Dissertation
an der
Fakultät für Mathematik, Informatik und Statistik
der
Ludwig–Maximilians–Universität München

Nonparametric Estimation of the Jump Component in Financial Time Series

vorgelegt von

Serkan Yener

München, den 06. Juni 2012

Dissertation
an der
Fakultät für Mathematik, Informatik und Statistik
der
Ludwig–Maximilians–Universität München

Nonparametric Estimation of the Jump Component in Financial Time Series

vorgelegt von

Serkan Yener

Abgabe: 06. Juni 2012
Disputation: 31. Juli 2012

Erstgutachter: Prof. Stefan Mittnik, PhD
Zweitgutachter: Prof. Dr. Svetlozar Rachev

Nonparametric Estimation of the Jump Component in Financial Time Series

Serkan Yener

Abstract

In this thesis, we analyze nonparametric estimation of Lévy-based models using wavelets methods. As the considered class is restricted to pure-jump Lévy processes, it is sufficient to estimate their Lévy densities. For implementing a wavelet density estimator, it is necessary to setup a preliminary histogram estimator. Simulation studies show that there is an improvement of the wavelet estimator by invoking an optimally selected histogram. The wavelet estimator is based on block-thresholding of empirical coefficients. We conclude with two empirical applications which show that there is a very high arrival rate of small jumps in financial data sets.

Zusammenfassung

Diese Arbeit untersucht nichparametrische Verfahren zur Schätzung von Modellen, welche auf Lévy-Prozessen basieren. Ausgehend von einer ökonomischen und statistischen Argumentation wird dabei die allgemeine Klasse der Lévy-Prozesse auf reine Lévy-Sprungprozesse beschränkt, welche eindeutig durch die entsprechende Lévy-Dichte charakterisiert sind. Zur nichtparametrischen Schätzung dieser Dichten wird ein zweistufiges Verfahren vorgeschlagen: In der ersten Stufe wird, basierenden auf statistischen Optimalitätsbetrachtungen, ein Schätzer für ein Histogramm entwickelt. Dieses wird für einen Wavelet-Schätzer der zweiten Stufe benötigt, welcher auf blockweisem "Thresholding" beruht. Simulationsstudien für zwei Lévy-basierte Modelle zeigen, dass der optimale gewählte Schätzer der ersten Stufe zu einer Verbesserung des Wavelet-Schätzers führt. In zwei empirischen Anwendungen deutet der Wavelet-Schätzer auf eine hohe Aktivität kleiner Sprünge hin.

Contents

Notations

$a := b$	a is defined as b
$a \approx b$	a is approximately equal to b
$a_n \asymp b_n$	For positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, there exists some constant C such that $C^{-1} \leqslant a_n/b_n \leqslant C$.
$a_n = O(b_n)$	a_n/b_n is bounded by some constant as $n \to \infty$
$a_n = O_P(b_n)$	a_n/b_n is bounded by some constant in probability as $n \to \infty$
$a_n = o(b_n)$	$a_n/b_n \to 0$ as $n \to \infty$
$a_n = o_P(b_n)$	$a_n/b_n \to 0$ in probability as $n \to \infty$
$a \wedge b$	$\mathsf{min}(a, b)$
$a \vee b$	$\mathsf{max}(a, b)$
$\overline{\mathbb{R}}$	extended real line, i.e., $\mathbb{R} \cup \{-\infty, +\infty\}$
$\overline{A}$	closure of set A
$\mathbb{1}_A$	indicator function of set A
$f^{(m)}$	mth derivative of function f
$\mathcal{W}_p^m$	Sobolev space with smoothness parameter m and integration parameter p
$X \overset{d}{=} Y$	X and Y are identically distributed
$\mathsf{sgn}(x)$	$\begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ +1 & \text{if } x > 0 \end{cases}$
$\nu \ll \mu$	measure ν is absolutely continuous with respect to measure μ
$X \rightsquigarrow Y$	X converges weakly to Y

List of Figures

List of Tables

Introduction

Why Lévy Processes?

Continuous-time models based on Brownian motions as background stochastic driving processes have a long tradition in mathematical finance. Indeed, their roots can be traced back to the doctoral thesis of Bachelier (1900) on the rational pricing of financial options. From a theoretical perspective, this long-lasting success is mostly due to an important result of Itō (1951) in the realm of stochastic calculus which forms the backbone of an elegant and powerful theory of risk-neural arbitrage pricing for options in continuous time. For an overview of this theory, see the monographs of Shiryaev (1999, Chapter VII) or Shreve (2004, Chapters 4 & 5).

Initially, Bachelier (1900) proposed a Brownian motion with drift to model the dynamics of an underlying stock price. Unfortunately, it turned out not to be a reasonable model for stock prices since it does not warrant non-negative prices. This shortcoming was eliminated by the model of Samuelson (1965) which is now known as the *geometric Brownian motion*. To be more precise, a stock price S_t is a geometric Brownian motion if its dynamics obeys the stochastic differential equation

$$\frac{\mathsf{d}S_t}{S_t} = \gamma\,\mathsf{d}t + \sigma\,\mathsf{d}W_t\ ,$$

for all $t \in [0,\infty)$, satisfying the solution

$$S_t = S_0\,\mathsf{exp}\left\{\left(\gamma - \frac{\sigma^2}{2}\right)t + \sigma W_t\right\} = S_0\,\mathsf{exp}(B_t)\ ,$$

where B_t is a Brownian motion with drift with *Wiener process* $W_t \overset{d}{=} \mathsf{N}(\,0, t\,)$ and constant drift $\gamma \in \mathbb{R}$ and diffusion coefficient $\sigma \geqslant 0$.

This model was used in the seminal works of Black and Scholes (1973) and Merton (1973) as the driving process of the underlying for deriving the price of an option based upon the principle of absence of arbitrage opportunities, and has several important im-

plications: Firstly, the Brownian motion (with drift) is obviously a continuous-time generalization of a random walk (with drift). Thus, a Brownian motion (with drift) is a Markov process which, in turn, means that the process satisfies the weak form of the *efficient market hypothesis* (Fama, 1970). Roughly speaking, this hypothesis postulates that there exists no trading strategy based upon stock market time series which provides an 'abnormal' profit in the long run. Secondly, the log-returns calculated over an investment horizon $\Delta t > 0$ are uncorrelated and normally distributed, i.e.,

$$r_{t+\Delta t} := \log \frac{S_{t+\Delta t}}{S_t} \stackrel{d}{=} \mathsf{N}\left(\left(\gamma - \frac{\sigma^2}{2} \right) \Delta t, \sigma^2 \Delta t \right) .$$

The persistent success and popularity of the Black-Merton-Scholes model, especially among practitioners, can be explained by the simple structure of the geometric Brownian motion. Note that, on the one hand, the normal distribution and the linearity of the process are easy to understand. On the other hand, it provides option pricing with a great deal of analytical tractability such that closed-form expressions for pricing formulae can often be derived. Moreover, continuous sample paths render security markets complete such that perfect hedging arguments can be applied.

As often in science, a beautiful theory loses much of its appeal when confronted with reality. The use of the geometric Brownian motion in modeling stock prices and in option pricing leads to theoretical predictions which are at odds with what is observed in empirical data. Astonishingly, these contradictions constitute phenomena which are consistently, observed across different financial markets, asset classes, and historical episodes. Consequently, they are often dubbed "the stylized facts of financial markets."

We now list the most important stylized facts and refer to, for example, Cont (2001), Cont and Tankov (2003, Chapter 7), Schoutens (2003, Chapter 4), or Shiryaev (1999, Chapter IV) for more detailed discussions.

(1) *Leptokurtic returns distribution:* The empirical distribution of asset returns calculated over short investment horizons Δt, like intra-daily or daily, is more peaked around the origin and has more probability mass in the tails than a fitted normal distribution. However, there seems to be aggregational Gaussianity as the marginal distribution of empirical returns calculated over longer investment horizons Δt, like quarterly or annual, tends to a normal distribution (Akgiray and Booth, 1988).

(2) *Jumps in asset prices:* The sample paths of asset prices exhibit substantial discontinuities, even for heavily traded, i.e., liquid, assets.

(3) *Volatility clustering:* Empirical asset returns exhibit distinct periods of low and high

volatility. Thus, financial markets pass through sustained phases of tranquility and turbulence.

