. By Theorem II.77.6 of [143], M is a martingale. And by our assumptions, $(M_t)_{t\geq 0}$ is uniformly integrable, so M is a uniformly integrable martingale. □

Lemma 4.2.4. *Let $x \geq 0$. It then holds that*

$$0 \leq \log \frac{1+\lambda x}{(1+x)^\lambda} \leq \frac{\lambda(1-\lambda)}{2}x^2 \quad \text{and} \tag{4.5}$$

$$0 \leq \log \frac{(1+x)^a}{1+ax} \leq \frac{a(a-1)}{2}x^2 \tag{4.6}$$

for $0 \leq \lambda \leq 1$ and $a \geq 1$.

Proof. We first prove (4.5). To prove the lower inequality, it suffices to argue that $(1+\lambda x)/(1+x)^\lambda \geq 1$, which is equivalent to $1+\lambda x-(1+x)^\lambda \geq 0$. Fix $0 \leq \lambda \leq 1$ and define $h_\lambda(x) = 1+\lambda x-(1+x)^\lambda$. Then $h'_\lambda(x) = \lambda - \lambda(1+x)^{\lambda-1} \geq 0$ and $h_\lambda(0) = 0$. This implies the first inequality in (4.5). In order to prove the second inequality, we define g_λ by putting $g_\lambda(x) = \frac{1}{2}\lambda(1-\lambda)x^2 - \log(1+\lambda x) + \lambda\log(1+x)$. We then need to prove $g_\lambda(x) \geq 0$. We obtain $g_\lambda(0) = 0$ and

$$\begin{aligned} g'(x) &= \lambda(1-\lambda)x - \frac{\lambda}{1+\lambda x} + \frac{\lambda}{1+x} \\ &= \frac{\lambda(1-\lambda)x(1+\lambda x)(1+x) - \lambda(1+x) + \lambda(1+\lambda x)}{(1+\lambda x)(1+x)} \\ &= \frac{(\lambda-\lambda^2)(x^2+\lambda x^2+\lambda x^3)}{(1+\lambda x)(1+x)} \geq 0, \end{aligned} \tag{4.7}$$

so $g_\lambda(x) \geq 0$ for all $0 \leq \lambda \leq 1$ and $x \geq 0$, yielding the second inequality in (4.5). Next, consider (4.6). For the lower inequality, note that $(1+x)^a - (1+ax) \geq 0$, so that $(1+x)^a/(1+ax) \geq 1$. For the upper inequality, we may apply (4.5) to obtain

$$\log \frac{(1+x)^a}{1+ax} = a \log \frac{1+x}{(1+ax)^{1/a}} \leq a \frac{\frac{1}{a}(1-\frac{1}{a})}{2}(ax)^2 = \frac{a(a-1)}{2}x^2, \tag{4.8}$$

for $a \geq 1$. □

Proof of Theorem 4.2.1 for the case $\alpha = 1$. In this case, we wish to show that when $\liminf_{\varepsilon \to 0} \varepsilon \log E \exp(((1-\varepsilon)/2)[M]_\infty)$ is finite, $\mathcal{E}(M)$ is a uniformly integrable martingale. We first prove that $\mathcal{E}(M)$ is a uniformly integrable martingale under the stronger condition that $\exp((1+\varepsilon)\frac{1}{2}[M]_\infty)$ is integrable for some $\varepsilon > 0$. Fix such an $\varepsilon > 0$, and let $a, r > 1$. Applying (4.6) of Lemma 4.2.4, we then have

$$\begin{aligned}
\mathcal{E}(M)_t^a &= \exp\left(aM_t - \frac{1}{2}a[M^c]_t + \sum_{0<s\leq t} \log(1+\Delta M_s)^a - a\Delta M_s \right) \\
&= \mathcal{E}(arM)_t^{1/r} \exp\left(\frac{a(ar-1)}{2}[M^c]_t + \sum_{0<s\leq t} \log \frac{(1+\Delta M_s)^a}{(1+ar\Delta M_s)^{1/r}} \right) \\
&\leq \mathcal{E}(arM)_t^{1/r} \exp\left(\frac{a(ar-1)}{2}[M]_t \right). \qquad (4.9)
\end{aligned}$$

Now let T be a bounded stopping time. Note that as arM has nonnegative jumps, $\mathcal{E}(arM)$ is a nonnegative supermartingale and so $0 \leq E\mathcal{E}(arM)_T \leq 1$. Let $y = ar$ and let s be the dual exponent to r, such that $s = r/(r-1)$. Applying Hölder's inequality in (4.9), we obtain

$$E\mathcal{E}(M)_T^a \leq \left(E \exp\left(\frac{y(y-1)}{2(r-1)}[M]_\infty \right) \right)^{1/s}. \tag{4.10}$$

Next, note that the mapping $y \mapsto y(y-1)$ is increasing for $y \geq 1$. Therefore, $\inf_{y>r>1} y(y-1)/(2(r-1)) = \inf_{r>1} r/2 = 1/2$, and so there exists $y > r > 1$ such that $y(y-1)/(2(r-1)) \leq (1+\varepsilon)/2$. Fixing such $y > r > 1$ and putting $a = y/r$, we obtain $a > 1$ and (4.10) allows us to conclude that with the supremum being over all bounded stopping times, we have

$$\sup_T E\mathcal{E}(M)_T^a \leq \left(E \exp\left((1+\varepsilon)\frac{1}{2}[M]_\infty \right) \right)^{1/s}, \tag{4.11}$$

where the right-hand side is finite by assumption. By Lemma 4.2.3, $\mathcal{E}(M)$ is a uniformly integrable martingale.

Next, we merely assume that $\liminf_{\varepsilon \to 0} \varepsilon \log E \exp(((1-\varepsilon)/2)[M]_\infty)$ is finite. In particular, for all $\varepsilon > 0$, $\exp(((1-\varepsilon)/2)[M]_\infty)$ is integrable. Therefore, $[M]_\infty$ is

integrable, so M is a square-integrable martingale and the limit M_∞ exists. Fix $0 < \lambda < 1$. As $[\lambda M]_t = \lambda^2 [M]_t$, we have by our earlier results that $\mathcal{E}(\lambda M)$ is a uniformly integrable martingale. Using (4.5) of Lemma 4.2.4, we have

$$\begin{aligned} 1 &= E \exp\left(\lambda M_\infty - \frac{\lambda^2}{2}[M^c]_\infty + \sum_{0<t} \log(1+\lambda \Delta M_t) - \lambda \Delta M_t \right) \\ &= E\mathcal{E}(M)_\infty^\lambda \exp\left(\frac{\lambda(1-\lambda)}{2}[M^c]_\infty + \sum_{0<t} \log \frac{1+\lambda \Delta M_t}{(1+\Delta M_t)^\lambda} \right) \\ &\leq E\mathcal{E}(M)_\infty^\lambda \exp\left(\frac{\lambda(1-\lambda)}{2}[M]_\infty \right). \end{aligned} \tag{4.12}$$

Now fix $\gamma \geq 0$. Applying Jensen's inequality in (4.12) with the concave function $x \mapsto x^\lambda$ as well as Hölder's inequality with the dual exponents $\frac{1}{\lambda}$ and $\frac{1}{1-\lambda}$, we obtain, with $F_\gamma = ([M]_\infty > \gamma)$, that

$$\begin{aligned} 1 &\leq E\mathcal{E}(M)_\infty^\lambda \exp\left(\frac{\lambda\gamma(1-\lambda)}{2} \right) + E\mathcal{E}(M)_\infty^\lambda \exp\left(\frac{\lambda(1-\lambda)}{2}[M]_\infty \right) 1_{F_\gamma} \\ &\leq (E\mathcal{E}(M)_\infty)^\lambda \exp\left(\frac{\lambda\gamma(1-\lambda)}{2} \right) + (E\mathcal{E}(M)_\infty 1_{F_\gamma})^\lambda \left(E \exp\left(\frac{\lambda}{2}[M]_\infty \right) \right)^{1-\lambda}. \end{aligned}$$

By our assumptions, we have that $\liminf_{\lambda \to 1} (E \exp((\lambda/2)[M]_\infty))^{1-\lambda}$ is finite. Let c denote the value of the limes inferior. By the above, we then obtain

$$1 \leq E\mathcal{E}(M)_\infty + cE\mathcal{E}(M)_\infty 1_{([M]_\infty > \gamma)}. \tag{4.13}$$

Letting γ tend to infinity, we obtain $1 \leq E\mathcal{E}(M)_\infty$, which by Lemma 4.2.2 shows that $\mathcal{E}(M)$ is a uniformly integrable martingale. □

For the remaining case of $0 \leq \alpha < 1$, we need the following further inequalities.

Lemma 4.2.5. *Let $x \geq 0$. It then holds that*

$$0 \leq (1+\lambda x) - (1+x)^\lambda \leq \frac{\lambda(1-\lambda)}{2} x^2 \quad \text{and} \tag{4.14}$$

$$0 \leq (1+x)^a - (1+ax) \leq \frac{a(a-1)}{2} x^2, \tag{4.15}$$

for $0 \leq \lambda \leq 1$ and $1 \leq a \leq 2$.

Proof. Fix $0 \leq \lambda \leq 1$. The lower inequality in (4.14) is equivalent to the statement that $(1+\lambda x)/(1+x)^\lambda \geq 1$, which follows from (4.5) of Lemma 4.2.4. Next, put $g_\lambda(x) = \frac{\lambda(1-\lambda)}{2} x^2 + (1+x)^\lambda - (1+\lambda x)$. In order to obtain the upper inequality, we need to prove $g_\lambda(x) \geq 0$. To this end, note that

$$g_\lambda'(x) = \lambda(1-\lambda)x + \lambda(1+x)^{\lambda-1} - \lambda \quad \text{and} \tag{4.16}$$

$$g_\lambda''(x) = \lambda(1-\lambda) - \lambda(1-\lambda)(1+x)^{\lambda-2}. \tag{4.17}$$

As $g''_\lambda(x) \geq 0$, $g'_\lambda(0) = 0$ and $g_\lambda(0) = 0$, we conclude that g_λ is nonnegative and thus (4.14) holds. Next, consider a with $1 \leq a \leq 2$. Using (4.6) of Lemma 4.2.4, we find that the lower inequality of (4.15) holds. For the upper inequality, define $h_a(x) = \frac{a(a-1)}{2}x^2 + 1 + ax - (1+x)^a$, we need to prove $h_a(x) \geq 0$. To do so, we note that

$$h'_a(x) = a(a-1)x + a - a(1+x)^{a-1} \quad \text{and} \tag{4.18}$$

$$h''_a(x) = a(a-1) - a(a-1)(1+x)^{a-2}, \tag{4.19}$$

such that $h''_a(x) \geq 0$, $h'_a(0) = 0$ and $h_a(0) = 0$, yielding as in the previous case that h_a is nonnegative and so we obtain (4.15). □

Lemma 4.2.6. *Let $x \geq 0$. It then holds that*

$$0 \leq \log \frac{1 + \lambda x + (1 + \sqrt{1-\alpha}x)^\lambda - (1 + \lambda\sqrt{1-\alpha}x)}{(1+x)^\lambda} \leq \alpha \frac{\lambda(1-\lambda)}{2}x^2 \tag{4.20}$$

for $\alpha, \lambda \in [0,1]$.

Proof. Let $\beta = \sqrt{1-\alpha}$, such that $\alpha = 1 - \beta^2$. We need to prove that for $x \geq 0$ and $\beta, \lambda \in [0,1]$, it holds that

$$0 \leq \log \frac{\lambda(1-\beta)x + (1+\beta x)^\lambda}{(1+x)^\lambda} \leq (1-\beta^2)\frac{\lambda(1-\lambda)}{2}x^2. \tag{4.21}$$

Consider the first inequality in (4.21). To prove this, it suffices to show that for $x \geq 0$ and $\beta, \lambda \in [0,1]$ it holds that

$$1 \leq \frac{\lambda(1-\beta)x + (1+\beta x)^\lambda}{(1+x)^\lambda}, \tag{4.22}$$

which is equivalent to $\lambda(1-\beta)x + (1+\beta x)^\lambda - (1+x)^\lambda \geq 0$. As this holds for all $x \geq 0$, $\lambda \in [0,1]$ and β equal to one, it suffices to prove that the derivative with respect to β of the left-hand side is nonpositive, meaning that we need to prove $\lambda x \geq x\lambda(1+\beta x)^{\lambda-1}$. However, this follows as $0 \leq (1+\beta x)^{\lambda-1} \leq 1$. Thus, the first inequality in (4.21) holds. Next, we consider the second inequality. We need to show that for $x \geq 0$ and $\lambda, \beta \in [0,1]$, it holds that

$$0 \leq (1-\beta^2)\frac{\lambda(1-\lambda)}{2}x^2 - \log \frac{\lambda(1-\beta)x + (1+\beta x)^\lambda}{(1+x)^\lambda}. \tag{4.23}$$

First note that the result holds when β is equal to one, $x \geq 0$ and $0 \leq \lambda \leq 1$. It therefore suffices to prove that the derivative with respect to β is nonpositive, meaning that we need to prove that for $x \geq 0$ and $\beta, \lambda \in [0,1]$,

$$0 \geq \frac{\lambda x - x\lambda(1+\beta x)^{\lambda-1}}{\lambda(1-\beta)x + (1+\beta x)^\lambda} - \beta\lambda(1-\lambda)x^2. \tag{4.24}$$

Multiplying by the denominator, which is positive, this is equivalent to

$$0 \leq \beta\lambda(1-\lambda)x^2(\lambda(1-\beta)x + (1+\beta x)^\lambda) - (\lambda x - x\lambda(1+\beta x)^{\lambda-1}). \tag{4.25}$$

which follows if we can show $1 \leq \beta(1-\lambda)x(\lambda(1-\beta)x + (1+\beta x)^\lambda) + (1+\beta x)^{\lambda-1}$. As $\beta(1-\beta)\lambda(1-\lambda)x^2 \geq 0$, it thus suffices to show that for $x \geq 0$ and $\lambda, \beta \in [0,1]$, we have $1 \leq \beta(1-\lambda)x(1+\beta x)^\lambda + (1+\beta x)^{\lambda-1}$. However, as this holds for any $\beta, \lambda \in [0,1]$ when x is zero, we find that it suffices to show that the derivative with respect to x is nonnegative, so that we need to show

$$0 \leq \beta(1-\lambda)((1+\beta x)^\lambda - x\beta\lambda(1+\beta x)^{\lambda-1}) + \beta(\lambda-1)(1+\beta x)^{\lambda-2} \tag{4.26}$$

for $x \geq 0$ and $\beta, \lambda \in [0,1]$. To this end, as $\beta(1-\lambda) \geq 0$, it suffices to show that $0 \leq (1+\beta x)^\lambda - x\beta\lambda(1+\beta x)^{\lambda-1} - (1+\beta x)^{\lambda-2}$ for $x \geq 0$ and $\beta, \lambda \in [0,1]$. To this end, simply note that

$$\begin{aligned} &(1+\beta x)^\lambda - x\beta\lambda(1+\beta x)^{\lambda-1} - (1+\beta x)^{\lambda-2} \\ =&(1+\beta x)^{\lambda-2}((1+\beta x)^2 - x\beta\lambda(1+\beta x) - 1) \\ =&(1+\beta x)^{\lambda-2}((1-\lambda)\beta^2x^2 + \beta(2-\lambda)x). \end{aligned} \tag{4.27}$$

As this is nonnegative, the result follows. □

The upper inequality in Lemma 4.2.6 is not obvious. However, an indication that the constant $\alpha\frac{\lambda(1-\lambda)}{2}$ is the right one may be obtained by a simple argument as follows. By the l'Hôpital rule, we have

$$\begin{aligned} &\lim_{x\to 0} \frac{1}{x^2} \log \frac{1+\lambda x + (1+\sqrt{1-\alpha}x)^\lambda - (1+\lambda\sqrt{1-\alpha}x)}{(1+x)^\lambda} \\ =&\lim_{x\to 0} \frac{1}{x^2} \log \frac{\lambda(1-\sqrt{1-\alpha})x + (1+\sqrt{1-\alpha}x)^\lambda}{(1+x)^\lambda} \\ =&\lim_{x\to 0} \frac{1}{2x} \left(\frac{\lambda(1-\sqrt{1-\alpha}) + \sqrt{1-\alpha}\lambda(1+\sqrt{1-\alpha}x)^{\lambda-1}}{\lambda(1-\sqrt{1-\alpha})x + (1+\sqrt{1-\alpha}x)^\lambda} - \frac{\lambda}{(1+x)} \right). \end{aligned} \tag{4.28}$$

Identifying a common divisor and applying the l'Hôpital rule again, we obtain that the above is equal to

$$\begin{aligned} &\frac{1}{2}\lim_{x\to 0}((1-\alpha)\lambda(\lambda-1)(1+\sqrt{1-\alpha}x)^{\lambda-2})(1+x) \\ +&\frac{1}{2}\lim_{x\to 0}(\lambda(1-\sqrt{1-\alpha}) + \sqrt{1-\alpha}\lambda(1+\sqrt{1-\alpha}x)^{\lambda-1}) \\ -&\frac{1}{2}\lim_{x\to 0}\lambda(\lambda(1-\sqrt{1-\alpha}) + \sqrt{1-\alpha}\lambda(1+\sqrt{1-\alpha}x)^{\lambda-1}), \end{aligned} \tag{4.29}$$

which by elementary calculations is equal to $\alpha\frac{\lambda(1-\lambda)}{2}$, the factor in front of x^2 in Lemma 4.2.6.

Proof of Theorem 4.2.1 for the case $0 \leq \alpha < 1$. We consider the case $0 < \alpha < 1$, the remaining case of $\alpha = 0$ follows by a similar method.

Fix $\varepsilon > 0$. We first prove that $\mathcal{E}(M)$ is a uniformly integrable martingale under the stronger condition that $\exp((1+\varepsilon)\frac{1}{2}(\alpha[M]_\infty + (1-\alpha)\langle M\rangle_\infty))$ is integrable. Assume given $a, r > 1$ with the property that $ar \leq 2$. Define a process U by putting $U_t = ar\sum_{0<s\leq t}\log(1+\Delta M_s) - \Delta M_s$. We have

$$\mathcal{E}(M)_t^a = \exp\left(arM_t - \frac{1}{2}[arM^c]_t + U_t\right)^{1/r} \exp\left(\frac{a(ar-1)}{2}[M^c]_t\right). \tag{4.30}$$

We wish to decompose the first factor in the right-hand side of (4.30) in two ways, one involving an optional increasing factor and one involving a predictable increasing factor. Put $N_t^o = arM_t$. For the optional decomposition, we note that

$$U_t = \left(\sum_{0<s\leq t}\log(1+\Delta N_s^o) - \Delta N_s^o\right) + \sum_{0<s\leq t}\log\frac{(1+\Delta M_s)^{ar}}{1+ar\Delta M_s}, \tag{4.31}$$

which yields

$$\exp\left(arM_t - \frac{1}{2}[arM^c]_t + U_t\right)^{\alpha/r} = \mathcal{E}(N^o)_t^{\alpha/r}\exp\left(\frac{\alpha}{r}\sum_{0<s\leq t}\log\frac{(1+\Delta M_s)^{ar}}{1+ar\Delta M_s}\right).$$

Next, for $1 \leq \beta \leq 2$, we define $W_t^\beta = \sum_{0<s\leq t}(1+\Delta M_s)^\beta - (1+\beta\Delta M_s)$. Note that the sum is well-defined, increasing and locally integrable by (4.15) of Lemma 4.2.5, as $[M]$ is locally integrable by our assumptions. Therefore, the compensator V^β of W^β is well-defined, and is increasing and locally integrable as well. Also note that $(1+\Delta M_s)^\beta = 1 + \beta\Delta M_s + \Delta W^\beta$. Further define two local martingales by putting $N_t^p = arM_t + W_t^{ar} - V_t^{ar}$ and $\bar{N}_t^p = \int_0^t (1+\Delta V_s^{ar})^{-1}\,\mathrm{d}N_s^p$, where $\bar{N}^p$ is well-defined as $\Delta V^{ar} \geq 0$ and $(1+\Delta V_s^{ar})^{-1}$ is predictable and locally bounded.

We begin by considering some properties of $\bar{N}^p$. First, we observe that

$$\begin{aligned}\Delta\bar{N}_t^p &= \frac{\Delta N_t^p}{1+\Delta V_t^{ar}} = \frac{ar\Delta M_t + \Delta W_t^{ar} - \Delta V_t^{ar}}{1+\Delta V_t^{ar}} \\ &= \frac{(1+\Delta M_t)^{ar}}{1+\Delta V_t^{ar}} - 1 > -1\end{aligned} \tag{4.32}$$

Furthermore, define $A_t^{ar} = \sum_{0<s\leq t}\Delta V_s^{ar}(1+\Delta V_s^{ar})^{-1}$. As ΔV^{ar} is predictable and nonnegative, the process A^{ar} is well-defined, and is also predictable, increasing and locally bounded, and $[A^{ar}, N^p]_t = \sum_{0<s\leq t}\Delta A_s^{ar}\Delta N_s^p$. By Proposition I.4.49 of [83], $[A^{ar}, N^p]$ is a local martingale. As the two local martingales $\int_0^t A_s^{ar}\,\mathrm{d}N_s^p$ and $[A^{ar}, N^p]$ are purely discontinuous and have the same jumps, they are equal by the uniqueness

part of Theorem 7.25 of [66], and we thus obtain

$$\begin{aligned}\bar{N}^p_t &= N^p_t - \int_0^t \frac{\Delta V^{ar}_s}{1+\Delta V^{ar}_s}\,\mathrm{d}N^p_s \\ &= arM_t + W^{ar}_t - V^{ar}_t - \sum_{0<s\le t}(1+\Delta V^{ar}_s)^{-1}\Delta V^{ar}_s \Delta N^p_s. \end{aligned} \tag{4.33}$$

Also, as the function $x \mapsto \log(1+x)-x$ is nonpositive for $x \ge 0$ and V^{ar} is increasing, we obtain $\log(1+\Delta V^{ar}) - \Delta V^{ar} \le 0$. Combining our observations, we get

$$\begin{aligned}&ar\log(1+\Delta M_s) - ar\Delta M_s - (\log(1+\Delta \bar{N}^p_s) - \Delta \bar{N}^p_s) \\ =&ar\log(1+\Delta M_s) - ar\Delta M_s - \left(\log\frac{(1+\Delta M_s)^{ar}}{1+\Delta V^{ar}_s} - \Delta \bar{N}^p_s\right) \\ =&\Delta W^{ar}_s - \frac{\Delta V^{ar}_s \Delta N^p_s}{1+\Delta V^{ar}_s} + \log(1+\Delta V^{ar}_s) - \Delta V^{ar}_s \le \Delta W^{ar}_s - \frac{\Delta V^{ar}_s \Delta N^p_s}{1+\Delta V^{ar}_s}, \end{aligned} \tag{4.34}$$

where the logarithm in first expression is well-defined by (4.32). This implies

$$\begin{aligned}U_t &\le \left(\sum_{0<s\le t}\log(1+\Delta\bar{N}^p_s) - \Delta\bar{N}^p_s\right) + \sum_{0<s\le t}\Delta W^{ar}_s - \frac{\Delta V^{ar}_s \Delta N^p_s}{1+\Delta V^{ar}_s} \\ &= \bar{N}^p_t - arM_t + \left(\sum_{0<s\le t}\log(1+\Delta\bar{N}^p_s) - \Delta\bar{N}^p_s\right) + V^{ar}_t. \end{aligned} \tag{4.35}$$

Also noting that $[(\bar{N}^p)^c]_t = [(N^p)^c]_t = [arM^c]_t$, we obtain the relationship

$$\exp\left(arM_t - \frac{1}{2}[arM^c]_t + U_t\right)^{(1-\alpha)/r} \le \mathcal{E}(\bar{N}^p)^{(1-\alpha)/r}_t \exp\left(\frac{1-\alpha}{r}V^{ar}_t\right).$$

Combining our results with (4.30), we obtain $\mathcal{E}(M)^a_t \le \mathcal{E}(N^o)^{\alpha/r}_t \mathcal{E}(N^p)^{(1-\alpha)/r} X_t$, where the process X is defined by

$$X_t = \exp\left(\frac{a(ar-1)}{2}[M^c]_t + \frac{\alpha}{r}\sum_{0<s\le t}\log\frac{(1+\Delta M_s)^{ar}}{1+ar\Delta M_s} + \frac{1-\alpha}{r}V_t\right). \tag{4.36}$$

Here, note that by (4.6) of Lemma 4.2.4 and (4.15) of Lemma 4.2.5, we have, as $1 \le ar \le 2$, that

$$\sum_{0<s\le t}\log\frac{(1+\Delta M_s)^{ar}}{1+ar\Delta M_s} \le \frac{ar(ar-1)}{2}[M^d]_t \quad \text{and} \tag{4.37}$$

$$V^{ar}_t \le \frac{ar(ar-1)}{2}\langle M^d\rangle_t, \tag{4.38}$$

leading to the inequality

$$\mathcal{E}(M)_t^a \leq \mathcal{E}(N^o)_t^{\alpha/r}\mathcal{E}(N^p)^{(1-\alpha)/r}\exp\left(\frac{a(ar-1)}{2}(\alpha[M]_t+(1-\alpha)\langle M\rangle_t\right). \quad (4.39)$$

Next, as $\Delta N_t^o \geq 0 > -1$ and $\Delta \bar{N}_t^p > -1$, $\mathcal{E}(N^o)$ and $\mathcal{E}(\bar{N}^p)$ are nonnegative supermartingales, and so for all bounded stopping times T, $0 \leq E\mathcal{E}(N^o)_T \leq 1$ and $0 \leq E\mathcal{E}(N^p)_T \leq 1$. Now let s be the dual exponent of r, such that $s = r/(r-1)$. Noting that $\frac{1}{r/\alpha} + \frac{1}{r/(1-\alpha)} + \frac{1}{s} = \frac{1}{r} + \frac{1}{s} = 1$,we may then apply Hölder's inequality for triples of functions to the inequality (4.39), yielding for any bounded stopping time T that

$$E\mathcal{E}(M)_T^a \leq \left(E\exp\left(\frac{y(y-1)}{2(r-1)}(\alpha[M]_\infty+(1-\alpha)\langle M\rangle_\infty\right)\right)^{1/s}, \quad (4.40)$$

where $y = ar$. This holds for all $a, r > 1$ such that $ar \leq 2$, and is a bound similar to (4.10). Proceeding as in the proof of the case $\alpha = 1$, we then obtain as a consequence of Lemma 4.2.3 that $\mathcal{E}(M)$ is a uniformly integrable martingale.

Next, assume that $\liminf_{\varepsilon\to 0}\varepsilon\log E\exp(((1-\varepsilon)/2)(\alpha[M]_\infty+(1-\alpha)\langle M\rangle_\infty)$ is finite. In particular, for $\varepsilon > 0$, $\exp(((1-\varepsilon)/2)\alpha[M]_\infty)$ is integrable, so $[M]_\infty$ is integrable. As a consequence, M is a square-integrable martingale, and the limit M_∞ exists.

Now fix $0 < \lambda < 1$ and define

$$W^\lambda(\alpha)_t = \sum_{0<s\leq t}(1+\sqrt{1-\alpha}\Delta M_s)^\lambda - (1+\lambda\sqrt{1-\alpha}\Delta M_s). \quad (4.41)$$

Note that by Lemma 4.2.5, the terms in the sum in (4.41) are nonpositive and bounded from below by $-(1-\alpha)\frac{1}{2}\lambda(1-\lambda)(\Delta M_s)^2$. In particular, we find that $W^\lambda(\alpha)$ is well-defined, decreasing and integrable. Letting $V^\lambda(\alpha)$ be the compensator of $W^\lambda(\alpha)$, $V^\lambda(\alpha)$ is then decreasing and integrable as well, and $W^\lambda(\alpha) - V^\lambda(\alpha)$ is a uniformly integrable martingale. We show that $V^\lambda(\alpha)$ is continuous. To this end, let T be some predictable stopping time. By Theorem VI.12.6 of [143] and its proof, we have $E\Delta M_T = 0$ and $E\Delta(W^\lambda(\alpha) - V^\lambda(\alpha))_T = 0$, so that

$$\begin{aligned} EV^\lambda(\alpha)_T &= EW^\lambda(\alpha)_T \\ &= E((1+\sqrt{1-\alpha}\Delta M_T)^\lambda - (1+\lambda\sqrt{1-\alpha}\Delta M_T)) \quad (4.42)\\ &= E(1+\sqrt{1-\alpha}\Delta M_T)^\lambda - 1 \geq 0, \quad (4.43)\end{aligned}$$

because of our assumption that $\Delta M \geq 0$. Thus, we know now that as V^λ is decreasing, $\Delta V_T^\lambda \leq 0$, and from the above, $EV^\lambda(\alpha)_T \geq 0$. We conclude that $\Delta V_T^\lambda = 0$ for all predictable stopping times. Lemma VI.19.2 of [143] then shows that $V^\lambda(\alpha)$ is continuous.

Let $L_t^\lambda = \lambda M_t + W^\lambda(\alpha) - V^\lambda(\alpha)$. By our previous observations, L^λ is a uniformly integrable martingale. In particular the limit L_∞^λ exists. Note that $(L^\lambda)^c = \lambda M^c$, so

it holds that $[(L^\lambda)^c]_t = \lambda^2[M^c]_t$. Also note that by continuity of $V^\lambda(\alpha)$, we have

$$\begin{aligned}\Delta L_t^\lambda &= \lambda\Delta M_t + (1+\sqrt{1-\alpha}\Delta M_t)^\lambda - (1+\lambda\sqrt{1-\alpha}\Delta M_t)\\ &= (\lambda - \lambda\sqrt{1-\alpha})\Delta M_t + (1+\sqrt{1-\alpha}\Delta M_t)^\lambda - 1\\ &\geq \lambda(1-\sqrt{1-\alpha})\Delta M_t \geq 0,\end{aligned} \tag{4.44}$$

and as $W^\lambda(\alpha)$ has nonpositive jumps, we also have $\Delta L_t^\lambda \leq \lambda\Delta M_t$. Combining these observations, we obtain $[L^\lambda]_t \leq \lambda^2[M]_t$, yielding that L^λ is square-integrable. We also obtain $\langle L^\lambda\rangle_t \leq \lambda^2\langle M\rangle_t$. This implies

$$\alpha[L^\lambda]_\infty + (1-\alpha)\langle L^\lambda\rangle_\infty \leq \lambda^2(\alpha[M]_\infty + (1-\alpha)\langle M\rangle_\infty), \tag{4.45}$$

so by what we already have shown, $\mathcal{E}(L^\lambda)$ is a uniformly integrable martingale. By elementary calculations, we obtain

$$\mathcal{E}(L^\lambda)_\infty = \mathcal{E}(M)_\infty^\lambda \exp\left(\frac{\lambda(1-\lambda)}{2}[M^c]_\infty + \sum_{0<t}\log\frac{1+\Delta L_t^\lambda}{(1+\Delta M_t)^\lambda} - V^\lambda(\alpha)_\infty\right).$$

By (4.14) of Lemma 4.2.5 and Lemma 4.2.6, we obtain the two inequalities

$$-V^\lambda(\alpha)_\infty \leq (1-\alpha)\frac{\lambda(1-\lambda)}{2}\langle M^d\rangle_\infty \tag{4.46}$$

$$\sum_{0<t}\log\frac{1+\Delta L_t^\lambda}{(1+\Delta M_t)^\lambda} \leq \alpha\frac{\lambda(1-\lambda)}{2}[M^d]_\infty, \tag{4.47}$$

so that combining our conclusions, we have

$$\begin{aligned}1 &= E\mathcal{E}(L^\lambda)_\infty\\ &\leq E\mathcal{E}(M)_\infty^\lambda \exp\left(\frac{\lambda(1-\lambda)}{2}([M^c]_\infty + \alpha[M^d]_\infty + (1-\alpha)\langle M^d\rangle_\infty)\right)\\ &= E\mathcal{E}(M)_\infty^\lambda \exp\left(\frac{\lambda(1-\lambda)}{2}(\alpha[M]_\infty + (1-\alpha)\langle M\rangle_\infty)\right),\end{aligned} \tag{4.48}$$

which is a bound similar to (4.12). Therefore, proceeding as in the proof of the case $\alpha = 1$, we obtain as a consequence of Lemma 4.2.2 that $\mathcal{E}(M)$ is a uniformly integrable martingale. □

We take a moment to reflect on the methods applied in the above proof, and make the following observations. First, while the proof of the case $0 \leq \alpha < 1$ is more complicated than the proof of the case $\alpha = 1$, both proofs follow very much the same plan: Use Hölder's inequality to argue that the result holds in a simple case where $\frac{1}{2}$ is exchanged with $(1+\varepsilon)\frac{1}{2}$ in the exponent, then use Hölder's inequality again to obtain the general proof. Also, note that the local martingale $\bar{N}^p$ used in the first part of the proof of the case $0 \leq \alpha < 1$ is related to general decompositions of exponential martingales, see Lemma II.1 of [117].

The comparatively simple structure of the proof is made possible by three main factors: The factor $\lambda(1-\lambda)$ present in the real analysis inequalities allows us to apply Hölder's inequality in the second parts of the proofs. Some of these inequalities have been noted earlier with a factor $1-\lambda$ instead of $\lambda(1-\lambda)$, compare for example (4.14) with (1.2) and (1.3) of [109], where the inequalities follow by a Taylor expansion argument. The more advanced triple-parameter inequality (4.20) allows us to obtain a criterion combining the quadratic variation and the predictable quadratic variation. Finally, the assumption $\Delta M \geq 0$, apart from making most of the real analysis inequalities applicable, also ensures that the compensator $V^{\lambda}(\alpha)$ in the second part of the proof of the case $0 \leq \alpha < 1$ is continuous.

5

An elementary proof that the first hitting time of an F_σ set by a jump process is a stopping time

Alexander Sokol

2010 Mathematics Subject Classification. Primary 60G44; Secondary 60G07.

Key words and phrases. Stopping time, Jump process, First hitting time.

Abstract. We give a short and elementary proof that the first hitting time of a F_σ set by the jump process of a càdlàg adapted process is a stopping time.

5.1 Introduction

For a stochastic process X and a subset B of the real numbers, the mapping T defined by $T = \inf\{t \geq 0 | X_t \in B\}$ is called the first hitting time of B by X. In [159], a short and elementary proof was given that the first hitting time of an open set by the jump process of a càdlàg adapted process is a stopping time. A similar result is proved by elementary means in [14], Proposition 1.3.14, where it is shown that the first hitting time of $[c, \infty)$ for $c > 0$ by the jump process of a càdlàg adapted process

is a stopping time. Using methods similar to both [14] and [159], we prove in this note that the hitting time of an F_σ set, meaning a countable union of closed sets, by the jump process of a càdlàg adapted process is a stopping time. As open sets are F_σ sets, this result covers both the case of hitting an open and a closed set.

5.2 Main result

We assume given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ such that the filtration $(\mathcal{F}_t)_{t\geq 0}$ is right-continuous in the sense that $\mathcal{F}_t = \cap_{s>t}\mathcal{F}_s$ for all $t \geq 0$. Also, we use the convention that $X_{0-} = X_0$, so that there is no jump at the timepoint zero.

Theorem 5.2.1. *Let X be a càdlàg adapted process, and let U be an F_σ set in $\mathbb{R}$. Define $T = \inf\{t \geq 0 | \Delta X_t \in U\}$. Then T is a stopping time.*

Proof. By the càdlàg property of X, ΔX is zero everywhere except for on a countable set. Therefore, T is identically zero if U contains zero, and so it is immediate that T is a stopping time in this case. We conclude that it suffices to prove the result in the case where U does not contain zero. Therefore, assume that U is an F_σ set not containing zero. By right-continuity of the filtration, it suffices to show $(T < t) \in \mathcal{F}_t$ for $t > 0$, see Theorem I.1 of [134]. Fix $t > 0$. Assume that $U = \cup_{n=1}^\infty F_n$, where F_n is closed. As $X_0 - X_{0-} = 0$ and U does not contain zero, we have

$$\begin{aligned}(T < t) &= (\exists\, s \in (0,t) : X_s - X_{s-} \in U) \\ &= \cup_{s\in(0,t)} \cup_{n=1}^\infty (X_s - X_{s-} \in F_n) \\ &= \cup_{n=1}^\infty \cup_{s\in(0,t)} (X_s - X_{s-} \in F_n) \\ &= \cup_{n=1}^\infty (\exists\, s \in (0,t) : X_s - X_{s-} \in F_n). \end{aligned} \tag{5.1}$$

Thus, it suffices to show that $(\exists\, s \in (0,t) : X_s - X_{s-} \in F) \in \mathcal{F}_t$ for all closed F. Assume given such a closed set F. We claim that

$$(\exists\, s \in (0,t) : X_s - X_{s-} \in F) = \cap_{n=1}^\infty \cup_{(p,q)\in\Theta_n} (X_q - X_p \in F_n)\,, \tag{5.2}$$

where $F_n = \{x \in \mathbb{R} \mid \exists\, y \in F : |x - y| \leq 1/n\}$ and Θ_n is the subset of $\mathbb{Q}^2$ defined by $\Theta_n = \{(p,q) \in \mathbb{Q}^2 | 0 < p < q < t, |p - q| \leq 1/n\}$.

To prove this, we first consider the inclusion towards the right. Assume that there is $0 < s < t$ such that $X_s - X_{s-} \in F$. Fix $n \geq 1$. By the path properties of X, we obtain that for $p, q \in \mathbb{R}$ with $0 < p < s < q < t$ and p and q close enough to s, $|X_q - X_s| \leq 1/2n$ and $|X_p - X_{s-}| \leq 1/2n$, yielding $|(X_q - X_p) - (X_s - X_{s-})| \leq 1/n$ and thus $X_q - X_p \in F_n$. By picking p and q in $\mathbb{Q}$ close enough to s, we obtain $(p,q) \in \Theta_n$ as well. This proves the inclusion towards the right.

Next, consider the inclusion towards the left. Assume that for all $n \geq 1$, there is $(p_n, q_n) \in \Theta_n$ such that $X_{q_n} - X_{p_n} \in F_n$. We then also have $\lim_n |p_n - q_n| = 0$. By

taking two consecutive subsequences and relabeling, we may assume that in addition to having $\lim_n |p_n - q_n| = 0$ and $0 < p_n < q_n < t$, both p_n and q_n are monotone. As (F_n) is decreasing, we then also obtain $X_{q_n} - X_{p_n} \in F_n$ for all $n \geq 1$. As p_n and q_n are bounded and monotone, they are convergent, and as $\lim_n |q_n - p_n| = 0$, it follows that the limit s is the same for both q_n and p_n.

We wish to argue that $0 < s < t$, that $X_{s-} = \lim_n X_{p_n}$ and that $X_s = \lim_n X_{q_n}$. First note that as both (p_n) and (q_n) are monotone, the limits $\lim_n X_{p_n}$ and $\lim_n X_{q_n}$ exist and are either equal to X_s or X_{s-}. As $X_{q_n} - X_{p_n} \in F_n$, we obtain

$$\lim_n X_{q_n} - \lim_n X_{p_n} = \lim_n X_{q_n} - X_{p_n} \in \cap_{n=1}^{\infty} F_n = F,$$

where the final equality follows as F is closed. As F does not contain zero, we conclude $\lim_n X_{q_n} - \lim_n X_{p_n} \neq 0$. From this, we immediately obtain $0 < s < t$, as if $s = 0$, we would obtain that both $\lim_n X_{q_n}$ and $\lim_n X_{p_n}$ were equal to X_s, and if $s = t$, both $\lim_n X_{q_n}$ and $\lim_n X_{p_n}$ would be equal to X_{s-}, in both cases yielding a contradiction. Also, we cannot have that both limits are X_s or that both limits are X_{s-}, and so only two cases are possible, namely that $X_s = \lim_n X_{q_n}$ and $X_{s-} = \lim_n X_{p_n}$ or that $X_s = \lim_n X_{p_n}$ and $X_{s-} = \lim_n X_{q_n}$. We wish to argue that the former holds. If $X_s = X_{s-}$, this is trivially the case. Assume that $X_s \neq X_{s-}$ and that $X_s = \lim_n X_{p_n}$ and $X_{s-} = \lim_n X_{q_n}$. If $q_n \geq s$ eventually or $p_n < s$ eventually, we obtain $X_s = X_{s-}$, a contradiction. Therefore, $q_n < s$ infinitely often and $p_n \geq s$ infinitely often. By monotonicity, $q_n < s$ and $p_n \geq s$ eventually, a contradiction with $p_n < q_n$. We conclude $X_s = \lim_n X_{q_n}$ and $X_{s-} = \lim_n X_{p_n}$, as desired.

From this, we conclude $X_s - X_{s-} = \lim_n X_{q_n} - \lim_n X_{p_n} \in F$. This proves the existence of $s \in (0, t)$ such that $X_s - X_{s-} \in F$, and so proves the inclusion towards the right.

We have now shown (5.2). Now, as X_s is $\mathcal{F}_t$ measurable for all $0 \leq s \leq t$, the set $\cap_{n=1}^{\infty} \cup_{(p,q)\in\Theta_n} (X_q - X_p \in F_n)$ is $\mathcal{F}_t$ measurable as well. We conclude that $(T < t) \in \mathcal{F}_t$ and so T is a stopping time. □

6

Proving existence results in martingale theory using a subsequence principle

Alexander Sokol

2010 Mathematics Subject Classification. Primary 60G07; Secondary 60G44; 60H05.

Key words and phrases. Martingale, Compensator, Quadratic variation, Stochastic integral.

Abstract. New proofs are given of the existence of the compensator (or dual predictable projection) of a locally integrable càdlàg adapted process of finite variation and of the existence of the quadratic variation process for a càdlàg local martingale. Both proofs apply a functional analytic subsequence principle. After presenting the proofs, we discuss their application in giving a simplified account of the construction of the stochastic integral of a locally bounded predictable process with respect to a semimartingale.

6.1 Introduction

Assume given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ satisfying the usual conditions, see [134], Section I.1, for the definition of this and other standard probabilistic

concepts. For a locally integrable càdlàg adapted process A with initial value zero and finite variation, the compensator, also known as the dual predictable projection, is the unique locally integrable càdlag predictable process $\Pi_p^* A$ with initial value zero and finite variation such that $A - \Pi_p^* A$ is a local martingale. For a càdlàg local martingale M with initial value zero, the quadratic variation process is the unique increasing càdlàg adapted process $[M]$ with initial value zero such that $M^2 - [M]$ is a local martingale and $\Delta[M] = (\Delta M)^2$. In both cases, uniqueness is up to indistinguishability.

For both the dual predictable projection and the quadratic variation, the proofs of the existence of these processes are among the most difficult in classical martingale theory, see for example [143], [66] or [134] for proofs. In this article, we give new proofs of the existence of these processes. The proofs are facilitated by the following lemma, first applied in this form to probability theory in [13]. We also give a short proof of the lemma.

Lemma 6.1.1. *Let (X_n) be sequence of variables bounded in $\mathcal{L}^2$. There exists a sequence (Y_n) such that each Y_n is a convex combination of a finite set of elements in $\{X_n, X_{n+1}, \ldots\}$ and (Y_n) is convergent in $\mathcal{L}^2$.*

Proof. Let α_n be the infimum of EZ^2, where Z ranges through all finite convex combinations of elements in $\{X_n, X_{n+1}, \ldots\}$, and define $\alpha = \sup_n \alpha_n$. If $Z = \sum_{k=n}^{K_n} \lambda_k X_k$ for some convex weights $\lambda_n, \ldots, \lambda_{K_n}$, we obtain $\sqrt{EZ^2} \leq \sup_n \sqrt{EX_n^2}$, in particular we have $\alpha_n \leq \sup_n EX_n^2$ and so $\alpha \leq \sup_n EX_n^2$ as well, proving that α is finite. For each n, there is a variable Y_n which is a finite convex combination of elements in $\{X_n, X_{n+1}, \ldots\}$ such that $E(Y_n)^2 \leq \alpha_n + \frac{1}{n}$. Let $m \geq n$, we then obtain

$$\begin{aligned} E(Y_n - Y_m)^2 &= 2EY_n^2 + 2EY_m^2 - E(Y_n + Y_m)^2 \\ &= 2EY_n^2 + 2EY_m^2 - 4E(\tfrac{1}{2}(Y_n + Y_m))^2 \\ &\leq 2(\alpha_n + \tfrac{1}{n}) + 2(\alpha_m + \tfrac{1}{m}) - 4\alpha_n \\ &= 2(\tfrac{1}{n} + \tfrac{1}{m}) + 2(\alpha_m - \alpha_n). \end{aligned} \tag{6.1}$$

As (α_n) is convergent, it is Cauchy. Therefore, the above shows that (Y_n) is Cauchy in $\mathcal{L}^2$, therefore convergent, proving the lemma. □

Lemma 6.1.1 may be seen as a combination of variants of the following two classical results: Every bounded sequence in a reflexive Banach space contains a weakly convergent subsequence (see Theorem 4.41-B of [165]), and every weakly convergent sequence in a reflexive Banach space has a sequence of convex combinations of its elements converging strongly to the weak limit (see Theorem 3.13 of [148]). In [13], an $\mathcal{L}^1$ version of Lemma 6.1.1 is used to give a simple proof of the Doob-Meyer theorem, building on the ideas of [84] and [136].

The remainder of the article is organized as follows. In Section 6.2, we give our proof of the existence of the compensator, and in Section 6.3, we give our proof of

the existence of the quadratic variation. In Section 6.4, we discuss how these results may be used to give a simplified account of the theory of stochastic integration with respect to semimartingales. In particular, the account proposed excludes the use of: The début theorem, the section theorems and the Doob-Meyer theorem. Section 6.5 contains auxiliary results which are needed in the main proofs.

6.2 The existence of the compensator

In this section, we will show that for any càdlàg adapted process A with initial value zero and paths of finite variation, locally integrable, there exists a càdlàg predictable process $\Pi_p^* A$ with initial value zero and paths of finite variation, locally integrable, unique up to indistinguishability, such that $A - \Pi_p^* A$ is a local martingale. We refer to $\Pi_p^* A$ as the compensator of A. The proofs will use some basic facts from the general theory of processes, some properties of monotone convergence for càdlàg increasing mappings, and Lemma 6.1.1. Essential for the results are the results on the limes superior of discrete approximations to the compensator, the proof of this is based on the technique developed in [84] and also applied in [13]. Note that as the existence of the compensator follows directly from the Doob-Meyer theorem, see for example Section I.3b of [83], the interest of the proofs given in this section is that if we restrict our attention to the compensator of a finite variation process instead of a submartingale, the uniform integrability arguments applied in [136] may be done away with, and furthermore we need only an $\mathcal{L}^2$ subsequence principle and not an $\mathcal{L}^1$ subsequence principle as in [13]. We begin by recalling some standard nomenclature and fixing our notation.

