NON-STATIONARY PROCESSES AND THEIR APPLICATION TO FINANCIAL HIGH-FREQUENCY DATA

Mailan Trinh

A thesis submitted for the degree of Doctor of Philosophy

University of Sussex

March 2018

UNIVERSITY OF SUSSEX

MAILAN TRINH

A THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NON-STATIONARY PROCESSES AND THEIR APPLICATION TO FINANCIAL HIGH-FREQUENCY DATA

SUMMARY

The thesis is devoted to non-stationary point process models as generalizations of the standard homogeneous Poisson process. The work can be divided in two parts.

In the first part, we introduce a fractional non-homogeneous Poisson process (FNPP) by applying a random time change to the standard Poisson process. We characterize the FNPP by deriving its non-local governing equation. We further compute moments and covariance of the process and discuss the distribution of the arrival times. Moreover, we give both finite-dimensional and functional limit theorems for the FNPP and the corresponding fractional non-homogeneous compound Poisson process. The limit theorems are derived by using martingale methods, regular variation properties and Anscombe's theorem. Eventually, some of the limit results are verified via a Monte-Carlo simulation.

In the second part, we analyze statistical point process models for durations between trades recorded in financial high-frequency trading data. We consider parameter settings for models which are non-stationary or very close to non-stationarity which is quite typical for estimated parameter sets of models fitted to financial data. Simulation, parameter estimation and in particular model selection are discussed for the following three models: a non-homogeneous normal compound Poisson process, the exponential autoregressive conditional duration model (ACD) and a Hawkes process model. In a Monte-Carlo simulation, we test the performance of the following information criteria for model selection: Akaike's information criterion, the Bayesian information criterion and the Hannan-Quinn information criterion. We are particularly interested in the relation between the rate of correct model selection and the underlying sample size. Our numerical results show that the model selection for the compound Poisson type model works best for small parameter numbers. Moreover, the results for Hawkes processes confirm the theoretical asymptotic distributions of model selection whereas for the ACD model the model selection exhibits adverse behavior in certain cases.

Declaration of Authorship

I hereby declare that this thesis has not been, and will not be, submitted in whole or in part to another University for the award of any other degree.

Mailan Trinh

28th March 2018

To my parents

Contents

Introduction

Motivation

Time series analysis as presented in standard textbooks like Brockwell and Davis 1991 or Hamilton 1994 assumes integer-indexed time series of the form $x_1, x_2, \ldots, x_n$. This assumption is most suitable for data and measurements that can be recorded at specific *equidistant* times or are already *aggregated*. For example, this is the case for daily, monthly or yearly stock market data.

A useful assumption for time series models is the concept of *stationarity*[1]: A time series is stationary if the autocorrelation function only depends on the lag, i.e. the time difference $h := t - s$ between two data points x_s and x_t, where $s < t$. Stationarity allows a form of dependence between data points that still ensures consistency and asymptotic results for parameter estimates of time series models such as ARMA (**a**uto**r**egressive **m**oving **a**verage) and GARCH (**g**eneralized **a**uto**r**egressive **c**onditional **h**eteroskedasticity).

An initially non-stationary time series can sometimes be transformed into a stationary one. This is usually done by detecting and removing deterministic trends and seasonality as well as differencing (see Section 1.4 in Brockwell and Davis 1991).

These two assumptions of regularly spaced and stationary data are called into question when moving to high-frequency level of financial data. As a consequence of technological advancement, it is possible to record all transaction of a trading day or as Engle 2000 termed it: financial data are increasingly available at "ultra-high-frequency". This kind of intra-day or tick-by-tick data are inherently *irregularly spaced*. One could aggregate the data to fit into the framework of integer-indexed time series, but this can be problematic as pointed out in Engle and Russell 1998: The choice of the time grid for aggregation is somewhat arbitrary and distorts the results of a subsequent statistical analysis. If time intervals are too small, some intervals are empty or just contain a single observation. If the intervals are too large, information on the time structure might get lost. A way to accommodate irregularly

[1] At this point, we refer to stationarity as second-order or weak stationarity as opposed to strict stationarity, where the finite dimensional marginals of the process do not depend on the lag. For an exact definition see Definition 1.3.2 and Definition 1.3.3 in Brockwell and Davis 1991.

spaced data is continuous-time point process models. Engle and Russell 1998 have proposed the ACD (**a**utoregressive **c**onditional **d**uration) model which will be discussed further in Section 5.3 of this thesis. As direct generalizations of the standard time series models, there are approaches in constructing continuous time analogues such as the CARMA (Brockwell 2001, 2004, 2014 and Section 11.5 in Brockwell and Davis 2016) and COGARCH (Klüppelberg, Lindner and Maller 2004, Brockwell, Chadraa and Lindner 2006) process. Slightly separate from the theory around time series models, doubly stochastic point processes are already established in actuarial risk theory, but their subclass self-exciting point processes has received attention in recent publications (see Section 5.4 in the thesis) and are viable alternatives to the ACD model.

Concerning the stationarity property, it is debatable whether this theoretically convenient property can be reconciled with stylized facts of empirical data. One of these stylized facts is long-range dependence which is closely related to non-stationarity. Long-range dependence usually describes the slow decay (slower than exponential) of the autocorrelation function for absolute or squared returns. When working with stationary processes, long-memory is for example achieved by fractional integration, which is a generalized differencing method (Hosking 1981). However, it is not always easy to distinguish whether data is stationary with long-memory or simply non-stationary and thus it is not clear which model approach to follow. Numerous tests for detection of non-stationarity have been developed and proposed (see Dette, Preuss and Sen 2017 and references therein). Empirical studies suggest the existence of structural changes or structural breaks (Rapach and Strauss 2008, Mikosch and Stărică 2004, Stărică and Granger 2005), i.e. market shocks after which estimated parameters of time series models need to be adjusted. A possible modeling choice in order to accommodate such changes are locally stationary models, i.e. in between structural changes we still assume a homogeneous process. The compound Poisson type model discussed in Section 5.2 can be categorized as a locally stationary point process. Moreover the fractional Poisson process is a non-stationary point process and Leonenko, Meerschaert, Schilling et al. 2014 has proved some long-memory properties. Furthermore, Biard and Saussereau 2014, 2016 also discuss long-range dependence and propose the fractional Poisson process for application in an actuarial modeling framework.

Structure of the thesis

The thesis is a collection of work during my PhD studies at the Department of Mathematics of the University of Sussex. The content of the thesis has appeared in the following publications and preprints:

J. Chen, A. G. Hawkes, E. Scalas and M. Trinh (2018). "Performance of information criteria for selection of Hawkes process models of financial data". In: *Quant. Finance* 18.2, pp. 225–235

N. Leonenko, E. Scalas and M. Trinh (2017b). "The fractional non-homogeneous Poisson process". In: *Statist. Probab. Lett.* 120, pp. 147–156

N. Leonenko, E. Scalas and M. Trinh (2017a). "Limit Theorems for the Fractional Non-homogeneous Poisson Process". In: *ArXiv e-prints*. arXiv: 1711.08768 [math.PR]

L. Ponta, M. Trinh, M. Raberto, E. Scalas and S. Cincotti (2012). "Modeling non-stationarities in high-frequency financial time series". In: *ArXiv e-prints*. arXiv: 1212.0479 [q-fin.ST]

Wherever it seemed necessary and appropriate, preliminaries are given for understanding the content of each chapter. We require the reader to have some basic knowledge of graduate level mathematics, especially in the area of probability and statistics. If in doubt, we refer the reader to standard textbooks on probability (including martingale theory and Lévy processes) and statistics, for example Durrett 2010, Applebaum 2009, Georgii 2007, Czado and Schmidt 2011.

Chapter 1 sets up the mathematical framework of one-dimensional point processes and their simulation methods. As a typical example, we present the Poisson process together with properties relevant for the following chapters. Both the fractional Poisson process and the financial models for durations between trades can be viewed as generalizations of certain aspects of the usual Poisson process.

The rest of the thesis is divided in two parts: The first part consists of Chapter 2 and 3 which discuss the fractional Poisson process and in particular the fractional non-homogeneous Poisson process (FNPP).
In Chapter 2, building on the standard Poisson process, we propose a construction of a fractional non-homogeneous Poisson process via a time-change of a non-homogeneous Poisson process using the inverse of an α-stable subordinator. Preliminaries for understanding the essential elements of this construction, like α-stable distributions and their associated Lévy processes and the Mittag-Leffler function, are provided in the first part of the chapter. In direct comparison with previous results on the fractional homogeneous Poisson process (FHPP), we derive the

one-dimensional marginals, moments and covariance and prove governing equations. Links to the FHPP and the standard Poisson process are discussed as special cases of the FNPP. The results on the FNPP in Chapter 2 are mainly based on work in Leonenko, Scalas and Trinh 2017b.
After deriving governing equations for the FNPP, we move on to the derivation of finite-dimensional and functional limit theorems for the fractional Poisson process in Chapter 3. First, we provide an introduction to the basic framework of weak convergence of probability measures and more specific convergence notions for stochastic processes with càdlàg paths (right-continuous with left limits). For the space of càdlàg functions we focus on the J_1 and M_1 topology and general techniques for proving functional convergence with respect to those topologies, in particular with applications to the FNPP. The results on limit theorems for the FNPP follows up on the work in Leonenko, Scalas and Trinh 2017b and can also be found in Leonenko, Scalas and Trinh 2017a

The second part of the thesis is devoted to the statistical analysis of performance of information criteria in selecting model orders of financial models for durations between trades. Recall that such models are typically applied for high-frequency intra-day trading data in order to directly model trade durations instead of losing this information by aggregation to equidistant time grids.
Chapter 4 serves as an introduction to information criteria for model selection, especially comparing the different motivations for the two most popular ones, Akaike's information criterion (AIC) and and the Bayesian information criterion BIC. The chapter is meant to be read together with Chapter 5, which contains numerical results of a Monte-Carlo experiment to test the performance of information criteria for model selection within three different model classes: a type of compound Poisson model, the ACD model and a Hawkes process model. Especially Hawkes processes have recently gained popularity as a model for financial data. For each model class, we first give their definition and stationarity conditions if necessary. Moreover, we discuss simulation and estimation methods as well as goodness-of-fit measures needed for the setup of the Monte-Carlo simulation. This work on model selection draws from work found in Chen et al. 2018 and Section 4 in Ponta et al. 2012. Preceding Section 4 in Ponta et al. 2012 is an empirical study of high-frequency trading data, taken from the Italian stock exchange Borsa Italiana on the FTSE MIB index, that led to the proposed compound Poisson type model. This material is not covered in this thesis.

Following the main text is an appendix containing some useful results from regular variation theory which are applied at several points in the thesis. Moreover, a

manual and the source code for key functions of the implementations done for the Monte-Carlo simulations are included in the appendix.

Acknowledgements

First and foremost, I would like to express my sincere gratitude towards my PhD supervisor Enrico Scalas who has supported me with his patience and guidance throughout the years leading up to the thesis and who gave me the freedom to work in my own way. Besides my supervisor, I would like to thank my examiners József Lörinczi and Nicholas Simm for their insightful comments, but also for the hard questions that motivated me to widen my research in various directions. Special thanks go to my second supervisor Nicos Georgiou for his opinion and advice. I also thank Maggie Chen, Nikolai Leonenko, Alan Hawkes and Linda Ponta who I had the privilege to work with.
I thank the members of the Department of Mathematics of the University of Sussex for the welcoming atmosphere and the enjoyable, encouraging and inspiring conversations. I am indebted to the Strategic Development Fund of the University of Sussex without their financial support this past work would not be possible.

I would like to thank everyone who has contributed to the completion of this thesis. Last, but not least, I thank my family and friends for the emotional support and for believing in me.

Guide to notation

s.t.	such that
w.r.t.	with respect to
w.l.o.g.	without loss of generality
i.i.d	independently and identically distributed
iff	if and only if
a.s.	almost surely
$\mathrm{Exp}(\lambda)$	exponential distribution with mean $1/\lambda$
$N(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$X, (X(t))_{t\geq 0}$	random variable/stochastic process with time index t, the index is omitted whenever the range of the index is clear.
$\mathbb{1}_A(x)$	indicator function: $\mathbb{1}_A(x) = 1$ if $x \in A$ and $\mathbb{1}_A(x) = 0$ otherwise
$\delta_x(y)$	delta distribution: $\delta_x(y) = 1$ if $x = y$, $\delta_x(y) = 0$ otherwise
FHPP	fractional homogeneous Poisson process
FNPP	fractional non-homogeneous Poisson process
càdlàg	right-continuous with left limits (continue à droite, limite à gauche)
$\mathcal{C}(X, Y)$	space of continuous functions defined on the space X, mapping to the space Y
$\mathcal{C}_c(X, Y)$	space of continuous functions with compact support
$\mathcal{C}_b(X, Y)$	space of continuous and bounded functions
$\mathcal{C}_0(X, Y)$	space of continuous functions vanishing at infinity
$\mathcal{D}(X, Y)$	space of càdlàg functions, Skorokhod space
$\overset{d}{=}$	equality in distribution
$\xrightarrow{\text{a.s.}}$	almost sure convergence
$\xrightarrow{\text{f.d.}}$, $\xrightarrow{d}$	convergence in finite-dimensional distributions, convergence in distribution
$\xrightarrow{P}$	convergence in probability
$\xrightarrow{w}$	weak convergence (of probability measures)
$\xrightarrow{J_1}$, $\xrightarrow{M_1}$	convergence w.r.t. the J_1 or M_1 topology

ACD	**a**utoregressive **c**onditional **d**uration model
GARCH	**g**eneralized **a**uto**r**egressive **c**onditional **h**eteroskedasticity
CLT	central limit theorem
DOA	domain of attraction (of a stable law)
MLE	maximum likelihood estimation/estimator
MSE	mean squared error
RMSE	root mean squared error
IC	information criterion/criteria
AIC	Akaike's information criterion
BIC	Bayesian information criterion
HQ	Hannan-Quinn information criterion

Chapter 1

Point processes

In this chapter, we provide some general results of point process theory and specify them for the Poisson process as a typical example. These concepts will be used in the subsequent chapters and are essential for the definition and discussion of point process models.

1.1 Point process theory and martingales

Definition 1. Let E be a complete separable metric space and $\mathcal{B}(E)$ be the σ-field of its Borel sets.

(i) A locally finite measure on $\mathcal{B}(E)$ is called a *Borel measure*.

(ii) A Borel measure μ on E is *boundedly finite* if $\mu(A) < \infty \ \forall A \in \mathcal{B}(E)$.

(iii) A boundedly finite, integer-valued measure is a *counting measure*.

(iv) A counting measure is *simple* if $N(\{x\}) \in \{0, 1\} \ \forall x \in E$.

We consider simple, boundedly finite, integer-valued measures and refer to them as *point processes*. Usually, we will set $E = \mathbb{R}_+$ which will have the interpretation of time in later applications. In this case, we do not need to primarily think of a point process as a random measure. Instead, the point process is literally random points, say $t_1, t_2, t_3, \ldots$ on the positive half line[1]. The associated counting measure is also a counting process N which has increasing càdlàg paths (see Figure 1.1).
Similar to Lévy's characterization of Brownian motion using its quadratic variation (see Theorem II.4.4 in Jacod and Shiryaev 2003), point processes are characterized by their compensators.

[1]Capital letters will usually refer to random variables, e.g. $T_1, T_2, \ldots$ for arrival times and lower case letters denote their realizations $t_1, t_2, \ldots$.

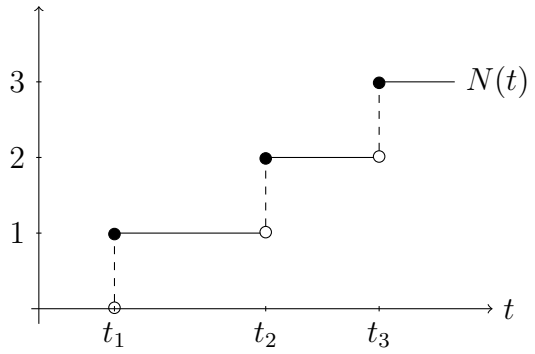

Figure 1.1: Typical càdlàg path of a counting process N associated with a (Poisson) point process with arrival times $t_1, t_2, t_3, \ldots$.

Definition 2. Let $(N(t))_{t\geq 0}$ be a counting process on $\mathbb{R}_+$ adapted to a filtration $(\mathcal{F}_t)_{t\geq 0}$. A *compensator* A of ξ w.r.t. $(\mathcal{F}_t)_{t\geq 0}$ is a monotonic non-decreasing, right-continuous predictable process such that $(M(t))_{t\geq 0}$, where $M(t) := N(t) - A(t)$, is a local martingale.

The Doob-Meyer decomposition (see Theorem 4.10 in Karatzas and Shreve 1988) of submartingales ensures the existence and uniqueness of the compensator of a point process. The following theorem will be useful in later chapters:

Theorem 1. Let $\mathbb{P}$ be a probability measure on a probability space $(\Omega, \mathcal{F})$. Then, there exists for a point process $(N(t))_{t\geq 0}$ a $(\mathcal{F}_t)$-predictable random measure $\mathrm{d}A(t)$ (or equivalently an increasing $(\mathcal{F}_t)$-predictable process $(A(t))_{t\geq 0}$) such that for all positive $(\mathcal{F}_t)$-predictable processes $(X(t))_{t\geq 0}$, it holds that

$$\mathbb{E}\left[\int_0^\infty X(t)\,\mathrm{d}N(t)\right] = \mathbb{E}\left[\int_0^\infty X(t)\,\mathrm{d}A(t)\right]$$

(see Theorem I.3.18 Jacod and Shiryaev 2003).

In the case that $t \mapsto A(t,\omega)$ is an absolutely continuous function which admits a density λ s.t.

$$A(t,\omega) = \int_0^t \lambda(\tau,\omega)\,\mathrm{d}\tau, \tag{1.1}$$

we call λ the $(\mathcal{F}_t)$-conditional *intensity* of $(N(t))_{t\geq 0}$. Intuitively, the conditional intensity can be interpreted as the instantaneous probability of an event at time t given the history $\mathcal{F}_t$. Formally, we can write for the intensity

$$\lambda(t) \approx \lim_{\Delta\to 0} \frac{1}{\Delta}\mathbb{E}[N(t+\Delta) - N(t)|\mathcal{F}_{t-}] = \lim_{\Delta\to 0} \frac{1}{\Delta}\mathbb{P}[N(t+\Delta) - N(t) = 1|\mathcal{F}_{t-}].$$

Although the existence of the compensator is guaranteed, this is not always the case for the intensity. Nevertheless, point process models can be defined using an intensity function as we will see in later chapters.

Let $T_1, T_2, \ldots$ be the events of the point process $(N(t))_{t\geq 0}$. The conditional probability density function (if it exists) of the nth event is denoted by $p_n(t|T_1, T_2, \ldots, T_{n-1})$ and the *survivor function* or *survival function* is given by

$$S_n(t|T_1, T_2 \ldots, T_{n-1}) = 1 - \int_{T_{n-1}}^{t} p_n(u|T_1, T_2, \ldots, T_{n-1})\,\mathrm{d}u, \qquad (t > T_{n-1})$$

and the *hazard function* is defined as

$$h_n(t|T_1, T_2, \ldots, T_{n-1}) = \frac{p_n(t|T_1, T_2, \ldots, T_{n-1})}{S_n(t|T_1, T_2 \ldots, T_{n-1})}. \tag{1.2}$$

Using the above terms we can express the conditional intensity via

$$\lambda(t) = \begin{cases} h_1(t) & \text{if } 0 < t \leq T_1, \\ h_n(t|T_1, T_2, \ldots, T_{n-1}) & \text{if } T_{n-1} < t \leq T_n, n \geq 2 \end{cases} \tag{1.3}$$

(for details see Section 7.2 in Daley and Vere-Jones 2003). Indeed, it can be shown that the conditional intensity functions in Equation (1.1) and the representation in (1.3) coincide a.s. (see Corollary 14.1.V. in Daley and Vere-Jones 2008).

1.2 The homogeneous Poisson process

The Poisson process is the archetypical stochastic process for random occurrences. There are various possible characterizations for the homogeneous Poisson process (see Vidmar 2016 and references therein). Nevertheless, we will restrict the presentation to the most commonly encountered definitions as seen especially in the actuarial risk theory context: Most of the results in this section can be found in Mikosch 2009. First, we choose the characterization of the Poisson process as a Lévy process as a definition.

Definition 3. Let $\lambda > 0$ and $(N(t))_{t\geq 0}$ be a stochastic process such that

(i) $N(0) = 0$ a.s.

(ii) N has independent increments

(iii) N has stationary increments with

$$N(t) - N(s) \sim \mathrm{Poi}(\lambda(t-s)), \quad 0 \leq s < t. \tag{1.4}$$

(iv) N has càdlàg paths.

Then N is a homogeneous Poisson process and λ is called the intensity parameter.

It follows from (i) and (iii) that the one-dimensional distributions of N are given by

$$p_{N(t)}(k) := \mathbb{P}(N(t) = k) = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad k \in \mathbb{N}.$$

Since the Poisson process is a Lévy process, there is a closed form expression for its characteristic function:

$$\varphi_{N(t)}(u) := \mathbb{E}[\mathrm{e}^{\mathrm{i}uN(t)}] = \exp(\lambda t(\mathrm{e}^{\mathrm{i}u} - 1)).$$

The homogeneous Poisson process can also be characterized by its compensator $A(t) = \lambda t$, which implies a constant intensity function $a(t) = \lambda$. This result is known as the *Watanabe characterisation* due to Watanabe 1964. Related to this is another characterization via an infinitesimal description:

Theorem 2. Let $(N(t))_{t\geq 0}$ be a counting process with stationary and independent increments, $N(0) = 0$ and increasing càdlàg paths. Then, $(N(t))_{t\geq 0}$ is a homogeneous Poisson process with parameter λ if and only if

$$\begin{aligned}\mathbb{P}(N(t+h) - N(t) = 1) &= \lambda h + o(h) \quad (h \to 0+)\\ \mathbb{P}(N(t+h) - N(t) > 1) &= o(h) \quad (h \to 0+),\end{aligned}$$

(see Theorem 2.2.III in Daley and Vere-Jones 2003).

The Poisson process is a Markov process and its semigroup is generated by the shift operator (see Example 3.3.6 in Applebaum 2009).

$$(Lf)(x) = \lambda(f(x+1) - f(x))$$

and the associated governing equation is

$$\begin{cases} f'(t) &= \lambda(f(t+1) - f(t)) \\ f(0) &= \delta_0. \end{cases} \tag{1.5}$$

The Poisson process can equivalently be defined as a renewal process with i.i.d. $\mathrm{Exp}(\lambda)$ distributed waiting times $(J_i)_{i\in\mathbb{N}}$, i.e.

$$1 - F_{J_i}(t) := \mathbb{P}(J_i > t) = \exp(-\lambda t).$$

The corresponding arrival times are given by $T_n = J_1 + J_2 + \ldots + J_n$ and their distribution is sometimes referred to as *Erlang distribution*. Let F_{T_n} and f_{T_n} be the

distribution function and probability density function of T_n respectively. Then

$$F_{T_n}(t) = 1 - \mathrm{e}^{\lambda t} \sum_{x=1}^{n-1} \frac{(\lambda t)^x}{x!}, \qquad f_{T_n}(t) = \mathrm{e}^{-\lambda t} \frac{\lambda^n t^{n-1}}{(n-1)!}.$$

Given the arrival times $T_1, T_2, \ldots$, the corresponding counting process $(N(t))_{t\geq 0}$ can be written as

$$N(t) := \sup\{n \in \mathbb{N} : T_n \leq t\}.$$

Instead of specifying the waiting time distribution, another possible characterization of the homogeneous Poisson process is the *order statistics property* (see Liberman 1985, Gan and Yang 1989, Gan and Yang 1990, Section 2.1.6 in Mikosch 2009):

Theorem 3 (Order statistics property)**.** A renewal process $(N(t))_{t\geq 0}$ is a homogeneous Poisson process if and only if the distribution of the interarrival times $T_1, T_2, \ldots, T_n$ given $N(t) = n$ is equal to the distribution of the (minimum) order statistics of i.i.d. samples $U_1, U_2, \ldots, U_n$ of a Uniform$(0, t)$ distribution, i.e. for $n \in \mathbb{N}$

$$(T_1, T_2, \ldots, T_n | N(t) = n) \stackrel{d}{=} (U_{(1)}, U_{(2)}, \ldots, U_{(n)}) \tag{1.6}$$

This turns out to be very useful for simulation purposes.

1.3 The inhomogeneous Poisson process

The homogeneous Poisson process is one of the simplest examples of a Lévy process. Similarly, one can consider the *inhomogeneous Poisson process* [2] as a simple example of an additive process. We require the process $(N(t))_{t\geq 0}$ to fulfill properties (i), (ii) and (iv) in Definition 3 and replace the stationary increment property (iii) by a more general one

(iii') $\quad N(t) - N(s) \sim \mathrm{Poi}(\Lambda(s,t)), \quad 0 \leq s < t,$

where the function Λ has to satisfy $\Lambda(r,s) + \Lambda(s,t) = \Lambda(r,t)$ for all $s, r, t \in \mathbb{R}_+$ with $r \leq s \leq t$. In the following, we assume Λ to be absolutely continuous, i.e. there is a positive function $\lambda : \mathbb{R}_+ \to (0, \infty)$ such that

$$\begin{aligned} \Lambda : \mathbb{R}_+^2 &\to [0, \infty) \\ (s,t) &\mapsto \int_s^t \lambda(\tau) \mathrm{d}\tau. \end{aligned} \tag{1.7}$$

The function Λ is sometimes called *rate function*. The idea of this generalization is to allow the intensity λ of the process to depend on time. Indeed, if we set $\lambda(\tau) \equiv \lambda$

[2] synonymously *non-homogeneous* Poisson process

constant we get

$$\Lambda(s,t) = \lambda(t-s) \text{ and } N(t) - N(s) \sim \text{Poi}(\lambda(t-s)), \quad 0 \leq s < t$$

and we obtain property (iii), i.e. N is a homogeneous Poisson process iff λ is constant.

Example 1. Rate functions can be constructed using the hazard function as described in Section 1.1.

(i) A popular extreme value distribution is the *Weibull distribution* with probability distribution function

$$F(t) = 1 - \mathrm{e}^{-(t/b)^c}, \quad c > 0, b > 0,$$

(see Embrechts, Klüppelberg and Mikosch 1997). The corresponding intensity function is given by

$$\lambda(t) = \frac{F'(t)}{1 - F(t)} = \frac{c}{b}\left(\frac{t}{b}\right)^{c-1}.$$

Integration yields Weibull's rate function:

$$\Lambda(t) := \Lambda(0,t) = \left(\frac{t}{b}\right)^c, \quad c > 0, b > 0$$

(ii) A commonly used distribution for mortality is the *Gompertz–Makeham law* with distribution function

$$F(t) = 1 - \exp\left(-\frac{c}{b}(\mathrm{e}^{bt} - 1) - \mu t\right), \quad c > 0, b > 0, \mu \geq 0,$$

(see Marshall and Olkin 2007). The corresponding rate function and intensity are given by

$$\Lambda(t) = \frac{c}{b}\mathrm{e}^{bt} - \frac{c}{b} + \mu t, \ \lambda(t) = c\mathrm{e}^{bt} + \mu, \quad c > 0, b > 0, \mu \geq 0.$$

In Brémaud 1975 it is shown that the compensator of the inhomogeneous Poisson process is given by $A(t) = \Lambda(0,t)$, i.e. $M(t) := N(t) - \Lambda(0,t)$ is a martingale.
It is possible to transform a non-homogeneous Poisson process into a homogeneous one via a time-change which is exactly given by the compensator:

Theorem 4 (Time-change theorem)**.** Let $(N(t))_{t\geq 0}$ be a simple point process adapted to a history $(\mathcal{F}_t)_{t\geq 0}$ with bounded, strictly positive conditional $\mathcal{F}_t$-intensity

$\lambda^*(t)$ and $\mathcal{F}_t$-compensator

$$\Lambda^*(t) = \int_0^t \lambda^*(u)\,\mathrm{d}u$$

that is not a.s. bounded. Under the random time-change $t \mapsto \Lambda^*(t)$, the transformed process $\tilde{N}(t) = N(\Lambda^*(t))$ is a Poisson process with unit rate.
Conversely (see Theorem 7.4.I. in Daley and Vere-Jones 2003).

The result goes back to Meyer 1971 and Papangelou 1972. In particular, the time change theorem implies that

$$N(t) = N_1(\Lambda(0,t)),$$

where N_1 is the homogeneous Poisson process with intensity parameter 1.
The characteristic function of the increments can be written as

$$\psi_{(N(t)-N(s))}(u) = \exp(\Lambda(s,t)(\mathrm{e}^{\mathrm{i}u} - 1)).$$

The inhomogeneous Poisson process is a time-inhomogeneous Markov process and the dynamics can be described by the Kolmogorov equations (see Section 3.5.3 in Applebaum 2009). Let g_k denote the one-dimensional marginals of N and w.l.o.g. $0 < s < t$:

$$g_k(s,t) := \mathbb{P}(N_1(\Lambda(s,t)) = k) = \mathrm{e}^{-\Lambda(s,t)}\frac{\Lambda(s,t)^k}{k!}.$$

and for the increment we write

$$\begin{aligned} p_x(t,v) &:= \mathbb{P}(N(t+v) - N(v) = x) \\ &= \frac{\mathrm{e}^{-\Lambda(v,t+v)}\Lambda(v,t+v)^x}{x!}, \quad x = 0,1,2,\ldots. \end{aligned} \tag{1.8}$$

For notational convenience, we write $p_k(t) = p_k(t,0)$. We can derive the Kolmogorov backward equation (sometimes also called master equation) as follows:

$$\begin{aligned} \frac{\partial}{\partial s}g_k(s,t) &= \lambda(s)\mathrm{e}^{-\Lambda(s,t)}\frac{\Lambda(s,t)^k}{k!} - \lambda(s)\mathrm{e}^{-\Lambda(s,t)}\frac{k\Lambda(s,t)^{k-1}}{k!} \\ &= \lambda(s)\mathrm{e}^{-\Lambda(s,t)}\left[\frac{\Lambda(s,t)^k}{k!} - \frac{\Lambda(s,t)^{k-1}}{(k-1)!}\right] \\ &= \lambda(s)[g_k(s,t) - g_{k-1}(s,t)]. \end{aligned}$$

In a similar way, the forward equation yields an equation of Fokker-Planck type

$$\begin{aligned}\frac{\partial}{\partial t}g_k(s,t) &= -\lambda(t)\mathrm{e}^{-\Lambda(s,t)}\frac{\Lambda(s,t)^k}{k!} + \lambda(t)\mathrm{e}^{-\Lambda(s,t)}\frac{k\Lambda(s,t)^{k-1}}{k!} \\ &= \lambda(t)\mathrm{e}^{-\Lambda(s,t)}\left[-\frac{\Lambda(s,t)^k}{k!} + \frac{\Lambda(s,t)^{k-1}}{(k-1)!}\right] \\ &= -\lambda(t)[g_k(s,t) - g_{k-1}(s,t)] \end{aligned} \tag{1.9}$$

or in terms of the increment

$$\frac{\mathrm{d}}{\mathrm{d}t}p_x(t,v) = -\lambda(t+v)p_x(t,v) + \lambda(t+v)p_{x-1}(t,v), \quad x = 0,1,2,\ldots, \tag{1.10}$$

with initial conditions

$$p_x(0,v) = \begin{cases} 1, & x = 0 \\ 0, & x \geq 1 \end{cases}$$

and $p_{-1}(t,v) \equiv 0$.
We see that for constant λ Equation (1.10) simplifies to Equation (1.5), the homogeneous case.

1.4 Simulation

1.4.1 Simulation methods for the homogeneous Poisson process

Since more general simulation algorithms for point processes rely on an effective simulation method for the homogeneous Poisson process, it is useful to discuss common approaches for that first. Due to the constant intensity parameter, several of the previously mentioned characterizations of the homogeneous Poisson process lend themselves to vectorized sampling algorithms.
The common approach to simulate Lévy processes is to sample the increments, due to their independence. If this is done on an equidistant grid, the increments are i.i.d. and the sampling can be vectorized in this step. Finally, one needs to calculate the cumulative sum to get the process values. However, this would give us merely an approximation of $(N(t))_{t\geq 0}$ because there would be no information on the exact event times $T_1, T_2, \ldots$.
The renewal representation of the Poisson process allows us to simulate the arrival times directly by drawing from $\mathrm{Exp}(\lambda)$. However, the procedure is not easily vectorized as it is not clear a priori how many samples are actually needed (assuming we are to simulate the process on a given time interval).
At this point, the order statistics characterization comes quite handy as we can

see from Formula (1.6) that we can draw the number of events in a given interval separately from the actual event times. This leaves us with Algorithm 1.1

Algorithm 1.1: Sampling of a homogeneous Poisson process

```
1   input:
2         T − time horizon, sampling interval [0,T)
3         λ − intensity parameter
4   output:
5         t = (t_1, t_2, ...) − event times of the simulated process
6   begin
7      1) Generate N ∼ Poi(λT)
8      2) Generate u_1, u_2, ..., u_N i.i.d. samples drawn from Uniform(0,1)
9      3) Apply the minimum order statistics ⇝ t ← (u_(1), ..., u_(N))
10  end
```

1.4.2 Thinning algorithms

Thinning algorithms are adaptions of the sampling approach via acceptance-rejection schemes. In their basic form of drawing i.i.d. samples from a given probability density function f the idea is to find an upper bound on f using a probability density g which is easy to sample from:

$$f \leq cg,$$

where c is a constant. The pseudocode in Algorithm 1.2 displays how such a method would be implemented.

Algorithm 1.2: Acceptance-rejection method

```
1   input:
2         f − probability density
3         g − probability density for which a simulation method is
4               already implemented
5         c − constant for upper bound on f
6   output:
7         X sample from distribution with density f
8   begin
9      do
10        1) draw X ∼ g
11        2) draw U ∼ Uniform(0,1)
12     while U > f(X)/cg(X)
13  end
```

In a similar way, the same technique works for point processes, where the acceptance-rejection rule is applied to a ratio of intensities instead of probability densities. In the next paragraphs we will discuss the two cases in which intensity functions can be globally or locally bounded.

The Lewis-Shedler algorithm If a time-varying and possibly random intensity function λ can be a.s. bounded from above by a constant M, we are able to obtain samples from the point process with underlying intensity λ: First, a sample from a homogeneous Poisson process with intensity parameter M can be drawn using for example Algorithm 1.1. This would give us a sample with too many events. We therefore need to "thin" out the sample proportional to $\lambda(t)/M$. This leads to the thinning algorithm by Lewis and Shedler 1979 (see Algorithm 1.3).

Algorithm 1.3: Lewis-Shedler algorithm

```
input:
      M – bound on the conditional intensity
      A – limit of the simulation interval [0,A)
      λ* – conditional intensity function
output:
      {t_1, t_2, ...} – event times of the simulated process
begin
   1) simulate x_1, x_2, ... as realizations of a Poisson process
      with intensity parameter M.
   2) simulate independent samples y_1, y_2, ... of Uniform(0,1)
   3) set  k ← 1, j ← 1, H = ∅
   4) if x_k > A ⇝ terminate
      else evaluate λ* = λ(x_k|H)
      end if
   5) if y_k ≤ λ*(x_k)/M ⇝ set t_j ← x_k, H = H ∪ {t_j}, j ← j + 1
      end if
   6) k ← k + 1, goto (4)
end
```

Ogata's modified algorithm The existence of a global upper bound on the intensity function is not always given. This is for example the case for Hawkes processes which we will discuss in more detail in Chapter 5. Hawkes processes belong to the class of self-exciting point processes, which essentially means that the occurrence of an event caused a positive jump in the intensity. Although additional requirements like stationarity might prevent a blow up in intensity, this is generally not enough to guarantee a deterministic global upper bound on the intensity.
Fortunately, it is possible to relax the requirement to a *local* upper bound on the intensity between events. For the algorithm to work, one needs to make sure that the current upper bound is updated whenever an event occurs. We first present Ogata's modified algorithm Ogata 1981 as described in Section 7.5 in Daley and Vere-Jones 2003 and then give a version specifically for the Hawkes process in Chapter 5.
For Algorithm 1.4 we assume the following: Let $(\mathcal{H}_t)_{t\geq 0}$ be a filtration. It typically contains the events of the point process for it to be adapted, but could also include additional information. Suppose that there exist functions $M(t|\mathcal{H}_t)$ and $L(t|\mathcal{H}_t)$

such that for all initial histories $\mathcal{H}_0$ and all $t \in [0, \infty)$, $n = 1, 2, \ldots$ and sequences $0 < t_1 < t_2 < \ldots < t_{n-1} < t$ the hazard functions satisfy

$$h_n(t + u|\mathcal{H}_0, t_1, \ldots, t_{n-1}) \leq M(t|\mathcal{H}_t), \quad 0 \leq u < L(t|\mathcal{H}_t).$$

Algorithm 1.4: Ogata's modified algorithm

```
1   input:
2         H_0 - initial history
3         M(t|H), L(t|H) - functions for local spatial and temporal
4                          bounds
5         λ* - conditional intensity function
6   output:
7         {t_1, t_2, ...} - event times of the simulated process
8   begin
9      1) set t ← 0, i ← 0, H ← H_0
10     2) while termination condition not met
11           evaluate M ← M(t|H) and L ← L(t|H)
12           generate T ~ Exp(M(t)) and U ~ Uniform(0,1)
13           if T > L ⇝ t ← t + L
14           else if λ*(t + T)/M > U ⇝ t ← t + T
15                else i ← i + 1, t_i ← t + T, t ← t_i  H ∪ {t_i}
16                end if
17           end if
18        end while
19  end
```

1.5 Summary

We have revised the general definition of point processes and in particular their relation to martingale theory. As a typical example, we presented the homogeneous and inhomogeneous Poisson process. This revision should prepare for the treatment of the fractional version of the Poisson process in the following chapters. We have concluded the chapter with a simulation algorithm for the Poisson process and the standard thinning algorithms for point processes with (locally) bounded intensity. These are basis for the implementation of simulation algorithms needed in Monte-Carlo experiments in Chapter 5.

Chapter 2

The fractional Poisson process

In this chapter, we introduce the fractional Poisson process and its relation to α-stable subordinators and the Caputo derivative. We propose a definition for the fractional non-homogeneous Poisson process and derive some first properties. First, we give some preliminaries on the Mittag-Leffler function and α-stable distributions.

2.1 Preliminaries

2.1.1 The Mittag-Leffler function

The one-parameter Mittag-Leffler function was proposed by Mittag-Leffler 1903a,b, 1905 for the summation of divergent series and is given by

$$E_\alpha(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(1+\alpha k)}, \quad \alpha \in \mathbb{C}, \mathrm{Re}(\alpha) > 0, z \in \mathbb{C}. \tag{2.1}$$

The Mittag-Leffler function can be viewed as a generalization of the exponential function. If we set $\alpha = 1$, the right hand side of Equation (2.1) reduces to the power series representation of the exponential function. In other words, $E_1(z) = \exp(z)$. The two parameter Mittag-Leffler function was first used by Wiman 1905a,b and is defined by

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\beta+\alpha k)}, \quad \alpha, \beta \in \mathbb{C}, \mathrm{Re}(\alpha), \mathrm{Re}(\beta) > 0, z \in \mathbb{C}.$$

In the most general form, we will consider the three parameter Mittag-Leffler function, which was introduced by Prabhakar 1971. To write it in compact form the notation for the rising and falling factorial is useful. Let $\gamma \in \mathbb{C}$ and $k \in \mathbb{N}$, then

define the falling and the rising factorial as

$$(\gamma)^{\underline{k}} = \gamma(\gamma-1)\dots(\gamma-k+1) = \frac{\Gamma(\gamma+1)}{\Gamma(\gamma-k+1)} \qquad \text{and}$$
$$(\gamma)^{\overline{k}} = \gamma(\gamma+1)\dots(\gamma+k-1) = \frac{\Gamma(\gamma+k)}{\Gamma(\gamma)}$$

respectively. The three parameter Mittag-Leffler function can now be defined as follows:

$$E^{\gamma}_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{(\gamma)^{\overline{k}} z^k}{k!\Gamma(\beta+\alpha k)} \quad \alpha,\beta,\gamma \in \mathbb{C}, \mathrm{Re}(\alpha), \mathrm{Re}(\beta) > 0, z \in \mathbb{C}.^{1}$$

Note that $E^{1}_{\alpha,\beta} = E_{\alpha,\beta}$ and $E_{\alpha,1} = E_{\alpha}$.
For an extensive review on Mittag-Leffler functions and their applications, the reader is referred to Erdélyi et al. 1981, the Appendix in Mainardi and Gorenflo 2000 and the review paper Haubold, Mathai and Saxena 2011. For the sake of completeness we will give useful relations of the Mittag-Leffler function that will be used later in the following proposition.

Proposition 5. Let $\alpha > 0$ and $k = 0, 1, 2, \dots$. Then

a) $\dfrac{\mathrm{d}^k}{\mathrm{d}z^k} E_{\alpha}(z) = k! E^{k+1}_{\alpha,\alpha k+1}(z)$

b) $\mathcal{L}\{t^{\alpha k} E^{(k)}_{\alpha}(-t^{\alpha}); s\} = \dfrac{k! s^{\beta-1}}{(1+s^{\beta})^{k+1}}.$

Proof. For part (a) we check via direct computation. As the power series of the Mittag-Leffler function converges absolutely, we can interchange differentiation and the limit of the series:

$$\begin{aligned}
\frac{\mathrm{d}^k}{\mathrm{d}z^k} E_{\alpha}(z) &= \frac{\mathrm{d}^k}{\mathrm{d}z^k} \sum_{j=0}^{\infty} \frac{z^j}{\Gamma(1+\alpha j)} = \sum_{j=k}^{\infty} \frac{(j)^{\underline{k}} z^{j-k}}{\Gamma(1+\alpha j)} = \sum_{j=0}^{\infty} \frac{(j+k)^{\underline{k}} z^j}{\Gamma(1+\alpha(j+k))} \\
&= \sum_{j=0}^{\infty} \frac{(j+k)!}{j!} \frac{z^j}{\Gamma(1+\alpha(j+k))} = k! \sum_{j=0}^{\infty} \frac{(j+k)!}{k!} \frac{z^j}{j!\Gamma(1+\alpha(j+k))} \\
&= k! \sum_{j=0}^{\infty} \frac{(j)^{\overline{k}} z^j}{j!\Gamma(1+\alpha(j+k))} = k! E^{k+1}_{\alpha,\alpha k+1}(z),
\end{aligned}$$

which yields the desired result.

[1] In the original notation given in Prabhakar's work $(\gamma)^{\overline{k}}$ was replaced by $(\gamma)_k$ for the rising factorial and is often referred to as Pochhammer symbol. However, we follow the recommendation given in Knuth 1992. Pochhammer himself actually used it as a notation for the binomial coefficient, i.e. $(\gamma)_k = \binom{\gamma}{k}$. Across the literature, $(\gamma)_k$ is used as both the rising and the falling factorial and can therefore lead to confusion.

In order to show (b), we again interchange differentiation and limit:

$$\begin{aligned}
\mathcal{L}\{t^{\alpha k}E_\alpha^{(k)}(-t^\alpha);s\} &= \int_0^\infty e^{-st}t^{\alpha k}\left(\frac{\mathrm{d}^k}{\mathrm{d}z^k}\sum_{j=0}^\infty \frac{z^j}{\Gamma(1+\alpha j)}\bigg|_{z=-t^\alpha}\right)\mathrm{d}t \\
&= \int_0^\infty e^{-st}t^{\alpha k}\left(\sum_{j=k}^\infty \frac{(j)^{\underline{k}}z^{j-k}}{\Gamma(1+\alpha j)}\bigg|_{z=-t^\alpha}\right)\mathrm{d}t = \int_0^\infty e^{-st}t^{\alpha k}\left(\sum_{j=k}^\infty \frac{(j)^{\underline{k}}(-t^\alpha)^{j-k}}{\Gamma(1+\alpha j)}\right)\mathrm{d}t \\
&= \sum_{j=k}^\infty \frac{(j)^{\underline{k}}}{\Gamma(1+\alpha j)}(-1)^{j-k}\underbrace{\int_0^\infty \mathrm{e}^{-st}t^{\alpha j}\mathrm{d}t}_{=\frac{1}{s}\int_0^\infty \mathrm{e}^{-u}\left(\frac{u}{s}\right)^{\alpha j}\mathrm{d}u = s^{-1-\alpha j}\Gamma(1+\alpha j)} = \sum_{j=k}^\infty (j)^{\underline{k}}(-1)^{j-k}s^{-1-\alpha j} \\
&= s^{-1-\alpha k}\sum_{j=k}^\infty (j)^{\underline{k}}\underbrace{(-1)^{j-k}s^{-\alpha(j-k)}}_{(-s^{-\alpha})^{j-k}} = s^{-1-\alpha k}\left(\frac{\mathrm{d}^k}{\mathrm{d}z^k}\sum_{j=0}^\infty z^j\bigg|_{z=-s^{-\alpha}}\right) \\
&= s^{-1-\alpha k}\,\frac{k!}{(1-z)^{k+1}}\bigg|_{z=-s^{-\alpha}} = s^{-1-\alpha k}\frac{k!}{(1+s^{-\alpha})^{k+1}} = \frac{k!s^{-1+\alpha}}{(1+s^\alpha)^{k+1}}
\end{aligned}$$

□

2.1.2 Stable distributions

Let us consider a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with mean $\mu \in \mathbb{R}$ and finite variance $\sigma^2 > 0$. Then the classic central limit theorem (CLT) states that under suitable centering and scaling, the sequence of partial sums $S_n = \sum_{k=1}^n X_k$ converges in distribution to a normal law[2], i.e.

$$\frac{1}{\sqrt{n}}\left(\sum_{k=1}^n X_k - \mu\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma^2).$$

In the case that the variance of X_1 is infinite, it is possible to generalize the classical CLT if we allow not only a normal distribution, but a broader class of distributions in the limit. This class of distributions is referred to as *stable laws* which can be defined by their characteristic function.

Definition 4 (Characteristic function of stable laws $S_\alpha(\sigma,\beta,\mu)$)**.** A random variable X is said to have a *stable distribution* if there are parameters $0 < \alpha \le 2$, $\sigma \ge 0$, $-1 \le \beta \le 1$ and $\mu \in \mathbb{R}$ such that its characteristic function has the following form:

$$\mathbb{E}[\exp(i\theta X)] = \begin{cases} \exp\left(-\sigma^\alpha|\theta|^\alpha\left[1-\mathrm{i}\beta\,\mathrm{sign}(\theta)\tan\left(\frac{\pi\alpha}{2}\right)\right]+\mathrm{i}\mu\theta\right) & \text{if } \alpha \neq 1, \\ \exp\left(-\sigma|\theta|\left[1+\mathrm{i}\beta\frac{2}{\pi}\,\mathrm{sign}(\theta)\ln(|\theta|)\right]+\mathrm{i}\mu\theta\right) & \text{if } \alpha = 1 \end{cases}.$$

The parameter α is called *index of stability* (see Definition 1.1.6 in Samorodnitsky

[2]see Theorem 5.29 in Georgii 2007

and Taqqu 1994).

The index α determines the existence of moments:

Proposition 6. Let $X \sim S_\alpha(\sigma, \beta, \mu)$ with $0 < \alpha < 2$. Then

$$\begin{aligned} \mathbb{E}[|X|^p] &< \infty \quad \text{for any } 0 < p < \alpha \\ \mathbb{E}[|X|^p] &= \infty \quad \text{for any } p \geq \alpha \end{aligned}$$

(see Property 1.2.16 in Samorodnitsky and Taqqu 1994).

Definition 5 (Domain of attraction of a stable law)**.** Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. The law of X_1 is said to be in the *domain of attraction* (DOA) of a stable law $S_\alpha(\sigma, \beta, \mu)$ if there exist sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ and a random variable $S \sim S_\alpha(\sigma, \beta, \mu)$ such that

$$a_n \sum_{k=1}^{n} X_k - b_n \xrightarrow[n\to\infty]{d} S. \tag{2.2}$$

Indeed, one can show that Definition 4 and Definition 5 are equivalent, i.e. the law of a random variable S is a stable law if and only if it arises from a generalized CLT as in (2.2) (see Chapter 1 in Samorodnitsky and Taqqu 1994 and references therein). The tail behaviour of the distribution of X_1 determines whether it lies in the DOA of a stable law. For a definition of the set of slowly varying functions $\mathcal{R}_0$ see Definition A.2 in the appendix.

Theorem 7. A distribution μ is in the DOA of a stable law with exponent $\alpha \in (0, 2)$ if and only if

i) $\lim\limits_{x\to\infty} \dfrac{\mu(x, \infty)}{\mu(-\infty, x) + \mu(x, \infty)} = M \in [0, 1]$

ii) If $M > 0$, then $\mu(x, \infty) = \dfrac{L^+(x)}{x^\alpha}$ for some slowly varying function L^+.

If $M > 0$, then $\mu(-\infty, -x) = \dfrac{L^-(x)}{x^\alpha}$ for some slowly varying function L^-.

The distribution μ is in the DOA of a normal law if and only if

$$\lim_{x\to\infty} \frac{x^2 \int_{|y|>x} \mathrm{d}\mu(y)}{\int_{|y|\leq x} y^2 \mathrm{d}\mu(y)} = 0$$

(see Theorem 9.34 and Theorem 9.41 in Breiman 1968).

A similar characterizing theorem for DOA of stable laws can be found in Section XVII.5 in Feller 1971 and also as Theorem 3.2 in Gut 2013 from which we will state the part relevant for our purposes:

Theorem 8. A random variable X belongs to the DOA of a stable distribution iff there exists $l \in \mathcal{R}_0$ such that

$$U(x) = \mathbb{E}[X^2 \mathbb{1}_{\{|X| \leq x\}}] \sim x^{2-\alpha} l(x), \quad (x \to \infty). \tag{2.3}$$

Remark 1. The function U can be interpreted as a truncated (second) moment function. The proof explicitly constructs the norming constants a_n in 2.2 as

$$(a_n)^{-1} = \inf \left\{ x : \frac{nU(x)}{x^2} \leq 1 \right\}.$$

From this together with 2.3 one can conclude that $a_n \in \mathcal{R}_{-1/\alpha}$ (as a function of n).