(4) *Smiles & smirks in implied volatility:* Contrary to what the Black-Merton-Scholes model predicts, the implied volatility computed from observed option prices is not constant (neither across strike prices nor across maturities). Note that this is partially attributed to stylized facts (1)–(3) which contradict the assumption of a geometric Brownian motion (Hull, 2000, Chapter 17).

(5) *Long memory in volatility:* Although empirical asset returns do not exhibit significant autocorrelation, the autocorrelation function of absolute returns decays slowly.

(6) *Leverage effect:* The future volatility of empirical stock returns is negatively correlated with past returns.

The failure of geometric Brownian motions with respect to (1) is evidently due to the normal distribution's inability to reproduce leptokurtosis. An obvious solution consists in replacing the normal distribution by some leptokurtic distribution. Following this line of reasoning, Mandelbrot (1963) suggested the stable distribution to improve the empirical fit compared to applying the normal distribution. For a general treatment of stable (Paretian) distributions with many financial applications, see the monograph of Rachev and Mittnik (2000).

However, they are 'stable in law' under time aggregation such that aggregational Gaussianity of returns is ruled out. Moreover, stable distributions fail to have finite second moments, since they are too heavy-tailed, which is argued to be an undesirable property by some practitioners. In order to tackle the later problem, the so-called *exponential Lévy process*

$$S_t = S_0 \exp(X_t)$$

was put forward, where the driving Brownian motion with drift is simply substituted by a Lévy process X_t. Recently, a very flexible (and yet mathematically tractable) sub-class of Lévy processes, known as generalized hyperbolic distributions, have been successfully fitted to stock market returns (Eberlein, 2001).

Notice that heavy tails of the marginal law corresponds to the rare occurrences of large returns, i.e., sudden changes in the price process, but which are much more frequent than under normality. The simplest way of generating this type of non-normality is to augment the continuous sample paths of a (geometric) Brownian motion by jumps of random sizes occurring at random times, as proposed by Merton (1976). Thus, stylized facts (1) and (2) may indeed be related issues.

Unfortunately, as Cont and Tankov (2003, pp. 319–351) discuss, option pricing in the presence of jumps becomes much less tractable as security markets become incomplete. This is a general problem in non-Gaussian option pricing and its resolution depends on advances in *semimartingale theory* which provides ways to ways with jump processes appropriately. For example, although there does not exist a unique pricing formula for an option with discontinuous sample paths of the underlying, an optimal pricing formula has been derived for hyperbolic processes.

On the empirical side, recent nonparametric studies seem to substantiate the relevance of including jumps in financial models. For example, Aït-Sahalia and Jacod (2009b) found evidence for the presence of jumps in stock prices, while Lee and Mykland (2006) showed that jumps play an important role in the S&P 500 index. See also Lee and Hannig (2010). Mancini and Renò (2011) found evidence for jumps in interest rate time series using a kernel-based nonparametric estimation method.

Based on nonparametric analysis of high-frequency data Barndorff-Nielsen and Shephard (2007) and Todorov and Tauchen (2011) provided evidence that jumps are present in both prices and volatility. A corresponding option pricing model in a double-jump setup was proposed by Duffie, Pan, and Singleton (2000) which is particularly appealing from a practitioner's point of view as its affine structure allows for closed-form solutions.

Up to now, jumps and Lévy processes were motivated to explain stylized facts (1)–(3) about the marginal distribution of asset returns. However, jump Lévy processes can also be used to model complicated dynamics of asset prices. In particular, Barndorff-Nielsen and Shephard (2001) introduced a model driven by Lévy processes and which potentially explains all of the above stylized facts. For more details on this model, see Section 4.2.

In sum, because of the potential impact of jumps on financial models and their empirical applications, it is of paramount importance for risk managers, traders, portfolio managers, and policy makers alike to obtain a thorough understanding of their true nature.

Why Nonparametric Estimation?

The price we have to pay for this gain in modeling flexibility of Lévy processes, compared to models based on Brownian motions, is the increased computational flexibility. To be more precise, Lévy-based models usually do not admit for an explicit closed-form solution of their returns densities which renders the corresponding likelihood functions intractable. Thus, direct application of maximum likelihood becomes infeasible. For implementing likelihood methods, we must resort to simulation techniques or Fourier inversions of the corresponding characteristic function in order to obtain a numerical approximation of the

returns density. Either solution may turn out to be extremely computationally intensive. In particular, as pointed out by Lo (1988), likelihood methods based on Fourier inversions require the inversion to be computed for every evaluation step of the likelihood function. This may become computationally expensive for numerical maximization of the likelihood function. Even if computable, numerical maximum likelihood may witness convergence problems and instability with respective to local maxima, in applications.

In contrast to parametric maximum likelihood, nonparametric estimators are straightforwardly and fast to compute for the Lévy processes we consider which, roughly speaking, are generalizations of inhomogeneous Poisson processes. Another, and maybe *the* major, advantage of nonparametric methods is that we do not have to settle for one particular model *a priori*. In the last two decades, research on Lévy processes has been buoyant leading to a tremendous surge of models for financial applications. This made it even harder to opt for a particular parametric model. Hence, there is always the danger to pick a model that is either too simple or too complex. Meaning that the estimation is misspecified or inefficient, respectively. A reasonable way out of this dilemma is to "let the data speak for themselves," which is exactly where nonparametric methods come into play. Moreover, nonparametric methods are able to detect features of the data which may remain undetected when applying parametric models, even when one applies the model with the highest degree of flexibility. This is especially relevant for financial models as they are often geared with a view towards applicability to mathematical finance.

Ultimately, due its explorative character, a suitable nonparametric estimator may lend itself to building the basis of goodness-of-fit testing for selecting the best parametric model. Nonparametric estimators are sometimes criticized for having convergence rates slower than the parametric one of $n^{-1/2}$ (or n^{-1} if measures in terms of the $\mathcal{L}_2$-risk) and, thus, have a unsatisfactory performance in small samples. Fortunately, this shortcoming of nonparametric methods is less relevant when using high-frequency data samples in financial markets.

Papers, which dealt with the (parametric or nonparametric) estimation of Lévy processes, are Akritas (1982), Akritas and Johnson (1981), Ball and Torous (1983), Basawa and Brockwell (1982), Gugushvili (2009), Masuda (2009), Neumann and Reiß (2009), Rubin and Tucker (1959), Shimizu (2006a,b, 2009a,b), and Shimizu and Yoshida (2006), among others.

What's New?

This thesis focusses on the nonparametric estimation of pure-jump Lévy processes via orthogonal projections based on discretely sampled observations.

The estimator we propose is a block-thresholded wavelet estimator put forward by Hall, Kerkyacharian, and Picard (1998), Chicken and Cai (2005), and Cai (1999) who derived is optimality properties such as adaptation in the minimax sense and oracle inequalities. We transfer their approach to the problem of estimating a Lévy density nonparametrically and discuss some of their optimality results in this context.

Very recently, Song (2010) considered the nonparametric estimation of a Lévy density using wavelet bases. However, this wavelet estimator is linear in contrast to our nonlinear wavelet estimator which allows to adapt to more general forms of the unknown Lévy densities.

Since we will be dealing with discretely sampled data, it is necessary to show consistency of an estimator which is intended for a discrete-time model. This was already accomplished by Figueroa-López (2009) and Figueroa-López and Houdré (2006) for nonparametric estimation of Lévy densities via piecewise polynomials. We discuss their result and point out another possibility for establishing weak convergence.

Note that the analytical results for wavelet estimators are derived in the context of the (Gaussian) nonparametric regression model which hampers its applicability to estimating frequency curves, such as densities or intensities. As a common practice, a preliminary estimator, i.e., a histogram, is computed on which the wavelet estimator is implemented. Unfortunately, there is no theoretically founded recipe for computing this preliminary estimator in an optimal way. Usually, the number of bins is selected arbitrarily. To this end, we adopt the approach of Birgé and Rozenholc (2006) to the problem of constructing a histogram estimator for a Lévy density. The resulting estimator satisfies a nonasymptotic optimality property such that it is expected to perform well in small samples.