By $\mathcal{A}$, we denote the set of processes which are càdlàg adapted and increasing with initial value zero. For $A \in \mathcal{A}$, the limit A_∞ of A_t for t tending to infinity always exists in $[0, \infty]$. We say that A is integrable if A_∞ is integrable. The subset of integrable processes in $\mathcal{A}$ is denoted by $\mathcal{A}^i$. For $A \in \mathcal{A}$, we say that A is locally integrable if there exists a localising sequence (T_n) such that $A^{T_n} \in \mathcal{A}^i$. The set of such processes is denoted by $\mathcal{A}_\ell^i$. By $\mathcal{V}$, we denote the set of processes which are càdlàg adapted with initial value zero and has paths of finite variation. For $A \in \mathcal{V}$, V_A denotes the process such that $(V_A)_t$ is the variation of A over $[0, t]$. V_A is then an element of $\mathcal{A}$. For $A \in \mathcal{V}$, we say that A is integrable if V_A is integrable, and we say that A is locally integrable if V_A is locally integrable. The corresponding spaces of stochastic processes are denoted by $\mathcal{V}^i$ and $\mathcal{V}_\ell^i$, respectively. By $\mathbb{D}_+$, we denote the set of nonnegative dyadic rationals, $\mathbb{D}_+ = \{k2^{-n} | k \geq 0, n \geq 0\}$. The space of square-integrable martingales with initial value zero is denoted by $\mathcal{M}^2$. Also, we say that two processes X and Y are indistinguishable if their sample paths are almost surely equal, and in this case, we say that X is a modification of Y and vice versa. We say that a process X is càdlàg if it is right-continuous with left limits, and we say that a process X is càglàd if it is left-continuous with right limits.

Our main goal in this section is to show that for any $A \in \mathcal{V}_\ell^i$, there is a predictable

element $\Pi_p^* A$ of $\mathcal{V}_\ell^i$, unique up to indistinguishability, such that $A - \Pi_p^* A$ is a local martingale. To prove the result, we first establish the existence of the compensator for some simple elements of $\mathcal{V}_\ell^i$, namely processes of the type $\xi 1_{[\![T,\infty[\![}$, where T is a stopping time with $T > 0$, ξ is bounded, nonnegative and $\mathcal{F}_T$ measurable and $[\![T,\infty[\![= \{(t,\omega) \in \mathbb{R}_+ \times \Omega \mid T(\omega) \leq t\}$. After this, we apply monotone convergence arguments and localisation arguments to obtain the general existence result.

Lemma 6.2.1. *Let T be a stopping time with $T > 0$ and let ξ be nonnegative, bounded and $\mathcal{F}_T$ measurable. Define $A = \xi 1_{[\![T,\infty[\![}$. A is then an element of $\mathcal{A}^i$, and there exists a predictable process $\Pi_p^* A$ in $\mathcal{A}^i$ such that $A - \Pi_p^* A$ is a uniformly integrable martingale.*

Proof. Let $t_k^n = k2^{-n}$ for $k, n \geq 0$. We define $A_t^n = A_{t_k^n}$ for $t_k^n \leq t < t_{k+1}^n$, as well as

$$B_t^n = \sum_{i=1}^{k+1} E(A_{t_i^n} - A_{t_{i-1}^n} | \mathcal{F}_{t_{i-1}^n}) \text{ for } t_k^n < t \leq t_{k+1}^n, \tag{6.2}$$

and $B_0^n = 0$. Note that both A^n and B^n have initial value zero, since $T > 0$. Also note that A^n is càdlàg adapted and B^n is càglàd adapted. Put $M^n = A^n - B^n$. Note that M^n is adapted, but not necessarily càdlàg or càglàd. Also note that, with the convention that a sum over an empty index set is zero, it holds that

$$A_{t_k^n}^n = A_{t_k^n} \quad \text{and} \quad B_{t_k^n}^n = \sum_{i=1}^{k} E(A_{t_i^n} - A_{t_{i-1}^n} | \mathcal{F}_{t_{i-1}^n}) \tag{6.3}$$

for $k \geq 0$. Therefore, $(B_{t_k^n})_{k \geq 0}$ is the compensator of the discrete-time increasing process $(A_{t_k^n})_{k \geq 0}$, see Theorem II.54 of [142], so $(M_{t_k^n}^n)_{k \geq 0}$ is a discrete-time martingale with initial value zero. We next show that each element in this sequence of discrete-time martingales is bounded in $\mathcal{L}^2$, and the limit variables constitute a sequence bounded in $\mathcal{L}^2$ as well, this will allow us to apply Lemma 6.1.1. To this end, note that since B^n has initial value zero,

$$\begin{aligned}
(B_{t_k^n}^n)^2 &= 2(B_{t_k^n}^n)^2 - \sum_{i=0}^{k-1} (B_{t_{i+1}^n}^n)^2 - (B_{t_i^n}^n)^2 \\
&= \sum_{i=0}^{k-1} 2B_{t_k^n}^n (B_{t_{i+1}^n}^n - B_{t_i^n}^n) - (B_{t_{i+1}^n}^n)^2 + (B_{t_i^n}^n)^2 \\
&= \sum_{i=0}^{k-1} 2(B_{t_k^n}^n - B_{t_i^n}^n)(B_{t_{i+1}^n}^n - B_{t_i^n}^n) - (B_{t_{i+1}^n}^n - B_{t_i^n}^n)^2 \\
&\leq \sum_{i=0}^{k-1} 2(B_{t_k^n}^n - B_{t_i^n}^n)(B_{t_{i+1}^n}^n - B_{t_i^n}^n).
\end{aligned} \tag{6.4}$$

Now let c be a bound for ξ. Applying that $B^n_{t^n_{i+1}}$ is $\mathcal{F}_{t^n_i}$ measurable, the martingale property of $(M^n_{t^n_k})_{k\geq 0}$ and the fact that A and B are increasing and A is bounded by c, we find

$$\begin{aligned} E(B^n_{t^n_k} - B^n_{t^n_i})(B^n_{t^n_{i+1}} - B^n_{t^n_i}) &= E(B^n_{t^n_{i+1}} - B^n_{t^n_i})E(B^n_{t^n_k} - B^n_{t^n_i}|\mathcal{F}_{t^n_i}) \\ &= E(B^n_{t^n_{i+1}} - B^n_{t^n_i})E(A^n_{t^n_k} - A^n_{t^n_i}|\mathcal{F}_{t^n_i}) \\ &\leq cE(B^n_{t^n_{i+1}} - B^n_{t^n_i}). \end{aligned} \tag{6.5}$$

All in all, $E(B^n_{t^n_k})^2 \leq 2c\sum_{i=0}^{k-1} E(B^n_{t^n_{i+1}} - B^n_{t^n_i}) = 2cEB^n_{t^n_k} = 2cEA^n_{t^n_k} \leq 2c^2$. Thus $E(M^n_{t^n_k})^2 \leq 4E(A^n_{t^n_k})^2 + 4E(B^n_{t^n_k})^2 \leq 12c^2$. We conclude that $(M^n_{t^n_k})_{k\geq 0}$ is bounded in $\mathcal{L}^2$, and so convergent almost surely and in $\mathcal{L}^2$ to a limit M^n_∞, and the sequence $(M^n_\infty)_{n\geq 0}$ is bounded in $\mathcal{L}^2$ as well.

By Lemma 6.1.1, there exists a sequence of naturals (K_n) with $K_n \geq n$ and for each n a finite sequence of reals $\lambda^n_n, \ldots, \lambda^n_{K_n}$ in the unit interval summing to one, such that $\sum_{i=n}^{K_n} \lambda^n_i M^i_\infty$ is convergent in $\mathcal{L}^2$ to some variable M_∞. Let M be a càdlàg version of the process $t \mapsto E(M_\infty|\mathcal{F}_t)$. Define $B = A - M$, we wish to argue that there is a modification of B satisfying the requirements of the lemma.

To do so, first note that as $(M^k_{t^n_k})_{k\geq 0}$ is a martingale, Doob's inequality yields

$$\lim_{n\to\infty} E\sup_{k\geq 0}\left(M_{t^n_k} - \sum_{i=n}^{K_n} \lambda^n_i M^i_{t^n_k}\right)^2 = 0, \tag{6.6}$$

and by picking a subsequence and relabeling, we may assume that the convergence is almost sure as well. In particular, $\sum_{i=n}^{K_n} \lambda^n_i M^i_q$ converges almost surely to M_q for all $q \in \mathbb{D}_+$. Now put $C^n = \sum_{i=n}^{K_n} \lambda^n_i B^i$. Note that C^n is càdlàg, adapted and increasing, and

$$\begin{aligned} \lim_{t\to\infty} C^n_t = \lim_{m\to\infty} C^n_m &= \lim_{m\to\infty} \sum_{i=n}^{K_n} \lambda^n_i B^i_m \\ &= \lim_{m\to\infty} A_m - \sum_{i=n}^{K_n} \lambda^n_i M^i_m = A_\infty - \sum_{i=n}^{K_n} \lambda^n_i M^i_\infty, \end{aligned} \tag{6.7}$$

showing that $C^n \in \mathcal{A}^i$ and that $(C^n_\infty)_{n\geq 0}$ is bounded in $\mathcal{L}^2$. Also note that for each $q \in \mathbb{D}_+$, it holds that $A_q = \lim_{n\to\infty} A^n_q$ almost surely. Therefore,

$$B_q = A_q - M_q = \lim_{n\to\infty} A^n_q - \sum_{i=n}^{K_n} \lambda^n_i M^i_q = \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda^n_i B^i_q = \lim_{n\to\infty} C^n_q, \tag{6.8}$$

almost surely. From this, we obtain that B is almost surely increasing on $\mathbb{D}_+$. As B is càdlàg, this shows that B is almost surely increasing on all of $\mathbb{R}_+$. Next, we

show that $B_t = \limsup_{n\to\infty} C^n_t$ almost surely, simultaneuously for all $t \geq 0$, this will allow us to show that B has a predictable modification. To this end, note that for $t \geq 0$ and $q \in \mathbb{D}_+$ with $q \geq t$, $\limsup_{n\to\infty} C^n_t \leq \limsup_{n\to\infty} C^n_q = B_q$. As B is càdlàg, this yields $\limsup_{n\to\infty} C^n_t \leq B_t$. This holds almost surely for all $t \in \mathbb{R}_+$ simultaneously. Similarly, $\liminf_{n\to\infty} C^n_t \geq B_{t-}$ almost surely, simultaneously for all $t \geq 0$. All in all, we conclude that almost surely, $B_t = \limsup_{n\to\infty} C^n_t$ for all continuity points t of B, simultaneously for all $t \geq 0$. As the jumps of B can be exhausted by a countable sequence of stopping times, we find that in order to show the desired result on the limes superior, it suffices to show for any stopping time S that $B_S = \limsup_{n\to\infty} C^n_S$.

Fixing a stopping time S, we first note that as $0 \leq C^n_S \leq C^n_\infty$, the sequence of variables $(C^n_S)_{n\geq 0}$ is bounded in $\mathcal{L}^2$ and thus in particular uniformly integrable. Therefore, by Lemma 6.5.1, $\limsup_{n\to\infty} EC^n_S \leq E\limsup_{n\to\infty} C^n_S \leq EB_S$. As $\limsup_{n\to\infty} C^n_S \leq B_S$ almost surely, we find that to show $\limsup_{n\to\infty} C^n_S = B_S$ almost surely, it suffices to show that EC^n_S converges to EB_S, and to this end, it suffices to show that EB^n_S converges to EB_S. Now define S_n by putting $S_n = \infty$ whenever $S = \infty$ and $S_n = t^n_k$ whenever $t^n_{k-1} < S \leq t^n_k$. (S_n) is then a sequence of stopping times taking values in $\mathbb{D}_+$ and infinity and converging downwards to S, and it holds that

$$B^n_S = \sum_{k=0}^{\infty} B^n_{t^n_{k+1}} 1_{(t^n_k < S \leq t^n_{k+1})} = \sum_{k=0}^{\infty} B^n_{t^n_{k+1}} 1_{(S_n = t^n_{k+1})} = B^n_{S_n}. \tag{6.9}$$

As $(M^n_{t^n_k})_{k\geq 0}$ is a uniformly integrable discrete-time martingale, the optional sampling theorem yields $EB^n_{S_n} = EA^n_{S_n}$, and similarly, $EB_S = EA_S$. As A is càdlàg and bounded and $A^n_{S_n} = A_{S_n}$, the dominated convergence theorem allows us to obtain

$$\begin{aligned} \lim_{n\to\infty} EB^n_S &= \lim_{n\to\infty} EB^n_{S_n} = \lim_{n\to\infty} EA^n_{S_n} \\ &= \lim_{n\to\infty} EA_{S_n} = \lim_{n\to\infty} EA_S = EB_S. \end{aligned} \tag{6.10}$$

Recalling our earlier observations, we may now conclude that $\limsup_{n\to\infty} C^n_t = B_t$ almost surely for all points of discontinuity of B, and so all in all, the result holds almost surely for all $t \in \mathbb{R}_+$ simultaneously.

We now apply this to show that B has a predictable modification. Let F be the almost sure set where $B = \limsup_{n\to\infty} C^n$. Theorem 3.33 of [66] then shows that $1_F C^n$ is a predictable càdlàg process, and $1_F B = \limsup_{n\to\infty} 1_F C^n$. Therefore, $1_F B$ is a predictable càdlàg process, and $1_F B$ is almost surely increasing as well. Now let $\Pi^*_p A$ be a modification of B such that $\Pi^*_p A$ is in $\mathcal{A}^i$. Again using Theorem 3.33 of [66], $\Pi^*_p A$ is predictable since B is predictable, and as $A - \Pi^*_p A$ is a modification of the uniformly integrable martingale $A - B$, we conclude that $\Pi^*_p A$ satisfies all the requirements to be the compensator of A. □

With Lemma 6.2.1 in hand, the remainder of the proof for the existence of the compensator merely consists of monotone convergence arguments.

Lemma 6.2.2. *Let A^n be a sequence of processes in $\mathcal{A}^i$ such that $\sum_{n=1}^{\infty} A^n$ converges pointwise to a process A. Assume for each $n \geq 1$ that B^n is a predictable element of $\mathcal{A}^i$ such that $A^n - B^n$ is a uniformly integrable martingale. A is then in $\mathcal{A}^i$, and $\sum_{n=1}^{\infty} B^n$ almost surely converges pointwise to a predictable process $\Pi_p^* A$ in $\mathcal{A}^i$ such that $A - \Pi_p^* A$ is a uniformly integrable martingale.*

Proof. Clearly, A is in $\mathcal{A}^i$. With $B = \sum_{n=0}^{\infty} B^n$, B is a well-defined process with values in $[0, \infty]$, since each B^n is nonnegative. We wish to argue that there is a modification of B which is the compensator of A. First note that as each B^n is increasing and nonnegative, so is B. Also, as $A^n - B^n$ is a uniformly integrable martingale, the optional sampling theorem and two applications of the monotone convergence theorem yields for any bounded stopping time T that

$$EB_T = \lim_{n\to\infty} \sum_{k=1}^{n} EB_T^k = \lim_{n\to\infty} \sum_{k=1}^{n} EA_T^k = EA_T, \tag{6.11}$$

which in particular shows that B almost surely takes finite values. Therefore, by Lemma 6.5.2, we obtain that B is almost surely nonnegative, càdlàg and increasing. Also, by another two applications of the monotone convergence theorem, we obtain for any stopping time T that $EB_T = \lim_{t\to\infty} EB_{T\wedge t} = \lim_{t\to\infty} EA_{T\wedge t} = EA_T$. This holds in particular with $T = \infty$, and therefore, the limit of B_t as t tends to infinity is almost surely finite and is furthermore integrable. Lemma 6.5.2 then also shows that $\sum_{k=1}^{n} B^k$ converges almost surely uniformly to B on $\mathbb{R}_+$.

We now let $\Pi_p^* A$ be a nonnegative càdlàg increasing adapted modification of B. Then $\Pi_p^* A$ is in $\mathcal{A}^i$, and $E(\Pi_p^* A)_T = EA_T$ for all stopping times T, so by Theorem 77.6 of [143], $A - \Pi_p^* A$ is a uniformly integrable martingale. Also, $\sum_{k=1}^{n} B^k$ almost surely converges uniformly to $\Pi_p^* A$ on $\mathbb{R}_+$. In order to complete the proof, it remains to show that $\Pi_p^* A$ is predictable. To this end, note that by uniform convergence, Lemma 6.5.3 shows that for any stopping time T, $\Delta(\Pi_p^* A)_T = \lim_n \sum_{k=1}^{n} \Delta B_T^k$. As B^k is predictable, we find by Theorem 3.33 of [66] that if T is totally inacessible, $\Delta(\Pi_p^* A)_T$ is zero almost surely, and if T is predictable, $\Delta(\Pi_p^* A)_T$ is $\mathcal{F}_{T-}$ measurable. Therefore, Theorem 3.33 of [66] shows that $\Pi_p^* A$ is predictable. □

Theorem 6.2.3. *Let $A \in \mathcal{V}_\ell^i$. There exists a predictable process $\Pi_p^* A$ in $\mathcal{V}_\ell^i$, unique up to indistinguishability, such that $A - \Pi_p^* A$ is a local martingale.*

Proof. We first consider uniqueness. If $A \in \mathcal{V}_\ell^i$ and B and C are two predictable processes in $\mathcal{V}_\ell^i$ such that $A - B$ and $A - C$ both are local martingales, we find that $B - C$ is a predictable local martingale with paths of finite variation. By Theorem 6.3 of [66], uniqueness follows.

As for existence, Lemma 6.2.1 establishes existence for the case where $A = \xi 1_{[\![T,\infty[\![}$ where ξ is nonnegative, bounded and $\mathcal{F}_T$ measurable. Using Lemma 6.2.2, this extends to the case where $\xi \in \mathcal{L}^1(\mathcal{F}_T)$. For general $A \in \mathcal{A}^i$, there exists by Theorem 3.32 of [66] a sequence of stopping times (T_n) covering the jumps of A. Put

$A^d = \sum_{n=1}^{\infty} \Delta A_{T_n} 1_{[\![T_n,\infty[\![}$. As $A \in \mathcal{A}^i$, A^d is a well-defined element of $\mathcal{A}^i$, and $A - A^d$ is a continuous element of $\mathcal{A}^i$. As we have existence for each $\Delta A_{T_n} 1_{[\![T_n,\infty[\![}$, Lemma 6.2.2 allows us to obtain existence for A. Existence for $A \in \mathcal{V}^i$ is then obtained by decomposing $A = A^+ - A^-$, where $A^+, A^- \in \mathcal{A}^i$, and extends to $A \in \mathcal{V}^i_\ell$ by a localisation argument. □

From the characterisation in Theorem 6.2.3, the usual properties of the compensator such as linearity, positivity, idempotency and commutation with stopping, can then be shown.

6.3 The existence of the quadratic variation

In this section, we will prove the existence of the quadratic variation process for a local martingale by a reduction to the cases of bounded martingales and martingales of integrable variation, applying Lemma 6.1.1 to obtain existence for bounded martingales. Apart from Lemma 6.1.1, the proofs will also use the fundamental theorem of local martingales as well as some properties of martingales with finite variation. Our method of proof is direct and is simpler than the methods employed in for example [87] or [83], where the quadratic covariation is defined through the integration-by-parts formula and requires the construction and properties of the stochastic integral for semimartingales.

Lemma 6.3.1. *Let M be a bounded martingale with initial value zero. There exists a process $[M]$ in $\mathcal{A}^i$, unique up to indistinguishability, such that $M^2 - [M] \in \mathcal{M}^2$ and $\Delta[M] = (\Delta M)^2$. We call $[M]$ the quadratic variation process of M.*

Proof. We first consider uniqueness. Assume that A and B are two processes in $\mathcal{A}^i$ such that $M^2 - A$ and $M^2 - B$ are in $\mathcal{M}^2$ and $\Delta A = \Delta B = (\Delta M)^2$. In particular, the process $A - B$ is a continuous element of $\mathcal{M}^2$ and has paths of finite variation, so Theorem 6.3 of [66] shows that $A - B$ is almost surely zero, such that A and B are indistinguishable. This proves uniqueness. Next, we consider the existence of the process. Let $t^n_k = k2^{-n}$ for $n, k \geq 0$, we then find

$$\begin{aligned} M_t^2 &= \sum_{k=1}^{\infty} M^2_{t\wedge t^n_k} - M^2_{t\wedge t^n_{k-1}} \\ &= 2\sum_{k=1}^{\infty} M^t_{t^n_{k-1}}(M^t_{t^n_k} - M^t_{t^n_{k-1}}) + \sum_{k=1}^{\infty}(M^t_{t^n_k} - M^t_{t^n_{k-1}})^2, \end{aligned} \tag{6.12}$$

where the terms in the sum are zero from a point onwards, namely for such k that $t^n_{k-1} \geq t$. Define $N^n_t = 2\sum_{k=1}^{\infty} M^t_{t^n_{k-1}}(M^t_{t^n_k} - M^t_{t^n_{k-1}})$. Our plan for the proof is to show that (N^n) is a bounded sequence in $\mathcal{M}^2$. This will allow us to apply Lemma 6.1.1 in order to obtain some $N \in \mathcal{M}^2$ which is the limit of appropriate convex combinations

of the (N^n). We then show that by putting $[M]$ equal to a modification of $M^2 - N$, we obtain a process with the desired qualities.

We first show that N^n is a martingale by applying Theorem II.77.6 of [143]. Clearly, N^n is càdlàg and adapted with initial value zero, and so it suffices to prove that N_T^n is integrable and that $EN_T^n = 0$ for all bounded stopping times T. To this end, note that as M is bounded, there is $c > 0$ such that $|M_t| \leq c$ for all $t \geq 0$. Then N_T^n is clearly integrable, as it is the sum of finitely many terms each bounded by $4c^2$, and we have

$$\begin{aligned} EN_T^n &= E\sum_{k=1}^{\infty} M_{T\wedge t_{k-1}^n}(M_{T\wedge t_k^n} - M_{T\wedge t_{k-1}^n}) \qquad (6.13)\\ &= \sum_{k=1}^{\infty} EM_{t_{k-1}^n}^T(M_{t_k^n}^T - M_{t_{k-1}^n}^T) = \sum_{k=1}^{\infty} EM_{t_{k-1}^n}^T E(M_{t_k^n}^T - M_{t_{k-1}^n}^T|\mathcal{F}_{t_{k-1}^n}), \end{aligned}$$

where the interchange of summation and expectation is allowed, as the only nonzero terms in the sum are for those k such that $t_{k-1}^n \leq t$, and there are only finitely many such terms. As M^T is a martingale, $E(M_{t_k^n}^T - M_{t_{k-1}^n}^T|\mathcal{F}_{t_{k-1}^n}) = 0$ by optional sampling, so the above is zero and N^n is a martingale by Theorem II.77.6 of [143]. Next, we show that (N^n) is bounded in $\mathcal{L}^2$. Fix $k \geq 1$, we first consider a bound for the second moment of $N_{t_k^n}^n$. To obtain this, note that for $i < j$,

$$\begin{aligned} &E(M_{t_{i-1}^n}(M_{t_i^n} - M_{t_{i-1}^n}))(M_{t_{j-1}^n}(M_{t_j^n} - M_{t_{j-1}^n}))\\ &= E(M_{t_{i-1}^n}(M_{t_i^n} - M_{t_{i-1}^n})E(M_{t_{j-1}^n}(M_{t_j^n} - M_{t_{j-1}^n})|\mathcal{F}_{t_i^n})\\ &= E(M_{t_{i-1}^n}(M_{t_i^n} - M_{t_{i-1}^n})M_{t_{j-1}^n}E(M_{t_j^n} - M_{t_{j-1}^n}|\mathcal{F}_{t_i^n}), \qquad (6.14) \end{aligned}$$

which is zero, as $E(M_{t_j^n} - M_{t_{j-1}^n}|\mathcal{F}_{t_i^n}) = 0$, and by the same type of argument, we obtain $E(M_{t_i^n} - M_{t_{i-1}^n})(M_{t_j^n} - M_{t_{j-1}^n}) = 0$. In other words, the variables are pairwisely orthogonal, and so

$$\begin{aligned} E(N_{t_k^n}^n)^2 &= E\left(\sum_{i=1}^{k} M_{t_{i-1}^n}(M_{t_i^n} - M_{t_{i-1}^n})\right)^2 = \sum_{i=1}^{k} E\left(M_{t_{i-1}^n}(M_{t_i^n} - M_{t_{i-1}^n})\right)^2\\ &\leq c^2\sum_{i=1}^{k} E(M_{t_i^n} - M_{t_{i-1}^n})^2 = c^2E\left(\sum_{i=1}^{k} M_{t_i^n} - M_{t_{i-1}^n}\right)^2 = c^2EM_{t_k^n}^2, \qquad (6.15) \end{aligned}$$

which yields $\sup_{t\geq 0} E(N_t^n)^2 = \sup_{k\geq 1} E(N_{t_k^n}^n)^2 \leq \sup_{k\geq 1} c^2EM_{t_k^n}^2 \leq 4c^2EM_\infty^2$, and this is finite. Thus, $N^n \in \mathcal{M}^2$, and $E(N_\infty^n)^2 = \lim_t E(N_t^n)^2 \leq 4c^2EM_\infty^2$, so $(N_\infty^n)_{n\geq 1}$ is bounded in $\mathcal{L}^2$.

Now, by Lemma 6.1.1, there exists a sequence of naturals (K_n) with $K_n \geq n$ and for each n a finite sequence of reals $\lambda_n^n, \ldots, \lambda_{K_n}^n$ in the unit interval summing to one, such that $\sum_{i=n}^{K_n} \lambda_i^n N_\infty^i$ is convergent in $\mathcal{L}^2$ to some variable N_∞. It then holds that

there is $N \in \mathcal{M}^2$ such that $E\sup_{t\geq 0}(N_t - \sum_{i=n}^{K_n} \lambda_i^n N_t^i)^2$ tends to zero. By picking a subsequence and relabeling, we may assume without loss of generality that we also have almost sure convergence. Define $A = M^2 - N$, we claim that there is a modification of A satisfying the criteria of the theorem.

To prove this, first note that as M^2 and N are càdlàg and adapted, so is A. We want to show that A is almost surely increasing and that $\Delta A = (\Delta M)^2$ almost surely. We first consider the jumps of A. To prove that $\Delta A = (\Delta M)^2$ almost surely, it suffices to show that $\Delta A_T = (\Delta M_T)^2$ almost surely for any bounded stopping time T. Let T be any bounded stopping time. As it holds that $\sup_{t\geq 0}(N_t - \sum_{i=n}^{K_n} \lambda_i^n N_t^i)^2$ converges to zero almost surely, we find

$$\begin{aligned} A_T = M_T^2 - N_T &= \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n (M_T^2 - N_T^i) \\ &= \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n \sum_{k=1}^{\infty} (M_T^{t_k^i} - M_T^{t_{k-1}^i})^2, \end{aligned} \tag{6.16}$$

almost surely. In particular, we obtain

$$\Delta A_T = \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n \left(\sum_{k=1}^{\infty} (M_T^{t_k^i} - M_T^{t_{k-1}^i})^2 - (M_{T-}^{t_k^i} - M_{T-}^{t_{k-1}^i})^2 \right). \tag{6.17}$$

Fix $i, k \geq 0$. Note that

$$\begin{aligned} (M_t^{t_k^i} - M_t^{t_{k-1}^i})^2 - (M_{t-}^{t_k^i} - M_{t-}^{t_{k-1}^i})^2 = 0 \qquad & \text{when } t \leq t_{k-1}^i \text{ or } t > t_k^i \\ (M_t^{t_k^i} - M_t^{t_{k-1}^i})^2 = (M_t - M_{t_{k-1}^i})^2 \qquad & \text{when } t_{k-1}^i < t \leq t_k^i \\ (M_{t-}^{t_k^i} - M_{t-}^{t_{k-1}^i})^2 = (M_{t-} - M_{t_{k-1}^i})^2 \qquad & \text{when } t_{k-1}^i < t \leq t_k^i. \end{aligned}$$

From these observations, we conclude that with $s(t,i)$ denoting the unique t_{k-1}^i such that $t_{k-1}^i < t \leq t_k^i$, we have

$$\Delta A_T = \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n ((M_T - M_{s(T,i)})^2 - (M_{T-} - M_{s(T,i)})^2). \tag{6.18}$$

Here, it holds that

$$\begin{aligned} &(M_T - M_{s(T,i)})^2 - (M_{T-} - M_{s(T,i)})^2 \\ =&M_T^2 - 2M_T M_{s(T,i)} + M_{s(T,i)}^2 - (M_{T-}^2 - 2M_{T-}M_{s(T,i)} + M_{s(T,i)}^2) \\ =&M_T^2 - M_{T-}^2 - 2\Delta M_T M_{s(T,i)} \\ =&(M_T - M_{T-})(M_T + M_{T-}) - 2\Delta M_T M_{s(T,i)} \\ =&(\Delta M_T)^2 + 2\Delta M_T (M_{T-} - M_{s(T,i)}), \end{aligned} \tag{6.19}$$

yielding $\Delta A_T = (\Delta M_T)^2 + 2\Delta M_T \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n (M_{T-} - M_{s(T,i)})$. Now, we always have $s(T,i) < T$ and $|s(T,i) - T| \leq 2^{-i}$. Therefore, given $\varepsilon > 0$, there is $n \geq 1$ such that for all $i \geq n$, $|M_{T-} - M_{s(T,i)}| \leq \varepsilon$. As the $(\lambda_i^n)_{n\leq i\leq K_n}$ are convex weights, we obtain for n this large that $|\sum_{i=n}^{K_n} \lambda_i^n (M_{T-} - M_{s(T,i)})| \leq \varepsilon$. This allows us to conclude that $\sum_{i=n}^{K_n} \lambda_i^n (M_{T-} - M_{s(T,i)})$ converges almost surely to zero. Combining this with our previous conclusions, we obtain $\Delta A_T = (\Delta M_T)^2$ almost surely. Since this holds for any arbitrary stopping time, we now obtain $\Delta A = (\Delta M)^2$ up to indistinguishability.

Next, we show that A is almost surely increasing. Consider elements $p, q \in \mathbb{D}_+$ with $p \leq q$, we will show that $A_p \leq A_q$ almost surely. There exists $j \geq 1$ and naturals $n_p \leq n_q$ such that $p = n_p 2^{-j}$ and $q = n_q 2^{-j}$. By our previous results, we have $A_p = \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n \sum_{k=1}^{\infty} (M_{p\wedge t_k^i} - M_{p\wedge t_{k-1}^i})^2$, and analogously for A_q. For $i \geq j$, $p \wedge t_k^i = n_p 2^{-j} \wedge k2^{-i} = (n_p 2^{i-j} \wedge k)2^{-i}$, and analogously for $q \wedge t_k^i$. Therefore, we obtain that almost surely,

$$\begin{aligned}
\lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n \sum_{k=1}^{\infty} (M_{p\wedge t_k^i} - M_{p\wedge t_{k-1}^i})^2 &= \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n \sum_{k=1}^{n_p 2^{i-j}} (M_{t_k^i} - M_{t_{k-1}^i})^2 \\
&\leq \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^n \sum_{k=1}^{n_q 2^{i-j}} (M_{t_k^i} - M_{t_{k-1}^i})^2 \\
&= \lim_{n\to\infty} \sum_{i=n}^{K_n} \lambda_i^m \sum_{k=1}^{\infty} (M_{q\wedge t_k^i} - M_{q\wedge t_{k-1}^i})^2, \quad (6.20)
\end{aligned}$$

allowing us to make the same calculations in reverse and conclude $A_p \leq A_q$ almost surely. As $\mathbb{D}_+$ is countable, we conclude that A is inceasing on $\mathbb{D}_+$ almost surely, and as A is càdlàg, we conclude that A is increasing almost surely. Furthermore, as we have that $A_\infty = M_\infty^2 - N_\infty$ and both M_∞^2 and N_∞ are integrable, we conclude that A_∞ is integrable.

Finally, let F be the almost sure set where A is increasing. Put $[M] = A1_F$. As all null sets are in $\mathcal{F}_t$ for $t \geq 0$, $[M]$ is adapted as A is adapted. Furthermore, $[M]$ is càdlàg, increasing and $[M]_\infty$ exists and is integrable. As $M^2 - [M] = N$ up to indistinguishability, we now have constructed a process $[M]$ which is in $\mathcal{A}^i$ such that $M^2 - [M]$ is in $\mathcal{M}^2$ and $\Delta[M] = (\Delta M)^2$ up to indistinguishability. This concludes the proof. □

Theorem 6.3.2. *Let M be a local martingale with initial value zero. There exists $[M] \in \mathcal{A}$ with the properties that $M^2 - [M]$ is a local martingale with initial value zero and $\Delta[M] = (\Delta M)^2$.*

Proof. We first consider the case where $M = M^b + M^i$, where M^b and M^i both are local martingales with initial value zero, M^b is bounded and M^i is of integrable variation. In this case, $\sum_{0<s\leq t} (\Delta M_t^i)^2$ is absolutely convergent for any $t \geq 0$, and

we may therefore define a process A^i in $\mathcal{A}$ by putting $A^i_t = \sum_{0<s\leq t}(\Delta M^i_t)^2$. As M^b is bounded, $\sum_{0<s\leq t} \Delta M^b_t \Delta M^i_t$ is almost surely absolutely convergent as well, and so we may define a process A^x in $\mathcal{V}$ by putting $A^x_t = \sum_{0<s\leq t} \Delta M^b_t \Delta M^i_t$. Finally, by Theorem 6.3.1, there exists a process $[M^b]$ in $\mathcal{A}^i$ such that $(M^b)^2 - [M^b]$ is in $\mathcal{M}^2$ and $\Delta[M^b] = (\Delta M^b)^2$. We put $A_t = [M^b]_t + 2A^x + A^i$ and claim that there is a modification of A satisfying the criteria in the theorem.

To this end, first note that A clearly is càdlàg adapted of finite variation, and for $0 \leq s \leq t$, we have $[M^b]_t \geq [M^b]_s + \sum_{s<u\leq t}(\Delta M^b_u)^2$ almost surely, so that we obtain $A_t - A_s \geq \sum_{s<u\leq t}(\Delta M^b_u + \Delta M^i_u)^2$ almost surely, showing that A is almost surely increasing. To show that $M^2 - A$ is a local martingale, note that

$$M^2 - A = (M^b)^2 - [M^b] + 2(M^b M^i - A^x) + (M^i)^2 - A^i. \tag{6.21}$$

Here, $(M^b)^2 - [M^b]$ is in $\mathcal{M}^2$ by Lemma 6.3.1, in particular a local martingale. By the integration-by-parts formula, we have $(M^i)^2_t - A^i_t = 2\int_0^t M^i_{s-}\,\mathrm{d}M^i_s$, where the integral is well-defined as M_{s-} is bounded on compacts. Using Theorem 6.5 of [66], the integral process $\int_0^t M^i_{s-}\,\mathrm{d}M^i_s$ is a local martingale, and so $(M^i)^2 - A^i$ is a local martingale. Therefore, in order to obtain that $M^2 - A$ is a local martingale, it suffices to show that $M^b M^i - A^x$ is a local martingale. By Theorem 5.32 of [66], $M^b_t M^i_t - \int_0^t M^b_s\,\mathrm{d}M^i_s$ is a local martingale, so it suffices to show that $\int_0^t M^b_s\,\mathrm{d}M^i_s - A^x_t$ is a local martingale. As ΔM^b is bounded, it is integrable, and so we have

$$\int_0^t M^b_s\,\mathrm{d}M^i_s = \int_0^t \Delta M^b_s\,\mathrm{d}M^i_s + \int_0^t M^b_{s-}\,\mathrm{d}M^i_s = A^x_t + \int_0^t M^b_{s-}\,\mathrm{d}M^i_s. \tag{6.22}$$

As $\int_0^t M^b_{s-}\,\mathrm{d}M^i_s$ is a local martingale, again by Theorem 6.5 of [66], we finally conclude that $M^b M^i - A^x$ is a local martingale. Thus, $M^2 - A$ is a local martingale. This proves existence in the case where $M = M^b + M^i$, where M^b is bounded and M^i has integrable variation.

Finally, we consider the case of a general local martingale M with initial value zero. By Theorem III.29 of [134], $M = M^b + M^i$, where M^b is locally bounded and M^i has paths of finite variation. With (T_n) a localising sequence for both M^b and M^i, our previous results then show the existence of a process $A^n \in \mathcal{A}$ such that $(M^{T_n})^2 - A^n$ is a local martingale and $\Delta A^n = (\Delta M^{T_n})^2$. By uniqueness, we may define $[M]$ by putting $[M]_t = A^n_t$ for $t \leq T_n$. We then obtain that $[M] \in \mathcal{A}$, $M^2 - [M]$ is a local martingale and $\Delta[M] = (\Delta M)^2$, and the proof is complete. □

6.4 Discussion

The results given in Sections 6.2 and 6.3 yield comparatively simple proofs of the existence of the compensator and the quadratic variation, two technical concepts essential to martingale theory in general and stochastic calculus in particular. We

will now discuss how these proofs may be used to give a simplified account of the development of the basic results of stochastic integration theory. Specifically, the question we ask is the following: How can one, starting from basic continuous-time martingale theory, construct the stochastic integral of a locally bounded predictable process with respect to a semimartingale, as simply as possible?

Since the publication of one of the first complete accounts of the general theory of stochastic integration in [36], several others have followed, notably [66, 143, 87, 83, 134], each contributing with simplified and improved proofs. The accounts in [66] and [143] make use of the predictable projection to prove the Doob-Meyer theorem, and to obtain the uniqueness of this projection, they apply the difficult section theorems. In [87] and [134], this dependence is removed, using the methods of, among others, [136] and [11], respectively. In general, however, the methods in [87] and [134] are not entirely comparable, as [87] follows the traditional path of starting with continuous-time martingale theory, developing some general theory of processes, and finally constructing the stochastic integral for semimartingales, while [134] begins by defining a semimartingale as a "good integrator" in a suitable sense, and develops the theory from there, in the end proving through the Bichteler-Dellacherie theorem that the two methods are equivalent. The developement of the stochastic integral we will suggest below follows in the tradition also seen in [87].

We suggest the following path to the construction of the stochastic integral:

1. Development of the predictable σ-algebra and predictable stopping times, in particular the equivalence between, in the language of [143], being "previsible" and being "announceable".

2. Development of the main results on predictable processes, in particular the characterization of predictable càdlàg processes as having jumps only at predictable times, and having the jump at a predictable time T being measurable with respect to the σ-algebra $\mathcal{F}_{T-}$.

3. Proof of the existence of the compensator, leading to the fundamental theorem of local martingales, meaning the decomposition of any local martingale into a locally bounded and a locally integrable variation part. Development of the quadratic variation process using these results.

4. Construction of the stochastic integral using the fundamental theorem of local martingales and the quadratic variation process.

The proofs given in Sections 6.2 and 6.3 help make this comparatively short path possible. We now comment on each of the points above, and afterwards compare the path outlined with other accounts of the theory.

As regards point 1, the equivalence between a stopping time being previsible (having a predictable graph) and being announceable (having an announcing sequence) is

proved in [143] as part of the PFA theorem, which includes the introduction of $\mathcal{F}_{T-}$. However, the equivalence between P (previsibility) and A (accessibility) may be done without any reference to $\mathcal{F}_{T-}$, and this makes for a pleasant separation of concerns.

The main result in point 2, the characterization of predictable càdlàg functions, can be found for example as Theorem 3.33 of [66]. The existence of the compensator in point 3 may now be obtained as in Section 6.2, and the fundamental theorem of local martingales may then be proven as in the proof of Theorem III.29 of [134]. After this, the existence of the quadratic variation may be obtained as in Section 6.3. Note that the traditional method for obtaining the quadratic variation is either as the remainder term in the integration-by-parts formula (as in [83]), or through a localisation to $\mathcal{M}^2$, applying the Doob-Meyer theorem. Our method removes the need for the application of the Doob-Meyer theorem.

Finally, in point 4, these results may be combined to obtain the existence of the stochastic integral of a locally bounded predictable process with respect to a semimartingale using the fundamental theorem of local martingales and a modification of the methods given in Chapter IX of [66].

As for comparisons of the approach outlined above with other approaches, for example [87], the main benefit of the above approach would be that the development of the compensator is obtained in a very simple manner, in particular not necessitating a decomposition into predictable and totally inaccessible parts, and without any reference to "naturality". Note, however, that the expulsion of "naturality" from the proof of the Doob-Meyer theorem in [136] already was obtained in [84] and [13]. In any case, focusing attention on the compensator instead of a general supermartingale decomposition simplifies matters considerably. Furthermore, developing the quadratic variation directly using the fundamental theorem of local martingales allows for a very direct construction of the stochastic integral, while the method given in [87] first develops a preliminary integral for local martingales which are locally in $\mathcal{M}^2$.

6.5 Auxiliary results

Lemma 6.5.1. *Let (X_n) be a uniformly integrable sequence of variables. It then holds that*

$$\limsup_{n\to\infty} EX_n \leq E \limsup_{n\to\infty} X_n. \tag{6.23}$$

Proof. Since (X_n) is uniformly integrable, it holds that $\lim_{\lambda\to\infty} \sup_n EX_n 1_{(X_n>\lambda)}$ is zero. Let $\varepsilon > 0$ be given, we may then pick λ so large that $EX_n 1_{(X_n>\lambda)} \leq \varepsilon$ for all n. Now, the sequence $(\lambda - X_n 1_{(X_n\leq\lambda)})_{n\geq 1}$ is nonnegative, and Fatou's lemma therefore

yields

$$\begin{aligned}\lambda - E\limsup_{n\to\infty} X_n 1_{(X_n\le\lambda)} &= E\liminf_{n\to\infty}(\lambda - X_n 1_{(X_n\le\lambda)}) \\ &\le \liminf_{n\to\infty} E(\lambda - X_n 1_{(X_n\le\lambda)}) \\ &= \lambda - \limsup_{n\to\infty} EX_n 1_{(X_n\le\lambda)}. \end{aligned} \tag{6.24}$$

The terms involving the limes superior may be infinite and are therefore a priori not amenable to arbitrary arithmetic manipulation. However, by subtracting λ and multiplying by minus one, we yet find

$$\limsup_{n\to\infty} EX_n 1_{(X_n\le\lambda)} \le E\limsup_{n\to\infty} X_n 1_{(X_n\le\lambda)}. \tag{6.25}$$

As we have ensured that $EX_n 1_{(X_n>\lambda)} \le \varepsilon$ for all n, this yields

$$\limsup_{n\to\infty} EX_n \le \varepsilon + E\limsup_{n\to\infty} X_n 1_{(X_n\le\lambda)} \le \varepsilon + E\limsup_{n\to\infty} X_n, \tag{6.26}$$

and as $\varepsilon > 0$ was arbitrary, the result follows. □

Lemma 6.5.2. *Let (f_n) be a sequence of nonnegative increasing càdlàg mappings from $\mathbb{R}_+$ to $\mathbb{R}$. Assume that $\sum_{n=1}^{\infty} f_n$ converges pointwise to some mapping f from $\mathbb{R}_+ \to \mathbb{R}$. Then, the convergence is uniform on compacts, and f is a nonnegative increasing càdlàg mapping. If $f(t)$ has a limit as t tends to infinity, the convergence is uniform on $\mathbb{R}_+$.*

Proof. Fix $t \ge 0$. For $m \ge n$, we have

$$\sup_{0\le s\le t}\left|\sum_{k=1}^{m} f_k(s) - \sum_{k=1}^{n} f_k(s)\right| = \sup_{0\le s\le t}\sum_{k=n+1}^{m} f_k(s) = \sum_{k=n+1}^{m} f_k(t), \tag{6.27}$$

which tends to zero as m and n tend to infinity. Therefore, $(\sum_{k=1}^{n} f_k)$ is uniformly Cauchy on $[0,t]$, and so has a càdlàg limit on $[0,t]$. As this limit must agree with the pointwise limit, we conclude that $\sum_{k=1}^{n} f_k$ converges uniformly on compacts to f, and therefore f is nonnegative, increasing and càdlàg.

It remains to consider the case where $f(t)$ has a limit $f(\infty)$ as t tends to infinity. In this case, we find that $\lim_t f_n(t) \le \lim_t f(t) = f(\infty)$, so $f_n(t)$ has a limit $f_n(\infty)$ as t tends to infinity as well. Fixing $n \ge 1$, we have

$$\sum_{k=1}^{n} f_k(\infty) = \sum_{k=1}^{n}\lim_{t\to\infty} f_k(t) = \lim_{t\to\infty}\sum_{k=1}^{n} f_k(t) \le \lim_{t\to\infty} f(t) = f(\infty). \tag{6.28}$$

Therefore, $(f_k(\infty))$ is absolutely summable. As we have

$$\sup_{t\ge 0}\left|\sum_{k=1}^{m} f_k(t) - \sum_{k=1}^{n} f_k(t)\right| = \sup_{t\ge 0}\sum_{k=n+1}^{m} f_k(t) = \sum_{k=n+1}^{m} f_k(\infty), \tag{6.29}$$

we find that $(\sum_{k=1}^{n} f_k)$ is uniformly Cauchy on $\mathbb{R}_+$, and therefore uniformly convergent. As the limit must agree with the pointwise limit, we conclude that f_n converges uniformly to f on $\mathbb{R}_+$. This concludes the proof. □

Lemma 6.5.3. *Let (f_n) be a sequence of bounded càdlàg mappings from $\mathbb{R}_+$ to $\mathbb{R}$. If (f_n) is Cauchy in the uniform norm, there is a bounded càdlàg mapping f from $\mathbb{R}_+$ to $\mathbb{R}$ such that $\sup_{t\geq 0} |f_n(t) - f(t)|$ tends to zero. In this case, it holds that $\sup_{t\geq 0} |f_n(t-) - f(t-)|$ and $\sup_{t\geq 0} |\Delta f_n(t) - \Delta f(t)|$ tends to zero as well.*

Proof. Assume that (f_n) is Cauchy in the uniform norm. As the space of bounded functions from $[0,\infty)$ to $\mathbb{R}$ is complete under the uniform norm, f_n converges uniformly to f. We show that f is càdlàg. Let $t \geq 0$, we will show that f is right-continuous at t. Take $\varepsilon > 0$ and take n so that $\sup_{t\geq 0} |f(t) - f_n(t)| \leq \varepsilon$. Let $\delta > 0$ be such that $|f_n(t) - f_n(s)| \leq \varepsilon$ for $s \in [t, t+\delta]$, then

$$|f(t) - f(s)| \leq |f(t) - f_n(t)| + |f_n(t) - f_n(s)| + |f_n(s) - f_n(t)| \leq 3\varepsilon \tag{6.30}$$

for such s. Therefore, f is right-continuous at t. Now let $t > 0$, we claim that f has a left limit at t. First note that for n and m large enough, it holds for any $t > 0$ that $|f_n(t-) - f_m(t-)| \leq \sup_{t\geq 0} |f_n(t) - f_m(t)|$. Therefore, the sequence $(f_n(t-))_{n\geq 1}$ is Cauchy, and so convergent to some limit $\xi(t)$. Now let $\varepsilon > 0$ and take n so that $\sup_{t\geq 0} |f(t) - f_n(t)| \leq \varepsilon$ and $|f_n(t-) - \xi(t)| \leq \varepsilon$. Let $\delta > 0$ be such that $t - \delta \geq 0$ and such that whenever $s \in [t-\delta, t)$, $|f_n(s) - f_n(t-)| \leq \varepsilon$. Then

$$|f(s) - \xi(t)| \leq |f(s) - f_n(s)| + |f_n(s) - f_n(t-)| + |f_n(t-) - \xi(t)| \leq 3\varepsilon \tag{6.31}$$

for any such s. Therefore, f has a left limit at t. This shows that f is càdlàg.