A subclass of stable distributions of particular interest in the next section are the totally right skewed stable distributions $S_\alpha(\sigma, 1, 0)$ with $\alpha \in (0, 1)$. The density function has support in $\mathbb{R}_+$. In this special case, the Laplace transform of $X \sim S_\alpha(\sigma, 1, 0)$ is

$$\mathbb{E}[\mathrm{e}^{-\gamma X}] = \exp\left(-\frac{\sigma^\alpha}{\cos(\pi\alpha/2)} \gamma^\alpha \right), \qquad \alpha \in (0, 1). \tag{2.4}$$

For convenience, we choose

$$\sigma^* = \left(\cos\left(\frac{\pi\alpha}{2} \right) \right)^{1/\alpha} \geq 0,$$

and $S_\alpha(\sigma^*, 1, 0)$ for $\alpha \in (0, 1)$ is the underlying infinitely divisible distribution of a strictly increasing Lévy process in the next section.

2.2 The stable subordinator and its inverse

Definition 6. Let $\alpha \in (0, 1)$, then the α-stable subordinator $(L_\alpha(t))_{t \geq 0}$ is a positive valued Lévy process with Laplace transform

$$\phi(u) = \mathbb{E}[\mathrm{e}^{-uL_\alpha(t)}] = \mathrm{e}^{-tu^\alpha}.$$

The *inverse α-stable subordinator* $(Y_\alpha(t))_{t \geq 0}$ is defined by

$$Y_\alpha(t) := \inf\{u \geq 0 : L_\alpha(u) > t\}. \tag{2.5}$$

The Laplace transform of the one-dimensional marginal distribution associated with the inverse α-stable subordinator can be expressed in terms of the Mittag-Leffler function: Let $h(t, \cdot)$ be the density function of the distribution of $Y_\alpha(t)$, then

$$\int_0^\infty \mathrm{e}^{-su} h_\alpha(t, u) \mathrm{d}u = E_\alpha(-st^\alpha)$$

(see Proposition 1(a) iii) in Bingham 1971).
Note that both L_α and Y_α are self-similar, i.e.

$$L_\alpha(t) \stackrel{d}{=} t^{1/\alpha} L_\alpha(1) \quad \text{and} \quad Y_\alpha(t) \stackrel{d}{=} t^\alpha Y_\alpha(1)$$

and in particular we have the relation

$$Y_\alpha(t) \stackrel{d}{=} \left(\frac{L_\alpha(1)}{t}\right)^{-\alpha}, \quad t > 0,$$

(see Corollary 3.1 (a) in Meerschaert and Scheffler 2004), which implies

$$h_\alpha(t,x) = \frac{t}{\alpha x^{1+\frac{1}{\alpha}}} g_\alpha\left(\frac{t}{x^{\frac{1}{\alpha}}}\right), \quad x \geq 0, t \geq 0, \tag{2.6}$$

where g_α is the probability density of $L_\alpha(1)$.
Using Equation (2.5), we can derive the following limit for the distribution function of Y_α:

$$\mathbb{P}(Y_\alpha(t) \leq u) = \mathbb{P}(D_\alpha(u) \geq t) \xrightarrow[t\to\infty]{} 0.$$

It is possible to say more about the asymptotic behaviour by invoking a Tauberian theorem.

Proposition 9. Let $(Y_\alpha(t))_{t\geq 0}$ be the inverse α-stable subordinator with distribution density $x \mapsto h_\alpha(t,x)$. Then for fixed $u \geq 0$, $\alpha \in (0,1)$

$$h_\alpha(t,x) \sim \frac{(1-\alpha)}{\Gamma(2-\alpha)} t^{-\alpha} \qquad (t \to \infty).$$

Proof. We use a Tauberian theorem which can be found in Simon 1979 (see Theorem A.4 in Appendix A) To this end, define the measure

$$\mu(\mathrm{d}t) = h_\alpha(t,x)\mathrm{d}t.$$

We denote its Laplace(-Stieltjes) transform by

$$G(s) := \int_0^\infty \mathrm{e}^{-ts}\mu(\mathrm{d}t) = \int_0^\infty \mathrm{e}^{-st} h_\alpha(t,x)\mathrm{d}t = s^{\alpha-1}\exp(-xs^\alpha)$$

Set $\gamma = 1-\alpha$, then $s^\gamma G(s) \to 1$ for $s \to 0$. Then it follows by Theorem A.4 that

$$b^{-\gamma}\mu([0,b)) \sim \frac{1}{\Gamma(2-\alpha)}$$

which is equivalent to

$$\int_0^b h_\alpha(t,x)\,\mathrm{d}t \sim \frac{1}{\Gamma(2-\alpha)} b^{1-\alpha}.$$

As both sides are differentiable w.r.t. b we may use l'Hospital's rule to derive asymptotics for the density h. On the one hand, it holds that

$$1 = \lim_{b\to\infty} \frac{\int_0^b h_\alpha(t,x)\,\mathrm{d}t}{\frac{1}{\Gamma(2-\alpha)} b^{1-\alpha}} = \lim_{b\to\infty} \frac{h_\alpha(b,x)}{\frac{1}{\Gamma(2-\alpha)}(1-\alpha)b^{-\alpha}}$$

and on the other hand we have

$$1 = \lim_{b\to\infty} \frac{\frac{1}{\Gamma(2-\alpha)} b^{1-\alpha}}{\int_0^b h_\alpha(t,x)\,\mathrm{d}t} = \lim_{b\to\infty} \frac{\frac{1}{\Gamma(2-\alpha)}(1-\alpha)b^{-\alpha}}{h_\alpha(b,x)}.$$

Thus it follows that

$$h_\alpha(t,x) \sim \frac{(1-\alpha)}{\Gamma(2-\alpha)} t^{-\alpha}.$$

□

2.3 Definition of the fractional Poisson process

The fractional homogeneous Poisson process (FHPP) can be defined in two equivalent ways. First, the *renewal approach* in defining the FHPP replaces the exponential waiting time distribution of the standard homogeneous Poisson process (see p. 17) with a Mittag-Leffler distribution.

$$N_\alpha(t) := \sup\{n \in \mathbb{N} : T_n < t\}, \qquad T_n = \sum_{k=1}^{n} J_k,$$

where (J_k) are i.i.d. distributed with survival function $\mathbb{P}(J_k > t) = E_\alpha(-\lambda t^\alpha)$. As shown in Theorem. 2.2 in Meerschaert, Nane and Vellaisamy 2011, one can also define the FHPP in an equivalent way via the *time-change approach*. Let $(N_\lambda(t))$ be a homogeneous Poisson process with parameter λ and $(Y_\alpha(t))$ be the inverse α-stable subordinator:

$$N_\alpha(t) := N_\lambda(Y_\alpha(t)).$$

Remark 2 (Mittag-Leffler distribution)**.** Note that the term "Mittag-Leffler distribution" can refer to two kinds of probability distributions which are not equivalent:

(i) The distribution of the inverse α-stable subordinator due to its relation to the

Mittag-Leffler function via its Laplace transform (see Section 2.2).

(ii) The distribution of the waiting times J_k in the renewal representation of the FHPP:

$$\mathbb{P}(J_k > t) = E_\alpha(-\lambda t^\alpha)$$

Analogous to the introduction of fractionality of the FHPP, we propose the construction of a fractional non-homogeneous Poisson process using the subordination approach. Incidentally, we cannot follow the renewal approach as the non-homogeneous Poisson process cannot be represented as a renewal process[3].

Definition 7. Let $(N_1(t))_{t\geq 0}$ be a Poisson process with parameter $\lambda = 1$, $(Y_\alpha(t))_{t\geq 0}$ an inverse α-stable subordinator and Λ a rate function. Then the *fractional non-homogeneous Poisson process of first kind* is defined as

$$N_\alpha(t) = N_1(\Lambda(Y_\alpha(t))), \quad t \geq 0, 0 < \alpha < 1 \tag{2.7}$$

and the *fractional non-homogeneous Poisson process of second kind* is given by

$$\tilde{N}_\alpha(t) := N_1(Y_\alpha(\Lambda(t))), \quad t \geq 0, 0 < \alpha < 1.$$

Remark 3. We will refer to the fractional non-homogeneous Poisson process (FNPP) as the first kind defined in (2.7) unless stated otherwise. For results on the second kind see for example Maheshwari and Vellaisamy 2017.

2.4 Governing equations

2.4.1 Fractional differential operators

The Caputo derivative is a possible way to define a fractional derivative (Caputo 1967). [4]

Definition 8. The Caputo derivative is defined as

$$D_t^\alpha f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\mathrm{d}f(\tau)}{\mathrm{d}\tau} \frac{\mathrm{d}\tau}{(t-\tau)^\alpha}, \quad 0 < \alpha < 1. \tag{2.8}$$

According to Theorem 2.1 in Kilbas, Srivastava and Trujillo 2006 the Caputo derivative exists and is well-defined for absolutely continuous functions. The following

[3] Although the non-homogeneous Poisson process cannot be represented as a classical renewal process, there exists a construction method for generalized renewal processes. For details see Gergely and Yezhow 1973.

[4] For other notable fractional derivatives such as the Grünwald and Riemann-Liouville fractional derivative see Chapter 2 in Meerschaert and Sikorskii 2012.

formula for the Laplace transform of the Caputo derivative is well known and we provide a proof for the sake of completeness.

Proposition 10. Let f be in the domain of the Caputo derivative with Laplace transform $\tilde{f}(s) := \mathcal{L}\{f; s\}$ and let $g(t) := D_t^\alpha f(t)$. Then, the Laplace transform of g is given by

$$\mathcal{L}\{D_t^\alpha f; s\} = s^\alpha \tilde{f}(s) - s^{\alpha-1} f(0+) \tag{2.9}$$

Proof. Using the definition in (2.8) we get

$$\begin{aligned}
\mathcal{L}\{D_t^\alpha f; s\} &= \int_0^\infty \mathrm{e}^{-st} D_t^\alpha f(t)\,\mathrm{d}t = \frac{1}{\Gamma(1-\alpha)} \int_0^\infty \mathrm{e}^{-st} \left(\int_0^t \frac{\mathrm{d}}{\mathrm{d}\tau} f(\tau) \frac{\mathrm{d}\tau}{(t-\tau)^\alpha} \right) \mathrm{d}t \\
&= \frac{1}{\Gamma(1-\alpha)} \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}\tau} f(\tau) \underbrace{\left(\int_\tau^\infty \mathrm{e}^{-st} \frac{\mathrm{d}t}{(t-\tau)^\alpha} \right)}_{= \mathrm{e}^{-\tau s} \frac{\Gamma(1-\alpha)}{s^{-\alpha+1}}} \mathrm{d}\tau \qquad (2.10) \\
&= s^{\alpha-1} \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}\tau} f(\tau) \mathrm{e}^{\tau s} \mathrm{d}\tau = s^{\alpha-1}(s\tilde{f}(s) - f(0+)) = s^\alpha \tilde{f}(s) - s^{\alpha-1} f(0+),
\end{aligned}$$

where in (2.10) we used Fubini's theorem. □

2.4.2 The homogeneous case

The governing equations of the one-dimensional marginal distribution of the FHPP can be derived using the Fourier-Laplace transform. The equations were first introduced by Laskin 2003 and Beghin and Orsingher 2009, 2010. The section follows the presentation found in Meerschaert and Sikorskii 2012.

Theorem 11. The one-dimensional marginal distributions

$$p_x^\alpha(t) := \mathbb{P}(N_\lambda(Y_\alpha(t)) = x), \quad x = 1, 2, \ldots$$

satisfy the following Cauchy problem

$$\begin{cases} D_t^\alpha p_x^\alpha(t) = \lambda(p_x^\alpha(t) - p_{x-1}^\alpha(t)) \\ p_x^\alpha(0) = \delta_0(x); \quad p_{-1}^\alpha \equiv 0. \end{cases} \tag{2.11}$$

Proof. By a conditioning argument we can write

$$p_x^\alpha(t) = \int_0^\infty \mathrm{e}^{-\lambda u} \frac{(\lambda u)^x}{x!} h_\alpha(t, u)\,\mathrm{d}u$$

Next, we apply the Fourier transform w.r.t. the state variable x and the Laplace

transform w.r.t. the time variable t, which yields

$$\begin{aligned}\overline{p}_y^\alpha(s) &= \int_0^\infty \exp(\lambda u(\mathrm{e}^{\mathrm{i}y}-1))s^{\alpha-1}\exp(-us^\alpha)\,\mathrm{d}u \\ &= \int_0^\infty s^{\alpha-1}\exp(u(\lambda(\mathrm{e}^{\mathrm{i}y}-1)-s^\alpha))\,\mathrm{d}u \\ &= \frac{s^{\alpha-1}}{\lambda(\mathrm{e}^{\mathrm{i}y}-1)-s^\alpha}\exp(u(\lambda(\mathrm{e}^{\mathrm{i}y}-1)-s^\alpha))\,\big|_{u=0}^\infty = \frac{s^{\alpha-1}}{s^\alpha-\lambda(\mathrm{e}^{\mathrm{i}y}-1)}. \end{aligned} \tag{2.12}$$

The limit in (2.12) can be calculated as follows:

$$\begin{aligned}\lim_{u\to\infty}\exp((\lambda(\mathrm{e}^{\mathrm{i}y}-1)-s^\alpha)u) &= \lim_{u\to\infty}\exp(\lambda u(\cos(y)+\mathrm{i}\sin(y)-1)-s^\alpha u) \\ &= \lim_{u\to\infty}\underbrace{\exp(\mathrm{i}\lambda\sin(y))}_{|\cdot|=1}\underbrace{\exp(\lambda(\cos(y)-1))}_{\leq 1}\exp(-s^\alpha u) = 0.\end{aligned}$$

Next, we apply the same Fourier-Laplace transform to Equation (2.11) in order to compare it with the result in (2.12). First, the Laplace transform of Equation (2.11) w.r.t. t is

$$s^\alpha\tilde{p}_x^\alpha(s) - s^{\alpha-1}\underbrace{p_x^\alpha(0+)}_{=\delta_0(x)} = \lambda(\tilde{p}_x^\alpha(s)-\tilde{p}_{x-1}^\alpha(s)), \tag{2.13}$$

where we have used the Formula (2.9) for the Laplace transform of the Caputo derivative according to Proposition 10 and the initial condition $p_x^\alpha(0+) = \delta_0(x)$. Next, the Fourier transform w.r.t. x of above equation yields

$$s^\alpha\overline{p}_y^\alpha(s) - s^{\alpha-1} = \lambda(\mathrm{e}^{\mathrm{i}y}-1)\overline{p}_y^\alpha(s).$$

Finally, we are able to solve for $\overline{p}_y^\alpha$ algebraically:

$$\overline{p}_y^\alpha(s) = \frac{s^{\alpha-1}}{s^\alpha-\lambda(\mathrm{e}^{\mathrm{i}y}-1)},$$

which coincides with (2.12). The assertion follows by uniqueness of the Fourier-Laplace transform. □

Remark 4. The above Cauchy problem is also solved in the proof of Theorem 2.1 in Beghin and Orsingher 2010 by using the Laplace transform only. The proof resolves the recursion in Equation (2.13) by using the initial conditions:

$$\tilde{p}_x^\alpha(s) = \frac{\lambda^x s^{\alpha-1}}{(s^\alpha+\lambda)^{x+1}}, \quad x = 0,1,2,\ldots$$

Using Proposition 5 we can invert the Laplace transform to obtain an expression for

p_x^a in terms of the Mittag-Leffler function:

$$p_x^\alpha(t) = (\lambda t^\alpha)^x E_{\alpha,\alpha x+1}^{x+1}(-\lambda t^\alpha).$$

This coincides with the marginals obtained from the renewal approach (see Mainardi, Gorenflo and Scalas 2004).

2.4.3 The non-homogeneous case

In the following, we derive a differential equation involving the distribution of the increments of the FNPP that generalizes the results in Theorem 11, Equation 1.5 and Equation 1.9. In other words, the governing equations of both the standard Poisson process and the fractional homogeneous Poisson process are contained as special cases.
To this end, first consider the process $(I(t,v))_{t\geq 0}$ for $v \geq 0$ defined as

$$I(t,v) = N_1(\Lambda(t+v)) - N_1(\Lambda(v)),$$

where $(N_1(t))_{t\geq 0}$ is a homogeneous Poisson process with intensity $\lambda = 1$ and Λ is a rate function as described on page 19. We will refer to $(I(t,v))_{t\geq 0}$ as *increment process* of the non-homogeneous Poisson process.
The fractional increment process of the NPP is given by

$$I_\alpha(t,v) := I(Y_\alpha(t),v) = N_1(\Lambda(Y_\alpha(t)+v)) - N_1(\Lambda(v)). \tag{2.14}$$

and its marginals will be denoted as

$$\begin{aligned} f_x^\alpha(t,v) &:= \mathbb{P}\{N_1(\Lambda(Y_\alpha(t)+v)) - N_1(\Lambda(v)) = x\}, \quad x = 0,1,2,\dots \\ &= \int_0^\infty p_x(u,v) h_\alpha(t,u)\mathrm{d}u \\ &= \int_0^\infty \frac{\mathrm{e}^{-\Lambda(v,u+v)}\Lambda(v,u+v)^x}{x!} h_\alpha(t,u)\mathrm{d}u. \end{aligned} \tag{2.15}$$

For the FNPP the marginal distributions are given by

$$\begin{aligned} f_x^\alpha(t,0) = \mathbb{P}\{N_\alpha(t) = x\} &= \int_0^\infty p_x(u) h_\alpha(t,u)\mathrm{d}u \\ &= \int_0^\infty \frac{\mathrm{e}^{-\Lambda(u)}\Lambda(u)^x}{x!} h_\alpha(t,u)\mathrm{d}u, \quad x = 0,1,2,\dots \end{aligned} \tag{2.16}$$

For shorthand notation we write $f_x^\alpha(t) := f_x^\alpha(t,0)$.

Theorem 12. Let $I_\alpha(t,v)$ be the fractional increment process defined in (2.14).

Then, its marginal distribution given in (2.15) satisfies the following fractional differential-integral equations

$$D_t^\alpha f_x^\alpha(t,v) = \int_0^\infty \lambda(u+v)[-p_x(u,v) + p_{x-1}(u,v)]h_\alpha(t,u)\mathrm{d}u, \quad x = 0,1,\ldots, \tag{2.17}$$

with initial condition

$$f_x^\alpha(0,v) = \begin{cases} 1, & x = 0, \\ 0, & x \geq 1 \end{cases} \tag{2.18}$$

and $f_{-1}^\alpha(0,v) \equiv 0$, where $p_x(u,v)$ is given by (1.8) (with $p_{-1}(u,v) = 0$) and $h_\alpha(t,u)$ is given by (2.6).

Proof. The initial conditions are easily checked using the fact that $Y_\alpha(0) = 0$ a.s and it remains to prove (2.17). Let f_x^α be defined as in Equation (2.15). Taking the characteristic function of f_x^α and the Laplace transform w.r.t. t yields

$$\begin{aligned} \bar{f}_y^\alpha(r,v) &= \int_0^\infty \hat{p}_y(u,v)\tilde{h}_\alpha(r,u)\mathrm{d}u \\ &= \int_0^\infty \exp(\Lambda(v,u+v)(\mathrm{e}^{\mathrm{i}y}-1))r^{\alpha-1}\mathrm{e}^{-ur^\alpha}\mathrm{d}u. \end{aligned}$$

Using integration by parts we get

$$\begin{aligned} \bar{f}_y^\alpha(r,v) = r^{\alpha-1}\Bigg[& -\frac{1}{r^\alpha}\underbrace{\mathrm{e}^{-ur^\alpha}\exp(\Lambda(v,u+v)(\mathrm{e}^{\mathrm{i}y}-1))\Big|_{u=0}^\infty}_{=1} \\ & + \frac{1}{r^\alpha}\int_0^\infty \left(\frac{\mathrm{d}}{\mathrm{d}u}\Lambda(v,u+v)\right)(\mathrm{e}^{\mathrm{i}y}-1)\exp(\Lambda(v,u+v)(\mathrm{e}^{\mathrm{i}y}-1))\mathrm{e}^{-ur^\alpha}\mathrm{d}u\Bigg] \\ = \frac{1}{r^\alpha}\Bigg[& r^{\alpha-1} + (\mathrm{e}^{\mathrm{i}y}-1)\int_0^\infty \lambda(u+v)\ \exp(\Lambda(v,u+v)(\mathrm{e}^{\mathrm{i}y}-1))r^{\alpha-1}\mathrm{e}^{-ur^\alpha}\mathrm{d}u\Bigg]. \end{aligned}$$

Now we are able to calculate the Caputo derivative in Laplace space using Equation (2.9). Note that $\hat{f}_y^\alpha(0^+,v) = 1$ as $Y_\alpha(0) = 0$ a.s.

$$\begin{aligned} & r^\alpha \bar{f}_y^\alpha(r,v) - r^{\alpha-1} \\ & \qquad = (\mathrm{e}^{\mathrm{i}y}-1)\int_0^\infty \lambda(u+v)\exp(\Lambda(v,u+v)(\mathrm{e}^{\mathrm{i}y}-1))r^{\alpha-1}\mathrm{e}^{-ur^\alpha}\mathrm{d}u \\ & \qquad = (\mathrm{e}^{\mathrm{i}y}-1)\int_0^\infty \lambda(u+v)\hat{p}_y(u,v)\tilde{h}_\alpha(r,u)\mathrm{d}u. \end{aligned}$$

Inversion of the Laplace transform yields

$$D_t^\alpha \hat{f}_y^\alpha(t,v) = (\mathrm{e}^{\mathrm{i}y}-1)\int_0^\infty \lambda(u+v)\hat{p}_y(u,v)h_\alpha(t,u)\mathrm{d}u$$

and finally, by inverting the characteristic function, we obtain

$$D_t^\alpha f_x^\alpha(t,v) = \int_0^\infty \lambda(u+v)[-p_x(u,v) + p_{x-1}(u,v)]h_\alpha(t,u)\mathrm{d}u. \tag{2.19}$$

which was to be shown. □

Directly from Theorem 12 setting $v = 0$ one gets

Corollary 1. Let $N_\alpha(t)$, $t \geq 0$, $0 < \alpha < 1$ be a FNPP given by (2.7). Then, its marginal distributions shown in (2.16) satisfy the following fractional differential-integral equations:

$$D_t^\alpha f_x^\alpha(t) = \int_0^\infty \lambda(u)[-p_x(u) + p_{x-1}(u)]h_\alpha(t,u)\mathrm{d}u, \tag{2.20}$$

with initial condition

$$f_x^\alpha(0) = \begin{cases} 1, & x = 0, \\ 0, & x \geq 1 \end{cases} \tag{2.21}$$

and $f_{-1}^\alpha(0) \equiv 0$, where $p_x(u)$ is given by (1.8) and $h_\alpha(t,u)$ is given by (2.6).

Special cases

It is useful to consider two special cases of the governing equations derived above, the FHPP and the NPP.

(i) To get back to the FHPP we choose $\lambda(t) = \lambda > 0$ as a constant to get

$$\begin{aligned} D_t^\alpha f_x^\alpha(t) &= \lambda \int_0^\infty [-p_x(u) + p_{x-1}(u)]h_\alpha(t,u)\mathrm{d}u \\ &= -\lambda f_x^\alpha(t) + \lambda f_{x-1}^\alpha(t) \end{aligned} \tag{2.22}$$

which is identical with (2.11). Indeed for constant λ in (2.16) we get

$$f_x^\alpha(t) = \int_0^\infty \frac{\mathrm{e}^{-u\lambda}(\lambda u)^x}{x!} h_\alpha(t,u)\mathrm{d}u = p_x^\alpha(t),$$

i.e. f_x^α coincides with the marginal probabilities of the FHPP.

(ii) To obtain the case of the NPP, we consider $\alpha = 1$ for which we have $\tilde{h}_1(s,u) = \mathrm{e}^{-us}$ and its inverse Laplace transform is the delta distribution: $\mathcal{L}^{-1}\{\tilde{h}\}(t,u) = \delta(t-u)$. By replacing this in Equation (2.15) we formally get

$$f_x^1(t,v) = \int_0^\infty p_x(u,v)\delta(t-u)\mathrm{d}u = p_x(t,v),$$

which means that f_x^1 coincides with the marginal probabilities p_x of the NPP.

Moreover, the proof of Theorem 12 is still valid and by substituting Dirac's delta distribution in Equation (2.17) we get for $t \geq 0$

$$\begin{aligned} D_t^1 p_x(t,v) &= D_t^1 f_x^1(t,v) \\ &= \int_0^\infty \lambda(u+v)[-p_x(u+v) + p_{x-1}(u,v)]\delta(t-u)\mathrm{d}u \\ &= \lambda(t+v)[-p_x(t,v) + p_{x-1}(t,v)] \end{aligned}$$

which coincides with (1.10). Related to this formal calculation, Proposition 33 shows a convergence result for the limit $\alpha \to 1$.

2.5 Moments and covariance structure

As a further characterization of the FNPP, we now give the first moments of its distribution, namely the expectation, the variance and the covariance.

2.5.1 Moments

For fixed $t > 0$, the moments of the Poisson distribution with rate $\Lambda(t)$ can be calculated via the derivatives of its characteristic function. However, the most explicit formula for higher moments of the Poisson distribution is given by

$$\mathbb{E}[[N(t)]^k] = \sum_{i=1}^{k} \Lambda(t)^i \left\{ \begin{array}{c} k \\ i \end{array} \right\}, \tag{2.23}$$

where $\left\{ \begin{array}{c} k \\ i \end{array} \right\}$ are the Stirling numbers of second kind:

$$\left\{ \begin{array}{c} k \\ i \end{array} \right\} = \frac{1}{i!} \sum_{j=0}^{i} (-1)^{i-j} \left(\begin{array}{c} i \\ j \end{array} \right) j^k.$$

Equation (2.23) is the non-homogeneous generalization of the Dobiński formula (see Dobiński 1877). Polynomials of the form:

$$q_k(x) = \sum_{i=1}^{k} x^i \left\{ \begin{array}{c} k \\ i \end{array} \right\}$$

are known as Touchard polynomials, exponential polynomials or Bell polynomials. Note that the first moment is

$$\mathbb{E}[N(t)] = \Lambda(t)$$

and the second moment is given by

$$\mathbb{E}[[N(t)]^2] = \Lambda(t) + \Lambda(t)^2,$$

which we will use later for the calculation of the expectation, variance and covariance.
Thus for the higher moments of the subordinated process we have

$$\begin{aligned}\mathbb{E}[[N(Y_\alpha(t))]^k] = \mathbb{E}[\mathbb{E}[[N(Y_\alpha(t))]^k | Y_\alpha(t)]] = & \int_0^\infty \mathbb{E}[[N(x)]^k] h_\alpha(t,x)\mathrm{d}x \\ = \int_0^\infty \sum_{i=1}^k \Lambda(x)^i \left\{ \begin{array}{c} k \\ i \end{array} \right\} h_\alpha(t,x)\mathrm{d}x = & \mathbb{E}\left[\sum_{i=1}^k \Lambda(Y_\alpha(t))^i \left\{ \begin{array}{c} k \\ i \end{array} \right\}\right]. \end{aligned} \tag{2.24}$$

Expectation and variance immediately follow from 2.24. The expectation is

$$\mathbb{E}[N(Y_\alpha(t))] = \mathbb{E}[\Lambda(Y_\alpha(t))]. \tag{2.25}$$

Then, using

$$\mathbb{E}[[N(Y_\alpha(t))]^2] = \mathbb{E}[\Lambda(Y_\alpha(t))] + \mathbb{E}[\Lambda(Y_\alpha(t))^2], \tag{2.26}$$

we find

$$\mathrm{Var}[N(Y_\alpha(t))] = \mathbb{E}[[N(Y_\alpha(t))]^2] - \mathbb{E}[[N(Y_\alpha(t))]]^2 = \mathbb{E}[\Lambda(Y_\alpha(t))] + \mathrm{Var}[\Lambda(Y_\alpha(t))]. \tag{2.27}$$

2.5.2 Covariance

Let $s, t \in \mathbb{R}_+$ and w.l.o.g. assume $s < t$. Then

$$\begin{aligned}\mathbb{E}[N(s)N(t)] &= \mathbb{E}[N(t) - N(s)]\mathbb{E}[N(s)] + \mathbb{E}[N(s)^2] \\ &= \Lambda(s,t)\Lambda(0,s) + \Lambda(0,s)^2 + \Lambda(0,s)\end{aligned}$$

and thus

$$\begin{aligned}\mathrm{Cov}(N(s), N(t)) &= \mathbb{E}[N(s)N(t)] - \mathbb{E}[N(s)]\mathbb{E}[N(t)] \\ &= \Lambda(s,t)\Lambda(0,s) + \Lambda(0,s)^2 + \Lambda(0,s) - \Lambda(0,s)\Lambda(0,t) \\ &= \Lambda(0,s)[\Lambda(s,t) + \underbrace{\Lambda(0,s) - \Lambda(0,t)}_{=-\Lambda(s,t)} + 1] = \Lambda(0,s).\end{aligned}$$

The same calculation can be done for the case $t < s$. In short, both cases can be summarized in the following way:

$$\mathrm{Cov}(N(s), N(t)) = \Lambda(0, s \wedge t). \tag{2.28}$$

Proposition 13. By the law of total covariance, one finds:

$$\begin{aligned}\mathrm{Cov}[N(Y_\alpha(s)), N(Y_\alpha(t))] &= \mathbb{E}[\mathrm{Cov}[N(Y_\alpha(s)), N(Y_\alpha(t))|Y_\alpha(s), Y_\alpha(t)]] \\ &\quad + \mathrm{Cov}[\mathbb{E}[N(Y_\alpha(s))|Y_\alpha(s), Y_\alpha(t)], \mathbb{E}[N(Y_\alpha(t))|Y_\alpha(s), Y_\alpha(t)]] \\ &= \mathbb{E}[\Lambda(0, Y_\alpha(s \wedge t))] + \mathrm{Cov}[\Lambda(Y_\alpha(s)), \Lambda(Y_\alpha(t))]\end{aligned} \tag{2.29}$$

Proof. For the first term, we have

$$\begin{aligned}&\mathbb{E}[\mathrm{Cov}[N(Y_\alpha(s)), N(Y_\alpha(t))|Y_\alpha(s), Y_\alpha(t)]] = \mathbb{E}[\mathbb{E}[N(Y_\alpha(s))N(Y_\alpha(t))]|Y_\alpha(s), Y_\alpha(t)] \\ &\quad - \mathbb{E}[N(Y_\alpha(s))|Y_\alpha(s), Y_\alpha(t)]\mathbb{E}[N(Y_\alpha(t))|Y_\alpha(s), Y_\alpha(t)] \\ &\quad = \int_0^\infty \int_0^\infty \mathbb{E}[N(x)N(y)] p_{(Y_\alpha(s), Y_\alpha(t))}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &\quad - \int_0^\infty \int_0^\infty \mathbb{E}[N(x)]\mathbb{E}[N(y)] p_{(Y_\alpha(s), Y_\alpha(t))}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &\quad = \int_0^\infty \int_0^\infty \mathrm{Cov}[N(x), N(y)] p_{(Y_\alpha(s), Y_\alpha(t))}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &\quad = \mathbb{E}[\Lambda(0, Y_\alpha(s) \wedge Y_\alpha(t))] = \mathbb{E}[\Lambda(Y_\alpha(s \wedge t))].\end{aligned}$$

Note that in the last step we have used that Y_α is an increasing process.
For the second term:

$$\begin{aligned}&\mathrm{Cov}[\mathbb{E}[N(Y_\alpha(s))|Y_\alpha(s), Y_\alpha(t)], \mathbb{E}[N(Y_\alpha(t))|Y_\alpha(s), Y_\alpha(t)]] \\ &\quad = \mathbb{E}[\mathbb{E}[N(Y_\alpha(s))|Y_\alpha(s), Y_\alpha(t)]\mathbb{E}[N(Y_\alpha(t))|Y_\alpha(s), Y_\alpha(t)]] \\ &\quad - \mathbb{E}[\mathbb{E}[N(Y_\alpha(s))|Y_\alpha(s), Y_\alpha(t)]]\mathbb{E}[\mathbb{E}[N(Y_\alpha(t))|Y_\alpha(s), Y_\alpha(t)]] \\ &\quad = \int_0^\infty \int_0^\infty \mathbb{E}[N(x)]\mathbb{E}[N(y)] p_{(Y_\alpha(s), Y_\alpha(t))}(x, y)\, \mathrm{d}x\, \mathrm{d}y - \mathbb{E}[N(Y_\alpha(s))]\mathbb{E}[N(Y_\alpha(t))] \\ &\quad = \mathbb{E}[\Lambda(Y_\alpha(s))\Lambda(Y_\alpha(t))] - \mathbb{E}[\Lambda(Y_\alpha(s))]\mathbb{E}[\Lambda(Y_\alpha(t))] \\ &\quad = \mathrm{Cov}[\Lambda(Y_\alpha(s)), \Lambda(Y_\alpha(t))],\end{aligned}$$

where $p_{(Y_\alpha(s), Y_\alpha(t))}(x, y)$ is the joint density of $Y_\alpha(s)$ and $Y_\alpha(t)$. □

Remark 5. The two-point cumulative distribution function of the inverse stable subordinator $Y_\alpha(t)$ can be computed using the fact that (see Leonenko, Meerschaert

and Sikorskii 2013)

$$\mathbb{P}(Y_\alpha(s) > x, Y_\alpha(t) > y) = \int_{v=0}^{t} \frac{\alpha}{v} y h_\alpha(s, y) \int_{u=0}^{s-v} \frac{\alpha}{u}(x-y) h_\alpha(t, x-y)\,\mathrm{d}u\,\mathrm{d}v. \quad (2.30)$$

Remark 6. For the homogeneous case $\Lambda(t) = \lambda t$, we get

$$\mathrm{Cov}[N(Y_\alpha(s)), N(Y_\alpha(t))] = \lambda \mathbb{E}[Y_\alpha(s \wedge t)] + \lambda^2 \mathrm{Cov}[Y_\alpha(s), Y_\alpha(t)],$$

which is consistent with the results in Leonenko, Meerschaert, Schilling et al. 2014.

Using a regular variation assumption on λ, we can derive asymptotics for the expectation of the FNPP.

Proposition 14. Let $\lambda \in \mathcal{R}_\rho$ and $(Y_\alpha(t))_{t\geq 0}$ be the inverse α-stable subordinator. Then

$$\mathbb{E}[\Lambda(Y_\alpha(t))] \in \mathcal{R}_{\alpha(\rho+1)}, \quad (t \to \infty). \quad (2.31)$$

Proof. Define

$$\phi(t) := \mathbb{E}[\Lambda(Y_\alpha(t))] = \int_0^\infty \Lambda(x) h(t, x)\,\mathrm{d}x.$$

Consider the Laplace transform of ϕ:

$$\begin{aligned}
\mathcal{L}\{\phi; s\} &= \int_0^\infty \mathrm{e}^{-st}\phi(t)\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-st}\left(\int_0^\infty \Lambda(x) h(t, x)\,\mathrm{d}x\right)\mathrm{d}t \\
&= \int_0^\infty \Lambda(x)\left(\int_0^\infty \mathrm{e}^{-st} h(t, x)\,\mathrm{d}t\right)\mathrm{d}x \quad (2.32) \\
&= \int_0^\infty \Lambda(x) s^{\alpha-1}\exp(-x s^\alpha)\,\mathrm{d}x = s^{\alpha-1}\int_0^\infty \Lambda(x)\exp(-x s^\alpha)\,\mathrm{d}x \\
&= s^{\alpha-1}\mathcal{L}\{\Lambda; s\}, \quad (2.33)
\end{aligned}$$

where we have used Fubini's theorem in (2.32). We are thus able to express the Laplace transform of ϕ in terms of the Laplace transform of Λ evaluated at s^α.
It follows from $\lambda \in \mathcal{R}_\rho$ that there exists a slowly varying function $g \in \mathcal{R}_0$ such that

$$\lambda(x) = x^\rho l(x), \qquad (x \to \infty).$$

By Proposition A.5 we get

$$\Lambda(t) \sim t^{\rho+1}\frac{l(t)}{1+\rho} \qquad (t \to \infty)$$

or in other words $\Lambda \in \mathcal{R}_{\rho+1}$. The Laplace-Stieltjes transform of Λ is given by

$$\tilde{\Lambda}(s) = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}\Lambda(t) = \int_0^\infty \mathrm{e}^{-st}\lambda(t)\,\mathrm{d}t.$$

Karamata's Tauberian theorem (Theorem A.4) yields

$$\tilde{\Lambda}(s) \sim s^{-\rho-1} \frac{l\left(\frac{1}{s}\right)}{(1+\rho)\Gamma(1+\rho)} \qquad (s \to 0+),$$

which implies

$$\begin{aligned}\mathcal{L}\{\Lambda; s\} &= \int_0^\infty \mathrm{e}^{-st} \int_0^t \lambda(\tau)\,\mathrm{d}\tau\,\mathrm{d}t = \underbrace{-\frac{1}{s}\mathrm{e}^{-st}\int_0^t \lambda(\tau)\,\mathrm{d}\tau\Big|_{t=0}^{\infty}}_{=0} + \frac{1}{s}\int_0^\infty \mathrm{e}^{-st}\lambda(t)\,\mathrm{d}t \\ &= \frac{1}{s}\tilde{\Lambda}(s) \sim s^{-\rho-2}\frac{l\left(\frac{1}{s}\right)}{(1+\rho)\Gamma(1+\rho)} \qquad (s \to 0+).\end{aligned}$$

By Proposition A.2 (ii) for the composition of regularly varying functions we may conclude that

$$\begin{aligned}\mathcal{L}\{\Lambda, s^\alpha\} &= \frac{1}{s^\alpha}\tilde{\Lambda}(s^\alpha) \sim s^{-\alpha}s^{\alpha(-\rho-1)}\frac{l\left(\frac{1}{s^\alpha}\right)}{(1+\rho)\Gamma(1+\rho)} \quad (s \to 0+) \\ &= s^{-1-\alpha(\rho+1)}\frac{l_1\left(\frac{1}{s}\right)}{(1+\rho)\Gamma(1+\rho)},\end{aligned}$$

where $l_1(s) = l(s^\alpha)$ is another slowly varying function. Plugging above into Equation (2.33) yields

$$\begin{aligned}\mathcal{L}\{\phi; s\} &= s^{\alpha-1}\mathcal{L}\{\Lambda; s\} \sim s^{\alpha-1-\alpha\rho-2\alpha}\frac{l_1(\frac{1}{s})}{(1+\rho)\Gamma(1+\rho)} \quad (s \to 0+) \\ &= s^{-1-\alpha(\rho+1)}\frac{l_1(\frac{1}{s})}{(1+\rho)\Gamma(1+\rho)}\end{aligned}$$

Using Karamata's Tauberian theorem again we get

$$\int_0^t \mathbb{E}[\Lambda(Y_\alpha(\tau))]\,\mathrm{d}\tau \sim t^{1+\alpha(\rho+1)}\frac{l_1(t)}{1+\rho}.$$

Since $t \mapsto \phi(t)$ is an increasing function, we may apply the monotone density theorem (Theorem A.6) to get

$$\mathbb{E}[\Lambda(Y_\alpha(t))] \sim t^{\alpha(\rho+1)}l_1(t), \quad (t \to \infty)$$

which yields the assertion. □

2.6 Summary

We have introduced a fractional non-homogeneous Poisson process as $N_\alpha(t) = N_1(\Lambda(Y_\alpha(t)))$ where $N_1(t)$ is the homogeneous Poisson process with $\lambda = 1$, $\Lambda(t)$ is the rate function and $Y_\alpha(t)$ is the inverse stable subordinator. It reduces to the usual non-homogeneous Poisson process in the case $\alpha = 1$ and additionally to the homogeneous Poisson process for a constant intensity λ. The calculations of moments for this process is a straightforward application of the rules for conditional expectations. The assumption of regular variation gives a result on the asymptotics of the mean.

Building on the ideas of this chapter, we will categorize the FNPP as a doubly stochastic process and derive limit theorems for the FNPP in the next chapter.

Chapter 3

Limit theorems

This chapter is devoted to the derivation of limit theorems for stochastic processes involving the fractional Poisson process, both the homogeneous and inhomogeneous one, as well as for the fractional compound Poisson process.

3.1 Preliminaries: Convergence in the Skorokhod space

First, we give a short introduction to weak convergence of probability measures on topological spaces, in particular with Polish (complete, separable and metrizable) spaces in mind. The primary aim is to provide the reader with the notation and terminology used in this chapter. Moreover, we would like to point out how the framework is linked to profound measure-theoretic, functional analytic and topological concepts as far as the scope of this section allows.

3.1.1 Weak convergence of probability measures and Riesz representation theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We consider a (continuous time) stochastic process $(X_t)_{t\geq 0}$ as a measurable map from Ω into a suitable function space E:

$$X : \Omega \to E, \quad \omega \mapsto X_\cdot(\omega),$$

i.e. $X_t(\omega) = x(t)$, $x \in E$. To ensure that the process is measurable w.r.t. $\mathcal{F}$ and adapted to a (right continuous) filtration $(\mathcal{F}_t)_{t\geq 0}$ we define

$$\mathcal{F}_t^0 := \sigma(X_s : s \leq t)$$
$$\mathcal{F} := \bigcup_{t\geq 0} \mathcal{F}_t^0, \quad \mathcal{F}_t := \bigcap_{s>t} \mathcal{F}_s^0.$$

The filtration $(\mathcal{F}_t)_{t\geq 0}$ and the σ-field $\mathcal{F}$ are thus generated by the cylinder sets associated with $(X_t)_{t\geq 0}$. The σ-field $\mathcal{F}$ is sometimes referred to as Kolmogorov σ-field due to its connection to Kolmogorov's extension theorem (see Theorem A.3.1. in Durrett 2010).

Typical paths of realizations of X belong to the space of càdlàg functions, which is also known as *Skorokhod space*, denoted by $\mathcal{D}$. The càdlàg path property allows for uniqueness results: for example Lévy processes are shown to have a unique càdlàg modification up to indistinguishability (see Theorem 2.1.8 in Applebaum 2009). The space of continuous functions $\mathcal{C}$ is an important subspace of $\mathcal{D}$ and we will discuss the cases $E = \mathcal{C}$ and $E = \mathcal{D}$. For our purposes, we need to equip E with a suitable metric d such that E becomes a Polish space. Additionally, the topology associated with d induces a Borel σ-field on E, which we will denote by $\mathcal{B}(E)$ or $\mathcal{B}$ for short. The Borel σ-field $\mathcal{B}$ has to be compatible with the Komogorov σ-field above to ensure consistent measurability.

The pushforward measure $\mathbb{P} \circ X^{-1}$ belongs to the set of probability measures over E, denoted by $\mathcal{P}(E)$. In order to make assertions about the limit of a sequence of stochastic processes $(X^n)_{n\in\mathbb{N}}$, $X^n = (X_t^n)_{t\geq 0}$, we first need to define a suitable convergence notion for the laws of (X^n), for which we write $\mathcal{L}(X_t^n)$.[1] In other words, when does a sequence of measures $\mu_n := \mathbb{P} \circ (X^n)^{-1}$ converge in $\mathcal{P}(E)$? There are several takes on how the choice of convergence can be motivated. We follow a measure-theoretic perspective which can be found in Chapter VIII in Elstrodt 2005. One might want to start from the idea that convergence of finite measures μ_n to a measure μ could be defined by demanding

$$\mu_n(A) \to \mu(A) \quad \forall\, A \in \mathcal{B}. \tag{3.1}$$

This is sometimes referred to as *strong convergence* of measures. Moreover, if we even require the above convergence to be uniform among the sets $A \in \mathcal{B}$, we arrive at the *convergence in total variation*. Both convergence concepts are too restrictive for our purposes as the following example shows.

Example 2 (see pp. 380-381 in Elstrodt 2005)**.** Let $E = \mathbb{R}$ and define the measures

$$\mu_n(A) := \mathbb{1}_A\left(\frac{1}{n}\right), \quad \mu(A) := \mathbb{1}_A(0), \quad n \in \mathbb{N},\, A \in \mathcal{B}.$$

Since the sequence $1/n$ converges to 0 for $n \to \infty$, We might want to have convergence of $\mu_n \to \mu$.

[1] It is possible for the processes (X_t^n) to exist on different probability spaces for different n, i.e. X^n is a map on $(\Omega^n, \mathcal{F}^n, \mathbb{P}^n)$. In this case, the law of X^n is to be understood under the respective measure $\mathbb{P}^n$. For simplicity and notational convenience we will assume that all stochastic processes in the sequence are on the same probability space.

However, for $A = (-\infty, 0]$ we have $\mu_n(A) = 0 \neq 1 = \mu(A)$ and for $A = (0, \infty)$ it holds that $\mu_n(A) = 1 \neq 0 = \mu(A)$. Both implie that μ_n does not converge to μ for $n \to \infty$ in the way defined by (3.1).

In order to relax the condition in (3.1) a plausible idea would be to require that it just holds for a subset of $\mathcal{B}$. In the way (3.1) is written, it is not clear how a reasonable restriction should look like. This is where following integral representation comes in handy.

Proposition 15. The convergence notion of (3.1) is equivalent to the convergence of the associated linear forms:

$$\int_E f \, \mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_E f \, \mathrm{d}\mu, \quad \forall \, f \in \mathcal{L}^\infty(E, \mathcal{B}(E), \mu). \tag{3.2}$$

Proof. "$\Leftarrow$" Let $A \in \mathcal{B}$. Choose $f = \mathbb{1}_A \in \mathcal{L}^\infty(E, \mathcal{B}(E), \mu)$, then

$$\lim_{n\to\infty} \mu_n(A) = \lim_{n\to\infty} \int_E \mathbb{1}_A \, \mathrm{d}\mu_n = \int_E \mathbb{1}_A \, \mathrm{d}\mu = \mu(A).$$

"$\Rightarrow$" Let $f \in \mathcal{L}^\infty(E, \mathcal{B}(E), \mu)$. Let $(f^m)_{m\in\mathbb{N}}$ be a sequence of step functions approximating f, i.e. there exist coefficients $a_1, \ldots, a_m$ and sets $A_1, \cdot, A_m$ such that f^m converges uniformly to f (see Corollary 4.14 in Elstrodt 2005).

$$\lim_{m\to\infty} \operatorname*{ess\,sup}_{x\in E} |f^m(x) - f| = \lim_{m\to\infty} \operatorname*{ess\,sup}_{x\in X} \left| \sum_{k=1}^{m} a_k \mathbb{1}_{A_k}(x) - f \right| = 0.$$

By dominated convergence we get

$$\lim_{n\to\infty} \int_E f \, \mathrm{d}\mu_n = \lim_{n\to\infty} \lim_{m\to\infty} \int_E f^m \, \mathrm{d}\mu_n. \tag{3.3}$$

According to the Moore-Osgood theorem (see Theorem 7.11 in Rudin 1976 or Theorem 2.1.4.1. in Gelbaum and Olmsted 1990), we need to show that at least one of the limits (either for $n \to \infty$ or $m \to \infty$) converges uniformly in order to exchange the limits. By assumption we have that $\mu_n(E) \to \mu(E)$ which implies that the sequence $(\mu_n(E))_n$ is bounded:

$$\sup_{n\in\mathbb{N}} \mu_n(E) < \infty.$$

Using that we derive

$$\begin{aligned}\lim_{m\to\infty}\sup_{n\in\mathbb{N}}\left|\int_E f^m - f\mathrm{d}\mu_n\right| &\leq \lim_{m\to\infty}\sup_{n\in\mathbb{N}}\int_E |f^m - f|\mathrm{d}\mu_n \\ &\leq \lim_{m\to\infty}\operatorname*{ess\,sup}_{x\in E}|f^m(x) - f(x)|\sup_{n\in\mathbb{N}}\mu_n(X) \\ &= 0,\end{aligned}$$

This proves that uniformly in n

$$\lim_{m\to\infty}\int_E f^m\,\mathrm{d}\mu_n = \int_E f\,\mathrm{d}\mu.$$

We are allowed to exchange the limits in (3.3):

$$\begin{aligned}\lim_{n\to\infty}\int_E f\,\mathrm{d}\mu_n = \lim_{m\to\infty}\lim_{n\to\infty}\int_E f^m\,\mathrm{d}\mu_n &= \lim_{m\to\infty}\lim_{n\to\infty}\sum_{k=1}^m a_k\mu_n(A_k) \\ &= \lim_{m\to\infty}\sum_{k=1}^m a_k\lim_{n\to\infty}\mu_n(A_k) = \lim_{m\to\infty}\sum_{k=1}^m a_k\lim_{n\to\infty}\mu(A_k) \\ &= \int_E f\,\mathrm{d}\mu,\end{aligned}$$

which proves the thesis. □

By the above proposition we are able to interpret the strong convergence of probability measures as convergence of positive linear forms on the function space $\mathcal{L}^\infty$. Instead of choosing a subset of $\mathcal{B}(E)$ for $\mu_n \to \mu$ to hold, we can alternatively replace $\mathcal{L}^\infty$ with a different function space. Such a function space for which the associated linear forms have to converge is sometimes referred to as *convergence determining class*[2]. Indeed, we are able to fully characterize positive linear forms on certain classes of continuous functions defined on Polish spaces. This is accomplished by the Riesz representation theorems, of which we will present the most relevant for probabilistic purposes.
Recall from Definition 1 on p. 15 that a locally finite measure on the Borel sets $\mathcal{B}$ is called a Borel measure.

Definition 9. Let $(E, \mathcal{B}, \mu)$ be a measure space.

(i) A measure μ is *inner regular* if for all $A \in \mathcal{B}$

$$\mu(A) = \sup\{\mu(K) : K \subset A,\ K \text{ compact}\}.$$

[2]In the finite dimensional case $\mathrm{E} = \mathbb{R}$, the characteristic functions of random variables are such a class due to the Cramér-World device.

(ii) An inner regular Borel measure is called *Radon measure*.

(iii) A measure μ is *outer regular* if for all $A \in \mathcal{B}$

$$\mu(A) = \inf\{\mu(U) : U \supset A,\, U \text{ open }\}$$

(iv) A measure μ is *regular* if it is inner and outer regular.

(v) A linear form $I : E \to \mathbb{R}$ is *positive*[3] if $I(x) \geq 0$ if $x \geq 0$.

On a Polish space every Borel measure is regular according to Ulam's theorem (see Theorem VIII.1.16 in Elstrodt 2005).