Finally, we use Monte Carlo simulations in order to evaluate how our proposed approach works in practice. The model, that we consider in simulations, are the variance gamma process of Madan and Seneta (1990) and the Lévy-driven Ornstein-Uhlenbeck process of Barndorff-Nielsen and Shephard (2001). Afterwards, we apply our approach to the nonparametric estimation of the Lévy densities of S&P 500 returns and of the subordinator driving the daily realized volatility of the S&P 500. To the best of our knowledge, nonparametric Lévy density estimation via wavelet methods has never been applied to financial data before.

The structure of the thesis is as follows: In Chapter 1, we introduce Lévy processes

and explain some fundamental properties. We also motivate why it might be sufficient to restrict the analysis to pure-jump Lévy processes. In Chapter 2, we introduce a general way of nonparametric estimation which will be the backbone of our preliminary histogram estimator. We also discuss optimality criteria which will be used to gauge the quality of our estimators. In Chapter 3, nonparametric estimation via wavelet block-thresholding for Lévy densities is introduced. Chapter 4 contains the implementations of our approach.

Chapter 1

Lévy Processes

1.1 Definition, Examples & Basic Properties

This section introduces the very definition of a Lévy process, which explains why they are often dubbed as 'processes with stationary and independent increments,' along with some basic results providing deeper insight. All of these can be found in monographs like Applebaum (2004), Bertoin (1996), and Sato (1999).

Definition 1.1 (Lévy Process) *The $\mathbb{R}$-valued stochastic process $X = \{X_t\}_{t\geqslant 0}$ defined on $\big(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geqslant 0}, P\big)$ is a Lévy process, if it satisfies the following conditions:*

C1 $X_0 = 0$ *(P-a.s.).*

C2 Independent increments: For any $t, s \geqslant 0$, the increment $X_{t+s} - X_t$ is independent of $\mathscr{F}_t$.

C3 Stationary increments: For any $t, s \geqslant 0$, the law of the increment $X_{t+s} - X_t$ does not depend on t, i.e., $X_{t+s} - X_t \stackrel{d}{=} X_s$.

C4 Càdlàg sample paths: For P-almost all $\omega \in \Omega$, the sample paths of X belong to the function space $\mathcal{D}_0[0,\infty)$, i.e., the map $t \mapsto X_t(\omega)$ is right-continuous with left limits (P-a.s.).

C5 Stochastic continuity: X is continuous in probability, i.e., for any $t \geqslant 0$ and for any $\epsilon > 0$,

$$\lim_{s\to 0} P\Big(\big|X_{t+s}(\omega) - X_t(\omega)\big| > \epsilon\Big) = 0\ .$$

Remark 1.2 *Condition C1 is merely a technical normalization which simplifies derivations and proofs without loss of generality. Condition C2 is often stated in a more operational form: For any $n \in \mathbb{N}$ and for any associated collection $0 \leqslant t_0 < t_1 < \cdots < t_{n-1} < t_n < \infty$, the increments $X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}$ are (pairwise and mutually) independent. Condition C3 states that the increments of X are time-homogeneous in the sense that the distribution of $X_{t+s} - X_t$ is shift-invariant. Taken together, Conditions C2 and C3 imply the famous* iid*-increments property of Lévy processes. Recalling the definition of the space $\mathcal{D}$ of càdlàg functions, Condition C4 postulates that the sample paths of X can have at most a countable number of jumps (Protter, 2004, Theorem 30). Condition C5 is a bit tricky, as it appears to contradict Condition C4 at first sight. However, it does not rule out discontinuous sample paths. (In fact, the continuity of sample paths implies stochastic continuity, but not vice versa.) What stochastic continuity essentially asserts is that, for any $t \geqslant 0$, jumps of X are not deterministic, but stochastic, i.e.,*

$\Delta X_t := X_t - X_{t-} = 0$ *(P-a.s.), where* $X_{t-} := \lim_{s \nearrow t} X_s$ *exists due to Condition C4.*[1]

Although Definition 1.1 is quite general, it is no convenient devise for modeling purposes because it does not impose sufficient structure. This sections's theorems are of fundamental importance as they shed some light on ways of characterizing Lévy processes.

The starting point for the first result is the notion of **infinite divisibility** of a probability law. A random variable X (and its probability distribution) is infinitely divisible if, for any $n \in \mathbb{N}$, there exists an iid collection of n random variables $X_1, \ldots, X_n$ such that $X \stackrel{d}{=} X_1 + \cdots + X_n$. Examples for infinitely divisible laws are the Gaussian, Poisson, and α-stable ones. It can easily be shown that any Lévy process is infinitely divisible: For any $t \geqslant 0$, let us fix an arbitrary $n \in \mathbb{N}$. Next, fix the time interval $\Delta := t/n$ at which random variables from $\{X_s : 0 \leqslant s \leqslant t\}$ are sampled, i.e., $\{X_{i\Delta} : 0 \leqslant i \leqslant n\}$. Using the latter to define the increments $\{\Delta X_i := X_{i\Delta} - X_{(i-1)\Delta} : 1 \leqslant i \leqslant n\}$, we arrive at

$$\begin{aligned} X_t := X_{n\Delta} &= X_{n\Delta} - X_0 \\ &= X_{n\Delta} - X_{(n-1)\Delta} + X_{(n-1)\Delta} - X_{(n-2)\Delta} + \cdots + X_\Delta - X_0 \\ &= \Delta X_n + \Delta X_{n-1} + \cdots + \Delta X_1 \ . \end{aligned}$$

Since $\{X_{i\Delta}\}_{i=1}^n$ are sampled on an equally spaced grid, the corresponding increments $\{\Delta X_i\}_{i=1}^n$ are iid due to Conditions C2 and C3 of Definition 1.1.

From this result, it is obvious that the law of any X_t is given by the convolution of the laws of its increments. However, since dealing with convolutions can be a daunting task, using characteristic functions in this specific instance seems to be more promising. In particular, if $P_t = P \circ X_t^{-1}$ is the law of X_t, then the Fourier transform of P_t defines its characteristic function

$$\Phi_{X_t}(u) := \mathsf{E}\big[\, \mathsf{e}^{\mathsf{i}uX_t} \,\big] = \int_{\mathbb{R}} \mathsf{e}^{\mathsf{i}ux} P_t(\mathsf{d}x) \ ,$$

for all $u \in \mathbb{R}$. From the infinite divisibility of P_t, it follows that, for any $n \in \mathbb{N}$, there exists some probability distribution $P_{t,n}$ with characteristic function $\Phi_{t,n}$ such that

$$\Phi_{X_t}(u) = \int_{\mathbb{R}} \mathsf{e}^{\mathsf{i}ux} P_{t,n}^t(\mathsf{d}x) = \underbrace{\Phi_{X_t,n}(u)\Phi_{X_t,n}(u)\cdots\Phi_{X_t,n}(u)}_{n \text{ times}} = \big[\Phi_{X_t,n}(u)\big]^n \ , \qquad (1.1.1)$$

for all $u \in \mathbb{R}$.

The following result's merits are twofold. On the one hand, it shows that the class

[1] It should be noted that this assumption might be questionable in practice since it ignores jumps originating from announcements of payroll or interest rate policy news.

of infinitely divisible distributions and the class of Lévy processes are connected via a one-to-one correspondence. On the other hand, it shows that the characterization of any Lévy process can be reduced to three parameters. The latter fact is of paramount importance to the statistical inference of Lévy processes. Theorem 1.3 is a summary of Theorems 7.10 and 8.1 of Sato (1999).