Finally, we have for any $t > 0$ and any sequence (s_n) converging strictly upwards to t that $|f(t-) - f_n(t-)| = \lim_m |f(s_m) - f_n(s_m)| \leq \sup_{t\geq 0} |f(t) - f_n(t)|$, so we conclude that $\sup_{t\geq 0} |f(t-) - f_n(t-)|$ converges to zero as well. As a consequence, we also obtain that $\sup_{t\geq 0} |\Delta f(t) - \Delta f_n(t)|$ converges to zero. □

7

Causal interpretation of stochastic differential equations

Alexander Sokol and Niels Richard Hansen

2010 Mathematics Subject Classification. Primary 60H10; Secondary 62A01.

Key words and phrases. Stochastic differential equation, Causality, Structural equation model, Identifiability, Lévy process, Weak conditional local independence.

Abstract. We give a causal interpretation of stochastic differential equations (SDEs) by defining the postintervention SDE resulting from an intervention in an SDE. We show that under Lipschitz conditions, the solution to the postintervention SDE is equal to a uniform limit in probability of postintervention structural equation models based on the Euler scheme of the original SDE, thus relating our definition to mainstream causal concepts. We prove that when the driving noise in the SDE is a Lévy process, the postintervention distribution is identifiable from the semigroup of the SDE. Also for the case of Lévy driving noise, we relate our results to the notion of weak conditional local independence (WCLI) by proving that if a coordinate X^i is locally unaffected by an intervention in another coordinate X^j, then X^i is WCLI of X^j.

7.1 Introduction

The notion of causality has long been of interest to both statisticians and scientists working in fields applying statistics. In general, causal models are models containing families of possible distributions of the variables observed as well as appropriate mathematical descriptions of causal structures in the data. Thus, claiming that a causal model is true amounts to claiming more than statements about the distribution of the variables observed. Causal modeling has several goals, prominent among them are:

1. Estimation of intervention effects from partially observed systems with a given causal structure.
2. Identification of the causal structure from observational data.

One of the most developed theories of causal inference is the approach based on directed acyclic graphs (DAGs) and finitely many variables with no explicit time component, descibed in [161] and [126]. In recent years, there have been efforts to develop similar notions of causality for stochastic processes, both in discrete time and in continuous time. For discrete time results, see for example [46, 47, 48, 49]. As discrete time models often are defined through explicit functional relationships between variables, as in for example autoregressive processes, such models fit directly into the DAG-based framework. As for continuous time, early discussions of causality can be found in [59, 54, 30]. One of the most recent frameworks for causality in continuous-time is based on the concept of weak conditional local independence. For results related to this, see [37, 27, 57, 144, 145]. An alternative notion of causality defined solely through filtrations is developed in [131, 130]. In Section 4.1 of [1] it is noted that both ordinary differential equations and stochastic differential equations (SDEs) allow for a natural interpretation in terms of "influence", and that interventions may be defined by substitutions in the differential equations. In this paper, we make these ideas precise. Our main contributions are:

1. For a given SDE, we give a precise definition of the postintervention SDE resulting from an intervention.
2. We show that under certain regularity assumptions, the solution of the postintervention SDE is the limit of a sequence of interventions in structural equation models based on the Euler scheme of the observational SDE.
3. We prove that for SDEs with a Lévy process as the driving semimartingale, the postintervention distribution is identifiable from the semigroup of the SDE.
4. We relate our results to weak conditional local independence (WCLI) by showing that for SDEs with a Lévy process as the driving semimartingale, X^i is WCLI of X^j if X^i is locally unaffected by an intervention in X^j.

Of particular note is that the identifiability result (3) in the list above corresponds to a case where the error variables are not all independent, as is otherwise often assumed to be the case when calculating intervention effects in the DAG-based framework. For the DAG-based framework, in the case of independent errors, parts of the causal structure may be learned from the observational distribution, as seen in [171], and intervention distributions may be calculated by a truncated factorization formula as in (3.10) of [126]. For dependent errors, such results are harder to come by. In our case, we take advantage of the Markov nature of the solutions to SDEs with Lévy noise in order to obtain our identifiability result for SDE models, and we are also able to obtain explicit descriptions of the resulting postintervention distributions. Also note that in many cases, the semigroup of the SDE is identifiable from the observational distribution. Our identifiability implies that in such cases, the postintervention distributions are identifiable from the observational distribution.

In matters of causality, it is important to distinguish clearly between definitions, theorems and interpretations. Our definition of postintervention SDEs will be a purely mathematical construct. It will, however, have a natural causal interpretation. Given an SDE model, in order to use the definition of intervention given here to predict the effects of real-world interventions, it is necessary that the SDE can be sensibly interpreted as a data-generating mechanism with certain properties: Specifically, as we will argue in Section 7.4, it is essentially sufficient that the driving semimartingales are autonomous in the sense that they may be assumed not to be directly affected by interventions. This is an assumption which is not testable from a statistical viewpoint. It is, nonetheless, an assumption which may be justified by other means in concrete cases.

The remainder of the paper is organized as follows. In Section 7.2, we motivate and introduce our notion of intervention for SDEs. In Section 7.3, we review the terminology of causal inference as developed in [126] and [161], based on structural equation models and directed acyclic graphs. Section 7.4 shows that under certain conditions, our notion of intervention is equivalent to taking a limit of interventions in the context of structural equation models based on the Euler scheme of the SDE. In Section 7.5, we give conditions for postintervention distributions to be identifiable from the semigroup of the SDE. Section 7.6 relates our work to weak conditional local independence. Finally, in Section 7.7, we discuss our results. Sections 7.8 and 7.9 contain proofs.

7.2 Interventions for stochastic differential equations

Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ satisfying the usual conditions, see [134] for the definition of this and other notions related to continuous-time stochastic processes. In this section, given an SDE, we define the notion of a postintervention SDE. This notion yields a causal interpretation of stochastic differential equations.

In general, the precise meaning of "causation" is a point of contemporary debate, see for example [33]. For our purposes, it suffices to take a practical standpoint: The causal structure of a system is sufficiently elucidated if we know the effects of making interventions in the system. To motivate our definition, we begin by investigating a simple example.

Example 7.2.1. Chemical kinetics is concerned with the dynamic evolution of the concentrations of chemicals given in terms of a number of coupled chemical reactions, see [173]. This example considers two chemicals and we derive a simple system of SDEs from the fundamental mechanisms of the chemical reactions. If the concentration of one chemical is fixed – as an alternative to letting it evolve according to the chemical reactions – the fundamental mechanisms allow us to obtain an SDE for the concentrations of the remaining chemicals. This equation then describes the effects of an intervention, and can be obtained from the original system by a purely mechanical deletion and substitution process.

The chemicals are denoted x and y and the corresponding concentrations are denoted X and Y, respectively. We assume that four reactions are possible, namely:

$$\begin{aligned} \emptyset &\xrightarrow{a} y \\ y &\xrightarrow{b_{12}} x \\ x &\xrightarrow{b_{11}} \emptyset \\ y &\xrightarrow{b_{22}} \emptyset \end{aligned}$$

Here, the first reaction denotes the creation or influx of chemical y with constant rate a, the second reaction denotes the change of y into x at rate $b_{12}Y$, and the third and fourth reactions denote degradation or outflux of x and y with rates $b_{11}X$ and $b_{22}Y$, respectively. We collect the rates into the vector

$$\lambda(X,Y) = \begin{bmatrix} a \\ b_{12}Y \\ b_{11}X \\ b_{22}Y \end{bmatrix}. \tag{7.1}$$

The so-called stoichiometric matrix

$$S = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{bmatrix} \tag{7.2}$$

collects the information about the number of molecules, for each of the two chemicals (rows), which are created or destroyed by each of the four reactions (columns). The rates $\lambda(X,Y)$ and the stoichiometric matrix S form the fundamental parameters of the system. We are interested in using $\lambda(X,Y)$ and S to construct a dynamical model for X and Y.

Several different stochastic and deterministic models are available. One stochastic model is obtained by considering a Markov jump process on $\mathbb{N}_0^2$, where each coordinate denotes the total number of molecules of each chemical x and y, and the

transition rates are given in terms of S and $\lambda(X, Y)$. A system of SDEs approximating the Markov jump process, see [6], is given by

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} X_0 \\ Y_0 + at \end{bmatrix} + \int_0^t B \begin{bmatrix} X_s \\ Y_s \end{bmatrix} \, \mathrm{d}s + \int_0^t \Sigma(X_s, Y_s) \, \mathrm{d}W_s \tag{7.3}$$

where W_s denotes a four-dimensional Wiener process,

$$\begin{aligned} \Sigma(X, Y) &= S \mathrm{diag} \sqrt{\lambda(X, Y)} \\ &= \begin{bmatrix} 0 & \sqrt{b_{12} Y} & -\sqrt{b_{11} X} & 0 \\ \sqrt{a} & -\sqrt{b_{12} Y} & 0 & -\sqrt{b_{22} Y} \end{bmatrix} \end{aligned} \tag{7.4}$$

and

$$B = \begin{bmatrix} -b_{11} & b_{12} \\ -b_{12} & -b_{22} \end{bmatrix}. \tag{7.5}$$

If we are able to fix the concentration Y_t at a level ζ, we effectively remove the first and last of the reactions and the second will have the constant rate $b_{12}\zeta$. By arguments as above we then derive the SDE

$$X_t = X_0 + t b_{12} \zeta - \int b_{11} X_s \, \mathrm{d}s + \int_0^t \sigma(X_s) \, \mathrm{d}\widetilde{W}_s \tag{7.6}$$

with $\widetilde{W}_s$ a two-dimensional Wiener process and $\sigma(x) = (\sqrt{b_{12}\zeta}, -\sqrt{b_{11} x})$. This SDE describes the dynamics of the system after the intervention. We observe that this SDE can be obtained from (7.3) by deleting the equation for Y_t and substituting Y_t with ζ in the remaining equation.

It should be noted that due to the square root in the diffusion coefficient, the SDEs in this example do not satisfy the usual Lipschitz conditions. To avoid technical issues we may cap all entries in $\lambda(X, Y)$ at a lower level c and an upper level C with $0 < c \leq C < \infty$. The resulting SDE will then have bounded Lipschitz coefficients. ∘

Example 7.2.1 illustrates how a model for the intervention in a system can be obtained from a model for the entire system. In this particular example, the resulting model for the intervention can be justified by reference to the fundamental mechanisms – the chemical reactions – that drive the system, and interventions result in SDEs modified by substitution and deletion. While noting that this correspondence between interventions and substitution and deletion in the original equations may not always be justified, we will use this principle as a general, purely probabilistic definition of interventions in SDEs. Note also that in the example above, the matrix

$$\Sigma(X, Y)\Sigma(X, Y)^t = \begin{bmatrix} b_{12} Y + b_{11} X & -b_{12} Y \\ -b_{12} Y & a + b_{12} Y + b_{22} Y \end{bmatrix} \tag{7.7}$$

is not diagonal, implying that the martingale parts of the semimartingale (X, Y) are not orthogonal. This shows that there are naturally occuring situations where it is necessary to consider models with non-orthogonal martingale parts – a situation excluded in the WCLI framework of [57] and motivating our definition.

In order to formalize our definition in a general framework, let Z be a d-dimensional semimartingale and assume that $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ is a continuous mapping, where $\mathbb{M}(p, d)$ denotes the space of real $p \times d$ matrices. We consider the stochastic differential equation

$$X_t^i = X_0^i + \sum_{j=1}^{d} \int_0^t a_{ij}(X_{s-}) \, \mathrm{d}Z_s^j, \qquad i \leq p, \tag{7.8}$$

which in matrix notation may be described more succintly as

$$\mathrm{d}X_t = a(X_{t-}) \, \mathrm{d}Z_t. \tag{7.9}$$

Definition 7.2.2. Consider some $m \leq p$ and $\zeta \in \mathbb{R}$. The stochastic differential equation arising from (7.8) under the intervention $X^m := \zeta$ is

$$Y_t^i = Y_0^i + \sum_{j=1}^{d} \int_0^t b_{ij}(Y_{s-}) \, \mathrm{d}Z_s^j, \qquad i \leq p, \tag{7.10}$$

where $Y_0^i = X_0^i$ for $i \neq m$ and $Y_0^m = \zeta$, and $b : \mathbb{R}^p \to \mathbb{M}(p, d)$ is given by letting $b_{ij}(x) = a_{ij}(x)$ for $i \neq m$ and $b_{mj}(x) = 0$ for all $x \in \mathbb{R}^p$ and $j \leq d$.

By Definition 7.2.2, intervening takes an SDE as its argument and yields another SDE. Note that existence and uniqueness of solutions are not required for Definition 7.2.2 to make sense, although we will mainly take interest in cases where both (7.8) and (7.10) have unique solutions. By Theorem V.7 of [134], this is the case whenever the mapping a is Lipschitz. Note also that the solution Y to (7.10) will satisfy that $Y_t^m = \zeta$ for all $t \geq 0$, while the process U consisting of the coordinates of Y excluding the m'th will satisfy the $(p-1)$-dimensional SDE

$$U_t^i = X_0^i + \sum_{j=1}^{d} \int_0^t c_{ij}(U_{s-}) \, \mathrm{d}Z_s^j, \qquad i \neq m, \tag{7.11}$$

where $c : \mathbb{R}^{p-1} \to \mathbb{M}(p-1, d)$ is defined by $c_{ij}(x) = a_{ij}(x_1, \ldots, \zeta, \ldots, x_p)$ for $i \neq m$ and $j \leq d$, and the ζ is on the m'th coordinate. This is the same type of postintervention structure as we obtained in Example 7.2.1 by reference to fundamental mechanisms.

Assume that (7.8) and (7.10) have unique solutions for all interventions. We refer to (7.8) as the observational SDE, to the solution of (7.8) as the observational process, and to the distribution of the solution of (7.8) as the observational distribution. We refer to (7.10) as the postintervention SDE, to the solution of (7.10) as the postintervention process and to the distribution of the solution to (7.10) as the postintervention distribution.

7.3 Terminology of SEMs, DAGs and interventions

In this section, we review the basic notions related to intervention calculus for structural equation models. For a detailed overview, see [126] or [161]. We will use these notions in Section 7.4 to interpret our definition of intervention for SDEs in terms of intervention calculus for structural equation models.

Let V be a finite set, and let E be a subset of $V \times V$. A directed graph G on V is a pair (V, E). We refer to V as the vertex set, and refer to E as the edge set. Note that by this definition, there can be at most one edge between any pair of vertices. A path is an unbroken series of vertices and edges such that no vertices are repeated except possibly the initial and terminal vertices. A directed cycle is a path with the same initial and terminal vertices and all arrows pointing in the same direction. We say that G is an acyclic directed graph (DAG) if G contains no directed cycles. Note that this in particular excludes that the graph contains an edge with the same initial and terminal vertex. For any graph G and $i \in V$, we write $\mathrm{pa}(i) = \{j \in V \mid (j, i) \in E\}$, and refer to $\mathrm{pa}(i)$ as the parents of the vertex i. If we wish to emphazise the graph G, we also write $\mathrm{pa}_G(i)$.

A structural equation model (SEM) consists of three components:

1. Two families $(X_i)_{i \in V}$ and $(U_i)_{i \in V}$ of random variables.
2. A directed acyclic graph G on V.
3. A set of functional relationships $X_i = f_i(X_{\mathrm{pa}_G(i)}, U_i)$.

We refer to $(X_i)_{i \in V}$ as the primary variables and $(U_i)_{i \in V}$ as the noise variables. Note that we do not a priori assume that the noise variables are independent. The idea behind a SEM is that the DAG provides the sequence in which the functional relationships are evaluated, thus yielding an algorithm for obtaining the values of $(X_i)_{i \in V}$ from $(U_i)_{i \in V}$. A SEM does not only yield the distribution of the variables $(X_i)_{i \in V}$, but also a description of the data-generating mechanism. This is made precise by the notion of an intervention, see Definition 3.2.1 of [126]. For completeness, we repeat the definition here.

Definition 7.3.1. Consider a SEM with primary variables $(X_i)_{i \in V}$, noise variables $(U_i)_{i \in V}$, DAG G and functional relationships $X_i = f_i(X_{\mathrm{pa}_G(i)}, U_i)$. Let A be a subset of V. The postintervention SEM obtained by doing $X_i := x_i$ for $i \in A$ is the SEM with primary variables $(X_i)_{i \in V}$, noise variables $(U_i)_{i \in V}$, DAG G' obtained by removing all edges with terminal vertices $i \in A$ from G and functional relationships obtained by substituting x_i for X_i in all functional relationships with $i \notin A$ as well as exchanging all equations corresponding to indicies $i \in A$ with the simple equations $X_i = x_i$.

The idea behind Definition 7.3.1 is that if the algorithm implicit in a SEM represents the data-generating mechanism for $(X_i)_{i \in V}$, then an intervention in the system re-

sulting in fixing X_i at the value x_i for $i \in A$ would yield a data-generating mechanism corresponding to substituting the value x_i in all functional relationships involving X_i for $i \in A$.

7.4 Interpretation of postintervention SDEs

In this section, we show that under Lipschitz conditions on the coefficients in (7.8), the solution to the postintervention SDE described in Definition 7.2.2 is the limit of a sequence of postintervention SEMs as described in Definition 7.3.1 based on the Euler scheme of (7.8). We use this to clarify the role of the driving semimartingales $Z^1, \ldots, Z^d$ in relation to the causal interpretation of their SDE.

Definition 7.4.1. The signature of the SDE (7.8) is the graph S with vertex set $\{1, \ldots, p\}$ and an edge from i to j if it holds that there is k such that the mapping a_{jk} is not independent of the i'th coordinate.

Letting $a_{j\cdot} = (a_{j1}, \ldots, a_{jd})$, another way of describing the signature S in Definition 7.4.1 is that there is an edge from i to j if $x_i \mapsto a_{j\cdot}(x)$ is not constant, or equivalently, there is no edge from i to j if it holds for all k that a_{jk} does not depend on the i'th coordinate. From an intuitive viewpoint, the signature S defined in Definition 7.4.1 describes which coordinates of the SDE (7.8) are causally dependent on each other in an infinitesimal sense: There is an edge from i to j if and only if X^i has an infinitesimal causal effect on X^j. This motivates the following definition.

Definition 7.4.2. We say that X^j is locally unaffected by X^i in the SDE (7.8) if there is no edge from i to j in the signature of (7.8).

Being locally unaffected is a property of two coordinates of an SDE. If there is no risk of ambiguity, we leave out the SDE and simply state that X^j is locally unaffected by X^i.

The signature is used in the following definition to define a SEM corresponding to the Euler scheme for (7.8). With a slight abuse of notation we choose in Definition 7.4.3 for convenience to consider the initial variables $X_0^1, \ldots, X_0^p$ as primary variables instead of noise variables. This is not a problem as it is nonetheless clear how interventions for the SEM given in Definition 7.4.3 should be understood.

Definition 7.4.3. Fix $T > 0$ and consider $\Delta > 0$ such that T/Δ is a natural number. Let $N = T/\Delta$ and $t_k = k\Delta$. The Euler SEM over $[0, T]$ with step size Δ for (7.8) consists of the following:

1. The primary variables are the $p(N+1)$ variables in the set $(X_{t_k}^\Delta)_{0 \leq k \leq N}$, indexed by $\{0, \ldots, N\} \times \{1, \ldots, p\}$.
2. The noise variable for the i'th coordinate of $X_{t_k}^\Delta$ is the d-dimensional variable $Z_{t_k} - Z_{t_{k-1}}$.

3. The DAG is the graph $G = (V, E)$ with vertex set $\{0, \ldots, N\} \times \{1, \ldots, p\}$ defined by having $((i_1, j_1), (i_2, j_2))$ be an edge of D if and only if $i_2 = i_1 + 1$ and either $j_2 = j_1$ or (j_1, j_2) is an edge in the signature of (7.8).

4. The functional relationships are given by:

$$(X_{t_k}^{\Delta})^i = (X_{t_{k-1}}^{\Delta})^i + \sum_{j=1}^{d} a_{ij}(X_{t_{k-1}}^{\Delta})(Z_{t_k}^j - Z_{t_{k-1}}^j). \tag{7.12}$$

A visualization of the DAG for the SEM of Definition 7.4.3 is shown in Figure 7.4.1. The figure shows how the signature S determines the DAG describing the algorithm for calculating the variables in the Euler SEMs. For convenience, we have also included the error variables of the Euler SEM in Figure 7.4.1, with dotted directed edges to distinguish them from the primary variables. Making the intervention $(X^{\Delta})_{t_k}^1 := \zeta$ for all k corresponds to removing all edges of the DAG in Figure 7.4.1 with terminal vertex in the top row.

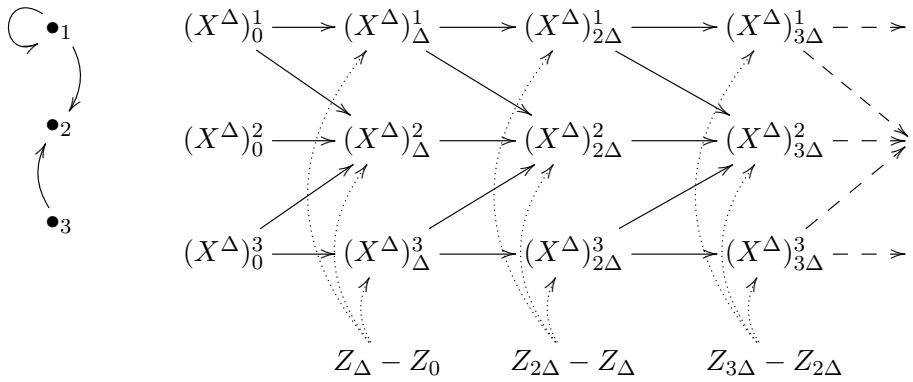


Figure 7.4.1: The signature for a three-dimensional SDE (left) and the DAG for the corresponding Euler SEM (right).

Combining the following two lemmas yields the main result of this section.

Lemma 7.4.4. *Fix $T > 0$ and let $(\Delta_n)_{n\geq 1}$ be a sequence of positive numbers converging to zero such that T/Δ_n is natural for all $n \geq 1$. For each n, there exists a pathwisely unique solution to the equation*

$$(X_t^n)^i = X_0^i + \sum_{j=1}^{d} \int_0^t a_{ij}(X_{\eta_n(s-)}^n)\, \mathrm{d}Z_s^j, \qquad i \leq p, \tag{7.13}$$

where $\eta_n(t) = k\Delta_n$ for $k\Delta_n \leq t < (k+1)\Delta_n$, satisfying that $((X^n)_{t_k})_{0\leq k\leq T/\Delta_n}$ are the primary variables in the Euler SEM for (7.8), and $\sup_{0\leq t\leq T} |X_t - X_t^n|$ converges in probability to zero, where X is the solution to (7.8).

Proof. By inspection, (7.13) has a unique solution, and $((X^n)_{t_k})_{k \leq T/\Delta_n}$ is the primary variables in the Euler SEM for (7.8). That $\sup_{0 \leq t \leq T} |X_t - X_t^n|$ converges in probability to zero is the corollary to Theorem V.16 of [134]. □

Lemma 7.4.5. *Fix $T > 0$ and consider $\Delta > 0$ such that T/Δ is a natural number. Fix $m \leq p$ and $\zeta \in \mathbb{R}$. The Euler SEM for (7.10) is equal to the postintervention SEM obtained by making the intervention $(X_{t_k}^\Delta)^m := \zeta$ for $0 \leq k \leq T/\Delta$ in the Euler SEM for (7.8).*

Proof. The functional relationships in the Euler SEM for (7.8) are

$$(X_{t_k}^\Delta)^i = (X_{t_{k-1}}^\Delta)^i + \sum_{j=1}^{d} a_{ij}(X_{t_{k-1}}^\Delta)(Z_{t_k}^j - Z_{t_{k-1}}^j), \tag{7.14}$$

while for (7.10) and $i \neq m$, they are

$$\begin{aligned}(Y_{t_k}^\Delta)^i &= (Y_{t_{k-1}}^\Delta)^i + \sum_{j=1}^{d} b_{ij}(Y_{t_{k-1}}^\Delta)(Z_{t_k}^j - Z_{t_{k-1}}^j) \\ &= (Y_{t_{k-1}}^\Delta)^i + \sum_{j=1}^{d} a_{ij}(Y_{t_{k-1}}^\Delta)(Z_{t_k}^j - Z_{t_{k-1}}^j),\end{aligned} \tag{7.15}$$

and $(Y_{t_k}^\Delta)^m = (Y_{t_{k-1}}^\Delta)^m$, which, since $(Y^\Delta)_0^m = \zeta$, yields $(Y^\Delta)_{t_k}^m = \zeta$ for all k. By inspection, (7.15) is the result of substituting ζ for $(X_{t_{k-1}}^\Delta)^m$ in (7.14). The result follows. □

Together, Lemma 7.4.4 and Lemma 7.4.5 states that the diagram in Figure 7.4.2 commutes: Defining interventions directly in terms of changing the terms in the stochastic differential equation has the same effect as intervening in the Euler SEM and taking the limit.

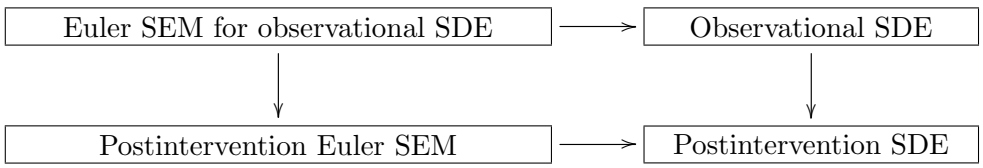


Figure 7.4.2: The interpretation of intervention in a stochastic differential equation understood as the limit of interventions in the Euler SEMs.

These results clarify what Definition 7.2.2 means: Intuitively, we consider the semimartingale Z as "autonomous" and assume that interventions do not directly influence this semimartingale. This autonomy is made concrete by assuming that the

family $(Z_{t_k} - Z_{t_{k-1}})_{k \leq N}$ are the noise variables in the Euler SEM, such that there are no arrows in the DAG for the SEM with terminal vertices in $(Z_{t_k} - Z_{t_{k-1}})_{k \leq N}$. The lemmas show that when this condition holds true, the notion of intervention given in Definition 7.2.2 is consistent with the result of intervention in the Euler SEM. Note that this does not constitute a proof of causality. Rather, it gives guidelines as to when it is reasonable to expect that our notion of intervention will reflect real-world interventions: namely, when none of the coordinates X^i have a direct effect on the driving semimartingales. Whether this is the case or not is in general not a testable assumption.

Also note that as we are not using the Euler SEMs to draw any conclusions about the distribution of the variables, we do not require independence of the noise variables $(Z_{t_k} - Z_{t_{k-1}})_{k \leq N}$. In particular, the variables in the Euler SEM do not need to be Markov with respect to the DAG in the sense of [126].

Concluding this section, we give two examples to illutrate the nature of interventions. In Example 7.4.6, we calculate the postintervention SDE for an intervention in an Ornstein-Uhlenbeck SDE, and in Example 7.4.7, we illustrate the necessity of a sharp division between autonomous and non-autonomous interpretations of processes.

Example 7.4.6. Let $x_0 \in \mathbb{R}^p$, $A \in \mathbb{R}^p$, $B \in \mathbb{M}(p,p)$ and $\sigma \in \mathbb{M}(p,d)$. The Ornstein-Uhlenbeck SDE with initial value X_0, mean reversion level A, mean reversion speed B, diffusion matrix σ and d-dimensional driving noise is

$$X_t = X_0 + \int_0^t B(X_s - A)\,\mathrm{d}s + \sigma W_t, \tag{7.16}$$

where W is a d-dimensional $(\mathcal{F}_t)$ Brownian motion, see Section II.72 of [142]. Fix a coordinate $m \leq p$ and $\zeta \in \mathbb{R}$. Making the intervention $X^m := \zeta$, we obtain that the postintervention process Y satisfies $Y_t^m = \zeta$, and for $i \neq m$,

$$Y_t^i = X_0^i + \int_0^t \sum_{j \neq m}^{p} B_{ij}(Y_s^j - A_j) + B_{im}(\zeta - A_m)\,\mathrm{d}s + \sum_{j=1}^{d} \sigma_{ij} W_t^j. \tag{7.17}$$

Now let $\tilde{B}$ be the submatrix of B obtained by removing the m'th row and column of B, and assume that $\tilde{B}$ is invertible. With Y^{-m} denoting the $p-1$ dimensional process obtained by removing the m'th coordinate from Y, we then obtain

$$Y_t^{-m} = Y_0 + \int_0^t \tilde{B}(Y_s^{-m} - \tilde{A})\,\mathrm{d}s + \tilde{\sigma} W_t, \tag{7.18}$$

where Y_0 is obtained by removing the m'th coordinate from X_0, $\tilde{\sigma}$ is obtained by removing the m'th row of σ and $\tilde{A} = \alpha - \tilde{B}^{-1}\beta$, where α and β are obtained by removing the m'th coordinate from A and from the vector whose i'th component is $B_{im}(\zeta - A_m)$, respectively. Thus, Y^{-m} solves an $(p-1)$-dimensional Ornstein-Uhlenbeck SDE with initial value Y_0, mean reversion level $\tilde{A}$, mean reversion speed $\tilde{B}$ and diffusion matrix $\tilde{\sigma}$. ∘

The next example shows that an SDE may not always be amenable to a causal interpretation of the type given in Definition 7.2.2.

Example 7.4.7. Let $X^1 = W$ be a one-dimensional Wiener process, consider a twice continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$ and let $X_t^2 = f(X_t^1)$ for $t \geq 0$. If this relation constitutes the actual causal relation between X^1 and X^2, the result of the intervention $X^1 := \zeta$ should be that $X_t^2 = f(\zeta)$. However, by Itô's lemma, we obtain

$$\begin{aligned} X_t^2 &= f(X_0^1) + \frac{1}{2}\int_0^t f''(X_s^1)\,\mathrm{d}[X^1]_s + \int_0^t f'(X_s^1)\,\mathrm{d}X_s^1 \qquad (7.19)\\ &= f(0) + \frac{1}{2}\int_0^t f''(X_s^1)\,\mathrm{d}s + \int_0^t f'(X_s^1)\,\mathrm{d}W_s, \end{aligned}$$

yielding the system of SDEs

$$X_t^1 = \int_0^t \mathrm{d}W_s \qquad (7.20)$$

$$X_t^2 = f(0) + \frac{1}{2}\int_0^t f''(X_s^1)\,\mathrm{d}s + \int_0^t f'(X_s^1)\,\mathrm{d}W_s, \qquad (7.21)$$

which is of the form given in (7.8). Applying Definition 7.2.2 to this SDE, the resulting postintervention SDE for X^2 under the intervention $X^1 := \zeta$ becomes

$$X_t^2 = f(0) + \frac{1}{2}\int_0^t f''(\zeta)\,\mathrm{d}s + \int_0^t f'(\zeta)\,\mathrm{d}W_s, \qquad (7.22)$$

which yields the result $X_t^2 = f(0) + \frac{1}{2}f''(\zeta)t + f'(\zeta)W_t$, in contradiction with the correct result, $X_t^2 = f(\zeta)$. The problem is that by substituting W for X^1 in the SDE derived from Itô's lemma, the resulting SDE loses its causal interpretation. The driving semimartingale W is not autonomous, and as a consequence, the postintervention SDE does not give the desired result. This example also shows that whether a process is autonomous or not is not something which may be determined simply by investigation of the SDE in question, but is a judgement to be made from case to case. ∘

Example 7.4.7 is not meant to put to question the appropriateness of Definition 7.2.2, but rather to show the limitations of this definition. We should note that it is not the use of Itô's lemma in itself which is the problem in Example 7.4.7, it is the subsequent substitution of X^1 by W. In fact, if we intervene directly in (7.19) by replacing X^1 by the constant ζ, the result would be that $X_t^2 = f(\zeta)$. We could thus say that (7.19) retains the causal interpretation. However, Definition 7.2.2 does not allow for such interventions on the integrators. To do so generally would complicate matters considerably, and we will not pursue this any further.

7.5 Identifiability of postintervention distributions

In this section, we formulate a result, Theorem 7.5.3, giving conditions for the postintervention distributions to be uniquely determined by the semigroup of the SDE. Our objective is to show that this uniqueness holds when the driving semimartingale for the SDE is a Lévy process and the coefficients of the SDE are Lipschitz and bounded. Before stating and proving our main result, we review some basic notions of Markov process theory and Lévy processes.

Recall from Chapter 4 of [51] that a family of transition probabilities on $\mathbb{R}^p$ is a family $(P_t(x,\cdot))_{x\in\mathbb{R}^p,t\geq 0}$ of probability measures on $\mathbb{R}^p$ such that $(t,x)\mapsto P_t(x,B)$ is measurable for all Borel measurable $B\subseteq\mathbb{R}^p$, $P_0(x,\cdot)$ is the Dirac measure in x and for all $t,s\geq 0$ it holds that $P_{t+s}(x,B)=\int_{\mathbb{R}^p}P_s(y,B)P_t(x,\mathrm{d}y)$. Given a càdlàg stochastic process X with values in $\mathbb{R}^p$, we say that X is an $(\mathcal{F}_t)$ Markov process if there is a family $P_t(x,\cdot)$ of transition probabilities on $\mathbb{R}^p$ such that for $s,t\geq 0$ and $B\in\mathcal{B}_p$, it holds that $P(X_{t+s}\in B|\mathcal{F}_t)=P_s(X_t,B)$ almost surely. In this case, we say that X has transition probabilities $(P_t(x,\cdot))$. If this holds with the filtration induced by the process itself, meaning that $\mathcal{F}_t$ is the σ-algebra generated by the variables $(X_s)_{s\leq t}$, we simply say that X is a Markov process.

Let $\mathbf{b}(\mathbb{R}^p)$ denote the space of bounded Borel measurable functions from $\mathbb{R}^p$ to $\mathbb{R}$. For a family of transition probabilities $P_t(x,\cdot)$, we define $P_t:\mathbf{b}(\mathbb{R}^p)\to\mathbf{b}(\mathbb{R}^p)$ by $P_tf(x)=\int f(y)P_t(x,\mathrm{d}y)$. The mapping P_t is then a linear operator on $\mathbf{b}(\mathbb{R}^p)$. Furthermore, P_0 is the identity operator, $\|P_t\|\leq 1$ for all $t\geq 0$ where $\|\cdot\|$ denotes the operator norm induced by the uniform norm on $\mathbf{b}(\mathbb{R}^p)$, and it holds that $P_{t+s}=P_tP_s$ for $t,s\geq 0$. All in all, this implies that (P_t) is a contraction semigroup of operators.

Next, let $C_0(\mathbb{R}^p)$ denote the Banach space of continuous mappings from $\mathbb{R}^p$ to $\mathbb{R}$ vanishing at infinity, endowed with the uniform norm, see Chapter 5 of [116]. Also, let $C_0^2(\mathbb{R}^p)$ denote the subset of $C_0(\mathbb{R}^p)$ which are twice continuously differentiable with all partial derivatives in $C_0(\mathbb{R}^p)$. We say that the semigroup (P_t) is Feller if P_t maps $C_0(\mathbb{R}^p)$ into itself and for all $f\in C_0(\mathbb{R}^p)$, the mapping $t\mapsto P_tf$ from $[0,\infty)$ to $C_0(\mathbb{R}^p)$ is continuous at zero. The restriction of (P_t) to $C_0(\mathbb{R}^p)$ is then a strongly continuous contraction semigroup. In this case, we let $\mathcal{D}(A)$ be the set of $f\in C_0(\mathbb{R}^p)$ where $\lim_{t\to 0}t^{-1}(P_tf-P_0f)$ exists as a limit in $C_0(\mathbb{R}^p)$, and when it exists, we let Af denote the limit. We refer to $\mathcal{D}(A)$ as the domain of A, and we refer to A as the generator of the semigroup. By Corollary 1.1.6 of [51], A is then a densely defined and closed linear operator on $C_0(\mathbb{R}^p)$. Finally, if X is a càdlàg Markov process with a Feller semigroup, we say that X is a Feller process, and we say that X has generator A, where A is the generator of the Feller semigroup.

Having reviewed the relevant notions of Markov process theory, we next recall some basic results for Lévy processes, see [7] or [151] for an overview. Recall that a Lévy measure on $\mathbb{R}^d$ is a measure ν assigning zero measure to $\{0\}$ such that $x\mapsto\min\{1,\|x\|^2\}$ is integrable with respect to ν. A d-dimensional Lévy triplet is a triplet (α,C,ν), where α is an element of $\mathbb{R}^d$, C is a positive semidefinite $d\times d$

matrix and ν is a Lévy measure on $\mathbb{R}^d$. Further recall by Theorem 1.2.14 of [7] that for any bounded neighborhood D of zero in $\mathbb{R}^d$ and any d-dimensional Lévy process X, there is a Lévy triplet (α, C, ν) such that

$$Ee^{iuX_1} = \exp\left(iu^t\alpha - \frac{1}{2}u^tCu - \int_{\mathbb{R}^d} e^{iu^tx} - 1 - iu^tx1_D(x)\,\mathrm{d}\nu(x)\right), \tag{7.23}$$

and this triplet uniquely determines the distribution of X. We refer to (α, C, ν) as the characteristics of X with respect to D, or as the D-characteristic triplet of X. Conversely, for any bounded neighborhood D of zero in $\mathbb{R}^d$ and any Lévy triplet (α, C, ν), there exists a Lévy process having (α, C, ν) as its D-characteristic triplet.

We are now ready to state our main result. Lemma 7.5.1 and Definition 7.5.2 introduce the semigroup of an SDE, and Theorem 7.5.3 is the identifiability result. Let D be a bounded neighborhood of zero in $\mathbb{R}^d$. Consider the SDE

$$X_t^i = X_0^i + \sum_{j=1}^d \int_0^t a_{ij}(X_{s-})\,\mathrm{d}Z_s^j, \qquad i \leq p, \tag{7.24}$$

where Z is a d-dimensional Lévy process with D-characteristic triplet (α, C, ν), the mapping $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ is Lipschitz and bounded and X_0 is some variable.

Lemma 7.5.1. *There exists a unique Feller semigroup (P_t) with the property that any solution of (7.24), independent of the initial distribution and the probability space on which the solution exists, is a Feller process with semigroup (P_t).*

Definition 7.5.2. We refer to the semigroup of Lemma 7.5.1 as the semigroup of the SDE (7.24).

Theorem 7.5.3. *Consider the SDEs*

$$X_t^i = X_0^i + \sum_{j=1}^d \int_0^t a_{ij}(X_s)\,\mathrm{d}Z_s^j, \qquad i \leq p, \tag{7.25}$$

and

$$Y_t^i = Y_0^i + \sum_{j=1}^d \int_0^t \tilde{a}_{ij}(Y_s)\,\mathrm{d}\tilde{Z}_s^j, \qquad i \leq p, \tag{7.26}$$

where Z is a d-dimensional Lévy process, $\tilde{Z}$ is a $\tilde{d}$-dimensional Lévy process and the mappings $a : \mathbb{R}^p \to \mathbb{M}(p, d)$ and $\tilde{a} : \mathbb{R}^p \to \mathbb{M}(p, \tilde{d})$ are Lipschitz and bounded. Assume that (7.25) and (7.26) have the same semigroup, and that the initial values have the same distribution. Then, the postintervention distributions of doing $X^m := \zeta$ in (7.25) and doing $Y^m := \zeta$ in (7.26) are equal for all m and ζ.

Lemma 7.5.1 and Theorem 7.5.3 are proven in Section 7.8. Note that the requirement that a and $\tilde{a}$ be bounded in Theorem 7.5.3 is only used to ensure the Feller property. Theorem 7.5.3 states that for SDEs with a Lévy process as the driving semimartingale, postintervention distributions are identifiable from the semigroup. In the remainder of this section, we discuss the content of Theorem 7.5.3.

First, recall that a main theme of the DAG-based framework for causal inference as in [161] and [126] is to identify conditions for when postintervention distributions are identifiable from the observational distribution. Theorem 7.5.3 gives a criterion for when postintervention distributions are identifiable from the semigroup of the SDE, which is not exactly the same. Nonetheless, in a large family of naturally occurring cases, the semigroup is identifiable from the observational distribution, for example, if the solutions to (7.25) and (7.26) are irreducible.

Next, we comment on the relationship between the result of Theorem 7.5.3 and the identifiability results of DAG-based causal inference. Consider the Euler SEM of Definition 7.4.3, illustrated in Figure 7.4.1. In the DAG of this SEM, the orientation of all arrows is assumed known: All orientations for arrows from primary variables point forward in time. If the error variables for each primary variable were independent, it would hold that the distribution of the variables would be Markov with respect to the DAG in the sense of [126]. In this case, by the results of [171], we would be able to identify the skeleton of the graph (that is, its undirected edges) from the observational distribution. As all orientations are given, this leads to identifiability of the entire graph. Using the truncated factorization (3.10) of [126], this leads to identifiability of intervention distributions from the observational distribution. Thus, in this case, identifiability would not be a surprising result.

However, when the driving semimartingale Z is a Lévy process, the error variables are independent across time, but are not independent across coordinates: For each k, the variables $X^1_{\Delta k}, \dots, X^p_{\Delta k}$ have the same d-dimensional error variable, namely $Z_{\Delta k} - Z_{\Delta(k-1)}$, and so the Euler SEM illustrated in Figure 7.4.1 is not Markov with respect to its DAG. Therefore, our scenario differs from the conventional causal modeling scenario of [126] in two ways: Both by considering a continuous-time model with uncountably many variables and by considering a particular type of dependent errors.

As the final result of this section, we give an example of a particularly simple case where identifiability can be seen explicitly from the transition probabilities.

Example 7.5.4. Let W and $\tilde{W}$ be d-dimensional and $\tilde{d}$-dimensional Brownian motions, let B and $\tilde{B}$ be $p \times p$ matrices, and let σ and $\tilde{\sigma}$ be $p \times d$ and $p \times \tilde{d}$ matrices. Consider two processes X and Y being the unique solutions to the Ornstein-Uhlenbeck SDEs

$$X_t = X_0 + \int_0^t BX_t \, \mathrm{d}t + \sigma W_t \tag{7.27}$$

and

$$Y_t = Y_0 + \int_0^t \tilde{B} X_t \, \mathrm{d}t + \tilde{\sigma} W_t. \tag{7.28}$$

We will show by a direct analysis that if the SDEs have the same semigroup and the initial distributions are are equal, then the postintervention distributions are equal as well. For notational simplicity, we consider intervening on the first coordinate, making the interventions $X^1 := \zeta$ and $Y^1 := \zeta$. It will suffice to show equality of distributions for the non-intervened coordinates in the postintervention distributions. Consider block decompositions of the form

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \qquad \text{and} \qquad \sigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix}, \tag{7.29}$$

where B_{11} is a 1×1 matrix and B_{22} is a $(p-1) \times (p-1)$ matrix and σ_1 is a $1 \times d$ matrix and σ_2 is a $(p-1) \times d$ matrix. Also consider corresponding decompositions of $\tilde{B}$ and $\tilde{\sigma}$.

Assume that the semigroups and the initial distributions are equal. In particular, the transition probabilities for X and Y are the same. With $P_t(x, \cdot)$ denoting the transition probability of moving from state x in time t for X, the results of [80] show that

$$P_t(x, \cdot) = \mathcal{N}\left(\exp(tB)x, \int_0^t \exp(sB)\sigma\sigma^t \exp(sB^t) \, \mathrm{d}s \right), \tag{7.30}$$

where the right-hand side denotes a Gaussian distribution, and similarly for the transition probabilities of Y. As these are equal for all $x \in \mathbb{R}^p$ and $t \geq 0$, we obtain $\exp(tB) = \exp(t\tilde{B})$ for all $t \geq 0$, so by differentiating, $B = \tilde{B}$ as well. Likewise, as $\int_0^t \exp(sB)\sigma\sigma^t \exp(sB^t) \, \mathrm{d}s = \int_0^t \exp(s\tilde{B})\tilde{\sigma}\tilde{\sigma}^t \exp(s\tilde{B}^t) \, \mathrm{d}s$ for all $t \geq 0$, we obtain $\sigma\sigma^t = \tilde{\sigma}\tilde{\sigma}^t$. Note that

$$\sigma\sigma^t = \begin{bmatrix} \sigma_1\sigma_1^t & \sigma_1\sigma_2^t \\ \sigma_2\sigma_1^t & \sigma_2\sigma_2^t \end{bmatrix}, \tag{7.31}$$

and similarly for $\tilde{\sigma}\tilde{\sigma}^t$. Therefore, we obtain in particular that $\sigma_2\sigma_2^t = \tilde{\sigma}_2\tilde{\sigma}_2^t$.