Theorem 16 (Riesz representation theorem for C_b)**.** Let E be a Polish space, $C_b(E)$ be the class of bounded continuous functions and $I : C_b(E) \to \mathbb{R}$ a positive linear form. Define the measure μ via

$$\mu_0(K) := \inf\{I(f) : f \in \mathcal{C}_b(E), f \geq \mathbb{1}_K\},\ K \text{ compact},$$
$$\mu(A) := \sup\{\mu_0(K) : K \subset A,\, K \text{ compact}\},\ A \in \mathcal{B}.$$

Then μ is a finite Radon measure and $C_b(E) \subset \mathcal{L}^1(\mu)$ and the following are equivalent:

(i) I can be represented by μ:

$$I(f) = \int_E f \,\mathrm{d}\mu, \quad \forall\, f \in \mathcal{C}_b(E). \tag{3.4}$$

(ii) $\mu(E) = I(\mathbb{1}_E)$

(iii) $\forall \varepsilon > 0\, \exists K$ compact s.t. $I(f) < \varepsilon \quad \forall f \in \mathcal{C}_b(E), 0 \leq f \leq 1, f|_K = 0.$

Moreover, if any of (i)-(iii) are fulfilled, μ is the only Radon measure on $\mathcal{B}$ to allow the representation in (3.4).[4]

We have seen that with any finite Borel measure μ we can associate a positive linear form I such that the representation in (3.4) holds. Conversely, the Riesz representation theorem tells us that on $C_b(E)$ for a positive linear form, under certain conditions, there exists a unique Borel measure μ such that it can represented as in (3.4).

[3] non-negative to be more precise (we follow the terminology in Elstrodt 2005)

[4] A more general formulation of this theorem for completely regular Hausdorff spaces and an additional fourth condition involving nets can be found in Elstrodt 2005

This provides the background and possible motivation for the definition of *weak convergence* of finite measures $\mu_n \to \mu$ as

$$\int_E f \,\mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_E f \,\mathrm{d}\mu, \quad \forall\, f \in \mathcal{C}_b(E).$$

If the measures are associated with a sequence of random variables (X_n) and X such that $\mu_n = \mathcal{L}(X_n)$ and $\mu = \mathcal{L}(X)$, then we can express the above convergence in terms of expectations:

$$\mathbb{E}[f(X_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(X)] \quad \forall\, f \in \mathcal{C}_b(E).$$

We write $X_n \xrightarrow[n\to\infty]{w} X$ and occasionally $X_n \xrightarrow[n\to\infty]{d} X$ if it is a sequence of random variables mapping in a finite-dimensional space (typically $\mathbb{R}^n$) instead of a function space. In the finite dimensional case, convergence in (finite-dimensional) distribution and weak convergence coincide. If E were a locally compact Hausdorff space, we could choose the continuous functions with compact support $\mathcal{C}_c(E)$ as the space of test functions. Then, there is a Riesz representation theorem that describes the one-to-one correspondence between positive linear forms on $\mathcal{C}_c$ and finite Borel measures.

Theorem 17 (Riesz representation for $\mathcal{C}_c$)**.** Let E be a a locally compact Hausdorff space with countable basis. Then, for every positive linear form $I : \mathcal{C}_c \to \mathbb{R}$ there exists a unique Borel measure μ such that

$$I(f) = \int_E f \,\mathrm{d}\mu, \quad \forall\, f \in \mathcal{C}_c(E).$$

This leads to the definition of *vague convergence*, i.e. a sequence of finite measures μ_n converges to μ if

$$\int_E f \,\mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_E f \,\mathrm{d}\mu, \quad \forall\, f \in \mathcal{C}_c(E) \tag{3.5}$$

holds.

Weak convergence and to some extent vague convergence are useful for the probabilistic setting. It is worth mentioning that formally the expression in (3.5) looks very similar to convergence of functionals, i.e. *continuous* linear forms, or in other words weak-$*$ convergence for the space of continuous functions. Indeed, a characterization of the dual space of $\mathcal{C}_c$ and $\mathcal{C}_0$ (the space of continuous functions vanishing at infinity) is motivated by the analysis of quantum mechanical Hamiltonians. Since there is no longer a restriction due to measurability issues, such a theory introduces the σ-algebra of Baire sets, which is smaller than the Borel σ-algebra and is associ-

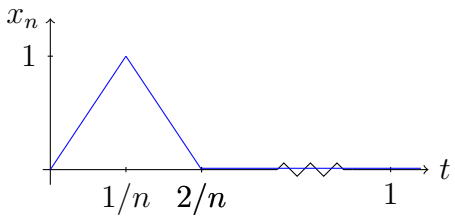

Figure 3.1: Shape of a function in the sequence $(x_n)_{n\in\mathbb{N}}$ of Example 3

ated with the Baire measure. This allows an identification of weak-$*$ convergence of functionals on $\mathcal{C}_c$ and $\mathcal{C}_0$ with vague convergence if E is a (locally) compact space. For details in this direction related to mathematical physics see Reed and Simon 1980.

3.1.2 Prokhorov's theorem

Returning now to stochastic processes with càdlàg paths, we can define following mode of convergence:

Definition 10 (Convergence in finite dimensions)**.** Let $(X_t^n)_{t\geq 0}$ be a sequence of stochastic processes. The sequence *convergences in finite dimensions* if its finite dimensional marginals converge, i.e. $X^n \xrightarrow[n\to\infty]{\text{f.d.}} X$ if

$$(X_{t_1}^n, X_{t_2}^n, \ldots, X_{t_N}^n) \xrightarrow[n\to\infty]{d} (X_{t_1}, X_{t_2}, \ldots, X_{t_N}), \quad \forall\, t_1, \ldots, t_N \geq 0,\ N \in \mathbb{N}.$$

However, convergence in finite dimensions does not imply convergence in the respective function space also called *functional convergence*. In other words, the path properties might not be preserved in the limit or the limit might not even exist.

Example 3 (see Example 11.6.1. in Whitt 2002)**.** Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of continuous functions which converges pointwise to a function $x = 0$. Such a sequence can be a linear interpolation between

$$x_n(0) = 0, \quad x_n\left(\frac{1}{n}\right) = 1 \quad \text{and} \quad x_n\left(\frac{2}{n}\right) = 0.$$

The distance is constant in the uniform norm $\|x_n - x\|_\infty = 1$. If we set for random variables X_n and X that $\mathbb{P}(X_n = x_n) = 1$ and $\mathbb{P}(X = x) = 1$, then the pointwise convergence $x_n \to x$ implies $X_n \xrightarrow{\text{f.d.}} X$, but $\mathbb{P}(\|X_n - X\|_\infty = 1) = 1$. See Figure 3.1.

Prokhorov's theorem uses the definition of tightness to connect the topology of the path space E with the weak topology in the space of measures $\mathcal{P}(E)$.

Definition 11. Let E be a metric space with Borel σ-algebra $\mathcal{B}$ and $\mathcal{P}(E)$ the set of probability measures on E.

(i) A finite Borel measure is *tight* if

$$\forall \varepsilon > 0 \; \exists K \subset E \text{ compact s.t. } \mu(K^c) < \varepsilon.$$

(ii) A set $M \subset \mathcal{P}(E)$ is tight[5] if

$$\forall \varepsilon > 0 \; \exists K \subset E \text{ compact s.t. } \mu(K^c) < \varepsilon \, \forall \mu \in M \tag{3.6}$$

(iii) A sequence $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(E)$ is tight if the set $M := \{\mu_n : n \in \mathbb{N}\}$ is tight.

We are now able to formulate Prokhorov's theorem which characterizes compact sets in $\mathcal{P}(E)$.

Theorem 18 (Prokhorov's theorem)**.** Let E be a Polish space. A set $M \in \mathcal{P}(E)$ is relatively compact iff it is bounded and tight.

For a proof see Section 5 in Billingsley 1999.

Method 1 (Compactness approach)**.** Given a sequence of stochastic processes (X^n), we can show that $(X^n_t)_{t\geq 0} \xrightarrow[n\to\infty]{w} (X_t)_{t\geq 0}$ as a functional limit by following two steps:

(i) The sequence of stochastic processes $(X^n)_{n\in\mathbb{N}}$ is tight, i.e. the sequence of laws $\mu_n := \mathcal{L}(X^n)$ is tight. Then Prokhorov's theorem implies that (μ_n) is relatively compact in $\mathcal{P}(E)$.

(ii) Prove that $\mu := \mathcal{L}(X)$ is the only possible limit point.

The above method is sometimes referred to as *compactness approach* to functional convergence. Point (ii) can be shown via convergence of the finite-dimensional marginals (on a dense subset of $\mathbb{R}_+$). However, one often does not have access to their analytic form. For example, Lévy processes are defined via their characteristic function and finite-dimensional convergence is usually shown via convergence of the corresponding characteristic functions. In other words, Lévy's continuity theorem is invoked:

Theorem 19 (Lévy's continuity theorem)**.** If $(\varphi_n)_{n\in\mathbb{N}}$ is a sequence of characteristic functions and there exists a function $\psi : \mathbb{R}^d \to \mathbb{C}$ such that, for all $u \in \mathbb{R}^d$, $\varphi_n(u) \to \psi(u)$ as $n \to \infty$ and ψ is continuous at 0, then ψ is the characteristic function of a probability distribution (see Theorem 1.1.15 in Applebaum 2009 and also compare with Theorem 3.3.6. in Durrett 2010).

[5]Some refer to (3.6) as uniform tightness since the choice of the set K is uniform throughout the set M.

More generally, conditions for functional convergence of semimartingales can also be expressed in terms of the characteristic triplet of the Lévy-Khinchine formula (see Jacod and Shiryaev 2003 for an extensive treatment).

3.1.3 Compactness in $\mathcal{C}$

A closer look at (3.6) reveals that the choice of the metric d on the space E determines the compact sets from which K can be chosen and affects the tightness condition. Therefore, it is useful to have a characterization of relatively compact sets on metric spaces. In the case of continuous functions, relatively compact sets are characterized by the Arzelà-Ascoli theorem, which we will state here along with the necessary notation and terminology. One possible formulation of the Arzelà-Ascoli theorem is for a subset $A \subset \mathcal{C}(S, \mathbb{R})$, where S is a compact space. Then A is relatively compact iff it is bounded and equicontinuous, a property which can be defined as follows:

Definition 12 (Equicontinuity)**.** Let (S, d) be a metric space. A subset $A \subset \mathcal{C}(S, \mathbb{R})$ is *equicontinuous* if

$$\forall \varepsilon > 0 \; \exists \delta > 0 \text{ s.t. } d(s,t) < \delta \Rightarrow |x(s) - x(t)| < \varepsilon \; \forall x \in A.$$

This especially holds true for $S = [0,1]$, $d(s,t) = |s-t|$ and the space $\mathcal{C}([0,1])$ equipped with the uniform topology induced by the supremum norm

$$\|x\|_\infty := \sup_{t \in [0,1]} |x(t)|.$$

Alternatively to the ε-δ-formulation, the equicontinuity condition can be expressed by using the *modulus of continuity*. As defined on p. 80 in Billingsley 1999:

$$w(x, \theta) = \sup\{|x(s) - x(t)| : |s - t| \leq \theta, \, \theta \in [0,1]\}.$$

It holds that

$$x \in \mathcal{C}([0,1], \mathbb{R}) \Leftrightarrow \lim_{\theta \to 0} w(x, \theta) = 0,$$

which is equivalent to requiring x to be uniformly continuous. A subset A is equicontinuous if

$$\lim_{\theta \to 0} \sup_{x \in A} w(x, \theta) = 0.$$

When we consider $S = \mathbb{R}_+$ instead of a bounded interval, this space is locally compact, but no longer compact. Jacod and Shiryaev 2003 propose the following

adjustment to the modulus of continuity: let $\theta > 0$ and $N \in \mathbb{N}$ and define

$$w_N(x, \theta) := \sup \left\{ \sup_{\tau \in [t, t+\theta]} x(\tau) : 0 \leq t \leq t + \theta \leq N \right\}.$$

Then, it holds that

$$x \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}) \Leftrightarrow \lim_{\theta \to 0} w_N(x, \theta) = 0 \quad \forall N \in \mathbb{N},$$

which is equivalent to demanding that x is locally uniformly continuous. Equicontinuity of a subset A is given by

$$\lim_{\theta \to 0} \sup_{x \in A} w_N(x, \theta) = 0 \quad \forall N \in \mathbb{N}.$$

A more general version of the Arzelà-Ascoli theorem is needed to cover the cases of non-compact spaces. Such a version can be found as Theorem 4.7.1. on p. 290 in Munkres 2000, which we will state here in a slightly modified form:

Theorem 20 (Arzelà-Ascoli for locally compact spaces)**.** Let S be a locally compact Hausdorff space, (F, d) be a complete metric space., $A \subset \mathcal{C}(S, F)$. Then, A is relatively compact in $\mathcal{C}(S, F)$ w.r.t. the topology of compact convergence iff A is equicontinuous and pointwise totally bounded.

Remark 7.

(i) Recall that in a locally compact space it holds that local uniform convergence is equivalent to compact convergence.

(ii) In a finite-dimensional metric space, e.g. $\mathbb{R}_+$, total boundedness is equivalent to usual boundedness.

We are able to apply the general theorem to the relevant spaces of continuous functions.

Corollary 2 (Arzelà-Ascoli theorem for $\mathcal{C}([0, 1], \mathbb{R})$)**.** A subset $A \subset \mathcal{C}([0, 1], \mathbb{R})$ is relatively compact for the uniform topology iff

a) $\sup_{x \in A} |x(0)| < \infty$

b) $\lim_{\theta \to 0} \sup_{x \in A} w(x, \theta) = 0$

(see Theorem 7.2. in Billingsley 1999).

Corollary 3 (Arzelà-Ascoli theorem for $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$)**.** A subset $A \subset \mathcal{C}([0,1], \mathbb{R})$ is relatively compact for the uniform topology iff

a) $\sup_{x \in A} |x(0)| < \infty$

b) $\lim_{\theta \to 0} \sup_{x \in A} w_N(x, \theta) = 0 \quad \forall N \in \mathbb{N}$

(see Theorem VI.1.5 in Jacod and Shiryaev 2003).

Since it is quite difficult to prove tightness directly via its definition the following characterization can be used which specifies tightness for the space of continuous functions.

Theorem 21 (Tightness for $\mathcal{C}([0,1], \mathbb{R})$)**.** A sequence X_n with paths in $\mathcal{C}([0,1], \mathbb{R})$ is $\mathcal{C}$-tight iff

(i) $\forall \varepsilon > 0 \; \exists c > 0, n_0 \in \mathbb{N}$ s.t.

$$n \geq n_0 \Rightarrow \mathbb{P}\left(\sup_{t \in [0,1]} |X_t^n| > c \right) < \varepsilon,$$

(ii) $\forall \eta > 0, \varepsilon > 0 \; \exists n_0 \in \mathbb{N}, \theta > 0$ s.t.

$$n \geq n_0 \Rightarrow \mathbb{P}(w'(X^n, \theta) > \eta) < \varepsilon$$

,

(see Theorem 7.3 in Billingsley 1999).

Theorem 22 (Tightness for $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$)**.** A sequence (X_n) with paths in $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$ is $\mathcal{C}$-tight iff

(i) $\forall N \in \mathbb{N}, \varepsilon > 0 \; \exists n_0 \in \mathbb{N}, c > 0$ s.t.

$$n \geq n_0 \Rightarrow \mathbb{P}\left(\sup_{t \leq N} |X_t^n| > c \right) \leq \varepsilon,$$

(ii) $\forall N \in \mathbb{N}, \varepsilon > 0, \eta > 0 \; \exists n_0 \in \mathbb{N}, \theta > 0$ s.t.

$$n \geq n_0 \Rightarrow \mathbb{P}\left(w_N(X^n, \theta) > \eta \right) \leq \varepsilon$$

(see Proposition VI.3.26 in Jacod and Shiryaev 2003).

Finally, it can be shown that the Kolmogorov σ-field coincides with the Borel σ-field on $\mathcal{C}$ (see Exercise 4.2 on p. 60 with solutions on p. 119 in Karatzas and Shreve

1988).

In summary, the choice of metric needs to address the following aspects:

(i) (E, d) is a Polish space.

(ii) The metric induces a topology on E and determines the weak topology on $\mathcal{P}(E)$ via Prokhorov's theorem. We are able to characterize relatively compact sets via the modulus of continuity.

(iii) The σ-field generated by the open sets in (E, d) coincides with the σ-field generated by the cylinder sets associated with the stochastic processes.

3.1.4 A Skorokhod topology: J_1 and compactness on $\mathcal{D}$

While $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$ equipped with the local uniform topology fulfills all requirements (i)-(iii), it is not clear how to choose the topology for $\mathcal{D}$. This larger space of càdlàg functions is non-separable under the uniform topology as can be seen from the following example.

Example 4. Consider the set $\{x_s\}_{s\in[0,1/2]}$ of functions in $\mathcal{D}([0,1], \mathbb{R})$ of the form $x_s(t) := \mathbb{1}_{[s,1]}(t)$ for $s \in [0, 1/2]$. The set contains uncountably many elements and under the uniform metric $\|x_{s_1} - x_{s_2}\| = 1$ for $s_1 \neq s_2$. If $\mathcal{D}$ were separable, $\{x_s\}_{s\in[0,1/2]}$ would contain a countable dense subset, which is not possible as all elements have constant distance.

Skorokhod 1956 proposed four different topologies for $\mathcal{D}$, namely J_1, J_2, M_1 and M_2 and Jakubowski 1997 introduced the non-metrizable S topology on $\mathcal{D}$. We give the definition of the J_1 topology and the corresponding modifications in the framework of separability, completeness, measurability and compactness. After that, the M_1 topology is briefly introduced together with the theorems which are relevant for our purposes later on in the chapter.
The requirements for the J_1 topology should be weaker than uniform convergence that accounts for the jumps of càdlàg paths.

Example 5. Consider the sequence $(x_n)_{n\in\mathbb{N}} \subset \mathcal{D}([0,1], \mathbb{R})$ of the form

$$x_n(t) := \left(1 + \frac{1}{n}\right) \mathbb{1}_{\left[\frac{1}{2}+\frac{1}{n},1\right]}(t), \quad n \geq 3$$

and the potential limit candidate

$$x(t) := \mathbb{1}_{\left[\frac{1}{2},1\right]}(t)$$

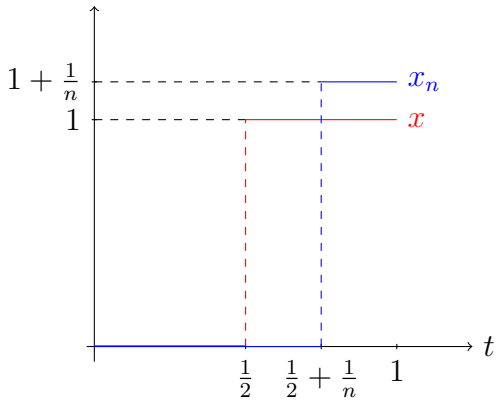

Figure 3.2: Shape of functions in Example 5

Similar to the examples seen before, $x_n \xrightarrow[n\to\infty]{} x$ pointwise but not uniformly as $\|x_n - x\|_\infty = 1 \ \forall n$. See Figure 3.2.

Intuitively, we want to match the jumps in Example 5 by using a small time-change. To this end, define the set of continuous time-changes:

$$\Lambda := \{\lambda : I \to I : \lambda \in \mathcal{C}(I) \text{ strictly increasing}\},$$

where $I = [0,1]$ or $I = \mathbb{R}_+$ depending on the domain of the functions of the space $E = \mathcal{D}$. The identity map is denoted by $\mathrm{id} : I \to I$. Then a J_1 metric is defined as

$$d_{J_1}(x_1, x_2) := \inf_{\lambda\in\Lambda} \{\|x_1 \circ \lambda - x_2\|_\infty \vee \|\lambda - \mathrm{id}\|_\infty\}.$$

Example 5 (Continued)**.** We can choose a $\lambda \in \Lambda$ such that

$$d_{J_1}(x_n, x) \xrightarrow[n\to\infty]{} 0.$$

Consider the functions

$$\lambda_n(t) = \begin{cases} \left(1+\frac{2}{n}\right)t, & t \in \left[0, \frac{1}{2}\right), \\ \left(1-\frac{2}{n}\right)t + \frac{2}{n}, & t \in \left[\frac{1}{2}, 1\right]. \end{cases}$$

It can be checked that $\lim_{t\to 1/2^-} \lambda_n(t) = 1/2 + 1/n = \lambda_n(1/2)$ and λ_n is strictly increasing for $n \geq 3$. Moreover

$$(x_n \circ \lambda_n)(t) = \left(1+\frac{1}{n}\right) \mathbb{1}_{\left[\frac{1}{2}+\frac{1}{n}, 1\right]}(\lambda_n(t)) = \left(1+\frac{1}{n}\right) \mathbb{1}_{\left[\frac{1}{2}, 1\right]}(t),$$

i.e. the jumps of $(x_n \circ \lambda_n)$ are now matched with the jump of x at $t = 1/2$. We

calculate

$$\|\lambda - \mathrm{id}\| = \frac{1}{n} \text{ and } \|x_n \circ \lambda_n - x\| = \frac{1}{n},$$

which implies

$$d_{J_1}(x_n, x) = \frac{1}{n} \xrightarrow[n\to\infty]{} 0.$$

Suitable modifications of the moduli of continuity for the spaces $\mathcal{D}([0,1],\mathbb{R})$ and $\mathcal{D}(\mathbb{R}_+,\mathbb{R})$, respectively, allow analogous formulations of the Arzelà-Ascoli theorem for those spaces. Let $\theta > 0$ and $N \in \mathbb{N}$. We define the two moduli and continuity

$$w'(x,\theta) := \inf\left\{\max_{i\le r} \sup_{s,t\in[t_{i-1},t_i)} |x(s)-x(t)| : 0 = t_0 < \ldots < t_n = 1, \inf_{i\le r}(t_i - t_{i-1}) \ge \theta\right\}$$

$$w'_N(x,\theta) := \inf\left\{\max_{i\le r} \sup_{s,t\in[t_{i-1},t_i)} |x(s)-x(t)| : 0 = t_0 < \ldots < t_n = N, \inf_{i< r}(t_i - t_{i-1}) \ge \theta\right\}.$$

Similar to the case of $E = \mathcal{C}$ it holds that

$$x \in \mathcal{D}([0,1],\mathbb{R}) \Leftrightarrow \sup_{s\in[0,1]} |x(s)| < \infty \text{ and } \lim_{\theta\to 0} w'(x,\theta) = 0$$

$$x \in \mathcal{D}(\mathbb{R}_+,\mathbb{R}) \Leftrightarrow \sup_{s\le N} |x(s)| < \infty \text{ and } \lim_{\theta\to 0} w'_N(x,\theta) = 0 \quad \forall N \in \mathbb{N}.$$

(see Lemma VI.1.11 in Jacod and Shiryaev 2003). The versions of the Arzelà-Ascoli theorem can be stated as follows.

Corollary 4 (Relative compactness in $\mathcal{D}([0,1],\mathbb{R})$)**.** A subset $A \in \mathcal{D}([0,1])$ is relatively compact iff

a) $\sup_{x\in A} \sup_{s\in[0,1]} |x(s)| < \infty$

b) $\lim_{\theta\to 0} \sup_{x\in A} w'(x,\theta) = 0$

(see Theorem 12.3 in Billingsley 1999).

Corollary 5 (Relative compactness in $\mathcal{D}(\mathbb{R}_+,\mathbb{R})$)**.** A subset $A \in \mathcal{D}(\mathbb{R}_+)$ is relatively compact iff

a) $\sup_{x\in A} \sup_{s\le N} |x(s)| < \infty$

b) $\lim_{\theta\to 0} \sup_{x\in A} w'_N(x,\theta) = 0 \quad \forall N \in \mathbb{N}$

(see Theorem VI.1.14 b) in Jacod and Shiryaev 2003).

Remark 8. Indeed, it can be shown that $(\mathcal{D}, d_{J_1})$ is a separable metric space. However, $\mathcal{D}$ is not complete under d_{J_1}. The metric d_{J_1} can be replaced by an equivalent metric under which $\mathcal{D}$ is complete. Concerning measurability, one can prove that also in this case the Borel σ-field coincides with the σ-field generated by the cylinder sets. For detailed proofs we refer to Billingsley 1999 and Theorem VI.1.14 in Jacod and Shiryaev 2003.

A characterization of tightness in $\mathcal{D}(\mathbb{R}_+, \mathbb{R})$ looks very much like Theorem 21 and 22 where the modulus of continuity is replaced by w'_N.

Theorem 23 (Tightness for $\mathcal{D}([0,1], \mathbb{R})$)**.** A sequence (X_n) with paths in $\mathcal{D}([0,1], \mathbb{R})$ is tight iff

(i) $\forall \varepsilon > 0\ \exists c > 0, n_0 \in \mathbb{N}$ s.t.

$$n \geq n_0 \Rightarrow \mathbb{P}\left(\sup_{t\in[0,1]} |X_t^n| > c\right) < \varepsilon,$$

(ii) $\forall \eta > 0, \varepsilon > 0\ \exists \theta > 0, n_0 \in \mathbb{N}$ s.t.

$$n \geq n \Rightarrow \mathbb{P}(w'(X_n, \theta) > \eta) < \varepsilon$$

(see Theorem 13.2 Billingsley 1999).

Theorem 24 (Tightness for $\mathcal{D}(\mathbb{R}_+, \mathbb{R})$)**.** A sequence (X_n) with paths in $\mathcal{D}(\mathbb{R}_+, \mathbb{R})$ is tight iff

(i) $\forall N \in \mathbb{N}, \varepsilon > 0\ \exists n_0 \in \mathbb{N}, c > 0$ s.t.

$$n \geq n_0 \Rightarrow \mathbb{P}\left(\sup_{t\leq N} |X_t^n| > c\right) \leq \varepsilon,$$

(ii) $\forall N \in \mathbb{N}, \varepsilon > 0, \eta > 0\ \exists n_0 \in \mathbb{N}, \theta > 0$ s.t.

$$n \geq n_0 \Rightarrow \mathbb{P}\left(w'_N(X^n, \theta) > \eta\right) \leq \varepsilon$$

(see Proposition VI.3.21 in Jacod and Shiryaev 2003).

The above theorems allow us to apply the compactness approach of Method 1 also for the space $\mathcal{D}$. If a sequence in $\mathcal{D}$ converges to a function in $\mathcal{C}$ in the J_1 topology, the convergence improves to (local) uniform convergence.
One of the rare cases in which tightness can be shown generally does indeed apply to some situations that we will encounter in this chapter.

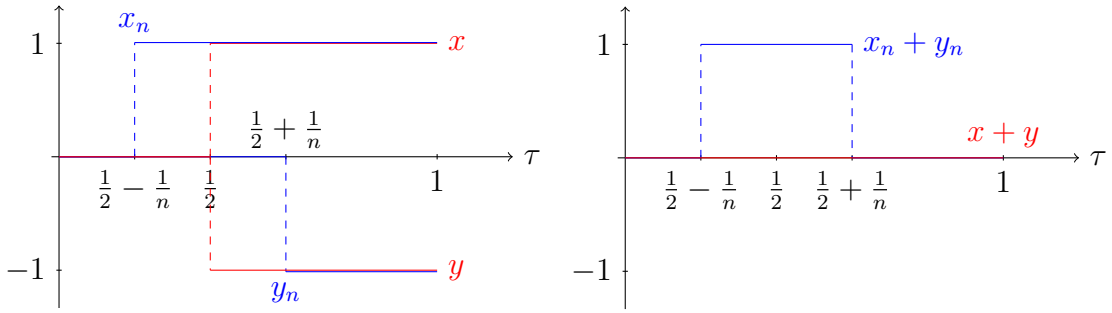

Figure 3.3: Illustration of the step functions of Example 6.

Theorem 25 (Tightness for increasing processes)**.** Let X_n, X be increasing processes such that

(a) either X is continuous

(b) or all X_n, X are point processes.

Then, if $X_n \xrightarrow[n\to\infty]{\text{f.d.}} X$ for some dense subset $D \subset \mathbb{R}_+$, we also have $X_n \xrightarrow[n\to\infty]{J_1} X$ (see Theorem VI.3.37 in Jacod and Shiryaev 2003).

3.1.5 Continuity of functions on $\mathcal{D} \times \mathcal{D}$: M_1 topology and continuous mapping approach

Note that $\mathcal{D}$ with the J_1 topology is not a topological vector space, i.e. addition is not continuous under the J_1 topology as can be seen in the following example:

Example 6. Define the following two sequences in the space $\mathcal{D}([0,1],\mathbb{R})$:

$$x_n := \mathbb{1}_{\left[\frac{1}{2}-\frac{1}{n},1\right]} \text{ and } y_n := -\mathbb{1}_{\left[\frac{1}{2}+\frac{1}{n},1\right]}.$$

Although the sequences converge in the J_1 topology individually to $x = \mathbb{1}_{\left[\frac{1}{2},1\right]}$ and $y = -\mathbb{1}_{\left[\frac{1}{2},1\right]}$ respectively, the sum $x_n + y_n = \mathbb{1}_{\left[\frac{1}{2}-\frac{1}{n},\frac{1}{2}+\frac{1}{n}\right]}$ does not converge to 0 (see Figure 3.3).

Another useful function on $\mathcal{D} \times \mathcal{D}$ is composition, which is only continuous w.r.t. J_1 on certain subsets of $\mathcal{D} \times \mathcal{D}$. To make this more precise we follow pp. 430-431 in Whitt 2002 and define the following subsets of $E = \mathcal{C}$ and $E = \mathcal{D}$ respectively:

$$\begin{aligned}
E_+ &:= \{x \in E : x(0) \geq 0\} \\
E_\uparrow &:= \{x \in E_+ : x \text{ non-decreasing}\} \\
E_{\uparrow\uparrow} &:= \{x \in E_+ : x \text{ strictly increasing}\} \\
E_m &:= \{x \in E_+ : x \text{ monotone}\}
\end{aligned}$$

Consider the composition map:

$$\begin{aligned} \circ : \mathcal{D} \times \mathcal{D}_{\uparrow} &\to \mathcal{D} \\ (x, y) &\mapsto x \circ y. \end{aligned}$$

We see that $y \in \mathcal{D}_{\uparrow}$ has the role of a time-change applied to $x \in \mathcal{D}$. Therefore, it makes sense for y to be non-negative and non-decreasing. We can now state the theorems for continuity of the composition under the J_1 topology.

Theorem 26. The composition map is measurable and continuous at $(x, y) \in \mathcal{C} \times \mathcal{C}_{\uparrow}$.

Theorem 27 (J_1-continuity of the composition)**.** The composition is continuous at

$$(x, y) \in (\mathcal{D} \times \mathcal{C}_{\uparrow\uparrow}) \cup (\mathcal{C} \times \mathcal{D}_{\uparrow}),$$

using the J_1-topology throughout.

The so-called *continuous mapping approach* makes use of theorems of this kind to derive limit theorems for time-changed processes. It can be anticipated from the previous chapter that we want to apply the above theorems to derive limit theorems for processes that are time-changed by the inverse α-stable subordinator $(Y_\alpha(t))_{t\geq 0}$. Unfortunately, the paths of $(Y_\alpha(t))_{t\geq 0}$ are only non-decreasing, but not strictly increasing. For this situation, there exists another topology on the Skorokhod space which is weaker than the J_1-topology and allows continuity of the composition on the relevant subspace of $\mathcal{D} \times \mathcal{D}_+$. To understand intuitively how the convergence notion of J_1 is weakened, consider following example.

Example 7. Consider the sequence of functions $(x_n)_{n\in\mathbb{N}}$ and a limit candidate x defined as

$$x_n := n\left(t - \frac{1}{2} + \frac{1}{n}\right) \mathbb{1}_{\left[\frac{1}{2}-\frac{1}{n}, \frac{1}{2}\right)}(t) + \mathbb{1}_{\left[\frac{1}{2}, 1\right]}(t), \quad x := \mathbb{1}_{\left[\frac{1}{2}, 1\right]}(t).$$

We have $\|x_n - x\|_\infty = 1$ and $d_{J_1}(x_n, x) = 1$ as the jump at $t = \frac{1}{2}$ cannot be matched (see Figure 3.4).

The examples show that jumps need to "match" under the J_1-convergence. However, there are many situations where this is not the case. A topology for "non-matched" jumps considers the graph of the function in $\mathbb{R}^2$:

For a function $x \in \mathcal{D}(I, \mathbb{R})$, $I = [0, 1]$ or $I = \mathbb{R}_+$, we can define the *completed graph* as

$$\Gamma_x := \{(t, z) \in I \times \mathbb{R} : z = \alpha x(t-) + (1 - \alpha)x(t) \text{ for some } \alpha \in [0, 1]\}.$$

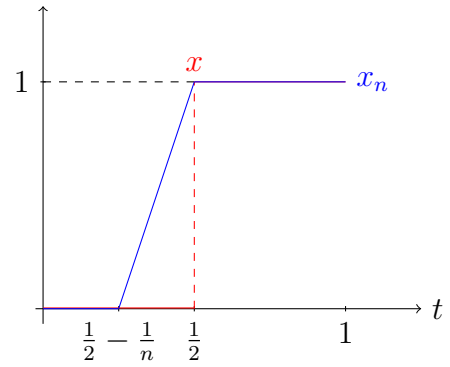

Figure 3.4: Sequence of continuous functions $(x_n)_{n\in\mathbb{N}}$ converging to a step function x (see Example 7).

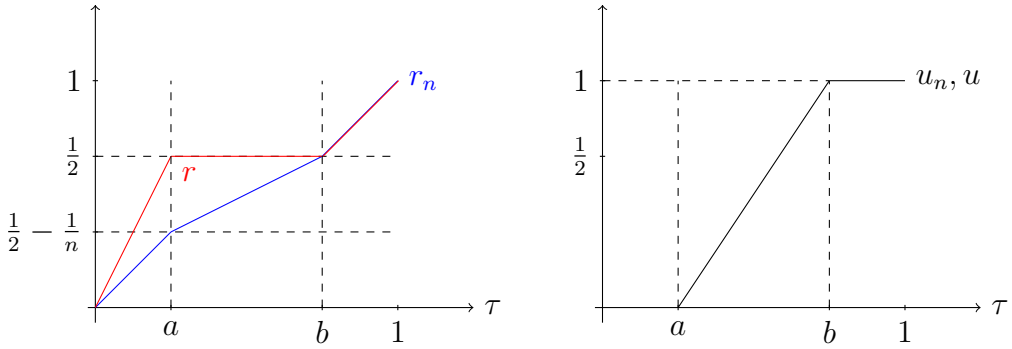

Figure 3.5: Illustration of the parametric representation of the graph of x in Example 7. The points a, b can be arbitrarily chosen from $(0, 1)$.

with an order on Γ_x: $(t_1, z_1) \leq (t_2, z_2)$ if either

(i) $t_1 < t_2$ or

(ii) $t_1 = t_2$ and $|x(t_1-) - z_1| \leq |x(t_2-) - z_2|$.

A parametric representation of Γ_x is a map

$$\begin{aligned} P : I &\to \Gamma_x \\ t &\mapsto (r(t), u(t)) \end{aligned}$$

which is non-decreasing w.r.t. the above order. Let Π_x denote the set of parametric representations corresponding to an element $x \in \mathcal{D}$. Then, the M_1-metric is defined as

$$d_{M_1}(x_1, x_2) := \inf_{\substack{(r_j,u_j)\in\Pi_{x_j} \\ j=1,2}} \{\|r_1 - r_2\| \vee \|u_1 - u_2\|\}.$$

Example 7 (Continued)**.** Figure 3.5 depicts possible parametric representations of Γ_x such that $d_{M_1}(x_n, x) \to 0$.

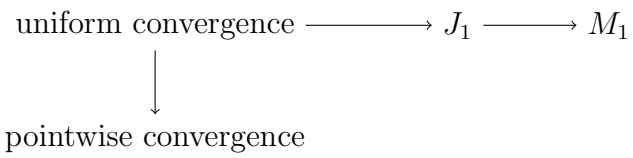

Figure 3.6: Overview of relations between the topologies. The relation $a \to b$ is to be read as a implies b.

For a proof of separability, completeness, compactness and measurability results of $\mathcal{D}$ under the M_1-topology, we refer to Chapter 12 in Whitt 2002. Figure 3.6 shows the relative strength of the modes of convergence discussed in this section. We omit the tightness criteria for the M_1-topology, since these will not be used in the scope of this work. We will arrive at M_1-convergence by using the following theorems of the continuous mapping approach.

Theorem 28 (M_1-continuity of the composition)**.** If $(x_n, y_n) \xrightarrow[n\to\infty]{} (x, y)$ in $\mathcal{D} \times \mathcal{D}_\uparrow$ and

$$(x, y) \in (\mathcal{D} \times \mathcal{C}_{\uparrow\uparrow}) \cup (\mathcal{C}_m \times \mathcal{D}_\uparrow),$$

then $x_n \circ y_n \xrightarrow[n\to\infty]{} x \circ y$ in $\mathcal{D}$ using the M_1-topology throughout.

The above theorem can be seen as an analogue to Theorem 27. It can be seen that we still need one of the processes to be continuous in the limit. If this is not the case, there is a different theorem which requires conditions on the jumping times of the processes.

Theorem 29 (M_1-continuity of the composition, discontinuity points)**.** Suppose that $(x_n, y_n) \xrightarrow[n\to\infty]{} (x, y)$ in $\mathcal{D} \times \mathcal{D}_\uparrow$. Let $\mathrm{Disc}(z)$ denote the set of discontinuity points of an element $z \in \mathcal{D}$. If

(i) y is continuous and strictly increasing at points $t \in \mathrm{Disc}(x)$ and

(ii) x is monotone on $[y(t-), y(t)]$ and $y(t-), y(t) \notin \mathrm{Disc}(x)$ whenever $t \in \mathrm{Disc}(y)$,

then $x_n \circ y_n \xrightarrow[n\to\infty]{} x \circ y$ in $\mathcal{D}$ using the M_1-topology throughout.

3.2 A martingale approach to limit theorems for the fractional Poisson process

In the previous section, alongside an introduction to finite-dimensional convergence and the various modes of functional convergence, we have already discussed the theorems underlying the the compactness and the continuous mapping approach to deriving limit theorems. Before applying these to the fractional Poisson process, we present a third method using martingale theory. Essentially, under suitable

conditions, the convergence of compensators can imply the convergence of the corresponding processes. Moreover, the compensator of a point process can be used as the centralizing and norming quantity of a central limit theorem.

3.2.1 The fractional Poisson process as a Cox process

In order to apply martingale theory to the FNPP, we need to identify its compensator. To this end, it is useful to first verify that the FNPP belongs to the class of Cox processes. Cox processes go back to Cox 1955 who proposed to replace the deterministic intensity of a Poisson process by a random one. In this section, we discuss the connection between FNPP and Cox processes.

Definition 13. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(N(t))_{t\geq 0}$ be a point process adapted to a filtration $(\mathcal{F}_t^N)_{t\geq 0}$. $(N(t))_{t\geq 0}$ is a *Cox process* if there exists a right-continuous, increasing process $(A(t))_{t\geq 0}$ such that, conditional on the filtration $(\mathcal{F}_t)_{t\geq 0}$, where

$$\mathcal{F}_t := \mathcal{F}_0 \vee \mathcal{F}_t^N, \quad \mathcal{F}_0 = \sigma(A(t), t \geq 0),$$

then $(N(t))_{t\geq 0}$ is a Poisson process with intensity $\mathrm{d}A(t)$.
In particular, we have $\mathbb{E}[N(t)|\mathcal{F}_t] = A(t)$ and

$$\mathbb{P}(N(t) = k|\mathcal{F}_t) = \mathrm{e}^{-A(t)}\frac{A(t)^k}{k!}, \qquad k = 0, 1, 2, \ldots.$$

Remark 9.

1. A Cox process N is said to be *directed* by A, if their relation is as in the above definition and A is called the *directing process* of N. Cox processes are also called $(\mathcal{F}_t)_{t\geq 0}$-Cox process, doubly stochastic processes, conditional Poisson processes or $(\mathcal{F}_t)_{t\geq 0}$-conditional Poisson process.

2. Definitions vary across the literature. The above definition can be compared to essentially equivalent definitions: in Brémaud 1981, 6.12 on p. 126 in Jacod and Shiryaev 2003, Definition 6.2.I on p. 169 in Daley and Vere-Jones 2008, where $\mathcal{X} = \mathbb{R}^+$ and Definition 6.6.2 on p.193 in Bielecki and Rutkowski 2002.

The FHPP as a Cox process It can be verified that the FHPP belongs to the class of Cox processes (or doubly stochastic Poisson processes or mixed Poisson processes). This can be done by using a result in Yannaros 1994 (see also Theorem 1 of Section 2.2 in Grandell 1976) which we will state here for the readers' convenience:

Lemma 1. An ordinary renewal process whose interarrival distribution function F_J satisfies

$$F_J(t) = 1 - \int_0^\infty \mathrm{e}^{-tx} \mathrm{d}V(x), \tag{3.7}$$

where V is a proper distribution function with $V(0) = 0$ is a Cox process.

The proof of this result uses a lemma due to Kingman 1964, which is formulated for the Laplace transform of F_J:

Lemma 2. An ordinary renewal process with interarrival distribution function F_J is a Cox process if and only if the Laplace transform $\hat{F}_J$ of F_J satisfies

$$\hat{F}_J(s) = \frac{1}{1 - \ln(\hat{G}(s))}, \tag{3.8}$$

where $\hat{G}$ is the Laplace transform of an infinitely divisible distribution function G.

Both lemmata can be used to check whether a renewal process also belongs to the class of Cox processes. Especially, Lemma 2 gives a full characterization of renewal Cox processes via the Laplace transform of the waiting time distribution. However, the theorem does not give any insight about the underlying filtration setting. This will become more evident from the following discussion concerning the general case of the FNPP. In the case of the FHPP the conditions of both lemmata can be verified. To this end, as presented in the the previous chapter, recall that the interarrival times J of the FHPP can be expressed by the one-parameter Mittag-Leffler function (we assume $\lambda = 1$ in this paragraph):

$$F_J(t) = 1 - E_\alpha(-t^\alpha).$$

Moreover, it can be found in Mainardi and Gorenflo 2000 that

$$\int_0^\infty \mathrm{e}^{-rt} K_\alpha(r) \mathrm{d}r = E_\alpha(-t^\alpha), \quad \text{where } K_\alpha(r) = \frac{1}{\pi} \frac{r^{\alpha-1} \sin(\alpha\pi)}{r^{2\alpha} + 2r^\alpha \cos(\alpha\pi) + 1}.$$

For $0 < \alpha < 1$ the function $K_\alpha(r)$ is positive and qualifies as a probability density as $\int_0^\infty K_\alpha(r)\mathrm{d}r = 1$. Therefore, the function $V(x) := \int_0^x K_\alpha(r)\mathrm{d}r$ fulfills the conditions of Lemma 1.

The FNPP as a Cox process In the non-homogeneous case, we cannot apply the theorems which characterize Cox renewal processes as the FNPP cannot be represented as a classic renewal process. Therefore, we need to resort to Definition 13 for verification. It can be shown that the FNPP is a Cox process under a suitably constructed filtration. We will follow the construction of doubly stochastic processes

given in Section 6.6 in Bielecki and Rutkowski 2002. Let $(\mathcal{F}_t^{N_\alpha})_{t\geq 0}$ be the natural filtration of the FNPP $(N_\alpha(t))_{t\geq 0}$

$$\mathcal{F}_t^{N_\alpha} := \sigma(\{N_\alpha(s) : s \leq t\}).$$

We assume the information on the inverse α-stable subordinator to be available at time $t = 0$, i.e.

$$\mathcal{F}_0 := \sigma(\{Y_\alpha(s), s \geq 0\}).$$

We refer to this choice of initial σ-algebra as *non-trivial initial history* as opposed to the case of *trivial initial history*, which is $\mathcal{F}_0 = \{\varnothing, \Omega\}$.
The overall filtration $(\mathcal{F}_t)_{t\geq 0}$ is then given by

$$\mathcal{F}_t := \mathcal{F}_0 \vee \mathcal{F}_t^{N_\alpha}, \tag{3.9}$$

which is sometimes referred to as *intrinsic history*. If we choose a trivial initial history, the intrinsic history will coincide with the natural filtration of the FNPP.

Proposition 30. Let the FNPP be adapted to the filtration $(\mathcal{F}_t)$ as in (3.9) with non-trivial initial history $\mathcal{F}_0 := \sigma(\{Y_\alpha(t), t \geq 0\})$. Then the FNPP is a $(\mathcal{F}_t)$-Cox process directed by $(\Lambda(Y_\alpha(t)))_{t\geq 0}$.

Proof. This follows from Proposition 6.6.7. on p. 195 in Bielecki and Rutkowski 2002. We give a similar proof: As $(Y_\alpha(t))_{t\geq 0}$ is $\mathcal{F}_0$-measurable we have for $s \leq t$

$$\begin{aligned}
&\mathbb{E}[\exp\{\mathrm{i}u(N_\alpha(t) - N_\alpha(s))\}|\mathcal{F}_s] \\
&\quad = \mathbb{E}\left[\exp\{\mathrm{i}u(N_\alpha(t) - N_\alpha(s))\}|\mathcal{F}_0 \vee \mathcal{F}_s^{N_\alpha}\right] \\
&\quad = \mathbb{E}\left[\exp\{\mathrm{i}u(N_1(\Lambda(Y_\alpha(t))) - N_1(\Lambda(Y_\alpha(s))))\}|\mathcal{F}_0 \vee \mathcal{F}_{\Lambda(Y_\alpha(s))}^{N_1}\right] \qquad (3.10) \\
&\quad = \mathbb{E}\left[\exp\{\mathrm{i}u(N_1(\Lambda(Y_\alpha(t))) - N_1(\Lambda(Y_\alpha(s))))\}|\mathcal{F}_0\right] \qquad (3.11) \\
&\quad = \exp[\Lambda(Y_\alpha(s), Y_\alpha(t))(e^{\mathrm{i}u} - 1)],
\end{aligned}$$

where in (3.10) we used the time-change theorem (see for example Theorem 7.4.I. p. 258 in Daley and Vere-Jones 2003) and in (3.11) the fact that the standard Poisson process has independent increments. This means, conditional on $(\mathcal{F}_t)_{t\geq 0}$, $(N_\alpha(t))$ has independent increments and

$$(N_\alpha(t) - N_\alpha(s))|\mathcal{F}_s \sim \mathrm{Poi}(\Lambda(Y_\alpha(s), Y_\alpha(t))) \stackrel{d}{=} \mathrm{Poi}(\Lambda(Y_\alpha(t)) - \Lambda(Y_\alpha(s))).$$

Thus, $(N(Y_\alpha(t)))$ is a Cox process directed by $\Lambda(Y_\alpha(t))$ by definition. □

3.2.2 The FNPP and its compensator

The identification of the FNPP as a Cox process in the previous section allows us to determine the compensator of the FNPP. In fact, the compensator of a Cox process coincides with its directing process. From Lemma 6.6.3. p.194 in Bielecki and Rutkowski 2002 we have the result

Proposition 31. Let the FNPP be adapted to the filtration $(\mathcal{F}_t)$ as in (3.9) with non-trivial initial history $\mathcal{F}_0 := \sigma(\{Y_\alpha(t), t \geq 0\})$. Assume $\mathbb{E}[\Lambda(Y_\alpha(t))] < \infty \; \forall t > 0$. Then the FNPP has $\mathcal{F}_t$-compensator $(A(t))_{t\geq 0}$, where $A(t) := \Lambda(Y_\alpha(t))$, i.e. the stochastic process $(M(t))_{t\geq 0}$ defined by $M(t) := N(Y_\alpha(t)) - \Lambda(Y_\alpha(t))$ is a $\mathcal{F}_t$-martingale.

A central limit theorem

Using the compensator of the FNPP, we can apply martingale methods in order to derive limit theorems for the FNPP. For the sake of completeness, we restate the definition of $\mathcal{F}$-stable convergence along with the theorem which will be used later.

Definition 14. If $(X_n)_{n\in\mathbb{N}}$ and X are $\mathbb{R}$-valued random variables on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and $\mathcal{F}$ is a sub-σ-algebra of $\mathcal{E}$, then $X_n \to X$ ($\mathcal{F}$-stably) in distribution if for all $B \in \mathcal{F}$ and all $A \in \mathcal{B}(\mathbb{R})$ with $\mathbb{P}(X \in \partial A) = 0$,

$$\mathbb{P}(\{X_n \in A\} \cap B) \xrightarrow[n\to\infty]{} \mathbb{P}(\{X \in A\} \cap B)$$

(see Definition A.3.2.III. in Daley and Vere-Jones 2003).

Note that $\mathcal{F}$-stable convergence implies weak convergence/convergence in distribution. We can derive a central limit theorem for the FNPP using Corollary 14.5.III. in Daley and Vere-Jones 2003 which we state here as a lemma for convenience.

Lemma 3. Let N be a simple point process on $\mathbb{R}_+$, $(\mathcal{F}_t)_{t\geq 0}$-adapted and with continuous $(\mathcal{F}_t)_{t\geq 0}$-compensator A. Suppose for each $T > 0$ an $(\mathcal{F}_t)_{t\geq 0}$-predictable process $f_T(t)$ is given such that

$$B_T^2 = \int_0^T [f_T(u)]^2 \mathrm{d}A(u) > 0$$

and define

$$X_T := \int_0^T f_T(u)[\mathrm{d}N(u) - \mathrm{d}A(u)].$$

Then the randomly normed integrals X_T/B_T converge $\mathcal{F}_0$-stably to a standard normal variable $W \sim N(0,1)$ for $T \to \infty$.

The above lemma allows us to show the following result for the FNPP.

Proposition 32. Let $(N(Y_\alpha(t)))_{t\geq 0}$ be the FNPP adapted to the filtration $(\mathcal{F}_t)_{t\geq 0}$ as defined in Section 3.2.1. Then,

$$\frac{N(Y_\alpha(T)) - \Lambda(Y_\alpha(T))}{\sqrt{\Lambda(Y_\alpha(T))}} \xrightarrow[T\to\infty]{} W \sim N(0,1) \qquad \mathcal{F}_0\text{-stably.} \tag{3.12}$$

Proof. First note that the compensator $A(t) := \Lambda(Y_\alpha(t))$ is continuous in t. Let $f_T(u) \equiv 1$ a constant, then

$$\begin{aligned} B_T^2 &= \int_0^T [f_T(u)]^2 \mathrm{d}A(u) = \Lambda(Y_\alpha(T)) - \Lambda(Y_\alpha(0)) \\ &= \Lambda(Y_\alpha(T)) > 0,\ \forall T > 0 \end{aligned}$$

and

$$\begin{aligned} X_T &:= \int_0^T [\mathrm{d}N(Y_\alpha(u)) - \mathrm{d}A(u)] \\ &= N(Y_\alpha(T)) - A(T) - N(Y_\alpha(0)) + A(0) \\ &= N(Y_\alpha(T)) - A(T)] = N(Y_\alpha(T)) - \Lambda(Y_\alpha(T)). \end{aligned}$$

It follows from Lemma 3 above that

$$\frac{X_T}{B_T} = \frac{N(Y_\alpha(T)) - \Lambda(Y_\alpha(T))}{\sqrt{\Lambda(Y_\alpha(T))}} \xrightarrow[T\to\infty]{} W \sim N(0,1) \qquad \mathcal{F}_0\text{-stably.}$$

□

Limit $\alpha \to 1$

In the following, we give a more rigorous proof for the limit $\alpha \to 1$ (point (ii) under special cases) in Section 2.4.3. We will use Theorem 3.36 in Jacod and Shiryaev 2003 which will be stated here as a lemma.