Theorem 1.3 (Lévy-Khintchine Representation) *The stochastic process $X = \{X_t\}_{t \geqslant 0}$ is a Lévy process if and only if its characteristic function has the form $\Phi_{X_t}(u) = \mathsf{E}\big[\, \mathsf{e}^{\mathsf{i}uX_t} \,\big] = \mathsf{e}^{t\Psi(u)}$ with characteristic exponent*

$$\Psi(u) = \mathsf{i}u\gamma - \frac{u^2\sigma^2}{2} + \int \big(\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}ux\mathbb{1}_{\{|x|\leqslant 1\}}\big)\, \nu(\mathsf{d}x) \;,$$

for all $u \in \mathbb{R}$, where $\gamma \in \mathbb{R}$, $\sigma^2 \geqslant 0$, and $\nu : \mathbb{R}\backslash\{0\} \to \mathbb{R}_+$ is a Borel measure satisfying

$$\int_{\mathbb{R}\backslash\{0\}} (1 \wedge x^2)\nu(\mathsf{d}x) < \infty \;.$$

Remark 1.4 *One immediate consequence of Theorem 1.3 is that the so-called **characteristic triplet** (γ, σ, ν) uniquely determines the probability law of X. As we will see later on, the components of the characteristic triplet can be interpreted as follows: γ is the center or drift parameter of X, σ is the diffusion parameter of X, and ν is the Lévy measure determining the jump behavior of X. For more on ν, see also Definition 1.9.*

Remark 1.5 *The integrability condition of the Lévy measure in Theorem 1.3 is often cast in the form of*

$$\nu\big(\{0\}\big) = 0 \qquad \textit{and} \qquad \int_{\mathbb{R}} (1 \wedge x^2)\nu(\mathsf{d}x) < \infty \;.$$

This is the reason why many authors replace $\mathbb{R}\backslash\{0\}$ by $\mathbb{R}$, while implicitly keeping in mind that the Lévy measure ν vanishes at the origin. Heuristically, $\nu\big(\{0\}\big) = 0$ can be interpreted as an identifiability condition disentangling the continuous part of a Lévy process from its jumps.

Before presenting the second fundamental result on Lévy processes in Theorem 1.6, we take a closer look at the characteristic exponent $\Psi(u)$ of Theorem 1.3, which also serves the purpose of partially motivating Theorem 1.6. The form of the characteristic exponent in Theorem 1.3 suggests that any Lévy process can be decomposed into the sum of four

independent Lévy processes:

$$\Psi(u) = \underbrace{\mathrm{i}u\gamma}_{①} - \underbrace{\frac{u^2\sigma^2}{2}}_{②} + \underbrace{\int_{\{|x|>1\}} \left(\mathrm{e}^{\mathrm{i}ux} - 1\right) \nu(\mathrm{d}x)}_{③} + \underbrace{\int_{\{0<|x|\leqslant 1\}} \left(\mathrm{e}^{\mathrm{i}ux} - 1 - \mathrm{i}ux\right) \nu(\mathrm{d}x)}_{④} \quad (1.1.2)$$

The first part of the characteristic exponent in Equation 1.1.2 leads to the characteristic function

$$\Phi^{①}_{X_t}(u) := \mathsf{exp}(\mathrm{i}u\gamma t)$$

which is associated to the random variable

$$X_t^{①} = \gamma t \ .$$

Obviously, this is not a genuine stochastic process but a **linear drift function** and is, as it turns out, the only deterministic Lévy process.

The second part of the characteristic exponent in Equation 1.1.2 leads to the characteristic function

$$\Phi^{②}_{X_t}(u) := \mathsf{exp}\left(-\frac{u^2\sigma^2}{2}\, t\right)$$

which is associated to the random variable

$$X_t^{②} = \sigma W_t \ ,$$

where $W_t \overset{d}{=} \mathsf{N}(\,0, t\,)$ is the well-known **Wiener process**, and $X_t^{②}$ is the **Brownian motion** with diffusion parameter σ. This result follows immediately by observing that

$$\Psi^{②}(u) = \mathsf{exp}\left(-\frac{u^2\sigma^2}{2}\right)$$

is the characteristic exponent of the random variable $\mathsf{N}(\,0, \sigma^2\,)$, whose characteristic function reads as

$$\Phi_{\mathsf{N}(0,\sigma^2)}(u) = \mathsf{exp}\left(-\frac{u^2\sigma^2}{2}\right) \ .$$

By combining parts ① and ②, we arrive at a more general stochastic process, whose characteristic function

$$\Phi_{B_t}(u) = \mathsf{exp}\left(\mathrm{i}u\gamma t - \frac{u^2\sigma^2}{2}\, t\right)$$

is associated to the random variable

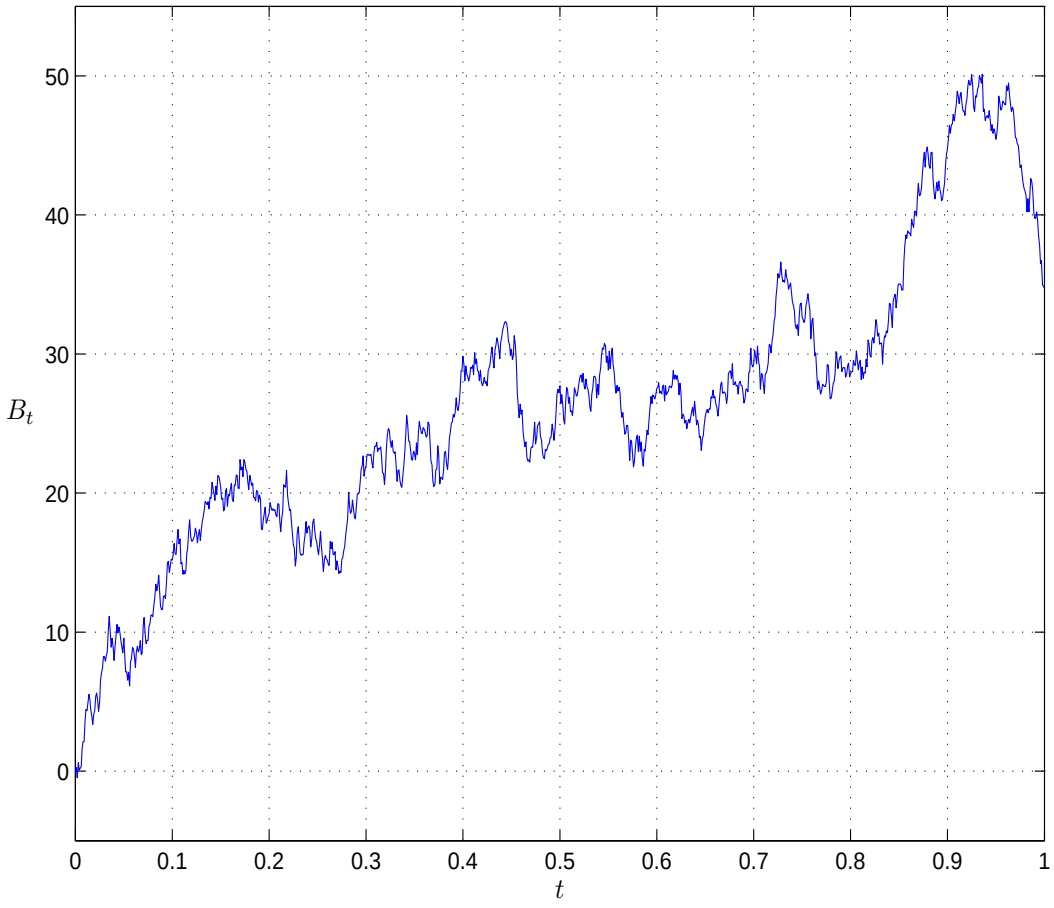
$$B_t := \gamma t + \sigma W_t \ , \quad (1.1.3)$$

i.e., the **Brownian motion with drift**. Again, this follows from the fact that

$$\Psi(u) = \mathrm{i}u\gamma - \frac{u^2\sigma^2}{2}$$

is the characteristic exponent of the random variable $\mathsf{N}(\gamma, \sigma^2)$. The Brownian motion with drift is used to describe the dynamics of the underlying (log) stochastic price process in Black and Scholes (1973)'s option pricing model. Figure 1.1 illustrates a simulated sample path of a Brownian motion with drift (1.1.3).

Figure 1.1: *Sample Path of Brownian Motion with Drift*
This figure depicts a simulated sample path of the Brownian motion with drift (1.1.3) with $\gamma = 0.02$ and $\sigma = 0.75$. The characteristic triplet of B_t reads as $(\gamma, \sigma, 0)$. The sampling interval is scaled down to $[0, 1]$.
This exemplifies the very defining feature of Brownian motion: continuous sample paths.