Now, applying Definition 7.2.2 and recalling Example 7.4.6, the intervened processes minus the first coordinate, $\tilde{X}^{-1}$ and $\tilde{Y}^{-1}$ (note that the superscripts do not denote reciprocals), are Ornstein-Uhlenbeck processes with initial values X_0^{-1} and Y_0^{-1}, mean reversion speeds B_{22} and $\tilde{B}_{22}$, mean reversion levels $-B_{22}^{-1}B_{21}\zeta$ and $-\tilde{B}_{22}^{-1}\tilde{B}_{21}\zeta$ and diffusion matrices σ_2 and $\tilde{\sigma}_2$. As X_0 and Y_0 have the same distribution, and we know that $B = \tilde{B}$ and $\sigma_2\sigma_2^t = \tilde{\sigma}_2\tilde{\sigma}_2^t$, we obtain that the distributions of $\tilde{X}^{-1}$ and $\tilde{Y}^{-1}$ must be equal. Thus, by direct calculation of transition probabilities, we see that for the Ornstein-Uhlenbeck SDE with zero mean reversion level, intervention distributions are identifiable from the observational distribution. ∘

7.6 Interventions and WCLI

In this section, we discuss the relationship between postintervention processes and weak conditional local independence (WCLI) of the observational process. We first review some results on random measures and semimartingale characteristics, see [83] for a detailed development of such results.

A random measure on $\mathbb{R}_+ \times \mathbb{R}^d$ is a family of nonnegative measures $(\mu(\omega, \cdot))_{\omega\in\Omega}$ such that $\mu(\omega, \{0\} \times \mathbb{R}^d) = 0$ for all ω. Put $\tilde{\Omega}_d = \Omega \times \mathbb{R}_+ \times \mathbb{R}^d$, $\tilde{\mathcal{O}}_d = \mathcal{O} \otimes \mathcal{B}_d$ and $\tilde{\mathcal{P}}_d = \mathcal{P} \otimes \mathcal{B}_d$, where $\mathcal{O}$ and $\mathcal{P}$ denote the optional and predictable σ-algebras on $\Omega \times \mathbb{R}^+$, respectively. A mapping from $\tilde{\Omega}_d$ to $\mathbb{R}$ which is $\tilde{\mathcal{O}}_d$ measurable is called an optional function, and a mapping from $\tilde{\Omega}_d$ to $\mathbb{R}$ which is $\tilde{\mathcal{P}}_d$ measurable is called a predictable function. If we wish to make the filtration $(\mathcal{F}_t)$ explicit, we refer to $(\mathcal{F}_t)$ optional and $(\mathcal{F}_t)$ predictable functions. Note that as $\tilde{\mathcal{O}}_d \subseteq \mathcal{F} \otimes \mathcal{B}_+ \otimes \mathcal{B}_d$, it holds that for any optional function W and any fixed $\omega \in \Omega$, $(t, x) \mapsto W(\omega, t, x)$ is $\mathcal{B}_+ \otimes \mathcal{B}_d$ measurable. Therefore, the integral $\int_{[0,t]\times\mathbb{R}^d} |W(\omega, s, x)| \, \mathrm{d}\mu(\omega, \mathrm{d}s, \mathrm{d}x)$ is always well-defined. We write $(|W| * \mu)_t(\omega)$ for this integral. When $(|W| * \mu)_t(\omega)$ is finite for all ω and $t \geq 0$, we furthermore define $(W * \mu)_t(\omega) = \int_{[0,t]\times\mathbb{R}^d} W(\omega, s, x) \, \mathrm{d}\mu(\omega, \mathrm{d}s, \mathrm{d}x)$. If $W * \mu$ is optional for all nonnegative bounded optional μ-integrable functions W, we say that μ is optional. If $W * \mu$ is predictable for all nonnegative bounded predictable μ-integrable functions W, we say that W is predictable. For any optional random measure μ, we say that μ is $\tilde{\mathcal{P}}_d$-σ-finite if there is a partition $(A_n)_{n\geq 1}$ of $\tilde{\mathcal{P}}_d$ measurable sets of $\tilde{\Omega}_d$ such that $E(1_{A_n} * \mu)_\infty$ is finite.

By Theorem II.1.8 of [83], for any optional $\tilde{\mathcal{P}}_d$-σ-finite random measure μ, there exists a predictable random measure ν, unique up to indistinguishability, such that for all nonnegative bounded $\tilde{\mathcal{P}}_d$ measurable functions W, $E(W * \nu)_\infty = E(W * \mu)_\infty$. We refer to ν as the compensator of μ. Furthermore, Theorem II.1.8 of [83] also shows that if $|W| * \mu$ is locally integrable, then $|W| * \nu$ is locally integrable as well.

We now introduce the characteristics of a d-dimensional semimartingale X. For such a semimartingale, we define the jump measure μ^X for X by letting $\mu^X(\omega)$ be the measure on $\mathcal{B}_+ \otimes \mathcal{B}_d$ defined by

$$\mu^X(\omega)(A) = \sum_{t\geq 0} 1_A(t, \Delta X_t(\omega)). \tag{7.32}$$

By Proposition II.1.16 of [83], μ^X is optional and $\tilde{\mathcal{P}}_d$-σ-finite. Therefore, the compensator of μ^X exists, we denote it by ν^X. Furthermore, we define a mapping $h^d : \mathbb{R}^d \to \mathbb{R}^d$ by letting $h^d(x) = x1_{(\|x\|_2\leq 1)}$, the canonical truncation function. Then $X_t - \sum_{0<s\leq t} \Delta X_s - h^d(\Delta X_s)$ is a special semimartingale, and we let B be its predictable finite variation part. Finally, we let C be the process with values in the real symmetric $d \times d$ matrices given by $C^{ij}_t = [(X^i)^c, (X^j)^c]_t$, where $(X^i)^c$ denotes the continuous martingale part of X^c, see Proposition I.4.27 of [83]. We then define the h^d-characteristics of X to be the triple (B, C, ν^X). For convenience, we will also just refer to (B, C, ν^X) as the characteristics of X, supressing the dependence on

h^d. By Remark II.2.8 of [83], for fixed d, the h^d-characteristics are unique up to indistinguishability.

We are now ready to state the definition of weak conditional local independence. In [57], the following definition of weak conditional local independence is made. Assume that Y is a d-dimensional special semimartingale with a decomposition of the form $Y = Y_0 + A + M$, where A is predictable and of finite variation and M is a local martingale. Let (B, C, ν) be the characteristics of Y. In [57] it is further assumed that the coordinates of M have zero quadratic covariation and that the characteristic C is deterministic. In this case, Definition 2 of [57] states that X^i is weakly conditionally locally independent (WCLI) of X^m if the characteristics B^i and ν^i of X^i are $(\mathcal{F}_t^{-m})$ predictable, where $(\mathcal{F}_t^{-m})_{t\geq 0}$ is the usual augmentation of the filtration induced by the processes $X^1, \ldots, X^p$ excluding X^m. This definition is well-posed whenever the characteristics (B, C, ν) are unique. Therefore, it can be extended to all special semimartingales. Making this extension, we obtain the following theorem.

Theorem 7.6.1. *Let X be the solution to (7.8). Assume that X is a special semimartingale and that Z is a Lévy process. If X^i is locally unaffected by X^m in (7.8), then X^i is WCLI of X^m.*

Theorem 7.6.1 is proven in Section 7.9. Intuitively, Theorem 7.6.1 states that under certain assumptions on the driving semimartingales, having X^i locally unaffected by X^m yields that X^i in a sense is locally independent of X^m, a notion made precise by having the characteristics of X^i $(\mathcal{F}_t^{-m})$ predictable.

7.7 Discussion

In this section, we will reflect on the results of the preceeding sections and discuss opportunities for further work.

The definition of the postintervention SDE, Definition 7.2.2, is certainly a natural way to define how interventions should affect stochastic dynamic systems. However, the definition reflects unstated assumptions about causality, and it is important to make precise when the definition can be assumed to reflect an actual real-world intervention and when the definition is simply a mathematical construct. This is clarified in Section 7.4, where we used the DAG-based intervention calculus to show that the postintervention SDE of Definition 7.2.2 can be assumed to reflect real-world interventions when the following hold:

1. The SDE reflects a data-generating mechanism in which the variables at a given timepoint are obtained as a function of the previous timepoints and the driving semimartingales.

2. The driving semimartingales are not directly affected by interventions, in the sense that they can be taken to be noise variables in the Euler SEMs.

In full generality, causal mechanisms of a model are not identifiable from the observational distribution, see [171]. However, when considering only restricted classes of structural equation models, the underlying causal mechanisms may often be identifiable, see for example [180, 73, 129]. In such cases, linearity of the functional relationships or Gaussianity of the noise variables often determine identifiability. In our case, as shown in Section 7.5, identifiability holds whenever the driving semimartingale is a Lévy process. This ensures practical applicability of our results. The proofs given in Section 7.5 use the Markov structure of the solution to the SDE. In the case where the driving semimartingale has independent, but not stationary, increments, the solution to the SDE will be a non-homogeneous Markov process, thus also amenable to operator methods, though requiring more powerful technical results. We expect that Theorem 7.5.3 extends to this case. Likewise, Theorem 7.6.1 also extends to the case of increments that are independent but not stationary, as can be seen by the fact that Theorem II.4.15 of [83] also holds for such processes.

It should also be noted that identifiability holds independently of the dimension of the driving Lévy process. This is useful, for instance, in relation to Example 7.2.1. We do not need to use the specific SDE driven by a four-dimensional Wiener process. We can replace the diffusion term in the SDE by a term involving the positive definite square root of the diffusion matrix and a two-dimensional Wiener process without affecting the postintervention distribution.

It is, however, important to be careful about the interpretation of the identifiability result. The result states that when using Definition 7.2.2 to model interventions, the postintervention distributions are identifiable. As discussed above, Definition 7.2.2 is not always useful as a notion of intervention: This requires that we are willing to interpret the SDE in a particular way. As Example 7.4.7 shows, not all SDEs are amenable to such an interpretation – this requires separate arguments, such as in Example 7.2.1.

A complete theory of interventions in continuous time stochastic processes should be able to cover cases such as Example 7.4.7. Our results should be seen as a step in the direction of a complete theory and encourage further generalizations. Another opportunity for further research concerns latent variables: In the DAG-based framework of [126], the back-door and front-door criteria shows how to calculate intervention effects from the observational distribution in the presence of latent variables. For an SDE, the causal structure is summarized in the signature, see Definition 7.4.1, which does not need to be acyclic, reflecting the possibility of feedback loops. It is an open question how to obtain similar results in terms of the signature in the case of, for example, a diffusion model with some coordinates being unobserved.

7.8 Proof of identifiability

In this section, we prove Lemma 7.5.1 and Theorem 7.5.3. We first consider Lemma 7.5.1, which will follow from the following two lemmas. Lemma 7.8.1 identifies the solutions to the relevant SDE as Feller processes and identifies the generator on $C_0^2(\mathbb{R}^p)$, and Lemma 7.8.2 shows that the semigroups of interest are identified uniquely by the values of their generators on $C_0^2(\mathbb{R}^p)$.

Lemma 7.8.1. *Let D be a bounded neighborhood of zero in $\mathbb{R}^d$. Consider the SDE (7.24), where Z is a d-dimensional Lévy process with D-characteristic triplet (α, C, ν), and $a : \mathbb{R}^p \to \mathbb{M}(p,d)$ is Lipschitz and bounded. The solution of (7.24) is a Feller process, and this Feller process has a generator whose domain includes $C_0^2(\mathbb{R}^p)$ and for any $f \in C_0^2(\mathbb{R}^p)$ and $x \in \mathbb{R}^p$, it holds that*

$$\begin{aligned} Af(x) = &\sum_{i=1}^{p}\sum_{j=1}^{d} a_{ij}(x)\alpha_j \frac{\partial f}{\partial x_i}(x) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(a(x)Ca(x)^t)_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &+ \int f(x + a(x)y) - f(x) - 1_D(y)\sum_{i=1}^{p}\frac{\partial f}{\partial x_i}(x)\sum_{j=1}^{d} a_{ij}(x)y_j \,\mathrm{d}\nu(y). \end{aligned} \tag{7.33}$$

Proof. Applying Theorem 2.4.16 of [7], we have

$$\begin{aligned} Z_t = \alpha t + BW_t &+ \int 1_{[0,t]\times D}(s,x)\,\mathrm{d}M(\,\mathrm{d}s,\,\mathrm{d}x) \\ &+ \int 1_{[0,t]\times D^c}(s,x)\,\mathrm{d}N(\,\mathrm{d}s,\,\mathrm{d}x), \end{aligned} \tag{7.34}$$

where $C = BB^t$ for some $B \in \mathbb{M}(d,d)$, W is a d-dimensional Brownian motion, N is a Poisson random measure on $\mathbb{R}_+ \times (\mathbb{R}^d \setminus \{0\})$ with intensity measure $m_+ \otimes \nu$, independent of W, and M is N minus its compensator. Here, m_+ denotes the Lebesgue measure on $\mathbb{R}_+$. We may then rewrite the SDE (7.24) as

$$\begin{aligned} X_t = X_0 &+ \int_0^t b(X_{s-})\,\mathrm{d}s + \int_0^t \sigma(X_{s-})\,\mathrm{d}W_s \\ &+ \int 1_{[0,t]\times D}(s,x)F(X_{s-},y)\,\mathrm{d}M(\,\mathrm{d}s,\,\mathrm{d}y) \\ &+ \int 1_{[0,t]\times D^c}(s,x)F(X_{s-},y)\,\mathrm{d}N(\,\mathrm{d}s,\,\mathrm{d}y) \end{aligned} \tag{7.35}$$

where $b(x) = a(x)\alpha$, $\sigma(x) = a(x)B$ and $F(x,y) = a(x)y$. Thus, the SDE is of the type given as (6.12) in [7]. By Theorem 6.4.5 of [7], X is therefore a Markov process, and by Theorem 6.7.4 of [7], it has a Feller transition semigroup with a generator A

whose domain includes $C_0^2(\mathbb{R}^p)$, and for $f \in C_0^2(\mathbb{R}^p)$, it holds that

$$\begin{aligned} Af(x) = &\sum_{i=1}^{p} b_i(x) \frac{\partial f}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (\sigma(x)\sigma(x)^t)_{ij} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &+ \int_D f(x + F(x,y)) - f(x) - \sum_{i=1}^{p} F_i(x,y) \frac{\partial f}{\partial x_i}(x) \, \mathrm{d}\nu(y) \\ &+ \int_{D^c} f(x + F(x,y)) - f(x) \, \mathrm{d}\nu(y). \end{aligned} \tag{7.36}$$

Substituting our expressions for b, σ and F, we obtain

$$\begin{aligned} Af(x) = &\sum_{i=1}^{p} \sum_{j=1}^{d} a_{ij}(x) \alpha_j \frac{\partial f}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (a(x) C a(x)^t)_{ij} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &+ \int f(x + a(x)y) - f(x) - 1_D(y) \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) \sum_{j=1}^{d} a_{ij}(x) y_j \, \mathrm{d}\nu(y), \end{aligned} \tag{7.37}$$

which is equal to (7.33). This proves the result. □

Lemma 7.8.2. *Let (P_t) and (Q_t) are two Feller semigroups with generators A and B. Assume that $\mathcal{D}(A)$ and $\mathcal{D}(B)$ both contain $C_0^2(\mathbb{R}^p)$ and that on $C_0^2(\mathbb{R}^p)$, A and B are of the type given in (7.33). If $Af = Bf$ for $f \in C_0^2(\mathbb{R}^p)$, then $(P_t) = (Q_t)$.*

Proof. We begin by considering the semigroup (P_t) and its generator A, and by introducing some definitions. Let A_0 denote the restriction of A to $C_c^2(\mathbb{R}^p)$. Let μ be some probability measure on $(\mathbb{R}^p, \mathcal{B}_p)$. As in Section 4.3 of [51], a càdlàg process X with values in $\mathbb{R}^p$, defined on some probability space, is said to be a solution of the martingale problem for (A, μ) if it holds that X_0 has distribution μ and the process

$$f(X_t) - \int_0^t A_0 f(X_s) \, \mathrm{d}s \tag{7.38}$$

is a martingale with respect to the filtration induced by X, for all $f \in C_c^2(\mathbb{R}^p)$. Also, as in [102], X is said to be a weak solution of (7.24) if there exists a filtered probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t)_{t \geq 0}, \tilde{P})$ endowed with processes $\tilde{X}$ and $\tilde{Z}$, where $\tilde{Z}$ is a d-dimensional Lévy process with D-characteristic triplet (α, C, ν) such that $\tilde{X}$ satisfies (7.24) and X and $\tilde{X}$ has the same distribution.

Now note that A_0 is of the type given as (13) in [102]. As C is positive semidefinite, we have $C = LL^t$ for some $L \in \mathbb{M}(d, d)$. The condition (14) in [102] then in our case translates into having

$$\sup_{x \in K} \|a(x)\alpha\|_2 + \|a(x)L\|_2^2 + \int_D \|a(x)u\|_2^2 \, \mathrm{d}\nu(u) + \int_{D^c} \|a(x)u\|_2 \wedge 1 \, \mathrm{d}\nu(u) \tag{7.39}$$

be finite for all compact K in $\mathbb{R}^p$, which holds as a is bounded on compacts and ν is a Lévy measure. Therefore, Theorem 2.3 of [102] is applicable and shows that any process X which is a solution to the martingale problem for (A_0, μ) is a weak solution of (7.24). However, due to our assumption that a is Lipschitz and bounded, Theorem 14.95 of [82] shows that for a fixed initial distribution, all solutions to the SDE (7.24) have the same distribution. Therefore, it also holds that all solutions to the martingale problem for (A_0, μ) will have the same distribution.

Now let B_0 be the restriction of B to $C_c^2(\mathbb{R}^p)$. By our assumptions, $A_0 = B_0$, so the martingale problem for (A_0, μ) has the same solutions as the martingale problem for (B_0, μ). Fix $x \in \mathbb{R}^p$ and let X^x and Y^x be solutions to two SDEs of the form (7.24) with initial value x and generators A and B, respectively, such processes exist by Theorem V.7 of [134]. By Lemma 7.8.1, X^x and Y^x are Feller processes with generators A and B, respectively, with respect to their own induced filtrations $(\mathcal{F}_t^X)$ and $(\mathcal{F}_t^Y)$, respectively. By (III.10.12) of [142], it then holds for all $f \in C_c^2(\mathbb{R}^p)$ and all bounded stopping times T with respect to $(\mathcal{F}_t^X)$ that

$$Ef(X_T) = E\int_0^T A_0 f(X_s)\,\mathrm{d}s, \tag{7.40}$$

which shows by Theorem II.77.6 of [142] that (7.38) is a martingale. Thus, X^x solves the martingale problem for (A_0, μ). Similarly, we find that Y^x solves the martingale problem for (B_0, μ). However, by what we already have seen, the two martingale problems have the same solutions, and moreover, all solutions have the same distribution. Therefore, X^x and Y^x have the same distribution. Letting P^x denote the common distribution, letting X° denote the identity mapping on the space of càdlàg paths from $[0, \infty)$ to $\mathbb{R}^p$ and letting ε_x denoting the Dirac measure in x, we therefore obtain by (4.1.10) of [51] for any $B \in \mathcal{B}_p$ and $t \geq 0$ that

$$\begin{aligned}(P_t 1_B)(x) &= \int (P_t 1_B)(y)\,\mathrm{d}\varepsilon_x(y) = P^x(X_t^\circ \in B) \\ &= \int (Q_t 1_B)(y)\,\mathrm{d}\varepsilon_x(y) = (Q_t 1_B)(x).\end{aligned} \tag{7.41}$$

We may now conclude that $(P_t 1_B)(x) = (Q_t 1_B)(x)$ for all $B \in \mathcal{B}_p$, $x \in \mathbb{R}^p$ and $t \geq 0$, and therefore, $(P_t) = (Q_t)$, as desired. □

Proof of Lemma 7.5.1. By Lemma 7.8.1, there exists a solution to (7.24), and the solution is a Feller process with a Feller semigroup (P_t) such that (P_t) has a generator whose domain includes $C_0^2(\mathbb{R}^p)$ and such that the generator on this space is given by (7.33).

We need to argue that all solutions to (7.24), independent of the initial distribution and the probability space on which the solution exists, are Feller processes with semigroup (P_t). To this end, let X be a solution of (7.24). By Lemma 7.8.1, the solution is a Feller process, and has a Feller semigroup (Q_t) whose generator agrees

with the generator of (P_t) on the set $C_0^2(\mathbb{R}^p)$. By Lemma 7.8.2, $(P_t) = (Q_t)$, and the lemma is proven. □

Next, we turn our attention to Theorem 7.5.3. In order to prove this result, we first state some technical lemmas.

Lemma 7.8.3. *Let E be a neighborhood of zero in $\mathbb{R}^p$. On $C_0^2(\mathbb{R}^p)$, the generator of the semigroup of (7.24) may be rewritten as*

$$\begin{aligned} Af(x) = &\sum_{i=1}^{p} \beta_i(x) \frac{\partial f}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (a(x)Ca(x)^t)_{ij} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &+ \int f(x+y) - f(x) - 1_E(y) \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) y_i \, \mathrm{d}T_x(\nu)(y), \end{aligned} \tag{7.42}$$

where $T_x : \mathbb{R}^d \to \mathbb{R}^p$ is defined by $T_x(y) = a(x)y$, and

$$\beta_i(x) = \sum_{j=1}^{d} a_{ij}(x)\alpha_j + \int (1_{T_x^{-1}(E)}(y) - 1_D(y)) \sum_{j=1}^{d} a_{ij}(x) y_j \, \mathrm{d}\nu(y), \tag{7.43}$$

whenever the integrals are well-defined and finite. This finiteness condition is in particular satisfied if E is bounded.

Proof. Assume that the integrals

$$\int f(x+y) - f(x) - 1_E(y) \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) y_i \, \mathrm{d}T_x(\nu)(y) \qquad \text{and} \tag{7.44}$$

$$\int (1_{T_x^{-1}(E)}(y) - 1_D(y)) \sum_{j=1}^{d} a_{ij}(x) y_j \, \mathrm{d}\nu(y) \tag{7.45}$$

are well-defined and finite. We then obtain

$$\begin{aligned} &\int f(x + a(x)y) - f(x) - 1_D(y) \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) \sum_{j=1}^{d} a_{ij}(x) y_j \, \mathrm{d}\nu(y) \\ &= \int f(x+y) - f(x) - 1_E(y) \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) y_i \, \mathrm{d}T_x(\nu)(y) \\ &+ \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) \int (1_{T_x^{-1}(E)}(y) - 1_D(y)) \sum_{j=1}^{d} a_{ij}(x) y_j \, \mathrm{d}\nu(y), \end{aligned} \tag{7.46}$$

which, when substituted in (7.33), yields the desired expression for the generator. It remains to prove that (7.44) and (7.45) are well-defined finite in the case where E is

bounded. As regards (7.44), note that by continuity of T_x, $T_x^{-1}(E)$ is a neighborhood of zero in $\mathbb{R}^d$. $T_x^{-1}(E)$ may be unbounded, but we nonetheless have

$$\begin{aligned}&\int \left| f(x+y) - f(x) - 1_E(y) \sum_{i=1}^p \frac{\partial f}{\partial x_i}(x) y_i \right| \mathrm{d}T_x(\nu)(y) \\ &= \int \left| f(x+T_x(y)) - f(x) - 1_E(T_x(y)) \sum_{i=1}^p \frac{\partial f}{\partial x_i}(x) T_x(y)_i \right| \mathrm{d}\nu(y). \end{aligned} \tag{7.47}$$

Here, the mapping $y \mapsto f(x+T_x(y)) - f(x)$ is bounded as $f \in C_0^2(\mathbb{R}^p)$, and the mapping $y \mapsto 1_E(T_x(y)) \sum_{i=1}^p \frac{\partial f}{\partial x_i}(x) T_x(y)_i$ is bounded as E is bounded. Therefore, the integrand is bounded. And a first order Taylor expansion shows that on $T_x^{-1}(E)$, the integrand is bounded by $y \mapsto C\|y\|^2$ for some $C > 0$. As $T_x^{-1}(E)$ is a neighborhood of zero in $\mathbb{R}^d$, we conclude that the integrability properties of ν yields that the integral is finite. As for (7.45), we obtain

$$\begin{aligned}&\int (1_{T_x^{-1}(E)}(y) - 1_D(y)) \left| \sum_{i=1}^p \frac{\partial f}{\partial x_i}(x) \sum_{j=1}^d a_{ij}(x) y_j \right| \mathrm{d}\nu(y) \\ &= \int 1_{T_x^{-1}(E)\setminus D}(y) \left| \sum_{i=1}^p \frac{\partial f}{\partial x_i}(x) T_x(y)_i \right| \mathrm{d}\nu(y) \\ &- \int 1_{D\setminus T_x^{-1}(E)}(y) \left| \sum_{i=1}^p \frac{\partial f}{\partial x_i}(x) \sum_{j=1}^d a_{ij}(x) y_j \right| \mathrm{d}\nu(y), \end{aligned} \tag{7.48}$$

where again, both the final integrals are finite due to the integrability properties of ν, and so the former integral is finite as well. This concludes the proof. □

Lemma 7.8.4. *Assume that X and Y are two Feller processes with generators whose domain both contain $C_0^2(\mathbb{R}^p)$ such that the generators on this set are of the type (7.33). If the generators are equal on $C_0^2(\mathbb{R}^p)$ the initial distributions of X and Y are equal, then X and Y have the same distribution.*

Proof. Let (P_t) and (Q_t) be Feller transition semigroups of X and Y, respectively. By Lemma 7.8.2, $(P_t) = (Q_t)$, yielding by Theorem 4.1.1 of [51] that X and Y have the same distribution. □

Lemma 7.8.5. *Fix $x \in \mathbb{R}^p$ and let D be a bounded neighborhood of zero in $\mathbb{R}^p$. Let $a, \tilde{a} \in \mathbb{R}^p$ and $b, \tilde{b} \in \mathbb{M}(p,p)$, and let ν and $\tilde{\nu}$ be two measures on $\mathbb{R}^p$ such that $x \mapsto \min\{1, \|x\|^2\}$ is integrable with respect to ν and $\tilde{\nu}$. Consider two linear*

functionals A and $\tilde{A}$ from $C_0^2(\mathbb{R}^p)$ to $\mathbb{R}$, where A is given by

$$Af = \sum_{i=1}^{p} a_i \frac{\partial f}{\partial x_i}(x) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} b_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j}(x) + \int f(x+y) - f(x) - 1_D(y)\sum_{i=1}^{p}\frac{\partial f}{\partial x_i}(x) y_i \,\mathrm{d}\nu(y), \tag{7.49}$$

and $\tilde{A}$ is given by the same expression, with $\tilde{a}$, $\tilde{b}$ and $\tilde{\nu}$ substituted for a, b and ν. It then holds that $A = \tilde{A}$ if and only if $a = \tilde{a}$, $b = \tilde{b}$ and $\nu = \tilde{\nu}$ on $\mathbb{R}^p \setminus \{0\}$.

Proof. It is immediate that if $a = \tilde{a}$, $b = \tilde{b}$ and $\nu = \tilde{\nu}$ on $\mathbb{R}^p \setminus \{0\}$, then $A = \tilde{A}$. We need to prove the converse. Thus, assume that $A = \tilde{A}$. Fix a bounded neighborhood B of x in $\mathbb{R}^p$. Assume that B contains the open ball in the Euclidean metric centered at x with radius $\delta > 0$. Using approximate units such as defined in [63], we may for $0 < \gamma < 1$ construct a family of mappings $(f_\gamma) \subseteq C_0^2(\mathbb{R}^p)$ with the following properties: f_γ is bounded by 1, f_γ converges pointwise to 1_B as γ tends to zero, and for $\gamma \leq \gamma_0$, where γ_0 is some positive number, f_γ is constant and equal to one on the open ball in the Euclidean metric centered at x with radius $\delta(1-\gamma)$. Now consider $\gamma \leq \min\{\gamma_0, 1/2\}$, and consider $y \in \mathbb{R}^p$ with $\|y\|_2 < \delta/2$. In particular, $\|y\|_2 \leq \delta(1-\gamma)$, yielding $f_\gamma(x+y) = f_\gamma(x) = 1$. Therefore, for such γ, we obtain

$$\begin{aligned} Af_\gamma &= \int f_\gamma(x+y) - f_\gamma(x)\,\mathrm{d}\nu(y) = \int 1_{(\|y\|_2 \geq \delta/2)}(f_\gamma(x+y) - f_\gamma(x))\,\mathrm{d}\nu(y) \\ &= \int 1_{(\|y\|_2 \geq \delta/2)}(f_\gamma(x+y) - 1)\,\mathrm{d}\nu(y), \end{aligned} \tag{7.50}$$

and similarly, $\tilde{A}f_\gamma = \int 1_{(\|y\|_2 \geq \delta/2)}(f_\gamma(x+y)-1)\,\mathrm{d}\tilde{\nu}(y)$. As $x \mapsto \{1, \|x\|^2\}$ is integrable with respect to ν and $\tilde{\nu}$, both these measures are bounded on $\{y \in \mathbb{R}^p \mid \|y\|_2 \geq \delta/2\}$. Therefore, we may apply the dominated convergence theorem and obtain

$$\begin{aligned} \lim_{\gamma\to 0} Af_\gamma &= \lim_{\gamma\to 0}\int 1_{(\|y\|_2 \geq \delta/2)}(f_\gamma(x+y) - 1)\,\mathrm{d}\nu(y) \\ &= \int 1_{(\|y\|_2 \geq \delta/2)}(1_B(x+y) - 1)\,\mathrm{d}\nu(y) \\ &= \int 1_B(x+y) - 1\,\mathrm{d}\nu(y) = -\int 1_{B^c}(x+y)\,\mathrm{d}\nu(y), \end{aligned} \tag{7.51}$$

and similarly, $\lim_{\gamma\to 0}\tilde{A}f_\gamma = -\int 1_{B^c}(x+y)\,\mathrm{d}\tilde{\nu}(y)$. We thus obtain

$$\int 1_{B^c}(x+y)\,\mathrm{d}\nu(y) = -\lim_{\gamma\to 0} Af_\gamma = -\lim_{\gamma\to 0}\tilde{A}f_\gamma = \int 1_{B^c}(x+y)\,\mathrm{d}\tilde{\nu}(y). \tag{7.52}$$

As B was an arbitrary bounded neighborhood of x, we conclude that ν and $\tilde{\nu}$ agree on all sets of the form B^c where B is a bounded neighborhood of zero. Therefore,

$\nu = \tilde{\nu}$ on $\mathbb{R}^p \setminus \{0\}$. This implies that for all $f \in C_0^2(\mathbb{R}^p)$, we have

$$\sum_{i=1}^{d}(a_i - \tilde{a_i})\frac{\partial f}{\partial x_i}(x) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(b_{ij} - \tilde{b}_{ij})\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = 0. \tag{7.53}$$

Fix $i \leq p$. Again applying the approximation results of Chapter 2 of [63], there exists $f \in C_0^2(\mathbb{R}^p)$ such that $f(y) = y_i$ in a neighborhood of x, implying $a_i - \tilde{a}_i = 0$. As i was arbitrary, we obtain $a = \tilde{a}$. This implies that for all $f \in C_0^2(\mathbb{R}^p)$, we have

$$\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(b_{ij} - \tilde{b}_{ij})\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = 0. \tag{7.54}$$

Fixing $i, j \leq p$, by Chapter 2 of [63], there exists a function $f \in C_0^2(\mathbb{R}^p)$ such that $f(y) = y_i y_j$ in a neighborhood of x, implying $b_{ij} - \tilde{b}_{ij} = 0$. This completes the proof. □

Note that in the statement of Lemma 7.8.5, the measures ν and $\tilde{\nu}$ are not required to be Lévy measures, as we do not require that the measures assign measure zero to $\{0\}$. This will be important, as we in the proof of Theorem 7.5.3 will use the lemma for linear transformations of Lévy measures. Such measures retain their integrability properties, but may assign non-zero measure to $\{0\}$ when the linear transformation is non-injective.

Proof of Theorem 7.5.3. Fix a bounded neighborhood D of zero in $\mathbb{R}^d$, a bounded neighborhood $\tilde{D}$ of zero in $\mathbb{R}^{\tilde{d}}$ and a bounded neighborhood E of zero in $\mathbb{R}^p$. Assume that Z has D-characteristics (α, C, ν) and that $\tilde{Z}$ has $\tilde{D}$-characteristics $(\tilde{\alpha}, \tilde{C}, \tilde{\nu})$. For $x \in \mathbb{R}^p$, define $T_x^a : \mathbb{R}^d \to \mathbb{R}^p$ by $T_x^a(y) = a(x)y$ and $T_x^{\tilde{a}} : \mathbb{R}^{\tilde{d}} \to \mathbb{R}^p$ by $T_x^{\tilde{a}}(y) = \tilde{a}(x)y$. Also define

$$\beta_i(x) = \sum_{j=1}^{d} a_{ij}(x)\alpha_j + \int (1_{(T_x^a)^{-1}(E)}(y) - 1_D(y))\sum_{j=1}^{d} a_{ij}(x)y_j \,\mathrm{d}\nu(y) \tag{7.55}$$

$$\tilde{\beta}_i(x) = \sum_{j=1}^{\tilde{d}} \tilde{a}_{ij}(x)\tilde{\alpha}_j + \int (1_{(T_x^{\tilde{a}})^{-1}(E)}(y) - 1_{\tilde{D}}(y))\sum_{j=1}^{\tilde{d}} \tilde{a}_{ij}(x)y_j \,\mathrm{d}\tilde{\nu}(y). \tag{7.56}$$

Let $A : C_0^2(\mathbb{R}^p) \to C_0(\mathbb{R}^p)$ be given by (7.42), except with T_x exchanged with T_x^a, and let $\tilde{A} : C_0^2(\mathbb{R}^p) \to C_0(\mathbb{R}^p)$ be given similarly, except with β, a, C, T_x^a, ν, D and α exchanged by $\tilde{\beta}$, $\tilde{a}$, $\tilde{C}$, $T_x^{\tilde{a}}$, $\tilde{\nu}$, $\tilde{D}$ and $\tilde{\alpha}$. By our assumptions and Lemma 7.8.3, $A = \tilde{A}$. As a consequence, by the uniqueness result of Lemma 7.8.5, we find that for all $x \in \mathbb{R}^p$ and $i \leq p$, we have

$$\beta_i(x) = \tilde{\beta}_i(x), \tag{7.57}$$

$$a(x)Ca(x)^t = \tilde{a}(x)\tilde{C}\tilde{a}(x)^t, \tag{7.58}$$

$$T_x^a(\nu) = T_x^{\tilde{a}}(\tilde{\nu}). \tag{7.59}$$

We will use these equalities to obtain equality of the postintervention distributions. To this end, now assume that X_0 and Y_0 have the same distribution. We need to show that the postintervention distributions of doing $X^m := \zeta$ in (7.25) and doing $Y^m := \zeta$ in (7.26) are equal for all m and ζ. To this end, fix $k \geq 1$ and define two mappings

$$\rho_k : \mathbb{M}(p,k) \to \mathbb{M}(p,k) \tag{7.60}$$

$$\tau_k : \mathbb{M}(k,p) \to \mathbb{M}(k,p) \tag{7.61}$$

with ρ_k being the mapping substituting the entries on the m'th row with zeroes, and τ_k being the mapping substituting the entries on the m'th column with zeroes. We then find that the postintervention SDEs for doing $X^m := \zeta$ and $Y^m := \zeta$ in (7.25) and (7.26), respectively, are

$$X_t^i = (X_0^*)^i + \sum_{j=1}^{d} \int_0^t b_{ij}(X_s)\,\mathrm{d}Z_s^j \tag{7.62}$$

$$Y_t^i = (Y_0^*)^i + \sum_{j=1}^{d} \int_0^t \tilde{b}_{ij}(Y_s)\,\mathrm{d}\tilde{Z}_s^j, \tag{7.63}$$

for $i \leq p$, where $b : \mathbb{R}^p \to \mathbb{M}(p,d)$ and $\tilde{b} : \mathbb{R}^p \to \mathbb{M}(p,\tilde{d})$ are given by the expressions $b(x) = \rho_d(a(x))$ and $\tilde{b} = \rho_{\tilde{d}}(\tilde{a}(x))$, and $(X_0^*)^i$ and $(Y_0^*)^i$ are equal to X_0^i and Y_0^i, respectively, for $i \neq m$, and equal to ζ on the m'th coordinate. By Lemma 7.8.1, the distribution of the first process has a generator B which on $C_0^2(\mathbb{R}^p)$ is equal to

$$\begin{aligned} Bf(x) = &\sum_{i=1}^{p} \gamma_i(x) \frac{\partial f}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (b(x)Cb(x)^t)_{ij} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &+ \int f(x+y) - f(x) - 1_E(y) \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) y_i \,\mathrm{d}T_x^b(\nu)(y), \end{aligned} \tag{7.64}$$

and the distribution of the second process has a generator $\tilde{B}$ which on $C_0^2(\mathbb{R}^p)$ is equal to

$$\begin{aligned} \tilde{B}f(x) = &\sum_{i=1}^{p} \tilde{\gamma}_i(x) \frac{\partial f}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} (\tilde{b}(x)\tilde{C}\tilde{b}(x)^t)_{ij} \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &+ \int f(x+y) - f(x) - 1_E(y) \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) y_i \,\mathrm{d}T_x^{\tilde{b}}(\tilde{\nu})(y), \end{aligned} \tag{7.65}$$

where

$$\gamma_i(x) = \sum_{j=1}^{d} b_{ij}(x)\alpha_j + \int (1_{(T_x^b)^{-1}(E)}(y) - 1_D(y)) \sum_{j=1}^{d} b_{ij}(x) y_j \,\mathrm{d}\nu(y), \tag{7.66}$$

$$\tilde{\gamma}_i(x) = \sum_{j=1}^{\tilde{d}} \tilde{b}_{ij}(x)\tilde{\alpha}_j + \int (1_{(T_x^{\tilde{b}})^{-1}(E)}(y) - 1_{\tilde{D}}(y)) \sum_{j=1}^{\tilde{d}} \tilde{b}_{ij}(x) y_j \,\mathrm{d}\tilde{\nu}(y). \tag{7.67}$$

As X_0 and Y_0 have the same distribution, we find that X_0^* and Y_0^* have the same distributions as well. Therefore, Lemma 7.8.4 shows that in order to prove the desired result, it suffices to show that $B = \tilde{B}$ on $C_0^2(\mathbb{R}^p)$. To this end, note that for all $y \in \mathbb{R}^d$, $T_x^b(y) = \rho_d(a(x))y = \rho_1(a(x)y) = \rho_1(T_x^a(y))$, and similarly, $T_x^{\tilde{b}}(y) = \rho_1(T_x^{\tilde{a}}(y))$. Therefore, (7.59) implies that for all $x \in \mathbb{R}^p$,

$$\begin{aligned} T_x^b(\nu) &= (\rho_1 \circ T_x^a)(\nu) = \rho_1(T_x^a(\nu)) \\ &= \rho_1(T_x^{\tilde{a}}(\nu)) = (\rho_1 \circ T_x^{\tilde{a}})(\tilde{\nu}) = T_x^{\tilde{b}}(\tilde{\nu}). \end{aligned} \tag{7.68}$$

Also,

$$\begin{aligned} b(x)Cb(x)^t &= \rho_d(a(x))C\rho_d(a(x))^t = \rho_d(a(x)C)\tau_d(a(x)^t) \\ &= \tau_p(\rho_d(a(x)C)a(x)^t) = \tau_p(\rho_p(a(x)Ca(x)^t)), \end{aligned} \tag{7.69}$$

meaning that $b(x)Cb(x)^t$ is equal to $a(x)Ca(x)^t$ with zeroes substituted for the entries on the m'th row and column. Similarly, $\tilde{b}(x)C\tilde{b}(x)^t = \tau_p(\rho_p(\tilde{a}(x)C\tilde{a}(x)^t))$. As a consequence, (7.58) allows us to conclude that for all $x \in \mathbb{R}^p$,

$$b(x)Cb(x)^t = \tilde{b}(x)\tilde{C}\tilde{b}(x)^t. \tag{7.70}$$

Finally, note that $\gamma_m(x) = 0 = \tilde{\gamma}_m(x)$ and for $i \neq m$, we have

$$\begin{aligned} \gamma_i(x) &= \sum_{j=1}^d b_{ij}(x)\alpha_j + \int (1_{(T_x^b)^{-1}(E)}(y) - 1_D(y)) \sum_{j=1}^d b_{ij}(x)y_j \,\mathrm{d}\nu(y) \\ &= \sum_{j=1}^d a_{ij}(x)\alpha_j + \int (1_{(T_x^a)^{-1}(\rho_1^{-1}(E))}(y) - 1_D(y)) \sum_{j=1}^d a_{ij}(x)y_j \,\mathrm{d}\nu(y), \end{aligned} \tag{7.71}$$

and similarly for $\tilde{\gamma}$. In particular, as $\rho_1^{-1}(E)$ is a neighborhood of zero in $\mathbb{R}^p$, calculating backwards, this implies in particular by Lemma 7.8.1 that the latter integral is finite even though $\rho_1^{-1}(E)$ may not be bounded. Again applying Lemma 7.8.1, (7.57) also holds with E substituted with $\rho_1^{-1}(E)$, and so (7.71) shows that

$$\gamma_i(x) = \tilde{\gamma}_i(x). \tag{7.72}$$

Combining (7.68), (7.70) and (7.72), Lemma 7.8.5 shows that B and $\tilde{B}$ agree on $C_0^2(\mathbb{R}^p)$, and thus Lemma 7.8.4 yields that the postintervention distributions are equal. □

7.9 Proof of WCLI properties

In this section, we prove Theorem 7.6.1. To this end, we first state two lemmas. We remark that the calculation of the characteristics in the proof of Lemma 7.9.1 is similar to the results given as Proposition IX.5.3 of [83] and Lemma 2.5 of [88].

Lemma 7.9.1. *Let K be a d-dimensional predictable and locally bounded process, and define $Y_t = \sum_{j=1}^d \int_0^t K_s^j \,\mathrm{d}Z_s^j$. Letting (B^Z, C^Z, ν^Z) be the h^d-characteristics of Z, it holds that the h^1-characteristics (B^Y, C^Y, ν^Y) of Y are given by*

$$B_t^Y = \sum_{j=1}^d \int_0^t K_s^j \,\mathrm{d}(B^Z)_s^j + ((h^1 \circ H - H \circ h^d) * \nu^Z)_t \tag{7.73}$$

$$C_t^Y = \sum_{j=1}^d \sum_{k=1}^d K_s^j K_s^k \,\mathrm{d}(C^Z)_s^{jk} \tag{7.74}$$

$$\nu^Y(\omega, A) = \int_{\mathbb{R}_+ \times \mathbb{R}^d} 1_A(t, H(x)_t(\omega)) \,\mathrm{d}\nu^Z(\omega, \,\mathrm{d}t, \,\mathrm{d}x), \tag{7.75}$$

where $A \in \mathcal{B}$ and $H(x)_t(\omega) = \sum_{j=1}^d K_t^j(\omega) x_j$.