Lemma 4. Let $(X_t)_{t\geq 0}$ be a Poisson process with compensator $(A_t)_{t\geq 0}$, i.e. $A_t = \mathbb{E}[X_t]$, and (X^n) a sequence of point processes $(X_t^n)_{t\geq 0}$ with compensator $(A_t^n)_{t\geq 0}$; let $D \subset \mathbb{R}_+$.

(i) The following condition implies $X^n \xrightarrow[n\to\infty]{\text{f.d.}} X$ on the set D:

$$A_t^n \xrightarrow[n\to\infty]{P} A \quad \forall t \in D. \tag{3.13}$$

(ii) If moreover D is dense in $\mathbb{R}_+$, then (3.13) implies the $X^n \xrightarrow[n\to\infty]{J_1} X$.

Proposition 33. Let the FNPP be adapted to the filtration $(\mathcal{F}_t)$ as in (3.9) with non-trivial initial history $\mathcal{F}_0 := \sigma(\{Y_\alpha(t), t \geq 0\})$. Let $(N_\alpha(t))_{t\geq 0}$ be the FNPP as defined in (2.7). Then, we have the limit

$$N_\alpha \xrightarrow[\alpha\to 1]{J_1} N \quad \text{in} \quad D([0,\infty)).$$

Proof. By Proposition 31 we see that $(\Lambda(Y_\alpha(t)))_{t\geq 0}$ is the compensator of $(N_\alpha(t))_{t\geq 0}$. According to Lemma 4 it suffices to show

$$\Lambda(Y_\alpha(t)) \xrightarrow[\alpha\to 1]{\mathcal{P}} \Lambda(t).$$

We can check that the Laplace transform of the density of the inverse α-stable subordinator converges to the Laplace transform of the delta distribution:

$$\mathcal{L}\{h_\alpha(\cdot, y)\}(s, y) = E_\alpha(-ys^\alpha) \xrightarrow[\alpha\to 1]{} \mathrm{e}^{-ys} = \mathcal{L}\{\delta_0(\cdot - y)\}(s, y). \tag{3.14}$$

We may take the limit as the power series representation of the (entire) Mittag-Leffler function is absolutely convergent. Thus (3.14) implies

$$Y_\alpha(t) \xrightarrow[\alpha\to 1]{d} t \qquad \forall t \in \mathbb{R}_+.$$

As convergence in distribution to a constant automatically improves to convergence in probability, we have

$$Y_\alpha(t) \xrightarrow[\alpha\to 1]{\mathcal{P}} t \qquad \forall t \in \mathbb{R}_+.$$

By the continuous mapping theorem, it follows that

$$\Lambda(Y_\alpha(t)) \xrightarrow[\alpha\to 1]{\mathcal{P}} \Lambda(t) \qquad \forall t \in \mathbb{R}_+,$$

which concludes the proof. □

3.3 Regular variation and scaling limits

In this section we will work with the trivial initial filtration setting ($\mathcal{F}_0 = \{\varnothing, \Omega\}$), i.e. $\mathcal{F}_t$ is assumed to be the natural filtration of the FNPP. In this setting, the FNPP can generally not be seen as a Cox process and although the compensator of the FNPP does exist, it is difficult to give a closed form expression for it.
Instead, we follow the approach of results given in Grandell 1976, Serfozo 1972a, Serfozo 1972b, which require conditions on the function Λ. Recall that a function Λ

is *regularly varying with index* $\beta \in \mathbb{R}$ if

$$\frac{\Lambda(xt)}{\Lambda(t)} \xrightarrow[t\to\infty]{} x^\beta, \qquad \forall x > 0. \tag{3.15}$$

Under the mild condition of measurability, one can show that the above condition is not too restrictive in the sense that if the quotient of the right hand side of (3.15) converges to a function $x \mapsto g(x)$, g has to be of the form x^β (see Theorem A.1).

Example 8. We check whether typical rate functions (taken from Example 1) fulfill the regular variation condition.

(i) Weibull's rate function

$$\Lambda(t) = \left(\frac{t}{b}\right)^c, \quad \lambda(t) = \frac{c}{b}\left(\frac{t}{b}\right)^{c-1}, \quad c \geq 0,\, b > 0$$

is regularly varying with index c. This can be seen as follows

$$\frac{\Lambda(xt)}{\Lambda(t)} = \frac{(xt)^c}{t^c} = x^c, \quad \forall x > 0.$$

(ii) Makeham's rate function

$$\Lambda(t) = \frac{c}{b}\mathrm{e}^{bt} - \frac{c}{b} + \mu t, \quad \lambda(t) = c\mathrm{e}^{bt} + \mu, \quad c > 0,\, b > 0,\, \mu \geq 0$$

is not regularly varying, since

$$\frac{\Lambda(xt)}{\Lambda(t)} = \frac{(c/b)\mathrm{e}^{bxt} - (c/b) + \mu xt}{(c/b)\mathrm{e}^{bt} - (c/b) + \mu t} = \frac{(c/b)\mathrm{e}^{bt(x-1)} - (c/b)\mathrm{e}^{-bt} + \mu x t\mathrm{e}^{-bt}}{(c/b) - (c/b)\mathrm{e}^{-bt} + \mu t\mathrm{e}^{-bt}}$$
$$\xrightarrow{t\to\infty} \begin{cases} 0 & \text{if } x < 1 \\ 1 & \text{if } x = 1 \\ +\infty & \text{if } x > 1 \end{cases}$$

does not fulfill (3.15). △

In the following, the condition that Λ is regularly varying is useful for proving limit results. We will first show a one-dimensional limit theorem before moving on to the functional analogue.

3.3.1 A one-dimensional limit theorem

For a one-dimensional limit, we first provide a self-contained proof which essentially evokes Lévy's continuity theorem and is a good excercise before using the same

method in multiple dimensions in the next section to show convergence in finite-dimensions.

Proposition 34. Let the FNPP $(N_\alpha(t))_{t\geq 0}$ be defined as in Equation (2.7). Suppose the function $t \mapsto \Lambda(t)$ is regularly varying with index $\beta \in \mathbb{R}$. Then the following limit holds for the FNPP:

$$\frac{N_\alpha(t)}{\Lambda(t^\alpha)} \xrightarrow[t\to\infty]{d} (Y_\alpha(1))^\beta. \tag{3.16}$$

Proof. We will first show that the characteristic function of the random variable on the left hand side of (3.16) converges to the characteristic function of the right hand side.
By self-similarity of Y_α we have

$$N_1(\Lambda(Y_\alpha(t))) \stackrel{d}{=} N_1(\Lambda(t^\alpha Y_\alpha(1))).$$

Therefore, it follows for the characteristic function of $Z(t) := \frac{N_\alpha(t)}{\Lambda(t^\alpha)}$ that

$$\begin{aligned}
\varphi(t) &:= \mathbb{E}[\exp(\mathrm{i}uZ(t))] = \mathbb{E}[\exp(\mathrm{i}u\Lambda(t^\alpha)^{-1}N_1(\Lambda(Y_\alpha(t))))] \\
&= \mathbb{E}[\exp(\mathrm{i}u\Lambda(t^\alpha)^{-1}N_1(\Lambda(t^\alpha Y_\alpha(1))))] \\
&= \int_0^\infty \mathbb{E}[\exp(\mathrm{i}u\Lambda(t^\alpha)^{-1}N_1(\Lambda(t^\alpha x)))]h_\alpha(1,x)\,\mathrm{d}x \qquad (3.17) \\
&= \int_0^\infty \exp(\Lambda(t^\alpha x)(\mathrm{e}^{\mathrm{i}u\Lambda(t^\alpha)^{-1}} - 1))h_\alpha(1,x)\,\mathrm{d}x, \qquad (3.18)
\end{aligned}$$

where we used a conditioning argument in (3.17), $x \mapsto h_\alpha(1,x)$ is the density function of the distribution of $Y_\alpha(1)$. In the last step in (3.18) we may insert the characteristic function of a Poisson distributed random variable with parameter $\Lambda(t^\alpha x)$ evaluated at the point $u\Lambda(t^\alpha)^{-1}$.
In order to pass to the limit, we need to justify that we may exchange integration and limit. It can be observed that the integrand is dominated by an integrable function independent of t:

$$\begin{aligned}
&\big|\mathbb{E}[\exp(\mathrm{i}u\Lambda(t^\alpha)^{-1}\ N_1(\Lambda(t^\alpha x)))]h_\alpha(1,x)| \\
&\qquad\qquad \leq \mathbb{E}[|\exp(\mathrm{i}u\Lambda(t^\alpha)^{-1}N_1(\Lambda(t^\alpha x)))|]h_\alpha(1,x) \leq h_\alpha(1,x)
\end{aligned}$$

This allows us to use the dominated convergence theorem to get

$$\begin{aligned}
\lim_{t\to\infty}\varphi(t) &= \lim_{t\to\infty}\int_0^\infty \exp(\Lambda(t^\alpha x)(\mathrm{e}^{\mathrm{i}u\Lambda(t^\alpha)^{-1}} - 1))h_\alpha(1,x)\,\mathrm{d}x \\
&= \int_0^\infty \Big[\lim_{t\to\infty}\exp(\Lambda(t^\alpha x)(\mathrm{e}^{\mathrm{i}u\Lambda(t^\alpha)^{-1}} - 1))\Big]\, h_\alpha(1,x)\,\mathrm{d}x. \qquad (3.19)
\end{aligned}$$

We are left with calculating the limit in the square bracket in (3.19). To this end,

consider a power series expansion of $\mathrm{e}^{\mathrm{i}u\Lambda(t^\alpha)^{-1}}$ to observe that

$$\begin{aligned}\exp\left(\Lambda(t^\alpha x)(\mathrm{e}^{\mathrm{i}u\Lambda(t^\alpha)^{-1}}-1)\right) &= \exp\left(\Lambda(t^\alpha x)\left(\sum_{k=1}^{\infty}\frac{(\mathrm{i}u)^k}{\Lambda(t^\alpha)^k k!}\right)\right)\\ &= \exp\left(\frac{\mathrm{i}u}{1!}\underbrace{\frac{\Lambda(t^\alpha x)}{\Lambda(t^\alpha)}}_{\xrightarrow[t\to\infty]{}x^\beta} + \underbrace{\Lambda(t^\alpha x)\mathcal{O}\left(\frac{1}{\Lambda(t^\alpha)^2}\right)}_{\xrightarrow[t\to\infty]{}0}\right),\end{aligned}$$

where we have used that Λ is regularly varying with index β in the last step. Inserting this result into (3.19) yields

$$\begin{aligned}\lim_{t\to\infty}\varphi(t) &= \int_0^\infty \exp\left(\mathrm{i}ux^\beta\right) h_\alpha(1,x)dx\\ &= \mathbb{E}[\mathrm{e}^{\mathrm{i}u(Y_\alpha(1))^\beta}].\end{aligned}$$

Applying Lévy's continuity theorem concludes the proof. □

A more general result can be found as Theorem 3.4 in Serfozo 1972a or Theorem 1 on pp. 69-70 in Grandell 1976, which we state here with slight modification of notation:

Theorem 35. Let N be a doubly stochastic process with directing process A. Suppose that there exist real numbers $(a_t)_{t\geq 0}$ and $(b_t)_{t\geq 0}$ with

$$a_t > 0 \quad \text{and} \quad \frac{b_t}{a_t} \xrightarrow[t\to\infty]{} \sigma^2,\ 0 \leq \sigma^2 < \infty$$

and a random variable S such that

$$\frac{A(t)}{a_t} - b_t \xrightarrow[n\to\infty]{d} S.$$

Then

$$\frac{N(A(t))}{a_t} - b_t \xrightarrow[n\to\infty]{d} S + \sigma W,$$

where W is a standard normal distributed random variable.

Remark 10. As a special case of the Proposition 34 we get for $\Lambda(t) = \lambda t$, for constant $\lambda > 0$

$$\frac{\Lambda(xt)}{\Lambda(t)} = x^1$$

which means Λ is regularly varying with index $\beta = 1$. It follows that

$$\frac{N_1(\lambda Y_\alpha(t))}{\lambda t^\alpha} \xrightarrow[t\to\infty]{d} Y_\alpha(1).$$

This is in accordance to the scaling limit given in Cahoy, Uchaikin and Woyczynski 2010 who showed

$$\frac{N_1(\lambda Y_\alpha(t))}{\mathbb{E}[N_1(\lambda Y_\alpha(t))]} = \frac{N_1(\lambda Y_\alpha(t))}{\frac{\lambda t^\alpha}{\Gamma(1+\alpha)}} \xrightarrow[t\to\infty]{d} \Gamma(1+\alpha) Y_\alpha(1).$$

3.3.2 A functional limit theorem

The one-dimensional result in Proposition 34 can be extended to a functional limit theorem. In the following, we consider the Skorokhod space $\mathcal{D}([0,\infty))$ endowed with a suitable topology introduced in Section 3.1.

Theorem 36. Let the FNPP $(N_\alpha(t))_{t\geq 0}$ be defined as in Equation (2.7). Suppose the function $t \mapsto \Lambda(t)$ is regularly varying with index $\beta \in \mathbb{R}$. Then the following limit holds for the FNPP:

$$\left(\frac{N_\alpha(t\tau)}{\Lambda(t^\alpha)}\right)_{\tau\geq 0} \xrightarrow[t\to\infty]{J_1} \left([Y_\alpha(\tau)]^\beta\right)_{\tau\geq 0}. \tag{3.20}$$

Remark 11. As the limit process has continuous paths the mode of convergence improves to local uniform convergence. Also in this section, we will denote the homogeneous Poisson process with intensity parameter $\lambda = 1$ with N_1.

In order to prove the theorem we need Theorem 2 on p. 81 in Grandell 1976, which we will state here for convenience.

Theorem 37. Let $\bar{\Lambda}$ be a stochastic process in $\mathcal{D}([0,\infty))$ with $\bar{\Lambda}(0) = 0$ and let $N = N_1(\bar{\Lambda})$ be the corresponding doubly stochastic process. Let $a \in \mathcal{D}([0,\infty))$ with $a(0) = 0$ and $t \mapsto b_t$ a positive regularly varying function with index $\rho > 0$ such that

$$\frac{a(t)}{b_t} \xrightarrow[t\to\infty]{} \kappa \in [0,\infty) \text{ and}$$
$$\left(\frac{\bar{\Lambda}(t\tau) - a(t\tau)}{b_t}\right)_{\tau\geq 0} \xrightarrow[t\to\infty]{J_1} (S(\tau))_{\tau\geq 0},$$

where S is a stochastic process in $\mathcal{D}([0,\infty))$. Then

$$\left(\frac{N(t\tau) - a(t\tau)}{b_t}\right)_{\tau\geq 0} \xrightarrow[t\to\infty]{J_1} (S(\tau) + h(B(\tau)))_{\tau\geq 0},$$

where $h(\tau) = \kappa\tau^{2\rho}$ and $(S(t))_{t\geq 0}$ and $(B(t))_{t\geq 0}$ are independent. $(B(t))_{t\geq 0}$ is the standard Brownian motion in $\mathcal{D}([0,\infty))$.

Proof of Theorem 36. We apply Theorem 37 and choose $a \equiv 0$ and $b_t = \Lambda(t^\alpha)$. Then it follows that $\kappa = 0$ and it can be checked that b_t is regularly varying with index

$\alpha\beta$:

$$\frac{b_{xt}}{b_t} = \frac{\Lambda(x^\alpha t^\alpha)}{\Lambda(t^\alpha)} \xrightarrow[t\to\infty]{} x^{\alpha\beta}$$

by the regular variation property in (3.15).
We are left to show that

$$\tilde{\Lambda}_t(\tau) := \left(\frac{\Lambda(Y_\alpha(t\tau))}{\Lambda(t^\alpha)}\right)_{\tau\geq 0} \xrightarrow[t\to\infty]{J_1} \left([Y_\alpha(\tau)]^\beta\right)_{\tau\geq 0}.$$

This can be done by following the usual technique of first proving convergence of the finite-dimensional marginals and then tightness of the sequence in the Skorokhod space $\mathcal{D}([0,\infty))$.
Concerning the convergence of the finite-dimensional marginals we show convergence of their respective characteristic functions. Let $t > 0$ be fixed at first, $\tau = (\tau_1, \tau_2, \ldots, \tau_n) \in \mathbb{R}^n_+$ and $\langle\cdot,\cdot\rangle$ denote the scalar product in $\mathbb{R}^n$. Then, we can write the characteristic function of the joint distribution of the vector

$$\frac{\Lambda(t^\alpha Y_\alpha(\tau))}{\Lambda(t^\alpha)} = \left(\frac{\Lambda(t^\alpha Y_\alpha(\tau_1))}{\Lambda(t^\alpha)}, \frac{\Lambda(t^\alpha Y_\alpha(\tau_2))}{\Lambda(t^\alpha)}, \ldots, \frac{\Lambda(t^\alpha Y_\alpha(\tau_n))}{\Lambda(t^\alpha)}\right) \in \mathbb{R}^n_+$$

as

$$\begin{aligned}
\varphi_t(u) &:= \mathbb{E}\left[\exp\left(\mathrm{i}\left\langle u, \frac{\Lambda(Y_\alpha(t\tau))}{\Lambda(t^\alpha)}\right\rangle\right)\right] = \mathbb{E}\left[\exp\left(\mathrm{i}\left\langle u, \frac{\Lambda(t^\alpha Y_\alpha(\tau))}{\Lambda(t^\alpha)}\right\rangle\right)\right] \\
&= \int_{\mathbb{R}^n_+} \exp\left(\mathrm{i}\left\langle u, \frac{\Lambda(t^\alpha x)}{\Lambda(t^\alpha)}\right\rangle\right) h_\alpha(\tau, x)\mathrm{d}x \\
&= \int_{\mathbb{R}^n_+} \left[\prod_{k=1}^n \exp\left(\mathrm{i}u_k \frac{\Lambda(t^\alpha x_k)}{\Lambda(t^\alpha)}\right)\right] h_\alpha(\tau_1, \ldots, \tau_n; x_1, \ldots, x_n)\mathrm{d}x_1 \ldots \mathrm{d}x_n
\end{aligned}$$

where $u \in \mathbb{R}^n$ and $h_\alpha(\tau, x) = h_\alpha(\tau_1, \tau_2, \ldots, \tau_n; x_1, x_2 \ldots, x_n)$ is the density of the joint distribution of $(Y_\alpha(\tau_1), Y_\alpha(\tau_2), \ldots, Y_\alpha(\tau_n))$. We can find a dominating function by the following estimate:

$$\left|\exp\left(\mathrm{i}\left\langle u, \frac{\Lambda(t^\alpha x)}{\Lambda(t^\alpha)}\right\rangle\right) h_\alpha(\tau, x)\right| \leq h_\alpha(\tau, x).$$

The upper bound is an integrable function which is independent of t. By dominated

convergence we may interchange limit and integration:

$$\begin{aligned}\lim_{t\to\infty}\varphi_n(u) &= \lim_{t\to\infty}\int_{\mathbb{R}^n_+}\exp\left(\mathrm{i}\left\langle u,\frac{\Lambda(t^\alpha x)}{\Lambda(t^\alpha)}\right\rangle\right)h_\alpha(\tau,x)\mathrm{d}x\\ &= \int_{\mathbb{R}^n_+}\lim_{t\to\infty}\exp\left(\mathrm{i}\left\langle u,\frac{\Lambda(t^\alpha x)}{\Lambda(t^\alpha)}\right\rangle\right)h_\alpha(\tau,x)\mathrm{d}x\\ &= \int_{\mathbb{R}^n_+}\exp\left(\mathrm{i}\left\langle u,x^\beta\right\rangle\right)h_\alpha(\tau,x)\mathrm{d}x = \mathbb{E}[\exp(i\langle u,(Y_\alpha(\tau))^\beta\rangle)],\end{aligned}$$

where in the last step we used the continuity of the exponential function and the scalar product to calculate the limit. By Lévy's continuity theorem we may conclude that for $n\in\mathbb{N}$

$$\left(\frac{\Lambda(Y_\alpha(t\tau_k))}{\Lambda(t^\alpha)}\right)_{k=1,\dots,n}\xrightarrow[t\to\infty]{d}\left([Y_\alpha(\tau_k)]^\beta\right)_{k=1,\dots,n}.$$

In order to show tightness, first observe that for fixed t both the stochastic process $\tilde{\Lambda}_t$ on the left hand side and the limit candidate $([Y_\alpha(\tau)]^\beta)_{\tau\geq 0}$ have increasing paths. Moreover, the limit candidate has continuous paths. Therefore we are able to invoke Theorem 25 to ensure tightness of the sequence $(\tilde{\Lambda}_t)_{t\geq 0}$ and thus the assertion follows. □

By applying the transformation theorem for probability densities to (2.6), we can write for the density $h^\beta_\alpha(t,\cdot)$ of the one-dimensional marginal of the limit process $([Y_\alpha(t)]^\beta)_{t\geq 0}$ as

$$\begin{aligned}h^\beta_\alpha(t,x) &= \frac{1}{\beta}x^{1/\beta-1}h_\alpha(t,x^{1/\beta})\\ &= \frac{1}{\beta}x^{1/\beta-1}\frac{t}{\alpha x^{1/\beta(1+1/\alpha)}}g_\alpha\left(\frac{t}{y^{1/(\alpha\beta)}}\right)\\ &= \frac{t}{\alpha\beta x^{1+1/(\alpha\beta)}}g_\alpha\left(\frac{t}{y^{1/(\alpha\beta)}}\right).\end{aligned}\tag{3.21}$$

Note that this is *not* the density of $Y_{\alpha\beta}(t)$.
A further limit result can be obtained for the FHPP via a continuous mapping argument.

Proposition 38. Let $(N_1(t))_{t\geq 0}$ be a homogeneous Poisson process and $(Y_\alpha(t))_{t\geq 0}$ be the inverse α-stable subordinator. Then

$$\left(\frac{N_1(Y_\alpha(t))-\lambda Y_\alpha(t)}{\sqrt{\lambda}}\right)_{t\geq 0}\xrightarrow[\lambda\to\infty]{J_1}(B(Y_\alpha(t)))_{t\geq 0},$$

where $(B(t))_{t\geq 0}$ is a standard Brownian motion.

Proof. The classic result

$$\left(\frac{N_1(t) - \lambda t}{\sqrt{\lambda}}\right)_{t\geq 0} \xrightarrow[\lambda\to\infty]{J_1} (B(t))_{t\geq 0}$$

can be shown by using that $(N_1(t) - \lambda t)_{t\geq 0}$ is a martingale. As $(B(t))_{t\geq 0}$ has continuous paths and $(Y_\alpha(t))_{t\geq 0}$ has increasing paths we may use Theorem 27 to obtain the result. □

The above proposition can be compared with Lemma 5 in the next section and a similar continuous mapping argument is applied in the proof of Theorem 41.

3.4 The fractional compound Poisson process

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. The fractional compound Poisson process is defined analogously to the standard compound Poisson process where the Poisson process is replaced by a fractional one:

$$Z_\alpha(t) := \sum_{k=1}^{N_\alpha(t)} X_k, \tag{3.22}$$

where $\sum_{k=1}^{0} X_k := 0$. The process N_α is not necessarily independent of the X_i's unless stated otherwise.
We will assume a limit result for the sequence of partial sums without time-change

$$S_n := \sum_{k=1}^{n} X_k, \tag{3.23}$$

usually a stable limit, i.e. there exist sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ and a random variable following a stable distribution S such that

$$\bar{S}_n := a_n S_n - b_n \xrightarrow[n\to\infty]{d} S.$$

(see also Section 2.1.2 for more on stable distributions). In other words the distribution of the X_k's is in the domain of attraction of a stable law.
In the following, we will derive limit theorems for the fractional compound Poisson process. In Section 3.4.2, we assume N_α to be independent of the X_k's and use a continuous mapping theorem argument to show functional convergence w.r.t. a suitable Skorokhod topology. A corresponding one-dimensional limit theorem would follow directly from the functional one: According to Proposition VI.3.14 in Jacod and Shiryaev 2003 or Theorem 11.6.6 in Whitt 2002 convergence in J_1 and M_1 implies that we also have convergence in finite dimensions on the time domain except

the set of fixed times of discontinuity. As stable processes are Lévy processes, this set is empty (see Lemma 2.3.2 in Applebaum 2009) and Y_α has continuous paths. However, in the special case of N_α being a FHPP, using Anscombe's theorem in the next section allows us to drop the independence assumption between N_α and the X_k's and thus strengthen the result for the one-dimensional limit.

3.4.1 A one-dimensional limit result

The following theorem is due to Anscombe 1952 and can be found slightly reformulated in Richter 1965.

Theorem 39. We assume that the following conditions are fulfilled:

(i) The sequence of random variables R_n such that

$$R_n \xrightarrow[n\to\infty]{d} R,$$

for some random variable R.

(ii) Let the family of integer-valued random variables $(N(t))_{t\geq 0}$ be relatively stable, i.e. for a real-valued function ψ with $\psi(t) \xrightarrow[t\to\infty]{} +\infty$ it holds that

$$\frac{N(t)}{\psi(t)} \xrightarrow[t\to\infty]{P} 1.$$

(iii) (Uniform continuity in probability) For every $\varepsilon > 0$ and $\eta > 0$ there exists a $c = c(\varepsilon, \eta)$ and a $t_0 = t_0(\varepsilon, \eta)$ such that for all $t \geq t_0$

$$\mathbb{P}\left(\max_{m:|m-t|<ct} |R_m - R_t| > \varepsilon\right) < \eta.$$

Then,

$$R_{N(t)} \xrightarrow[t\to\infty]{d} R.$$

Concerning the condition (ii), note that the required convergence in probability is stronger than the convergence in distribution we have derived in the previous sections for the FNPP. Nevertheless, in the special case of the FHPP, we can prove the following lemma.

Lemma 5. Let N_α be a FHPP, i.e. $\Lambda(t) = \lambda t$. Then with $C := \frac{\lambda}{\Gamma(1+\alpha)}$ it holds that

$$\frac{N_\alpha(t)}{Ct^\alpha} \xrightarrow[t\to\infty]{P} 1.$$

Proof. According to Proposition 4.1 from Di Crescenzo, Martinucci and Meoli 2016 we have the result that for fixed $t > 0$ the convergence

$$\frac{N_1(\lambda Y_\alpha(t))}{\mathbb{E}[N_1(\lambda Y_\alpha(t))]} = \frac{N_1(\lambda Y_\alpha(t))}{\frac{\lambda t^\alpha}{\Gamma(1+\alpha)}} \xrightarrow[\lambda\to\infty]{L^1} 1 \tag{3.24}$$

holds and therefore also in probability.
It can be shown by using the fact that the moments and the waiting time distribution of the FHPP can be expressed in terms of the Mittag-Leffler function.
Let $\varepsilon > 0$. We have

$$\lim_{t\to\infty} \mathbb{P}\left(\left|\frac{N_1(\lambda Y_\alpha(t))}{Ct^\alpha} - 1\right| > \varepsilon\right) = \lim_{t\to\infty} \mathbb{P}\left(\left|\frac{N_1(\lambda t^\alpha Y(1))}{\frac{\lambda t^\alpha}{\Gamma(1+\alpha)}} - 1\right| > \varepsilon\right) \tag{3.25}$$

$$= \lim_{\tau\to\infty} \mathbb{P}\left(\left|\frac{N_1(\tau Y(1))}{\frac{\tau\cdot 1^\alpha}{\Gamma(1+\alpha)}} - 1\right| > \varepsilon\right) = 0, \tag{3.26}$$

where in (3.25) we used the self-similarity property of Y_α and in (3.26) we applied (3.24) with $t = 1$. □

As a direct application of Theorem 39 we can prove the following lemma.

Lemma 6. Let N_α be a FHPP and $X_1, X_2, \ldots$ be a sequence of i.i.d. variables in the DOA of a stable law μ. Then, for the partial sums S_n defined in (3.23) there exist sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ such that

$$a_{N_\alpha(t)} S_{N_\alpha(t)} - b_{N_\alpha(t)} \xrightarrow[t\to\infty]{d} S,$$

where $S \sim \mu$.

Proof. We would like to use the above theorem for $R_n = \bar{S}_n$. Indeed, condition (i) follows from the assumption that the law of X_1 lies in the domain of attraction of a stable law and condition (ii) follows from Lemma 5. It is readily proven in Theorem 3 in Anscombe 1952 that $(\bar{S}_n)$ satisfies the condition (iii), if condition (i) and (ii) are fulfilled. Therefore, it follows from Theorem 39 that

$$\bar{S}_{N_\alpha(t)} = a_{N_\alpha(t)} \sum_{k=1}^{N_\alpha(t)} X_k - b_{N_\alpha(t)} \xrightarrow[t\to\infty]{d} S. \tag{3.27}$$

□

Finally, we would like to replace $N_\alpha(t)$ with $\lfloor Ct^\alpha \rfloor$ in the index of a and b. This requires additional conditions. The following theorem is a slight modification of Theorem 3.6 in Chapter 9 of Gut 2013.

Theorem 40. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}[X_1] = 0$ and set

$$S_n := \sum_{k=1}^{n} X_k, \quad n \geq 1.$$

Suppose that $(a_n)_{n\geq 0}$ is a sequence of positive norming constants such that

$$\frac{S_n}{a_n} \xrightarrow[n\to\infty]{d} S,$$

where S follows a stable law with index $\alpha \in (1, 2]$. Let $(N(t))_{t\geq 0}$ be a sequence of integer-valued random variables such that (ii) in Theorem 39 is fulfilled. Then,

$$a_{\lfloor Ct^\alpha \rfloor} \sum_{k=1}^{N_\alpha(t)} X_k = a_{\lfloor Ct^\alpha \rfloor} Z_\alpha(t) \xrightarrow[t\to\infty]{d} S.$$

Idea of proof. By Lemma 6 we have

$$a_{N_\alpha(t)} \sum_{k=1}^{N_\alpha(t)} X_k \xrightarrow[t\to\infty]{d} S,$$

as $b_n = 0$ by assumption. In order to replace $N_\alpha(t)$ with $\lfloor Ct^\alpha \rfloor$ in the index of a one has to show that

$$\frac{N_\alpha(t)}{Ct^\alpha} \xrightarrow[t\to\infty]{P} 1$$

implies

$$\frac{a_{N_\alpha(t)}}{a_{\lfloor Ct^\alpha \rfloor}} \xrightarrow[t\to\infty]{P} 1.$$

The derivation of suitable estimates relies on the fact that $n \mapsto a_n$ is regularly varying (for details see Lemma 2.9 (a) in Gut 1974). □

Remark 12.

(i) The conditions restrict to the centered, symmetric case (i.e. $\mathbb{E}[X_1] = 0$, $b_n = 0$) and $\alpha \in (1, 2]$ as the mean exists (see Proposition 6). While it can be shown that $a_n \in \mathcal{R}_{-1/\alpha}$, in the non-symmetric case (see also Remark 1), we generally do not have a regular variation property for b_n.

(ii) Note that this convergence result does not require N_α to be independent of the X_k's. The above derivation also works for mixing sequences $X_1, X_2, \ldots$ instead of i.i.d. (see Csörgő and Fischler 1973 for a generalization of Anscombe's theorem for mixing sequences).

3.4.2 A functional limit theorem

Theorem 41. Let the FNPP $(N_\alpha(t))_{t\geq 0}$ be defined as in Equation (2.7) and suppose the function $t \mapsto \Lambda(t)$ is regularly varying with index $\beta \in \mathbb{R}$. Moreover let $X_1, X_2, \ldots$ be i.i.d. random variables independent of N_α. Assume that the law of X_1 is in the domain of attraction of a stable law, i.e. there exist sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ and a stable Lévy process $(S(t))_{t\geq 0}$ such that the partial sums S_n definced in (3.23) satisfy

$$\left(a_n S_{\lfloor nt \rfloor} - b_n\right)_{t\geq 0} \xrightarrow[n\to\infty]{J_1} (S(t))_{t\geq 0}. \tag{3.28}$$

Then the fractional compound Poisson process Z_α defined in (3.22) fulfills the following limit:

$$(c_n Z_\alpha(nt) - d_n)_{t\geq 0} \xrightarrow[n\to\infty]{M_1} \left(S\left([Y_\alpha(t)]^\beta\right)\right)_{t\geq 0},$$

where $c_n := a_{\lfloor \Lambda(n) \rfloor}$ and $d_n := b_{\lfloor \Lambda(n) \rfloor}$.

Proof. The proof follows the technique proposed by Meerschaert and Scheffler 2004: By Theorem 36 we have

$$\left(\frac{N_\alpha(t\tau)}{\Lambda(t^\alpha)}\right)_{\tau\geq 0} \xrightarrow[t\to\infty]{J_1} \left([Y_\alpha(\tau)]^\beta\right)_{\tau\geq 0}.$$

By the independence assumptions we can combine this with (3.28) to get

$$\left(a_{\lfloor \Lambda(n^\alpha) \rfloor} S_{\lfloor \Lambda(n^\alpha)t \rfloor} - b_{\lfloor \Lambda(n^\alpha) \rfloor}, [\Lambda(n^\alpha)]^{-1} N_\alpha(nt)\right)_{t\geq 0} \xrightarrow[n\to\infty]{J_1} (S(t), [Y_\alpha(t)]^\beta)_{t\geq 0}$$

in the space $\mathcal{D}([0,\infty), \mathbb{R}\times[0,\infty))$. Note that $([Y_\alpha(t)]^\beta)_{t\geq 0}$ is non-decreasing. Moreover, due to independence the Lévy processes $(S(t))_{t\geq 0}$ and $(D_\alpha(t))_{t\geq 0}$ do not have simultaneous jumps (for details see Becker-Kern, Meerschaert and Scheffler 2004 and more generally Cont and Tankov 2004). This allows us to apply Theorem 29 to get the assertion by a continuous mapping argument since the composition mapping is continuous in this setting. □

3.5 Some numerical examples

Figure 3.7 shows the shape and time-evolution of the densities for different values of α. As Y_α is an increasing process, the densities spread to the right hand side as time passes.

We conducted a small Monte-Carlo simulation in order to illustrate the one-dimensional convergence results of Proposition 32 and Proposition 34. In Figures 3.8, 3.9 and 3.10, we can see that the simulated values for the probability density $x \mapsto \varphi_\alpha(t, x)$

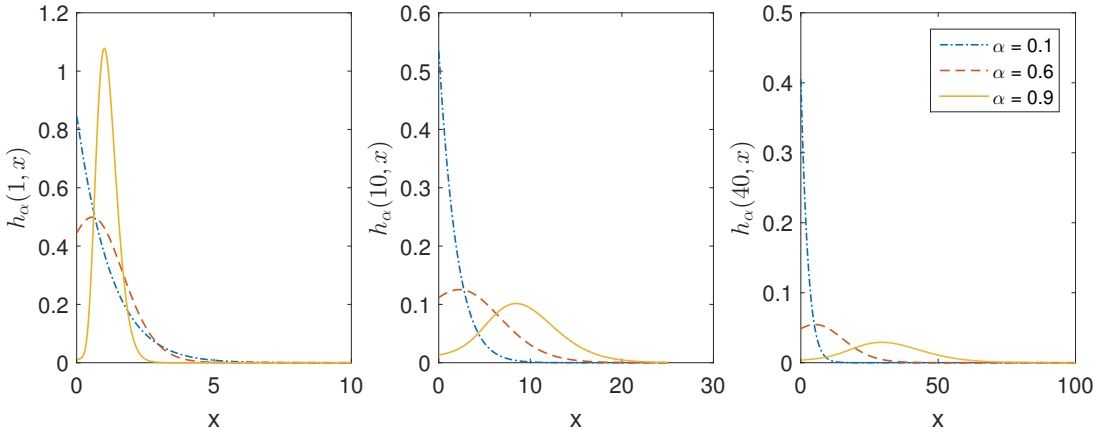

Figure 3.7: Plots of the probability densities $x \mapsto h_\alpha(t,x)$ of the distribution of the inverse α-stable subordinator $Y_\alpha(t)$ for different parameter $\alpha = 0.1, 0.6, 0.9$ indicating the time-evolution: the plot on the left is generated for $t = 1$, the plot in the middle for $t = 10$ and the plot on the right for $t = 40$.

of $[N(Y_\alpha(t)) - \Lambda(Y_\alpha(t))]/\sqrt{\Lambda(Y_\alpha(t))}$ approximate the density of a standard normal distribution for increasing time t. In a similar manner, Figure 3.11 depicts how the probability density function $x \mapsto \phi_\alpha(t,x)$ of $N_\alpha(t)/\Lambda(t^\alpha)$ approximates the density of $(Y_\alpha(t))^\beta$ given in (3.21), where Λ has regular variation index $\beta = 0.7$.

3.6 Summary

We have given a short review of weak convergence of probability measures and in particular in the case of continuous and càdlàg path spaces. Continuing with the FHPP of Chapter 2, we identified the FHPP as a Cox process under a suitable choice of filtration. Thus we were able to address the case $\alpha = 1$ for the FNPP as a functional limit. In the case of a trivial initial filtration, we made reasonable assumptions for the rate function Λ in order to derive scaling limits. Finally, we derive limit results for the fractional compound Poisson process. This concludes the first part of the thesis on fractional Poisson processes.

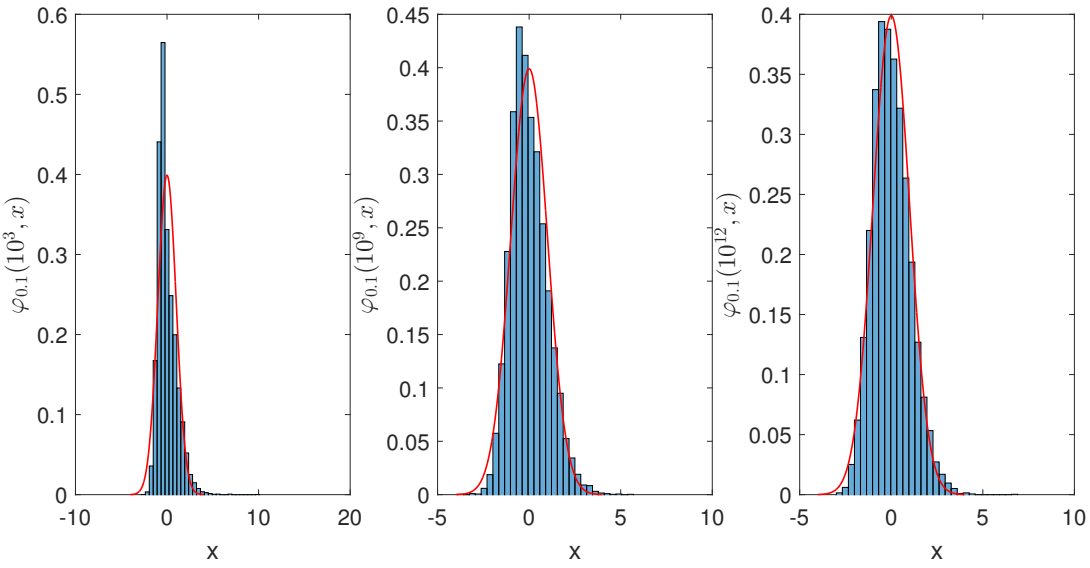

Figure 3.8: The red line shows the probability density function of the standard normal distribution, the limit distribution according to Proposition 32. The blue histograms depict samples of size 10^4 of the right hand side of (3.12) for different times $t = 10, 10^9, 10^{12}$ to illustrate convergence to the standard normal distribution.

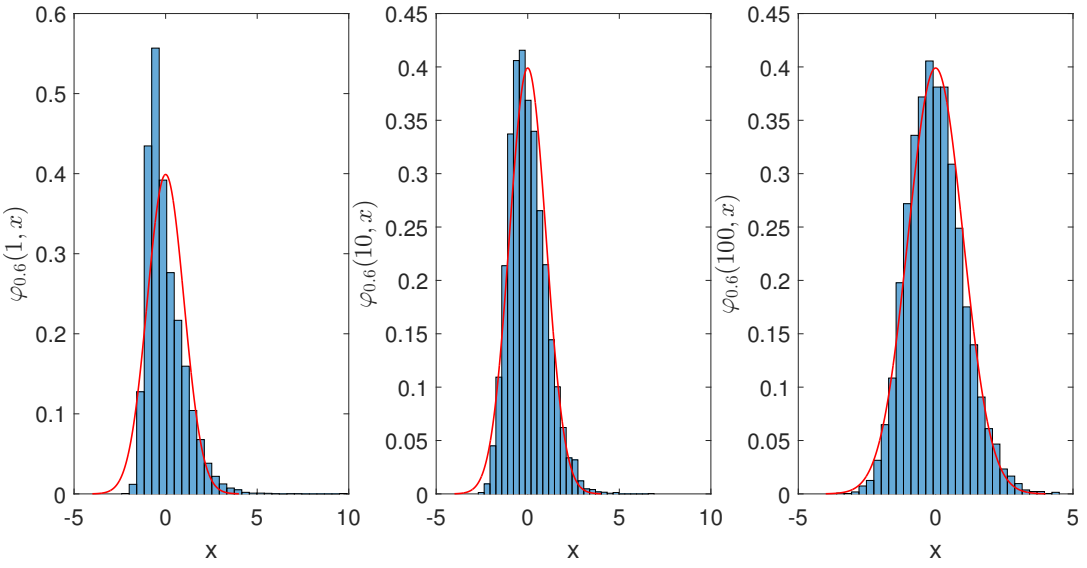

Figure 3.9: The red line shows the probability density function of the standard normal distribution, the limit distribution according to Proposition 32. The blue histograms depict samples of size 10^4 of the right hand side of (3.12) for different times $t = 1, 10, 100$ to illustrate convergence to the standard normal distribution.

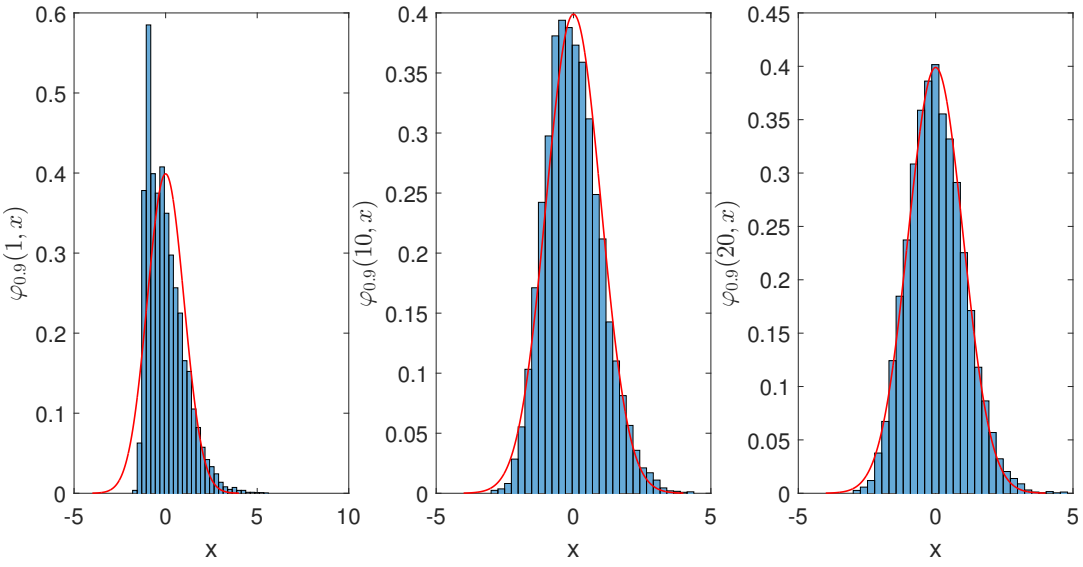

Figure 3.10: The red line shows the probability density function of the standard normal distribution, the limit distribution according to Proposition 32. The blue histograms depict samples of size 10^4 of the right hand side of (3.12) for different times $t = 1, 10, 20$ to illustrate convergence to the standard normal distribution.

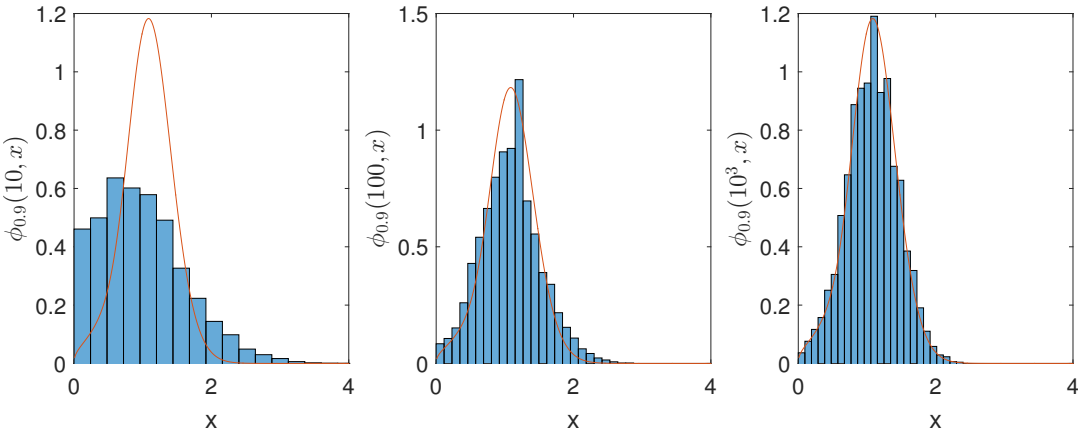

Figure 3.11: Red line: probability density function ϕ of the distribution of the random variable $(Y_{0.9}(1))^{0.7}$, the limit distribution according to Proposition 34. The blue histogram is based on 10^4 samples of the random variables on the right hand side of (3.16) for time points $t = 10, 100, 10^3$ to illustrate the convergence result.

Chapter 4

Information criteria and model selection

4.1 The model selection problem

In a broader sense, the principles that govern the process of selecting a single model or a subset of models from a given set of candidate models or working hypotheses in order to describe an observed phenomenon are deeply rooted in the in the idea of scientific method. Essentially, the decision to be made between a number of competing models is driven by two contradicting desires: fit and simplicity.
On the one hand, we cannot hope for a model, how complex it may be, to fully describe reality. From a philosophical standpoint it is questionable whether our perception and subsequent description of reality is able to reflect reality and truth as such. Is there such a thing as a true model? Moreover, independently from the notion of truth and true model, many phenomena that we seek to understand and analyze in sciences such as physics, biology and economic and social sciences exhibit a high degree of complexity. As a result, we consider models only as approximations of certain key aspects of reality in order to infer underlying principles and laws. As a quote attributed to George Box put it: "All models are wrong, but some are useful".
On the other hand, in contrary to the previous argument that a model can never be complex enough to match reality as a whole, pragmatism gives rise to the principle of parsimony or in its looser form sometimes referred to as Ockham's razor. [1] The idea to keep a model as simple as possible while still capturing the underlying phenomenon is not only mentioned in Ockham's work, but is stated in various forms by other scientists.

[1] Sometimes also written Occam.

The fact that models are required to be good approximations, i.e. to fit the data, while at the same time to be as simple as possible shows that a trade-off has to be made. Depending on the discipline there are various methods of model selection with varying degree of subjectivity. The context of this chapter is model selection in statistical sciences for which the model selection problem can be formulated as follows: Given observed (and appropriately collected) data and a set of candidate models arising from the scientific context where the data is coming from we want to quantify the suitability of a model relative to its competitors. Model parameters are estimated using statistical methods like maximum likelihood or least squares estimation. The trade-off between fit and simplicity can be seen as a trade-off between bias and variance: A model containing only few parameters may not be able to capture the underlying phenomenon whereas a complex model might contain too many parameters with little descriptive power. In other words, a good model should be able to separate the information from the noise within the data. "We are not trying to model the data, we are trying to model the information in the data." (p. 275, Burnham and Anderson 2004).

4.2 Information criteria

Information criteria (IC) are tools for model selection which aim to quantify the trade-off structure of the model selection problem. We first give an overview of the IC before presenting the ideas and derivations that justify the respective formulas. This section follows introductory work which can be found in Claeskens and Hjort 2008 and references therein.

Definition 15. For a given model fitted to data via MLE let $\mathcal{L}$ be the maximal log-likelihood value, k the number of parameters and n be the sample size of the data set. Then, we define:

1. **Akaike's information criterion (AIC)**

$$\mathrm{AIC} = -2\mathcal{L} + 2k \tag{4.1}$$

2. **Bayesian information criterion (BIC)**

$$\mathrm{BIC} = -2\mathcal{L} + k\ln(n) \tag{4.2}$$

3. **Hannan and Quinn information criterion (HQ)**

$$\mathrm{HQ} = -2\mathcal{L} + 2k\ln(\ln(n)) \tag{4.3}$$

The above information criteria are of the form

$$\mathrm{IC}(M_k) = -2\mathcal{L} + c(k, n) \tag{4.4}$$

where M_k is a model associated with parameter number k and $c(k, n)$ is a suitably chosen penalty term that accounts for the complexity of the model, i.e. the number of parameters. Within a given set of models to choose from, the "best" model is the one which minimizes the IC value. In other words, the selected model should give the best fit to the data, i.e. have a large log-likelihood value, while being as parsimonious as possible, i.e. use few parameters. Therefore, formula (4.4) represents the trade-off situation we have discussed previously.
The specific form of the penalty term for the AIC and BIC was initially derived from the respective frequentist and Bayesian parameter estimation framework. Whereas the AIC derives from the Kullback-Leibler or entropy distance and can be related to the likelihood ratio, the BIC is related to the Bayes factors. However, as pointed out in Burnham and Anderson 2004, the AIC can be interpreted in a Bayesian framework and vice versa, therefore negating a strict categorization of AIC and BIC as frequentist and Bayesian respectively. We will revisit this point in Section 4.2.3.

4.2.1 Akaike's information criterion

In his original work Akaike 1973, Akaike uses the connection between maximum likelihood estimation method and the Kullback-Leibler distance between probability distributions in order to derive a rule for model selection that is widely known as Akaike's information criterion (AIC).