The third and forth part of the characteristic exponent in Equation (1.1.2) define continuous-time stochastic processes which are known as compound Poisson processes. To be more precise, let $\{N_t\}_{t\geqslant 0}$ be a Poisson process with jump intensity $0 < \lambda < \infty$ and $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of iid copies of a random variable X with probability distribution

function F. If all of these random variables are mutually independent, then the (random) partial sum

$$S_{N_t} := X_1 + \cdots + X_{N_t} = \sum_{k=1}^{N_t} X_k \ , \tag{1.1.4}$$

where $S_0 = 0$ or $X_0 = 0$ is assumed for $N_t = 0$, defines a **compound Poisson process**. As can be shown (Cont and Tankov, 2003), any continuous-time stochastic process is a Lévy process with piecewise constant sample paths. However, in contrast to the underlying Poisson process N_t which governs the number of discontinuities of the sample paths of S_{N_t}, the jump sizes of S_{N_t} are not necessarily equal to one, but follow the jump distribution F. In Appendix A.1, we derive the characteristic function of the compound Poisson process S_{N_t} in (1.1.4),

$$\Phi_{S_{N_t}}(u) = \mathsf{E}\big[\, \mathsf{e}^{\mathsf{i}uS_{N_t}} \,\big] = \mathsf{exp}\left\{ t \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1)\lambda F(\mathsf{d}x) \right\} = \mathsf{exp}\big\{ t\Psi(u) \big\} \ ,$$

where $\Psi(u) = \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1)\lambda F(\mathsf{d}x)$ is the characteristic exponent of S_{N_t}. Figure 1.2 illustrates a simulated sample path of a compound Poisson process (1.1.4).

If we restrict $X_k = 1$ for all $k \in \mathbb{N}$, then the compound Poisson process S_{N_t} collapses o the special case of a Poisson process N_t with intensity λ,

$$S_{N_t} = \sum_{k=1}^{N_t} X_k = \sum_{k=1}^{N_t} \delta_1(x) = \sum_{k=1}^{N_t} 1 = N_t \ , \tag{1.1.5}$$

with characteristic function

$$\Phi_{S_{N_t}}(u) = \mathsf{exp}\left\{ t \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1)\lambda \delta_1(\mathsf{d}x) \right\} = \mathsf{exp}\big\{ t\lambda(\mathsf{e}^{\mathsf{i}u} - 1) \big\} = \mathsf{exp}\big\{ t\Psi(u) \big\} \ ,$$

where δ_1 is the (Dirac) point mass at 1, and $\Psi(u) = \lambda(\mathsf{e}^{\mathsf{i}u} - 1)$ is the characteristic exponent of a Poisson process. The characteristic function of a **compensated compound Poisson process**

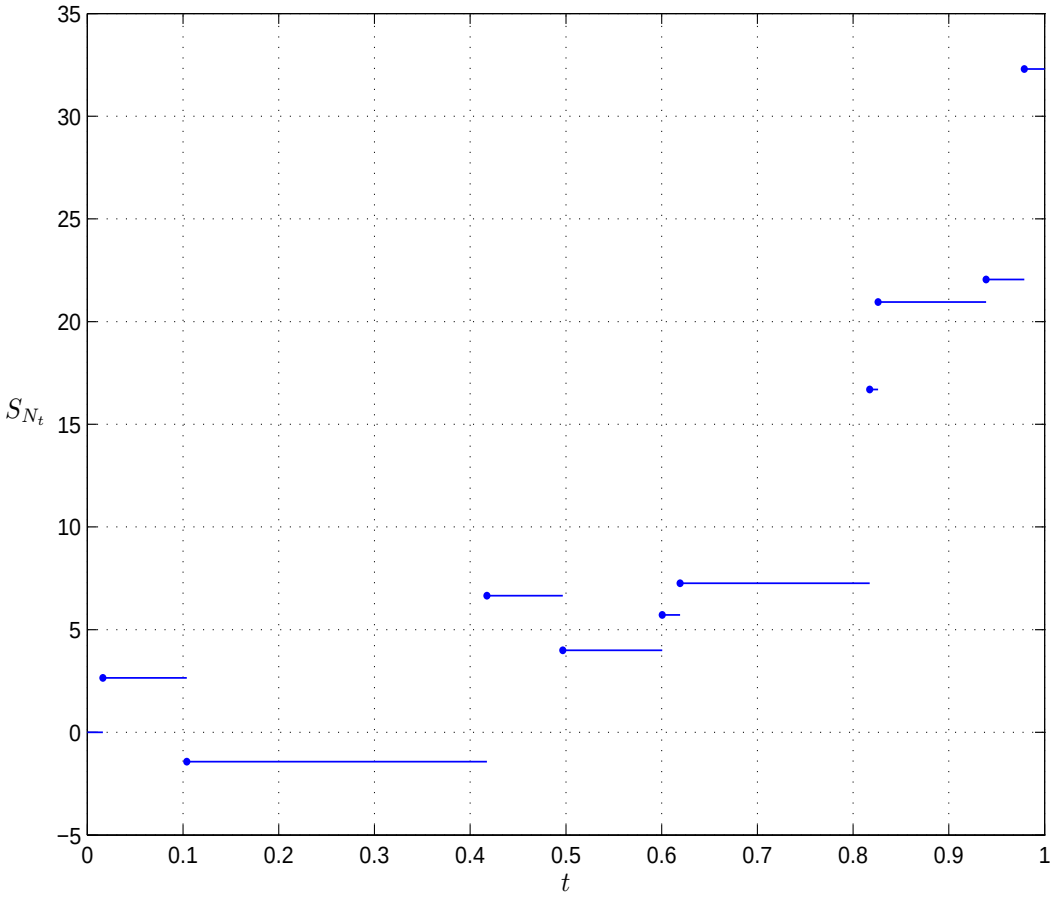
$$\tilde{S}_{N_t} := S_{N_t} - \mathsf{E}[\, S_{N_t} \,] \tag{1.1.6}$$

is derived in Appendix A.1:

$$\Phi_{\tilde{S}_{N_t}}(u) = \mathsf{E}\Big[\, \mathsf{e}^{\mathsf{i}u\tilde{S}_{N_t}} \,\Big] = \mathsf{exp}\left\{ t \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}ux)\lambda F(\mathsf{d}x) \right\} = \mathsf{exp}\big\{ t\Psi(u) \big\} \ ,$$

where $\Psi(u) = \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}ux)\lambda F(\mathsf{d}x)$ is the characteristic exponent of $\tilde{S}_{N_t}$. Figure 1.3 illustrates a simulated sample path of a compensated compound Poisson process (1.1.6).

Figure 1.2: *Sample Path of Compound Poisson Process*
This figure depicts a simulated sample path of the compound Poisson process (1.1.4) with intensity $\lambda = 0.01$ and whose jumps are normally distributed with $\mu_X = 1$ and $\sigma_X = 5$. The characteristic triplet of S_{N_t} reads as $\big(0, 0, \lambda \mathsf{N}\big(\,\mu_X, \sigma_X^2\,\big)\big)$. The sampling interval is scaled down to $[0, 1]$.



If we proceed by combining a Brownian motion B_t with drift γ and a compound Poisson process with intensity λ and jump size distribution F, we arrive at the **jump-diffusion Lévy process**