Proof. We begin by calculating an expression for the first characteristic, B^Y. To do so, we identify the predictable finite variation part of the special semimartingale $Y_t - \sum_{0<s\leq t} \Delta Y_s - h^1(\Delta Y_s)$. Note that $(\omega, t, x) \mapsto H(x)_t(\omega)$ is predictable. By the definition of B^Z, there exists a d-dimensional local martingale M with the property that $Z_t = Z_0 + B_t^Z + M_t + \sum_{0<s\leq t} \Delta Z_s - h^d(\Delta Z_s)$. We then obtain the decomposition $Y_t = A_t + \sum_{j=1}^d \int_0^t K_s^j \,\mathrm{d}M_s^j$, where the latter is a local martingale and

$$\begin{aligned} A_t &= \sum_{j=1}^d \int_0^t K_s^j \,\mathrm{d}(B^Z)_s^j + \sum_{j=1}^d \sum_{0<s\leq t} K_s^j (\Delta Z_s^j - h_j^d(\Delta Z)_s) \\ &= \sum_{j=1}^d \int_0^t K_s^j \,\mathrm{d}(B^Z)_s^j + \sum_{0<s\leq t} H(\Delta Z_s)_s - H(h^d(\Delta Z_s))_s \\ &= \sum_{j=1}^d \int_0^t K_s^j \,\mathrm{d}(B^Z)_s^j + ((H - H \circ h^d) * \mu^Z)_t, \end{aligned} \tag{7.76}$$

understanding that $H - H \circ h^d$ here denotes $(\omega, t, x) \mapsto H(x)_t(\omega) - H(h^d(x))_t(\omega)$, which is a predictable function, and the integral with respect to μ^Z is finite by taking absolute values and calculating backwards. With similar notation, we also obtain

$$\begin{aligned} \sum_{0<s\leq t} \Delta Y_s - h^1(\Delta Y_s) &= \sum_{0<s\leq t} \sum_{j=1}^d K_s^j \Delta Z_s^j - h^1\left(\sum_{j=1}^d K_s^j \Delta Z_s^j\right) \\ &= \sum_{0<s\leq t} H(\Delta Z_s) - h^1(H(\Delta Z_s)) = ((H - h^1 \circ H) * \mu^Z)_t. \end{aligned} \tag{7.77}$$

Therefore, we obtain $Y_t - \sum_{0<s\leq t} \Delta Y_s - h^1(\Delta Y_s) = \tilde{A}_t + \sum_{j=1}^d \int_0^t K_s^j \,\mathrm{d}M_s^j$, where $\tilde{A}$

is the finite variation process given by

$$\tilde{A}_t = \sum_{j=1}^{d} \int_0^t K_s^j \, \mathrm{d}(B^Z)_s^j + ((h^1 \circ H - H \circ h^d) * \mu^Z)_t. \tag{7.78}$$

Now define $B_t^Y = \sum_{j=1}^d \int_0^t K_s^j \, \mathrm{d}(B^Z)_s^j + ((h^1 \circ H - H \circ h^d) * \nu^Z)_t$, where the integral with respect to ν^Z is well-defined as integrability with respect to μ^Z implies integrability with respect to ν^Z. As $h^1 \circ H - H \circ h^d$ is a predictable function, the latter term is predictable. And as B^Z is predictable, the process $\int_0^t K_s^j \, \mathrm{d}(B^Z)_s^j$ only jumps at predictable times T, and the jump is $K_T^j \Delta(B^Z)_T^j$, which is $\mathcal{F}_{T-}$ measurable by Corollary 3.23 of [66]. Therefore, Theorem 3.33 of [66] shows that $\int_0^t K_s^j \, \mathrm{d}(B^Z)_s^j$ is predictable, and thus B^Y is predictable. Thus, B^Y is the predictable finite variation part of the process $Y_t - \sum_{0<s\leq t} \Delta Y_s - h^1(\Delta Y_s)$ and is therefore the first characteristic of Y. As regards the process C^Y, note that by Theorem 9.3 of [66], $Y_t^c = \sum_{j=1}^d \int_0^t K_s^j \, \mathrm{d}(Z^j)_s^c$. Thus, we immediately obtain $C_t^Y = \sum_{j=1}^d \sum_{k=1}^d K_s^j K_s^k \, \mathrm{d}(C^Z)_s^{jk}$. It remains to calculate the third characteristic. For all $A \in \mathcal{B}_+ \otimes \mathcal{B}$, we have

$$\begin{aligned} \mu^Y(\omega, A) &= \sum_{0<t} 1_A(t, \Delta Y_t(\omega)) = \sum_{0<t} 1_A\left(t, \sum_{j=1}^d K_s^j(\omega) \Delta Z_t^j(\omega)\right) \\ &= \sum_{0<t} 1_A(t, H(\Delta Z(\omega)_t)_t(\omega)) = \int_{\mathbb{R}_+ \times \mathbb{R}^d} 1_A(t, H(x)_t(\omega)) \, \mathrm{d}\mu^Z(\omega, \, \mathrm{d}t, \, \mathrm{d}x). \end{aligned} \tag{7.79}$$

Now define $\nu^Y(\omega, A) = \int_{\mathbb{R}_+ \times \mathbb{R}^d} 1_A(t, H(x)_t(\omega)) \, \mathrm{d}\nu^Z(\omega, \, \mathrm{d}t, \, \mathrm{d}x)$. We wish to argue that ν^Y is the compensator of the jump measure of Y. To this end, we first show that ν^Y is predictable. By Section VI.16 of [143], $\mathcal{P}$ is generated by the family $[\![T, \infty[\![$ for T a predictable stopping time. Therefore, $\tilde{\mathcal{P}}_1$ is generated by sets of the form $[\![T, \infty[\![\times C$, where $C \in \mathcal{B}$. By a monotone convergence argument, we then obtain that in order to prove that ν^Y is predictable, it suffices to show that $1_{[\![T,\infty[\![\times C} * \nu^Y$ is predictable for all predictable stopping times T and all $C \in \mathcal{B}$. To do so, fix a predictable stopping time T and a set $C \in \mathcal{B}$, we then have

$$\begin{aligned} (1_{[\![T,\infty[\![\times C} * \nu^Y)_t(\omega) &= \int_{[0,t] \times \mathbb{R}^d} 1_{[T(\omega),\infty) \times C}(t, H(x)_t(\omega)) \, \mathrm{d}\nu^Z(\omega, \, \mathrm{d}t, \, \mathrm{d}x) \\ &= \int_{[0,t] \times \mathbb{R}^d} 1_{[\![T,\infty[\![\times C}(\omega, t, H(x)_t(\omega)) \, \mathrm{d}\nu^Z(\omega, \, \mathrm{d}t, \, \mathrm{d}x). \end{aligned} \tag{7.80}$$

Now note that the mapping $(\omega, t, x) \mapsto H(x)_t(\omega)$ is $\mathcal{P} \otimes \mathcal{B}_d$-$\mathcal{B}$ measurable. From this, we conclude that $(\omega, t, x) \mapsto (\omega, t, H(x)_t(\omega))$ is $\mathcal{P} \otimes \mathcal{B}_d$-$\mathcal{P} \otimes \mathcal{B}$ measurable. As $[\![T, \infty[\![\times C \in \mathcal{P} \otimes \mathcal{B}$, $(\omega, t, x) \mapsto 1_{[\![T,\infty[\![\times C}$ is $\mathcal{P} \otimes \mathcal{B}$-$\mathcal{B}$ measurable. We conclude that $(\omega, t, x) \mapsto 1_{[\![T,\infty[\![\times C}(\omega, t, H(x)_t(\omega))$ is $\tilde{\mathcal{P}}_d$-$\mathcal{B}$ measurable, thus a predictable function, so as ν^Z is predictable, $1_{[\![T,\infty[\![\times C} * \nu^Y$ is predictable, so ν^Y is predictable. It remains

to prove that $E(W * \nu^Y)_\infty = E(W * \mu^Y)_\infty$ for all nonnegative bounded predictable functions W. Again, it suffices to consider predictable functions of the form $1_{[\![T,\infty[\![\times C} \cdot$ However, rewriting the integrand as a predictable function as in (7.80), this follows immediately from the fact that ν^Z is the compensator of μ^Z. □

Lemma 7.9.2. *Let X be the solution to (7.8). Assume that Z is a Lévy process and assume that X^i is locally unaffected by X^m in (7.8). Let (B, C, ν) be the semimartingale characteristics of X^i. Let $(\mathcal{F}_t^{-m})_{t\geq 0}$ be the usual augmentation of the filtration induced by the processes $X^1, \ldots, X^p$ excluding X^m. Then B, C and ν are $(\mathcal{F}_t^{-m})$ predictable.*

Proof. By Theorem II.4.15 of [83], Z being a Lévy process implies the existence of a deterministic version (B^Z, C^Z, ν^Z) of the characteristics of Z. In particular, B^Z, C^Z and ν^Z are all $(\mathcal{F}_t^{-m})$ predictable. And by the assumptions we have made, $X_t^i = x_0^i + \sum_{j=1}^d \int_0^t K_s^j \, \mathrm{d}Z_s^j$, where $K_s^j = a_{ij}(Y_{s-})$ and a_{ij} does not depend on the m'th coordinate for all j. In particular, K_s^j is $(\mathcal{F}_t^{-m})$ predictable and locally bounded. With $H(x)_t(\omega) = \sum_{j=1}^d K_t^j(\omega) x_j$, we then find that $(\omega, t, x) \mapsto H(x)_t(\omega)$ is a $(\mathcal{F}_t^{-m})$ predictable function. By (7.73) and Theorem 3.33 of [66], we then obtain that B is $(\mathcal{F}_t^{-m})$ predictable. As regards the second characteristic, (7.74) shows that C is continuous and $(\mathcal{F}_t^{-m})$ adapted, therefore $(\mathcal{F}_t^{-m})$ predictable. Finally, by the same argument as in the proof of Lemma 7.9.1, we find that for any $(\mathcal{F}_t^{-m})$ predictable stopping time and $C \in \mathcal{B}$, $1_{[\![T,\infty[\![\times C} * \nu$ is $(\mathcal{F}_t^{-m})$ predictable, and so ν is $(\mathcal{F}_t^{-m})$ predictable. □

Proof of Theorem 7.6.1. As we have assumed that X is a special semimartingale, this is immediate from Lemma 7.9.2. □

Acknowledgements. The authors would like to thank Marloes Maathuis for fruitful discussions and comments.

8

Intervention in Ornstein-Uhlenbeck SDEs

Alexander Sokol

2010 Mathematics Subject Classification. Primary 60G15.

Key words and phrases. Causality, Intervention, SDE, Ornstein-Uhlenbeck process, Stationary distribution.

Abstract. We introduce a notion of intervention for stochastic differential equations and a corresponding causal interpretation. For the case of the Ornstein-Uhlenbeck SDE, we show that the SDE resulting from a simple type of intervention again is an Ornstein-Uhlenbeck SDE. We discuss criteria for the existence of a stationary distribution for the solution to the intervened SDE. We illustrate the effect of interventions by calculating the mean and variance in the stationary distribution of an intervened process in a particularly simple case.

8.1 Introduction

Causal inference for continuous-time processes is a field in ongoing development. Similar to causal inference for graphical models, see [126], one of the primary objectives for causal inference for continuous-time processes is to identify the effect of an intervention given assumptions on the distribution and causal structure of the observed continuous-time process.

Several flavours of causal inference are available for continuous-time processes, see for example [54, 57, 130]. In this paper, we outline a causal interpretation of stochastic differential equations and a corresponding notion of intervention, we calculate the distribution of an intervened Ornstein-Uhlenbeck SDE, and we calculate analytical expressions for the mean and variance of the stationary distribution of the resulting process for particular examples of interventions.

8.2 Causal interpretation of SDEs

Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ satisfying the usual conditions, see [134] for the definition of this and other notions related to continuous-time stochastic processes. Let Z be a d-dimensional semimartingale and assume that $a : \mathbb{R}^p \to \mathbb{M}(p,d)$ is a Lipschitz mapping, where $\mathbb{M}(p,d)$ denotes the space of real $p \times d$ matrices. Consider the stochastic differential equation (SDE)

$$X_t^i = x_0^i + \sum_{j=1}^{d} \int_0^t a_{ij}(X_{s-}) \, \mathrm{d}Z_s^j, \qquad i \leq p. \tag{8.1}$$

By the Lipschitz property of a, it holds by Theorem V.7 of [134] that there exists a pathwisely unique solution to (8.1). The following definition yields a causal interpretation of (8.1) based on simple substitution and inspired by ideas outlined in Section 4.1 of [1].

Definition 8.2.1. Consider some $m \leq p$ and $c \in \mathbb{R}$. The $(p-1)$-dimensional intervened SDE arising from the intervention $X^m := c$ is defined to be

$$U_t^i = x_0^i + \sum_{j=1}^{d} \int_0^t b_{ij}(U_{s-}) \, \mathrm{d}Z_s^j \text{ for } i \leq p \text{ with } i \neq m, \tag{8.2}$$

where $b_{ij}(y_1, \ldots, y_{m-1}, y_{m+1}, \ldots, y_p) = a_{ij}(y_1, \ldots, c, \ldots, y_p)$, and the c is on the m'th coordinate. Letting U be the unique solution to the SDE and defining Y by putting $Y = (U^1, \ldots, U^{m-1}, c, U^{m+1}, \ldots, U^p)$, we refer to Y as the intervened process and write $(X|X^m := c)$ for Y.

By Theorem V.16 and Theorem V.5 of [134], the solutions to both (8.1) and (8.2) may be approximated by the Euler schemes for their respective SDEs. Making these approximations and applying Pearl's notion of intervention in an appropriate sense, see [126], we may interpret Definition 8.2.1 as intervening in the system (8.1) under the assumption that the driving semimartingales $Z^1, \ldots, Z^d$ are noise processes unaffected by interventions, while the processes $X^1, \ldots, X^p$ are affected by interventions. Note that the operation of making an intervention takes a p-dimensional SDE as its input and yields a $(p-1)$-dimensional SDE as its output, and this operation is crucially dependent on the coefficients in the SDE: These coefficients in a sense

corresponds to the directed acyclic graphs of [126]. A major benefit of causality in systems such as (8.1) as compared to the theory of [126] is the ability to capture feedback systems and interventions in such feedback systems.

As the solutions to (8.1) and (8.2) are defined on the same probability space, we may even consider the process $Y - X$, where $Y = (X|X^m := c)$, allowing us to calculate for example the variance of the effect of the intervention. As Y and X are never observed simultaneously in practice, however, we will concentrate on analyzing the differences between the laws of Y and X separately.

8.3 Intervention in Ornstein-Uhlenbeck SDEs

Recall that for an $\mathcal{F}_0$ measurable variable X_0 and for $A \in \mathbb{R}^p$, $B \in \mathbb{M}(p,p)$ and $\sigma \in \mathbb{M}(p,d)$, the Ornstein-Uhlenbeck SDE with initial value X_0, mean reversion level A, mean reversion speed B, diffusion matrix σ and d-dimensional driving noise is

$$X_t = X_0 + \int_0^t B(X_s - A)\,\mathrm{d}s + \sigma W_t, \tag{8.3}$$

where W is a d-dimensional $(\mathcal{F}_t)$ Brownian motion, see Section II.72 of [142]. The unique solution to this equation is

$$X_t = \exp(tB)\left(X_0 - \int_0^t \exp(-sB)BA\,\mathrm{d}s + \int_0^t \exp(-sB)\sigma\,\mathrm{d}W_s\right) \tag{8.4}$$

where the matrix exponential is defined by $\exp(A) = \sum_{n=0}^{\infty} A^n/n!$, see [68]. This is a Gaussian homogeneous Markov process with continuous sample paths. The following lemma shows that making an intervention in an Ornstein-Uhlenbeck SDE yields an SDE whose nontrivial coordinates solve another Ornstein-Uhlenbeck SDE.

Lemma 8.3.1. *Consider the Ornstein-Uhlenbeck SDE (8.3) with initial value x_0. Fix $m \leq p$ and $c \in \mathbb{R}$, and let X be the unique solution to (8.3). Furthermore, let $Y = (X|X^m := c)$ and let Y^{-m} be the $p-1$ dimensional process obtained by removing the m'th coordinate from Y. Let $\tilde{B}$ be the submatrix of B obtained by removing the m'th row and column of B, and assume that $\tilde{B}$ is invertible. Then Y^{-m} solves*

$$Y_t^{-m} = y_0 + \int_0^t \tilde{B}(Y_s^{-m} - \tilde{A})\,\mathrm{d}s + \tilde{\sigma}W_t, \tag{8.5}$$

where y_0 is obtained by removing the m'th coordinate from x_0, $\tilde{\sigma}$ is obtained by removing the m'th row of σ and $\tilde{A} = \alpha - \tilde{B}^{-1}\beta$, where α and β are obtained by removing the m'th coordinate from A and from the vector whose i'th component is $b_{im}(c - a_m)$, respectively, where b_{im} is the entry corresponding to the i'th row and the m'th column of B, and a_m is the m'th element of A.

Proof. By Definition 8.2.1, we have

$$Y_t^i = y_0 + \int_0^t b_{im}(c - a_m) + \sum_{j \neq m} b_{ij}(Y_s^j - a_j)\,\mathrm{d}s + \sum_{j=1}^{p} \sigma_{ij} W_t^j \tag{8.6}$$

for $i \neq m$. Note that for any vector y, the system of equations in $\tilde{a}$

$$b_{im}(c - a_m) + \sum_{j \neq m} b_{ij}(y_j - a_j) = \sum_{j \neq m} b_{ij}(y_j - \tilde{a}_j) \text{ for } i \neq m, \tag{8.7}$$

is equivalent to the system of equations

$$\sum_{j \neq m} b_{ij}\tilde{a}_j = \left(\sum_{j \neq m} b_{ij} a_j \right) - b_{im}(c - a_m) \text{ for } i \neq m. \tag{8.8}$$

Since we have assumed $\tilde{B}$ to be invertible, this system of equations has the unique solution $\tilde{A} = \tilde{B}^{-1}(\tilde{B}\alpha - \beta) = \alpha - \tilde{B}^{-1}\beta$. For $i \neq m$, we therefore obtain that $Y_t^i = y_0 + \int_0^t \sum_{j \neq m} b_{ij}(Y_s^j - \tilde{a}_j)\,\mathrm{d}s + \sum_{j=1}^{p} \sigma_{ij} W_t^j$, proving the result. □

Recall that a principal submatrix of a matrix is a submatrix with the same rows and columns removed. In words, Lemma 8.3.1 states that if a particular principal submatrix $\tilde{B}$ of the mean reversion speed is invertible, then making the intervention $X^m := c$ in an Ornstein-Uhlenbeck SDE results in a new Ornstein-Uhlenbeck SDE with mean reversion speed $\tilde{B}$ and modified mean reversion level involving the inverse of $\tilde{B}$. Now assume that an Ornstein-Uhlenbeck SDE is given such that the solution has a stationary initial distribution. A natural question to ask is what interventions will yield intervened processes where stationary initial distributions also exist. In the following, we consider this question.

Recall that a square matrix is called stable if its eigenvalues have negative real parts and semistable if its eigenvalues have nonpositive real parts, see [31]. Theorem 4.1 of [178] yields necessary and sufficient criteria for the existence of a stationary probability measure for the solution of (8.3). One criterion is expressed in terms of the controllability subspace of of the matrix pair (B, σ), which is the span of the columns in the matrices $\sigma, B\sigma, \ldots, B^{p-1}\sigma$. In the case where σ has full column span, meaning that the columns of σ span all of $\mathbb{R}^p$, the controllability subspace is all of $\mathbb{R}^p$, and Theorem 4.1 of [178] shows that the existence of a stationary probability measure is equivalent to B being stable. The case where σ is not required to have full column span is more involved.

In the following, we will restrict our attention to Ornstein-Uhlenbeck processes with σ having full column span. By Theorem 4.1 of [178], it then holds that there exists a stationary distribution if and only if B is stable. Furthermore, applying Theorem 2.4 and Theorem 2.12 of [80], it holds in the affirmative case that the stationary distribution is the normal distribution with mean μ and variance Γ solving $B\mu = BA$ and $\sigma\sigma^t + B\Gamma + \Gamma B^t = 0$. Note that as B is stable, zero is not an eigenvalue of B, thus

B is invertible and $\mu = A$. Also, stability of B yields that $\Gamma = \int_0^\infty e^{sB}\sigma\sigma^t e^{sB^t}\,\mathrm{d}s$. For the $(p-1)$-dimensional Ornstein-Uhlenbeck process resulting from an intervention according to Lemma 8.3.1, the diffusion matrix $\tilde{\sigma}$ is obtained by removing the m'th row of σ. As the columns of σ span $\mathbb{R}^p$, the columns of $\tilde{\sigma}$ span $\mathbb{R}^{p-1}$. Therefore, it also holds for the intervened process that there exists a stationary distribution if and only if the mean reversion speed is stable. We conclude that for diffusion matrices with full column span, the existence of stationary distributions for both the original and the intervened SDE is determined solely by stability of the mean reversion speed matrix B and the corresponding principal submatrices.

Consider a stable matrix B. It then holds that if all principal submatrices of B are stable, all interventions will preseve stability of the system. We are thus lead to the question of when a principal submatrix of a matrix is stable. That stability does not in general lead to stability of principal submatrices may be seen from the following example. Define B by putting

$$B = \begin{bmatrix} 1 & 7 \\ -1 & -3 \end{bmatrix}.$$

The matrix B has eigenvalues $-1 \pm i\sqrt{3}$ and is thus stable, while the principal submatrix obtained by removing the second row and second column trivially has the single eigenvalue 1 and thus is not stable, in fact not even semistable. Conversely, $-B$ has eigenvalues $1 \pm i\sqrt{3}$ and thus is neither stable nor semistable, while the principal submatrix obtained by removing the second row and second column of $-B$ is stable.

There are classes of matrices satisfying that all principal submatrices are stable. For example, by the inclusion principle for symmetric matrices, see Theorem 4.3.15 of [72], it follows that a principal submatrix of any symmetric stable matrix again is stable. In general, though, it is difficult to ensure that all principal submatrices are stable. However, there are criteria ensuring that all principal submatices are semistable. For example, Lemma 2.4 of [67] shows that if B is stable and sign symmetric, then all principal submatrices of B is semistable. Here, sign symmetry is a somewhat involved matrix criterion, it does however hold that any stable symmtric matrix also is sign symmetric. Furthermore, by Theorem 1 of [31], either of the follow three properties are also sufficient for having all principal submatrices being semistable:

1. $A - D$ is stable for all nonnegative diagonal D.

2. DA is stable for all positive diagonal D.

3. There is positive diagonal D such that $AD + DA^t$ is negative definite.

8.4 An example of a particular intervention

Consider now a three-dimensional Ornstein-Uhlenbeck process X with σ being the identity matrix of order three and upper diagonal mean reversion speed matrix B, and assume that the diagonal elements of B all are negative. As the diagonal elements of B in this case also are the eigenvalues, B is then stable, and all principal submatrices are stable as well. The interpretation of having B upper diagonal is that the levels of both X^1, X^2 and X^3 influence the average change in X^1, while only the levels of X^2 and X^3 influence the average change in X^2 and only X^3 influences the average change in X^3. Figure 8.4.1 illustrates this, as well as the changes to the dependence structure obtained by making interventions $X^2 := c$ or $X^3 := c$.

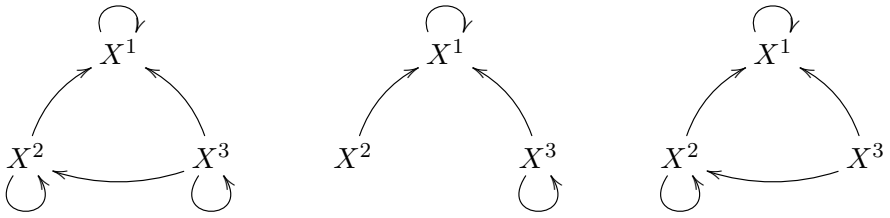


Figure 8.4.1: Graphical illustrations of the dependence structures of (X^1, X^2, X^3) (left), of the dependence when making the intervention $X^2 := c$ (middle) and of the dependence when making the intervention $X^3 := c$ (right).

We will investigate the details of what happens to the system when making the intervention $X^2 := c$ or $X^3 := c$. To this end, we calculate the mean and variance in the stationary distribution for the nontrivial coordinates in each of the intervened processes. Consider first the case of the intervention $X^2 := c$. Let μ and Γ denote the mean and variance in the stationary distribution after intervention. Applying Lemma 8.3.1, the SDE resulting from making this intervention is a two-dimensional Ornstein-Uhlenbeck SDE with mean reversion speed and mean reversion level

$$\begin{bmatrix} b_{11} & b_{13} \\ 0 & b_{33} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_1 \\ a_3 \end{bmatrix} - \begin{bmatrix} b_{11} & b_{13} \\ 0 & b_{33} \end{bmatrix}^{-1} \begin{bmatrix} b_{12}(c - a_2) \\ 0 \end{bmatrix}. \tag{8.9}$$

As we have

$$\begin{bmatrix} b_{11} & b_{13} \\ 0 & b_{33} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{b_{11}} & -\frac{b_{13}}{b_{11}b_{33}} \\ 0 & \frac{1}{b_{33}} \end{bmatrix}, \tag{8.10}$$

this immediately yields that

$$\mu = \begin{bmatrix} a_1 - \frac{b_{12}}{b_{11}}(c - a_2) \\ a_3 \end{bmatrix}. \tag{8.11}$$

As for the variance, recall that we have the representation

$$\Gamma = \int_0^\infty \exp\left(s \begin{bmatrix} b_{11} & b_{13} \\ 0 & b_{33} \end{bmatrix}\right) \exp\left(s \begin{bmatrix} b_{11} & 0 \\ b_{13} & b_{33} \end{bmatrix}\right) \mathrm{d}s. \tag{8.12}$$

In order to calculate this integral, first consider the case $b_{11} = b_{33}$. By Theorem 4.11 of [68], we in this case obtain

$$\exp\left(s \begin{bmatrix} b_{11} & b_{13} \\ 0 & b_{33} \end{bmatrix}\right) = e^{sb_{11}} \begin{bmatrix} 1 & sb_{13} \\ 0 & 1 \end{bmatrix}, \tag{8.13}$$

and similarly for the transpose. Applying that $\int_0^\infty x^\alpha e^{\beta x}\,\mathrm{d}x = \Gamma(\alpha+1)/(-\beta)^{\alpha+1}$ for all $\alpha > -1$ and $\beta < 0$, we conclude

$$\begin{aligned} \Gamma &= \int_0^\infty e^{2sb_{11}} \begin{bmatrix} 1 & sb_{13} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sb_{13} & 1 \end{bmatrix} \mathrm{d}s \\ &= \int_0^\infty e^{2sb_{11}} \begin{bmatrix} 1+s^2b_{13}^2 & sb_{13} \\ sb_{13} & 1 \end{bmatrix} \mathrm{d}s = \begin{bmatrix} -\frac{1}{2b_{11}} - \frac{b_{13}^2}{4b_{11}^3} & \frac{b_{13}}{4b_{11}^2} \\ \frac{b_{13}}{4b_{11}^2} & -\frac{1}{2b_{11}} \end{bmatrix}. \end{aligned} \tag{8.14}$$

In the case $b_{11} \neq b_{33}$, we put $\zeta = b_{13}/(b_{11} - b_{33})$ and Theorem 4.11 of [68] yields

$$\exp\left(s \begin{bmatrix} b_{11} & b_{13} \\ 0 & b_{33} \end{bmatrix}\right) = \begin{bmatrix} e^{sb_{11}} & \zeta(e^{sb_{11}} - e^{sb_{33}}) \\ 0 & e^{sb_{33}} \end{bmatrix}, \tag{8.15}$$

and we then obtain

$$\begin{aligned} &\exp\left(s \begin{bmatrix} b_{11} & b_{13} \\ 0 & b_{33} \end{bmatrix}\right) \exp\left(s \begin{bmatrix} b_{11} & 0 \\ b_{13} & b_{33} \end{bmatrix}\right) \\ &= \begin{bmatrix} e^{sb_{11}} & \zeta(e^{sb_{11}} - e^{sb_{33}}) \\ 0 & e^{sb_{33}} \end{bmatrix} \begin{bmatrix} e^{sb_{11}} & 0 \\ \zeta(e^{sb_{11}} - e^{sb_{33}}) & e^{sb_{33}} \end{bmatrix} \\ &= \begin{bmatrix} (1+\zeta^2)e^{2sb_{11}} - 2\zeta^2 e^{s(b_{11}+b_{33})} + \zeta^2 e^{2sb_{33}} & \zeta e^{s(b_{11}+b_{33})} - \zeta e^{2sb_{33}} \\ \zeta e^{s(b_{11}+b_{33})} - \zeta e^{2sb_{33}} & e^{2sb_{33}} \end{bmatrix}, \end{aligned} \tag{8.16}$$

implying that

$$\begin{aligned} \Gamma &= \begin{bmatrix} -\frac{(1+\zeta^2)}{2b_{11}} + \frac{2\zeta^2}{b_{11}+b_{33}} - \frac{\zeta^2}{2b_{33}} & -\frac{\zeta}{b_{11}+b_{33}} + \frac{\zeta}{2b_{33}} \\ -\frac{\zeta}{b_{11}+b_{33}} + \frac{\zeta}{2b_{33}} & -\frac{1}{2b_{33}} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{2b_{11}} - \zeta^2\left(\frac{1}{2b_{11}} + \frac{2}{b_{11}+b_{33}} - \frac{1}{2b_{33}}\right) & \frac{\zeta(b_{11}-b_{33})}{2b_{33}(b_{11}+b_{33})} \\ \frac{\zeta(b_{11}-b_{33})}{2b_{33}(b_{11}+b_{33})} & -\frac{1}{2b_{33}} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{2b_{11}} - \frac{b_{13}^2}{2b_{11}b_{33}(b_{11}+b_{33})} & \frac{b_{13}}{2b_{33}(b_{11}+b_{33})} \\ \frac{b_{13}}{2b_{33}(b_{11}+b_{33})} & -\frac{1}{2b_{33}} \end{bmatrix}. \end{aligned} \tag{8.17}$$

Note in particular that (8.17) also yields the correct result in the case $b_{11} = b_{33}$. Next, considering the intervention $X^3 := c$, we let ν and Σ denote the mean and variance

in the stationary distribution of the nontrivial coordinates after intervention. By Lemma 8.3.1, the result of making this interveniton is an Ornstein-Uhlenbeck SDE with mean reversion speed and mean reversion level

$$\begin{bmatrix} b_{11} & b_{12} \\ 0 & b_{22} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} - \begin{bmatrix} b_{11} & b_{12} \\ 0 & b_{22} \end{bmatrix}^{-1} \begin{bmatrix} b_{13}(c - a_3) \\ b_{23}(c - a_3) \end{bmatrix}, \tag{8.18}$$

yielding by calculations similar to the previous case that

$$\nu = \begin{bmatrix} a_1 - \left(\frac{b_{13}}{b_{11}} - \frac{b_{12}b_{23}}{b_{11}b_{22}}\right)(c - a_3) \\ a_2 - \frac{b_{23}}{b_{22}}(c - a_3) \end{bmatrix} \tag{8.19}$$

and

$$\Sigma = \begin{bmatrix} -\frac{1}{2b_{11}} - \frac{b_{12}^2}{2b_{11}b_{22}(b_{11}+b_{22})} & \frac{b_{12}}{2b_{22}(b_{11}+b_{22})} \\ \frac{b_{12}}{2b_{22}(b_{11}+b_{22})} & -\frac{1}{2b_{22}} \end{bmatrix}. \tag{8.20}$$

We have now calculated the mean and variance in the stationary distribution for both intervened processes. We next take a moment to interpret our results.

In the original system, all of X^1, X^2 and X^3 negatively influenced themselves, and in addition to this, X^2 influenced X^1 and X^3 influenced X^1 both directly and through its influence on X^2. Based on this, we would expect that making the intervention $X^2 := c$, the steady state of X^3 would not be changed, while the steady state of X^1 would change, depending on the level of influence b_{12} of X^2 on X^1. This is what we see in (8.11). When making the intervention $X^3 := c$, however, we obtain a change in the steady state of X^1 based both on the direct influence of X^3 on X^1, depending on b_{13}, but also on the indirect influence of X^3 on X^1 through X^2, depending also on b_{23} and b_{12}. Furthermore, the steady state of X^2 also changes. These results show themselves in (8.19).

As for the steady state variance, the changes resulting from interventions are in both cases of the same type, yielding moderately complicated analytical expressions, both independent of c. This implies that while we in most cases will be able to obtain any steady state mean for, say, X^1, by picking c suitably, the steady state variance can be influenced only by the type of intervention made, that is, on which parts of the system the interventions are made. Furthermore, by considering explicit formulas for the steady state variance in the original system, it may be seen that for example positive covariances may turn negative and vice versa when making interventions.

Acknowledgements. The development of the notion of intervention for SDEs is joint work with my thesis advisor, Niels Richard Hansen, whom I also thank for valuable discussions and advice.

9

Quantifying identifiability in independent component analysis

Alexander Sokol, Marloes H. Maathuis and Benjamin Falkeborg

2010 Mathematics Subject Classification. Primary 62F12; Secondary 62F35.

Key words and phrases. Independent Component Analysis, Identifiability, Kolmogorov norm, Contaminated distribution, Asymptotic statistics, Empirical process.

Abstract. We are interested in identifiability of the mixing matrix in the ICA model, when the error distribution is close to (but different from) Gaussian. In particular, we consider n independent samples from the ICA model $X = A\varepsilon$, where we assume that the coordinates of ε are independent and identically distributed according to a contaminated Gaussian distribution, and the amount of contamination is allowed to depend on n. We then investigate how identifiability of the mixing matrix depends on the amount of contamination. Our results suggest that identifiability becomes problematic if the amount of contamination decreases at rate $1/\sqrt{n}$ or faster.

9.1 Introduction

We consider the p-dimensional independent component analysis (ICA) model

$$X = A\varepsilon, \tag{9.1}$$

where A is a $p \times p$ mixing matrix, ε is a p-dimensional error (or source) variable with independent coordinates of mean zero, and X is a p-dimensional observational variable. Based on observations of X, ICA aims to identify the mixing matrix A and the distribution of the error variables ε. Theory and algorithms for ICA can be found in, e.g., [23, 28, 74, 75, 76, 122]. ICA has applications in many different disciplines, including blind source separation (e.g., [29]), face recognition (e.g., [10]), medical imaging (e.g., [12, 86, 172]) and causal discovery using the LiNGAM method (e.g., [156, 157]).

Our focus is on identifying the mixing matrix. Identifiability is an issue, since two different mixing matrices A and B may yield the same distribution of X, for example if the distribution of ε is multivariate Gaussian. In this case, the mixing matrix cannot be identified from X. In [28], it was shown that whenever at most one of the components of ε is Gaussian, the mixing matrix is asymptotically identifiable up to scaling and permutation of columns, see also Theorem 4 of [50]. In order to illustrate the relevance of identifying the mixing matrix in (9.1), we give an example based on causal inference.

Example 9.1.1. Consider a two-dimensional linear structucal equation model with additive noise of the form

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}, \tag{9.2}$$

see e.g. [156]. We assume that the coordinates of ε are independent, nondegenerate and have second moment, and assume that B is strictly triangular, meaning that all entries of B are zero except either B_{12} or B_{21}. In the first case, X_1 is a function of X_2 and vice versa for the second case. In the context of linear structural equation models, identifying which row of B is zero corresponds to identifying whether X_1 is a cause of X_2 or vice versa.

As B is strictly triangular, $I - B$ is invertible. Letting $A = (I - B)^{-1}$, we obtain

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}, \tag{9.3}$$

where A is upper or lower triangular according to whether the same holds for B. Thus, we have arrived at an ICA model of the form (9.1). By standardization, we may assume that X and ε both have mean zero and unit variance. Let $\alpha \in \mathbb{R}$ denote the covariance between X_1 and X_2. From independence of ε_1 and ε_2, we obtain

$$\begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} = V \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} A_{11}^2 + A_{12}^2 & A_{11}A_{21} + A_{12}A_{22} \\ A_{21}A_{11} + A_{22}A_{12} & A_{21}^2 + A_{22}^2 \end{bmatrix}. \tag{9.4}$$

In the case where A is lower triangular, meaning that $A_{12} = 0$, this yields

$$\begin{bmatrix} A_{11}^2 & A_{11}A_{21} \\ A_{21}A_{11} & A_{21}^2 + A_{22}^2 \end{bmatrix} = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}, \tag{9.5}$$

so that

$$A = \begin{bmatrix} 1 & 0 \\ \alpha & \sqrt{1-\alpha^2} \end{bmatrix}, \tag{9.6}$$

and similarly, for the case where A is upper triangular, we obtain

$$A = \begin{bmatrix} \sqrt{1-\alpha^2} & \alpha \\ 0 & 1 \end{bmatrix}. \tag{9.7}$$

Thus, A satisfies either (9.6) or (9.7). In the case where ε is jointly Gaussian, it is immediate that we cannot distinguish whether A satisfies (9.6) or (9.7) from the distribution of X alone. By the results of [28], distinguishing the cases (9.6) or (9.7) from the distribution of X is possible when ε has non-Gaussian coordinates. Thus, in this case, we may infer causal relationships from estimation of the mixing matrix in an ICA model. However, if the distribution of ε is close to Gaussian, it may be expected that based on samples from the distribution of X, identification of A and thus identification of the causal relationship becomes difficult. ∘

Motivated by the above, we study asymptotic identifiability of the mixing matrix under an asymptotic scenario where the distribution of ε depends on the sample size n, and tends to a Gaussian distribution as n tends to infinity. In fact, we will consider a general mean zero distribution ζ and an asymptotic scenario where the distribution of ε tends to ζ. Results on asymptotic identifiability for the case of a limiting Gaussian distribution then follow as a corollary.

Specifically, let ζ and ξ denote nondegenerate mean zero probability distributions on $(\mathbb{R}, \mathcal{B})$ such that $\xi \neq \zeta$. Fix $p \in \mathbb{N}$ and let A be a $p \times p$ matrix. Let (ε_n) be a sequence of p-dimensional variables such that for each n, the coordinates of $\varepsilon_n = (\varepsilon_{n1}, \ldots, \varepsilon_{np})$ are independent and identically distributed according to the contaminated ζ distribution

$$P_e(\beta_n) = \beta_n \xi + (1 - \beta_n)\zeta. \tag{9.8}$$

Here, scalar multiplication and addition of probability measures, or more generally of signed measures, is defined pointwisely as in [150]. We investigate asymptotic identifiability of the mixing matrix A based on n independent samples from the distribution of $X = A\varepsilon_n$, where β_n is allowed to tend to zero as n tends to infinity. Our results suggest that when $\beta_n \in o(1/\sqrt{n})$, asymptotic identifiability is determined solely by the properties of the limiting distribution ζ of $P_e(\beta_n)$. In particular, in the case where ζ is a Gaussian distribution, asymptotic identifiability becomes problematic if $\beta_n \in o(1/\sqrt{n})$. We also prove results showing, subject to certain regularity conditions, that in the scenario with $\beta_n = n^{-\rho}$ for some $0 < \rho < 1/2$, asymptotic identifiability properties remain as in the classical case where the distribution of ε does not depend on n.

9.2 Problem statement and main results

ICA can be used to estimate A when the distribution of ε is unknown. In this case, we may think of the statistical model corresponding to ICA as the collection of probability measures

$$\{L_A(R) \mid A \in \mathbb{M}(p,p), R \in \mathcal{P}(p)\}, \tag{9.9}$$

where $\mathbb{M}(p,p)$ denotes the space of $p \times p$ matrices, $L_A : \mathbb{R}^p \to \mathbb{R}^p$ is given by $L_A(x) = Ax$, $L_A(R)$ denotes the image measure of R under the transformation L_A, and $\mathcal{B}_p$ denotes the Borel-σ-algebra on $\mathbb{R}^p$. Also, $\mathcal{P}(p)$ denotes the set of product probability measures on $(\mathbb{R}^p, \mathcal{B}_p)$, indicating that the error distribution has independent coordinates. In other words, it is assumed that the distribution of X in (9.1) is equal to $L_A(R)$ for some $A \in \mathbb{M}(p,p)$ and $R \in \mathcal{P}(p)$. This is a semiparametric model, where A is the parameter of interest and R is a nuisance parameter. Asymptotic distributions of estimates of the mixing matrix in this type of set-up are derived in e.g. [5, 23, 77]. The difficulty of identifying A can then be appraised by considering for example the asymptotic variance of the estimates.

Alternatively, one can consider estimation of A for a given error distribution. This is the approach we take in this paper. When ε has the distribution of some fixed $R \in \mathcal{P}(p)$, the statistical model corresponding to the ICA model (9.1) is the collection of probability measures

$$\{L_A(R) \mid A \in \mathbb{M}(p,p)\}. \tag{9.10}$$

Asymptotic identifiability of A in (9.10) follows from the results of [28] and [50]. In particular, if no two coordinates of R are jointly Gaussian, the mixing matrix A is asymptotically identifiable up to scaling and permutation of columns, in the sense that $L_A(R) = L_B(R)$ implies $A = B\Lambda P$ for some diagonal matrix Λ and permutation matrix P.

We are interested in identifiability of the mixing matrix in (9.10) when the error distributions are different from Gaussian but close to Gaussian. Some results in this direction can be found in [125], where the authors calculated the Crámer-Rao lower bound for the model (9.10), under the assumption that the coordinates of the error distribution have certain regularity criteria such as finite variance and differentiable Lebesgue densities. These results indicate how the minimum variance of an unbiased estimator of the mixing matrix depends on the error distribution.

We consider the problem from the following different perspective. For $p \geq 1$ and any signed measure μ on $(\mathbb{R}, \mathcal{B})$, let $\mu \otimes \mu$ denote the product measure of μ with itself, and let $\mu^{\otimes p} = \otimes_{i=1}^{p} \mu$ denote the p-fold product measure. Fix two mean zero probability measures ξ and ζ with $\xi \neq \zeta$, and let $P_e(\beta)$ be the contaminated distribution given by

$$P_e(\beta) = \beta\xi + (1-\beta)\zeta. \tag{9.11}$$

Also, we write F^A for the cumulative distribution function of $L_A(\zeta^{\otimes p})$, and we write F^A_β for the cumulative distribution function of $L_A(P_e(\beta)^{\otimes p})$. In Section 9.3, we will show that F^A_β tends uniformly to F^A at an asymptotically linear rate in β as β tends to zero. As a consequence, whenever $F^A = F^B$, the distance $\|F^A_\beta - F^B_\beta\|_\infty$ tends to zero at an asymptotically linear rate as well. In Theorem 9.4.3, we use this result to show that when $F^A = F^B$ and $\beta \in o(1/\sqrt{n})$, identifiability of the mixing matrix is determined by the properties of F^A and not F^A_β. In particular, we argue in Corollary 9.4.4 that when ζ is a Gaussian distribution, $\beta \in o(1/\sqrt{n})$ and $AA^t = BB^t$, distinguishing between the candidates A and B for the mixing matrix becomes problematic. Finally, in Theorem 9.4.5, under suitable regularity conditions, we prove that in the case of sufficiently slow convergence to the limiting error distribution, meaning that $\beta_n = n^{-\rho}$ for some $0 < \rho < 1/2$, the asymptotic identifiability issues of the previous results do not manifest themselves, even when $F^A = F^B$. All proofs are given in Section 9.6.

9.3 An upper asymptotic distance bound

We begin by introducing some notation. For any measure μ on $(\mathbb{R}^p, \mathcal{B}_p)$, let $|\mu|$ denote the total variation measure of μ, see, e.g., [150]. We define two norms by

$$\|\mu\|_\infty = \sup_{x \in \mathbb{R}^p} |\mu((-\infty, x_1] \times \cdots \times (-\infty, x_p])|, \tag{9.12}$$

$$\|\mu\|_{tv} = |\mu|(\mathbb{R}^p), \tag{9.13}$$

and refer to these as the uniform and the total variation norms, respectively. The uniform norm for measures is also known as the Kolmogorov norm. Note that if P and Q are two probability measures on $(\mathbb{R}^p, \mathcal{B}_p)$ with cumulative distribution functions F and G, it holds that $\|P - Q\|_\infty = \|F - G\|_\infty$. Finally, we use the notation $f(s) \sim g(s)$ for $s \to s_0$ when $\lim_{s \to s_0} f(s)/g(s) = 1$. As in the previous section, let ξ and ζ be two mean zero probability distributions on $(\mathbb{R}, \mathcal{B})$ with $\xi \neq \zeta$. We aim to bound the distance

$$\|F^A_\beta - F^B_\beta\|_\infty = \|L_A(P_e(\beta)^{\otimes p}) - L_B(P_e(\beta)^{\otimes p})\|_\infty \tag{9.14}$$

for matrices $A, B \in \mathbb{M}(p, p)$ with $F^A = F^B$. The following theorem is a first step towards this goal.

Theorem 9.3.1. *Let $P_e(\beta) = \beta\xi + (1-\beta)\zeta$ for $\beta \in (0,1)$, and let $A \in \mathbb{M}(p,p)$. Then*

$$\lim_{\beta \to 0} \frac{L_A(P_e(\beta)^{\otimes p}) - L_A(\zeta^{\otimes p})}{\|P_e(\beta) - \zeta\|_\infty} = \sum_{k=1}^{p} L_A(\zeta^{\otimes(k-1)} \otimes \nu \otimes \zeta^{\otimes(p-k)}), \tag{9.15}$$

where $\nu = (\xi - \zeta)/\|\xi - \zeta\|_\infty$ and convergence is in $\|\cdot\|_\infty$. In particular, F^A_β tends uniformly to F^A as β tends to zero.

The proof of Theorem 9.3.1 exploits properties of the contaminated distributions $P_e(\beta)$ for $\beta \in (0,1)$, in particular that $\|P_e(\beta) - \zeta\|_\infty$ is nonzero and linear in β, and that $(P_e(\beta) - \zeta)/\|P_e(\beta) - \zeta\|_\infty$ is constant in β. As Lemma 9.3.2 shows, only contaminated distributions have these properties. This is our main reason for working with this family of distributions.

Lemma 9.3.2. *Let $\beta \mapsto Q(\beta)$ be a mapping from $(0,1)$ to the space of probability measures on $(\mathbb{R}, \mathcal{B})$ with the properties that $\|Q(\beta) - \zeta\|_\infty$ is nonzero and linear in β and $(Q(\beta) - \zeta)/\|Q(\beta) - \zeta\|_\infty$ is constant in β. Then $Q(\beta)$ can be written as a contaminated ζ distribution, in the sense that $Q(\beta) = \beta\xi + (1-\beta)\zeta$ for some probability measure ξ on $(\mathbb{R}, \mathcal{B})$.*

Due to the properties of contaminated distributions, Theorem 9.3.1 in fact also holds for other norms than the uniform norm. However, the choice of the norm is important when we wish to bound the norm of the right-hand side of (9.15). Such a bound is achieved in Lemma 9.3.3.

Lemma 9.3.3. *Let $A \in \mathbb{M}(p,p)$. Then*

$$\left\| \sum_{k=1}^{p} L_A \left(\zeta^{\otimes(k-1)} \otimes \frac{\xi - \zeta}{\|\xi - \zeta\|_\infty} \otimes \zeta^{\otimes(p-k)} \right) \right\|_\infty \leq 2p. \tag{9.16}$$

Combining Theorem 9.3.1 and Lemma 9.3.3 yields the following corollary, which we give without proof.

Corollary 9.3.4. *Let $A, B \in \mathbb{M}(p,p)$ be such that $F^A = F^B$. Define*

$$\begin{aligned} \Gamma(A, B, \nu) = &\sum_{k=1}^{p} L_A \left(\zeta^{\otimes(k-1)} \otimes \nu \otimes \zeta^{\otimes(p-k)} \right) \\ &- \sum_{k=1}^{p} L_B \left(\zeta^{\otimes(k-1)} \otimes \nu \otimes \zeta^{\otimes(p-k)} \right). \end{aligned} \tag{9.17}$$

Then we have, for $\beta \to 0$,

$$\|F_\beta^A - F_\beta^B\|_\infty \sim \left\| \Gamma \left(A, B, \frac{\xi - \zeta}{\|\xi - \zeta\|_\infty} \right) \right\|_\infty \|P_e(\beta) - \zeta\|_\infty \leq 4p\beta \|\xi - \zeta\|_\infty. \tag{9.18}$$

Corollary 9.3.4 shows that in the case where $F^A = F^B$, as β tends to zero and the error distributions $P_e(\beta)$ become closer to ζ, the distance between the observational distributions F_β^A and F_β^B decreases asymptotically linearly in β. Heuristically, this suggests that when $F^A = F^B$ and β is close to zero, the distributions F_β^A and F_β^B are hard to distinguish.

Corollary 9.3.4 is stated under the condition that $F^A = F^B$. For later use, we characterize the occurrence of this in the next lemma, in terms of ζ, A and B, for

the case where A and B are invertible and ζ is non-degenerate, meaning that ζ is not a Dirac measure. Recall that a probability distribution Q on $(\mathbb{R}, \mathcal{B})$ is said to be symmetric if, for every random variable Y with distribution Q, Y and $-Y$ have the same distribution. The proof of Lemma 9.3.5, given in Section 9.6, is a simple consequence of Theorem 4 of [50].