Definition 16. Let f and g be two probability density functions defined on $\mathbb{R}^n$. Then the *Kullback-Leibler distance* (KL-distance) between f and g is given by

$$\mathrm{KL}(g, f) = \int_{\mathbb{R}^n} g(y) \log\left(\frac{g(y)}{f(y)}\right) \mathrm{d}y. \tag{4.5}$$

The KL-distance originated in Kullback and Leibler 1951 and is also known as entropy distance or relative entropy due to its relation to the physical notion of Boltzmann entropy (see p. 266 in Burnham and Anderson 2004). In general, the KL-distance is not symmetric, i.e. $\mathrm{KL}(g, f) \neq \mathrm{KL}(f, g)$, thus is not a bona fide distance. However, it does hold that $\mathrm{KL}(g, f) \geq 0$ and $\mathrm{KL}(g, f) = 0$ iff $g = f$.

We assume that the data $y = (y_1, y_2, \ldots, y_n)$ is coming from a distribution with probability density function g which is unknown. Moreover, we suppose that the context of the data gives reason to believe that the underlying distribution may be

well described by a member of a family of probability distributions with probability density functions $\{f(\cdot,\theta)\}_{\theta\in\Theta_p}$, where θ is a parameter vector and $\Theta_p \subset \mathbb{R}^p$ the space of admissible configurations of θ. The model selection problem in this setting often reduces to the optimal choice of the dimension of θ or in other words the order p of the model. Usually, due to practical reasons we only allow a finite set of orders to choose from, for example $p \in \{1, 2, \ldots, P\}$. However, the derivation does not necessarily require a set of nested models. In that case, p can be seen more generally as an index instead of an order. As a shorthand, we will refer to model p as "the model with order/index p and parameter vector $\theta \in \Theta_p$".
For fixed model order p we can estimate the parameter θ via maximum likelihood estimation.

$$\hat{\theta} := \arg\max_{\theta\in\Theta_p} \log(f(y_1, y_2, \ldots, y_n, \theta)). \tag{4.6}$$

where $l_n(\theta) := \log(f(y_1, y_2, \ldots, y_n, \theta))$ is the log-likelihood function. When solely looking at maximum likelihood values $l_n(\hat{\theta})$, as a measure of fit to compare different indices p, the most complex models, i.e. those with the highest possible number of parameters, will dominate the simpler ones.
For simplicity and illustration of key ideas of the derivation we assume that the data $(y_1, y_2, \ldots, y_n)$ consist of realizations of i.i.d. random variables $Y = (Y_1, Y_2, \ldots, Y_n)$ and the usual conditions are given for the consistency of the MLE $\hat{\theta}$ of θ and asymptotic normality (see for example Thm. 7.30 in Georgii 2007, which assumes identifiability and unimodality of the probability densities).
Indeed, the connection between KL-distance and MLE can be seen in the proof of consistency of MLE in the i.i.d. case: Due to independence, the log-likelihood function takes the form

$$l_n(\theta) = \sum_{k=1}^{n} \log(f(y_i, \theta)).$$

By the law of large numbers

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n} \log(f(y_i,\theta)) \xrightarrow[n\to\infty]{P} \mathbb{E}_g[\log(f(Y_1,\theta))] &= \int_{\mathbb{R}^n} g(y)\log(f(y,\theta))\,\mathrm{d}y \\ &= \int_{\mathbb{R}^n} g(y)\log(g(y))\,\mathrm{d}y - \mathrm{KL}(g, f(\cdot,\theta))\end{aligned}$$

Heuristically, we may reason that maximizing of the log-likelihood function is more or less equivalent to minimizing the Kullback-Leibler distance. Of course, convergence of the objective functions generally does not imply convergence of the respective optimizers. To ensure that, we need to require the functions in the sequence to

be unimodal or concave. The proof of consistency therefore implies:

$$\hat{\theta} \xrightarrow[n\to\infty]{P} \theta_0 := \arg\min_{\theta\in\Theta} \mathrm{KL}(g, f(\cdot,\theta)),$$

where θ_0 is the best or least false parameter.
If the true distribution g were known, the KL-distance would be a potential quantity for comparisons of different models with probability density $f(\cdot,\theta)$:

$$\mathrm{KL}(g, f(\cdot,\theta)) = \int_{\mathbb{R}^n} g(y)\log(g(y))\,\mathrm{d}y - \underbrace{\int_{\mathbb{R}^n} g(y)\log(f(y,\theta))\,\mathrm{d}y}_{=:R_n(\theta)}. \tag{4.7}$$

The first term of (4.7) is constant w.r.t. θ and thus constant for all p. Therefore, in order to discriminate among the candidate models, we only need to analyze the term R_n. Since we cannot compute R_n directly due to a lack of knowledge of g, Akaike reasons that it is possible to *estimate* the expected value of R_n instead: Define

$$Q_n := \mathbb{E}_g[R_n] = \mathbb{E}[\log(f(X,\hat{\theta}(Y)))], \tag{4.8}$$

where X and Y are to be understood as two independent random vectors with underlying distribution g. The expectation is therefore to be taken under their joint distribution, in other words the product measure. A possible estimator for Q_n can be obtained by using the empirical distribution

$$\hat{Q}_n := \frac{1}{n}\sum_{i=1}^{n} \log(f(Y_i,\hat{\theta})) = \frac{1}{n} l_n(\hat{\theta}). \tag{4.9}$$

In order to get an unbiased estimator in first approximation, we will compare the Taylor expansions of R_n and $\hat{Q}_n$. At this point, it is useful to revisit quantities and notation related to the consistency and asymptotic normality of MLE and their relation to the KL-distance.
Under suitable regularity conditions on f we can write down the optimality conditions of the minimizer θ_0 of the KL-distance. To this end, we define the first derivative the *score vector* and the *information matrix* as the first derivative and the Hessian of the function $\theta \mapsto \log(f(y,\theta))$ respectively:

$$u(y,\theta) := \frac{\partial}{\partial\theta}\log(f(y,\theta)) \qquad I(y,\theta) := \frac{\partial^2 \log(f(y,\theta))}{\partial\theta\partial\theta^t}$$

The necessary first order optimality condition of the KL-distance is then given by

$$0 = \left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta_0} \mathrm{KL}(g, f(\cdot,\theta)) = -\left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta_0} \int_{\mathbb{R}^n} g(y)\log(f(y,\theta))\,\mathrm{d}y \tag{4.10}$$

For conditions that allow us to interchange the order of differentiation and integration, for example dominated convergence, we require the existence of a dominating function $h \in L^1$ such that

$$\frac{\partial}{\partial\theta}\log(f(y,\theta)) = u(y,\theta) \leq h(y)$$

holds for all y and independent of θ, we may further simplify (4.10) to obtain

$$0 = -\int_{\mathbb{R}^n} g(y)u(y,\theta_0)\mathrm{d}y = -\mathbb{E}_g[u(Y,\theta_0)]. \tag{4.11}$$

In order to state the asymptotic normality result for MLE, we need to define following second order quantities:

$$J := -\mathbb{E}_g[I(Y,\theta_0)], \qquad K := \mathrm{Var}_g[u(Y,\theta_0)].$$

Under suitable conditions for asymptotic normality (see Section 4.4. in Czado and Schmidt 2011) we have the limit

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow[n\to\infty]{d} J^{-1}U' \sim N_p(0, J^{-1}KJ^{-1}),$$

where $U' \sim N_p(0,K)$. Let $V_n := \sqrt{n}(\hat{\theta}-\theta)$.
Having discussed the first and second derivative of $\theta \mapsto f(Y,\theta)$, we can use above quantities to write down its two-term Taylor expansion evaluated at the MLE $\hat{\theta}$:

$$\begin{aligned}\log(f(Y,\hat{\theta})) = \log(f(Y,\theta_0)) &+ u(Y,\theta_0)^t(\hat{\theta}-\theta_0) \\ &+ \frac{1}{2}(\hat{\theta}-\theta_0)^t I(Y,\theta_0)(\hat{\theta}-\theta_0) + S(\hat{\theta}),\end{aligned} \tag{4.12}$$

where the equality holds a.s. and S is a residual term that vanishes for $\hat{\theta} \to \theta_0$.
Taking expectations yields

$$\begin{aligned}R_n(\hat{\theta}) &= \int_{\mathbb{R}^n} g(y)\left[\log(f(y,\theta_0)) + u(y,\theta_0)^t(\hat{\theta}-\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)^t I(y,\theta_0)(\hat{\theta}-\theta_0)\right.\\ &\qquad \left. + S(\hat{\theta})\right]\mathrm{d}y \\ &= \underbrace{\int_{\mathbb{R}^n} g(y)\log(f(y,\theta_0))\,\mathrm{d}y}_{=:Q_0} + \underbrace{\left(\int_{\mathbb{R}^n} g(y)u(y,\theta_0)\,\mathrm{d}y\right)^t}_{=0}(\hat{\theta}-\theta_0) \\ &\qquad + \frac{1}{2}(\hat{\theta}-\theta_0)^t\left(\int_{\mathbb{R}^n} g(y)I(y,\theta_0)\,\mathrm{d}y\right)(\hat{\theta}-\theta_0) + \tilde{S}(\hat{\theta}) \\ &= Q_0 - \frac{1}{2}n^{-1}V_n^t J V_n + o_P(1),\end{aligned} \tag{4.13}$$

where it is assumes that the residual term after integration is of order $o_P(1)$ for $\hat{\theta} \to \theta_0$. [2]

Analogously, we may plug (4.12) into (4.9) to obtain the Taylor expansion of $\hat{Q}_n$:

$$\begin{aligned}
\hat{Q}_n &= \frac{1}{n}\sum_{i=1}^{n}\Big[\log(f(Y_i,\theta_0)) + u(Y_i,\theta_0)^t(\hat{\theta}-\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)^t I(Y_i,\theta_0)(\hat{\theta}-\theta_0) \\
&\qquad + T(\hat{\theta})\Big] \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\log(f(Y_i,\theta_0))}_{=:\bar{Z}_n} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}u(Y_i,\theta_0)\right)}_{=:\bar{U}_n}^t(\hat{\theta}-\theta_0) \\
&\qquad + \frac{1}{2}(\hat{\theta}-\theta_0)^t\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}I(Y_i,\theta_0)\right)}_{=:\bar{J}_n}(\hat{\theta}-\theta_0) + o_P(1) \\
&= \bar{Z}_n + \bar{U}_n^t(\hat{\theta}-\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)^t\bar{J}_n(\hat{\theta}-\theta_0) + o_P(1). \qquad (4.14)
\end{aligned}$$

By the law of large numbers and the central limit theorem respectively we have

$$Q_0 - \bar{Z}_n \xrightarrow[n\to\infty]{P} 0, \quad \sqrt{n}\bar{U}_n \xrightarrow[n\to\infty]{d} U' \sim N_p(0,K) \quad \text{and} \quad J - J_n \xrightarrow[n\to\infty]{P} 0. \qquad (4.15)$$

Using (4.13) and (4.14) we further compute

$$\hat{Q}_n - R_n = \underbrace{(Z_n - Q_0)}_{\to 0} + \frac{1}{n}\underbrace{\sqrt{n}\bar{U_n}^t}_{\to (U')^t}\underbrace{V_n}_{J^{-1}U} + o_P(n^{-1}).$$

In first approximation, for large n, we may conclude that the difference $\hat{Q}_n - R_n$ behaves approximately like $n^{-1}(U')^t J^{-1}U'$. The convergence is in distribution and from a rigorous standpoint it is not possible to justify that the convergence also holds in mean. Nevertheless, in a last rather heuristic step we approximate:

$$\begin{aligned}
\mathbb{E}_g[\hat{Q}_n - Q_n] &= \mathbb{E}_g[\hat{Q}_n - R_n] - \underbrace{\mathbb{E}_g[R_n - Q_n]}_{=0 \text{ by def. in (4.8)}} \\
&\approx \frac{1}{n}\mathbb{E}[(U')^t J^{-1}U'] = \frac{1}{n}\mathbb{E}[(J^{-\frac{1}{2}}U')^t J^{-\frac{1}{2}}U'] \\
&= \frac{1}{n}\mathrm{Var}[J^{-\frac{1}{2}}U'] = \frac{1}{n}\mathrm{Tr}(J^{-\frac{1}{2}}KJ^{-\frac{1}{2}}) = \frac{1}{n}\mathrm{Tr}(J^{-1}K).
\end{aligned}$$

[2] The Taylor expansion of R_n implicitly interchanges integration and derivatives and would again require regularity assumptions to allow that.

This leads us to the bias corrected estimator

$$\hat{Q}_n - \frac{p^*}{n}, \quad \text{with } p^* = \mathrm{Tr}(J^{-1}K). \tag{4.16}$$

Remark 13. In the special case that the true density is contained in the set of candidate models and is attainable by θ_0, i.e. $g = f(\cdot, \theta_0)$, then $J = K$, i.e.

$$J(\theta_0) = -\int_{\mathbb{R}^n} f(y, \theta_0) I(y, \theta_0)\, \mathrm{d}y = \int_{\mathbb{R}^n} f(y, \theta_0) u(y, \theta_0) u(y, \theta_0)^t\, \mathrm{d}y$$

and J is called *Fisher information matrix*. This can be seen by taking the derivative in the first oder condition in 4.11 and rearranging terms. Moreover, p^* in (4.16) simplifies to

$$p^* = \mathrm{Tr}(J^{-1}K) = \mathrm{Tr}(\mathrm{Id}) = \mathrm{length}(\hat{\theta}) = p.$$

This leads to the classic AIC formula by multiplying by $-2n$

$$\mathrm{AIC} = -2n \left(\hat{Q}_n - \frac{p}{n} \right) = -2l_n(\hat{\theta}) + 2p,$$

which coincides with Formula (4.1). In practice, this formula is also used when it is not clear whether the true model is actually attainable.
In his original work, Akaike first proposes a mean discrimination functional of the form

$$I(\theta_1, \theta_0, \varphi) := \int_{\mathbb{R}} f(x, \theta_0) \varphi(\tau(x))\, \mathrm{d}x, \tag{4.17}$$

where φ is a function to be specified and τ is the likelihood ratio of the respective distributions corresponding to the parameters $\theta_0, \theta_1 \in \Theta$:

$$\tau(x) = \frac{f(x, \theta_1)}{f(x, \theta_0)}.$$

The likelihood ratio is commonly used as a test statistic for hypothesis tests on the model parameter θ: Let $0 \leq m \leq d$, $\Theta \subset \mathbb{R}^d$ and $\Theta_0 \subset \Theta$ such that the first m coordinate entries of the parameter vector coincide with a fixed θ_0, i.e.

$$\Theta_0 := \{\theta \in \Theta : \theta^{(1)} = \theta_0^{(1)}, \theta^{(2)} = \theta_0^{(2)}, \ldots, \theta^{(m)} = \theta_0^{(m)}\}$$

Define for a likelihood function $\mathcal{L}$ the quantities

$$L_0 := \sup\{\mathcal{L}(\theta), \theta \in \Theta_0\}, \quad L_1 := \sup\{\mathcal{L}(\theta), \theta \in \Theta\}.$$

Then the test statistic

$$T_n = 2\log\left(\frac{L_1}{L_0}\right)$$

corresponds to the hypothesis test

$$\begin{cases} H_0 : \theta^{(1)} = \theta_0^{(1)}, \theta^{(2)} = \theta_0^{(2)}, \ldots, \theta^{(m)} = \theta_0^{(m)} \\ H_1 : \theta \text{ unrestricted} \end{cases}.$$

Due to Wilks' theorem the test statistic follows a chi-square distribution in the large sample case: $T_n \sim \chi^2_m$ (Wilks 1938, see also p.132 in Young and Smith 2005).
Akaike gives justification for the choice of $\varphi(x) = -2\log(x)$ in Equation (4.17) and thus draws a connection to the Kullback-Leibler distance and subsequently uses Wilks' theorem, Taylor expansions and geometric arguments in the derivation of the IC.
Although the likelihood ratio and its associated hypothesis test could be used as a means to discriminate between models, it has several drawbacks: As mentioned before, likelihood values are merely a measure of fit and do not account for complexity. Moreover, the hypothesis test is limited to nested models which is generally not required for IC model selection. In addition, Burnham and Anderson point out that "[h]ypothesis testing is particularly limited in model selection" (p. 266 in Burnham and Anderson 2004). Instead, they advocate the application of IC for quantification of "strength of evidence" and especially methods of model weighting and averaging (see Section 4.2.3).

Additionally, in the case of small samples, Hurvich and Tsai 1989 proposed a correction of the AIC:

$$\text{AICc} = -2\mathcal{L} + \frac{2kn}{n-k-1}. \tag{4.18}$$

We shall follow the recommendation in Burnham and Anderson 2004 and use the AICc whenever $n < 40k_{\max}$ as a rule of thumb, where $k_{\max}$ is the maximal number of parameters used among the candidate models.

4.2.2 The Bayesian information criterion

The Bayesian information criterion goes back to Schwarz 1978, who applied the approach to exponential families, and is derived by applying a suitable approximation to the posterior probabilities. In the following derivation, we do not require the probability densities to belong to a specific distribution class. Instead, only a certain degree of regularity is assumed. We will slightly change the notation of the previous section to accommodate the Bayesian argument:
Let $y = (y_1, y_2, \ldots, y_n)$ denote the data and $M_1, M_2, \ldots, M_k$ be the list of candidate

models. For a model M_j, $j \in \{1, 2, \ldots, k\}$, there is an associated parameter vector $\theta_j \in \Theta_j$, where $\Theta_j \subset \mathbb{R}^{d_j}$ is the set admissible parameter configurations.
The definition of conditional probability implies

$$\mathbb{P}(y, M_j) = \mathbb{P}(M_j|y)\mathbb{P}(y) = \mathbb{P}(y|M_j)\mathbb{P}(M_j)$$

and rearranging of terms yields the well known Bayes' rule

$$\mathbb{P}(M_j|y) = \frac{1}{\mathbb{P}(y)}\mathbb{P}(y|M_j)\mathbb{P}(M_j). \tag{4.19}$$

In Bayesian statistics, $\mathbb{P}(M_j|y)$ is referred to as *posterior* probability. The unconditional likelihood, normalization constant or *evidence* $\mathbb{P}(y)$ can be written as

$$\mathbb{P}(y) = \sum_{j=1}^{k} \mathbb{P}(y|M_j)\mathbb{P}(M_j) = \sum_{j=1}^{k} \lambda_{n,j}(y)\mathbb{P}(M_j), \tag{4.20}$$

where $\lambda_{n,j}(y) := \mathbb{P}(y|M_j)$ is the marginal likelihood for model M_j. Further conditioning on the parameter vector θ_j yields

$$\lambda_{n,j}(y) := \mathbb{P}(y|M_j) = \int_{\theta_j \in \Theta_j} \mathbb{P}(y|M_j, \theta_j)\, \mathrm{d}\mathbb{P}(\theta_j|M_j). \tag{4.21}$$

For simplicity, we assume that $\mathbb{P}(\cdot|M_j)$ is absolutely continuous and admits a density which will be denoted by $\pi_{n,j}(\cdot|M_j)$. Equation (4.21) then simplifies to

$$\begin{aligned}\lambda_{n,j}(y) := \mathbb{P}(y|M_j) &= \int_{\theta_j \in \Theta_j} \mathbb{P}(y|M_j, \theta_j)\pi(\theta_j|M_j)\, \mathrm{d}\theta_j \\ &= \int_{\theta_j \in \Theta_j} \mathcal{L}_{n,j}(\theta_j)\pi(\theta_j|M_j)\, \mathrm{d}\theta_j. \end{aligned} \tag{4.22}$$

The expression $\mathcal{L}_{n,j}(\theta_j) := \mathbb{P}(y|M_j, \theta_j)$ coincides with the likelihood function in the maximum likelihood setting and $l_{n,j}(\theta_j) := \log(\mathcal{L}_{n,j}(\theta_j))$ can be identified with Equation (4.6).

Lemma 7 (Multidimensional Laplace approximation)**.** Let $\Omega \subset \mathbb{R}^d$. Assume $f \in \mathcal{C}^2(\Omega)$, with unique maximizer $x_0 \in \mathrm{int}\Omega$ and $g \in \mathcal{C}^\infty(\Omega)$. Then

$$\int_\Omega g(x)\mathrm{e}^{nf(x)}\, \mathrm{d}x \approx \mathrm{e}^{nf(x_0)} \left(\frac{2\pi}{n}\right)^{\frac{d}{2}} g(x_0)|D^2 f(x_0)|^{-\frac{d}{2}}$$

(for details see Section IX.5 in Wong 2001).

In order to apply the Laplace approximation to the integral in (4.22), we rearrange

the terms

$$\begin{aligned}\lambda_{n,j}(y) &= \int_{\theta_j \in \Theta_j} \pi(\theta_j|M_j) \exp\left(n\frac{1}{n}\log(\mathcal{L}_{n,j}(\theta_j))\right) \mathrm{d}\theta_j \\ &= \int_{\theta_j \in \Theta_j} \pi(\theta_j|M_j) \exp\left(n\frac{1}{n} l_{n,j}(\theta_j)\right) \mathrm{d}\theta_j.\end{aligned}$$

Since the MLE $\hat{\theta}_j$ maximizes $l_{n,j}$ by definition, we get the approximation

$$\lambda_{n,j}(y) \approx \mathcal{L}_{n,j}(\hat{\theta}_j) \left(\frac{2\pi}{n}\right)^{\frac{d_j}{2}} \pi(\hat{\theta}_j|M_j)|J_{n,j}(\hat{\theta}_j)|^{\frac{1}{2}},$$

where $J_{n,j}(\hat{\theta}_j)$ is the Fisher information matrix associated with the log-likelihood $l_{n,j}$. In particular we have convergence of $J_{n,j}$ for $n \to \infty$ due to the law of large numbers (compare with (4.15) in the previous section). The above result implies

$$\begin{aligned}-2\log(\lambda_{n,j}(y)) &\approx -\underbrace{2l_{n,j}(\hat{\theta}_j)}_{=\mathcal{O}_P(n)} + \underbrace{d_j \log(n)}_{=\mathcal{O}(\log(n))} \\ &\quad - \underbrace{d_j\log(2\pi) - \log(|J_{n,j}(\hat{\theta}_j)|) - 2\log(\pi(\hat{\theta}_j|M_j))}_{=\mathcal{O}_p(1)} \\ &\approx -2l_{n,j}(\hat{\theta}_j) + d_j\log(n) =: \mathrm{BIC}_{n,j},\end{aligned}$$

where we have dropped all lower order terms for $n \to \infty$ in the second step.

4.2.3 Model weighting and model averaging

Similar to the likelihood ratio in Section 4.2.1 in the frequentist context, the Bayesian approach allows a comparison of models using the Bayes factors. If we divide the Bayes formula for the different models M_{j_1} and M_{j_2} we get

$$\underbrace{\frac{\mathbb{P}(M_{j_2}|y)}{\mathbb{P}(M_{j_1}|y)}}_{\text{posterior odds}} = \underbrace{\frac{\mathbb{P}(y|M_{j_2})}{\mathbb{P}(y|M_{j_1})}}_{\text{Bayes factor}} \underbrace{\frac{\mathbb{P}(M_{j_2})}{\mathbb{P}(M_{j_1})}}_{\text{prior odds}}$$

Using the previous approximation, the Bayes factor can be approximated by the ratio of BIC values:

$$\frac{\mathbb{P}(y|M_{j_2})}{\mathbb{P}(y|M_{j_1})} = \frac{\lambda_{n,j_2}(y)}{\lambda_{n,j_1}(y)} \approx \frac{\exp(-\frac{1}{2}\mathrm{BIC}_{n,j_2})}{\exp(-\frac{1}{2}\mathrm{BIC}_{n,j_1})}.$$

If the model probabilities are equal (especially in the case of a uniform prior), the posterior odds coincide with the Bayes factor.

We can re-write the posterior probabilities in Equation (4.19) where $\lambda_{n,j}$ from Equa-

tion (4.22) denotes the marginal likelihood of model j:

$$\mathbb{P}(M_j|y) = \frac{\mathbb{P}(M_j)\lambda_{n,j}(y)}{\sum_{j'=1}^{k}\mathbb{P}(M_{j'})\lambda_{n,j'}(y)}.$$

Substituting the $\mathrm{BIC}_{n,j}$ as an approximation for $-2\log(\lambda_{n,j})$ yields

$$\mathbb{P}(M_j|y) \approx \frac{\mathbb{P}(M_j)\exp(-\frac{1}{2}\mathrm{BIC}_{n,j})}{\sum_{j'=1}^{k}\mathbb{P}(M_{j'})\exp(-\frac{1}{2}\mathrm{BIC}_{n,j'})}. \tag{4.23}$$

The posterior probabilities allow a direct interpretation as the probability of the model M_j given the data y and is conveniently normed on a scale from 0 to 1 by definition. In a similar way, the Akaike weights are defined as

$$w_j^{\mathrm{AIC}} := \frac{\exp(-\frac{1}{2}\mathrm{AIC}_{n,j})}{\sum_{j'=1}^{k}\exp(-\frac{1}{2}\mathrm{AIC}_{n,j'})} = \frac{\mathcal{L}_{n,j}\exp(-\frac{1}{2}d_j)}{\sum_{j'=1}^{k}\mathcal{L}_{n,j'}\exp(-\frac{1}{2}d_{j'})},$$

where $d_j = dim(\theta)$. When we consider the ratio of the Akaike weights for models with the same number of parameters, it reduces to the likelihood ratio as discussed in the context of Wilks' theorem (see p. 92).
In regard to the posterior probabilities in Equation (4.23), if we choose a suitable prior, we are able to recover the Akaike weights. Substituting

$$\mathbb{P}(M_j) = C\exp\left(\frac{1}{2}\mathrm{BIC}_{n,j}\right)\exp\left(-\frac{1}{2}\mathrm{AIC}_{n,j}\right),$$

with norming constant

$$C = \left(\sum_{j=1}^{k}\exp\left(\frac{1}{2}\mathrm{BIC}_{n,j}\right)\exp\left(-\frac{1}{2}\mathrm{AIC}_{n,j}\right)\right)^{-1},$$

in (4.23) we obtain

$$\mathbb{P}(M_j|y) \approx \frac{\exp(-\frac{1}{2}\mathrm{AIC}_{n,j})}{\sum_{j'=1}^{k}\exp(-\frac{1}{2}\mathrm{AIC}_{n,j'})} = w_j.$$

In other words, we are able to interpret the AIC in the Bayesian context.
More generally for any IC of the form

$$\mathrm{IC}_{n,j} = -2\mathcal{L} + c(d_j, n)$$

Buckland, Burnham and Augustin 1997 propose corresponding weights

$$w_j^{\mathrm{IC}} = \frac{\exp(-\frac{1}{2}\mathrm{IC}_{n,j})}{\sum_{j'=1}^{k}\exp(-\frac{1}{2}\mathrm{IC}_{n,j'})}.$$

The weights cannot only be used for direct comparison between candidate models, but are also useful for inference of key quantities. The idea is to design estimators for quantities of interest $\hat{\mu}$, e.g. mean or variance, as a weighted average of estimates $\hat{\mu}_j$ coming from each individual model M_j, i.e.

$$\hat{\mu} = \sum_{j=1}^{k} w_j \hat{\mu}_j.$$

According to Burnham and Anderson 2004 such multimodel inference has proven to be more accurate than selecting a single "best" model first and ignoring the risk of having selected an unsuitable model for the subsequent inference.

4.2.4 The consistency property

Similar to the consistency property of the MLE, it is a desirable property to have the IC selecting the correct model order with high probability when the underlying sample size increases. To be more precise:

Definition 17. Let n be the underlying sample size, $\mathcal{J}$ be the set of models among all competing models that minimize the Kullback-Leibler distance to the true model and let $\mathcal{J}_0 \subset \mathcal{J}$ be the subset of models with minimal (parameter) dimension. Then, an IC is said to be *consistent* if there is a $j_0 \in \mathcal{J}_0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{\min_{l\in\mathcal{J}\setminus\mathcal{J}_0}(\mathrm{IC}(M_{j_0}) - \mathrm{IC}(M_l)) > 0\right\} = 1, \tag{4.24}$$

i.e. the probability that the IC will choose a model with smallest dimension minimizing the Kullback-Leibler distance converges to 1.
An IC is *strongly consistent* if the assertion in (4.24) holds almost surely:

$$\mathbb{P}\left\{\min_{l\in\mathcal{J}\setminus\mathcal{J}_0}(\mathrm{IC}(M_{j_0}) - \mathrm{IC}(M_l)) > 0, \quad \text{for almost all } n\right\} = 1 \tag{4.25}$$

Remark 14. The above definition follows the notation in Claeskens and Hjort 2008, p. 101, but the original proof of sufficient conditions for consistency and strong consistency are shown in Sin and White 1996, (where consistency actually goes under the name of weak consistency).

The HQ is designed to have the slowest growing penalty term that still renders

the IC to be strongly consistent (see later for a more precise definition). The proof makes use of the law of iterated logarithm. Besides, the HQ was originally defined more generally as

$$\mathrm{HQ}' = -2\mathcal{L} + 2ck\ln(\ln(n)), \quad c > 1, \tag{4.26}$$

but c was chosen to be 1 in a subsequent example. Claeskens and Hjort 2008 point out that the choice of c is not clear and renders the information criterion less relevant for practitioners.

As a matter of fact, the AIC fails to be consistent as the penalty term does not depend on the sample size. The asymptotic distribution of the associated model selection was analyzed for autoregressive models for example in Shibata 1976. BIC and HQ on the other hand are found to be strongly consistent. As a consequence, their asymptotic distribution of model selection is bound to converge to a delta on the most parsimonious Kullback-Leibler minimizing model. The respective convergence rates for AIC and BIC were analyzed in Zhang 1993 for another regression model.

Consistency from a practical perspective From the previous section one might conclude that the non-consistent AIC would be inferior to the consistent BIC and HQ. However, the situation is more complicated: We have to keep in mind that consistency is an asymptotic property. This means that in theory the consistent IC will eventually outperform the AIC for almost all cases if the sample size is sufficiently large. Unfortunately, practitioners just have a limited amount of data available and it is very difficult to judge whether the sample size belongs to the asymptotic region. Indeed, empirical studies suggest for various statistical models that the AIC outperforms the BIC in small sample cases [3]: As an example among regression models, Hurvich and Tsai 1989; Hurvich and Tsai 1990 compared different IC on simulated data especially to promote the (still inconsistent) AICc as a modification of the AIC for smaller samples. More recently, Javed and Mantalos 2013 applied IC (AIC, BIC, HQ, AICc) in a MC simulation of (nonlinear) GARCH models. Their results suggest that the AIC outperforms the BIC and HQ for higher-order GARCH processes.

As a consequence of the above discussion, we can make the idea and objective of our Monte Carlo experiment in the next chapter more precise: First, we need to point out that the numerical results of the simplistic setting of our Monte Carlo experiment do not directly translate to how empirical data should be handled. IC

[3] By "small samples" we refer to the situation that the sample size is not sufficiently large enough for the asymptotic consistency results to hold, but large enough such that effects similar to the paradox discussed by Freedman 1983 can be safely excluded.

are one of many tools for model-selection and cross-validation. We do not expect to find a "best" IC, but rather want to verify the theoretical properties of the different IC for point process models. In particular, due to the fact that most theoretical results have been derived for regression models only, our work may help to shed light on asymptotic regions and convergence rates of consistent IC and the asymptotic distribution of selected orders of the AIC for this model class.

4.3 Summary

In this chapter, we have introduced information criteria for model selection. In particular, we have discussed the idea and derivation of the AIC and BIC as well as their connection. The concept of consistency lays the theoretical foundation for the application of IC in the next chapter, where we use simulated data to verify the asymptotic behaviour of the IC for growing sample size.

Chapter 5

Models for durations between trades and model selection

5.1 The Monte-Carlo setup

In this chapter, we will discuss three models for durations between trades in high-frequency financial data: a compound Poisson type model with time-varying deterministic intensity, the exponential ACD model and Hawkes processes. In particular, we will briefly describe simulation and estimation methods for each model class.
We will then perform the following Monte-Carlo experiment for each model: First, we simulate data from the respective model, where we use parameter sets and sample sizes that reflect typical observations in empirical studies of financial data whenever it is appropriate. This follows the advice given in Burnham and Anderson 2004. In order to test the performance of information criteria from Chapter 4 in finding the correct order of a model, we then fit several candidate models of different order from the same model class as the true underlying model via maximum likelihood estimation (MLE).
Although we are primarily interested in the performance of model selection, we must make sure that the MLE gives reasonable results. As we can see from the IC formulas in Definition 15, the IC are based on the maximized likelihood value. Therefore, we may conclude that there is a close connection between the quality of the MLE and the subsequent model selection result. In any case, a poor MLE due to numerical problems or lack of data is likely to compromise the model selection. For example, a correctly selected model order can be meaningless if the estimated model itself fails to describe and predict key features or quantities of the data we are interested in.
We ensure that the estimates are reasonably good by using the mean squared error (MSE) as a measure for the goodness of fit: Let $\theta \in \mathbb{R}$ be a generic model para-

meter to be estimated and $\hat{\theta}$ the corresponding estimator. Given N samples and $\hat{\theta}^{(k)}$, $k = 1, \ldots, N$, the estimates for each sample, the MSE can be calculated as

$$\mathrm{MSE}(\theta) = \mathbb{E}\left[|\theta - \hat{\theta}|^2\right] = \frac{1}{N}\sum_{k=1}^{N} |\theta - \hat{\theta}^{(k)}|^2 \tag{5.1}$$

(see Czado and Schmidt 2011). The root mean squared error is given by

$$\mathrm{RMSE}(\theta) = \sqrt{\frac{1}{N}\sum_{k=1}^{N} |\theta - \hat{\theta}^{(k)}|^2} \tag{5.2}$$

and the relative root mean squared error by

$$\mathrm{RMSE}_{\mathrm{rel}}(\theta) = \frac{1}{\theta}\sqrt{\frac{1}{N}\sum_{k=1}^{N} |\theta - \hat{\theta}^{(k)}|^2}. \tag{5.3}$$

It is easy to calculate the above quantities as the true model values are known in our mock data setting.
After assessing goodness-of-fit, we calculate the information criteria for model selection. As the model underlying the data is known, we can easily calculate the success rate of each IC, which allows a direct comparison.

5.2 A normal compound Poisson model

This model follows the idea of a locally stationary model for tick-by-tick data. It can be seen as a simplified form of the model proposed in Scalas 2007 and mimics the U-shaped trade intensity often observed in intra-day trading data[1] (see for instance Bertram 2004).

5.2.1 Definition

The compound Poisson model with discrete intensity (Dλ)-model

We suppose that high-frequency data is given over a time interval $[t_0, T]$. First, set a time grid $\{t_i\}_{i\in\{1,\ldots,n\}}$ such that $t_0 < t_1 < t_2 < \ldots < t_n = T$. On each time interval $[t_{i-1}, t_i)$ we have a compound Poisson process

$$X_i(t) := \sum_{k=1}^{N_i(t)} R_k^{(i)}, \tag{5.4}$$

[1] In a more extensive analysis, it is observed that the intra-day seasonalities depend on the market, the institutional setting and the time zone (see Section 3.4 in Hautsch 2012).

where $\{R_k^{(i)}\}_{k\in\mathbb{N}}$ is an i.i.d. sequence of $N(\mu_i, \sigma_i^2)$ distributed random variables and $(N_i(t))_{t\geq 0}$ is a homogeneous Poisson process with parameter λ_i. Further, $\{R_k^{(i)}\}_{k\in\mathbb{N}}$ are all independent of $(N_i(t))_{t\geq 0}$.
For a fixed time interval $[t_{i-1}, t_i)$ the log-likelihood function is given by

$$\mathcal{L}_i^D(\lambda_i, \mu_i, \sigma_i) = -\lambda_i(t_i - t_{i-1}) + \ln(\lambda_i)N_i(t_i) + \sum_{k=1}^{N_i(t_i)} \ln(p_{\mu_i,\sigma_i}(R_k^{(i)})), \tag{5.5}$$

where p_{μ_i,σ_i} denotes the probability density function of the $N(\mu_i, \sigma_i^2)$ distribution. Due to the independence assumptions the overall log-likelihood is given by the sum of all $\mathcal{L}_i$. Equation (5.5) can be derived from the general expression for the sample density function given on page 200 in Snyder and Miller 1991 by substituting a constant λ.
The maximum likelihood estimators are therefore:

$$\hat{\lambda}_i = N_i/w_i, \quad \hat{\mu}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} r_i, \quad \hat{\sigma}_i^2 = \frac{1}{N_i}\sum_{k=1}^{N_i}(r_i - \hat{\mu}_i)^2, \tag{5.6}$$

where N_i is the number of trades in the ith interval and $w_i = t_i - t_{i-1}$.
Note that the maximum likelihood estimator for σ^2 is biased and the bias can be corrected by using

$$\tilde{\sigma}_i^2 = \frac{1}{N_i - 1}\sum_{k=1}^{N_i}(r_i - \hat{\mu}_i)^2 \tag{5.7}$$

instead. We shall use either the biased or unbiased estimator in the following sections when appropriate.

The compound Poisson model with parametrized intensity (Pλ)-model

This model will be used for simulation later on as well as serve as a benchmark model when testing model selection criteria. As empirical results about the trading intensity suggest a daily seasonality, this model assumes that the step function in the (Dλ) model is parametrized by a quadratic function:

$$\lambda_{a,b,c}(t) = at^2 + bt + c, \quad t \in [0, 1]. \tag{5.8}$$

Of course, this parametrization can be easily replaced by a different function. Since λ needs to be positive and convex, we also have the conditions

$$a > 0 \text{ and } c > \frac{b^2}{4a}. \tag{5.9}$$

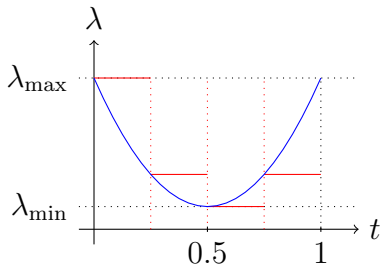

Figure 5.1: Deterministic intensity function (blue) and a step function approximation (red)

Similar to the (Dλ)-model, the log-likelihood for the (Pλ)-model is given by

$$\mathcal{L}_i^P(a,b,c,\mu_i,\sigma_i) = -\lambda_{a,b,c}(t_{i-1})(t_i - t_{i-1}) + \ln(\lambda_{a,b,c}(t_{i-1}))N_i(t_i) + \sum_{k=1}^{N_i(t_i)} \ln(p_{\mu_i,\sigma_i}(R_k^{(i)})). \tag{5.10}$$

While the maximum likelihood estimators for μ_i and σ_i are the same as for the (Dλ) case, the maximum likelihood estimators for a, b, c, which determine the form of λ, cannot be obtained in closed form. As a consequence, a numerical optimization method needs to be applied to estimate those parameters.

5.2.2 Simulation

The simulation algorithm essentially uses the (Pλ)-model. For simplicity we will choose the time interval $[t_0, T]$ to be $[0, 1]$. For the simulation we set an equidistant grid $0 = t_0 < t_1 < t_2 < \ldots < t_n = 1$ on the time interval. Thus, the interval $[0, 1]$ is divided into n subintervals. For $i \in \{1, \ldots, n\}$ the parameters μ_i, σ_i and λ_i on the subinterval $[t_{i-1}, t_i)$ are chosen to be

$$\begin{gathered} \mu_i = 0,\ \sigma_i = 1 \text{ and } \lambda_i = \lambda(t_{i-1})\ \forall i \in \{1, \ldots, n\}, \\ \text{where } \lambda(t) := 4(\lambda_{\max} - \lambda_{\min})(t - 0.5)^2 + \lambda_{\min}, \\ \forall t \in [0, 1] \text{ and } \lambda_{\min}, \lambda_{\max} > 0 \text{ constant.} \end{gathered} \tag{5.11}$$

The functional form of λ is inspired by the empirical findings in the previous sections and should account for the observed seasonality in a simple way. The form of λ can be easily replaced by more complex functions. We have chosen $\lambda_{\min} = 100$ and $\lambda_{\max} = 10000$. Note that the $\{\lambda_i\}$ form a step function approximation of the parabola in Equation (5.11), which is also depicted in Figure 5.1. For different grid sizes, we simulate with sample size 1000 each.

5.2.3 Fitting

The parameter estimation will be carried out using different grid sizes. Note that the grid size to be used in fitting is bounded from above by the length of the entire time interval (in our case 1). However, we would like to emulate the behavior of the intensity which was observed in empirical data, i.e. high intensity at the beginning and at the end of the trading day and relatively low intensity in the middle of the day. Consequently, we need at least 3 subintervals to have a piecewise constant function that fulfills these conditions on the time interval. Further, the smallest eligible grid size is bounded from below by the maximal distance between neighboring data points within the data set. Otherwise, there are subintervals which do not contain any data points. In such cases, the estimation formulas in (5.6) would fail.
More precisely, for the maximal distance $\Delta_{\max}$ between two consecutive data points within a given sample, the finest valid equidistant grid has at most $\left\lfloor \frac{1}{\Delta_{\max}} \right\rfloor$ subintervals. Therefore, we will consider a list of candidate models on grids which correspond to $n = 3, 4, \ldots, \left\lfloor \frac{1}{\Delta_{\max}} \right\rfloor$ subintervals on the interval $[0, 1]$.

For the (Dλ) model, the estimators are given in closed form in (5.6) and the likelihood value is easily calculated via Equation (5.5) and subsequently used for the calculation of the IC. We decide to use the biased estimator $\hat{\sigma}_i^2$: Since we are mainly interested in model selection, we would like to ensure that we work with the optimal value of the log-likelihood when calculating the IC.
In order to fit the (Pλ) model, we assume that the estimates for $\{\mu_i\}, \{\sigma_i\}$ and $\{\lambda_i\}$ for the (Dλ)-algorithm are already calculated and can be used as an input for the estimation of the (Pλ)-model. As mentioned previously, the estimators for μ_i and σ_i coincide in both models and no further calculation is needed for these parameters. It remains to solve the following minimization problem:

$$(\hat{a}, \hat{b}, \hat{c}) = \underset{a,b,c \in \mathbb{R}}{\arg\min} \left[-\sum_{i=1}^{n} \mathcal{L}_i^P(a, b, c, \mu_i, \sigma_i) \right]$$
$$\text{s.t.} \quad a > 0 \text{ and } c > \frac{b^2}{4a} \tag{5.12}$$

A reasonable choice of the starting value for the minimization algorithm can be easily obtained by the least-squares fit of the parabola to the $\{\lambda_i\}$ values of the (Dλ) case, which already gives a fairly good approximation of the parabola. In case the initial values obtained by this method do not lie in the admissible set, a change of signs for a or a shift of the parabola may be applied.
Note that the estimation of the (Pλ)-model requires a grid with at least 4 grid points, i.e. 3 subintervals on which $\lambda_1, \lambda_2, \lambda_3$ are estimated using the (Dλ)-model.

This ensures that the parabola is well determined. However, as mentioned before, this condition is not restrictive and covers all models on which we would like to run model selection.

5.2.4 Numerical results

Goodness of fit

Depending on the number of subintervals the total number of parameters can be quite large. Thus it is difficult compare the MSE separately for parameters. Therefore, we make a slight modification to the MSE formula: The distance in Equation (5.1) is understood as a functional distance. To be more precise, we choose the L^2-distance between the true step function intensity and the estimated one:

$$\mathbb{E}\left[|\theta - \hat{\theta}|^2\right] = \mathbb{E}\left[\|\theta - \hat{\theta}\|_{L^2}^2\right] \tag{5.13}$$

The cases of μ and σ^2 are the easier ones, as we just need to calculate the distance between a step function and a constant: For the step functions with values $\{\mu_i\}$ on the fitting grid $t_1 < t_2 < \ldots < t_n$ Equation (5.13) can be further written as

$$\begin{aligned}
\mathbb{E}\left[\|\mu - \hat{\mu}\|_{L^2}^2\right] &= \frac{1}{N}\sum_{k=1}^{N} \|\mu - \hat{\mu}^{(k)}\|_{L^2}^2 \\
&= \frac{1}{N}\sum_{k=1}^{N} \int_0^T (\mu(t) - \hat{\mu}^{(k)}(t))^2 \mathrm{d}t \\
&= \frac{1}{N}\sum_{k=1}^{N}\sum_{i=2}^{n} (\mu - \hat{\mu}_i^{(k)})^2 (t_i - t_{i-1}).
\end{aligned} \tag{5.14}$$

and in the same way for σ^2.

Concerning the intensity function, we have to merge the simulation grid $t_1^s < t_2^s < \ldots < t_m^s$ with the fitting grid $t_1^f < t_2^f < \ldots < t_r^f$. After reordering and relabeling, we can calculate the MSE on the merged grid $t_1 < t_2 < \ldots < t_n$ via

$$\mathbb{E}\left[\|\lambda - \hat{\lambda}\|_{L^2}^2\right] = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=2}^{n} (\lambda_i - \hat{\lambda}_i^{(k)})^2 (t_i - t_{i-1}). \tag{5.15}$$

The numerical results that we present as an example are for $N = 1000$ samples of data simulated from a grid containing 30 subintervals: Table 5.1 shows summary statistics of μ and σ^2, where the summary statistics were calculated over the set of fitting grids. The MSE for the μ and σ^2 are comparably small.

For the intensity function λ we plot the MSE against the number of subintervals used for fitting in Figure 5.2. Starting from a small number of subintervals, the

MSE decreases sharply before it reaches its optimum at 30, the true number of subintervals from the simulation. Number of subintervals above 30 give a larger MSE and, in the case of the (Dλ) model, instabilities of over parametrization even lead to an increasing MSE.

Table 5.1: Table of summary statistics of the MSE of the parameters μ and σ^2 of the compound Poisson type model. The analysis is based on 1000 samples generated from a simulation grid containing 30 subintervals.

	mean	min	max	std
μ	0.0545	0.0026	0.1049	0.0212
σ^2	0.1038	0.0049	0.1757	0.0439

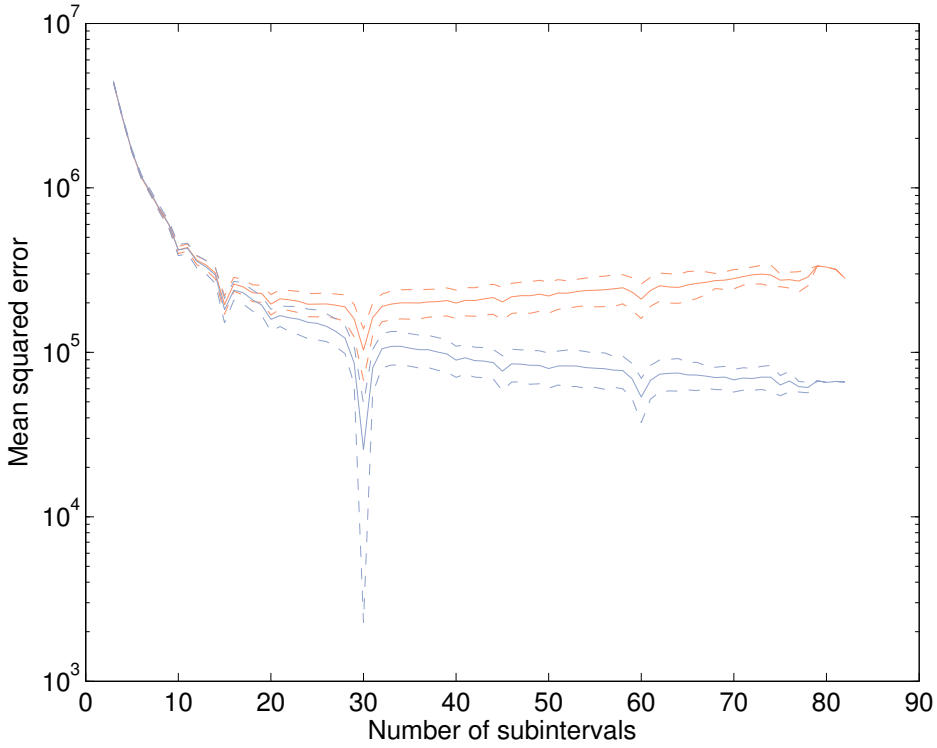

Figure 5.2: Plot of the mean squared error (MSE) of the estimation of the intensity function for the (Dλ) model (orange lines) and for the (Pλ) model (blue lines) respectively. The graph shows the MSE together with dashed lines indicating the size of the first standard deviation from the mean as a function of the underlying number of intervals of the fitting grid. The true model for the simulation originally used 30 subintervals. The MSE is calculated as a squared L^2 distance between the estimated and the true intensity function (see also Eq. (5.15)).

Concerning goodness of fit, we can see that the MSE of the (Pλ) model is consistently smaller than the MSE of the (Dλ) model. This is to be expected as, by construction of the experiment, the (Pλ) model is the true model and gives a better fit to the data.

Moreover, we can observe that apart from the optimum at 30 there are "preferred"

numbers of subintervals at 10, 20, 45, 60. This is crucial for the explanation of the behavior of model selection as the relationship between goodness of fit and number of subintervals in the region below the optimal number is not monotone.

The size of the MSE can be estimated from the expected fluctuations of the estimator $\hat{\lambda}$. The MSE can be estimated from below by means of the ideal situation when the simulation and fitting grid are identical. Without loss of generality, we assume an equidistant simulation grid with grid size $w = t_i - t_{i-1}$ and rewrite Equation (5.15):

$$\begin{aligned}\mathbb{E}\left[\|\lambda - \hat{\lambda}\|_{L^2}^2\right] &\geq w \sum_{i=2}^{n} \mathbb{E}\left[(\lambda_i - \hat{\lambda}_i)^2\right] \\ &= w \sum_{i=2}^{n} \operatorname{Var}\left[\hat{\lambda}_i\right] = \frac{1}{w} \sum_{i=2}^{n} \operatorname{Var}\left[N_i\right], \end{aligned} \tag{5.16}$$

where we have used the definition of the estimator in (5.6) and that the number of events in an interval of size w is Poisson distributed: $N_i \sim \mathrm{Poi}(\lambda w)$. We finally get that

$$\begin{aligned}\mathbb{E}\left[\|\lambda - \hat{\lambda}\|_{L^2}^2\right] &\geq \frac{1}{w} \sum_{i=2}^{n} \operatorname{Var}\left[N_i\right] \\ &= \frac{1}{w} \sum_{i=2}^{n} \lambda_i w \approx \frac{1}{w} \int_0^1 \lambda(t)\, \mathrm{d}t, \end{aligned} \tag{5.17}$$

where we approximate the integral of the step function by the integral of the smooth intensity parametrization in Equation (5.11). For our numerical example we have $\frac{1}{w} = 30$ and $\lambda_{\min} = 100$ and $\lambda_{\max} = 10000$. An explicit calculation of above integral gives the rough estimate

$$\mathbb{E}\left[\|\lambda - \hat{\lambda}\|_{L^2}^2\right] \gtrsim 30 \cdot 3400 = \mathcal{O}(10^5), \tag{5.18}$$

which is of about the same order of magnitude observable in Figure 5.2.