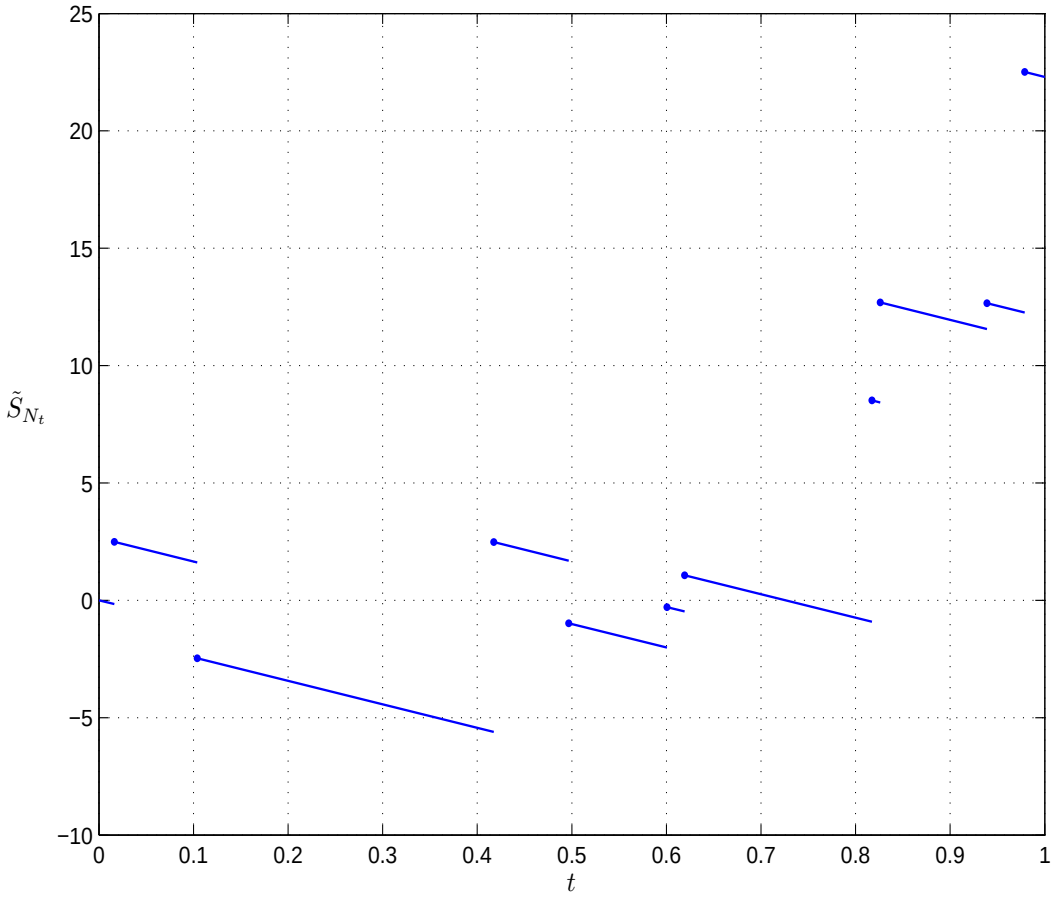
$$X_t = B_t + \sum_{k=1}^{N_t} X_k = \gamma t + \sigma W_t + \sum_{k=1}^{N_t} X_k \;, \tag{1.1.7}$$

whose characteristic exponent follows directly from the above parts:

$$\Psi(u) = \mathsf{i}u\gamma - \frac{u^2\sigma^2}{2} + \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1)\lambda F(\mathsf{d}x) \;.$$

In the literature on financial economics, the idea of merging compound Poisson and Brownian components was first introduced by Press (1967). Later on, Merton (1976) proposed his jump-diffusion Lévy processes by augmenting the Black and Scholes (1973)

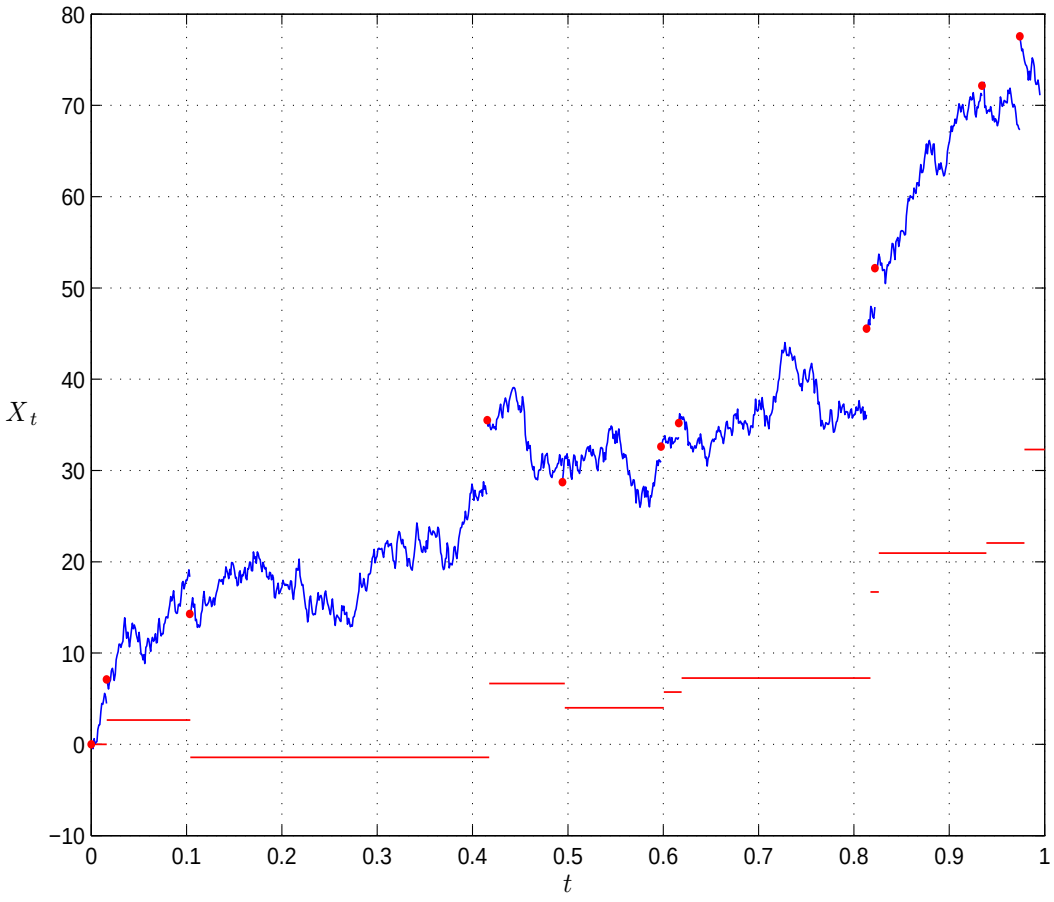
Figure 1.3: *Sample Path of Compensated Compound Poisson Process*
This figure depicts a simulated sample path of the compensated version (1.1.6) of the compound Poisson process (1.1.4) in Figure 1.2. The characteristic triplet of $\tilde{S}_{N_t}$ reads as $\big(\gamma_1, 0, \lambda \mathsf{N}\big(\mu_X, \sigma_X^2\big)\big)$ where $\gamma_1 := -\lambda\mu_X$ follows from (1.1.9) and κ_1 in Section 1.2. The sampling interval is scaled down to $[0, 1]$.



model by Gaussian jumps. Kou (2002) introduced a jump-diffusion Lévy process with double-exponentially distributed jumps. Figure 1.4 illustrates a simulated sample path of a jump-diffusion Lévy process (1.1.7).

Although the Merton (1976) was a major landmark for mathematical finance, its roots date back to 1930 when Kolmogorov and de Finetti erroneously suggested the jump-diffusion Lévy model to be the most general form of a Lévy process (Sato, 1999, p. 37). In order to move into the direction of the general form of the characteristic exponent of Theorem 1.3, we first replace the compound Poisson process by its compensated version. Using the above ingredients, a **jump-diffusion Lévy process with compensated**

Figure 1.4: *Sample Path of Jump-Diffusion Lévy Process*
This figure depicts a simulated sample path of the jump-diffusion Lévy process (1.1.7) which corresponds to the superposition of the processes of Figures 1.1 and 1.2. The characteristic triplet of X_t reads as $\big(\gamma, \sigma, \lambda \mathsf{N}\big(\,\mu_X, \sigma_X^2\,\big)\,\big)$. The sampling interval is scaled down to $[0, 1]$.