Lemma 9.3.5. *Assume that ζ is a non-degenerate mean zero probability measure on $(\mathbb{R}, \mathcal{B})$. Let $A, B \in \mathbb{M}(p, p)$ be invertible. Then the following hold:*

1. *If ζ is Gaussian, then $F^A = F^B$ if and only if $AA^t = BB^t$.*
2. *If ζ is non-Gaussian and symmetric, then $F^A = F^B$ if and only if $A = B\Lambda P$ for some permutation matrix P and a diagonal matrix Λ satisfying $\Lambda^2 = I$.*
3. *If ζ is non-symmetric, then $F^A = F^B$ if and only if $A = BP$ for some permutation matrix P.*

9.4 Asymptotic identifiability

We now turn to asymptotic properties of ICA models. We will need some basic facts about random fields in order to formulate our results, see [97, 111] for an overview. Recall that a mapping $R : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is said to be symmetric if $R(x, y) = R(y, x)$ for all $x, y \in \mathbb{R}^p$, and is said to be positive semidefinite if for all $n \geq 1$ and for all $x_1, \ldots, x_n \in \mathbb{R}^p$ and $\xi_1, \ldots, \xi_n \in \mathbb{R}$, it holds that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i R(x_i, x_j) \xi_j \geq 0. \tag{9.19}$$

For any symmetric and positive semidefinite function $R : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, there exists a mean zero Gaussian random field W with covariance function R, taking its values in $\mathbb{R}^{\mathbb{R}^p}$. In general, W will not have continuous paths. For a general random field W, we associate with W its intrinsic pseudometric ρ on $\mathbb{R}^p$, given by

$$\rho(x, y) = \sqrt{E(W(x) - W(y))^2}. \tag{9.20}$$

If the metric space $(\mathbb{R}^p, \rho)$ is separable, we say that W is separable. In this case, $\|W\|_\infty = \sup_{x \in D} |W(x)|$ with probability one, for any countable subset D of $\mathbb{R}^p$ which is dense under the pseudometric ρ.

The following lemma describes some important properties of a class of Gaussian fields particularly relevant to us. The result is well known, see for example [43]. For completeness, we outline a short proof in Section 9.6 based on a strong approximation result from the literature.

Lemma 9.4.1. *Let F be a cumulative distribution function on $\mathbb{R}^p$. There exists a p-dimensional separable mean zero Gaussian field W which has covariance function $R : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ given by $R(x, y) = F(x \wedge y) - F(x)F(y)$ for $x, y \in \mathbb{R}^p$, where $x \wedge y$ is the coordinate wise minimum of x and y. With $\mathbb{Q}$ denoting the rationals, it holds that $\|W\|_\infty = \sup_{x \in \mathbb{Q}^p} |W(x)|$ and $\|W\|_\infty$ is almost surely finite.*

For a fixed cumulative distribution function F, we refer to the Gaussian field described in Lemma 9.4.1 as an F-Gaussian field. We are now ready to formulate our results on asymptotic identifiability in ICA models. We first state a result, Theorem 9.4.2, concerning the classical asymptotic scenario, where the error distribution is not contaminated and does not depend on the sample size n. Fix a mean zero probability distribution ζ on $(\mathbb{R}, \mathcal{B})$ and a matrix $A \in \mathbb{M}(p, p)$. As in the previous section, we let F^A denote the cumulative distribution function of $L_A(\zeta^{\otimes p})$, corresponding to the distribution of $A\varepsilon$ when ε is a p-dimensional variable with independent coordinates having distribution ζ. Consider a probability space $(\Omega, \mathcal{F}, P)$ endowed with $(X_k)_{k \geq 1}$ be independent variables with cumulative distribution function F^A. Let $\mathbb{F}_n^A$ be the empirical distribution function of $X_1, \ldots, X_n$. Also assume that we are given an F^A-Gaussian field W on $(\Omega, \mathcal{F}, P)$.

Theorem 9.4.2. *Let $c \geq 0$ be a continuity point of the distribution of $\|W\|_\infty$. Then*

$$\lim_{n \to \infty} P(\sqrt{n}\|\mathbb{F}_n^A - F^A\|_\infty > c) = P(\|W\|_\infty > c), \tag{9.21}$$

while in the case where $F^A \neq F^B$, it holds that

$$\lim_{n \to \infty} P(\sqrt{n}\|\mathbb{F}_n^A - F^B\|_\infty > c) = 1. \tag{9.22}$$

The equations (9.21) and (9.22) roughly state that in the classical asymptotic scenario, $\sqrt{n}\|\mathbb{F}_n^A - F^A\|_\infty$ converges in distribution to $\|W\|_\infty$, while $\sqrt{n}\|\mathbb{F}_n^A - F^B\|_\infty$ is not bounded in probability if $F^A \neq F^B$. Note that Lemma 9.3.5 gives us conditions for $F^A = F^B$ and $F^A \neq F^B$ depending on ζ.

Next, we consider an asymptotic scenario where the distribution of the error variable is contaminated and the amount of contamination depends on the sample size n. As in Section 9.3, ξ and ζ are fixed mean zero probability measures on $(\mathbb{R}, \mathcal{B})$ with $\xi \neq \zeta$, $P_e(\beta) = \beta\xi + (1 - \beta)\zeta$, $A \in \mathbb{M}(p, p)$ is a fixed matrix, F^A is the cumulative distribution function of $L_A(\zeta^{\otimes p})$ and F_β^A is the cumulative distribution function of $L_A(P_e(\beta)^{\otimes p})$. Thus, F_β^A is the cumulative distribution function of $A\varepsilon$, where ε is a p-dimensional variable with independent coordinates having distribution $P_e(\beta)$. Consider a sequence (β_n) in $(0, 1)$, and consider a probability space $(\Omega, \mathcal{F}, P)$ endowed with a triangular array $(X_{nk})_{1 \leq k \leq n}$ such that for each n, the variables $X_{n1}, \ldots, X_{nn}$ are independent variables with cumulative distribution function $F_{\beta_n}^A$. Let $\mathbb{F}_{\beta_n}^A$ be the empirical distribution function of $X_{n1}, \ldots, X_{nn}$. Also assume that we are given an F^A-Gaussian field W on $(\Omega, \mathcal{F}, P)$. We are interested in the asymptotic properties of $\mathbb{F}_{\beta_n}^A$. Theorem 9.4.3 is our main result on this type of asymptotic scenarios.

Theorem 9.4.3. *Let* $\lim_n \sqrt{n}\beta_n = k$ *for some* $k \geq 0$*. If* $F^A = F^B$*, then*

$$\begin{aligned} P(\|W\|_\infty > c + 4pk\|\xi - \zeta\|_\infty) &\leq \liminf_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \\ &\leq \limsup_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \\ &\leq P(\|W\|_\infty \geq c - 4pk\|\xi - \zeta\|_\infty). \end{aligned} \tag{9.23}$$

In particular, if $k = 0$ *and* c *is a continuity point of the distribution of* $\|W\|_\infty$*, we have*

$$\lim_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) = P(\|W\|_\infty > c). \tag{9.24}$$

Theorem 9.4.3 essentially shows that for the asymptotic scenario considered, the convergence of $F^A_{\beta_n}$ to F^A is fast enough to ensure that the asymptotic properties of $\mathbb{F}^A_{\beta_n}$ are determined by F^A instead of $F^A_{\beta_n}$. Corollary 9.4.4 applies this result to the case where the error distributions become close to Gaussian without being Gaussian.

Corollary 9.4.4. *Assume that* $\lim_n \sqrt{n}\beta_n = 0$*. Let* $A, B \in \mathbb{M}(p,p)$ *be invertible. Assume that* $AA^t = BB^t$ *while* $A \neq B\Lambda P$ *for all diagonal* Λ *with* $\Lambda^2 = I$ *and all permutation matrices* P*. Let* ζ *be a nondegenerate Gaussian distribution and let* ξ *be such that* $P_e(\beta)$ *is non-Gaussian for all* $\beta \in (0,1)$*. Let* c *be a point of continuity for the distribution of* $\|W\|_\infty$*, with* W *an* F^A*-Gaussian field. It then holds that:*

1. $F^A_{\beta_n} \neq F^B_{\beta_n}$ *for all* $n \geq 1$*.*
2. $\lim_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) = P(\|W\|_\infty > c)$*.*

Statement (1) of Corollary 9.4.4 shows that for any finite n, we are in the case where, were the error distribution not changing with n, it would be possible to asymptotically distinguish $F^A_{\beta_n}$ and $F^B_{\beta_n}$ at rate $1/\sqrt{n}$ as in (9.22) of the classical case. However, statement (2) shows that as n increases and the error distributions becomes closer to a Gaussian distribution, distinguishing $F^A_{\beta_n}$ and $F^B_{\beta_n}$ at rate $1/\sqrt{n}$ is nonetheless not possible, with a limit result similar to (9.21). Note that by Lemma 9.3.5, having $A \neq B\Lambda P$ for all diagonal Λ with $\Lambda^2 = I$ and all permutation matrices P, as in Corollary 9.4.4, is the minimum requirement for non-Gaussian error distributions to be able to asymptotically distinguish F^A and F^B in the classical scenario.

Theorem 9.4.3 and Corollary 9.4.4 cover the case $\beta_n = o(1/\sqrt{n})$, in particular the case $\beta_n = n^{-\rho}$ for $\rho > 1/2$. We end the section with a result showing that under some further regularity conditions, distinguishing $F^A_{\beta_n}$ and $F^B_{\beta_n}$ at rate $1/\sqrt{n}$ is possible when $0 < \rho < 1/2$.

Theorem 9.4.5. *Let* $\rho \in (0, 1/2)$ *and let* $\beta_n = n^{-\rho}$*. For all* $A \in \mathbb{M}(p,p)$*, define*

$$\Gamma_1(A) = \sum_{k=1}^{p} L_A\left(\zeta^{\otimes(k-1)} \otimes \frac{\xi - \zeta}{\|\xi - \zeta\|_\infty} \otimes \zeta^{\otimes(p-k)}\right). \tag{9.25}$$

If either $F^A \neq F^B$ *or* $F^A = F^B$ *and* $\Gamma_1(A) \neq \Gamma_1(B)$*, then*

$$\lim_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) = 1. \tag{9.26}$$

9.5 Discussion

We studied identifiability of the ICA model for error distributions which have independent coordinates following contaminated distributions. We argued in particular that for contaminated Gaussian distributions, it holds that if the level of contamination decreases at rate $1/\sqrt{n}$ or faster, then asymptotic identifiability is determined by the Gaussian limiting distribution rather than by the non-Gaussian contaminated distribution. Combining this with Lemma 9.3.5, we obtain as a consquence that distinguishing A and B becomes difficult when $AA^t = BB^t$, rather than when A and B are equal up to sign reversion and permutation of columns. The consequence of this is that if we have n observations from an ICA model with a contaminated Gaussian error distribution with contamination level on the order of $1/\sqrt{n}$, we expect that identifying the mixing matrix will be difficult.

The proof of our main theoretical result, Theorem 9.4.3, rests on two partial results:

1. That when F_n is a sequence of cumulative distribution functions converging uniformly to F, then $\sqrt{n}(\mathbb{F}_n - F_n)$ converges weakly in $\mathcal{L}_\infty(\mathbb{R}^p)$, where $\mathbb{F}_n$ is an empirical process based on n independent observations of variables with cumulative distribution function F_n.

2. That convergence of the distribution function of $X\varepsilon$ for ε having distribution $P_e(\beta_n)^{\otimes p}$, with $P_e(\beta) = \beta\xi + (1-\beta)\zeta$, is asymptotically linear in β, implying that rate $1/\sqrt{n}$ convergence of error distributions ε_n translates into rate $1/\sqrt{n}$ convergence of observational distributions $A\varepsilon_n$.

In Theorem 9.4.5, we also considered the case of slower rates of decrease in the level of contamination, namely rates $n^{-\rho}$ for $0 < \rho < 1/2$. Our results here indicate that in such asymptotic scenarios, identifiability of the mixing matrix at rate $1/\sqrt{n}$ will be possible, subject to some regularity conditions related to the Γ_1 signed measures of (9.25).

We have conducted numerical experiments to assess our results. We considered the case where $p = 2$, ξ is the standard exponential distribution, ζ is the standard normal distribution and

$$A = \begin{bmatrix} 1 & 0 \\ \alpha & \sqrt{1-\alpha^2} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} \sqrt{1-\alpha^2} & \alpha \\ 0 & 1 \end{bmatrix}, \tag{9.27}$$

where $\alpha \in (0,1)$ is fixed, see also Example 9.1.1. It then holds that $AA^t = BB^t$ while $A \neq B\Lambda P$ for diagonal Λ with $\Lambda^2 = I$ and permutation matrices P. Combining

Theorem 9.4.3 and Theorem 9.4.5, we would expect that with

$$p(\rho) = \lim_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \tag{9.28}$$

for $\beta_n = n^{-\rho}$, we should have $p(\rho) = 1$ $p(\rho) = P(\|W\|_\infty > c)$ for $0 < \rho < 1/2$ and $1/2 < \rho$, respectively. By Monte Carlo simulations, we found that $p(\rho)$ appears to be constant for $\rho > 1/2$, in accordance with Theorem 9.4.3. However, our results did not satisfactorily indicate $p(\rho) = 1$ for $0 < \rho < 1/2$, as Theorem 9.4.5 would suggest. We suggest that this is because we in our simulations considered too low a level of n, namely $5 \cdot 10^4$. For large n, we would expect that for some constant $k > 0$,

$$P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \approx P(\sqrt{n}\|F^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \approx 1_{(\sqrt{n}k\beta_n > c)}, \tag{9.29}$$

and thus, we obtain $P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \approx 1$ when $n^{1/2-\rho} = \sqrt{n}\beta_n > c/k$, corresponding to $n > \exp((\log c/k)/(1/2 - \rho))$. For $\rho < 1/2$ close to $1/2$, this latter number grows extremely large, and so detection of the limiting value of 1 for $p(\rho)$ becomes difficult.

Our results also leave unanswered research questions, for example:

1. Is it possible to characterize the matrices A and B such that the regularity condition $\Gamma_1(A) \neq \Gamma_2(B)$ of Theorem 9.4.5 holds?

2. Together, Theorem 9.4.3 and Theorem 9.4.5 describe the behaviour of the empirical process $\mathbb{F}^A_{\beta_n}$ for asymptotic scenarios of the form $\beta_n = n^{-\rho}$ for $\rho > 0$, in particular describing the difficulty of using $\mathbb{F}^A_{\beta_n}$ to distinguish $F^A_{\beta_n}$ and $F^B_{\beta_n}$. Is it possible to obtain finite-sample bounds instead of limiting behaviours in these results?

3. How do Theorem 9.4.3 and Theorem 9.4.5 translate into results on the ability of practical algorithms such as the fastICA algorithm, see [74], to distinguish the correct mixing matrix?

4. Is it possible to use similar techniques to analyze identifiability of the mixing matrix in asymptotic scenarios where p tends to infinity?

5. Do the present results extend to cases where the coordinates of the error distributions are not contaminated normal distributions, or when the coordinates are not identically distributed?

In light of these unanswered questions, our presents results should be seen as a small step towards a better understanding of the identifiability of the mixing matrix for ICA for error distributions which are close to Gaussian but not Gaussian.

9.6 Proofs

9.6.1 Proofs for Section 9.3

Proof of Theorem 9.3.1. First note that we have $P_e(\beta)-\zeta = \beta(\xi-\zeta)$. Taking norms, this implies $\|P_e(\beta)-\zeta\|_\infty = \beta\|\xi-\zeta\|_\infty$ and

$$\frac{P_e(\beta)-\zeta}{\|P_e(\beta)-\zeta\|_\infty} = \frac{\xi-\zeta}{\|\xi-\zeta\|_\infty}. \tag{9.30}$$

With $\nu = (\xi-\zeta)/\|\xi-\zeta\|_\infty$, we then also have $P_e(\beta) = \zeta + \beta\|\xi-\zeta\|_\infty\nu$. We begin by analyzing $P_e(\beta)^{\otimes p}$. For Borel subsets $C_1,\ldots,C_p$ of $\mathbb{R}$, we have

$$\begin{aligned} P_e(\beta)^{\otimes p}(C_1\times\cdots\times C_p) &= (\zeta+\beta\|\xi-\zeta\|_\infty\nu)^{\otimes p}(C_1\times\cdots\times C_p) \\ &= \prod_{k=1}^{p}(\zeta(C_k)+\beta\|\xi-\zeta\|_\infty\nu(C_k)) \\ &= \sum_{k=0}^{p}\beta^k\|\xi-\zeta\|_\infty^k \sum_{\alpha\in S_k}\prod_{i=1}^{p}\zeta(C_i)^{1-\alpha_i}\nu(C_i)^{\alpha_i}, \end{aligned} \tag{9.31}$$

where $S_k = \{\alpha\in\{0,1\}^p \,|\sum_{i=1}^p \alpha_i = k\}$, and the last equality follows since

$$\prod_{k=1}^{p}(a_k+\gamma b_k) = \sum_{k=0}^{p}\gamma^k\sum_{\alpha\in S_k}\prod_{i=1}^{p}a_i^{1-\alpha_i}b_i^{\alpha_i}, \qquad \text{for } a,b\in\mathbb{R}^p \text{ and } \gamma\in\mathbb{R}. \tag{9.32}$$

Defining $\mu_0=\zeta$ and $\mu_1=\nu$, we then obtain

$$P_e(\beta)^{\otimes p}(C_1\times\cdots\times C_p) = \sum_{k=0}^{p}\beta^k\|\xi-\zeta\|_\infty^k\sum_{\alpha\in S_k}(\otimes_{i=1}^p\mu_{\alpha_i})(C_1\times\cdots\times C_p). \tag{9.33}$$

Letting $\Gamma_k = \sum_{\alpha\in S_k} L_A(\otimes_{i=1}^p\mu_{\alpha_i})$, this yields

$$L_A(P_e(\beta)^{\otimes p}) = \sum_{k=0}^{p}\beta^k\|\xi-\zeta\|_\infty^k\Gamma_k. \tag{9.34}$$

Next, note that $\Gamma_0 = L_A(\zeta^{\otimes p})$, so that

$$\begin{aligned} \lim_{\beta\to0}\frac{L_A(P_e(\beta)^{\otimes p})-L_A(\zeta^{\otimes p})}{\|P_e(\beta)-\zeta\|_\infty} &= \lim_{\beta\to0}\sum_{k=1}^{p}\beta^{k-1}\|\xi-\zeta\|_\infty^{k-1}\Gamma_k \\ &= \Gamma_1 = \sum_{k=1}^{p}L_A(\zeta^{\otimes(k-1)}\otimes\nu\otimes\zeta^{\otimes(p-k)}). \end{aligned} \tag{9.35}$$

In particular, this shows that for any $\eta > 0$,

$$\begin{aligned}\limsup_{\beta\to 0} \|L_A(P_e(\beta)^{\otimes p}) - L_A(\zeta^{\otimes p})\|_\infty &\leq \limsup_{\beta\to 0}(1+\eta)\|\Gamma_1\|_\infty \|P_e(\beta) - \zeta\|_\infty \\ &\leq \limsup_{\beta\to 0}(1+\eta)\|\Gamma_1\|_\infty \beta\|\xi - \zeta\|_\infty = 0, \quad (9.36)\end{aligned}$$

so F_β^A converges uniformly to F^A. □

Proof of Lemma 9.3.2. Let α be such that $\|Q(\beta) - \zeta\|_\infty = \alpha\beta$ for some $\alpha > 0$. Let $\xi = Q(1)$. With $\beta > 0$, we then have

$$\frac{Q(\beta) - \zeta}{\|Q(\beta) - \zeta\|_\infty} = \frac{Q(\beta) - \zeta}{\alpha\beta}, \tag{9.37}$$

while

$$\frac{Q(1) - \zeta}{\|Q(1) - \zeta\|_\infty} = \frac{\xi - \zeta}{\alpha}. \tag{9.38}$$

By our assumptions, the right-hand sides in (9.37) and (9.38) are equal. This implies $Q(\beta) = \beta\xi + (1-\beta)\zeta$. □

To prove Lemma 9.3.3, we first present a lemma relating the uniform norm of certain measures on $(\mathbb{R}^p, \mathcal{B}_p)$ to the uniform and total variation norms of some measures on $(\mathbb{R}, \mathcal{B})$.

Lemma 9.6.1. *Let $\mu_1, \ldots, \mu_p$ be signed measures on $(\mathbb{R}, \mathcal{B})$, and let $A \in \mathbb{M}(p,p)$. Then for any $i \in \{1, \ldots, p\}$, it holds that*

$$\|L_A(\mu_1 \otimes \cdots \otimes \mu_p)\|_\infty \leq 2\|\mu_i\|_\infty \prod_{k\neq i}^{p} \|\mu_k\|_{tv}. \tag{9.39}$$

Proof. For any permutation $\pi : \{1, \ldots, p\} \to \{1, \ldots, p\}$ and corresponding permutation matrix P, we have $L_A(\mu_1 \otimes \cdots \otimes \mu_p) = L_{AP^{-1}}(\mu_{\pi(1)} \otimes \cdots \otimes \mu_{\pi(p)})$. Hence, it suffices to consider $i = p$. Let $x \in \mathbb{R}^p$ and define $I_x = (-\infty, x_1] \times \cdots \times (-\infty, x_p]$. Then Fubini's theorem for signed measures yields

$$\begin{aligned}&|L_A(\mu_1 \otimes \cdots \otimes \mu_p)(I_x)| \\ &= \left| \int \cdots \int 1_{I_x}(L_A(y)) \, \mathrm{d}\mu_p(y_p) \cdots \mathrm{d}\mu_1(y_1) \right| \\ &\leq \int \cdots \int \left| \int 1_{I_x}(L_A(y)) \, \mathrm{d}\mu_p(y_p) \right| \mathrm{d}|\mu_{p-1}|(y_{p-1}) \cdots \mathrm{d}|\mu_1|(y_1), \quad (9.40)\end{aligned}$$

where we have also used the triangle inequality for integrals with respect to signed measures, which follows for example from Theorem 6.12 of [150]. We now analyze

the innermost integral of (9.40). For fixed $y_1, \ldots, y_{p-1}$, we have

$$\begin{aligned}
&\{y_p \in \mathbb{R} \mid 1_{I_x}(L_A(y)) = 1\} \\
=&\{y_p \in \mathbb{R} \mid \ \forall\, i \leq p : (Ay)_i \leq x_i\} \\
=&\cap_{i=1}^{p} \{y_p \in \mathbb{R} \mid a_{ip} y_p \leq x_i - (a_{i1} y_1 + \cdots + a_{i(p-1)} y_{p-1})\}. \qquad (9.41)
\end{aligned}$$

Hence, $\{y_p \in \mathbb{R} \mid 1_{I_x}(L_A(y)) = 1\}$ is a finite intersection of intervals, and is therefore itself an interval. This yields

$$|\mu_p(\{y_p \in \mathbb{R} \mid 1_{I_x}(L_A(y)) = 1\})| \leq 2\|\mu_p\|_\infty. \qquad (9.42)$$

This inequality is immediate when the interval is of the form $(-\infty, a]$ for some $a \in \mathbb{R}$. If the interval is of the form $[a, \infty)$, we have

$$\begin{aligned}
|\mu_p([a,\infty))| \leq& |\mu_p(\mathbb{R})| + |\mu_p(-\infty, a)| \\
=& \lim_{b\to\infty} |\mu_p((-\infty, b])| + |\mu_p((-\infty, a - 1/b])| \leq 2\|\mu_p\|_\infty, \qquad (9.43)
\end{aligned}$$

and similarly for other types of intervals, whether bounded or unbounded, open, half-open or closed. Combining (9.40) and (9.42) yields

$$\begin{aligned}
&|L_A(\mu_1 \otimes \cdots \otimes \mu_p)(I_x)| \\
\leq& \int \cdots \int 2\|\mu_p\|_\infty \, \mathrm{d}|\mu_{p-1}|(y_{p-1}) \cdots \mathrm{d}|\mu_1|(y_1) = 2\|\mu_p\|_\infty \prod_{k=1}^{p-1} \|\mu_k\|_{tv}. \qquad (9.44)
\end{aligned}$$

□

Proof of Lemma 9.3.3. Let $\nu = (\xi - \zeta)/\|\xi - \zeta\|_\infty$. By Lemma 9.6.1, we have

$$\|L_A(\zeta^{\otimes(k-1)} \otimes \nu \otimes \zeta^{\otimes(p-k)})\|_\infty \leq 2\|\nu\|_\infty \|\zeta\|_{tv}^{p-1} = 2. \qquad (9.45)$$

Applying the triangle inequality, we therefore obtain

$$\left\| \sum_{k=1}^{p} L_A\left(\zeta^{\otimes(k-1)} \otimes \frac{\xi - \zeta}{\|\xi - \zeta\|_\infty} \otimes \zeta^{\otimes(p-k)}\right) \right\|_\infty \leq 2p. \qquad (9.46)$$

□

Proof of Lemma 9.3.5. **Proof of (1)**. With ζ Gaussian with mean zero and variance σ^2, $L_A(\zeta^{\otimes p})$ is Gaussian with mean zero and variance $\sigma^2 AA^t$, and so the result is immediate for this case.

Proof of (3). Now consider the case where ζ is not a symmetric distribution. As $L_P(\zeta^{\otimes p}) = \zeta^{\otimes p}$ holds for any permutation matrix P, we obtain that if $A = BP$, then $L_A(\zeta^{\otimes p}) = L_B(\zeta^{\otimes p})$ and so $F^A = F^B$, proving one implication.

Conversely, assume that $F^A = F^B$, meaning that $L_A(\zeta^{\otimes p}) = L_B(\zeta^{\otimes p})$. As ζ is non-degenerate and non-Gaussian and A and B are invertible, Theorem 4 of [50] shows

that $A = B\Lambda P$, where $\Lambda \in \mathbb{M}(p,p)$ is an invertible diagonal matrix and $P \in \mathbb{M}(p,p)$ is a permutation matrix. This yields

$$\zeta^{\otimes p} = L_{B^{-1}}(L_B(\zeta^{\otimes p})) = L_{B^{-1}}(L_A(\zeta^{\otimes p})) = L_{\Lambda P}(\zeta^{\otimes p}) = L_\Lambda(\zeta^{\otimes p}). \tag{9.47}$$

Now let Z be a random variable with distribution ζ. The above then yields that for all i, $\Lambda_{ii}Z$ and Z have the same distribution. In particular, $|\Lambda_{ii}||Z|$ and $|Z|$ have the same distribution, so $P(|Z| \leq z/|\Lambda_{ii}|) = P(|Z| \leq z)$ for all $z \in \mathbb{R}$. As Z is not almost surely zero, there is $z \neq 0$ such that $P(|Z| \leq z - \varepsilon) < P(|Z| \leq z + \varepsilon)$ for all $\varepsilon > 0$. This yields $|\Lambda_{ii}| = 1$. Next, let φ denote the characteristic function of Z. We then have $\varphi(\Lambda_{ii}\theta) = \varphi(\theta)$ for all $\theta \in \mathbb{R}$. As Z is not symmetric, there is a $\theta \in \mathbb{R}$ such that $\varphi(\theta) \neq \varphi(-\theta)$. Therefore, $\Lambda_{ii} = -1$ cannot hold, so we must have $\Lambda_{ii} = 1$. We conclude that Λ is the identity matrix and thus $A = BP$, as required.

Proof of (2). Finally, consider a symmetric probability measure ζ. It is then immediate that when Λ and P are as in the statement of the lemma, it holds that $L_{\Lambda P}(\zeta^{\otimes p}) = \zeta^{\otimes p}$ and thus $F^A = F^B$ whenever $A = B\Lambda P$. The converse implication follows as in the proof of (3). □

9.6.2 Proofs for Section 7.5

Proof of Lemma 9.4.1. The existence of the process W follows from the results cited at the beginning of Section 7.5. To show separability, note that there exists for any $x \in \mathbb{R}^p$ a sequence $(x_n) \subseteq \mathbb{Q}^p$ such that $F(x) = \lim_{n\to\infty} F(x_n)$. Therefore, $\mathbb{R}^p$ endowed with the intrinsic pseudometric ρ of W is separable and $\mathbb{Q}^p$ is a countable dense subset. As a consequence, $\|W\|_\infty = \sup_{x\in\mathbb{Q}^p} |W(x)|$ almost surely holds. In particular, completing the underlying probability space, we may take $\|W\|_\infty$ to be measurable.

In order to see that $\|W\|_\infty$ is almost surely finite, note that by Theorem B of [32], there exists a probability space $(\Omega, \mathcal{F}, P)$ endowed with a sequence of variables (X_k), independent and with common cumulative distribution function F, as well as a sequence of p-dimensional separable Gaussian fields (W_k) with the same finite-dimensional distribution as W, such that with $\mathbb{F}_n$ denoting the empirical distribution function of $X_1, \ldots, X_n$, it holds that

$$P\left(\|\sqrt{n}(\mathbb{F}_n - F) - W_n\|_\infty \geq C_1 n^{-1/(2(2p-1))} \log n\right) \leq \frac{C_2}{n^2}, \tag{9.48}$$

for some $C_1, C_2 > 0$. As all the W_n have the same distribution, this yields in particular that

$$\begin{aligned} 1 - \frac{C_2}{n^2} \leq & P\left(\|\sqrt{n}(\mathbb{F}_n - F) - W_n\|_\infty \leq C_1 n^{-1/(2(2p-1))} \log n\right) \\ \leq & P(\|W\|_\infty < \infty). \end{aligned} \tag{9.49}$$

Letting n tend to infinity, this implies $P(\|W\|_\infty < \infty) = 1$, as required. □

Before proving Theorem 9.4.2 and Theorem 9.4.3, we show a result on empirical processes. Recall that for a metric space (M, d), the ε-covering number $N(\varepsilon, M, d)$ is the minimum number of open balls of radius ε which is required to cover (M, d), see e.g. Section 2.1.1 of [170].

Lemma 9.6.2. *Fix a cumulative distribution function F. Define $\rho : \mathbb{R}^p \times \mathbb{R}^p$ by*

$$\rho(x, y) = \sqrt{F(x) + F(y) - 2F(x \wedge y)}, \tag{9.50}$$

and let $I_x = (-\infty, x_1] \times \cdots \times (-\infty, x_p]$. Let Z be a variable with cumulative distribution function F. Then, the following holds:

1. *ρ is a pseudometric.*
2. *$\rho(x, y) = \sqrt{E(1_{I_x}(Z) - 1_{I_y}(Z))^2}$.*
3. *$(\mathbb{R}^p, \rho)$ is totally bounded.*

Proof. First note that

$$\begin{aligned} \rho(x, y)^2 =& F(x) + F(y) - 2F(x \wedge y) \\ =& E1_{I_x}(Z) + E1_{I_y}(Z) - 2E1_{I_x}(Z)1_{I_y}(Z) \\ =& E(1_{I_x}(Z) - 1_{I_y}(Z))^2, \end{aligned} \tag{9.51}$$

proving claim (2). It is then immediate that ρ is a pseudometric, proving claim (1). Next, it holds that $(\mathbb{R}^p, \rho)$ is totally bounded if and only if $N(\varepsilon, \mathbb{R}^p, \rho)$ is finite for all positive ε. Let Q be the distribution corresponding to the cumulative distribution function F, and let $\mathcal{L}^2(\mathbb{R}^p, \mathcal{B}_p, Q)$ be the space of Borel measurable functions from $\mathbb{R}^p$ to $\mathbb{R}$ which are square-integrable with respect to Q. Let $\|\cdot\|_{2,Q}$ denote the usual seminorm on $\mathcal{L}^2(\mathbb{R}^p, \mathcal{B}_p, Q)$. Applying claim (2), it is immediate that

$$N(\varepsilon, \mathbb{R}^p, \rho) = N(\varepsilon, (1_{I_x})_{x \in \mathbb{R}^p}, \|\cdot\|_{2,Q}). \tag{9.52}$$

Combining Example 2.6.1 and Exercise 2.6.9 of [170], we find that $(1_{I_x})_{x \in \mathbb{R}^p}$ is a Vapnik-Cervonenkis (VC) subgraph class with VC dimension $p + 1$. Furthermore, $(1_{I_x})_{x \in \mathbb{R}^p}$ has envelope function constant and equal to one. Therefore, Theorem 2.6.7 of [170] shows that $N(\varepsilon, (1_{I_x})_{x \in \mathbb{R}^p}, \|\cdot\|_{2,Q})$ and thus $N(\varepsilon, \mathbb{R}^p, \rho)$ is finite, and so $(\mathbb{R}^p, \rho)$ is totally bounded. □

Lemma 9.6.3. *Let (F_n) be a sequence of cumulative distribution functions on $\mathbb{R}^p$, and let F be a cumulative distribution function on $\mathbb{R}^p$. Let $(X_{nk})_{1 \leq k \leq n}$ be a triangular array such that for each n, $X_{n1}, \ldots, X_{nn}$ are independent with distribution F_n. Let $\mathbb{F}_n$ be the empirical distribution function of $X_{n1}, \ldots, X_{nn}$. If F_n converges uniformly to F, then $\sqrt{n}(\mathbb{F}_n - F_n)$ converges weakly in $\mathcal{L}_\infty(\mathbb{R}^p)$ to an F-Gaussian field.*

Proof. For $x, y \in \mathbb{R}^p$ and $n \geq 1$, let $R_n(x,y) = F_n(x \wedge y) - F_n(x)F_n(y)$ and also define $R(x,y) = F(x \wedge y) - F(x)F(y)$. Let ρ be the pseudometric of Lemma 9.6.2 corresponding to the cumulative distribution function F. Let Z_{nk} be the random field indexed by $\mathbb{R}^p$ given by $Z_{nk}(x) = 1_{I_x}(X_{nk})/\sqrt{n}$, where we as usual define $I_x = (-\infty, x_1] \times \cdots \times (-\infty, x_p]$. We then have

$$\begin{aligned}\sum_{k=1}^{n} Z_{nk}(x) - EZ_{nk}(x) =& \frac{1}{\sqrt{n}} \sum_{k=1}^{n} 1_{I_x}(X_{nk}) - F_n(x) \\ =& \sqrt{n}(\mathbb{F}_n(x) - F_n(x)). \end{aligned} \tag{9.53}$$

We will apply Theorem 2.11.1 of [170] to prove that $\sum_{k=1}^{n} Z_{nk} - EZ_{nk}$ and thus $\sqrt{n}(\mathbb{F}_n - F_n)$ converges weakly in $\mathcal{L}_\infty(\mathbb{R}^p)$. We may assume without loss of generality that all variables are defined on a product probability space as described in Section 2.11.1 of [170], and as the fields (Z_{nk}) can be constructed using only countably many variables, the measurability requirements in Theorem 2.11.1 of [170] can be ensured. In order to apply Theorem 2.11.1 of [170], first note that by Lemma 9.6.2, $(\mathbb{R}^p, \rho)$ is totally bounded and so can be applied in Theorem 2.11.1 of [170]. Also, the covariance function of $\sum_{k=1}^{n} Z_{nk} - EZ_{nk}$ is

$$\begin{aligned}&\operatorname{Cov}\left(\sum_{k=1}^{n} Z_{nk}(x) - EZ_{nk}(x), \sum_{k=1}^{n} Z_{nk}(y) - EZ_{nk}(y)\right) \\ =& \sum_{k=1}^{n}\sum_{i=1}^{n} EZ_{nk}(x)Z_{ni}(y) - EZ_{nk}(x)EZ_{ni}(y) \\ =& \frac{1}{n}\sum_{k=1}^{n} E1_{I_x}(X_{nk})1_{I_y}(X_{nk}) - E1_{I_x}(X_{nk})E1_{I_y}(X_{nk}) \\ =& F_n(x \wedge y) - F_n(x)F_n(y) = R_n(x,y). \end{aligned} \tag{9.54}$$

Note that

$$\begin{aligned}|R(x,y) - R_n(x,y)| \leq& |F(x \wedge y) - F_n(x \wedge y)| + |F(x)F(y) - F_n(x)F_n(y)| \\ \leq& |F(x \wedge y) - F_n(x \wedge y)| + |F(x) - F_n(x)| + |F_n(y) - F_n(y)|, \end{aligned} \tag{9.55}$$

so as F_n converges uniformly to F, R_n converges uniformly to R. Thus, the covariance functions of $\sum_{k=1}^{n} Z_{nk} - EZ_{nk}$ converge to R. Therefore, in order to apply Theorem 2.11.1 of [170], it only remains to confirm that the conditions of (2.11.2) in [170] hold. Fixing $\eta > 0$, we have

$$\begin{aligned}\sum_{k=1}^{n} E\|Z_{nk}\|_\infty^2 1_{(\|Z_{nk}\|_\infty > \eta)} =& \frac{1}{n}\sum_{k=1}^{n} E1_{I_x}(X_{nk})1_{(1_{I_x}(X_{nk}) > \sqrt{n}\eta)} \\ \leq& P(1_{I_x}(X_{n1}) > \sqrt{n}\eta), \end{aligned}$$

and so it is immediate that the first condition of (2.11.2) in [170] holds. Next, define $d_n^2(x,y) = \sum_{k=1}^n (Z_{nk}(x) - Z_{nk}(y))^2$. We then also have for $x, y \in \mathbb{R}^p$ that

$$d_n^2(x,y) = \frac{1}{n} \sum_{k=1}^n (1_{I_x}(X_{nk}) - 1_{I_y}(X_{nk}))^2, \tag{9.56}$$

and therefore, $Ed_n(x,y)^2 = F_n(x) + F_n(y) - 2F_n(x \wedge y)$. Thus, $(x,y) \mapsto Ed_n(x,y)^2$ converges uniformly to ρ^2 on $\mathbb{R}^p \times \mathbb{R}^p$. Therefore, we conclude that for any sequence (δ_n) of positive numbers tending to zero, it holds for all $\eta > 0$ that

$$\begin{aligned} \limsup_{n\to\infty} \sup_{x,y:\rho(x,y)\le\delta_n} Ed_n^2(x,y) &\le \limsup_{n\to\infty} \sup_{x,y:\rho(x,y)\le\delta_n} \rho(x,y)^2 \\ &\le \limsup_{n\to\infty} \delta_n^2 = 0. \end{aligned} \tag{9.57}$$

Hence, the second condition of (2.11.2) in [170] holds. In order to verify the final condition of (2.11.2) in [170], first note that by (9.56), $d_n(x,y)^2 = E_{\mathbb{P}_n}(1_{I_x} - 1_{I_y})^2$, where $E_{\mathbb{P}_n}$ denotes integration with respect to $\mathbb{P}_n$ and $\mathbb{P}_n$ is the empirical measure on $(\mathbb{R}^p, \mathcal{B}_p)$ in $X_{n1}, \ldots, X_{nn}$. Thus, $d_n(x,y)$ is the $\mathcal{L}^2(\mathbb{R}^p, \mathcal{B}_p, \mathbb{P}_n)$ distance between the mappings I_x and I_y, and so

$$N(\varepsilon, \mathbb{R}^p, d_n) = N(\varepsilon, (1_{I_x})_{x\in\mathbb{R}^p}, \|\cdot\|_{2,\mathbb{P}_n}) \le \sup_Q N(\varepsilon, (1_{I_x})_{x\in\mathbb{R}^p}, \|\cdot\|_{2,Q}), \tag{9.58}$$

where $\|\cdot\|_{2,Q}$ denotes the norm on $\mathcal{L}^2(\mathbb{R}^p, \mathcal{B}_p, Q)$ and the supremum is over all probability measures Q on $(\mathbb{R}^p, \mathcal{B}_p)$. Thus, the third condition of (2.11.2) in [170] is satisfied if only it holds that for all sequences (δ_n) of positive numbers tending to zero,

$$\lim_{n\to\infty} \int_0^{\delta_n} \sup_Q \sqrt{\log N(\varepsilon, (1_{I_x})_{x\in\mathbb{R}^p}, \|\cdot\|_{2,Q})}\, \mathrm{d}\varepsilon = 0. \tag{9.59}$$

However, Theorem 2.6.7 of [170] yields a constant $K > 0$ such that for $0 < \varepsilon < 1$,

$$\sup_Q N(\varepsilon, (1_{I_x})_{x\in\mathbb{R}^p}, \|\cdot\|_{2,Q}) \le K(p+1)(16e)^{p+1}\varepsilon^{-2p}. \tag{9.60}$$

As a consequence, again for $0 < \varepsilon < 1$,

$$\sup_Q \sqrt{\log N(\varepsilon, (1_{I_x})_{x\in\mathbb{R}^p}, \|\cdot\|_{2,Q})} \le \sqrt{\log K(p+1)(16e)^{p+1} - 2p\log\varepsilon}. \tag{9.61}$$

By elementary calculations, we obtain for $0 < c < d < 1$ and $a, b > 0$ that

$$\int_c^d \sqrt{a - b\log x}\, \mathrm{d}x = \left[x\sqrt{a - b\log x} - \frac{e^{a/b}\sqrt{\pi b}}{2} \mathrm{erf}\left(\frac{\sqrt{a - b\log x}}{\sqrt{b}} \right) \right]_c^d, \tag{9.62}$$

where erf denotes the error function, $\mathrm{erf}(x) = (2/\sqrt{\pi}) \int_0^x \exp(-y^2)\, \mathrm{d}y$. Therefore, we conclude that for all $0 < \eta < 1$, the mapping $x \mapsto \sqrt{a - b\log x}$ is integrable over $[0, \eta]$. Thus, (9.59) holds. Recalling (9.53), Theorem 2.11.1 of [170] now shows that $\sqrt{n}(\mathbb{F}_n - F_n)$ converges weakly in $\mathcal{L}_\infty(\mathbb{R}^p)$. By uniqueness of the finite-dimensional distributions of the limit, we find that the limit is an F-Gaussian field. □

Proof of Theorem 9.4.2. By Lemma 9.6.3 and the continuous mapping theorem, $\sqrt{n}\|\mathbb{F}_n^A - F^A\|_\infty$ converges weakly to $\|W\|_\infty$. Therefore, equation (9.21) follows. In order to prove equation (9.22), consider A and B such that $F^A \neq F^B$ and let $\|F^A - F^B\|_\infty = \alpha$. Whenever $\|\mathbb{F}_n^A - F^A\|_\infty \leq \alpha/2$, the reverse triangle inequality yields

$$\begin{aligned}\|\mathbb{F}_n^A - F^B\|_\infty =& \|\mathbb{F}_n^A - F^A - (F^B - F^A)\|_\infty \\ \geq& |\|\mathbb{F}_n^A - F^A\|_\infty - \|F^B - F^A\|_\infty| \\ =& |\|\mathbb{F}_n^A - F^A\|_\infty - \alpha| \geq \alpha/2. \end{aligned} \tag{9.63}$$

Since $\lim_{n\to\infty} P(\|\mathbb{F}_n^A - F^A\|_\infty \leq \alpha/2) = 1$ by Lemma 9.6.3, we obtain

$$\begin{aligned}&\limsup_{n\to\infty} P(\sqrt{n}\|\mathbb{F}_n^A - F^B\|_\infty \leq c) \\ &\quad = \limsup_{n\to\infty} P(\|\mathbb{F}_n^A - F^B\|_\infty \leq c/\sqrt{n}, \|\mathbb{F}_n^A - F^A\|_\infty \leq \alpha/2) \\ &\quad \leq \limsup_{n\to\infty} P(\|\mathbb{F}_n^A - F^B\|_\infty \leq c/\sqrt{n}, \|\mathbb{F}_n^A - F^B\|_\infty \geq \alpha/2) = 0. \end{aligned} \tag{9.64}$$

Hence, $\lim_{n\to\infty} P(\sqrt{n}\|\mathbb{F}_n^A - F^B\|_\infty \leq c) = 0$ and so (9.22) holds. □

Proof of Theorem 9.4.3. First note that the triangle inequality yields

$$|\sqrt{n}\|\mathbb{F}_{\beta_n}^A - F_{\beta_n}^B\|_\infty - \sqrt{n}\|\mathbb{F}_{\beta_n}^A - F_{\beta_n}^A\|_\infty| \leq \sqrt{n}\|F_{\beta_n}^A - F_{\beta_n}^B\|_\infty. \tag{9.65}$$

Therefore, we have the inequalities

$$\begin{aligned} &P(\sqrt{n}\|\mathbb{F}_{\beta_n}^A - F_{\beta_n}^A\|_\infty - \sqrt{n}\|F_{\beta_n}^B - F_{\beta_n}^A\|_\infty > c) \\ \leq& P(\sqrt{n}\|\mathbb{F}_{\beta_n}^A - F_{\beta_n}^B\|_\infty > c) \\ \leq& P(\sqrt{n}\|\mathbb{F}_{\beta_n}^A - F_{\beta_n}^A\|_\infty + \sqrt{n}\|F_{\beta_n}^B - F_{\beta_n}^A\|_\infty > c). \end{aligned} \tag{9.66}$$

Let $\eta > 0$. By Corollary 9.3.4, we can choose $N \geq 1$ such that for $n \geq N$,

$$\sqrt{n}\|F_{\beta_n}^B - F_{\beta_n}^A\|_\infty \leq 4p(1+\eta)\sqrt{n}\beta_n\|\xi - \zeta\|_\infty. \tag{9.67}$$

By our assumptions, $\lim_n \sqrt{n}\beta_n = k$. Letting $\gamma > 0$, we then find for n large that

$$\sqrt{n}\|F_{\beta_n}^B - F_{\beta_n}^A\|_\infty \leq 4p(1+\eta)(k+\gamma)\|\xi - \zeta\|_\infty. \tag{9.68}$$

For such n, the first inequality of (9.66) yields

$$\begin{aligned} &P(\sqrt{n}\|\mathbb{F}_{\beta_n}^A - F_{\beta_n}^B\|_\infty > c) \\ \geq& P(\|\sqrt{n}(\mathbb{F}_{\beta_n}^A - F_{\beta_n}^A)\|_\infty > c + \sqrt{n}\|F_{\beta_n}^B - F_{\beta_n}^A\|_\infty) \\ \geq& P(\|\sqrt{n}(\mathbb{F}_{\beta_n}^A - F_{\beta_n}^A)\|_\infty > c + 4p(1+\eta)(k+\gamma)\|\xi - \zeta\|_\infty). \end{aligned} \tag{9.69}$$

Now recall from Theorem 9.3.1 that $F^A_{\beta_n}$ converges uniformly to F^A. Therefore, Lemma 9.6.3 and the continuous mapping theorem show that $\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^A_{\beta_n}\|_\infty$ converges weakly to $\|W\|_\infty$. As a consequence, (9.69) yields

$$\begin{aligned}&\liminf_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c)\\ \geq& P(\|W\|_\infty > c + 4p(1+\eta)(k+\gamma)\|\xi-\zeta\|_\infty).\end{aligned} \tag{9.70}$$

Letting η and then γ tend to zero, we obtain

$$\liminf_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \geq P(\|W\|_\infty > c + 4pk\|\xi-\zeta\|_\infty). \tag{9.71}$$

Similarly, the second inequality of (9.66) yields

$$\begin{aligned}&P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c)\\ \leq& P(\|\sqrt{n}(\mathbb{F}^A_{\beta_n} - F^A_{\beta_n})\|_\infty > c - \sqrt{n}\|F^B_{\beta_n} - F^A_{\beta_n}\|_\infty)\\ \leq& P(\|\sqrt{n}(\mathbb{F}^A_{\beta_n} - F^A_{\beta_n})\|_\infty \geq c - 4p(1+\eta)(k+\gamma)\|\xi-\zeta\|_\infty),\end{aligned} \tag{9.72}$$

and by similar arguments as previously, we obtain

$$\limsup_{n\to\infty} P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c) \leq P(\|W\|_\infty \geq c - 4pk\|\xi-\zeta\|_\infty). \tag{9.73}$$

Combining our results, we obtain (9.23). □

Proof of Corollary 9.4.4. As we have assumed that $P_e(\beta_n)$ is non-Gaussian, it follows from Lemma 9.3.5 that $F^A_\beta \neq F^B_\beta$, since $A \neq B\Lambda P$ for all diagonal Λ with $\Lambda^2 = I$ and all permutation matrices P. This shows (1). And as $AA^t = BB^t$ and ζ is Gaussian, Lemma 9.3.5 yields $F^A = F^B$, so Theorem 9.4.3 yields (2). □

Proof of Theorem 9.4.5. Note that for any $x \in \mathbb{R}^p$, we have

$$\begin{aligned}&P(\sqrt{n}\|\mathbb{F}^A_{\beta_n} - F^B_{\beta_n}\|_\infty > c)\\ \geq& P(\sqrt{n}|\mathbb{F}^A_{\beta_n}(x) - F^B_{\beta_n}(x)| > c)\\ =& P(|\sqrt{n}(\mathbb{F}^A_{\beta_n}(x) - F^A_{\beta_n}(x)) + \sqrt{n}(F^A_{\beta_n}(x) - F^B_{\beta_n}(x))| > c).\end{aligned} \tag{9.74}$$

We first consider the case $F^A \neq F^B$. Let $x \in \mathbb{R}^p$ be such that $F^A(x) \neq F^B(x)$. Then $\lim_n F^A_{\beta_n}(x) - F^B_{\beta_n}(x) \neq 0$, so $|\sqrt{n}(F^A_{\beta_n}(x) - F^B_{\beta_n}(x)|$ tends to infinity as n tends to infinity. By the central limit theorem, $\sqrt{n}(\mathbb{F}^A_{\beta_n}(x) - F^A_{\beta_n}(x))$ converges in distribution. Therefore, (9.74) yields the result.