Model selection

Figures 5.3, 5.4 and 5.5 show box plots of the model selection results of the AIC, BIC and HQ respectively. In each box plot, the orange and blue box plot correspond to the results of the (Dλ)- and (Pλ)-model respectively. The horizontal axis shows the number of subintervals used in the simulation grid. On the vertical axis are the selected number of parameters after the parameter estimation of the (Dλ)- and (Pλ)-models using different discretizations of $[0, 1]$. A single box in the box plots extends from the 25th percentile to the 75th percentile and the dot indicates the median.

The whiskers have a maximum length of 1.5 times the box length and extend to the outermost point which is not considered as outlier. The crosses indicate outliers.

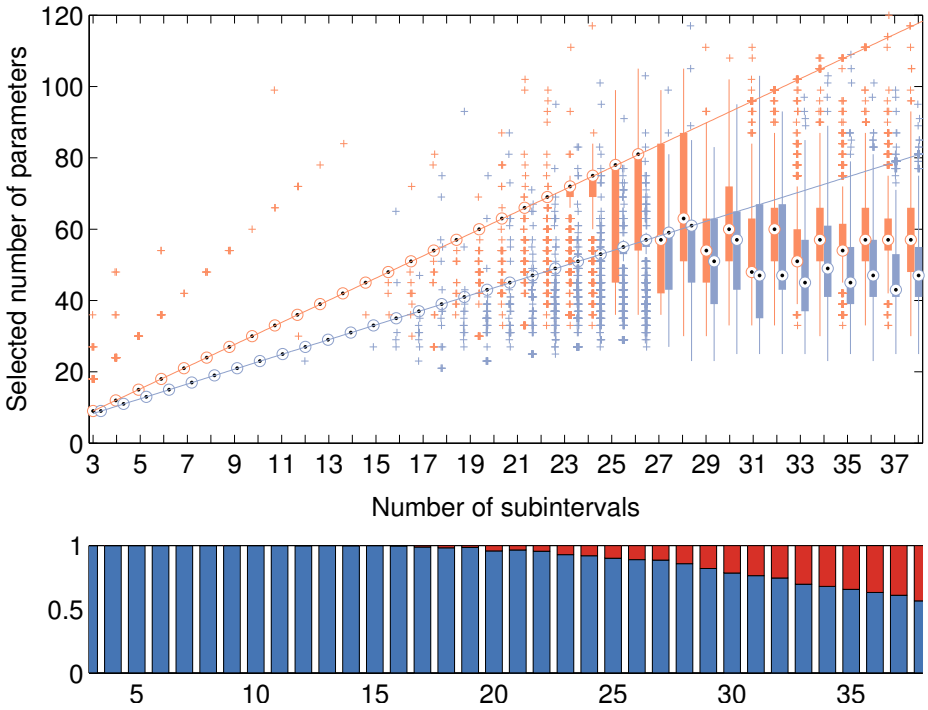

Figure 5.3: The lower plot shows the ratio of samples which allow the true model to be among the set of models from which the IC may choose from, in other words there is no misspecification (blue areas). This ratio decreases and for finer discretization there are more cases of model selection under misspecification (red areas). The sum of blue and red areas is 100%.
The upper plot shows that the model selection using the AIC for the (Dλ)-model (orange box plot) closely follows the reference line indicating $3n$ (n = number of subintervals) for small n, before deviating for larger n. The same holds for the (Pλ)-model (blue box plot) and its corresponding reference line $2n + 1$. The number of subintervals for which both box plots deviate from their respective reference lines is around $n = 25$ to $n = 27$. In the region $n < 15$, there are several outliers which are almost all overestimates.

Below the box plots, bars indicate the ratio of samples which allow model selection under correct specification (blue) and under misspecification (red): In our setting, we speak of model selection under misspecification if the correct model is not contained in the set of selectable models and cannot be chosen by the IC. If this is not the case, i.e. the correct model can potentially be chosen by the IC, we call it model selection under correct specification.

The results for the (Dλ) and (Pλ) model are very similar. Common for all three IC is that for small parameter numbers below 15 the model selection works well: the distributions of the selected orders are concentrated and closely follow the $3n$

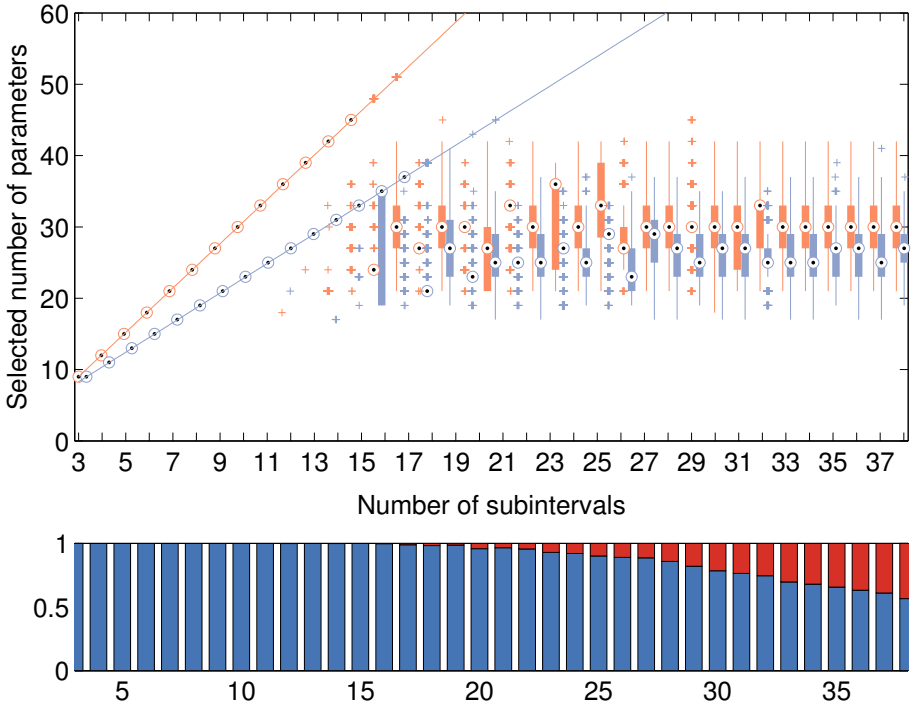

Figure 5.4: The lower plot shows the ratio of samples which allow the true model to be among the set of models from which the IC may choose from, in other words there is no misspecification (blue areas). This ratio decreases and for finer discretization there are more cases of model selection under misspecification (red areas). The sum of blue and red areas is 100%.
The upper plot shows that the model selection using the BIC for the (Dλ)-model (orange box plot) closely follows the reference line indicating $3n$ (n = number of subintervals) for small n before deviating for larger n. The same holds for the (Pλ)-model (blue box plots) and its corresponding reference line $2n + 1$. The number of subintervals for which both box plots deviate from their respective reference lines is around $n = 15$ to $n = 17$.

or $2n + 3$ reference line respectively, where n is the number of subintervals. For very large parameter numbers one can observe that the selected model orders remain distributed around a maximum model order and stop to follow the linear trend of the reference line. This is rather due to a limitation of our Monte-Carlo setup than an inherent property of the IC: As described in Section 5.2.2, we only work with equidistant grids when applying the model selection procedure. The finest grid which can be used for fitting is determined by the maximal distance $\Delta_{\max}$ between two consecutive points within a sample. On the other hand, $\Delta_{\max}$ is related to the minimal value of λ in the middle of the interval, depending on how small we choose the simulation grid size Δ_{sim}. This means that whenever $\Delta_{\max} > \Delta_{\text{sim}}$, the true model is not contained in the pool of models from which the IC may choose from. In other words, we have a case of model selection under misspecification. The bar plots show that first cases occur at around $n = 20$ and go up to a ratio of about 50% for the finest grid in the analysis.

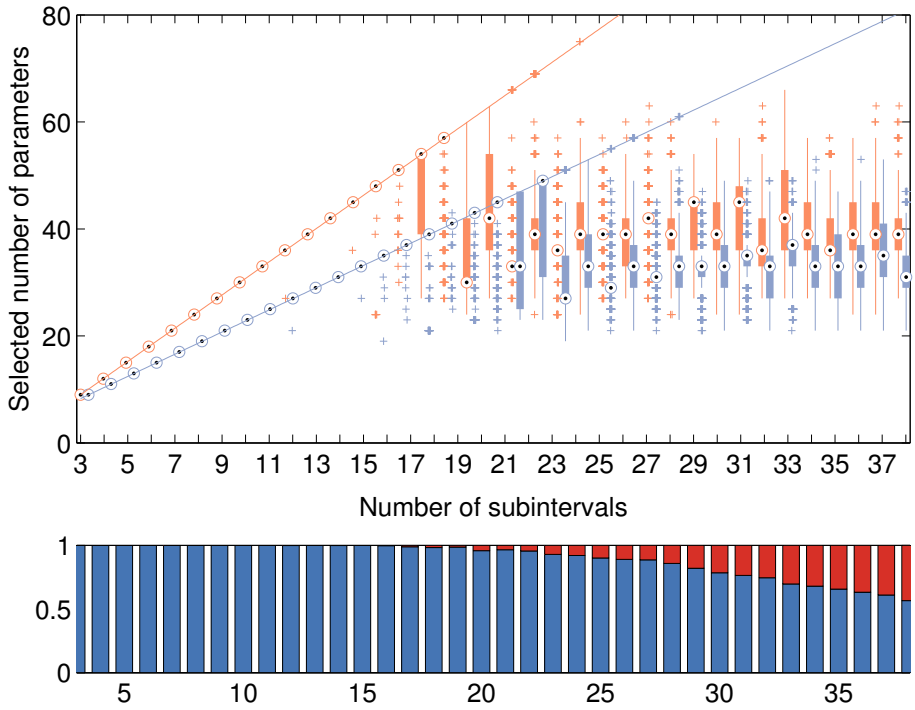

Figure 5.5: (Color online) The lower plot shows the ratio of samples which allow the true model to be among the set of models from which the IC may choose from, in other words there is no misspecification (blue areas). This ratio decreases and for finer discretization there are more cases of model selection under misspecification (red areas). The sum of blue and red areas is 100%.
The upper plot shows that the model selection using the HQ for the (Dλ)-model (orange box plot) closely follows the reference line indicating $3n$ (n = number of subintervals) for small n before deviating for larger n. The same holds for the (Pλ)-model (blue box plots) and its corresponding reference line $2n + 1$. The number of subintervals for which both box plots deviate from their respective reference lines is around $n = 18$ to $n = 20$.

Another look at Figure 5.2 hints that the general rule "the more parameters, the better the fit" is not entirely true: we can observe that the relation between grid size and MSE is not entirely monotone. This is due to the fact that the fit of the specific model does not only depend on the number of parameters, but also to some extent on the position of the grid. As a consequence, under misspecification, the selected order does not necessarily correspond to the finest available grid size above Δ_{sim}. This might explain the "plateaus" on the model selection results for large parameters.

Between the region of very small and very large parameters the IC exhibit quite different behaviors according to their intrinsic tendency of under- and overfitting, which will be described in the following:
The AIC tends to overestimate the number of parameters. It allows outliers (in the region of $n \leq 22$) as well as a larger number of cases of the model selection

to lie above the reference line (in the region of $n \geq 23$). In contrast, the selected model orders of the BIC and HQ are either on the reference line or strictly below the reference line. In other words BIC and HQ tend to underestimate. Additionally, we can see that, starting from around $n = 25$ to $n = 27$, the boxplot for the AIC deviates from the reference line and the BIC and HQ deviate earlier around $n = 15$ and $n = 20$ respectively. Especially, for $n < 27$ the underestimation in the BIC and HQ case is not attributable to the behavior of model selection under misspecification, as the ratio of model selection under misspecification is rather low. Based on our results, if the IC were to be ordered by their parsimonious character, the BIC would be the most parsimonious whereas the AIC the least.
The above observations show that the model selection using any of the three IC works quite well as long as the true model is actually retrievable. The AIC tends to overestimate, but the model selection results are closest to the reference line of true parameters compared to the other two IC.

5.3 The autoregressive conditional duration (ACD) model

The autoregressive conditional duration model was first proposed by Engle and Russell 1998. We will consider a model for the durations between events only, i.e. without marks:

Definition 18. Let $(\varepsilon_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with mean 1. The autoregressive conditional duration (ACD) model is defined as

$$x_i = \psi_i \varepsilon_i \tag{5.19}$$

$$\psi_i \equiv \psi_i(x_{i-1}, \ldots, x_1; \theta) := \mathbb{E}\left[x_i | x_{i-1}, \ldots, x_1\right]. \tag{5.20}$$

If we assume a point process $(N(t))_{t\geq 0}$ with arrival times $t_1, t_2, \ldots$ to correspond to an ACD process, then the relation between durations x_n, $n \in \{1, 2, \ldots\}$ and the arrival times is given by $x_n = t_{n+1} - t_n$, $t_0 = 0$.
The conditional intensity of the ACD model can be derived using the construction seen in Section 1.1 on page 15. The hazard function for ε with probability density function p_ε is given by

$$\lambda_\varepsilon(t) = \frac{p_\varepsilon(t)}{\mathbb{P}(\varepsilon > t)} = \frac{p_\varepsilon(t)}{\int_t^\infty p_\varepsilon(\tau)\mathrm{d}\tau}, \quad t \geq 0$$

and is referred to as *baseline hazard* in Engle and Russell 1998. Using the relation

in (5.19) and the fact that $(\varepsilon_i)_{i\in\mathbb{N}}$ are i.i.d., we obtain for $t \in [t_n, t_{n+1})$

$$\mathbb{P}(x_{n+1} \leq t | t_1, \dots t_n) = \mathbb{P}\left(\varepsilon_{n+1} \leq \frac{t}{\psi_{n+1}}\right).$$

Writing the above expression in terms of the probability densities $p_{x_{n+1}}$ and p_ε respectively yields

$$\int_0^t p_{x_{n+1}}(\tau | t_1, \dots, t_n)\, \mathrm{d}\tau = \int_0^{\frac{t}{\psi_{n+1}}} p_\varepsilon(\tau)\, \mathrm{d}\tau.$$

Differentiating w.r.t. t we get

$$p_{x_{n+1}}(t | t_1, \dots, t_n) = \frac{1}{\psi_{n+1}} p_\varepsilon\left(\frac{t}{\psi_{n+1}}\right).$$

Plugging this into the formula for the hazard function (see Equation (1.2) on p. 17) yields

$$\begin{aligned} h_{n+1}(t | t_1, \dots, t_n) &= \frac{p_{x_{n+1}}(t | t_1, \dots, t_n)}{\mathbb{P}(x_{n+1} > t | t_1, \dots, t_n)} = \frac{1}{\psi_{n+1}} \frac{p_\varepsilon\left(\frac{t}{\psi_{n+1}}\right)}{\mathbb{P}\left(\varepsilon > \frac{t}{\psi_{n+1}}\right)} \\ &= \frac{1}{\psi_{n+1}} \lambda_\varepsilon\left(\frac{1}{\psi_{n+1}}\right), \quad t \in [t_n, t_{n+1}), \end{aligned}$$

which piecewisely defines the conditional intensity function as seen in Section 1.1. The choice of ψ and the distribution of ε specifies the kind of ACD process. For the remainder of this chapter, we assume $\varepsilon \sim \mathrm{Exp}(1)$ which simplifies the formula for the hazard function to

$$h_{n+1}(t | t_1, \dots, t_n) = \frac{1}{\psi_{n+1}}.$$

We will work with the following representation for ψ_i:

$$\psi_i := \omega + \sum_{j=0}^{m} \alpha_j x_{i-j} + \sum_{k=0}^{q} \beta_k \psi_{i-k}, \tag{5.21}$$

where $\omega > 0$, $\alpha_i \geq 0$ and $\beta_i \geq 0$ for all i. We will call this model ACD(m, q).
The ACD model is closely related to the GARCH model proposed by Bollerslev 1986 as a volatility model that allows clustering. Indeed, it was stated that the ACD model "surprisingly turns out to be isomorphic to the GARCH model" (p. 428 in Engle 2002). Sufficient stationarity conditions are given by

$$\sum_{i=1}^{m} \alpha_i + \sum_{k=1}^{q} \beta_k < 1,$$

which is also the case for an analogous GARCH(m, q) model. For proof of weak and strict stationarity conditions for GARCH models which translate to ACD see Bollerslev 1986, Bougerol and Picard 1992, Bollerslev, Engle and Nelson 1994, Francq and Zakoïan 2010.
For given duration data $\{x_1, \ldots, x_n\}$ the log-likelihood function is given by

$$\mathcal{L}^{\text{ACD}}(\omega, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_q) = -\sum_{i=1}^{n} \left[\ln \psi_i + \frac{x_i}{\psi_i} \right] \tag{5.22}$$

(see p. 104 in Hautsch 2012).

5.3.1 Simulation and fitting

For both simulation and MLE of ACD models we use the `R` package `ACDm` written by Belfrage 2016. The model selection analysis for the ACD model follows the Monte-Carlo experiment conducted in Javed and Mantalos 2013. We consider model orders $m, q \in \{1, 2\}$ and Table 5.2 shows the choice of parameters for the simulation.

Table 5.2: Parameter settings for the simulation of ACD data

	ω	α_1	α_2	β_1	β_2
ACD(1,1)	1	0.089	–	0.85	–
ACD(1,2)	1	0.1	–	0.45	0.4
ACD(2,1)	1	0.15	0.15	0.65	–
ACD(2,2)	1	0.1	0.1	0.42	0.35

5.3.2 Numerical results

Goodness-of-fit

In the ACD case we have a simple parameter vector $(\omega, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_q) \in \mathbb{R}^{1+m+q}$. Therefore, we can use the formula given in Equation (5.1) for each scalar valued parameter. The results can be seen in Table B.1. The largest sample size ensures that the MSE are comparably low for each model. The largest contribution to the MSE comes from the ω parameter. A closer look shows that the MSE of the β parameter(s) is of different order depending on the model order q. In the case $q = 1$, the MSE of the β parameter is of the same size as the α parameter(s). However, in the case of $q = 2$, the order of the MSE of the β parameters are significantly larger than the MSE of the α parameters (by a factor of 10 in the ACD$(1, 2)$ case and by a factor of 100 in the ACD$(2, 2)$ case).

Model selection

The results of the model selection experiment can be found in Tables B.2 to B.5. The numbers are success rates in percent of the respective IC in selecting the correct model from which the simulation data was generated from. The qualitative behavior of the IC are unsurprisingly similar to the findings for the GARCH model in Javed and Mantalos 2013.
A closer look at Table B.2 shows that the success rate of the IC is exceptionally good in the case of ACD$(1,1)$ data. Even for a small sample size all information criteria are able to detect the correct model order in the majority of cases. The tendency to underfit works in favor for the BIC and to some extent also for the HQ. For the same reason, the success rates for the AIC are relatively low due to its overfitting property.
A similar behavior can be observed for ACD$(2,1)$ in Table B.4: Although the IC underestimate the model for smaller sample sizes as a ACD$(1,1)$ model, they improve for large sample sizes.
In both the ACD$(1,1)$ and the ACD$(2,1)$ case, i.e. the cases for $q = 1$, the behaviour of the model selection is acceptable: a reasonably large sample size, which is of the order of a typical intra day trading data sample, ensures a sufficiently large success rate in detecting the correct model. Unfortunately, this cannot be said about the case $q = 2$:
In the first example of ACD$(1,2)$ data in Table B.3, we see that the correct model order is never detected in the majority of cases even for large sample sizes. The best success rates are the ones of the AIC again due to its overfitting tendency. This may be concerning, as this shows that despite the fact that ACD$(1,2)$ and ACD$(2,1)$ have the same number of parameters the model selection behavior is far from comparable.
In comparison, the results for the ACD$(2,2)$, the most complex model in our experiment, are even more critical: Not only are the IC unable to detect the correct model in most of the cases even with large samples, but the best success rates, again from the AIC, are below 20%.

In relation with our observations in the previous section, the cases where model selection fails align with relatively high MSE of the β parameters for $q = 2$: The contribution of the MSE of the ω parameter is not as important, as this parameter is included in all models. However, the increase in MSE when moving from $q = 1$ to $q = 2$ might be one of the factors explaining the discrepancy in model selection between $q = 1$ and $q = 2$. This part of our Monte-Carlo experiment suggests that parameters which are harder to estimate compared to other model parameters (in

our case α vs. β parameters or in other words moving average vs. autoregressive parameters in Equation (5.21)) might also be less likely to be detected by model selection.

5.4 Hawkes processes

Hawkes processes are a class of self-exciting point processes that is popular in various areas of scientific research. This stochastic model allows dependence between random events and associated clustering effects while still offering an acceptable degree of tractability. Therefore, Hawkes process models, which were initially applied to seismic data (Hawkes and Adamopoulos 1973, Ogata 1988), also find application in areas such as neuroscience as models for spike trains (Pernice et al. 2012, Reynaud-Bouret, Rivoirard et al. 2014), genetics (Reynaud-Bouret and Schbath 2010), criminology (Mohler et al. 2011) and social networks (Crane and Sornette 2008, Blundell, Beck and Heller 2012). Moreover, Hawkes processes have been incorporated into economic and financial models for different asset classes and risk types. The property of self-excitation and event clustering were found to match empirically observed stylized facts of intra-day financial data better than models with independent events. Bowsher 2007 was among the early works to propose Hawkes processes for a financial model for mid-price changes, and also Bauwens and Hautsch 2004 and Hewlett 2006 followed similar approaches for trade duration models. Bacry, Mastromatteo and Muzy 2015 give an excellent overview of Hawkes processes in finance which can be complemented by Hawkes 2018. Nevertheless, in order to indicate the scope in which Hawkes processes are used in financial modeling, let us mention a few recent developments: Filimonov and Sornette 2012, Hardiman, Bercot and Bouchaud 2013, Hardiman and Bouchaud 2014 investigate whether price changes in futures markets are mainly endogenously driven or influenced by exogenous factors like news events or other market shocks. The key quantity is the so-called branching ratio. Rambaldi, Pennesi and Lillo 2015 analyze foreign exchange data and propose a combination of a stationary Hawkes process with either double exponential or power-law kernel for endogenous market activity and an additional term similar to a non-stationary Hawkes process. This additional term represents macroeconomic news that arrive at fixed times of the day and influence the market. Muni Toke and Pomponio 2012 and Achab et al. 2018 give applications for order books and Chavez-Demoulin, Davison and McNeil 2005, Chavez-Demoulin and McGill 2012 address Hawkes processes for value-at-risk (VaR) calculations and peaks-over-threshold (POT) methods. Among credit risk models, Errais, Giesecke and Goldberg 2010 incorporate Hawkes processes into the affine process framework for portfolio credit risk and Aït-Sahalia, Laeven and Pelizzon 2014 propose a valuation technique for CDS. Schneider, Lillo

and Pelizzon 2018 analyze illiquidity spillovers in sovereign bond markets via self-exciting point processes.

5.4.1 Definition and some properties

Within the scope of this presentation, we will consider one-dimensional Hawkes processes which are defined by specifying the conditional intensity function (see also Section 1.1). For an overview of more general spatio-temporal point processes see for example Reinhart 2017.

Definition 19. A *Hawkes process* is a point process $(N(t))_{t\geq 0}$ with (conditional) intensity function

$$\lambda(t) = \mu + \int_a^t g(t-\tau)\,\mathrm{d}N(\tau), \tag{5.23}$$

where $a = 0$ (finite past) or $a = -\infty$ (infinite past) and g is the *response function* with $g(\tau) \geq 0\ \forall \tau \in \mathbb{R}^+$ and $\mu > 0$ is the *baseline intensity*.

The function g is sometimes also referred to as *kernel* due to the integral in (5.23). The choice of g determines how past events affect the intensity function and thus the probability of future events. Essentially, for parametric estimation, there are two kernels which are widely used in the literature to fit financial data: the exponential kernel and the power-law kernel.
The early seismologic model which was proposed by Ogata 1988 is of power law type:

$$g(\tau) = \frac{K}{(\tau + c)^{1+\omega}},$$

where $K, c, \omega > 0$. Plugging this into (5.23) yields

$$\lambda(t) = \mu + K \sum_{i=1}^{k} \frac{1}{(t - t_i + c)^{1+\omega}},$$

where $\{t_1, \ldots, t_k\}$ are the jump times of $N(t)$ up to time t. This model is also called ETAS (epidemic-type aftershock sequence) model.
For an exponential kernel, choose g to be of the form

$$g(\tau) = \sum_{m=1}^{P} \alpha_m \mathrm{e}^{-\beta_m \tau},$$

with $\mu, \alpha_m, \beta_m > 0$. The resulting conditional intensity is given by

$$\lambda(t) = \mu + \sum_{m=1}^{P} \alpha_m \sum_{i=1}^{k} \mathrm{e}^{-\beta_m (t - t_i)}, \tag{5.24}$$

where $\{t_1, \ldots, t_k\}$ are again the jump times of $N(t)$ up to time t. In short, we will call this process exponential Hawkes process of order P or exponential Hawkes-P for short. The exponential Hawkes-P model will be the model used for numerical experiments on the model selection.

Remark 15 (Double exponential versus power-law, empirical findings in the literature)**.** In order to enhance the practical relevance of our experiments and results we would also like to use parameter settings which allow intensities which can also be observed in empirical studies. This is why we include a parameter set that was estimated in Lallouache and Challet 2016 for our Monte-Carlo experiment (see Table 5.4).
When applying the above models to financial data, the results allow following comparison: whereas the power law asymptotics are additionally supported by results from non-parametric estimation literature as in Bacry, Dayri and Muzy 2012, the exponential kernel case is analytically more tractable and is still applied in recent literature. Of course this is just a rough division and there have been variations and hybrid models proposed across the literature. For example, in Hardiman, Bercot and Bouchaud 2013 and Lallouache and Challet 2016, power-law kernels are approximated with sums of exponential functions with power-law weights.
Concerning the exponential Hawkes P-model, Hardiman, Bercot and Bouchaud 2013 found that the use of the single exponential intensity function might give misleading results, which is also confirmed by Rambaldi, Pennesi and Lillo 2015. However, this does not necessarily hold for exponential Hawkes processes of higher order: Lallouache and Challet 2016 found that Hawkes models with exponential intensity kernels of order $P = 2, 3$, but not greater than 4 perform better than the single exponential model and comparably well to power law models when applied to FX data. Due to computational inefficiency and lack of empirical evidence in the corresponding literature, we just consider the model orders up to $P = 3$. Moreover, the findings in Lallouache and Challet 2016 were corroborated by Omi, Hirata and Aihara 2017 who used data from the Japanese futures market and considered similar models with up to four exponential terms to later restrict their studies to the best fitting ones with two and three exponential terms.

Figure 5.6 shows a typical path of the counting process N of an exponential Hawkes-P process and its underlying (conditional) intensity process λ. Whenever an event occurs, there is a jump in λ of the size $\alpha := \sum_m^P \alpha_m$. In between events, the intensity function decays exponentially. The more events occurred in the recent past, the higher the intensity and thus the probability of future events. Therefore, the

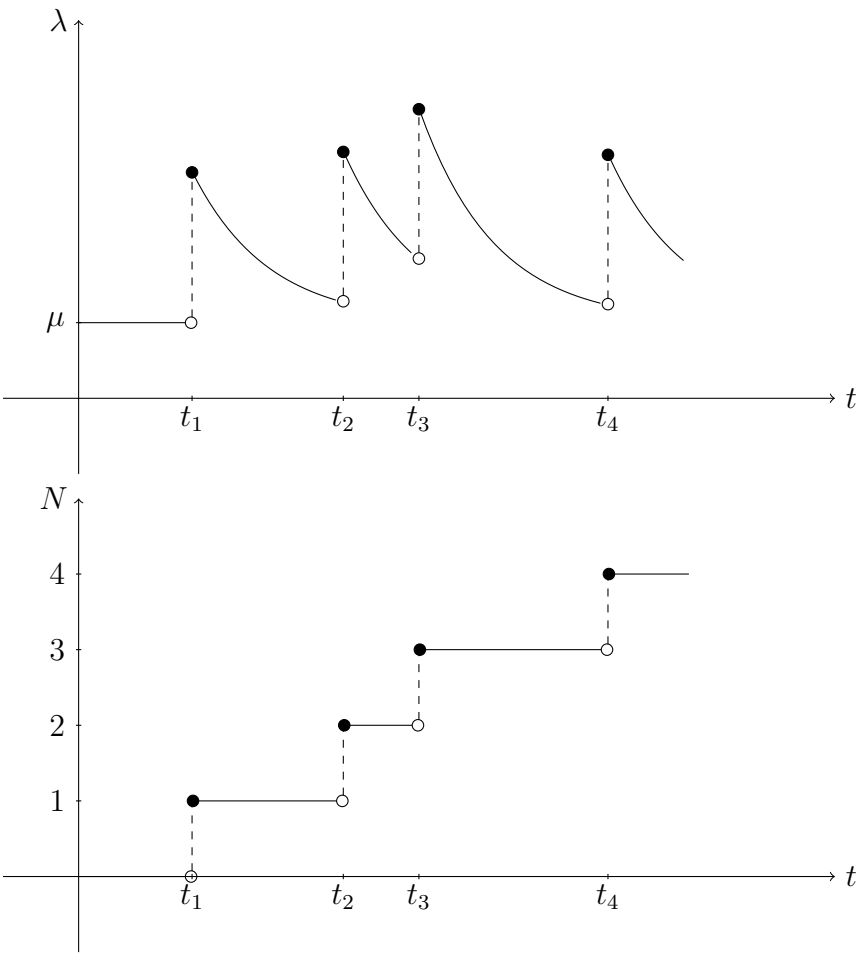

Figure 5.6: Depiction of a typical path of the counting process N corresponding to a Hawkes process with underlying intensity process λ.

process exhibits clustering and self-exciting effects.

Proposition 42. Let $(N(t))_{t\geq 0}$ be a stationary Hawkes process with infinite past and intensity λ as given in Equation (5.23). Then, the average intensity of N can be calculated to be

$$\Lambda := \mathbb{E}[\lambda(t)] = \frac{\mu}{1 - \int_0^\infty g(\nu)\,\mathrm{d}\nu},$$

Proof. Let $A(t) = \int_{-\infty}^t \lambda(u)\mathrm{d}u$ denote the compensator of N. Taking expectations on both sides of Equation (5.23) yields

$$\mathbb{E}[\lambda(t)] = \mu + \mathbb{E}\left[\int_{-\infty}^t g(t-u)\,\mathrm{d}N(u)\right] = \mu + \mathbb{E}\left[\int_{-\infty}^t g(t-u)\,\mathrm{d}A(u)\right] \tag{5.25}$$

$$= \mu + \mathbb{E}\left[\int_{-\infty}^t g(t-u)\lambda(u)\,\mathrm{d}u\right] = \mu + \int_{-\infty}^t g(t-u)\mathbb{E}[\lambda(u)]\,\mathrm{d}u, \tag{5.26}$$

where in (5.25) we used Theorem 1 and in (5.26) we applied Fubini's theorem since the integrand is positive. Due to stationarity, the average intensity is constant and

setting $\Lambda := \mathbb{E}[\lambda(t)]$ simplifies the above equation to

$$\Lambda = \mu + \Lambda \int_{-\infty}^{t} g(t-u)\,\mathrm{d}u$$

Solving for Λ and changing coordinates ($t - u =: v$) in the integral yields

$$\Lambda = \frac{\mu}{1 - \int_0^\infty g(v)\,\mathrm{d}v}.$$

□

Remark 16 (Branching ratio and stationarity)**.** The quantity $n := \int_0^\infty g(\nu)\,\mathrm{d}\nu$ is called the branching ratio. In particular, for the exponential kernel we have $n = \sum_{m=1}^{P} \alpha_m/\beta_m$ and the stationarity condition is $n < 1$. The special case of $n = 1$ also allows stationary processes which are treated in Brémaud and Massoulié 2001.

In the case of exponential Hawkes processes with finite past, we follow a different approach using Laplace transforms: Let $\varphi(t) := \mathbb{E}[\lambda(t)]$ now be the average intensity function of a non-stationary Hawkes process. Then, taking expectations as in the proof of Proposition 42 yields

$$\varphi(t) = \mu + \sum_{m=1}^{P} \int_0^t \alpha_m \mathrm{e}^{-\beta_m(t-u)} \varphi(u)\,\mathrm{d}u. \tag{5.27}$$

The Laplace transform of φ is given by

$$\begin{aligned}
\tilde{\varphi}(s) &= \int_0^\infty \mathrm{e}^{-st}\varphi(t)\,\mathrm{d}t \\
&= \int_0^\infty \mathrm{e}^{-st}\mu\,\mathrm{d}t + \sum_{m=1}^{P} \alpha_m \int_{t=0}^\infty \mathrm{e}^{-st} \int_{u=0}^t \mathrm{e}^{-\beta_m(t-u)}\varphi(u)\,\mathrm{d}u\,\mathrm{d}t \\
&= \frac{\mu}{s} + \sum_{m=1}^{P} \alpha_m \int_{u=0}^\infty \mathrm{e}^{-su}\varphi(u) \int_{t=u}^\infty \mathrm{e}^{-(s+\beta_m)(t-u)}\,\mathrm{d}t\,\mathrm{d}u \qquad (5.28) \\
&= \frac{\mu}{s} + \sum_{m=1}^{P} \frac{\alpha_m}{s+\beta_m} \int_{u=0}^\infty \mathrm{e}^{-su}\varphi(u)\,\mathrm{d}u = \frac{\mu}{s} + \left(\sum_{m=1}^{P} \frac{\alpha_m}{s+\beta_m}\right)\tilde{\varphi}(s),
\end{aligned}$$

where in (5.28) we are able to apply Fubini's theorem since the integrand is positive. Finally, we have an algebraic equation which can be solved for $\tilde{\varphi}$:

$$\tilde{\varphi}(s) = \frac{\frac{\mu}{s}}{1 - \sum_{m=1}^{P} \frac{\alpha_m}{s+\beta_m}}. \tag{5.29}$$

For $P > 1$ we could write alternatively:

$$\tilde{\varphi}(s) = \frac{\mu}{s} \frac{\prod_{m=1}^{P}(s+\beta_m)}{\prod_{m=1}^{P}(s+\beta_m) - \sum_{m=1}^{P} \alpha_m \prod_{k\neq m}(s+\beta_k)}. \tag{5.30}$$

This gives an analytic expression for the Laplace transform of the intensity function. From Equation (5.29) we can see that it is reasonable to demand the usual stationarity condition $\sum_{m=1}^{P} \alpha_m/\beta_m < 1$ in order to ensure that the right hand side term is well defined.
In general, the evaluation of the average intensity function can be done by (numerical) Laplace inversion. However, for lower model orders (up to $P = 4$) it is possible to invert the Laplace transform analytically. We will show this for first and second order in the following examples.

Example 9 (Formula for the average intensity in the case $P = 1$)**.** For $P = 1$ the expression in (5.29) simplifies to

$$\tilde{\varphi}(s) = \frac{\mu(s+\beta_1)}{s(s+\beta_1-\alpha_1)} = \frac{\mu}{\beta_1-\alpha_1}\left(\frac{\beta_1}{s} - \frac{\alpha_1}{s+\beta_1-\alpha_1}\right), \tag{5.31}$$

where we used a partial fractions decomposition in the last step. This allows us to analytically invert the Laplace transform:

$$\varphi(t) = \frac{\mu}{\beta_1-\alpha_1}\left(\beta_1 - \alpha_1 \mathrm{e}^{-(\beta_1-\alpha_1)t}\right), \quad t > 0. \tag{5.32}$$

Example 10 (Formula for the average intensity in the case $P = 2$)**.** For $P = 2$ we have

$$\tilde{\varphi}(s) = \frac{\mu(s+\beta_1)(s+\beta_2)}{s[(s+\beta_1)(s+\beta_2) - \alpha_1(s+\beta_2) - \alpha_2(s+\beta_1)]} \tag{5.33}$$

Starting from order $P = 2$, the explicit formulas can be quite complicated.
By standard partial fractions, the inverse Laplace transform can be calculated to be

$$\varphi(t) = \mu\left(A_1 + A_2\mathrm{e}^{s_2 t} + A_3\mathrm{e}^{s_3 t}\right). \tag{5.34}$$

Let R and Q denote the polynomial in the numerator and the denominator of the right hand side expression in (5.33) respectively. Then, assuming Q has only real valued roots of single multiplicity denoted by s_1, s_2, s_3, the partial fractions decomposition is given by

$$\tilde{\varphi}(s) = \frac{P(s)}{Q(s)} = \sum_{i=1}^{3} \frac{P(s_i)}{Q'(s_i)(s-s_i)} = \mu\left(\frac{A_1}{s} + \frac{A_2}{s-s_2} + \frac{A_3}{s-s_3}\right), \tag{5.35}$$

where

$$s_1 = 0, \quad s_2 = \frac{1}{2}(\gamma - \xi), \quad s_3 = \frac{1}{2}(\gamma + \xi) \tag{5.36}$$

$$\text{with} \quad \gamma = \alpha_1 + \alpha_2 - \beta_1 - \beta_2 \quad \text{and} \quad \xi = \sqrt{\gamma^2 - 4(\beta_1\beta_2 - \alpha_1\beta_2 - \alpha_2\beta_1)}. \tag{5.37}$$

The partial fractions decomposition implies that

$$A_1(s - s_2)(s - s_3) + A_2 s(s - s_3) + A_3 s(s - s_2) \stackrel{!}{=} (s + \beta_1)(s + \beta_2) \tag{5.38}$$

and comparing coefficients of s^2, s and 1 on both sides of the equation yields

$$A_1 + A_2 + A_3 = 1 \tag{5.39}$$

$$-A_1(s_1 + s_2) - A_2 s_3 - A_3 s_2 = \beta_1 + \beta_2 \tag{5.40}$$

$$A_1 s_1 s_2 = \beta_1 \beta_2. \tag{5.41}$$

Then we get

$$A_1 = \frac{\beta_1\beta_2}{s_1 s_2} = \frac{\beta_1\beta_2}{(\gamma^2 - \xi^2)/4} = \frac{\beta_1\beta_2}{\beta_1\beta_2 - \alpha_1\beta_2 - \alpha_2\beta_1} \tag{5.42}$$

by solving (5.41) for A_1 and inserting (5.36).
Now multiply (5.39) by s_2 and add (5.40) to get

$$-A_1 s_3 + A_2(s_2 - s_3) = \beta_1 + \beta_2 + s_2. \tag{5.43}$$

Solving for A_2 we get

$$\begin{aligned} A_2 &= \frac{\beta_1\beta_2/s_2 + \beta_1 + \beta_2 + s_2}{s_2 - s_3} = \frac{\beta_1\beta_2 + s_2(\beta_1 + \beta_2) + s_2^2}{s_2(s_2 - s_3)} \\ &= \frac{4\beta_1\beta_2 - 2(\xi - \gamma)(\beta_1 + \beta_2) + (\xi - \gamma)^2}{2\xi(\xi - \gamma)} = \frac{(\xi - \gamma - 2\beta_2)(\xi - \gamma - 2\beta_1)}{2\xi(\xi - \gamma)} \\ &= \frac{(\xi - \alpha_1 - \alpha_2 + \beta_1 - \beta_2)(\xi - \alpha_1 - \alpha_2 - \beta_1 + \beta_2)}{2\xi(\xi - \gamma)}. \end{aligned} \tag{5.44}$$

Multiplying (5.39) by s_3, adding (5.40) and following similar steps as for A_2 yield

$$A_3 = \frac{(\xi + \gamma + 2\beta_1)(\xi + \gamma + 2\beta_2)}{2\xi(\xi + \gamma)} = \frac{(\xi + \alpha_1 + \alpha_2 + \beta_1 - \beta_2)(\xi + \alpha_1 + \alpha_2 - \beta_1 + \beta_2)}{2\xi(\xi + \gamma)}. \tag{5.45}$$

The Laplace inversion gives the required result.
Note that with the condition $\sum_{m=1}^{P} \alpha_m/\beta_m < 1$ it follows that the roots s_2 and s_3 are real and negative.

From both examples, we can see that for large times t the exponential terms in Equations (5.32) and (5.34) become negligible and the remaining expressions co-

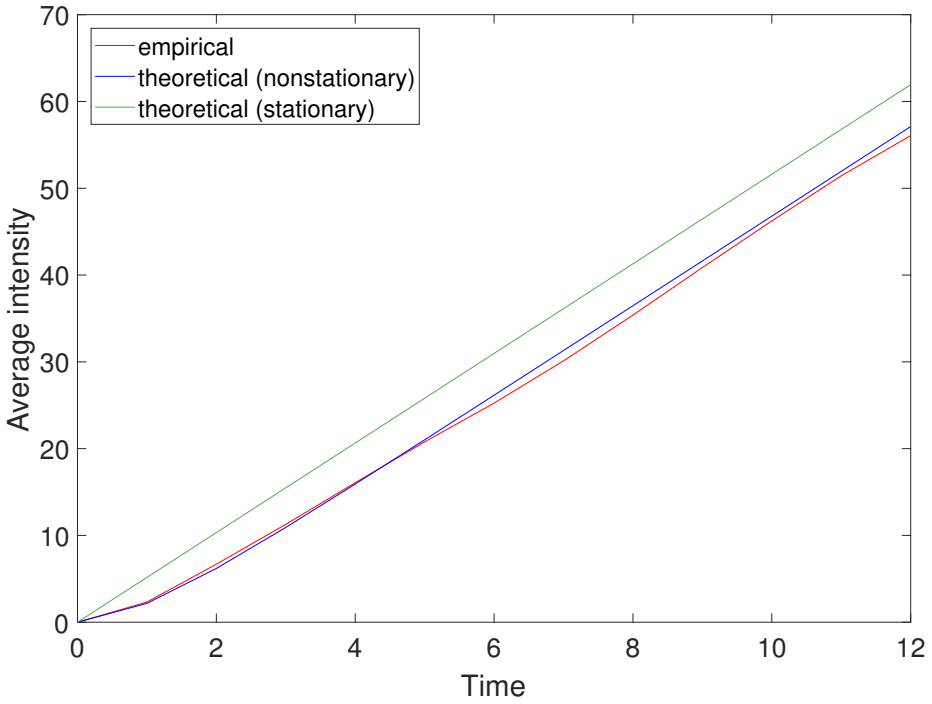

Figure 5.7: A comparison between the average number of events from a Monte-Carlo simulation and the theoretical values. For the parameter values $\mu = 0.5$, $\alpha_1 = 3.1, \alpha_2 = 5.9, \beta_1 = 9.9$ and $\beta_2 = 10$, we simulated an exponential Hawkes process of order $P = 2$ with finite past and plotted the empirical average number of events (red curve) against the theoretical values of the expected number of events in Eq (5.46). In the non-stationary case, we integrate the average intensity function in Eq (5.34) which corresponds to the blue curve. The stationary case is shown via the green curve.

incide with the intensity function of the stationary case. In a small Monte-Carlo experiment, we simulated 1000 paths of a Hawkes process with 1000 events (see also `empirAgg2.m`). The parameters are $\mu = 0.5$, $\alpha_1 = 3.1$, $\alpha_2 = 5.9$, $\beta_1 = 9.9$ and $\beta_2 = 10$. Figure 5.7 shows a plot of the empirically observed average number of events against the theoretically expected number of events. Plotting such figures might be useful for validation of a simulations algorithm. Recall the relation between average intensity function φ and expected number of events of a point process $(N(t))_{t\geq 0}$:

$$\mathbb{E}[N(t)] = \int_0^t \varphi(\tau)\,\mathrm{d}\tau \tag{5.46}$$

For small times we can observe the transient exponential behavior which vanishes for large times. In particular, the slope of the two theoretical functions are approximately equal for large times and indicate that the intensity function of the non-stationary case converges to the stationary case. Also, we can verify the edge effect when simulating a Hawkes process with finite past, which will be briefly discussed in the next section.

5.4.2 Simulation

Apart from the thinning approach Ogata 1981, there are several alternatives: the time-change approach based on the random time-change theorem Meyer 1971 and applied specifically to Hawkes processes for instance in Ozaki 1979, exact simulation Dassios and Zhao 2013 and perfect simulation Møller and Rasmussen 2005. We will use a variation of Algorithm 1.4, which follows the thinning approach. As seen in the previous section, simulating a Hawkes process with finite past in order to approximate a Hawkes process with infinite past will cause the simulated process to be non-stationary at the beginning of the simulation time. This phenomenon is also known as edge effect as offspring of events that might have occurred in the past are omitted. For further details on this see Møller and Rasmussen 2005, 2006.
However, similar to Dassios and Zhao 2013, we explicitly want to work with a Hawkes process with finite past. Therefore, we view the edge effect as an inherent property of the model rather than an artifact of the simulation. Besides, the exact simulation algorithm in Dassios and Zhao 2013, though applicable to multidimensional exponential models, does not directly apply to our proposed model due to the lack of identification of the exogenous and endogenous part of the intensity. This leaves us with the popular thinning algorithm going back to Lewis and Shedler 1979 and Ogata 1981. We used an implementation of the thinning algorithm to simulate the process on a time interval $[0, T]$ (see `hawkesThinning.m`) and compare models up to order 3. We first generate sample data that serve as a technical example for the estimation and model selection methods. The parameter settings are given in Table 5.3.

Algorithm 5.1: Ogata's algorithm for Hawkes processes

input:
  μ – baseline intensity
  α – upper bound on jumps of the intensity function
  λ – conditional intensity function
output:
  $\{t_1, t_2, \ldots\}$ – event times of the simulated process
begin
  1) set $M \leftarrow \mu$, $t \leftarrow 0$, $H \leftarrow \varnothing$
  2) generate $T \sim \text{Exp}(M)$
    if $T \leq A \rightsquigarrow t_1 \leftarrow T$
    else $\rightsquigarrow$ ***terminate***
    end if
  3) set $i \leftarrow k \leftarrow 1$
  4) update $M \leftarrow \lambda(t_k|H) + \alpha$
  5) generate $T \sim \text{Exp}(M)$, $U \sim \text{Uniform}(0, 1)$
  6) ***if*** $t + T > A \rightsquigarrow$ ***terminate***
    else if $\lambda(t + T|H)/M > U \rightsquigarrow t \leftarrow t + T$, $M \leftarrow \lambda(t|H)$, ***goto*** (5)

```
18                 else  k ← k + 1 ,  t_k ← t + T ,  t ← t_k ,  H ← H ∪ {t_k} ,
19                       goto (4)
20                 end if
21           end if
22     end
```

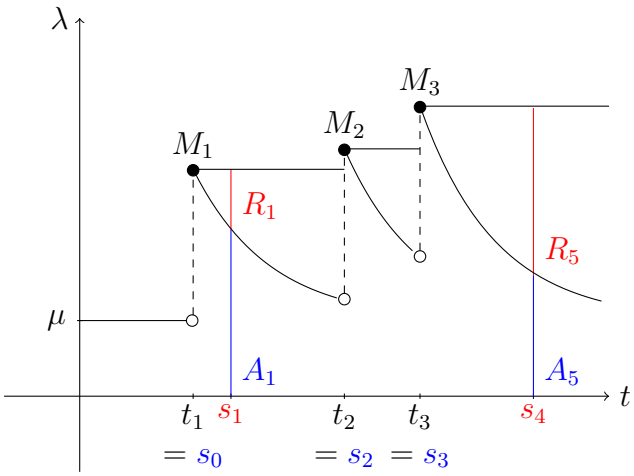

Figure 5.8: Thinning algorithm, an example: From simulated samples $\{s_1, s_2, s_3, s_4\}$ the points $\{s_1, s_4\}$ were rejected and the rest are accepted as samples of the Hawkes process (s_0 is the initial event which is Exp(μ) distributed). The acceptance-rejection ratio for s_k is A_k/R_k where $A_k = \lambda(s_k)$ and $R_k = M - \lambda(s_k)$, where $k = 1, 2, 3, 4$ and M is the current local upper bound on the intensity at the point s_k.

5.4.3 Fitting

The fitting algorithm follows the theory in Ozaki 1979 which is a standard maximum likelihood procedure. For a self-exciting point process with intensity λ the log-likelihood for data $0 < t_1 < \ldots < t_n < T$ is given by

$$\log \mathcal{L}(t_1, \ldots, t_n|\theta) = -\int_0^T \lambda(t|\theta)\,\mathrm{d}t + \int_0^T \log(\lambda(t|\theta))\mathrm{d}N(t). \tag{5.47}$$

Let $\theta = (\mu, \alpha_1, \ldots, \alpha_P, \beta_1 \ldots, \beta_P)$ be the vector of parameters for the Hawkes P-model. Inserting Equation (5.24) into Equation (5.47) gives

$$\begin{aligned}\log \mathcal{L}(t_1, \ldots, t_n|\theta) = -\mu T - \sum_{m=1}^{P} \left[\frac{\alpha_m}{\beta_m} \sum_{t_i<T} \left(1 - \mathrm{e}^{-\beta_m(T-t_i)}\right) \right] \\ + \sum_{t_k<T} \log\left(\mu + \sum_{m=1}^{P} \alpha_m \sum_{t_i<t_k} \mathrm{e}^{-\beta_m(t_k-t_i)} \right).\end{aligned} \tag{5.48}$$

Moreover, Ozaki 1979 shows that the log-likelihood can be calculated recursively, which reduces the computational burden from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$: Assume that $T = t_n$,

i.e. the last event is the last time point of observation. Then

$$\log \mathcal{L}(t_1, \ldots, t_n|\theta) = -\mu t_n - \sum_{m=1}^{P} \left[\frac{\alpha_m}{\beta_m} \sum_{t_i \leq t_n} \left(1 - \mathrm{e}^{-\beta_m (t_n - t_i)}\right) \right] + \sum_{t_k \leq t_n} \log \left(\mu + \sum_{m=1}^{P} \alpha_m A_m(k) \right), \tag{5.49}$$

$$\text{where } A_m(1) = 0 \quad \forall\, m = 1, \ldots, P$$

$$A_m(k) = \sum_{t_i < t_k} \mathrm{e}^{-\beta_m (t_k - t_i)} = (1 + A_m(k-1))\, \mathrm{e}^{-\beta_m (t_k - t_{k-1})}.$$

To obtain the MLE of the parameters we maximize the log-likelihood function with respect to the parameters subject to the stationarity condition:

$$\underset{\mu, \alpha_1, \ldots, \alpha_P, \beta_1, \ldots, \beta_P}{\arg\max} \log \mathcal{L}(t_1, \ldots, t_n|\mu, \alpha_1, \ldots, \alpha_P, \beta_1, \ldots, \beta_P) \tag{5.50}$$

$$\text{s.t. } \mu, \alpha_1, \ldots, \alpha_P, \beta_1, \ldots, \beta_P > 0, \quad \beta_1 < \ldots < \beta_P \quad \text{and} \quad \sum_{m=1}^{P} \frac{\alpha_m}{\beta_m} < 1.$$

We assume the β parameters to be ordered to avoid identification problems (see page 285 in Hautsch 2012). The maximization (or rather the minimization of the negative log-likelihood) is typically done numerically as the estimators are not available in closed form. We used the standard MATLAB function `fmincon` for constrained problems. The optimization routine can be found in the supplementary files `fitting.m`, `conditions.m` and `LogLik_iter.m`.