jumps,

$$X_t = \gamma t + \sigma W_t + \underbrace{\left(\sum_{k=1}^{N_t} X_k - \lambda t \mu_X \right)}_{=\tilde{S}_{N_t}},$$

has characteristic function

$$\begin{aligned}
\Phi_{X_t}(u) = \mathsf{E}\big[\,\mathrm{e}^{\mathrm{i}uX_t}\,\big] &= \mathsf{E}\left[\,\exp\left\{\mathrm{i}u\left(\gamma t + \sigma W_t + \sum_{k=1}^{N_t} X_k - \lambda t \mu_X\right)\right\}\right] \\
&= \mathsf{E}\big[\,\exp\big\{\mathrm{i}u(\gamma t + \sigma W_t)\big\}\,\big]\,\mathsf{E}\left[\,\exp\left\{\mathrm{i}u\left(\sum_{k=1}^{N_t} X_k - \lambda t \mu_X\right)\right\}\right] \\
&= \exp\left\{t\left(\mathrm{i}u\gamma - \frac{u^2\sigma^2}{2}\right)\right\} \exp\left\{t \int_{\mathbb{R}\setminus\{0\}} (\mathrm{e}^{\mathrm{i}ux} - 1 - \mathrm{i}ux)\lambda F(\mathrm{d}x)\right\}
\end{aligned}$$

$$= \mathsf{exp}\big\{t\Psi(u)\big\} \ ,$$

where the characteristic exponent

$$\Psi(u) = \mathsf{i}u\gamma - \frac{u^2\sigma^2}{2} + \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}ux)\lambda F(\mathsf{d}x) \tag{1.1.8}$$

now resembles the characteristic exponent of Theorem 1.3 more closely, although there are still two crucial differences. Interestingly, both of them originate from the same cause: the Lévy measure ν. The Lévy measure implied by the characteristic exponent $\Psi(u)$ in (1.1.8) is a finite measure, i.e.,

$$\nu\big(\mathbb{R}\setminus\{0\}\big) = \int_{\mathbb{R}\setminus\{0\}} \nu(\mathsf{d}x) < \infty \ .$$

Despite of being finite, the Lévy measure of a jump-diffusion Lévy process is generally not a probability measure since

$$\nu\big(\mathbb{R}\setminus\{0\}\big) = \int_{\mathbb{R}\setminus\{0\}} \nu(\mathsf{d}x) = \int_{\mathbb{R}\setminus\{0\}} \lambda F(\mathsf{d}x) = \lambda \int_{\mathbb{R}\setminus\{0\}} F(\mathsf{d}x) = \lambda \neq 1 \ .$$

Nevertheless, the jump size distribution constitutes a probability measure:

$$0 \leqslant F(\mathsf{d}x) = \frac{\nu(\mathsf{d}x)}{\lambda} = \frac{\nu(\mathsf{d}x)}{\nu\big(\mathbb{R}\setminus\{0\}\big)} \leqslant 1 \ .$$

Put differently, for a finite Lévy measure, there exists a factorization of ν into the expected number of jumps per unit of time, $\lambda = \mathsf{E}[\,X_t\,]/t$, and the jump size distribution, F. Generally, this factorization holds for models, where the jump component corresponds to a compound Poisson process only which, by definition, has finite jump intensity, i.e., has a finite number of jumps on any finite time interval.

This observation deepens our understanding of component ③ in (1.1.8). As already mentioned,

$$\Psi^{③}(u) = \int_{\{|x|>1\}} (\mathsf{e}^{\mathsf{i}ux} - 1)\nu(\mathsf{d}x) = \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1)\mathbb{1}_{\{|x|>1\}}\nu(\mathsf{d}x)$$

is the characteristic exponent of a compound Poisson process, but it is actually a compound Poisson process restricted to have only 'large' jumps, i.e., jumps larger than 1. The corresponding jump size distribution,

$$F(\mathsf{d}x) = \frac{\nu(\mathsf{d}x)}{\nu\big(\mathbb{R}\setminus[-1,1]\big)}\,\mathbb{1}_{\{|x|>1\}} \ ,$$

is derived from $\nu(\mathsf{d}x)\,\mathbb{1}_{\{|x|>1\}} = \lambda F(\mathsf{d}x)$ by plugging in the expression $\lambda = \nu\big(\mathbb{R}\backslash[-1,1]\big) > 0$. Recall that from the properties of any Lévy measure $\lambda = \nu\big(\mathbb{R}\backslash[-1,1]\big) < \infty$ such that there is only a finite number of jumps in any finite time interval.

These neat considerations all break down when we expect an infinite number of jumps in a finite time interval, i.e., $\lambda = \nu\big(\mathbb{R}\backslash\{0\}\big) = \infty$, which is perfectly possible since Theorem 1.3 allows for $\nu\big([-1,1]\backslash\{0\}\big) = \infty$. If this happens, the Lévy measure ν is an infinite (but σ-finite) measure satisfying additional integrability conditions laid out in Theorem 1.3. The problem resulting from this complication is that the integrand $(\mathsf{e}^{\mathsf{i}ux} - 1)$ may not be ν-integrable, and there arises the need for adjusting the integrand by compensating for the 'small' jumps of the Lévy process. After compensating, the integrand of the characteristic exponent $\Psi(u)$ in Theorem 1.3 becomes integrable with respect to the Lévy measure ν since, on the one hand, ν is bounded outside of some neighborhood of 0 due to $\int_{\{|x|>1\}} \nu(\mathsf{d}x) < \infty$ and, on the other hand,

$$\begin{aligned} \mathsf{e}^{\mathsf{i}ux} &= 1 + \mathsf{i}ux + \frac{(\mathsf{i}ux)^2}{2} + \cdots \\ \mathsf{e}^{\mathsf{i}ux} + 1 + \mathsf{i}ux &= O\big(|x|^2\big) \qquad \text{as } |x| \to 0, \end{aligned}$$

for all $u \in \mathbb{R}$, such that the compensated integrand satisfies $\int_{\{0<|x|\leqslant 1\}} x^2 \nu(\mathsf{d}x) < \infty$.

A question, which naturally arises at this point, is whether choosing 1 as the threshold level in $\mathbb{1}_{\{0<|x|\leqslant 1\}}$, for separating the 'small' and 'large' jumps, is innocuous. Indeed, this choice is somewhat arbitrary. It is possible to replace the indicator function by alternative truncation functions $h(x)$ as long as they warrant that the integrand of $\Psi(u)$ in Theorem 1.3 is integrable with respect to the Lévy measure ν. It is actually sufficient that $h : \mathbb{R} \to \mathbb{R}$ is a bounded, measurable function such that

$$\begin{aligned} h(x) &= 1 + o\big(|x|\big) \qquad \text{as } |x| \to 0 \\ h(x) &= O\big(1/|x|\big) \qquad \text{as } |x| \to \infty. \end{aligned}$$

Then, the characteristic function in Theorem 1.3 can be rewritten as

$$\Phi_{X_t}(u) = \mathsf{exp}\left\{\mathsf{i}u\gamma_h - \frac{u^2\sigma^2}{2} + \int_{\mathbb{R}\backslash\{0\}} \big[\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}uxh(x)\big]\nu(\mathsf{d}x)\right\} ,$$

where

$$\gamma_h = \gamma + \int_{\mathbb{R}\backslash\{0\}} x\big[h(x) - \mathbb{1}_{\{0<|x|\leqslant 1\}}\big]\nu(\mathsf{d}x) \, .$$

For example, if $\int_{\{0<|x|\leqslant 1\}} |x|\nu(\mathsf{d}x) < \infty$ holds, then it is sufficient for the integrand to

satisfy

$$\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}uxh(x) = O\big(|x|\big) \qquad \text{as } |x| \to 0,$$

which suggests to set $h(x) = 0$. Thus,

$$\Phi_{X_t}(u) = \mathsf{exp}\left\{\mathsf{i}u\gamma_0 - \frac{u^2\sigma^2}{2} + \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1)\nu(\mathsf{d}x)\right\} ,$$

with $\gamma_0 = \gamma - \int_{\{0<|x|\leqslant 1\}} x\nu(\mathsf{d}x)$, which corresponds to the characteristic function of a (compound) Poisson process. Moreover, assuming $\int_{\mathbb{R}\setminus\{0\}} |x|\nu(\mathsf{d}x) < \infty$, we can set $h(x) = 1$ such that

$$\Phi_{X_t}(u) = \mathsf{exp}\left\{\mathsf{i}u\gamma_1 - \frac{u^2\sigma^2}{2} + \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}ux)\nu(\mathsf{d}x)\right\} , \tag{1.1.9}$$

with $\gamma_1 = \gamma - \int_{\{|x|>1\}} x\nu(\mathsf{d}x)$, is the characteristic function of a compensated (compound) Poisson process. Furthermore, it is noteworthy that σ and ν are invariant to the choice of $h(x)$. For more details, see Sato (1999, pp. 38-39) and Shiryaev (1999, pp. 196-197).

We postpone an in-depth analysis of the jump component to the Section 1.3. For now, we simply present the second fundamental result on Lévy processes which follows directly from (1.1.2).