Next, consider the case $F^A = F^B$ and $\Gamma_1(A) \neq \Gamma_1(B)$. Let $x \in \mathbb{R}^p$ be such that $\Gamma_1(A)(I_x) \neq \Gamma_1(B)(I_x)$. Similarly to the proof of Theorem 9.3.1, we define measures $\mu_0 = \zeta$, $\mu_1 = (\xi - \zeta)/\|\xi - \zeta\|_\infty$ and also define $S_k = \{\alpha \in \{0,1\}^p \,|\sum_{i=1}^p \alpha_i = k\}$ and $\Gamma_k(A) = \sum_{\alpha\in S_k} L_A(\otimes_{i=1}^p \mu_{\alpha_i})$. Note that $\Gamma_k(A)$ with $k = 1$ corresponds to (9.25). Then, we have

$$L_A(P_e(\beta)^{\otimes p}) = \sum_{k=0}^{\infty} \beta^k \|\xi-\zeta\|_\infty^k \Gamma_k(A), \tag{9.75}$$

see (9.34). In particular, we obtain

$$\begin{aligned} F_\beta^A(x) - F_\beta^B(x) =& L_A(P_e(\beta)^{\otimes p})(I_x) - L_B(P_e(\beta)^{\otimes p})(I_x) \\ =& \sum_{k=1}^{p} \beta^k \|\xi - \zeta\|_\infty^k (\Gamma_k(A)(I_x) - \Gamma_k(B)(I_x)), \end{aligned} \tag{9.76}$$

where we have used that $\Gamma_0(A) = \Gamma_0(B)$, since $F^A = F^B$. Since $\beta_n = n^{-\rho}$, we obtain

$$\sqrt{n}(F_{\beta_n}^A(x) - F_{\beta_n}^B(x)) = \sum_{k=1}^{p} n^{1/2-k\rho} \|\xi - \zeta\|_\infty^k (\Gamma_k(A)(I_x) - \Gamma_k(B)(I_x)). \tag{9.77}$$

As $\rho < 1/2$, $1/2 - \rho > 0$. As $\|\xi - \zeta\|_\infty(\Gamma_1(A)(I_x) - \Gamma_1(B)(I_x)) \neq 0$, we conclude that as n tends to infinity, the term corresponding to $k = 1$ in the above tends to infinity in absolute value. Since the right hand side of (9.77) is a sum with finitely many terms, where the remaining terms are of lower degree in n, we conclude that $|\sqrt{n}(F_{\beta_n}^A(x) - F_{\beta_n}^B(x))|$ tends to infinity as n tends to infinity. As in the previous case, since the ordinary central limit theorem shows that $\sqrt{n}(\mathbb{F}_{\beta_n}^A(x) - F_{\beta_n}^A(x))$ converges in distribution, (9.74) yields the result. □

10

On degrees of freedom in nonlinear regression

Niels Richard Hansen and Alexander Sokol

2010 Mathematics Subject Classification. Primary 62A01; Secondary 62F30.

Key words and phrases. Nonlinear regression, Model selection, Covariance penalty.

Abstract. The degrees of freedom of an estimator can be used to estimate the generalization error corresponding to the estimator. For the linear regression model, the degrees of freedom are known for several families of estimators, such as the ridge regression and LASSO estimators. We consider constrained and L^1-penalized estimators in nonlinear regression models and prove SURE-type results on the degrees of freedom. In particular, for the case of L^1-penalized estimation, we obtain an explicit candidate for an unbiased estimator of the degrees of freedom, allowing for sparse model selection in nonlinear regression.

10.1 Introduction

Consider the nonlinear regression model

$$Y = \varphi(\beta) + \varepsilon \tag{10.1}$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}^n$, $\beta \in \mathbb{R}^p$ and ε is multivariate normal with mean zero and variance matrix $\sigma^2 I_n$, where I_n denotes the identity matrix of order n. This is an

extension of the ordinary linear regression model in the sense that this latter model is recovered in the case $\varphi(\beta) = X\beta$ for some design matrix $X \in \mathbb{M}(n,p)$, where $\mathbb{M}(n,p)$ is the space of real $n \times p$ matrices. The motivation for our results is the problem of sparse estimation of β in the model (10.1) when observing the n-dimensional random variable Y and $\sigma^2 > 0$ is known.

Assume that we are in possession of a family of predictors or fitting procedures $g_\lambda : \mathbb{R}^n \to \mathbb{R}^n$ for $\lambda \geq 0$, meaing that $g_\lambda(Y)$ is an estimator of the mean of Y based on the variable Y. We imagine that λ is a tuning parameter measuring the level of complexity of our fitting procedure, in the sense that the complexity of g_λ is monotone in λ. We write g or g_λ instead of $g_\lambda(Y)$ when convenient. The mean squared prediction error of the predictor g_λ is

$$\mathrm{MSPE}_\beta(g_\lambda) = E_\beta \|Y - g_\lambda(Y)\|_2^2, \tag{10.2}$$

where $\|\cdot\|_2$ denotes the Euclidean norm and E_β denotes expectation given that β is the true parameter. The mean squared prediction error is a measure of the quality of the predictor g_λ. It can be estimated by the training error, given by

$$\mathrm{err}(g_\lambda) = \|Y - g_\lambda(Y)\|_2^2. \tag{10.3}$$

However, a more useful measure of the quality of the predictor g_λ is the generalization error, which is given by

$$\mathrm{Err}_\beta(g_\lambda) = E_\beta \|Y^* - g_\lambda(Y)\|_2^2, \tag{10.4}$$

where Y^* is independent of Y and has the same distribution as Y. The notation applied here is taken from [64]. The generalization error measures the expected performance of our predictor g when used for prediction in a new sample. In general, err is an downwards biased estimator of Err. This is known as the optimism of the training error, see Section 7.4 of [64]. Heuristically, this is because $g_\lambda(Y)$ is fit to the data set Y, and so err only measures the in-sample error and not the extra-sample error. We are interested in bias correction for the estimation of the generalization error using the training error. It is convenient to write

$$\mathrm{Err}_\beta(g_\lambda) = E_\beta \mathrm{err}(g_\lambda) + 2\sigma^2 \mathrm{df}_\beta(g_\lambda), \tag{10.5}$$

where $\mathrm{df}_\beta(g_\lambda)$ is known as the degrees of freedom of g. Estimating the degrees of freedom thus allows for bias correction for estimation of Err_β. To investigate the degrees of freedom, we employ both the covariance form of the degrees of freedom, see [44], as well as Stein's Unbiased Risk Estimate (SURE), introduced by [163].

To understand the usefulness of degrees of freedom, consider predictors based on two particular families of estimators of β. The first family is a family of constrained least squares estimators, given by the requirement that

$$\hat{\beta}_\lambda = \operatorname*{argmin}_{\beta \in B_\lambda} \|Y - \varphi(\beta)\|_2^2, \tag{10.6}$$

where B_λ is the closed ball of radius λ in $\mathbb{R}^p$ with center at the origin in the L^1-norm. The second family of estimators is the family of L^1-penalized least squares estimators, given by

$$\hat{\beta}_\lambda \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - \varphi(\beta)\|_2^2 + \lambda \|\beta\|_1, \tag{10.7}$$

where $\|\cdot\|_1$ is the L^1-norm on $\mathbb{R}^p$. The use of the set inclusion operator in (10.7) is necessary, as the argument minimum may not be unique. Both of these families of estimators are applicable for model selection, in the sense that they yield estimates where particular coordinates of β are exactly zero, see Chapter 3 of [64]. For any estimator $\hat{\beta}_\lambda$, we may obtain a predictor g_λ by putting $g_\lambda = \varphi \circ \hat{\beta}_\lambda$. Estimating the generalization error for these predictors will allow for choosing the λ minimizing the generalization error and thus yielding a method for model selection. Our objective is to derive expressions for the degrees of freedom of g_λ.

Our main results are two theorems, Theorem 10.3.1 and Theorem 10.3.2. Theorem 10.3.1 yields an abstract expression for the degrees of freedom in the case of constrained least squares estimators. Theorem 10.3.2 yields, under some regularity conditions, an expression for the degrees of freedom in the L^1-penalized case which is easily adaptable to practical calculations.

The remainder of the article is organized as follows. In Section 10.2, we introduce the degrees of freedom and its covariance and divergence forms, and review known results. In Section 10.3, we state and discuss our main results. Sections 10.4 and 10.5 contain proofs of the main results.

10.2 Calculation of the degrees of freedom

In this section, we review known results on the degrees of freedom. We first give a well-known lemma about the bias of the training error, see Section 2 of [44] for a proof.

Lemma 10.2.1. *Assume that Y and Y^* are independent n-dimensional variables with the same distribution and with finite variance. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be some measurable mapping. Assume that $g(Y)$ has finite variance. It then holds that*

$$E\|Y^* - g(Y)\|_2^2 = E\|Y - g(Y)\|_2^2 + 2\sum_{i=1}^{n} \operatorname{Cov}(Y_i, g_i(Y)). \tag{10.8}$$

To understand the meaning of Lemma 10.2.1, consider the model (10.1). Assume given a predictor $g : \mathbb{R}^n \to \mathbb{R}^n$ satisfying the regularity criteria of Lemma 10.2.1. Following for example [176, 168], if we now define the degrees of freedom of the

predictor g as

$$\mathrm{df}_\beta(g) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Cov}_\beta(Y_i, g_i(Y)), \tag{10.9}$$

then Lemma 10.2.1 shows that

$$\mathrm{Err}_\beta(g) = E_\beta\mathrm{err}(g) + 2\sigma^2\mathrm{df}_\beta(g). \tag{10.10}$$

The formula (10.10) shows that the bias of the training error when estimating the generalization error can be expressed through the degrees of freedom (10.9). For particular models and particular predictors, the degrees of freedom can be calculated explicitly. For example, by elementary calculations in the linear regression model with $\varphi(\beta) = X\beta$, it holds for $S \in \mathbb{M}(n,n)$ and a linear predictor of the form $g(y) = Sy$ that

$$\mathrm{df}_\beta(g) = \mathrm{tr}\ S, \tag{10.11}$$

where tr denotes the trace of a matrix. In particular, in the case of $p \leq n$ and g being the ordinary least squares predictor $g(y) = X(X^tX)^{-1}X^ty$, see Section 3.2 of [64], we obtain $\mathrm{df}_\beta(g) = \mathrm{tr}\ X(X^tX)^{-1}X^t = p$. For more complicated models, it is helpful to express the degrees of freedom in a different way. To this end, we follow [163] and call a function $g : \mathbb{R}^p \to \mathbb{R}$ almost differentiable if there exists a function $\nabla h : \mathbb{R}^p \to \mathbb{R}$ such that each ∇h_i is locally integrable and for all $y \in \mathbb{R}^p$, it holds that

$$h(x+y) - h(x) = \sum_{i=1}^{p}\int_0^1 y_i\nabla h_i(x+ty)\,\mathrm{d}t. \tag{10.12}$$

for Lebesgue almost all $x \in \mathbb{R}^p$. In this case, we refer to ∇h as the gradient of h, and we define the divergence of h as

$$\mathrm{div}\, h = \sum_{i=1}^{p}(\nabla h)_i, \tag{10.13}$$

The following result is shown in Lemma 2 of [163].

Lemma 10.2.2. *Let Y be a Gaussian variable with variance $\sigma^2 I_n$ and consider a measurable mapping $g : \mathbb{R}^n \to \mathbb{R}^n$ which is almost differentiable with $(\mathrm{div}\, g)(Y)$ having finite mean. It then holds that*

$$\mathrm{df}_\beta(g) = E_\beta(\mathrm{div}\, g)(Y). \tag{10.14}$$

The divergence form (10.14) of the degrees of freedom is known as Stein's Unbiased Risk Estimate (SURE). The following example shows that the requirement that g is almost differentiable in Lemma 10.2.2 in general cannot be exchanged with a simpler requirement such as for example simply having g be differentiable Lebesgue almost everywhere.

Example 10.2.3. Consider the case $n = 1$ and let Y be a Gaussian variable with mean zero and variance $\sigma^2 > 0$. Define g by

$$g(y) = \begin{cases} 0 & y < 0 \\ c(y) & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases} \tag{10.15}$$

where $c : [0, 1] \to [0, 1]$ is the Cantor function, see Chapter 2 of [21]. It then holds that g is continuous and Lebesgue almost everywhere differentiable with g' equal to zero. Therefore, $E_\beta(\operatorname{div} g)(Y) = E_\beta g'(Y) = 0$, while

$$\frac{1}{\sigma^2}\operatorname{Cov}_\beta(Y, g(Y)) = \frac{1}{\sigma^2}E_\beta Y g(Y) \geq \frac{1}{\sigma^2}P_\beta(Y \geq 1) > 0. \tag{10.16}$$

In particular, the divergence formula (10.14) does not hold for g. ∘

Now consider the case where, instead of only a predictor, we have an estimator $\hat{\beta}$ of β at our disposal. In this case, we may construct a predictor g by putting $g = \varphi \circ \hat{\beta}$. When the regularity criteria of Lemma 10.2.1 and Lemma 10.2.2 are satisfied, this leads to the relationship

$$\operatorname{Err}(\varphi \circ \hat{\beta}) = E_\beta \operatorname{err}(\varphi \circ \hat{\beta}) + 2\sigma^2 E_\beta \operatorname{div}(\varphi \circ \hat{\beta})(Y). \tag{10.17}$$

The formula (10.17) implies that we may use $\operatorname{err}(\varphi \circ \hat{\beta}) + 2\sigma^2 \operatorname{div}(\varphi \circ \hat{\beta})$ to estimate the generalization error. The challenge in this respect is to obtain a simple expression for $\operatorname{div}(\varphi \circ \hat{\beta})$ which may be used for practical calculations.

In general, when we have an estimator $\hat{\beta}$ at our disposal, we write

$$\operatorname{df}_\beta(\hat{\beta}) = \operatorname{df}_\beta(\varphi \circ \hat{\beta}), \tag{10.18}$$

meaning that the degrees of freedom of the estimator is the degrees of freedom of the corresponding predictor. As $\hat{\beta}$ takes its values in $\mathbb{R}^p$ and $\varphi \circ \hat{\beta}$ takes its values in $\mathbb{R}^n$, this notational ambiguity should not cause any confusion.

Next, we review previous work relevant to us. On an abstract level, our focus is essentially model selection. By model selection, we mean the process of selecting among a set of possible models. For a statistical model, any sparse estimator yields a submodel of the original model corresponding to the reduced model where a set of the original parameters are set to zero. Our interest in the generalization error will often center around being able to perform model selection by selecting a sparse estimator minimizing the generalization error across a family of estimators. The subject of model selection is large, see [26, 20] for monographs giving an overview of the field. The field covers both for example early criteria for linear regression such as Mallow's C_p, see [114], as well as general information criteria such as AIC and BIC, introduced in [4] and [152] respectively. Of greater interest to us are criteria derived from the training error optimism and its covariance and divergence representations,

see for example [40, 44, 110, 118, 181, 168, 176]. We are mostly interested in results concerning the explicit calculation of the the degrees of freedom (10.9) in particular models.

Explicit expressions for the degrees of freedom are known in several models. Consider for example a linear regression model of the form

$$Y = X\beta + \varepsilon \tag{10.19}$$

where ε follows a normal distribution with mean zero and known variance $\sigma^2 I_n$, and $\beta \in \mathbb{R}^p$. As mentioned earlier, for a linear predictor of the form $g(y) = Sy$, where $S \in \mathbb{M}(n, n)$, it holds that

$$\mathrm{df}_\beta(g) = \mathrm{tr}\ S, \tag{10.20}$$

see Chapter 3 of [65]. In particular, this yields an explicit expression for the degrees of freedom for the ridge regression predictor. Ridge regression is a shrinkage method, with the ridge regression estimator being defined as

$$\hat{\beta}_\lambda^{rid} = \underset{\beta \in \mathbb{R}^p}{\mathrm{argmin}} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2, \tag{10.21}$$

for some $\lambda \geq 0$. Section 3.4 of [64] shows that

$$\mathrm{df}_\beta(\hat{\beta}_\lambda^{rid}) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}, \tag{10.22}$$

where d_j are the singular values of the design matrix X. More difficult is the problem of calculating degrees of freedom corresponding to the LASSO estimator, given by

$$\hat{\beta}_\lambda^{L} = \underset{\beta \in \mathbb{R}^p}{\mathrm{argmin}} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1. \tag{10.23}$$

The LASSO is an estimation method designed for variable selection by yielding sparse estimates, see [167]. In this context, therefore, model selection through choice of the shrinkage parameter λ is of particular interest. The degrees of freedom for $\hat{\beta}_\lambda^L$ has been calulcated in [181, 168]. To review the results, we introduce the active set of $\hat{\beta}_\lambda^L$, given by

$$\mathcal{A} = \{i \leq p \mid (\hat{\beta}_\lambda^L)_i \neq 0\}, \tag{10.24}$$

that is, the coordinates of the LASSO estimate which are not zero. Note that $\mathcal{A}$ depends on both λ, X and most importantly the response variable Y. In particular, the dependence on Y renders $\mathcal{A}$ a random set. In [181], it is shown that when X has full column rank, then

$$\mathrm{df}_\beta(\hat{\beta}_\lambda^L) = E_\beta |\mathcal{A}|, \tag{10.25}$$

where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$. Note that as X is an $n \times p$ matrix, the assumption of full column rank forces $p \leq n$. In [168], the case of general p and n is considered. To state their results, we introduce the following notation: For an $n \times p$ matrix X, $X_{(R,C)}$ denotes the submatrix of X corresponding to the rows $R \subseteq \{1, \ldots, n\}$ and the columns $C \subseteq \{1, \ldots, p\}$. The notation $X_{(R,\cdot)}$ and $X_{(\cdot,C)}$ corresponds to submatrices with all colums and all rows, respectively. In [168], the authors show that

$$\mathrm{df}_\beta(\hat{\beta}^L_\lambda) = E_\beta \operatorname{rank} X_{(\cdot,\mathcal{A})}. \tag{10.26}$$

This result is particularly remarkable because of what is not stated: In the case $p > n$, which is allowed in (10.26), the LASSO estimates may not be unique, and therefore, the active set $\mathcal{A}$ is not a priori well-defined. However, in [168] it is shown that almost surely, $\mathcal{A}$ is in fact independent of the LASSO solution considered. In the case of X having full column rank, (10.26) reduces to the result (10.25).

Less explicit results are known in another type of models, namely shape-restricted regression. In [118], simple linear regression models of the form (10.19) are considered, but the allowed parameters β are restricted to some set $\Omega \subseteq \mathbb{R}^p$ and the least squares estimate

$$\hat{\beta}^{res} = \operatorname*{argmin}_{\beta \in \Omega} \|Y - X\beta\|_2^2 \tag{10.27}$$

is analyzed. The authors of [118] consider the case where Ω is closed and convex, and show that in this case, $\hat{\beta}^{res}$ is an almost differentiable function of y, so that Lemma 10.2.2 may be invoked. In particular, the case where Ω is a convex polyhedron is considered in detail.

10.3 Main results

In this section, we state and discuss our main results, Theorem 10.3.1 and Theorem 10.3.2. While our movitating remarks in Section 10.1 and our review of known results in Section 10.2 were made in the context of a statistical model, such a model is in fact not necessary for the statement of our results. Rather, all we need is an n-dimensional Gaussian variable Y with variance $\sigma^2 I_n$, and a continuous mapping $\varphi : \mathbb{R}^p \to \mathbb{R}^n$. We therefore dismiss our previous model and simply consider given Y and φ. As in Section 10.2, the degrees of freedom of a predictor $g : \mathbb{R}^n \to \mathbb{R}^n$ is then defined to be

$$\mathrm{df}(g) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(Y_i, g_i(Y)). \tag{10.28}$$

We are now ready to present our results. Theorem 10.3.1 concerns constrained estimators for nonlinear regression, where the constraint is that β is restricted to a

compact set. In practical cases, this set could be chosen for example to be a closed ball centered at the origin in the L^1 or L^2 norm. Recall that a Radon measure on $(\mathbb{R}^p, \mathcal{B}_p)$, where $\mathcal{B}_p$ denotes the Borel-σ-algebra, is a positive measure which is finite on compact sets, see [52]. By $C_c^\infty(\mathbb{R}^n)$, we denote the space of mappings from $\mathbb{R}^n$ to $\mathbb{R}$ with compact support which are differentiable infinitely often.

Theorem 10.3.1. *Let K be a compact subset of $\mathbb{R}^p$. Assume that for each $y \in \mathbb{R}^n$, it holds that $\hat{\beta}(y) \in \operatorname{argmin}_{\beta \in K} \|y - \varphi(\beta)\|_2^2$ in such a way that $\hat{\beta}(y)$ is measurable function of y. Then, there exists a unique family of nonnegative Radon measures $(\mu^i)_{i \leq n}$ such that*

$$\operatorname{df}(\hat{\beta}) = \sum_{i=1}^{n} \int_{\mathbb{R}^n} \psi(y) \, \mathrm{d}\mu^i(y), \tag{10.29}$$

where ψ is the density of Y. The family of Radon measures is determined by having

$$\int_{\mathbb{R}^n} (\varphi \circ \hat{\beta})(y)_i \frac{\partial \phi}{\partial y_i}(y) \, \mathrm{d}y = -\int_{\mathbb{R}^n} \phi(y) \, \mathrm{d}\mu^i(x) \tag{10.30}$$

for all $\phi \in C_c^\infty(\mathbb{R}^n)$.

We take a moment to understand the content of Theorem 10.3.1. For any nonempty compact set $K \subseteq \mathbb{R}^n$, introduce the metric projection onto K as

$$\operatorname{pr}_K(x) = \operatorname*{argmin}_{y \in K} \|x - y\|_2^2. \tag{10.31}$$

The function pr_K is generally a multifunction, meaning that its values are sets in $\mathbb{R}^n$. As K is nonempty and compact, $\operatorname{pr}_K(x)$ is nonempty for all $x \in \mathbb{R}^n$. In Theorem 10.3.1, $(\varphi \circ \hat{\beta})(y) \in \operatorname{pr}_{\varphi(K)}(y)$ for all $y \in \mathbb{R}^n$. If sufficient regularity conditions were available, Lemma 10.2.2 would suggest that we could write

$$\operatorname{df}(\hat{\beta}) = E(\operatorname{div} \operatorname{pr}_{\varphi(K)})(Y). \tag{10.32}$$

Two obstacles to this exist. First, while the results of [8] show that, heuristically speaking, $\operatorname{pr}_{\varphi(K)}$ is differentiable Lebesgue almost everywhere, this is not formally well-defined as $\operatorname{pr}_{\varphi(K)}$ is not even a single-valued function. Second, even if $\operatorname{pr}_{\varphi(K)}$ were single-valued and differentiable Lebesgue almost everywhere, Lemma 10.2.2 requires almost differentiability, which is not implied by simply being differentiable Lebesgue almost everywhere, as Example 10.2.3 shows. The content of Theorem 10.3.1 is that by letting Radon measures take the place of ordinary derivatives, see Chapter 7 of [148] for details on this, we may surmount the two obstacles outlined and obtain a divergence-type expression for the degrees of freedom.

Our other main result, Theorem 10.3.2, is concerned with L^1-penalized estimation.

In order to state the theorem succintly, define $B \in \mathbb{M}(p,p)$ and $A \in \mathbb{M}(n,p)$ by

$$B_{ij}(y,\beta) = \sum_{k=1}^{n} \frac{\partial \varphi_k}{\partial \beta_i}(\beta)\frac{\partial \varphi_k}{\partial \beta_j}(\beta) - (y_k - \varphi(\beta)_k)\frac{\partial^2 \varphi_k}{\partial \beta_i \partial \beta_j}(\beta), \tag{10.33}$$

$$A_{ki}(\beta) = \frac{\partial \varphi_k}{\partial \beta_i}(\beta), \tag{10.34}$$

and, as in Section 10.2, let $\mathcal{A}$ denote the active set of an estimator.

Theorem 10.3.2. *Assume that φ is twice continuously differentiable. Fix $y \in \mathbb{R}^n$ and $\lambda \geq 0$. Assume that there exists a neighborhood U of y such that when defining $h : U \times \mathbb{R}^p \to \mathbb{R}$ by $h(y,\beta) = \frac{1}{2}\|y - \varphi(\beta)\|_2^2 + \lambda\|\beta\|_1$, we have that:*

1. *For all $z \in U$, there is a unique argument minimum $\hat{\beta}_\lambda(z)$ of $\beta \mapsto h(z,\beta)$.*
2. *The mapping $\hat{\beta}_\lambda : U \to \mathbb{R}^p$ is differentiable.*
3. *The active set $\mathcal{A} = \{i \leq p \mid g(z)_i \neq 0\}$ does not depend on z.*
4. *$H_{(\mathcal{A},\mathcal{A})}(y, \hat{\beta}_\lambda(y))$ is invertible, where $H(y,\beta)$ is the Hessian of the mapping $\beta \mapsto \|y - \varphi(\beta)\|_2^2$.*

Then $\varphi \circ \hat{\beta}_\lambda$ is differentiable at y, and

$$\operatorname{div}(\varphi \circ \hat{\beta}_\lambda)(y) = \operatorname{tr}\, A(y)_{(\cdot,\mathcal{A})} B(y, \hat{\beta}_\lambda)^{-1}_{(\mathcal{A},\mathcal{A})} A(y)^t_{(\cdot,\mathcal{A})}. \tag{10.35}$$

If the regularity conditions of Theorem 10.3.2 are satisfied for Lebesgue almost all y, and if Lemma 10.2.2 can be applied, (10.35) yields that

$$\operatorname{df}(\hat{\beta}_\lambda) = E_\beta \operatorname{tr}\, A(Y)_{(\cdot,\mathcal{A})} B(Y, \hat{\beta}_\lambda)^{-1}_{(\mathcal{A},\mathcal{A})} A(Y)^t_{(\cdot,\mathcal{A})}, \tag{10.36}$$

implying that

$$\widehat{\operatorname{Err}} = \operatorname{err} + 2\sigma^2 \operatorname{tr}\, A(Y)_{(\cdot,\mathcal{A})} B(Y, \hat{\beta}_\lambda)^{-1}_{(\mathcal{A},\mathcal{A})} A(Y)^t_{(\cdot,\mathcal{A})} \tag{10.37}$$

yields an unbiased estimate of the generalization error. Furthermore, (10.37) can be calculated even when the regularity conditions are not satisfied, and so the virtue of Theorem 10.3.2 is not just that it provides sufficient conditions for (10.35) to hold, but also that it generally provides a candidate for the degrees of freedom. Minimizing (10.37) for varying $\lambda \geq 0$ allows for sparse model selection for L^1-penalized estimation in nonlinear regression.

In order to relate Theorem 10.3.2 to the results of [181, 168] discussed in Section 10.2, consider the ordinary linear regression case of $\varphi(\beta) = X\beta$ for $X \in \mathbb{M}(n,p)$, and let $\hat{\beta}_\lambda \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$ be the LASSO estimator. In this case, we obtain $B(y,u) = X^tX$ and $A(u) = X$, yielding

$$\begin{aligned} \operatorname{df}(\hat{\beta}_\lambda) &= E \operatorname{tr}\, X_{(\cdot,\mathcal{A})}(X^tX)^{-1}_{(\mathcal{A},\mathcal{A})} X^t_{(\cdot,\mathcal{A})} \\ &= E \operatorname{tr}\, (X^tX)^{-1}_{(\mathcal{A},\mathcal{A})} X^t_{(\cdot,\mathcal{A})} X_{(\cdot,\mathcal{A})} = E|\mathcal{A}|, \end{aligned} \tag{10.38}$$

Note that for the calculation (10.38) to make sense, it is necessary that $(X^tX)_{(\mathcal{A},\mathcal{A})}$ is invertible. This is equivalent to having $(X^tX)_{(\mathcal{A},\mathcal{A})}$ be of full rank, which is equivalent to having $X_{(\cdot,\mathcal{A})}$ be of full rank. Therefore, (10.38) is in accordance with both (10.25) and (10.26).

10.4 Constrained nonlinear regression

In this section, we prove Theorem 10.3.1. Let $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ be continuous and consider the constrained least squares estimator

$$\hat{\beta}_\lambda(y) \in \operatorname*{argmin}_{\beta \in K} \|y - \varphi(\beta)\|_2^2, \tag{10.39}$$

where K is some compact subset of $\mathbb{R}^p$. It holds that $\|y - \varphi(\hat{\beta}(y))\|_2^2 \leq \|y - \varphi(\beta)\|_2^2$ for all $\beta \in K$. It then also holds that $\|y - \varphi(\hat{\beta}(y))\|_2^2 \leq \|x - y\|_2^2$ for all $x \in \varphi(K)$, so that

$$\varphi(\hat{\beta}) \in \operatorname*{argmin}_{x \in \varphi(K)} \|x - y\|_2^2, \tag{10.40}$$

where $\varphi(K)$ is a compact subset of $\mathbb{R}^n$ as K is compact and φ is continuous. As in Section 10.3, for any nonempty compact set $K \subseteq \mathbb{R}^p$, we now introduce the metric projection onto K as the multifunction

$$\mathrm{pr}_K(x) = \operatorname*{argmin}_{y \in K} \|x - y\|_2^2. \tag{10.41}$$

In special cases, pr_K is single-valued and thus corresponds to an ordinary function. This is for example the case when K in addition to being compact also is convex, see Theorem 4.10 of [150]. It is immediate that $\varphi(\hat{\beta}_\lambda)$ is an element of the metric projection of Y onto $\varphi(K)$. Therefore, we may understand the properties of $\varphi(\hat{\beta}_\lambda)$ using the properties of metric projections.

In addition to the metric projection, we also define the metric distance function of a compact set K by

$$d_K(x) = \inf_{y \in K} \|x - y\|_2. \tag{10.42}$$

Metric distance functions and metric projections are connected and well-studied objects. Main subjects of interest are questions of when the metric projection is single-valued, see for example [69, 106, 138, 179, 177], and questions of continuity and differentiability for both the metric distance and the metric projection, see for example [2, 53, 71, 100, 153, 154]. One of the main difficulties of working with the metric projection is that it generally is a multifunction. We will work with measurable selections of the multifunction, defined below. We will argue that any measurable selection of the metric projection is differentiable in a distributional sense, and that the derivative is a tempered distribution induced by a Radon measure. This will allow us to prove Theorem 10.3.1.

Definition 10.4.1. By a selection prs_K of pr_K, we mean a function $\mathrm{prs}_K : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{prs}_K(x) \in \mathrm{pr}_K(x)$ for all $x \in \mathbb{R}^n$. If prs_K is measurable, we say that prs_K is a measurable selection of pr_K.

In order to obtain our results, we first state a known result, first noted in [8]. For completeness, we include a proof. Recall that for $f : \mathbb{R}^n \to [-\infty, \infty]$ convex, the subdifferential at $x \in \mathbb{R}^n$ is $\partial f(x) = \{y \in \mathbb{R}^n \mid \forall\, z \in \mathbb{R}^n : f(z) - f(x) \geq y^t(z-x)\}$, see Section 23 of [140].

Lemma 10.4.2. *Let K be a nonempty compact set, put $f(x) = \|x\|_2/2 - d_K(x)^2/2$. The function f is convex, and its subdifferential satisfies $\mathrm{pr}_K(x) \subseteq \partial f(x)$.*

Proof. We first show that f is convex. To do so, first note that for all $x, y \in \mathbb{R}^n$, it holds that

$$\tfrac{1}{2}\|x\|_2^2 - \tfrac{1}{2}\|x-y\|_2^2 = \tfrac{1}{2}x^t x - \tfrac{1}{2}(x-y)^t(x-y) = x^t y - \tfrac{1}{2}y^t y, \tag{10.43}$$

and so it follows that

$$\begin{aligned} f(x) &= \tfrac{1}{2}\|x\|_2^2 - \tfrac{1}{2}\inf_{y\in K}\|x-y\|_2^2 \\ &= \sup_{y\in K} \tfrac{1}{2}\|x\|_2^2 - \tfrac{1}{2}\|x-y\|_2^2 \\ &= \sup_{y\in K} x^t y - \tfrac{1}{2}y^t y. \end{aligned} \tag{10.44}$$

Thus, f is a pointwise supremum of affine functions. Therefore, by Theorem 5.5 of [140], f is convex. Next, in order to prove the result on the subdifferential, let $y \in \mathrm{pr}_K(x)$. It then holds that $\|x-y\|_2 = d_K(x)$. Applying this and (10.43), we then have

$$\begin{aligned} f(z) &= \sup_{u\in K} z^t u - \tfrac{1}{2}u^t u = \sup_{u\in K} x^t u - \tfrac{1}{2}u^t u + (z-x)^t u \\ &\geq x^t y - \tfrac{1}{2}y^t y + (z-x)^t y = \tfrac{1}{2}\|x\|_2^2 - \tfrac{1}{2}\|x-y\|_2^2 + y^t(z-x) \\ &= f(x) + y^t(z-x), \end{aligned} \tag{10.45}$$

so $y \in \partial f(x)$, as required. □

We also make the following simple observation. For any subset A of $\mathbb{R}^n$, we let $\mathrm{diam}\, A$ denote the diameter of A, given by $\mathrm{diam}\, A = \sup_{x\in A}\|x\|_2$.

Lemma 10.4.3. *Let K be a nonempty compact set. Then*

$$\mathrm{diam}\, \mathrm{pr}_K(x) \leq 2\|x\|_2 + \inf_{u\in K}\|u\|_2. \tag{10.46}$$

Proof. Let $y \in \mathrm{pr}_K(x)$. We then have

$$\begin{aligned} \|y\|_2 &\leq \|x\|_2 + \|y - x\|_2 = \|x\|_2 + \inf_{u \in K} \|u - x\|_2 \\ &\leq 2\|x\|_2 + \inf_{u \in K} \|u\|_2 \end{aligned} \tag{10.47}$$

and the result follows. □

Applying Lemma 10.4.2, we obtain the following differentiability result. Let $C_c(\mathbb{R}^n)$ denote the set of continuous mappings from $\mathbb{R}^n$ to $\mathbb{R}$ with compact support, let $C_c^k(\mathbb{R}^n)$ denote the subset of $C_c(\mathbb{R}^n)$ of mappings which are k times continuously differentiable and let $C_c^\infty(\mathbb{R}^n)$ denote the subset of $C_c(\mathbb{R}^n)$ of mappings which are differentiable infinitely often. Note that for any measurable selection prs_K, Lemma 10.4.3 shows that prs_K has polynomial growth, and therefore in particular is Lebesgue integrable on compact sets. From this, we find that all integrals used below are well-defined.

Lemma 10.4.4. *Let prs_K be a measurable selection of pr_K. There exists a unique family of nonnegative Radon measures $(\mu^i)_{i \leq n}$ such that for $\phi \in C_c^2(\mathbb{R}^n)$, it holds that*

$$\int_{\mathbb{R}^n} \mathrm{prs}_K(x)_i \frac{\partial \phi}{\partial x_i}(x)\, \mathrm{d}x = -\int_{\mathbb{R}^n} \phi(x)\, \mathrm{d}\mu^i(x). \tag{10.48}$$

Proof. We first show uniqueness. Assume that there exists two families (μ^i) and (ν^i) of nonnegative Radon measures satisfying the properties of the theorem. Then

$$\int_{\mathbb{R}^n} \phi(x)\, \mathrm{d}\mu^i(x) = \int_{\mathbb{R}^n} \phi(x)\, \mathrm{d}\nu^i(x) \tag{10.49}$$

for all $\phi \in C_c^2(\mathbb{R}^n)$. Approximation with smooth functions yields that (10.49) also holds for $\phi \in C_c(\mathbb{R}^n)$. Theorem 2.14 of [150] then shows that $\mu^i = \nu^i$.

In order to prove existence, let f be the convex function from Lemma 10.4.2. By Theorem 6.3.2 of [52], there exists a unique family of nonnegative Radon measures μ^i such that for $\phi \in C_c^2(\mathbb{R}^n)$, it holds that

$$\int_{\mathbb{R}^n} f(x) \frac{\partial^2 \phi}{\partial x_i^2}\, \mathrm{d}x = \int_{\mathbb{R}^n} \phi(x)\, \mathrm{d}\mu^i(x). \tag{10.50}$$

Next, let A be the set of points $x \in \mathbb{R}^n$ where f is differentiable. By Theorem 6.3.1 of [52], f is Lipschitz continuous on all compact subsets of $\mathbb{R}^n$. Therefore, Theorem 3.1.2 of [52] shows that the complement of A has Lebesgue measure zero. Note that with e_i being the i'th unit vector of $\mathbb{R}^n$, we have for any $\psi \in C_c^1(\mathbb{R}^n)$ that

$$\begin{aligned} \int_A \left(\frac{f(x + he_i) - f(x)}{h} \right) \psi(x)\, \mathrm{d}x &= \int_{\mathbb{R}^n} \left(\frac{f(x + he_i) - f(x)}{h} \right) \psi(x)\, \mathrm{d}x \\ &= -\int_{\mathbb{R}^n} f(x) \frac{\psi(x) - \psi(x - he_i)}{h}\, \mathrm{d}x. \end{aligned} \tag{10.51}$$

Now let K' be a compact set such that ψ is zero outside of K'. Recalling that f is Lipschitz continuous on all compact subsets of $\mathbb{R}^n$, let C_1 be the Lipschitz constant of f over $\{x \in \mathbb{R}^n \mid d(K', x) \leq 1\}$. We then obtain for $h \leq 1$ that

$$\sup_{x \in K'} \left| \frac{f(x + he_i) - f(x)}{h} \right| \leq C_1. \tag{10.52}$$

Likewise, as $\psi \in C_c^1(\mathbb{R}^n)$, ψ is Lipschitz on all of $\mathbb{R}^n$. With C_2 denoting the Lipschitz constant, we have

$$\sup_{x \in K'} \left| \frac{\psi(x) - \psi(x - he_i)}{h} \right| \leq C_2. \tag{10.53}$$

Therefore, inserting the i'th partial derivative of ϕ in (10.51) and applying the dominated convergence theorem, we obtain

$$\int_A \frac{\partial f}{\partial x_i}(x) \frac{\partial \phi}{\partial x_i}(x) \, \mathrm{d}x = -\int_{\mathbb{R}^n} f(x) \frac{\partial^2 \phi}{\partial x_i^2}(x) \, \mathrm{d}x. \tag{10.54}$$

Now, for $x \in A$, it holds that $\partial f(x)$ consists of a single point, the gradient $\nabla f(x)$ of f at x. As $\mathrm{pr}_K(x) \subseteq \partial f(x)$ by Lemma 10.4.2, this implies that $\mathrm{pr}_K(x)$ consists of a single point, and thus $\mathrm{prs}_K(x) = \nabla f(x)$. Therefore,

$$\int_A \frac{\partial f}{\partial x_i}(x) \frac{\partial \phi}{\partial x_i} \, \mathrm{d}x = \int_A \mathrm{prs}_K(x)_i \frac{\partial \phi}{\partial x_i} \, \mathrm{d}x = \int_{\mathbb{R}^n} \mathrm{prs}_K(x)_i \frac{\partial \phi}{\partial x_i} \, \mathrm{d}x. \tag{10.55}$$

Combining (10.50), (10.54) and (10.55), we obtain

$$\int_{\mathbb{R}^n} \mathrm{prs}_K(x)_i \frac{\partial \phi}{\partial x_i} \, \mathrm{d}x = \int_{\mathbb{R}^n} \phi(x) \, \mathrm{d}\mu^i(x), \tag{10.56}$$

as required. □

Lemma 10.4.5. *Let prs_K be a measurable selection of pr_K, and let (μ^i) be the family of nonnegative Radon measures of Lemma 10.4.4. For each i, it holds that*

$$\int_{\mathbb{R}^n} \frac{1}{(1 + \|x\|_2^2)^N} \, \mathrm{d}\mu^i(x) < \infty, \tag{10.57}$$

when $N \geq 1 + (n-1)/2$.

Proof. First note that

$$\int_{\mathbb{R}^n} \frac{1}{(1 + \|x\|_2^2)^N} \, \mathrm{d}x = \int_0^\infty \frac{A_d r^{n-1}}{(1 + r^2)^N} \, \mathrm{d}r, \tag{10.58}$$

where A_d is the area of the unit sphere in n dimensions. This integral is finite whenever $2N - (n-1) \geq 2$, corresponding to $N \geq (n-1)/2$. Also recall from the proof of Lemma 10.4.4 that the measures μ^i satisfies

$$\int_{\mathbb{R}^n} f(x) \frac{\partial^2 \phi}{\partial x_i^2} \, \mathrm{d}x = \int_{\mathbb{R}^n} \phi(x) \, \mathrm{d}\mu^i(x), \tag{10.59}$$

for all $\phi \in C_c^2(\mathbb{R}^n)$, where f is the convex function from Lemma 10.4.2. Now fix $N \geq 1$ and define $h : \mathbb{R}^n \to \mathbb{R}$ by $h(x) = (1 + \|x\|_2^2)^{-N}$. Note that

$$\frac{\partial h}{\partial x_i}(x) = \frac{-2Nx_i}{(1 + \|x\|_2^2)^{N+1}}, \tag{10.60}$$

so that we obtain

$$\left|\frac{\partial h}{\partial x_i}(x)\right| \leq \frac{2N\|x\|_2}{(1 + \|x\|_2^2)^{N+1}} \leq \frac{2N}{(1 + \|x\|_2^2)^N}. \tag{10.61}$$

From this, we also obtain that

$$\frac{\partial^2 h}{\partial x_i^2} = -\frac{\partial}{\partial x_i}\frac{2Nx_i}{(1 + \|x\|_2^2)^{N+1}} = -\frac{2N}{(1 + \|x\|_2^2)^{N+1}} + \frac{4N(N+1)x_i^2}{(1 + \|x\|_2^2)^{N+2}}, \tag{10.62}$$

yielding

$$\left|\frac{\partial^2 h}{\partial x_i^2}\right| \leq \frac{2N}{(1 + \|x\|_2^2)^{N+1}} + \frac{4N(N+1)x_i^2}{(1 + \|x\|_2^2)^{N+2}} \leq \frac{4N(N+2)}{(1 + \|x\|_2^2)^N}. \tag{10.63}$$

Combining (10.61) and (10.63), we conclude that both h and its first-order and second-order partial derivatives are bounded by $x \mapsto 4N(N+2)(1 + \|x\|_2^2)^{-N}$. Now let $\psi \in C_c^\infty(\mathbb{R}^n)$ be such that for all $x \in \mathbb{R}^n$, $\psi(x/k)$ increases to one as k tends to infinity. Also define $\psi_k(x) = \psi(x/k)$. Then $x \mapsto h\psi_k$ is in $C_c^2(\mathbb{R}^n)$, where the multiplication is pointwise, and thus

$$\int_{\mathbb{R}^n} h\psi_k(x)\,\mathrm{d}\mu^i(x) = \int_{\mathbb{R}^n} f(x)\frac{\partial^2 h\psi_k}{\partial x_i^2}(x)\,\mathrm{d}x. \tag{10.64}$$

Next, note that

$$\begin{aligned}\frac{\partial^2 h\psi_k}{\partial x_i^2}(x) &= \frac{\partial}{\partial x_i}\left(\frac{\partial h}{\partial x_i}(x)\psi(x/k) + h(x)\frac{1}{k}\frac{\partial \psi}{\partial x_i}(x/k)\right)\\ &= \frac{\partial^2 h}{\partial x_i^2}(x)\psi(x/k) + \frac{\partial h}{\partial x_i}(x)\frac{1}{k}\frac{\partial \psi}{\partial x_i}(x/k)\\ &\quad + \frac{\partial h}{\partial x_i}(x)\frac{1}{k}\frac{\partial \psi}{\partial x_i}(x/k) + h(x)\frac{1}{k^2}\frac{\partial^2 \psi}{\partial x_i^2}(x/k).\end{aligned} \tag{10.65}$$

Also note that from the definition of f, there is $c, M > 0$ such that whenever x satisfies $\|x\|_2 \geq M$, it holds that $|f(x)| \leq c\|x\|_2^2$. Therefore, whenever $\|x\|_2 \geq M$, we obtain

$$\left|\frac{4N(N+2)f(x)}{(1 + \|x\|_2^2)^2}\right| \leq \frac{4N(N+2)c\|x\|_2^2}{(1 + \|x\|_2^2)^N} \leq \frac{4N(N+2)c}{(1 + \|x\|_2)^{N-1}}. \tag{10.66}$$

Combining (10.65) and (10.66) with our previous results from (10.61) and (10.63), we then obtain for $\|x\|_2 \geq M$ that

$$\left|f(x)\frac{\partial^2 h\psi_k}{\partial x_i \partial x_j}(x)\right| \leq \frac{C}{(1 + \|x\|_2^2)^{N-1}}, \tag{10.67}$$

where

$$C = 4N(N+2)c\left(\|\psi\|_\infty + 2\left\|\frac{\partial\psi}{\partial x_i}\right\|_\infty + \left\|\frac{\partial^2\psi}{\partial x_i^2}\right\|_\infty\right). \tag{10.68}$$

Finally, assume that $N \geq 1 + (n-1)/2$. By (10.67), we then find that

$$\int_{\mathbb{R}^n} 1_{(\|x\|_2 \geq M)} h\psi_k(x)\,\mathrm{d}\mu^i(x) \leq \int_{\mathbb{R}^n} \frac{C}{(1+\|x\|_2^2)^{(n-1)/2}}\,\mathrm{d}x, \tag{10.69}$$

which is finite by our earlier observations. As the monotone convergence theorem yields

$$\int_{\mathbb{R}^n} 1_{(\|x\|_2 \geq M)} h(x)\,\mathrm{d}\mu^i(x) = \lim_{k\to\infty} \int_{\mathbb{R}^n} 1_{(\|x\|_2 \geq M)} h\psi_k(x)\,\mathrm{d}\mu^i(x), \tag{10.70}$$

we conclude that h is integrable with respect to μ^i. □

The next theorem is our main result on differentiability of prs_K. We introduce some notation from the theory of distributions. We let $\mathcal{D}(\mathbb{R}^n)$, the space of test functions, denote $C_c^\infty(\mathbb{R}^n)$, and we let $\mathcal{S}(\mathbb{R}^n)$, the space of Schwartz functions, denote the space of those $\phi \in C^\infty(\mathbb{R}^n)$ such that

$$\sup_{|\alpha|\leq N} \sup_{x\in\mathbb{R}^n} (1+\|x\|_2^2)^N |(D_\alpha f)(x)| < \infty, \tag{10.71}$$

where we use standard multi-index notation and D_α denotes the partial derivative operator corresponding to the multi-index α. The spaces $\mathcal{D}(\mathbb{R}^n)$ and $\mathcal{S}(\mathbb{R}^n)$ are endowed with particular topologies whose details are not important to us here, see Chapter 7 of [148] for precise definitions.