Important asymptotic properties of the MLE for Hawkes processes have been studied and proven by Ogata 1978 (see also in the appendix in Rambaldi, Pennesi and Lillo 2015 for a brief summary). In particular, we may assume the MLE to be consistent, i.e. with sample size tending to infinity the MLE converge to the true values of the parameters. In order to verify these results with our Monte-Carlo experiment, we use the RMSE as a measure for the goodness of fit. Nevertheless, especially for empirical analysis, testing the distribution of the residuals is an important step of model diagnostics. As a consequence of the time-change theorem (see Theorem 4) for point processes, we may expect that the residuals

$$\theta_k = \int_{t_k}^{t_{k+1}} \hat{\lambda}(s)\, \mathrm{d}s \tag{5.51}$$

are i.i.d. exponentially distributed with mean 1, where $\{t_1, t_2, \ldots, t_n\}$ are the recorded event times of the process (see also Section 4.1.5 in Hautsch 2012 and Section 4.1 in Bowsher 2007). In order to calculate the residuals for the exponential Hawkes-P model, it is useful to rewrite the intensity in the following way: Let $t \in [t_k, t_{k+1})$,

$k \in \{1, 2, \ldots, n\}$. Then

$$\begin{aligned}
\lambda(t) &= \mu + \sum_{m=1}^{P} \alpha_m \sum_{i=1}^{k} \mathrm{e}^{-\beta(t-t_i)} \\
&= \mu + \sum_{m=1}^{P} \left(\alpha_m \mathrm{e}^{-\beta_m(t-t_k)} + \alpha_m \sum_{i=1}^{k-1} \mathrm{e}^{-\beta_m(t-t_i)} \right) \\
&= \mu + \sum_{m=1}^{P} \left(\alpha_m \mathrm{e}^{-\beta_m(t-t_k)} + \alpha_m \sum_{i=1}^{k-1} \mathrm{e}^{-\beta_m(t_k-t_i+t-t_k)} \right) \\
&= \mu + \sum_{m=1}^{P} \alpha_m \left(1 + \sum_{i=1}^{k-1} \mathrm{e}^{-\beta_m(t_k-t_i)} \right) \mathrm{e}^{-\beta_m(t-t_k)}.
\end{aligned}$$

This representation also appears in Dassios and Zhao 2013 following from an ODE approach. Using this expression we are able to calculate the integral in (5.51):

$$\begin{aligned}
\theta_k &= \int_{t_k}^{t_{k+1}} \hat{\lambda}(s)\, \mathrm{d}s \\
&= \hat{\mu}(t_{k+1} - t_k) + \sum_{m=1}^{P} \hat{\alpha}_m \left(1 + \sum_{i=1}^{k-1} \mathrm{e}^{-\hat{\beta}_m(t_k-t_i)} \right) \int_{t_k}^{t_{k+1}} \mathrm{e}^{-\hat{\beta}_m(t-t_k)} \mathrm{d}s \\
&= \hat{\mu}(t_{k+1} - t_k) + \sum_{m=1}^{P} \hat{\alpha}_m \left(1 + \sum_{i=1}^{k-1} \mathrm{e}^{-\hat{\beta}_m(t_k-t_i)} \right) \left(-\frac{1}{\hat{\beta}_m} \mathrm{e}^{-\hat{\beta}_m(t-t_k)} \right) \Bigg|_{t_k}^{t_{k+1}} \\
&= \hat{\mu}(t_{k+1} - t_k) + \sum_{m=1}^{P} \frac{\hat{\alpha}_m}{\hat{\beta}_m} \left(1 + \sum_{i=1}^{k-1} \mathrm{e}^{-\hat{\beta}_m(t_k-t_i)} \right) \left(1 - \mathrm{e}^{-\hat{\beta}_m(t_{k+1}-t_k)} \right).
\end{aligned}$$

As a first visual test, one can plot the quantiles of the residuals against the theoretical quantiles of the exponential distribution. Figure 5.9 shows some QQ-plots of exponential Hawkes-P models fitted to simulated data with varying model order. While underfitted cases can be recognized (since the plotted line diverges from the diagonal), overfitted situations are not as easy to identify. After a visual check via QQ-plot, one can also apply statistical tests like the Kolmogorov-Smirnov test or excess dispersion test (see Engle and Russell 1998). Independence is difficult to test for, indeed, Lallouache and Challet 2016 use a Ljung-Box test to check whether the residuals are uncorrelated.

5.4.4 Numerical results

Using the thinning algorithm described in Algorithm 5.1 we simulated four different data sets containing 1000 samples. Three of them correspond to each row of Parameter Set 1 in Table 5.3 and one data set consists of samples of an exponential Hawkes 2-model with parameter values from Parameter Set 2 shown in Table 5.4.

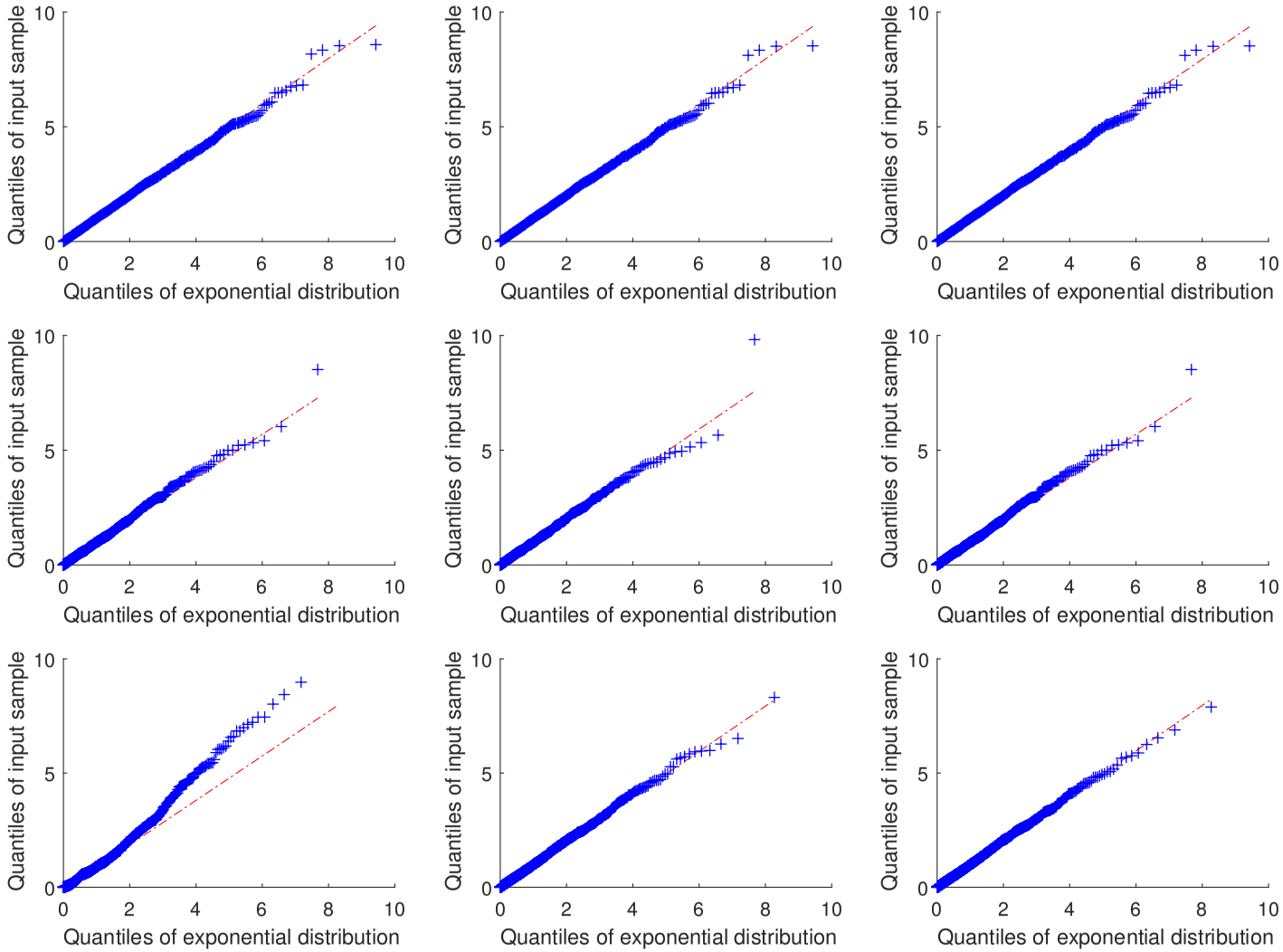

Figure 5.9: A plot in the k-th row and the j-th colum shows a QQ-plot of the residuals of an exponential Hawkes-j model fitted to data of sample size 1000 coming from an exponential Hawkes-k model.

Especially for Parameter Set 2 the time horizon T can be assumed to be given in seconds. It ranges from 10 min to 6 h to reflect typical intra-day financial data sets. In order to check how well the estimation method works for our parameter sets, we first assume that the correct model order P is known and run a MLE of the parameters of the true model underlying each data set. Subsequently, we are able to calculate the RMSE as a measure of distance between the true and the estimated parameter values. The absolute and relative RMSE values for Parameter Set 1 can be found in Table B.6 and Table B.7 respectively. For Parameter Set 2, see Tables B.11 and B.12. We observe that the RMSE decreases with increasing sample size. This is to be expected as the MLE is known to be consistent.
Finally, we assume that the true model order is not known, but needs to be selected by the IC. Consequently, for each data set we have to fit all possible model orders $P = 1, 2, 3$ and to calculate the associated IC values. In the following we discuss the results of the model selection.

Table 5.3: Simulation parameters (Parameter Set 1)

	μ	α_1	α_2	α_3	β_1	β_2	β_3
P=1	0.5	9	–	–	10	–	–
P=2	0.5	0.00066	100	–	0.001	300	–
P=3	0.5	0.00033	3.3	100	0.001	10	300

Table 5.4: Parameter set taken from Lallouache and Challet 2016 (Parameter Set 2)

μ	α_1	α_2	β_1	β_2
0.05	0.01761905	0.28	0.04761905	0.6666667

We first consider Parameter Set 1. For simulated data with model order $P = 1$ we can see in Table B.7 that the relative RMSE is comparably low even for the smallest samples corresponding to the time horizon $T = 500$. The model selection in Table B.8 confirms that the smallest sample size might already be enough to guarantee high success rates (over 90%) of all IC. Nevertheless, already in the lowest order case, we can observe the different behavior of consistent and inconsistent IC. For BIC and HQ, the success rate improves with increasing average sample size. In particular, the relation seems to be monotone and, in the case of the BIC, the success rate reaches 100% already for $T = 1000$, The HQ performs slightly worse than BIC, but is still well over 90% and very close to 100% for $T = 5000$. However, the AIC behaves in a more concerning manner. Even for large sample sizes the model

selection using the AIC allows a comparably large probability ($> 6\%$) to select a higher order than $P = 1$. As the AIC is not a consistent IC, we cannot exclude the possibility that these results already approximate the asymptotic distribution of model selection of the AIC. As mentioned earlier this asymptotic distribution is typically different from the delta distribution with mass one on the true model order. Additionally, the numerical results show that increasing the average sample size does not necessarily increase the success rate of model selection. For instance, moving from $T = 500$ to $T = 1000$ we can observe a decrease in success rate in the AIC case.
In the case of model order $P = 2$, there is the possibility of both over- and underestimation. We observe quite large RMSE for the parameters α_1 and β_1, especially for smaller samples corresponding to $T = 500$ and $T = 1000$. This could be one of the factors affecting the model selection for $T = 500$ in Table B.9: there is a significant proportion of underestimation among all IC, most notably the high underestimation rate of almost 95% of the BIC. The AIC seems to perform best in this setting for $T = 500$ with success rates slightly above 50%, but also with 48% underestimation. For larger samples, the BIC and HQ select the correct model order with very high probability (around 90% or even larger) and the BIC reaches 100% success rate at $T = 2000$. Again, we have the adverse effect that the success rates of the AIC decrease with growing average sample size. Even for the largest average sample size for $T = 5000$ there is a relatively large probability of overestimation of over 6%.
For data simulated with $P = 3$ we have a similar behavior as with $P = 2$. Again, Table B.7 reports large RMSE for the parameters α_1 and β_1 in small sample cases. As $P = 3$ is the highest selectable model order, this excludes cases of overspecification. This means that the we can observe the same pattern in model selection of the AIC as for the BIC and HQ: Starting at $T = 500$, there are mostly cases of underestimates followed by improving success rates as the sample size increases. All IC reach 100% success rate for $T = 2000$. However, it is very likely that we would be able to observe the tendency of the AIC to overestimate if we included higher orders $P > 3$ in the model selection set.

When working with Paramerter Set 2, we chose the time horizons 10 min, 15 min, 30 min, 1 h, 3 h and 6 h. At first, there are large RMSE values for $T = 600$ and $T = 900$ (see Tables B.11 and B.12), which shows that the sample sizes are so small that we cannot ensure good estimates of the MLE method. Especially estimates of α_2 and β_2 have large RMSE. This situation corresponds to the case $T = 500$ in the setting of Parameter Set 1. When we compare with the corresponding model selection in Table B.13, we observe the same phenomenon of underestimation is most severe for the BIC, less for the HQ and least for the AIC. As samples are quite small

for these cases and may fulfill the rule of thumb discussed at the end of Section 4.2.1, we included the combined model selection rule AICc/AIC in the table. It applies the AICc whenever $n < 7 \cdot 40 = 280$ and the AIC otherwise. The numerical results for the combined AICc/AIC selection rule are very similar to the standalone AIC and even slightly worse for T=600 and T=900.
When we move on to larger samples from 30 min to 1 h, there is a noticeable change in the RMSE values. More precisely, the RMSE values decrease faster for the second exponential term, i.e. α_2 and β_2, which leads to the first exponential term with α_1 and β_1 to contribute more to the overall estimation error. There is a noticeable increase in the rate of correct model selection among all IC ranging over 90% for $T = 3600$.
Finally, for large samples with time horizons from 3 h up to 6 h represent data of half up to an entire trading day respectively. The relative RMSE of each parameter is less than 20% and the rate of correct model selection for the consistent IC (BIC and HQ) is close to 100%. However, the success rate of the AIC decreases to about 94% with a 6% probability of overestimation.

5.5 Summary

We have analyzed the performance of IC for model selection for three model classes. In the case of the compound Poisson type process, the IC are able to detect the correct number especially for small and moderately many parameters. The model selection procedure for the ACD model is adversely affected by the differing estimation quality for different model orders and can especially be seen in higher orders of the model. The Hawkes process gives satisfying results concerning model selection and the results give hints to the asymptotic distribution of model selection of the AIC.
In all cases we see slight overestimation of the AIC whereas BIC and HQ are more parsimonious and tend to underfit in small sample cases.

Conclusion and outlook

We have seen several point processes that can be used as models for irregular spaced and potentially non-stationary financial data.
In the first half of the thesis we have introduced a non-homogeneous generalization of the fractional homogeneous Poisson process and derived first properties such as governing equations and moments and covariance as well as limit theorems. Using this theoretical basis, we can identify several points for potential further research. First, there is no limitation to the dimension of the Poisson process subject to the time-change. Indeed, Leonenko and Merzbach 2015, Aletti, Leonenko and Merzbach 2018 discuss fractional Poisson fields, whose non-stationary analogues can be explored. Second, further research can be directed towards the implications of the limit results for estimation techniques. The suitability of pre-existing estimation techniques for doubly stochastic models and a modification for the FNPP could be further investigated.
The second part of the thesis comprises a Monte-Carlo experiment to assess the performance of IC using simulated data. Concerning information criteria and model selection, a natural step would be an application to empirical data. This implies that we perform model selection under misspecification and probably need additional model diagnostic tools in order to assess the quality of model selection. The inclusion of robust IC and IC specifically designed for model selection under misspecification could be considered.
Again, all point processes encountered in Chapter 5, which were discussed as one-dimensional processes, have multidimensional generalizations and spatio-temporal point processes are natural directions for further research.

Bibliography

Achab, M. et al. (2018). "Analysis of order book flows using a non-parametric estimation of the branching ratio matrix". In: *Quant. Finance* 18.2, pp. 199–212.

Aït-Sahalia, Y., R. J. A. Laeven and L. Pelizzon (2014). "Mutual excitation in Eurozone sovereign CDS". In: *J. Econometrics* 183.2, pp. 151–167.

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle". In: *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. Akadémiai Kiadó, Budapest, pp. 267–281.

Aletti, G., N. Leonenko and E. Merzbach (2018). "Fractional Poisson Fields and Martingales". In: *J. Stat. Phys.* 170.4, pp. 700–730.

Anscombe, F. J. (1952). "Large-sample theory of sequential estimation". In: *Proc. Cambridge Philos. Soc.* 48, pp. 600–607.

Applebaum, D. (2009). *Lévy processes and stochastic calculus*. Second. Vol. 116. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, pp. xxx+460.

Bacry, E., K. Dayri and J.-F. Muzy (2012). "Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data". In: *The European Physical Journal B* 85.5, pp. 1–12.

Bacry, E., I. Mastromatteo and J.-F. Muzy (2015). "Hawkes Processes in Finance". In: *Market Microstructure and Liquidity* 01.01, p. 1550005. eprint: `http://www.worldscientific.com/doi/pdf/10.1142/S2382626615500057`.

Bauwens, L. and N. Hautsch (2004). "Dynamic Latent Factor Models for Intensity Processes". In: *CORE Discussion Paper*.

Becker-Kern, P., M. M. Meerschaert and H.-P. Scheffler (2004). "Limit theorems for coupled continuous time random walks". In: *Ann. Probab.* 32.1B, pp. 730–756.

Beghin, L. and E. Orsingher (2009). "Fractional Poisson processes and related planar random motions". In: *Electron. J. Probab.* 14, no. 61, 1790–1827.

— (2010). "Poisson-type processes governed by fractional and higher-order recursive differential equations". In: *Electron. J. Probab.* 15, no. 22, 684–709.

Belfrage, M. (2016). *ACDm: Tools for Autoregressive Conditional Duration Models*. R package version 1.0.4.

Bertram, W. K. (2004). "An empirical investigation of Australian Stock Exchange data". In: *Physica A: Statistical Mechanics and its Applications* 341, pp. 533–546.

Biard, R. and B. Saussereau (2014). "Fractional Poisson process: long-range dependence and applications in ruin theory". In: *J. Appl. Probab.* 51.3, pp. 727–740.

— (2016). "Correction: "Fractional Poisson process: long-range dependence and applications in ruin theory" [MR3256223]". In: *J. Appl. Probab.* 53.4, pp. 1271–1272.

Bielecki, T. R. and M. Rutkowski (2002). *Credit risk: modelling, valuation and hedging*. Springer Finance. Springer-Verlag, Berlin, pp. xviii+500.

Billingsley, P. (1999). *Convergence of probability measures*. Second. Wiley Series in Probability and Statistics: Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, pp. x+277.

Bingham, N. H. (1971). "Limit theorems for occupation times of Markov processes". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 17, pp. 1–22.

Bingham, N. H., C. M. Goldie and J. L. Teugels (1989). *Regular variation*. Vol. 27. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, pp. xx+494.

Blundell, C., J. Beck and K. A. Heller (2012). "Modelling reciprocating relationships with Hawkes processes". In: *Advances in Neural Information Processing Systems*, pp. 2600–2608.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". In: *J. Econometrics* 31.3, pp. 307–327.

Bollerslev, T., R. F. Engle and D. B. Nelson (1994). "Arch models". In: *Handbook of econometrics, Vol. IV*. Vol. 2. Handbooks in Econom. North-Holland, Amsterdam, pp. 2959–3038.

Bougerol, P. and N. Picard (1992). "Stationarity of GARCH processes and of some nonnegative time series". In: *J. Econometrics* 52.1-2, pp. 115–127.

Bowsher, C. G. (2007). "Modelling security market events in continuous time: intensity based, multivariate point process models". In: *J. Econometrics* 141.2, pp. 876–912.

Breiman, L. (1968). *Probability*. Addison-Wesley Publishing Company, Reading, Mass.-London-Don Mills, Ont., pp. ix+421.

Brémaud, P. (1975). "An extension of Watanabe's theorem of characterization of Poisson processes over the positive real half line". In: *J. Appl. Probability* 12, pp. 396–399.

— (1981). *Point processes and queues*. Martingale dynamics, Springer Series in Statistics. Springer-Verlag, New York-Berlin, pp. xviii+354.

Brémaud, P. and L. Massoulié (2001). "Hawkes branching point processes without ancestors". In: *J. Appl. Probab.* 38.1, pp. 122–135.

Brockwell, P. J. (2001). "Lévy-driven CARMA processes". In: *Ann. Inst. Statist. Math.* 53.1. Nonlinear non-Gaussian models and related filtering methods (Tokyo, 2000), pp. 113–124.

— (2004). "Representations of continuous-time ARMA processes". In: *J. Appl. Probab.* 41A. Stochastic methods and their applications, pp. 375–382.

— (2014). "Recent results in the theory and applications of CARMA processes". In: *Ann. Inst. Statist. Math.* 66.4, pp. 647–685.

Brockwell, P. J., E. Chadraa and A. Lindner (2006). "Continuous-time GARCH processes". In: *Ann. Appl. Probab.* 16.2, pp. 790–826.

Brockwell, P. J. and R. A. Davis (1991). *Time series: theory and methods*. Second. Springer Series in Statistics. Springer-Verlag, New York, pp. xvi+577.

— (2016). *Introduction to time series and forecasting*. Third. Springer Texts in Statistics. Springer, [Cham], pp. xiv+425.

Buckland, S. T., K. P. Burnham and N. H. Augustin (1997). "Model selection: an integral part of inference". In: *Biometrics*, pp. 603–618.

Burnham, K. P. and D. R. Anderson (2004). "Multimodel inference: understanding AIC and BIC in model selection". In: *Sociol. Methods Res.* 33.2, pp. 261–304.

Cahoy, D. O., V. V. Uchaikin and W. A. Woyczynski (2010). "Parameter estimation for fractional Poisson processes". In: *J. Statist. Plann. Inference* 140.11, pp. 3106–3120.

Caputo, M. (1967). "Linear models of dissipation whose Q is almost frequency independent—II". In: *Geophysical Journal International* 13.5, pp. 529–539.

Chavez-Demoulin, V., A. C. Davison and A. J. McNeil (2005). "Estimating value-at-risk: a point process approach". In: *Quant. Finance* 5.2, pp. 227–234.

Chavez-Demoulin, V. and J. McGill (2012). "High-frequency financial data modeling using Hawkes processes". In: *Journal of Banking & Finance* 36.12, pp. 3415–3426.

Chen, J. et al. (2018). "Performance of information criteria for selection of Hawkes process models of financial data". In: *Quant. Finance* 18.2, pp. 225–235.

Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, pp. xviii+312.

Cont, R. and P. Tankov (2004). *Financial modelling with jump processes*. Chapman & Hall/CRC Financial Mathematics Series. Chapman & Hall/CRC, Boca Raton, FL, pp. xvi+535.

Cox, D. R. (1955). "Some statistical methods connected with series of events". In: *J. Roy. Statist. Soc. Ser. B.* 17, 129–157, discussion, 157–164.

Crane, R. and D. Sornette (2008). "Robust dynamic classes revealed by measuring the response function of a social system". In: *Proceedings of the National Academy of Sciences* 105.41, pp. 15649–15653.

Csörgő, M. and R. Fischler (1973). "Some examples and results in the theory of mixing and random-sum central limit theorems". In: *Period. Math. Hungar.* 3. Collection of articles dedicated to the memory of Alfréd Rényi, II, pp. 41–57.

Czado, C. and T. Schmidt (2011). *Mathematische Statistik*. Springer-Verlag.

Daley, D. J. and D. Vere-Jones (2003). *An introduction to the theory of point processes. Vol. I*. Second. Probability and its Applications (New York). Elementary theory and methods. Springer-Verlag, New York, pp. xxii+469.

— (2008). *An introduction to the theory of point processes. Vol. II*. Second. Probability and its Applications (New York). General theory and structure. Springer, New York, pp. xviii+573.

Dassios, A. and H. Zhao (2013). "Exact simulation of Hawkes process with exponentially decaying intensity". In: *Electron. Commun. Probab.* 18.62, pp. 1–13.

Dette, H., P. Preuss and K. Sen (2017). "Detecting long-range dependence in nonstationary time series". In: *Electron. J. Stat.* 11.1, pp. 1600–1659.

Di Crescenzo, A., B. Martinucci and A. Meoli (2016). "A fractional counting process and its connection with the Poisson process". In: *ALEA Lat. Am. J. Probab. Math. Stat.* 13.1, pp. 291–307.

Dobiński, G. (1877). "Summierung der Reihe $\sum \frac{n^m}{n!}$ für $m = 1, 2, 3, 4, 5, \ldots$". In: *Archiv der Mathematik und Physik* 61, pp. 333–336.

Durrett, R. (2010). *Probability: theory and examples*. Fourth. Vol. 31. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, pp. x+428.

Elstrodt, J. (2005). *Maß- und Integrationstheorie*. Fourth. Springer-Lehrbuch. [Springer Textbook]. Grundwissen Mathematik. [Basic Knowledge in Mathematics]. Springer-Verlag, Berlin, pp. xvi+434.

Embrechts, P., C. Klüppelberg and T. Mikosch (1997). *Modelling extremal events*. Vol. 33. Applications of Mathematics (New York). For insurance and finance. Springer-Verlag, Berlin, pp. xvi+645.

Engle, R. (2002). "New frontiers for ARCH models". In: *Journal of Applied Econometrics* 17.5, pp. 425–446.

Engle, R. F. (2000). "The Econometrics of Ultra-high-frequency Data". In: *Econometrica* 68.1, pp. 1–22.

Engle, R. F. and J. R. Russell (1998). "Autoregressive conditional duration: a new model for irregularly spaced transaction data". In: *Econometrica* 66.5, pp. 1127–1162.

Erdélyi, A. et al. (1981). *Higher transcendental functions. Vol. III.* Based on notes left by Harry Bateman, Reprint of the 1955 original. Robert E. Krieger Publishing Co., Inc., Melbourne, Fla., pp. xvii+292.

Errais, E., K. Giesecke and L. R. Goldberg (2010). "Affine point processes and portfolio credit risk". In: *SIAM J. Financial Math.* 1.1, pp. 642–665.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney, pp. xxiv+669.

Filimonov, V. and D. Sornette (2012). "Quantifying reflexivity in financial markets: Toward a prediction of flash crashes". In: *Physical Review E* 85.5, p. 056108.

Francq, C. and J.-M. Zakoïan (2010). *GARCH models.* Structure, statistical inference and financial applications. John Wiley & Sons, Ltd., Chichester, pp. xiv+489.

Freedman, D. A. (1983). "A note on screening regression equations". In: *Amer. Statist.* 37.2, pp. 152–155.

Gan, G. and S.-S. Yang (1989). "An elementary proof of the order-statistics characterization of the Poisson process". In: *Amer. Statist.* 43.1, pp. 45–46.

— (1990). "Correction: "An elementary proof of the order-statistics characterization of the Poisson process" [Amer. Statist. **43** (1989), no. 1, 45–46]". In: *Amer. Statist.* 44.1, p. 65.

Gelbaum, B. R. and J. M. H. Olmsted (1990). *Theorems and counterexamples in mathematics.* Problem Books in Mathematics. Springer-Verlag, New York, pp. xxxiv+305.

Georgii, H.-O. (2007). *Stochastik.* expanded. de Gruyter Lehrbuch. [de Gruyter Textbook]. Einführung in die Wahrscheinlichkeitstheorie und Statistik. [Introduction to probability theory and statistics]. Walter de Gruyter & Co., Berlin, pp. xii+378.

Gergely, T. and I. I. Yezhow (1973). "On a construction of ordinary Poisson processes and their modelling". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 27, pp. 215–232.

Grandell, J. (1976). *Doubly stochastic Poisson processes.* Lecture Notes in Mathematics, Vol. 529. Springer-Verlag, Berlin-New York, pp. x+234.

Gut, A. (1974). "On the moments and limit distributions of some first passage times". In: *Ann. Probability* 2, pp. 277–308.

— (2013). *Probability: a graduate course.* Second. Springer Texts in Statistics. Springer, New York, pp. xxvi+600.

Hamilton, J. D. (1994). *Time series analysis.* Princeton University Press, Princeton, NJ, pp. xvi+799.

Hardiman, S. J., N. Bercot and J.-P. Bouchaud (2013). "Critical reflexivity in financial markets: a Hawkes process analysis". In: *The European Physical Journal B* 86.442.

Hardiman, S. J. and J.-P. Bouchaud (2014). "Branching ratio approximation for the self-exciting Hawkes process". In: 90, p. 062807.

Haubold, H. J., A. M. Mathai and R. K. Saxena (2011). "Mittag-Leffler functions and their applications". In: *J. Appl. Math.* Art. ID 298628, 51.

Hautsch, N. (2012). *Econometrics of financial high-frequency data*. Springer, Heidelberg, pp. xiv+371.

Hawkes, A. and L. Adamopoulos (1973). "Cluster models for earthquakes-regional comparisons". In: *Bull. Int. Statist. Inst* 45.3, pp. 454–461.

Hawkes, A. G. (2018). "Hawkes processes and their applications to finance: a review". In: *Quant. Finance* 18.2, pp. 193–198.

Hewlett, P. (2006). "Clustering of order arrivals, price impact and trade path optimisation". In: *Workshop on Financial Modeling with Jump processes, Ecole Polytechnique*, pp. 6–8.

Hosking, J. R. M. (1981). "Fractional differencing". In: *Biometrika* 68.1, pp. 165–176.

Hurvich, C. M. and C.-L. Tsai (1989). "Regression and time series model selection in small samples". In: *Biometrika* 76.2, pp. 297–307.

Hurvich, C. M. and C.-L. Tsai (1990). "The impact of model selection on inference in linear regression". In: *The American Statistician* 44.3, pp. 214–217.

Jacod, J. and A. N. Shiryaev (2003). *Limit theorems for stochastic processes*. Second. Vol. 288. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, pp. xx+661.

Jakubowski, A. (1997). "A non-Skorohod topology on the Skorohod space". In: *Electron. J. Probab.* 2, no. 4, 21 pp.

Javed, F. and P. Mantalos (2013). "GARCH-type models and performance of information criteria". In: *Comm. Statist. Simulation Comput.* 42.8, pp. 1917–1933.

Karatzas, I. and S. E. Shreve (1988). *Brownian motion and stochastic calculus*. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, pp. xxiv+470.

Kilbas, A. A., H. M. Srivastava and J. J. Trujillo (2006). *Theory and applications of fractional differential equations*. Vol. 204. North-Holland Mathematics Studies. Elsevier Science B.V., Amsterdam, pp. xvi+523.

Kingman, J. (1964). "On doubly stochastic Poisson processes". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 60. 04. Cambridge Univ Press, pp. 923–930.

Klüppelberg, C., A. Lindner and R. Maller (2004). "A continuous-time GARCH process driven by a Lévy process: stationarity and second-order behaviour". In: *J. Appl. Probab.* 41.3, pp. 601–622.

Knuth, D. E. (1992). "Two notes on notation". In: *Amer. Math. Monthly* 99.5, pp. 403–422.

Kullback, S. and R. A. Leibler (1951). "On information and sufficiency". In: *Ann. Math. Statistics* 22, pp. 79–86.

Lallouache, M. and D. Challet (2016). "The limits of statistical significance of Hawkes processes fitted to financial data". In: *Quant. Finance* 16.1, pp. 1–11.

Laskin, N. (2003). "Fractional Poisson process". In: *Commun. Nonlinear Sci. Numer. Simul.* 8.3-4. Chaotic transport and complexity in classical and quantum dynamics, pp. 201–213.

Leonenko, N., M. M. Meerschaert, R. L. Schilling et al. (2014). "Correlation structure of time-changed Lévy processes". In: *Commun. Appl. Ind. Math.* 6.1, e–483, 22.

Leonenko, N., M. M. Meerschaert and A. Sikorskii (2013). "Correlation structure of fractional Pearson diffusions". In: *Comput. Math. Appl.* 66.5, pp. 737–745.

Leonenko, N. and E. Merzbach (2015). "Fractional Poisson fields". In: *Methodol. Comput. Appl. Probab.* 17.1, pp. 155–168.

Leonenko, N., E. Scalas and M. Trinh (2017a). "Limit Theorems for the Fractional Non-homogeneous Poisson Process". In: *ArXiv e-prints*. arXiv: `1711.08768 [math.PR]`.

Leonenko, N., E. Scalas and M. Trinh (2017b). "The fractional non-homogeneous Poisson process". In: *Statist. Probab. Lett.* 120, pp. 147–156.

Lewis, P. A. W. and G. S. Shedler (1979). "Simulation of nonhomogeneous Poisson processes by thinning". In: *Naval Res. Logist. Quart.* 26.3, pp. 403–413.

Liberman, U. (1985). "An order statistic characterization of the Poisson renewal process". In: *J. Appl. Probab.* 22.3, pp. 717–722.

Maheshwari, A. and P. Vellaisamy (2017). "Fractional Poisson Process Time-Changed by Lévy Subordinator and Its Inverse". In: *Journal of Theoretical Probability*.

Mainardi, F. and R. Gorenflo (2000). "Fractional calculus: special functions and applications". In: *Advanced special functions and applications (Melfi, 1999)*. Vol. 1. Proc. Melfi Sch. Adv. Top. Math. Phys. Aracne, Rome, pp. 165–188.

Mainardi, F., R. Gorenflo and E. Scalas (2004). "A fractional generalization of the Poisson processes". In: *Vietnam J. Math.* 32.Special Issue, pp. 53–64.

Marshall, A. W. and I. Olkin (2007). *Life distributions*. Springer Series in Statistics. Structure of nonparametric, semiparametric, and parametric families. Springer, New York, pp. xx+782.

Meerschaert, M. M., E. Nane and P. Vellaisamy (2011). "The fractional Poisson process and the inverse stable subordinator". In: *Electron. J. Probab.* 16, no. 59, 1600–1620.

Meerschaert, M. M. and H.-P. Scheffler (2004). "Limit theorems for continuous-time random walks with infinite mean waiting times". In: *J. Appl. Probab.* 41.3, pp. 623–638.

Meerschaert, M. M. and A. Sikorskii (2012). *Stochastic models for fractional calculus*. Vol. 43. de Gruyter Studies in Mathematics. Walter de Gruyter & Co., Berlin, pp. x+291.

Meyer, P. A. (1971). "Démonstration simplifiée d'un théorème de Knight". In: *Séminaire de Probabilités, V (Univ. Strasbourg, année universitaire 1969–1970), Lecture Notes in Math.* 191, pp. 191–195.

Mikosch, T. (2009). *Non-life insurance mathematics*. Second. Universitext. An introduction with the Poisson process. Springer-Verlag, Berlin, pp. xvi+432.

Mikosch, T. and C. Stărică (2004). "Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects". In: *Review of Economics and Statistics* 86.1, pp. 378–390.

Mittag-Leffler, G. M. (1903a). "Sur la nouvelle fonction $E_\alpha(x)$". In: *CR Acad. Sci. Paris* 137.2, pp. 554–558.

— (1903b). "Une généralisation de l'intégrale de Laplace-Abel". In: *CR Acad. Sci. Paris (Ser. II)* 137, pp. 537–539.

— (1905). "Sur la représentation analytique d'une branche uniforme d'une fonction monogène". In: *Acta Math.* 29.1. cinquième note, pp. 101–181.

Mohler, G. O. et al. (2011). "Self-exciting point process modeling of crime". In: *J. Amer. Statist. Assoc.* 106.493, pp. 100–108.

Møller, J. and J. G. Rasmussen (2005). "Perfect simulation of Hawkes processes". In: *Adv. in Appl. Probab.* 37.3, pp. 629–646.

— (2006). "Approximate simulation of Hawkes processes". In: *Methodol. Comput. Appl. Probab.* 8.1, pp. 53–64.

Muni Toke, I. and F. Pomponio (2012). "Modelling Trades-Through in a Limit Order Book Using Hawkes Processes". In: *Economics: The Open-Access, Open-Assessment E-Journal* 6.2012-22.

Munkres, J. R. (2000). *Topology*. Second edition of [MR0464128]. Prentice Hall, Inc., Upper Saddle River, NJ, pp. xvi+537.

Ogata, Y. (1978). "The asymptotic behaviour of maximum likelihood estimators for stationary point processes". In: *Ann. Inst. Statist. Math.* 30.2, pp. 243–261.

— (1981). "On Lewis' simulation method for point processes". In: *Information Theory, IEEE Transactions on* 27.1, pp. 23–31.

— (1988). "Statistical models for earthquake occurrences and residual analysis for point processes". In: *Journal of the American Statistical association* 83.401, pp. 9–27.

Omi, T., Y. Hirata and K. Aihara (2017). "Hawkes process model with a time-dependent background rate and its application to high-frequency financial data". In: *Physical Review E* 96.1, p. 012303.

Ozaki, T. (1979). "Maximum likelihood estimation of Hawkes' self-exciting point processes". In: *Ann. Inst. Statist. Math.* 31.1, pp. 145–155.

Papangelou, F. (1972). "Integrability of expected increments of point processes and a related random change of scale". In: *Trans. Amer. Math. Soc.* 165, pp. 483–506.

Pernice, V. et al. (2012). "Recurrent interactions in spiking networks with arbitrary topology". In: *Physical review E* 85.3, p. 031916.

Ponta, L. et al. (2012). "Modeling non-stationarities in high-frequency financial time series". In: *ArXiv e-prints*. arXiv: `1212.0479 [q-fin.ST]`.

Prabhakar, T. R. (1971). "A singular integral equation with a generalized Mittag Leffler function in the kernel". In: *Yokohama Math. J.* 19, pp. 7–15.

Rambaldi, M., P. Pennesi and F. Lillo (2015). "Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach". In: *Physical Review E* 91.1, p. 012819.

Rapach, D. E. and J. K. Strauss (2008). "Structural breaks and GARCH models of exchange rate volatility". In: *J. Appl. Econometrics* 23.1, pp. 65–90.

Reed, M. and B. Simon (1980). *Methods of modern mathematical physics. I.* Second. Functional analysis. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, pp. xv+400.

Reinhart, A. (2017). "A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications". In: *ArXiv e-prints*. arXiv: `1708.02647 [stat.ME]`.

Reynaud-Bouret, P., V. Rivoirard et al. (2014). "Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis". In: *Journal of Mathematical Neuroscience*, 4:3.

Reynaud-Bouret, P. and S. Schbath (2010). "Adaptive estimation for Hawkes processes; application to genome analysis". In: *Ann. Statist.* 38.5, pp. 2781–2822.

Richter, W. (1965). "Übertragung von Grenzaussagen für Folgen von zufälligen Grössen auf Folgen mit zufälligen Indizes". In: *Teor. Verojatnost. i Primenen* 10. This article has appeared in English translation [Theor. Probability Appl. 10 (1965), 74–84], pp. 82–94.

Rudin, W. (1976). *Principles of mathematical analysis*. Third. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, pp. x+342.

Samorodnitsky, G. and M. S. Taqqu (1994). *Stable non-Gaussian random processes*. Stochastic Modeling. Stochastic models with infinite variance. Chapman & Hall, New York, pp. xxii+632.

Scalas, E. (2007). "Mixtures of compound Poisson processes as models of tick-by-tick financial data". In: *Chaos Solitons Fractals* 34.1, pp. 33–40.

Schneider, M., F. Lillo and L. Pelizzon (2018). "Modelling illiquidity spillovers with Hawkes processes: an application to the sovereign bond market". In: *Quant. Finance* 18.2, pp. 283–293.

Schwarz, G. (1978). "Estimating the dimension of a model". In: *Ann. Statist.* 6.2, pp. 461–464.

Serfozo, R. F. (1972a). "Conditional Poisson processes". In: *J. Appl. Probability* 9, pp. 288–302.

— (1972b). "Processes with conditional stationary independent increments". In: *J. Appl. Probability* 9, pp. 303–315.

Shibata, R. (1976). "Selection of the order of an autoregressive model by Akaike's information criterion". In: *Biometrika* 63.1, pp. 117–126.

Simon, B. (1979). *Functional integration and quantum physics*. Vol. 86. Pure and Applied Mathematics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, pp. ix+296.

Sin, C.-Y. and H. White (1996). "Information criteria for selecting possibly misspecified parametric models". In: *J. Econometrics* 71.1-2, pp. 207–225.

Skorokhod, A. V. (1956). "Limit theorems for stochastic processes". In: *Teor. Veroyatnost. i Primenen.* 1, pp. 289–319.

Snyder, D. L. and M. I. Miller (1991). "Random Point Processes in Time and Space". In: Springer. Chap. 4.4.

Stărică, C. and C. Granger (2005). "Nonstationarities in stock returns". In: *Review of economics and statistics* 87.3, pp. 503–522.

Vidmar, M. (2016). "Another characterization of homogeneous Poisson processes". In: *ArXiv e-prints*. arXiv: `1610.07147 [math.PR]`.

Watanabe, S. (1964). "On discontinuous additive functionals and Lévy measures of a Markov process". In: *Japan. J. Math.* 34, pp. 53–70.

Whitt, W. (2002). *Stochastic-process limits*. Springer Series in Operations Research. An introduction to stochastic-process limits and their application to queues. Springer-Verlag, New York, pp. xxiv+602.

Wilks, S. S. (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses". In: *The Annals of Mathematical Statistics* 9.1, pp. 60–62.

Wiman, A. (1905a). "Über den Fundamentalsatz in der Teorie der Funktionen $E^a(x)$". In: *Acta Math.* 29.1, pp. 191–201.

— (1905b). "Über die Nullstellen der Funktionen $E^a(x)$". In: *Acta Math.* 29.1, pp. 217–234.

Wong, R. (2001). *Asymptotic approximations of integrals*. Vol. 34. Classics in Applied Mathematics. Corrected reprint of the 1989 original. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xviii+543.

Yannaros, N. (1994). "Weibull renewal processes". In: *Ann. Inst. Statist. Math.* 46.4, pp. 641–648.

Young, G. A. and R. L. Smith (2005). *Essentials of statistical inference*. Vol. 16. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, pp. x+225.

Zhang, P. (1993). "On the convergence rate of model selection criteria". In: *Comm. Statist. Theory Methods* 22.10, pp. 2765–2775.

Appendix A

Regular variation and Tauberian theorems

The following definitions and results can be found in Bingham, Goldie and Teugels 1989.

Definition A.1 (Slow variation)**.** Let l be a positive measurable function, defined on some neighborhood $[X, \infty)$ of infinity and satisfying

$$\frac{l(\lambda x)}{l(x)} \longrightarrow 1 \quad \text{for} \quad x \longrightarrow \infty, \qquad \forall\, \lambda > 0.$$

Then l is said to be *slowly varying* (in Karamata's sense). We write $l \in \mathcal{R}_0$.

Definition A.2 (Regular variation)**.** A measurable function $f > 0$ satisfying

$$\frac{f(\lambda x)}{f(x)} \longrightarrow \lambda^\rho \quad \text{for} \quad x \longrightarrow \infty, \qquad \forall\, \lambda > 0 \tag{A.1}$$

is called *regular varying* of index ρ. We write $f \in \mathcal{R}_\rho$.

Theorem A.1 (Characterization theorem)**.** If $f > 0$ is measurable and (A.1) holds for all λ in a set of positive measure, then

(i) The limit in (A.1) holds for all $\lambda > 0$.

(ii) There exists a real number ρ with $g(\lambda) \equiv \lambda^\rho\ \forall \lambda > 0$.

(iii) $f(x) = x^\rho l(x)$ with l slowly varying.

Proposition A.2 (Properties of regular variation)**.**

(i) If $f \in \mathcal{R}_\rho$, then $f^\alpha \in \mathcal{R}_{\alpha\rho}, \quad \alpha \in \mathbb{R}$.

(ii) If $f_i \in \mathcal{R}_{\rho_i}\,(i = 1, 2)$ and $f_2(x) \to \infty$ for $x \to \infty$, then $f_1(f_2(x)) \in \mathcal{R}_{\rho_1\rho_2}$.

(iii) If $f_i \in \mathcal{R}_{\rho_i}\,(i = 1, 2)$, then $f_1(x) + f_2(x) \in \mathcal{R}_\rho$; $\rho = \max(\rho_1, \rho_2)$.

(iv) If $f_i \in \mathcal{R}_{\rho_i}\,(i = 1, \ldots, k)$ and $r(x_!, \ldots, x_k)$ a rational function with positive coefficients, then $r(f_1(x), \ldots, f_k(x)) \in \mathcal{R}$.

Theorem A.3 (Karamata's Tauberian theorem)**.** Let U be a non-decreasing right-continuous function on $\mathbb{R}$ with $U(x) = 0$, $\forall x < 0$ and let

$$\tilde{U}(s) := \int_0^\infty \mathrm{e}^{-sx} \mathrm{d}U(x)$$

If l varies slowly and $c \geq 0, \rho \geq 0$, the following are equivalent:

$$U(x) \sim cx^\rho l(x) \frac{1}{\Gamma(1+\rho)}, \qquad (x \to \infty) \tag{A.2}$$

$$\tilde{U}(s) \sim cs^{-\rho} l\left(\frac{1}{s}\right), \qquad (s \to 0+). \tag{A.3}$$

When $c = 0$, (A.2) is to be interpreted as $U(x) = o(x^\rho l(x))$; similarly for (A.3).

Theorem A.4 (A Tauberian theorem)**.** Let μ be a (positive) Borel measure on $[0, \infty)$ and suppose

$$\int \mathrm{e}^{-tx} \mathrm{d}\mu < \infty \quad \forall\, t > 0$$

and that for some $\gamma \geq 0$ and $D \geq 0$ it holds that

$$\lim_{t \to 0+} t^\gamma \int \mathrm{e}^{-tx} \mathrm{d}\mu(x) = D.$$

Then

$$\lim_{a \to \infty} a^{-\gamma} \mu([0, a)) = \frac{D}{\Gamma(\gamma + 1)}$$

(pp. 108-110 in Simon 1979).

Proposition A.5 (Integrating asymptotic relations)**.** If l is slowly varying, X is so large that $l(x)$ is locally bounded in $[X, \infty)$ and $\alpha > -1$, then

$$\int_X^x t^\alpha l(t)\, \mathrm{d}t \sim x^{\alpha+1} \frac{l(x)}{1+\alpha}, \quad (x \to \infty).$$

Theorem A.6 (Monotone density theorem)**.** Let $U(x) = \int_0^\infty u(y)\, \mathrm{d}y$. If $U(x) \sim cx^\rho l(x)$ for $x \to \infty$, where $c \in \mathbb{R}$, $\rho \in \mathbb{R}$, $l \in \mathcal{R}_\rho$ and if u is ultimately monotone,

then $u(x) \sim c\rho x^{\rho-1} l(x)$.

Remark A.1. Note that positivity of c and ρ is needed to say that U and u are regular varying.

Appendix B

Tables

Table B.1: Results of the MSE calculations for the ACD model

		MSE(ω)	MSE(α_1)	MSE(α_2)	MSE(β_1)	MSE(β_2)
ACD(1,1)	T=250	3.7508	0.0023	–	0.0231	–
	T=500	1.8887	0.0010	–	0.0108	–
	T=1000	0.3591	0.0005	–	0.0025	–
	T=2000	0.1245	0.0002	–	0.0010	–
ACD(1,2)	T=250	14.5255	0.0036	–	0.4748	0.4282
	T=500	3.7468	0.0019	–	0.3039	0.2681
	T=1000	0.6259	0.0010	–	0.1869	0.1606
	T=2000	0.1905	0.0005	–	0.0809	0.0681
ACD(2,1)	T=250	0.8491	0.0063	0.0108	0.0130	–
	T=500	0.2664	0.0032	0.0050	0.0053	–
	T=1000	0.0916	0.0014	0.0026	0.0023	–
	T=2000	0.0418	0.0007	0.0012	0.0011	–
ACD(2,2)	T=250	6.4135	0.0067	0.0102	0.3165	0.2445
	T=500	1.1077	0.0032	0.0061	0.2722	0.2031
	T=1000	0.3730	0.0014	0.0041	0.2086	0.1526
	T=2000	0.1512	0.0006	0.0026	0.1612	0.1181

Table B.2: Model selection results based on ACD(1,1) data samples: Given 1000 samples of size $T \in \{250, 500, 1000, 2000\}$ each column gives the percentage of cases in which the different IC selected the models ACD(1,1), ACD(1,2), ACD(2,1) and ACD(2,2) respectively. The bold numbers give the largest percentage per row.

		ACD(1,1)	ACD(1,2)	ACD(2,1)	ACD(2,2)
T=250	AIC	**58.7**	23.6	9.9	7.8
	BIC	**90.2**	7	2.1	0.7
	HQ	**77.9**	14.6	4.8	2.7
T=500	AIC	**62.9**	20.4	10.9	5.8
	BIC	**93.6**	4.7	1.6	0.1
	HQ	**82.6**	11.5	4.9	1
T=1000	AIC	**67.5**	16.4	11	5.1
	BIC	**97.4**	1.8	0.8	0
	HQ	**87.2**	7.5	4.8	0.5
T=2000	AIC	**71.3**	13.1	9.7	5.9
	BIC	**97.7**	1.6	0.6	0.1
	HQ	**91.5**	4.4	3	1.1

Table B.3: Model selection results based on ACD(1,2) data samples: Given 1000 samples of size $T \in \{250, 500, 1000, 2000\}$ each column gives the percentage of cases in which the different IC selected the models ACD(1,1), ACD(1,2), ACD(2,1) and ACD(2,2) respectively. The bold numbers give the largest percentage per row.