Theorem 10.4.6. *Let* prs_K *be a measurable selection of* pr_K*. There exists a unique family of nonnegative Radon measures* (μ^i) *such that for* $\phi \in \mathcal{S}(\mathbb{R}^n)$*, it holds that*

$$\int_{\mathbb{R}^n} \mathrm{prs}_K(x)_i \frac{\partial\phi}{\partial x_i}(x)\,\mathrm{d}x = -\int_{\mathbb{R}^n} \phi(x)\,\mathrm{d}\mu^i(x). \tag{10.72}$$

Proof. Uniqueness follows as in the proof of Lemma 10.4.4. As for existence, note that by Lemma 10.4.4, there exists a family of nonnegative Radon measures μ^i such that (10.72) holds whenever $\phi \in C_c^2(\mathbb{R}^n)$, in particular, it holds for all elements $\phi \in \mathcal{D}(\mathbb{R}^n)$. It suffices to extend this to the case $\phi \in \mathcal{S}(\mathbb{R}^n)$. To this end, define mappings $S_i, T_i : \mathcal{D}(\mathbb{R}^n) \to \mathbb{R}$ by

$$S_i(\phi) = \int_{\mathbb{R}^n} \mathrm{prs}_K(x)_i \phi(x)\,\mathrm{d}x \tag{10.73}$$

$$T_i(\phi) = \int_{\mathbb{R}^n} \phi(x)\,\mathrm{d}\mu^i(x). \tag{10.74}$$

Now, by Lemma 10.4.3, prs_K has polynomial growth. Therefore, Example 7.12(c) of [148] shows that S_i has a continuous extension to $\mathcal{S}(\mathbb{R}^n)$. Likewise, by Lemma 10.4.5 it holds that that each measure μ^i satisfies that $\int_{\mathbb{R}^n}(1+\|x\|_2^2)^{-N}\,\mathrm{d}\mu^i(x)$ is finite for sufficiently large N. Therefore, by Example 7.12(b) of [148], it holds that T_i has a continuous extension to $\mathcal{S}(\mathbb{R}^n)$ as well.

Now let $D_i : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}(\mathbb{R}^n)$ denote the i'th first-order partial derivative operator. By Theorem 7.4(b) of [148], D_i is continuous. Therefore, defining $\Lambda_i : \mathcal{S}(\mathbb{R}^n) \to \mathbb{R}$ by

$$\Lambda_i = S_i \circ D_i + T_i, \tag{10.75}$$

we find that Λ_i is continuous, and by (10.72) for $\phi \in \mathcal{D}(\mathbb{R}^n)$, Λ_i is zero on $\mathcal{D}(\mathbb{R}^n)$. By Theorem 7.10 of [148], $\mathcal{D}(\mathbb{R}^n)$ is dense in $\mathcal{S}(\mathbb{R}^n)$. As a consequence, Λ_i is zero on all of $\mathcal{S}(\mathbb{R}^n)$ and so (10.72) holds for all $\phi \in \mathcal{S}(\mathbb{R}^n)$. □

The meaning of Theorem 10.4.6 is best understood through the theory of distributions. A distribution, or generalized function, is a continuous linear functional T on $\mathcal{D}(\mathbb{R}^n)$. If T can be extended to a continuous linear functional on $\mathcal{S}(\mathbb{R}^n)$, we say that T is a tempered distribution. As in [148], we may associate a distribution $T_i : \mathcal{D}(\mathbb{R}^n) \to \mathbb{R}$ to $(\mathrm{prs}_K)_i$ by defining

$$T_i(\phi) = \int_{\mathbb{R}^n} \mathrm{prs}_K(x)_i \phi(x)\,\mathrm{d}x, \tag{10.76}$$

where the integral is well-defined and T_i is continuous because of the polynomial growth property of Lemma 10.4.3. The derivative of a distribution is always well-defined and is given by the distribution D_jT_i defined by

$$D_jT_i(\phi) = -\int_{\mathbb{R}^n} \mathrm{prs}_K(x)_i \frac{\partial\phi}{\partial x_j}(x)\,\mathrm{d}x. \tag{10.77}$$

Furthermore, we may also associate a distribution S_i to any Radon measure μ^i by defining

$$S_i(\phi) = \int_{\mathbb{R}^n} \phi(x)\,\mathrm{d}\mu^i(x). \tag{10.78}$$

Succintly put, Theorem 10.4.6 then states that the i'th partial derivative derivative of the distribution corresponding to $\mathrm{prs}_K(x)_i$ is a tempered distribution corresponding to a nonnegative Radon measure μ^i. Using this result, we can prove Theorem 10.3.1.

Proof of Theorem 10.3.1. We have $\varphi \circ \hat{\beta} \in \mathrm{argmin}_{x\in\varphi(K)} \|x-y\|_2^2$, so $\varphi \circ \hat{\beta}$ is a measurable selection of $\mathrm{pr}_{\varphi(K)}$. By Theorem 10.4.6, there exists a unique family of nonnegative Radon measures (μ^i) such that

$$\int_{\mathbb{R}^n} (\varphi\circ\hat{\beta})(x)_i \frac{\partial\phi}{\partial x_i}(x)\,\mathrm{d}x = -\int_{\mathbb{R}^n} \phi(x)\,\mathrm{d}\mu^i(x) \tag{10.79}$$

for all $\phi \in \mathcal{S}(\mathbb{R}^n)$. With ψ denoting the density of Y and ξ denoting the mean of Y, we have $\psi \in \mathcal{S}(\mathbb{R}^n)$ and

$$\frac{\partial}{\partial y_i}\psi(y) = \frac{\partial}{\partial y_i}\frac{1}{\sqrt{2\pi\sigma^2}^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\xi_i)^2\right) = -\frac{y_i-\xi_i}{\sigma^2}\psi(y). \tag{10.80}$$

Therefore, (10.79) yields

$$\begin{aligned}\operatorname{Cov}(Y_i, (\varphi\circ\hat\beta)_i(Y)) &= \operatorname{Cov}(Y_i-\xi_i, (\varphi\circ\hat\beta)_i(Y))\\ &= \int_{\mathbb{R}^n} (\varphi\circ\hat\beta)_i(y)(y_i-\xi_i)\psi(y)\,\mathrm{d}y\\ &= -\sigma^2\int_{\mathbb{R}^n}(\varphi\circ\hat\beta)_i(y)\frac{\partial}{\partial y_i}\psi(y)\,\mathrm{d}y\\ &= \sigma^2\int_{\mathbb{R}^n}\psi(y)\,\mathrm{d}\mu^i(y)\end{aligned} \tag{10.81}$$

and so Lemma 10.2.1 allows us to conclude that

$$\operatorname{df}(g) = \frac{1}{\sigma^2}\sum_{i=1}^n \operatorname{Cov}(Y_i, (\varphi\circ\hat\beta)_i(Y)) = \sum_{i=1}^n\int_{\mathbb{R}^n}\psi(y)\,\mathrm{d}\mu^i(y), \tag{10.82}$$

as required. □

10.5 L^1-penalized nonlinear regression

In this section, we prove Theorem 10.3.2. As in Section 10.2, for $X \in \mathbb{M}(n,p)$, $X_{(R,C)}$ denotes the submatrix of X corresponding to the rows $R \subseteq \{1,\dots,n\}$ and the columns $C \subseteq \{1,\dots,p\}$. The following lemma is a type of implicit differentiation theorem for solution mappings of L^1-penalized minimization problems, compare with the classical results of Chapter 9 of [149].

Lemma 10.5.1. *Fix $y \in \mathbb{R}^n$ and $\lambda \geq 0$. Consider a neighborhood U of y and a mapping $f : U \times \mathbb{R}^p \to \mathbb{R}$. Define $h : U \times \mathbb{R}^p \to \mathbb{R}$ by $h(y,\beta) = f(y,\beta) + \lambda\|\beta\|_1$. Assume that the following hold:*

1. *For $y \in U$, the mapping $\beta \mapsto f(y,\beta)$ is twice continuously differentiable.*
2. *For $y \in U$, the argument minimum $\hat\beta(y)$ of $\beta \mapsto h(y,\beta)$ is unique.*
3. *The mapping $\hat\beta : U \to \mathbb{R}^p$ is differentiable.*
4. *The active set $\mathcal{A} = \{i \leq p \mid \hat\beta_i(y) \neq 0\}$ does not depend on y.*
5. *$H_{(\mathcal{A},\mathcal{A})}(y,\hat\beta(y))$ is invertible, where $H(y,\beta)$ is the Hessian of $\beta \mapsto f(y,\beta)$.*

Let $M(y)$ be the $\mathcal{A} \times n$ matrix whose (i,k) entry is

$$M(y)_{ik} = \frac{\partial^2 f}{\partial y_k \partial \beta_i}(y, \hat{\beta}(y)). \tag{10.83}$$

Then, the derivative $D\hat{\beta}(y) \in \mathbb{M}(p,n)$ of $\hat{\beta}$ at y satisfies:

$$D\hat{\beta}(y)_{ik} = -(H_{(\mathcal{A},\mathcal{A})}(y, \hat{\beta}(y))^{-1} M(y))_{ik} \text{ for } i \in \mathcal{A}, \tag{10.84}$$

$$D\hat{\beta}(y)_{ik} = 0 \text{ for } i \in \mathcal{A}^c. \tag{10.85}$$

Proof. The relationship (10.85) is immediate from our assumption that the active set $\mathcal{A}$ does not depend on y. To prove (10.84), fix $y \in U$. As minima of differentiable functions are stationary points, we find that as $\hat{\beta}(y)$ is the argument minimum of $\beta \mapsto h(y,\beta)$, it holds that for all coordinates i such that $\beta \mapsto h(y,\beta)$ is differentiable in the i'th coordinate at $\hat{\beta}(y)$, the derivative is zero. As $\beta \mapsto \|\beta\|_1$ is differentiable in the i'th coordinate precisely if that coordinate is nonzero, we obtain for $i \in \mathcal{A}$ the two relationships

$$\frac{\partial h}{\partial \beta_i}(y, \hat{\beta}(y)) = 0, \tag{10.86}$$

$$\frac{\partial h}{\partial \beta_i}(y, \hat{\beta}(y)) = \frac{\partial f}{\partial \beta_i}(y, \hat{\beta}(y)) + \lambda \mathrm{sgn}(\hat{\beta}(y)_i). \tag{10.87}$$

Using the differentiability of f, the chain rule allows us to conclude that for all $i \in \mathcal{A}$ and $k \leq n$,

$$\begin{aligned} 0 =& \frac{\partial}{\partial y_k}\left(\frac{\partial f}{\partial \beta_i}(y, \hat{\beta}(y)) + \lambda \mathrm{sgn}(\hat{\beta}(y)_i)\right) \\ =& \frac{\partial^2 f}{\partial y_k \partial \beta_i}(y, \hat{\beta}(y)) + \sum_{j=1}^{p} \frac{\partial^2 f}{\partial \beta_j \partial \beta_i}(y, \hat{\beta}(y)) \frac{\partial \hat{\beta}_j}{\partial y_k}(y). \end{aligned} \tag{10.88}$$

Next, by our assumptions, $\hat{\beta}_i(y)$ is zero for all $y \in U$ and $i \in \mathcal{A}^c$. Therefore, the derivative of $\hat{\beta}_i$ at y is zero for all $i \in \mathcal{A}^c$, and so (10.88) yields

$$0 = \frac{\partial^2 f}{\partial y_k \partial \beta_i}(y, \hat{\beta}(y)) + \sum_{j \in \mathcal{A}} \frac{\partial^2 f}{\partial \beta_j \partial \beta_i}(y, \hat{\beta}(y)) \frac{\partial \hat{\beta}_j}{\partial y_k}(y) \tag{10.89}$$

for all $k \leq n$ and $i \in \mathcal{A}$. Thus, $0 = M(y) + H_{(\mathcal{A},\mathcal{A})}(y, \hat{\beta}(y))(D\hat{\beta}(y))_{(\mathcal{A},\cdot)}$, and so

$$(D\hat{\beta}(y))_{(\mathcal{A},\cdot)} = -\, H_{(\mathcal{A},\mathcal{A})}(y, \hat{\beta}(y))^{-1} M(y), \tag{10.90}$$

proving (10.84). □

The proof of Theorem 10.3.2 now simply proceeds by applying Lemma 10.5.1 to a particular choice of f. As in Section 10.3, define matrices $B(y,u) \in \mathbb{M}(p,p)$ and $A(u) \in \mathbb{M}(n,p)$ by

$$B_{ij}(y,u) = \sum_{k=1}^{n} \frac{\partial \varphi_k}{\partial \beta_i}(u) \frac{\partial \varphi_k}{\partial \beta_j}(u) - (y_k - \varphi(u)_k) \frac{\partial^2 \varphi_k}{\partial \beta_i \partial \beta_j}(u), \tag{10.91}$$

$$A_{ki}(u) = \frac{\partial \varphi_k}{\partial \beta_i}(u). \tag{10.92}$$

Proof of Theorem 10.3.2. Let $f(y,\beta) = \|y - \varphi(\beta)\|_2^2$. By Lemma 10.5.1, the formulas (10.84) and (10.85) hold for $\hat{\beta}$ at y. In particular, $\varphi \circ \hat{\beta}$ is differentiable at y, and

$$\begin{aligned} \operatorname{div}(\varphi \circ \hat{\beta})(y) =& \sum_{k=1}^{n} \frac{\partial}{\partial y_k} (\varphi \circ \hat{\beta})_k(y) = \sum_{k=1}^{n} \sum_{i=1}^{p} \frac{\partial \varphi_k}{\partial \beta_i}(\hat{\beta}(y)) \frac{\partial \hat{\beta}_i}{\partial y_k}(y) \\ =& \sum_{k=1}^{n} \sum_{i \in \mathcal{A}} \frac{\partial \varphi_k}{\partial \beta_i}(\hat{\beta}(y)) \frac{\partial \hat{\beta}_i}{\partial y_k}(y) = \sum_{k=1}^{n} \sum_{i \in \mathcal{A}} A(y)_{ki} \frac{\partial \hat{\beta}_i}{\partial y_k}(y). \end{aligned} \tag{10.93}$$

Fix $i, j \in \mathcal{A}$, we then have

$$\begin{aligned} \frac{\partial^2 f}{\partial \beta_i \partial \beta_j} f(y,\beta) =& \frac{\partial^2 f}{\partial \beta_i \partial \beta_j} \sum_{k=1}^{n} (y_k - \varphi(\beta)_k)^2 \\ =& -2 \frac{\partial f}{\partial \beta_i} \sum_{k=1}^{n} (y_k - \varphi(\beta)_k) \frac{\partial \varphi_k}{\partial \beta_j}(\beta) \\ =& 2 \sum_{k=1}^{n} \frac{\partial \varphi_k}{\partial \beta_i}(\beta) \frac{\partial \varphi_k}{\partial \beta_j}(\beta) - (y_k - \varphi(\beta)_k) \frac{\partial^2 \varphi_k}{\partial \beta_i \partial \beta_j}(\beta), \end{aligned} \tag{10.94}$$

and for $i \in \mathcal{A}$ and $k \le n$,

$$\frac{\partial^2 f}{\partial y_k \partial \beta_i}(y, \hat{\beta}(y)) = -2 \frac{\partial}{\partial y_k} \sum_{m=1}^{n} (y_m - \varphi(\beta)_m) \frac{\partial \varphi_m}{\partial \beta_j}(\beta) = -2 \frac{\partial \varphi_k}{\partial \beta_j}(\beta). \tag{10.95}$$

Therefore, by Lemma 10.5.1, we have for $k \le n$ and $i \in \mathcal{A}$ that

$$\frac{\partial \hat{\beta}_i}{\partial y_k}(y) = (B(y)^{-1} A(y)^t)_{ik}. \tag{10.96}$$

Using this in (10.93), we obtain

$$\begin{aligned} \operatorname{div}(\varphi \circ \hat{\beta})(y) =& \sum_{k=1}^{n} \sum_{i \in \mathcal{A}} A(y)_{ki} (B(y)^{-1} A(y)^t)_{ik} \\ =& \sum_{k=1}^{n} (A(y) B(y)^{-1} A(y)^t)_{kk} = \operatorname{tr} A(y) B(y)^{-1} A(y)^t, \end{aligned} \tag{10.97}$$

as desired. □

Bibliography

[1] O. O. Aalen, K. Røysland, J. M. Gran, and B. Ledergerber, *Causality, mediation and time: A dynamic viewpoint*, J. Roy. Statist. Soc. Ser. A **175** (2012), no. 4, 831–861.

[2] T. J. Abatzoglou, *The minimum norm projection on C^2-manifolds in $\mathbf{R}^n$*, Trans. Amer. Math. Soc. **243** (1978), 115–122.

[3] M. Abramowitz and I. A. Stegun (eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1984, Reprint of the 1972 edition, Selected Government Publications.

[4] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automatic Control **AC-19** (1974), 716–723, System identification and time-series analysis.

[5] S.-I. Amari and J.-F. Cardoso, *Blind source separation – semiparametric statistical approach*, IEEE Transactions on Signal Processing **45** (1997), no. 11, 2692–2700.

[6] D. Anderson and T. G. Kurtz, *Continuous time Markov chain models for chemical reaction networks*, Design and Analysis of Biomolecular Circuits, Springer, Heidelberg, 2011, pp. 3–42.

[7] D. Applebaum, *Lévy processes and stochastic calculus*, second ed., Cambridge Studies in Advanced Mathematics, vol. 116, Cambridge University Press, Cambridge, 2009.

[8] E. Asplund, *Differentiability of the metric projection in finite-dimensional Euclidean space*, Proc. Amer. Math. Soc. **38** (1973), 218–219.

[9] S. Azizpour, K. Giesecke, and G. Schwenkler, *Exploring the sources of default clustering*, Preprint (2012), 1–28.

[10] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, *Face recognition by independent component analysis*, IEEE Transactions on Neural Networks **13** (2002), no. 6, 1450–1464.

[11] R. F. Bass, *The Doob-Meyer decomposition revisited*, Canad. Math. Bull. **39** (1996), no. 2, 138–150.

[12] C. F. Beckmann and S. M. Smith, *Probabilistic independent component analysis for functional magnetic resonance imaging*, IEEE Transactions on Medical Imaging **23** (2004), no. 2, 137–152.

[13] M. Beiglböck, W. Schachermayer, and B. Veliyev, *A short proof of the Doob-Meyer theorem*, Stochastic Process. Appl. **122** (2012), no. 4, 1204–1209.

[14] K. Bichteler, *Stochastic integration with jumps*, Encyclopedia of Mathematics and its Applications, vol. 89, Cambridge University Press, Cambridge, 2002.

[15] T. Björk, *Arbitrage theory in continuous time*, Oxford University Press, 2009.

[16] J. M. Borwein and J. D. Vanderwerff, *Convex functions: constructions, characterizations and counterexamples*, Encyclopedia of Mathematics and its Applications, vol. 109, Cambridge University Press, Cambridge, 2010.

[17] P. Brémaud, *Point processes and queues*, Springer-Verlag, New York, 1981, Martingale dynamics, Springer Series in Statistics.

[18] A. N. Burkitt, *A review of the integrate-and-fire neuron model. I. Homogeneous synaptic input*, Biol. Cybernet. **95** (2006), no. 1, 1–19.

[19] ———, *A review of the integrate-and-fire neuron model. II. Inhomogeneous synaptic input and network properties*, Biol. Cybernet. **95** (2006), no. 2, 97–112.

[20] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference*, second ed., Springer-Verlag, New York, 2002, A practical information-theoretic approach.

[21] N. L. Carothers, *Real analysis*, Cambridge University Press, Cambridge, 2000.

[22] N. Cartwright, *What is wrong with Bayes nets?*, The Monist **84** (2001), no. 2, 242–264.

[23] A. Chen and P. J. Bickel, *Efficient independent component analysis*, Ann. Statist. **34** (2006), no. 6, 2825–2855.

[24] A. Cherny and A. N. Shiryaev, *On criteria for the uniform integrability of Brownian stochastic exponentials*, Optimal Control and Partial Differential Equations, IOS Press, 2001, pp. 80–92.

[25] J. R. Choksi, *Inverse limits of measure spaces*, Proc. London Math. Soc. (3) **8** (1958), 321–342.

[26] G. Claeskens and N. L. Hjort, *Model selection and model averaging*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2008.

[27] D. Commenges and A. Gégout-Petit, *A general dynamical statistical model with causal interpretation*, J. R. Stat. Soc. Ser. B Stat. Methodol. **71** (2009), no. 3, 719–736.

[28] P. Comon, *Independent component analysis, a new concept?*, Signal Processing **36** (1994), 287–314.

[29] Pierre Comon and Christian Jutten, *Handbook of blind source separation: Independent component analysis and applications*, Elsevier, Oxford, 2010.

[30] F. Comte and E. Renault, *Noncausality in continuous time models*, Econometric Theory **12** (1996), no. 2, 215–256.

[31] G. W. Cross, *Three types of matrix stability*, Linear Algebra and Appl. **20** (1978), no. 3, 253–263.

[32] M. Csörgő and L. Horváth, *A note on strong approximations of multivariate empirical processes*, Stochastic Process. Appl. **28** (1988), no. 1, 101–109.

[33] A. P. Dawid, *Causal inference without counterfactuals*, J. Amer. Statist. Assoc. **95** (2000), no. 450, 407–448, With comments and a rejoinder by the author.

[34] C. Dellacherie, *Quelques exemples familiers, en probabilités, d'ensembles analytiques, non Boréliens*, Séminaire de Probabilités, XII (Univ. Strasbourg, Strasbourg, 1976/1977), Lecture Notes in Math., vol. 649, Springer, Berlin, 1978, pp. 746–756.

[35] C. Dellacherie and P.-A. Meyer, *Probabilities and potential*, North-Holland Mathematics Studies, vol. 29, North-Holland Publishing Co., Amsterdam, 1978.

[36] ———, *Probabilities and potential. B*, North-Holland Mathematics Studies, vol. 72, North-Holland Publishing Co., Amsterdam, 1982, Theory of martingales, Translated from the French by J. P. Wilson.

[37] V. Didelez, *Graphical models for marked point processes based on local independence*, J. R. Stat. Soc. Ser. B Stat. Methodol. **70** (2008), no. 1, 245–264.

[38] F. Dinuzzo, M. Neve, G. De Nicolao, and U. P. Gianazza, *On the representer theorem and equivalent degrees of freedom of SVR*, J. Mach. Learn. Res. **8** (2007), 2467–2495.

[39] S. Ditlevsen and P. Greenwood, *The Morris–Lecar neuron model embeds a leaky integrate-and-fire model*, J. Math. Biol. **67** (2013), no. 2, 239–259.

[40] D. L. Donoho and I. M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc. **90** (1995), no. 432, 1200–1224.

[41] J. L. Doob, *Classical potential theory and its probabilistic counterpart*, Classics in Mathematics, Springer-Verlag, Berlin, 2001, Reprint of the 1984 edition.

[42] C. Dossal, M. Kachour, J. M. Fadili, G. Peyré, and C. Chesneau, *The degrees of freedom of the LASSO for general design matrix*, Preprint (2012), 1–17.

[43] R. M. Dudley, *Weak convergences of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces*, Illinois J. Math. **10** (1966), 109–126.

[44] B. Efron, *The estimation of prediction error: covariance penalties and cross-validation*, J. Amer. Statist. Assoc. **99** (2004), no. 467, 619–642, With comments and a rejoinder by the author.

[45] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist. **32** (2004), no. 2, 407–499, With discussion, and a rejoinder by the authors.

[46] M. Eichler, *Granger causality and path diagrams for multivariate time series*, J. Econometrics **137** (2007), no. 2, 334–353.

[47] ———, *Graphical modelling of multivariate time series*, Probab. Theory Related Fields **153** (2012), no. 1-2, 233–268.

[48] ———, *Causal inference with multiple time series: Principles and problems*, Preprint (2013), 1–19.

[49] M. Eichler and V. Didelez, *On Granger causality and the effect of interventions in time series*, Lifetime Data Anal. **16** (2010), no. 1, 3–32.

[50] J. Eriksson and V. Koivunen, *Identifiability, separability and uniqueness of linear ICA models*, IEEE Signal Processing Letters **11** (2004), no. 7, 601–604.

[51] S. N. Ethier and T. G. Kurtz, *Markov processes*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986, Characterization and convergence.

[52] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.

[53] S. Fitzpatrick, *Differentiation of real-valued functions and continuity of metric projections*, Proc. Amer. Math. Soc. **91** (1984), no. 4, 544–548.

[54] J.-P. Florens and D. Fougere, *Noncausality in continuous time*, Econometrica **64** (1996), no. 5, 1195–1212.

[55] H. Föllmer, *The exit measure of a supermartingale*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **21** (1972), 154–166.

[56] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, *Pathwise coordinate optimization*, Ann. Appl. Stat. **1** (2007), no. 2, 302–332.

[57] A. Gégout-Petit and D. Commenges, *A general definition of influence between stochastic processes*, Lifetime Data Anal. **16** (2010), no. 1, 33–44.

[58] K. Giesecke and G. Schwenkler, *Filtered likelihood for point processes*, Preprint (2011), 1–38.

[59] J. B. Gill and L. Petrović, *Causality and stochastic dynamic systems*, SIAM J. Appl. Math. **47** (1987), no. 6, 1361–1366.

[60] H. K. Gjessing, K. Røysland, E. A. Pena, and O. O. Aalen, *Recurrent events and the exploding Cox model*, Lifetime Data Anal. **16** (2010), no. 4, 525–546.

[61] P. Glasserman, *Monte Carlo methods in financial engineering*, Applications of Mathematics (New York), vol. 53, Springer-Verlag, New York, 2004, Stochastic Modelling and Applied Probability.

[62] C. W. J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica **37** (1969), no. 3, 424–38.

[63] G. Grubb, *Distributions and operators*, Graduate Texts in Mathematics, vol. 252, Springer, New York, 2009.

[64] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, second ed., Springer Series in Statistics, Springer, New York, 2009, Data mining, inference, and prediction.

[65] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*, Monographs on Statistics and Applied Probability, vol. 43, Chapman and Hall Ltd., London, 1990. MR 1082147 (92e:62117)

[66] S. W. He, J. G. Wang, and J. A. Yan, *Semimartingale theory and stochastic calculus*, Kexue Chubanshe (Science Press), Beijing, 1992.

[67] D. Hershkowitz and N. Keller, *Positivity of principal minors, sign symmetry and stability*, Linear Algebra Appl. **364** (2003), 105–124.

[68] N. J. Higham, *Functions of matrices*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008, Theory and computation.

[69] J.-B. Hiriart-Urruty, *Unsolved problems: At what points is the projection mapping differentiable?*, Amer. Math. Monthly **89** (1982), no. 7, 456–458.

[70] P. W. Holland, *Statistics and causal inference*, J. Amer. Statist. Assoc. **81** (1986), no. 396, 945–970, With discussion and a reply by the author.

[71] R. B. Holmes, *Smoothness of certain metric projections on Hilbert space*, Trans. Amer. Math. Soc. **184** (1973), 87–100.

[72] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985.

[73] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, *Nonlinear causal discovery with additive noise models*, Advances in Neural Information Processing Systems 21 (NIPS), MIT Press, 2009, pp. 689–696.

[74] A Hyvärinen, *Fast and robust fixed-point algorithms for independent component analysis*, IEEE Transactions on Neural Networks **10** (1999), no. 3, 626–634.

[75] ———, *Independent component analysis: Recent advances*, Phil. Trans. Roy. Soc. Ser. A **371** (2013), 1–19.

[76] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley-Blackwell, New York, 2001.

[77] P. Ilmonen and D. Paindaveine, *Semiparametrically efficient inference based on signed ranks in symmetric independent component models*, Ann. Statist. **39** (2011), no. 5, 2448–2476.

[78] K. Itô, *Stochastic integral*, Proc. Imp. Acad. Tokyo **20** (1944), 519–524.

[79] M. Izumisawa, T. Sekiguchi, and Y. Shiota, *Remark on a characterization of BMO-martingales*, Tôhoku Math. J. (2) **31** (1979), no. 3, 281–284.

[80] M. Jacobsen, *A brief account of the theory of homogeneous Gaussian diffusions in finite dimensions*, Frontiers in Pure and Applied Probability 1, TVP Science Publishers, 1993, pp. 86–94.

[81] ———, *Point process theory and applications*, Probability and its Applications, Birkhäuser Boston Inc., Boston, MA, 2006, Marked point and piecewise deterministic processes.

[82] J. Jacod, *Calcul stochastique et problèmes de martingales*, Lecture Notes in Mathematics, vol. 714, Springer, Berlin, 1979.

[83] J. Jacod and A. N. Shiryaev, *Limit theorems for stochastic processes*, second ed., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 288, Springer-Verlag, Berlin, 2003.

[84] A. Jakubowski, *An almost sure approximation for the predictable process in the Doob-Meyer decomposition theorem*, Séminaire de Probabilités XXXVIII, Lecture Notes in Math., vol. 1857, Springer, Berlin, 2005, pp. 158–164.

[85] R. Jarrow and P. E. Protter, *A short history of stochastic integration and mathematical finance: The early years, 1880–1970*, A festschrift for Herman Rubin, IMS Lecture Notes Monogr. Ser., vol. 45, Inst. Math. Statist., 2004, pp. 75–91.

[86] T. P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T.-W. Lee, and T. J. Sejnowski, *Imaging brain dynamics using independent component analysis*, Proceedings of the IEEE **89** (2001), no. 7, 1107–1122.

[87] O. Kallenberg, *Foundations of modern probability*, second ed., Probability and its Applications (New York), Springer-Verlag, New York, 2002.

[88] J. Kallsen and A. N. Shiryaev, *Time change representation of stochastic integrals*, Teor. Veroyatnost. i Primenen. **46** (2001), no. 3, 579–585.

[89] ———, *The cumulant process and Esscher's change of measure*, Finance Stoch. **6** (2002), no. 4, 397–428.

[90] R. L. Karandikar, *On pathwise stochastic integration*, Stochastic Process. Appl. **57** (1995), no. 1, 11–18.

[91] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*, Graduate Texts in Mathematics, vol. 113, Springer-Verlag, New York, 1988.

[92] A. F. Karr, *Point processes and their statistical inference*, second ed., Probability: Pure and Applied, vol. 7, Marcel Dekker Inc., New York, 1991.

[93] K. Kato, *On the degrees of freedom in shrinkage estimation*, J. Multivariate Anal. **100** (2009), no. 7, 1338–1352.

[94] N. Kazamaki, *Continuous exponential martingales and BMO*, Lecture Notes in Mathematics, vol. 1579, Springer-Verlag, Berlin, 1994.

[95] N. Kazamaki and T. Sekiguchi, *Uniform integrability of continuous exponential martingales*, Tohoku Math. J. (2) **35** (1983), no. 2, 289–301.

[96] R. W. Keener, *Theoretical statistics*, Springer Texts in Statistics, Springer, New York, 2010, Topics for a core course.

[97] D. Khoshnevisan, *Multiparameter processes*, Springer Monographs in Mathematics, Springer-Verlag, New York, 2002, An introduction to random fields.

[98] J. Komlós, *A generalization of a problem of Steinhaus*, Acta Math. Acad. Sci. Hungar. **18** (1967), 217–229.

[99] N. Krämer and M. Sugiyama, *The degrees of freedom of partial least squares regression*, J. Amer. Statist. Assoc. **106** (2011), 697–705.

[100] J. B. Kruskal, *Two convex counterexamples: A discontinuous envelope function and a nondifferentiable nearest-point mapping*, Proc. Amer. Math. Soc. **23** (1969), 697–703.

[101] N. Krylov, *A simple proof of a result of A. Novikov.*, Preprint (2009), 1–3.

[102] T. G. Kurtz, *Equivalence of stochastic equations and martingale problems*, Stochastic analysis 2010, Springer, Heidelberg, 2011, pp. 113–130.

[103] Y. A. Kutoyants, *Statistical inference for ergodic diffusion processes*, Springer Series in Statistics, Springer-Verlag London Ltd., London, 2004.

[104] P. Lancaster and M. Tismenetsky, *The theory of matrices*, second ed., Computer Science and Applied Mathematics, Academic Press Inc., Orlando, FL, 1985.

[105] P. Lansky and S. Ditlevsen, *A review of the methods for signal estimation in stochastic diffusion leaky integrate-and-fire neuronal models*, Biol. Cybernet. **99** (2008), no. 4-5, 253–262.

[106] K. S. Lau, *Almost Chebyshev subsets in reflexive Banach spaces*, Indiana Univ. Math. J. **27** (1978), no. 5, 791–795.

[107] S. L. Lauritzen, *Graphical models*, Oxford Statistical Science Series, vol. 17, The Clarendon Press Oxford University Press, New York, 1996, Oxford Science Publications.

[108] ———, *Causal inference from graphical models*, Complex stochastic systems (Eindhoven, 1999), Monogr. Statist. Appl. Probab., vol. 87, Chapman & Hall/CRC, Boca Raton, FL, 2001, pp. 63–107.

[109] D. Lépingle and J. Mémin, *Sur l'intégrabilité uniforme des martingales exponentielles*, Z. Wahrsch. Verw. Gebiete **42** (1978), no. 3, 175–203.

[110] K.-C. Li, *From Stein's unbiased risk estimates to the method of generalized cross validation*, Ann. Statist. **13** (1985), no. 4, 1352–1377.

[111] M. A. Lifshits, *Gaussian random functions*, Mathematics and its Applications, vol. 322, Kluwer Academic Publishers, Dordrecht, 1995.

[112] M. H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann, *Predicting causal effects in large-scale systems from observational data*, Nature Methods (2010), no. 4, 247248.

[113] M. H. Maathuis, M. Kalisch, and P. Bühlmann, *Estimating high-dimensional intervention effects from observational data*, Ann. Statist. **37** (2009), no. 6A, 3133–3164.

[114] C. L. Mallows, *Some comments on C_p*, Technometrics **15** (1973), 661–675.

[115] M. S. Masud and R. Borisyuk, *Statistical technique for analysing functional connectivity of multiple spike trains*, Journal of Neuroscience Methods **196** (2011), no. 1, 201–219.

[116] R. Meise and D. Vogt, *Introduction to functional analysis*, Oxford Graduate Texts in Mathematics, vol. 2, The Clarendon Press Oxford University Press, New York, 1997, Translated from the German by M. S. Ramanujan and revised by the authors.

[117] J. Mémin, *Décompositions multiplicatives de semimartingales exponentielles et applications*, Séminaire de Probabilités, XII (Univ. Strasbourg, Strasbourg, 1976/1977), Lecture Notes in Math., vol. 649, Springer, Berlin, 1978, pp. 35–46.

[118] M. Meyer and M. Woodroofe, *On the degrees of freedom in shape-restricted regression*, Ann. Statist. **28** (2000), no. 4, 1083–1104.

[119] A. Mijatović and M. Urusov, *On the martingale property of certain local martingales*, Probab. Theory Related Fields **152** (2012), no. 1-2, 1–30.

[120] J. M. Mooij and T. Heskes, *Cyclic causal discovery from continuous equilibrium data*, Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI-13), 2013, pp. 431–439.

[121] J. R. Norris, *Markov chains*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 2, Cambridge University Press, Cambridge, 1998, Reprint of 1997 original.

[122] M. Novey and T. Adah, *Complex ICA by negentropy maximization*, IEEE Transactions on Neural Networks **19** (2008), no. 4, 596–609.

[123] A. A. Novikov, *On an identity for stochastic integrals*, Theor. Probab. Appl. **17** (1972), 717–720.

[124] T. Okada, *A criterion for uniform integrability of exponential martingales*, Tôhoku Math. J. (2) **34** (1982), no. 4, 495–498.

[125] E. Ollila, H.-J. Kim, and V. Koivunen, *Compact Cramér-Rao bound expression for independent component analysis*, IEEE Transactions on Signal Processing **56** (2008), no. 4, 1421–1428.

[126] J. Pearl, *Causality*, second ed., Cambridge University Press, Cambridge, 2009, Models, reasoning, and inference.

[127] ———, *An introduction to causal inference*, Int. J. Biostat. **6** (2010), no. 2, Art. 7, 61.

[128] J. Peters, *Restricted structural equation models for causal inference*, ETH Zürich, Zürich, 2012.

[129] J. Peters and P. Bühlmann, *Identifiability of Gaussian structural equation models with same error variances*, Preprint (2012), 1–11.

[130] L. Petrović and S. Dimitrijević, *Invariance of statistical causality under convergence*, Statist. Probab. Lett. **81** (2011), no. 9, 1445–1448.

[131] L. Petrović and D. Stanojević, *Statistical causality, extremal measures and weak solutions of stochastic differential equations with driving semimartingales*, J. Math. Model. Algorithms **9** (2010), no. 1, 113–128.

[132] U. Picchini, S. Ditlevsen, A. De Gaetano, and P. Lansky, *Parameters of the diffusion leaky integrate-and-fire neuronal model for a slowly fluctuating signal*, Neural Computation **20** (2008), 2696–2714.

[133] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilinisky, and E. P. Simoncelli, *Spatio-temporal correlations and visual signaling in a complete neuronal population*, Nature **454** (2008), no. 7206, 995–999.

[134] P. E. Protter, *Stochastic integration and differential equations*, Stochastic Modelling and Applied Probability, vol. 21, Springer-Verlag, Berlin, 2005, Second edition. Version 2.1, Corrected third printing.

[135] P. E. Protter and K. Shimbo, *No arbitrage and general semimartingales*, Markov processes and related topics: a Festschrift for Thomas G. Kurtz, Inst. Math. Stat. Collect., vol. 4, Inst. Math. Statist., Beachwood, OH, 2008, pp. 267–283.

[136] K. M. Rao, *On decomposition theorems of Meyer*, Math. Scand. **24** (1969), 66–78.

[137] A. Rau, F. Jaffrézic, and G. Nuel, *Joint estimation of causal effects from observational and intervention gene expression data*, Preprint (2013), 1–18.

[138] S. Reich and A. J. Zaslavski, *Well-posedness and porosity in best approximation problems*, Topol. Methods Nonlinear Anal. **18** (2001), no. 2, 395–408.

[139] J. Robins, *A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect*, Math. Modelling **7** (1986), no. 9-12, 1393–1512, Mathematical models in medicine: diseases and epidemics, Part 2.

[140] R. T. Rockafellar, *Convex analysis*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1997, Reprint of the 1970 original, Princeton Paperbacks.

[141] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 317, Springer-Verlag, Berlin, 1998.

[142] L. C. G. Rogers and D. Williams, *Diffusions, Markov processes, and martingales. Vol. 1*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 2000, Foundations, Reprint of the second (1994) edition.

[143] ———, *Diffusions, Markov processes, and martingales. Vol. 2*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 2000, Itô calculus, Reprint of the second (1994) edition.

[144] K. Røysland, *A martingale approach to continuous-time marginal structural models*, Bernoulli **17** (2011), no. 3, 895–915.

[145] ———, *Counterfactual analyses with graphical models based on local independence*, Ann. Statist. **40** (2012), no. 4, 2162–2194.

[146] D. B. Rubin, *Estimating causal effects of treatments in randomized and nonrandomized studies*, Journal of Educational Psychology **66** (1974), 688–701.

[147] ———, *Bayesian inference for causal effects: The role of randomization*, Ann. Statist. **6** (1978), no. 1, 34–58.

[148] W. Rudin, *Functional analysis*, McGraw-Hill Book Co., New York, 1973, McGraw-Hill Series in Higher Mathematics.

[149] ———, *Principles of mathematical analysis*, third ed., McGraw-Hill Book Co., New York, 1976, International Series in Pure and Applied Mathematics.

[150] ———, *Real and complex analysis*, third ed., McGraw-Hill Book Co., New York, 1987.

[151] K.-I. Sato, *Lévy processes and infinitely divisible distributions*, Cambridge Studies in Advanced Mathematics, vol. 68, Cambridge University Press, Cambridge, 1999, Translated from the 1990 Japanese original, Revised by the author.

[152] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6** (1978), no. 2, 461–464.

[153] A. Shapiro, *On differentiability of metric projections in* $\mathbf{R}^n$*. I. Boundary case*, Proc. Amer. Math. Soc. **99** (1987), no. 1, 123–128.

[154] ———, *Existence and differentiability of metric projections in Hilbert spaces*, SIAM J. Optim. **4** (1994), no. 1, 130–141.

[155] X. Shen and J. Ye, *Adaptive model selection*, J. Amer. Statist. Assoc. **97** (2002), no. 457, 210–221.

[156] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, *A linear non-Gaussian acyclic model for causal discovery*, J. Mach. Learn. Res. **7** (2006), 2003–2030.

[157] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, *DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model*, J. Mach. Learn. Res. **12** (2011), 1225–1248.

[158] S. E. Shreve, *Stochastic calculus for finance. II*, Springer Finance, Springer-Verlag, New York, 2004, Continuous-time models.

[159] A. Sokol, *An elementary proof that the first hitting time of an open set by a jump process is a stopping time*, Séminaire de Probabilités, XLV, Springer, Berlin, 2013, pp. 301–304. Lecture Notes in Math., Vol. 2078.

[160] ———, *Optimal Novikov-type criteria for local martingales with jumps*, Elec. Comm. Prob. **18** (2013), 1–8.

[161] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*, second ed., Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2000, With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

[162] D. Steel, *Homogeneity, selection and the faithfulness condition*, Minds and Machines **16** (2006), 303–317.

[163] C. M. Stein, *Estimation of the mean of a multivariate normal distribution*, Ann. Statist. **9** (1981), no. 6, 1135–1151.

[164] D. J. Stekhoven, I. Moraes, G. Sveinbjörnsson, L. Hennig, M. H. Maathuis, and P. Bühlmann, *Causal stability ranking*, Bioinformatics **28** (2012), no. 21, 2819–2823.

[165] A. E. Taylor, *Introduction to functional analysis*, John Wiley & Sons Inc., New York, 1958.

[166] J. Tian, C. Kang, and J. Pearl, *A characterization of interventional distributions in semi-Markovian causal models*, Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, 2006, pp. 1239–1244.

[167] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 1, 267–288.

[168] R. J. Tibshirani and J. Taylor, *Degrees of freedom in LASSO problems*, Ann. Statist. **40** (2012), no. 2, 1198–1232.

[169] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, *A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects*, Journal of Neurophysiology **93** (2005), no. 2, 1074–1089.

[170] A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996, With applications to statistics.

[171] T. Verma and J. Pearl, *Equivalence and synthesis of causal models*, Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Elsevier, 1991, pp. 255–268.

[172] R. Vigario, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja, *Independent component approach to the analysis of EEG and MEG recordings*, IEEE Transactions on Biomedical Engineering **47** (2000), no. 5, 589–593.

[173] D. J. Wilkinson, *Stochastic modelling for systems biology*, Chapman & Hall/CRC Mathematical and Computational Biology Series, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[174] J. Woodward, *Causal independence and faithfulness*, Mulivariate Behavioural Research **33** (2003), 129–148.

[175] J. A. Yan, *À propos de l'intégrabilité uniforme des martingales exponentielles*, Seminar on Probability, XVI, Lecture Notes in Math., vol. 920, Springer, Berlin, 1982, pp. 338–347.

[176] J. Ye, *On measuring and correcting the effects of data mining and model selection*, J. Amer. Statist. Assoc. **93** (1998), no. 441, 120–131.

[177] L. Zajíček, *Differentiability of the distance function and points of multivaluedness of the metric projection in Banach space*, Czechoslovak Math. J. **33(108)** (1983), no. 2, 292–308.

[178] M. Zakai and J. Snyders, *Stationary probability measures for linear differential equations driven by white noise*, J. Differential Equations **8** (1970), 27–33.

[179] T. Zamfirescu, *The nearest point mapping is single valued nearly everywhere*, Arch. Math. (Basel) **54** (1990), no. 6, 563–566.

[180] K. Zhang and A. Hyvärinen, *On the identifiability of the post-nonlinear causal model*, Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Elsevier, 2009, pp. 647–655.

[181] H. Zou, T. Hastie, and R. Tibshirani, *On the "degrees of freedom" of the LASSO*, Ann. Statist. **35** (2007), no. 5, 2173–2192.

[182] D. Zwillinger, *CRC standard mathematical tables and formulae*, thirty-first ed., CRC Press, Boca Raton, FL, 2012.