		ACD(1,1)	ACD(1,2)	ACD(2,1)	ACD(2,2)
T=250	AIC	**58.6**	24.7	9.6	7.1
	BIC	**91.5**	6.5	1.3	0.7
	HQ	**78.6**	14.8	3.7	2.9
T=500	AIC	**60.6**	25.1	10.3	4
	BIC	**94.7**	4.3	0.7	0.3
	HQ	**81.2**	13.5	4.5	0.8
T=1000	AIC	**52.7**	27.8	15.2	4.3
	BIC	**92.6**	5.1	2.3	0
	HQ	**76**	14.7	8.8	0.5
T=2000	AIC	**41.5**	35.6	18	4.9
	BIC	**88.4**	6.7	4.9	0
	HQ	**67.6**	20.4	11.6	0.4

Table B.4: Model selection results based on ACD(2,1) data samples: Given 1000 samples of size $T \in \{250, 500, 1000, 2000\}$ each column gives the percentage of cases in which the different IC selected the models ACD(1,1), ACD(1,2), ACD(2,1) and ACD(2,2) respectively. The bold numbers give the largest percentage per row.

		ACD(1,1)	ACD(1,2)	ACD(2,1)	ACD(2,2)
T=250	AIC	**36.2**	20.9	31.8	11.1
	BIC	**73.7**	8.9	16.8	0.6
	HQ	**52.4**	16.3	28.1	3.2
T=500	AIC	19.1	20.7	**50**	10.2
	BIC	**59.9**	10.5	29	0.6
	HQ	36.5	16.4	**43.8**	3.3
T=1000	AIC	7.4	16.7	**64.8**	11.1
	BIC	35.6	11.9	**52.1**	0.4
	HQ	17.1	15.7	**63.7**	3.5
T=2000	AIC	1.2	12.7	**74.2**	11.9
	BIC	6.8	12.9	**80.1**	0.2
	HQ	2.2	14.2	**81.6**	2

Table B.5: Model selection results based on ACD(2,2) data samples: Given 1000 samples of size $T \in \{250, 500, 1000, 2000\}$ each column gives the percentage of cases in which the different IC selected the models ACD(1,1), ACD(1,2), ACD(2,1) and ACD(2,2) respectively. The bold numbers give the largest percentage per row.

		ACD(1,1)	ACD(1,2)	ACD(2,1)	ACD(2,2)
T=250	AIC	**56.7**	15.8	18.8	8.7
	BIC	**89.7**	5.3	4.5	0.5
	HQ	**74**	11.5	11.7	2.8
T=500	AIC	**57.2**	13.6	19.1	10.1
	BIC	**92.1**	2.9	4.6	0.4
	HQ	**78.4**	8	11.4	2.2
T=1000	AIC	**48.4**	13.1	23.4	15.1
	BIC	**91.5**	2.7	5.7	0.1
	HQ	**74**	6.9	16.1	3
T=2000	AIC	34.2	9.7	**37.2**	18.9
	BIC	**86.1**	1.8	11.5	0.6
	HQ	**59.7**	6.8	26.5	7

Table B.6: Absolute RMSE values for MLE of the exponential Hawkes models of order $P \in \{1, 2, 3\}$ using Parameter Set 1 with varying time horizons $T \in \{500, 1000, 2000, 5000\}$. The order of the model which was used for simulation coincides with the model used for fitting. Thus, the true parameter values are known and the RMSE is expected to decrease as the MLE improves.

		μ	α_1	α_2	α_3	β_1	β_2	β_3	Average sample size
P=1	T=500	0.039664	0.42256	–	–	0.44731	–	–	2483
	T=1000	0.02763	0.32473	–	–	0.32928	-	–	5019
	T=2000	0.018738	0.21756	–	–	0.22197	-	–	9977
	T=5000	0.011276	0.13846	–	–	0.14544	-	–	24962
P=2	T=500	0.071796	6.0214	12.399	–	37.322	44.92	–	470
	T=1000	0.061258	0.00085434	7.9865	–	0.0023803	18.956	–	1121
	T=2000	0.049989	0.00020347	4.7732	–	0.00045077	11.415	–	2977
	T=5000	0.042938	0.00010551	2.3918	–	0.00018339	5.9634	–	12883
P=3	T=500	0.07713	0.3232	1.3408	9.8602	1.5085	9.7124	30.678	929
	T=1000	0.061804	0.00036118	0.29469	6.2401	0.0021189	0.76784	18.334	2207
	T=2000	0.051527	0.0001279	0.19655	3.8663	0.00058096	0.49288	11.77	5840
	T=5000	0.045562	0.000055755	0.10766	1.8741	0.00019317	0.27921	5.6723	25017

Table B.7: Relative MSE values for MLE of the exponential Hawkes models of order $P \in \{1, 2, 3\}$ using Parameter Set 1 with varying time horizons $T \in \{500, 1000, 2000, 5000\}$. The order of the model which was used for simulation coincides with the model used for fitting. Thus, the true parameter values are known and the RMSE is expected to decrease as the MLE improves. The values are given in percent.

		μ	α_1	α_2	α_3	β_1	β_2	β_3	Average sample size
P=1	T=500	7.9328	4.6951	–	–	4.4731	–	–	2483
	T=1000	5.526	3.6082	–	–	3.2928	–	–	5019
	T=2000	3.7475	2.4173	–	–	2.2197	–	–	9977
	T=5000	2.2551	1.5384	–	–	1.4544	–	–	24962
P=2	T=500	14.359	912330	12.399	–	3732200	14.973	–	470
	T=1000	12.252	129.45	7.9865	–	238.03	6.3186	–	1121
	T=2000	9.9978	30.829	4.7732	–	45.077	3.805	–	2977
	T=5000	8.5877	15.986	2.3918	–	18.339	1.9878	–	12883
P=3	T=500	15.426	97939	40.63	9.8602	150850	97.124	10.226	929
	T=1000	12.361	109.45	8.9301	6.2401	211.89	7.6784	6.1113	2207
	T=2000	10.305	38.758	5.956	3.8663	58.096	4.9288	3.9233	5840
	T=5000	9.1124	16.895	3.2624	1.8741	19.317	2.7921	1.8908	25017

Table B.8: Model selection for simulated data of an exponential Hawkes model of order P=1 using Parameter Set 1 with varying time horizons $T \in \{500, 1000, 2000, 5000\}$. The numbers indicate how often the model order $P \in \{1, 2, 3\}$ is selected among the 1000 samples and are given in percent. Bold numbers show which model was selected most often.

	Time horizon	P=1	P=2	P=3	Average sample size
AIC	T=500	**92.8**	6.9	0.3	2483
	T=1000	**91.6**	7.9	0.5	5019
	T=2000	**92.1**	7.6	0.3	9977
	T=5000	**93.7**	6.1	0.2	24962
BIC	T=500	**99.8**	0.2	0	2483
	T=1000	**100**	0	0	5019
	T=2000	**100**	0	0	9977
	T=5000	**100**	0	0	24962
HQ	T=500	**98.9**	1.1	0	2483
	T=1000	**98.6**	1.2	0.2	5019
	T=2000	**99.2**	0.8	0	9977
	T=5000	**99.7**	0.3	0	24962

Table B.9: Model selection for simulated data of an exponential Hawkes model of order P=2 using Parameter Set 1 with varying time horizons $T \in \{500, 1000, 2000, 5000\}$. The numbers indicate how often the model order $P \in \{1, 2, 3\}$ is selected among the 1000 samples and are given in percent. Bold numbers show which model was selected most often.

	Time horizon	P=1	P=2	P=3	Average sample size
AIC	T=500	48.2	**50.3**	1.5	470
	T=1000	0.2	**99**	0.8	1121
	T=2000	0	**96.9**	3.1	2977
	T=5000	0	**93.7**	6.3	12883
BIC	T=500	**94.5**	5.4	0.1	470
	T=1000	10.3	**89.7**	0	1121
	T=2000	0	**100**	0	2977
	T=5000	0	**100**	0	12883
HQ	T=500	**76.9**	22.9	0.2	470
	T=1000	2.1	**97.8**	0.1	1121
	T=2000	0	**99.4**	0.6	2977
	T=5000	0	**99.6**	0.4	12883

Table B.10: Model selection for simulated data of an exponential Hawkes model of order P=3 using Parameter Set 1 with varying time horizons $T \in \{500, 1000, 2000, 5000\}$. The numbers indicate how often the model order $P \in \{1, 2, 3\}$ is selected among the 1000 samples and are given in percent. Bold numbers show which model was selected most often.

	Time horizon	P=1	P=2	P=3	Average sample size
AIC	T=500	0	**53.7**	46.3	929
	T=1000	0	0.2	**99.8**	2207
	T=2000	0	0	**100**	5840
	T=5000	0	0	**100**	25017
BIC	T=500	0	**96.5**	3.5	929
	T=1000	0	25	**75**	2207
	T=2000	0	0	**100**	5840
	T=5000	0	0	**100**	25017
HQ	T=500	0	**81.6**	18.4	929
	T=1000	0	4.7	**95.3**	2207
	T=2000	0	0	**100**	5840
	T=5000	0	0	**100**	25017

Table B.11: Absolute MSE values for MLE of the exponential Hawkes models of order $P = 2$ using Parameter Set 2 with varying time horizons $T \in \{600, 900, 1800, 3600, 7200, 21600\}$. The order of the model which was used for simulation coincides with the model used for fitting. Thus, the true parameter values are known and the RMSE is expected to decrease as the MLE improves.

	μ	α_1	α_2	β_1	β_2	Average sample size
T=600	0.030176	0.076208	12489000	0.17692	27063000000	135
T=900	0.025411	0.052971	1704000	0.11054	15861000000	205
T=1800	0.016031	0.023916	0.078183	0.049322	4.0458	417
T=3600	0.010572	0.0095225	0.039458	0.020001	0.15516	853
T=7200	0.0070848	0.0055058	0.025844	0.011919	0.088505	1708
T=21600	0.0039548	0.0030036	0.014737	0.006448	0.051022	5144

Table B.12: Relative MSE values for MLE of the exponential Hawkes models of order $P = 2$ using Parameter Set 2 with varying time horizons $T \in \{600, 900, 1800, 3600, 7200, 21600\}$. The order of the model which was used for simulation coincides with the model used for fitting. Thus, the true parameter values are known and the RMSE is expected to decrease as the MLE improves. The values are given in percent.

	μ	α_1	α_2	β_1	β_2	Average sample size
T=600	60.353	432.53	4460300000	371.53	4059400000000	135
T=900	50.822	300.64	608590000	232.14	2379200000000	205
T=1800	32.061	135.74	27.923	103.58	606.87	417
T=3600	21.144	54.047	14.092	42.003	23.275	853
T=7200	14.17	31.249	9.2299	25.03	13.276	1708
T=21600	7.9096	17.047	5.263	13.541	7.6533	5144

Table B.13: Model selection for simulated data of an exponential Hawkes model of order P=2 using Parameter Set 2 with varying time horizons $T \in \{600, 900, 1800, 3600, 7200, 21600\}$. The numbers indicate how often the model order $P \in \{1, 2, 3\}$ is selected among the 1000 samples and are given in percent. Bold numbers show which model was selected most often.

	Time horizon	P=1	P=2	P=3	Average sample size
AICc/AIC	T=600	**52.4**	47	0.6	135
	T=900	36.6	**62.5**	0.9	205
	T=1800	6.8	**90.2**	3	417
	T=3600	0	**96.9**	3.1	853
	T=7200	0	**94.7**	5.3	1708
	T=21600	0	**94.1**	5.9	5144
AIC	T=600	49.3	**50.1**	0.6	135
	T=900	34.8	**64.1**	1.1	205
	T=1800	6.7	**90.2**	3.1	417
	T=3600	0	**96.9**	3.1	853
	T=7200	0	**94.7**	5.3	1708
	T=21600	0	**94.1**	5.9	5144
BIC	T=600	**86.5**	13.5	0	135
	T=900	**79.8**	20.2	0	205
	T=1800	42.7	**57.2**	0.1	417
	T=3600	5	**95**	0	853
	T=7200	0.1	**99.9**	0	1708
	T=21600	0	**100**	0	5144
HQ	T=600	**69**	30.8	0.2	135
	T=900	**55.6**	44.4	0	205
	T=1800	17.5	**81.8**	0.7	417
	T=3600	1	**98.7**	0.3	853
	T=7200	0	**98.9**	1.1	1708
	T=21600	0	**99.2**	0.8	5144

Appendix C

Code manual

C.1 Compound Poisson type models

`compPoissonInterval` **Compound Poisson process on an interval**

Usage

```
function [ res ] = compPoissonInterval( T, lambda, mu, sigma2 )
```

Description

Simulates a compound Poisson process with parameters λ, μ and σ^2 in the interval [0,T]

Inputs

`T`	time horizon
`lambda`	λ parameter of the Exp(λ) distributed waiting times
`mu`	μ parameter of the $\mathcal{N}(\mu, \sigma^2)$ distributed jumps
`sigma2`	σ^2 parameter of the $\mathcal{N}(\mu, \sigma^2)$ distributed jumps

Value

`res`	matrix containing jump times in the first column and the corresponding values of the process (cumulative sum of the jump heights) in the second column

nonhomPoisson — Compound Poisson type process with varying lambda

Usage

```
function [ res, trueLambda ] = nonhomPoisson( grid, lambdaRange, mu,
                                              sigma2 )
```

Description

Simulates a compound Poisson type model with varying λ on the time interval $[0, 1]$

Inputs

`grid`	vector containing time points of the grid on which the process is simulated
`lambdaRange`	vector of the form $[\lambda_{\min}, \lambda_{\max}]$ containing smallest and largest possible value of lambda, which is parametrized by a parabola: $\lambda(t) := 4(\lambda_{\max} - \lambda_{\min})(t - 0.5)^2 + \lambda_{\min}$, $\forall t \in [0, 1]$ and $\lambda_{\min}, \lambda_{\max} > 0$ constant.
`mu`	μ parameter of the $\mathcal{N}(\mu, \sigma^2)$ distributed jumps
`sigma2`	σ^2 parameter of the $\mathcal{N}(\mu, \sigma^2)$ distributed jumps

Value

`res`	table consisting of two columns containing jump times and corresponding values of the process
`trueLambda`	table consisting of two columns containing the simulation grid and the corresponding true intensity values per interval.

ICSimInMem — Simulation of data for information criteria experiment

Usage

```
function [ data, trueLambda ] = ICSimInMem( N, gridSim, lambdaRange)
```

Description

Simulates N samples of the compound Poisson type model on the time interval $[0, 1]$

Inputs

`N`	number of samples
`gridSim`	vector containing time points of the grid on which the process is simulated
`lambdaRange`	vector of the form $[\lambda_{\min}, \lambda_{\max}]$ containing smallest and largest possible value of λ. (See also `nonhomPoisson` for parametrization.)

Value

`data`	cell of N sample paths
`trueLambda`	cell of N tables containing true intensity values

`likelihoodNP` Evaluation of the log-likelihood function of the (Dλ)-model

---

Usage

```
function [res] = likelihoodNP( lambda, mu, sigma2, marks,
                                                interval_length )
```

Description

Calculates for a time interval $[t_{i-1}, t_i)$

$$\mathcal{L}_i^D(\lambda_i, \mu_i, \sigma_i) = -\lambda_i(t_i - t_{i-1}) + \ln(\lambda_i)N_i(t_i) + \underbrace{\sum_{k=1}^{N_i(t_i)} \ln(p_{\mu_i,\sigma_i}(R_k^{(i)}))}_{:=\mathcal{L}_i^{\text{part}}}. \tag{C.1}$$

Remark: This function is written in C (MATLAB mex function) and needs to be compiled first using the command `mex likelihoodNP.c`.

Inputs

`lambda`	λ parameter of the Exp(λ) distributed waiting times
`mu`	μ parameter of the $\mathcal{N}(\mu, \sigma^2)$ distributed jumps
`sigma2`	σ^2 parameter of the $\mathcal{N}(\mu, \sigma^2)$ distributed jumps
`marks`	vector containing jump sizes
`interval_length`	length of the interval: $t_i - t_{i-1}$

Value

`res`	vector with two entries: res(1) contains the value of $\mathcal{L}_i^D$ and res(2) contains $\mathcal{L}_i^{\text{part}}$.

`fitInterval` **Parameter estimation for compound Poisson process**

Usage

```
function [ res ] = fitInterval( marks, interval_length )
```

Description

Estimates the parameters of the compound Poisson process on an interval: Estimators are given by closed formulas.

Inputs

`marks`	vector containing jump sizes
`interval_length`	length of the interval

Value

`res`	vector containing three entries for the parameters λ, μ and σ^2 respectively.

`fit` **Fitting algorithm for the (Dλ)-model**

Usage

```
function [ res ] = fit( grid, input )
```

Description

Estimates parameters and calculates log-likelihood value of the compound Poisson type model (discrete lambda) on a given grid

Inputs

`grid`	vector containing time points of the grid used for fitting
`input`	matrix containing jump times in the first column and the corresponding values of the process (cumulative sum of the jump heights) in the second column

Value

`res`	matrix containing the parameter estimates for λ_i, μ_i and σ_i in the first three columns followed by the likelihood values $\mathcal{L}_i^D$ and $\mathcal{L}_i^{\text{part}}$ in the fourth and fifth column respectively.

`grids` Generation of time grids

Usage

```
function [ res ] = grids( startTime, endTime, deltaMax )
```

Description

Generates cell array containing grids of different grid sizes

Inputs

`startTime`	start value of the time grid(s) for estimation
`endTime`	final value of the time grid(s) for estimation
`deltaMax`	finest eligible grid size

Value

`res`	cell array containing vectors representing the time grids

---

`gridAIC` Calculation of information criteria for the (Dλ)-model

Usage

```
function [ IC, numParam, fits ] = gridAIC( startTime, endTime, deltaMax,
data )
```

Description

Calculates information criteria for the (Dλ)-model after running a maximum likelihood estimation by calling `fit`

Inputs

`startTime, endTime, deltaMax`	inputs to generate cell array of grids by calling `grids`
`data`	matrix containing jump times and process value in two columns

Value

`IC`	1-by-3 cell containing the values of information criteria AIC, BIC HQ
`numParam`	number of parameters corresponding to the number of subintervals used in the fitting grid
`fits`	cell of results of the MLE for each grid configuration: each cell entry contains a matrix with 6 columns: the first column contains the right end interval values of the fitting grid, the second to fourth column the parameter estimates for λ, μ and σ^2 for each subinterval, the fifth columns gives the total value of the maximum log-likelihood value and the last column contains the part of the maximum log-likelihood attributable too the marks, i.e. the parameters μ and σ^2, which is needed later on.

`IChist` Model selection for the (Dλ)-model

Usage

```
function [ IC, numParam, fits ] = IChist( data )
```

Description

Model selection for samples given in `data`

Inputs

`data`	one data sample from previously generated data set using `ICSim`

Value

`IC`	vector containing number of parameters chosen by the AIC, BIC and HQ in each entry
`numParam`	see `gridAIC`
`fits`	see `gridAIC`

`partialLogLikNP` **Calculation of partial log-likelihood of the (Pλ)-model**

Usage

```
function [ res ] = partialLogLikNP( a, b, c, jumpTimes )
```

Description

Calculates for a time interval $[t_{i-1}, t_i)$ the part of the log-likelihood function attributable to the parametrized intensity λ: $-\mathcal{L}_i^{\text{param}}$ (negative sign for convenience when using minimization algorithm)

$$\mathcal{L}_i^P(a,b,c,\mu_i,\sigma_i) = \underbrace{-\lambda_{a,b,c}(t_{i-1})(t_i - t_{i-1}) + \ln(\lambda_{a,b,c}(t_{i-1}))N_i(t_i)}_{:=\mathcal{L}_i^{\text{param}}} + \underbrace{\sum_{k=1}^{N_i(t_i)} \ln(p_{\mu_i,\sigma_i}(R_k^{(i)}))}_{=\mathcal{L}_i^{\text{part}}}$$

with

$$\lambda_{a,b,c}(t) = at^2 + bt + c, \quad t \in [0,1].$$

Remark: This function is written in C (MATLAB mex function) and needs to be compiled first using the command `mex partialLogLikNP.c`. Call this function in MATLAB using `partialLogLikWrapper`.

Inputs

`a, b, c`	parameters of the parabola representation of λ
`jumpTimes`	vector containing jump times given in the data

Value

`res`	vector containing the function value of the partial log-likelihood as well as the gradient: $-(\mathcal{L}_i^{\text{param}}, \partial_a\mathcal{L}_i^{\text{param}}, \partial_b\mathcal{L}_i^{\text{param}}, \partial_c\mathcal{L}_i^{\text{param}})$

`partialLogLikWrapper` — Wrapper to call partialLogLikNP

Usage

```
function [ f, g ] = partialLogLikWrapper( a, b, c, jumpTimes )
```

Description

Calls the pre-compiled function partialLogLikNP.

Inputs

See `partialLogLikNP`

Value

`f`	negative function value $-\mathcal{L}_i^{\text{param}}$
`g`	negative gradient $-(\partial_a \mathcal{L}_i^{\text{param}}, \partial_b \mathcal{L}_i^{\text{param}}, \partial_c \mathcal{L}_i^{\text{param}})$

`constraints` — Constraints for the minimization routine

Usage

```
function [ c, ceq ] = constraints( p )
```

Description

Defines inequality and equality constraints for the MATLAB minimization function `fmincon`. It partly ensures that the parabola fitted to the empirical data for λ is non-negative by demanding that the extremal point of the parabola is non-negative. Additional constraints are implemented in the function `fit2`.
Remark: This function is called by `fmincon` and is not intended as a standalone function.

Inputs

`p`	parameter vector

Value

c	inequality constraint
ceq	equality constraint

`fitLambda` Fitting of λ in the (Pλ)-model

Usage

```
function [ x, value, exitflag ] = fitLambda( p0, grid, numJumps, mode )
```

Description

Fits polynomially parametrized step function to (intensity) data depending on case specified in the variable `order`

Inputs

`p0`	initial parameter values for `a, b, c`
`grid`	vector containing nodes of the fitting grid
`numJumps`	vector containing number of events per subinterval in the corresponding fitting grid
`order`	options: 3: fits a quadratic function using `fmincon` for constraint minimization any other value: fits a quadratic function using `fminunc` for unconstrained optimization

Value

`x`	(potential) minimizer
`value`	(potential) minimum value
`exitflag`	flag given by MATLAB minimization functions `fminunc` and `fmincon` indicating whether the minimization procedure finished successfully.

`fit2` Fitting algorithm for the (Pλ)-model

Usage

```
function [ res ] = fit2( prefit, input )
```

Description

Estimates parameters and calculates log-likelihood value of the (Pλ)-model on a given grid

Inputs

`prefit`	estimation values given by `fit` on the (Dλ)-model
`input`	data matrix containing jump times in the first column.

Value

`res`	cell of results of the MLE for each grid configuration: each cell entry contains a matrix with 9 columns: the first column contains the right end interval values of the fitting grid, the second to seventh column the parameter estimates for a, b, c, λ, μ and σ^2 for each subinterval, the fifth columns gives the total value of the partial log-likelihood value for the λ parameter and the last column contains the part of the maximum log-likelihood attributable too the marks, i.e. the parameters μ and σ^2,which is needed later on.

`gridAIC2` **Calculation of information criteria**

Usage

```
function [ IC, numParam, fits ] = gridAIC2( prefit, data )
```

Description

Calculates information criteria for the (Pλ)-model based on (Dλ) fitting results

Inputs

`prefit`	data cell containing (Dλ) fitting results, each cell corresponding to a certain fitting grid
`data`	matrix containing jump times and process value in two columns

Value

`IC`	1-by-3 cell containing the values of information criteria AIC, BIC HQ
`numParam`	number of parameters corresponding to the number of subintervals
`fits`	see output of `fit2`

`IChist2` **Model selection**

Usage

```
function [ IC, numParam, fits] = IChist2( prefit, data )
```

Description

Model selection for samples given in `data`

Inputs

`data`	one data sample from previously generated data set using `ICSim`

Value

`IC`	vector containing number of parameters chosen by the AIC, BIC and HQ in each entry
`numParam`	see `gridAIC2`
`fits`	see `gridAIC2`

C.2 Sample code for compound Poisson type models

Listing C.1: Sample code for simulation of data

```
1  sampleSize=1000;
2  numSpaces=40;
3  N=length(sampleSize);
4  trueLambda=cell(numSpaces-1,N); %store matrix of grid and true
       values (lambda, mu, sigma)
5  data=cell(numSpaces-1,N);
6  lambdaRange=[100  10000];
7
8  for k=2:1:40
9      gridSim = linspace(0, 1, k);
10     fprintf('k = %d\n', k);
11     [data{k-1,1}, trueLambda{k-1,1}] = ICSimInMem(sampleSize,
           gridSim, lambdaRange);
12 end
```

```
save('sim_01');
```

Listing C.2: Sample code for MLE of the (Dλ)-model

```
load('sim_01')

sampleSize=1000;
numSpaces=40;
IC=cell(numSpaces-1,1);
NP=cell(numSpaces-1,sampleSize);
fits=cell(numSpaces-1,sampleSize);
numParam=cell(numSpaces-1,sampleSize);

for k=3:1:39
    ICtemp=zeros(sampleSize,3);
    for n=1:sampleSize
        fprintf('k = %d, \t n = %d \n', k, n);
        dataCell=data{k,1}{n,1};
        [ICtemp(n,:), numParam{k,n}, fits{k,n}] = IChist(
            dataCell);
    end
    IC{k,1} = ICtemp;
end
save('fit_01');
```

Listing C.3: Sample code for MLE of the (Pλ)-model: It assumes that the script for (Dλ) has already been executed.

```
load('sim_01')
%load results from D-Lambda fitting
load('fit_01')

sampleSize=1000;
numSpaces=40;
IC2=cell(numSpaces-1,1);
fits2=cell(numSpaces-1,sampleSize);
numParam2=cell(numSpaces-1,sampleSize);

for k=3:1:39
    ICtemp=zeros(sampleSize,3);
    for n=1:sampleSize
        fprintf('k = %d, \t n = %d \n', k, n);
        dataCell=data{k,1}{n,1};
```

```
16          prefit = fits{k,n};
17          [ICtemp(n,:), numParam2{k,n}, fits2{k,n}] = IChist2(
                prefit, dataCell);
18      end
19      IC2{k,1} = ICtemp;
20  end
21  save('fitp_01');
```

The data contained in the data cells `IC` and `IC2` can be used to generate boxplots for the model selection which can be found in the paper by applying the MATLAB function `boxplot`.

C.3 Using the ACDm package

The `R` package `ACDm` by Markus Belfrage (`https://CRAN.R-project.org/package=ACDm`) provides functions for simulation and fitting of ACD models. The sample code in Listing C.4 shows wrapper functions for a repeated use in a Monte Carlo simulation.

Listing C.4: Sample code for wrapper functions to use `ACDm` library functions

```
simData <- function(n,m, p,q, omega, alpha, beta){
  data <- matrix(0,n,m)
  for (k in 1:n){
    cat(paste("..", k, ".."));
    data[k,] <- sim_ACD(N=m, param=c(omega, alpha, beta),
       order=c(p,q), Nburn=500);
  }
  return(data);
}

fitData <- function(p, q, omega, alpha, beta, data){
  out <- tryCatch(
      {
        temp <- acdFit(data, order=c(p,q), startPara=c(omega,
            alpha, beta));
        c(temp$mPara, temp$goodnessOfFit$value[1]);
      },
      error=function(e){
        rep(NaN, p+q+2);
      },
      warning=function(e){
```

```
20          rep(NaN, p+q+2);
21        },
22        finally={}
23      )
24    return(out)
25  }
```

Note that the library function `acdFit` has to be modified to actually quit with an error message for the `tryCatch` to be able to catch the exception. Listing C.5, which uses a combination of `cat` and `return` to give the error information and to end the function, should be replaced by `stop`, which throws an actual exception.

Listing C.5: Original exit sequence in the library function `acdFit`

```
1  ...
2  cat("\n\nError: Oops, seems like the the optimization
     function failed. Changing the 'optimFnc' or/and its
     settings, or starting from a diffrent 'startPara' might
     work. You can also trace the MLE search path by adding
     the argument 'control = list(trace = 1)'. \n\n")
3  return()
4  ...
```

```
1  ...
2  stop("\n\nError: Oops, seems like the the optimization
     function failed. Changing the 'optimFnc' or/and its
     settings, or starting from a diffrent 'startPara' might
     work. You can also trace the MLE search path by adding
     the argument 'control = list(trace = 1)'. \n\n")
3  ...
```

Listing C.6: Replacement code for `acdFit` to allow exception handling in higher level functions

C.4 Hawkes processes

`intensity`	**Evaluation of the intensity function of a Hawkes process**

Usage

```
function [ res ] = intensity( mu, alpha, beta, t, tp )
```

Description

Evaluates the intensity function of a Hawkes process

Inputs

`mu, alpha, beta`	parameters of the Hawkes process (alpha and beta can be vectors)
`t`	time point of evaluation of the intensity function
`tp`	event history prior to `t`

Value

`res`	value of the intensity function at t

`hawkesThinning` Simulation of a Hawkes process up to a specified time

Usage

```
function [ t ] = hawkesThinning( mu, ialpha, ibeta, T )
```

Description

Simulates a Hawkes process with up to a specified time horizon (contains starting value 0 at time 0)

Inputs

`mu, iapha, ibeta`	parameters of the Hawkes process
`T`	time horizon when to stop the simulation

Value

`t`	first entry is 0, followed by vector of simulated events

`hawkesThinning2` Simulation of a Hawkes process with specified sample size

Usage

```
function [ t ] = hawkesThinning2( mu, ialpha, ibeta, T )
```

Description

Simulates a Hawkes process with up to a specified sample size (contains starting value 0 at time 0)

Inputs

`mu, iapha, ibeta`	parameters of the Hawkes process
`T`	sample size (has to be at least 2)

Value

`t`	first entry is 0, followed by vector of simulated events

`empirAgg2` Calculation of average number of events

Usage

```
function [ res, grid ] = empirAgg2( mu, alpha, beta, M, delta, T )
```

Description

Calculates average number of events based on simulated paths of a Hawkes process (calls `hawkesThinning2.m`)

Inputs

`mu, alpha, beta`	parameters of the Hawkes process (alpha and beta can be vectors)
`M`	number of paths
`delta`	grid size for evaluation grid
`T`	end of observation value

Value

`grid`	time points at which avarage number of events are calculated
`res`	average number of events for each time entry in grid

LogLik_iter — Evaluation of the log-likelihood function of a Hawkes process

Usage

```
function [ L ] = LogLik_iter( p, t )
```

Description

evaluates the log-likelihood function of a Hawkes process for given parameters and data estimators using the MATLAB™ routine `fmincon` (calls `LogLik_iter.m` and `constraints.m`)

Inputs

`p`	vector of parameters of the Hawkes process [mu, alpha, beta]
`t`	vector of recorded events

Value

`L`	log-likelihood value

constrains — Parameter constrains for optimization

Usage

```
function [ c, ceq ] = constraints( p )
```

Description

Parameter constrains passed on to the optimization algorithm `fmincon`

Inputs

`p`	parameter vector [mu, alpha, beta]

Value

`c`	inequality constraint $c \leq 0$
`ceq`	ceq: equality constraint ceq ≤ 0

`fitting` Fitting of Hawkes process model to data

Usage

```
function [ x, value, exitflag ] = fitting( p0, data )
```

Description

Maximizes log-likelihood function to obtain maximum likelihood

Inputs

`p0`	initial values
`data`	given data set used for estimation

Value

`x`	(possible) maximizer of the log-likelihood function
`value`	value of the log-likelihood function evaluated at `x`
`exitflag`	indicator whether optimization routine was successful

`IC` Calculation of information criteria

Usage

```
function [ IC ] = IC( L, P, data )
```

Description

Calculates the values of AIC, BIC and HQ

Inputs

`L`	contains log-likelihood values
`P`	order of the model fitted
`data`	cell containing data sets corresponding to each entry of L

Value

`IC`	matrix containing the IC values; each row of the matrix has entries `[AIC, BIC, HQ]` for each entry of `L`

Appendix D

Source code

D.1 Poisson process

Listing D.1: compPoissonInterval

```
function [ res ] = compPoissonInterval( T, lambda, mu, sigma2 )
numJumps = poissrnd(lambda*T);
cum = [];
jumpTimes =[];
if numJumps > 0
    uniform = T*rand(numJumps,1);
    jumpTimes = sort(uniform);
    jumpSize = normrnd(mu, sqrt(sigma2), numJumps, 1);
    cum = cumsum(jumpSize);
end
res = [jumpTimes cum];
end
```

Listing D.2: nonhomPoisson

```
function [ res, trueLambda] = nonhomPoisson( grid, lambdaRange,
    mu, sigma2 )
n = length(grid);
cum = 0;
time = [];

lambdaVector=zeros((n-1),1);
for k=1:(n-1)
    startValue = cum(length(cum),1);
    startTime = grid(1,k);
    lambda = 4*(lambdaRange(1,2) - lambdaRange(1,1))*(grid(1,k)
```

```
        -0.5)^2 + lambdaRange(1,1);
    lambdaVector(k,1)=lambda;
    tProcess = compPoissonInterval(grid(1,k+1)-grid(1,k), lambda,
        mu, sigma2);
    if(length(tProcess) > 0)
        tProcess(:,1) = tProcess(:,1) + startTime;
        tProcess(:,2) = tProcess(:,2) + startValue;
        time = [time; tProcess(:,1)];
        cum = [cum; tProcess(:,2)];
    end
end
res = [time, cum(2:length(cum))];
trueLambda = [grid(2:end)', lambdaVector];
end
```

Listing D.3: ICSimInMem

```
function [ data, trueLambda ] = ICSimInMem( N, gridSim,
    lambdaRange)
k=1;
data = cell(N, 1);
trueLambda = cell(N, 1);
while k <= N
    fprintf('simk = %d\n',k);
    [data{k,1} , trueLambda{k,1}] = nonhomPoisson(gridSim,
        lambdaRange, 0, 1);
    if(~isempty(data))
        k=k+1;
    end
end
end
```

D.2 Dλ-model

Listing D.4: fit

```
function [ res ] = fit( grid, input )
jumpTimes = input(:,1);
m=length(input(:,2));
marks = input(:, 2) - [0 ; input(1:(m-1), 2)];
data = [jumpTimes, marks];
```

```
n = length(grid);
par = zeros((n-1), 3);
L = zeros((n-1), 1);
Lpart = zeros((n-1), 1);

for k=1:(n-1)
    tmarks = data(data(:,1) >= grid(1,k) & data(:,1) < grid(1,k
        +1),2);
    tlength = grid(1, k+1) - grid(1, k);
    if length(tmarks) > 1
        par(k,:) = fitInterval(tmarks, tlength);
        temp=likelihoodNP(par(k,1), par(k,2),par(k,3),tmarks',
            tlength);
        L(k,:)=temp(1);
        Lpart(k,:) = temp(2);
    end
end
res = [par L Lpart];
end
```

Listing D.5: fitInterval

```
function [ res ] = fitInterval( marks, interval_length )
lambda = length(marks)/interval_length;
mu = mean(marks);
sigma2 = sum((marks - mu).^2)/length(marks);
res = [lambda mu sigma2];
end
```

Listing D.6: gridAIC

```
function [ IC, numParam, fits ] = gridAIC( startTime, endTime,
    deltaMax, data )
g = grids(startTime, endTime, deltaMax);
n=length(g);
numParam = zeros(1,n);
fits = cell(1,n);
IC = cell(1,4);
AIC = zeros(1,n);
BIC = zeros(1,n);
HQ = zeros(1,n);
AICc = zeros(1,n);

```

```
for k=1:n
    fits{1,k} = fit(g{1,k}, data);
    numParamTemp=3*(length(g{1,k}) - 1);
    numParam(1,k) = numParamTemp;
    sampleSize = length(data(:,1));
    logLik=sum(fits{1,k}(:,4));

    AIC(1,k) = -2*logLik + 2*numParamTemp;
    BIC(1,k) = -2*logLik + numParamTemp*log(sampleSize);
    HQ(1,k) = -2*logLik + 2*numParamTemp*log(log(sampleSize));
    AICc(1,k) = -2*logLik + 2*numParamTemp...
        *sampleSize/(sampleSize - numParamTemp - 1);
    fits{1,k} = [g{1,k}(2:end)' fits{1,k}];
end
IC{1,1} = AIC;
IC{1,2} = BIC;
IC{1,3} = HQ;
IC{1,4} = AICc;
end
```

Listing D.7: grids

```
function [ res ] = grids( startTime, endTime, deltaMax )
n = max([floor((endTime-startTime)/deltaMax), 1]);
gridList = cell(1, n);
for k=1:n
   gridList{1,k} = linspace(startTime, endTime, 2+(k-1));
end
res = gridList;
end
```

Listing D.8: IChist

```
function [ IC, numParam, fits ] = IChist( data )
IC = zeros(1, 3);
    m = length(data(:,1));
    d = data(:,1) - [0 ; data(1:(m-1), 1)];
    deltaMax = max(d);
    [infoCrit, numParam, estimates] = gridAIC(0,1,deltaMax, data
        );
    fits=estimates;
    for q=1:3
        [M, I] = min(infoCrit{1,q});
```

```
10          IC(1,q) = numParam(1,I);
11      end
12  end
```

Listing D.9: likelihoodNP

```
1  #include <math.h>
2  #include "mex.h"
3  #define M_PI 3.14159265358979323846
4
5  void mexFunction( int nlhs, mxArray *plhs[], int nrhs, const
       mxArray *prhs[] )
6  {
7    double *prOut, *A, *lambda, *mu, *sigma2, *marks, *
       interval_length;
8    double res1, res2;
9    int m, q, n;
10   lambda=mxGetPr(prhs[0]);
11   mu = mxGetPr(prhs[1]);
12   sigma2 = mxGetPr(prhs[2]);
13   n = mxGetN(prhs[3]);
14   marks = mxGetPr(prhs[3]);
15   interval_length = mxGetPr(prhs[4]);
16
17   res1 = -lambda[0]*interval_length[0] + log(lambda[0])*n -
       0.5*log(2.0*M_PI*sigma2[0])*n;
18   res2 = 0.0;
19   for(int k=0; k < n; k++){
20     res2 = res2 - (marks[k] - mu[0])*(marks[k] - mu[0])/(2.0*
         sigma2[0]);
21   }
22
23   plhs[0] = mxCreateDoubleMatrix(2, 1, mxREAL);
24   prOut = mxGetPr(plhs[0]);
25   prOut[0] = res1 + res2;
26   prOut[1] = res2 - 0.5*log(2.0*M_PI*sigma2[0])*n;
27
28 }
```

D.3 Pλ-model

Listing D.10: constraints

```
function [ c, ceq ] = constraints( p )
c=-(p(1)*p(2)^2)/(4*p(1)^2) + (p(2)^2)/(2*p(1)) - p(3);
ceq=0;

end
```

Listing D.11: fit2

```
function [ res ] = fit2( prefit, input )
jumpTimes = input(:,1);
grid=[0;prefit(:,1)];
n=length(grid);
numJumps=zeros(1, n-1);
for k=1:(n-1)
    numJumps(k) = length(jumpTimes(jumpTimes >= grid(k) &
        jumpTimes < grid(k+1)));
end
lambda = prefit(:,2);
if length(lambda) < 3 %excluded
 error('Number of data points for lambda less than 4!');
else
    %first least squares fit of parabola to lambda values
    lambda0 = polyfit(grid(1:end-1)', lambda', 2);
    if lambda0(1) < 0 %parabola opens downwards
        lambda0(1) = -lambda0(1);
    end
    test=(lambda0(1)*lambda0(2)^2)/(4*lambda0(1)^2) - (lambda0
        (2)^2)/(2*lambda0(1)) + lambda0(3);
    if test <= 0 %min of parabola below zero
        lambda0(3) = lambda0(3) - test +1;
    end

    %first MLE (unconstraint) using above LS estmate as initial
        value
    [coef, LLambda, exitflag]=fitLambda(lambda0, grid, numJumps,
        2);

    %check if lambda is positive fct. -> if not do constraint
```

```
        optimization
27      test2=(coef(1)*coef(2)^2)/(4*coef(1)^2) - (coef(2)^2)/(2*
            coef(1)) + coef(3);
28      if coef(1)< 0 || test2 <= 0 || exitflag==0
29          [coef, LLambda, exitflag]=fitLambda(lambda0, grid,
                numJumps, 3);
30          fprintf('C ');
31      else
32          fprintf('U ');
33      end
34  end
35
36  lambdaPartLogLik = partialLogLik2(coef(1), coef(2), coef(3),
        grid, numJumps);
37  lambdaDiscrete = polyval(coef, prefit(:,1));
38  res = [prefit(:,1) , zeros(n-1,1)+coef(1), zeros(n-1,1)+coef(2),
        zeros(n-1,1)+coef(3),...
39      lambdaDiscrete , prefit(:,3:4) , lambdaPartLogLik , prefit
            (:,6)];
40  end
```

Listing D.12: fitLambda

```
1   function [ x, value, exitflag ] = fitLambda( p0, grid, numJumps,
        mode )
2   if mode ==3
3       fun = @(p) partialLogLikWrapper(p(1), p(2), p(3), grid,
            numJumps);
4       options=optimoptions(@fmincon, 'Display', 'notify');
5       [x, value, exitflag] = fmincon(fun,p0,[],[],[],[],[0, -Inf,
            -Inf],[],@constraints, options);
6   else
7       fun = @(p) partialLogLikWrapper(p(1), p(2), p(3), grid,
            numJumps);
8       options=optimoptions(@fminunc, 'Display', 'notify', 'GradObj
            ','on'); %
9       [x, value, exitflag] = fminunc(fun, p0, options);
10  end
11  end
```

Listing D.13: gridAIC2

```
1   function [ IC, numParam, fits ] = gridAIC2( prefit, data )
```

```
n=length(prefit);
numParam = zeros(1,n);
fits = cell(1,n);
IC = cell(1,4);
AIC = zeros(1,n);
BIC = zeros(1,n);
HQ = zeros(1,n);
AICc = zeros(1,n);

for k=3:n
    fits{1,k} = fit2(prefit{1,k}, data);

    N=length(fits{1,k}(:,1));
    numParamTemp=3+2*N;
    numParam(1,k) = numParamTemp;
    sampleSize = length(data(:,1));
    logLik=sum(fits{1,k}(:,8)) + sum(fits{1,k}(:,9));

    AIC(1,k) = -2*logLik + 2*numParamTemp;
    BIC(1,k) = -2*logLik + numParamTemp*log(sampleSize);
    HQ(1,k) = -2*logLik + 2*numParamTemp*log(log(sampleSize));
    AICc(1,k) = -2*logLik + 2*numParamTemp...
        *sampleSize/(sampleSize - numParamTemp - 1);
end
IC{1,1} = AIC;
IC{1,2} = BIC;
IC{1,3} = HQ;
IC{1,4} = AICc;
end
```

Listing D.14: IChist2

```
function [ IC, numParam, fits] = IChist2( prefit, data )
IC = zeros(1, 3);

[infoCrit, numParam, estimates] = gridAIC2(prefit, data);
fprintf('\n');
fits=estimates;
for q=1:3
    [M, I] = min(infoCrit{1,q});
    IC(1,q) = numParam(1,I);
end
```

```
11 end
```

Listing D.15: partialLogLikNP

```
1  #include <math.h>
2  #include "mex.h"
3  #define M_PI 3.14159265358979323846
4
5  void mexFunction( int nlhs, mxArray *plhs[], int nrhs, const
       mxArray *prhs[] )
6  {
7    double *prOut1,*prOut2, *a, *b, *c, *grid, *numJumps;
8    double f, g1, g2, g3, lambda, t1, t2;
9    int m, q, n;
10   a=mxGetPr(prhs[0]);
11   b = mxGetPr(prhs[1]);
12   c = mxGetPr(prhs[2]);
13   n = mxGetN(prhs[3]);
14   grid = mxGetPr(prhs[3]);
15   numJumps = mxGetPr(prhs[4]);
16
17   f=0.0;
18   g1=0.0;
19   g2=0.0;
20   g3=0.0;
21   for(int k=0; k < (n-1); k++){
22     t2=grid[k+1];
23     t1=grid[k];
24     lambda = (a[0]*t2*t2 + b[0]*t2 + c[0]);
25     f = f + log(lambda)*numJumps[k] - (lambda)*(t2-t1);
26     g1 = g1 + t2*t2/(lambda)*numJumps[k] - t2*t2*(t2-t1);
27     g2 = g2 + t2/(lambda)*numJumps[k] - t2*(t2-t1);
28     g3 = g3 + 1.0/(lambda)*numJumps[k];
29   }
30
31   g3=g3 - (grid[n-1]-grid[0]);
32
33   plhs[0] = mxCreateDoubleMatrix(4, 1, mxREAL);
34   prOut1 = mxGetPr(plhs[0]);
35   prOut1[0] = (-1.0)*f;
36   prOut1[1] = (-1.0)*g1;
37   prOut1[2] = (-1.0)*g2;
```

```
38    prOut1[3] = (-1.0)*g3;
39 }
```

D.4 Hawkes processes

Listing D.16: constraints

```
1  function [ c, ceq ] = constraints( p )
2  P = (length(p)-1)/2;
3
4  alpha = p(2:(2+P-1));
5  beta = p((2+P):end);
6  if(P==1)
7      c=sum(alpha./beta)-1;
8  else
9      %stationarity condition and increasing order of betas
10     c=[sum(alpha./beta)-1, - beta(2:end) + beta(1:(end-1))];
11 end
12 ceq=0;
13 end
```

Listing D.17: empirAgg2

```
1  function [ res, grid ] = empirAgg2( mu, alpha, beta, M, delta, T
       )
2  data=zeros(M, T);
3
4  for k=1:M
5      fprintf('%s%d\n', 'k = ', k);
6      datatemp = hawkesThinning2(mu, alpha, beta, T);
7      data(k,:) = datatemp(2:end);
8  end
9  grid=0:delta:min(max(data));
10 n = length(grid);
11 N = zeros(1, n);
12 for k=1:M
13     for m=1:n
14         N(m) = N(m) + sum(data(k,:) <= grid(m));
15     end
16 end
17 res = N/M;
```

```
end
```

Listing D.18: fitting

```
function [ x, value, exitflag ] = fitting( p0, data )
fun = @(p) -LogLik_iter(p, data);
options=optimoptions(@fmincon, 'Display', 'notify');
[x, value, exitflag] = fmincon(fun,p0,[],[],[],[],zeros(1,
    length(p0)),[],@constraints, options);
end
```

Listing D.19: hawkesThinning

```
function [ t ] = hawkesThinning( mu, ialpha, ibeta, T )
alpha = sum(ialpha);
lambdaStar = mu;
t = 0;
jump = 0;
u = exprnd(1/lambdaStar, 1, 1);

if u <= T
    t = [t, u];
    s = u;
    jump=1;
else
    return
end

tp = [];
while 1
    if jump == 1
        lambdaStar = intensity(mu, ialpha, ibeta, s, tp) + alpha
            ;
    else
        lambdaStar = intensity(mu, ialpha, ibeta, s, t(2:end));
    end
    u = exprnd(1/lambdaStar, 1, 1);
    s = s + u;
    if s > T
        break;
    else
        U = rand;
        if U <= intensity(mu, ialpha, ibeta, s, t(2:end))/
```

```
            lambdaStar
30              tp = t(2:end);
31              t = [t, s];
32              jump = 1;
33          else
34              jump = 0;
35          end
36      end
37  end
38  end
```

Listing D.20: hawkesThinning2

```
1   function [ t ] = hawkesThinning2( mu, ialpha, ibeta, T )
2
3   alpha = sum(ialpha);
4   lambdaStar = mu;
5   t = zeros(1, T+1);
6   u = exprnd(1/lambdaStar, 1, 1);
7
8   t(2) = u;
9   s = u;
10  jump=1;
11  index=2;
12
13  tp = [];
14  while 1
15      if jump == 1
16          lambdaStar = intensity(mu, ialpha, ibeta, s, tp) + alpha
                ;
17      else
18          lambdaStar = intensity(mu, ialpha, ibeta, s, t(2:end));
19      end
20      u = exprnd(1/lambdaStar, 1, 1);
21      s = s + u;
22      if index > T
23          break;
24      else
25          U = rand;
26          if U <= intensity(mu, ialpha, ibeta, s, t(2:end))/
                lambdaStar
27              tp = t(2:end);
```

```
28              t(index+1) = s;
29              jump = 1;
30              index=index+1;
31          else
32              jump = 0;
33          end
34      end
35  end
36  end
```

Listing D.21: IC

```
1  function [ IC ] = IC( L, P, data )
2  numParam = 1+2*P;
3  n = length(L);
4  IC = zeros(n,3); %AIC, BIC, HQ
5  for k=1:n
6      N = length(data{k});
7
8      IC(k, 1) = -2*L(k) + 2*numParam;      %AIC
9
10 %replace above line by this bloc to include AICc
11 %      if N/7 > 40   %rule of thumb: k_max=7
12 %          IC(k, 1) = -2*L(k) + 2*numParam;      %AIC
13 %      else
14 %          IC(k, 1) = -2*L(k) + (2*numParam*N)/(N-numParam-1); %
       AICc
15 %      end
16
17     IC(k, 2) = -2*L(k) + numParam*log(N); %BIC
18     IC(k, 3) = -2*L(k) + 2*numParam*log(log(N)); %HQ
19 end
20 end
```

Listing D.22: intensity

```
1  function [ res ] = intensity( mu, alpha, beta, t, tp )
2  res = mu;
3  order = length(alpha);
4  if ~isempty(tp)
5      for i=1:order
6          res = res + alpha(i)*sum(exp(-beta(i).*(t-tp)));
7      end
```

```
end
end
```

Listing D.23: LogLik_iter

```
function [ L ] = LogLik_iter( p, t )
P = (length(p)-1)/2;
mu = p(1);
alpha = p(2:(2+P-1));
beta = p((2+P):end);
n = length(t);
T = t(end);
L = -mu*T;
A = zeros(P,n);
for m=1:P
    for k=2:n
        A(m,k) = (1 + A(m,k-1))*exp(-beta(m)*(t(k) - t(k-1)));
    end
end

for m=1:P
    temp2 = 0;
    for k=1:n
        temp2 = temp2 + (1 - exp(-beta(m)*(T - t(k))));
    end
    L = L - alpha(m)/beta(m)*temp2;
end

for k=1:n
    temp1=0;
    for m=1:P
        temp1 = temp1 + alpha(m)*A(m,k);
    end
    L = L + log(mu + temp1);
end
end
```

List of Figures

Name Index

Subject Index