Theorem 1.6 (Lévy-Itō Decomposition) *The sample paths of any Lévy process* $X = \{X_t\}_{t\geqslant 0}$ *can be decomposed into:*

$$X_t = \underbrace{B_t}_{\substack{\textit{continuous}\\ \textit{part}}} + \underbrace{\sum_{0<s\leqslant t} \Delta X_s}_{\textit{jump part}} ,$$

where $B_t = \gamma t + \sigma W_t$ *is a Brownian motion with drift* γ *and Wiener process* W_t*, and* $\Delta X_s = X_s - X_{s-}$ *are continuous-time increments with* $X_{s-} = \mathsf{lim}_{u\nearrow s} X_u$*.*

The rest of this section is devoted to the connection between Lévy processes and the **theory of semimartingales**, and serves as a motivation of next section's main theme: Poisson random measures. Note that any compensated Lévy process $\big\{X_t - \mathsf{E}[\,X_t\,]\big\}_{t\geqslant 0}$ is a martingale such that it can be decomposed into a process with a linear drift and a martingale:

$$\begin{aligned} X_t &= X_t - \mathsf{E}[\,X_t\,] + \mathsf{E}[\,X_t\,] \\ &= \mathsf{E}[\,X_1\,]\,t + \big(X_t - \mathsf{E}[\,X_t\,]\big) , \end{aligned}$$

where $\mathsf{E}[\,X_1\,] = \gamma$ and the second summand comprises the Brownian and jump component,

in general. This representation is reminiscent of the general theory of semimartingales which allows for the decomposition

$$X_t = X_0 + V_t + M_t \ , \tag{1.1.10}$$

where $\{V_t\}_{t\geqslant 0}$ is process with finite variation on any finite time interval and $\{M_t\}_{t\geqslant 0}$ is a local martingale. Although this decomposition is generally not unique, this problem can be resolved by eliminating the 'big' jumps of X_t, say, jumps with absolute magnitude larger than 1:

$$\sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|>1\}} \ .$$

The resulting semimartingale is a so-called **special semimartingale**, with bounded jump sizes, which is unique. Furthermore, the semimartingale decomposition (1.1.10) can be casted in the **canonical decomposition**:

$$X_t = X_0 + V_t + M_t^c + M_t^d + \sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|>1\}}$$

$$\underbrace{X_t - X_0 - \sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|>1\}}}_{\text{special semimartingale}} = V_t + M_t^c + M_t^d \ ,$$

where V_t is now a predictable process with finite variation, and the local martingale M_t in (1.1.10) has been (uniquely) decomposed into two orthogonal, local martingales: a local martingale M_t^c with continuous paths and a purely discontinuous, local martingale M_t^d. For further details, see Jacod and Shiryaev (2003, pp. 40–44) or Prakasa Rao (1999, pp. 71–73).

1.2 Further Properties & Classification

As already mentioned, all information on the jump behavior of a Lévy process is comprised in its Lévy measure. In this subsection, we will take a closer look at the decomposition of jumps and at how it is connected to the so-called **distributional and path properties** of a Lévy process.

Let us first focus on the 'large' jumps whose frequency is characterized by the tails of the Lévy measure ν. An interesting result states that the finiteness of moments solely depends on the $\mathbb{1}_{\{|x|>1\}}$–part of ν (Sato, 1999, Theorem 25.3). To be more precise, the pth moment of a Lévy process $X = \{X_t\}_{t\geqslant 0}$ exists, i.e., $\mathsf{E}[\,|X_t|^p\,] < \infty$, for all $p \geqslant 1$, if

and only if

$$\int_{\{|x|>1\}} |x|^p \nu(\mathsf{d}x) < \infty$$

holds for *some* $t > 0$ (due to infinite divisibility of X_t). At first glance, this appears to be surprising when recalling the integrability condition on ν in Theorem 1.3. There, the condition $\int_{\{0<|x|\leqslant 1\}} x^2 \nu(\mathsf{d}x) < \infty$ was put forward as a defining property for all Lévy measures. On the one hand, this tells us that the 'small' jumps around the origin do not cause any convergence problems for moments of order greater than 1. On the other hand, this condition does not rule out the possibility of $\int_{\{0<|x|\leqslant 1\}} x\nu(\mathsf{d}x) = \infty$ which is a problem for the existence of the first moment. In particular, if this holds, then

$$\int_{\mathbb{R}\setminus\{0\}} x\nu(\mathsf{d}x) = \int_{\{|x|>1\}} x\nu(\mathsf{d}x) + \int_{\{0<|x|\leqslant 1\}} x\nu(\mathsf{d}x) = \infty \,,$$

even when assuming convergence in the tails, i.e., $\int_{\{|x|>1\}} x\nu(\mathsf{d}x) < \infty$. However, this issue can easily be clarified by deriving the cumulants of X_t via plugging the cumulant generating function $\ln\Phi(u) = t\Psi(u)$ in (A.1.1). From the characteristic exponent in Theorem 1.3, it is possible to obtain the following cumulants:

$$\begin{aligned}
\kappa_1 &= \frac{t}{\mathsf{i}}\,\Psi'(0) = \frac{t}{\mathsf{i}}\left\{\mathsf{i}\gamma - u\sigma^2 + \int (\mathsf{i}x\mathsf{e}^{\mathsf{i}ux} - \mathsf{i}x\mathbb{1}_{\{0<|x|\leqslant 1\}})\nu(\mathsf{d}x)\right\}\Big|_{u=0} \\
&= \frac{t}{\mathsf{i}}\left\{\mathsf{i}\gamma + \int (\mathsf{i}x - \mathsf{i}x\mathbb{1}_{\{0<|x|\leqslant 1\}})\nu(\mathsf{d}x)\right\} = \frac{t}{\mathsf{i}}\left\{\mathsf{i}\gamma + \int_{\{|x|>1\}} \mathsf{i}x\nu(\mathsf{d}x) + \int_{\{0<|x|\leqslant 1\}} 0\nu(\mathsf{d}x)\right\} \\
&= t\left\{\gamma + \int_{\{|x|>1\}} x\nu(\mathsf{d}x)\right\} \,, \\
\kappa_2 &= \frac{t}{\mathsf{i}^2}\,\Psi''(0) = -t\left\{-\sigma^2 + \int (\mathsf{i}x)^2\mathsf{e}^{\mathsf{i}ux}\nu(\mathsf{d}x)\right\}\Big|_{u=0} = t\left\{\sigma^2 + \int x^2\nu(\mathsf{d}x)\right\} \,, \\
\kappa_p &= \frac{t}{\mathsf{i}^p}\,\Psi^{(p)}(0) = \frac{t}{\mathsf{i}^p}\int (\mathsf{i}x)^p\mathsf{e}^{\mathsf{i}ux}\nu(\mathsf{d}x)\Big|_{u=0} = t\int x^p\nu(\mathsf{d}x) \,,
\end{aligned}$$

for all $p \geqslant 3$. Finally, mean, variance, skewness, and excess kurtosis of X_t follow from (A.1.2)–(A.1.5). These computations show two things: First, the existence of moments does indeed only depend upon the tail behavior of ν. Second, all cumulants are linearly increasing with time t. As usual, cumulants are a very convenient devise for describing deviations from normality (or from a Brownian motion B_t with drift). It is noteworthy that all Lévy processes with jump component are leptokurtic, i.e.,

$$\kappa_4 = t\int x^4\nu(\mathsf{d}x) > 0 \,,$$

if κ_4 exists at all. Moreover, properties of the marginal distribution of X_t can directly

be translated into properties of the marginal distribution of its increments. This is a straightforward consequence of Condition C3 of Definition 1.1, i.e., let $s = 1$ such that $X_{t+1} - X_t \stackrel{d}{=} X_1$. Put differently, the properties of the marginal law of one-period returns are equal to those of the marginal law of its level series at time 1. Thus, the increments of all Lévy processes with jump component are leptokurtic (if κ_4 exists), which shows their potential gains for explaining some of the stylized facts of financial time series.

The Lévy measure is not only informative about the moments (or, more broadly, the distributional properties) of a Lévy process, but it is also responsible for a diverse array of **path properties** of Lévy processes. This is sometimes dubbed as the **fine structure** of Lévy processes for which the 'small' jumps turn out to be crucial.

A very useful starting point is a **classification** scheme for Lévy processes due to Sato (1999, Definition 11.9 & Theorem 21.9).

Definition 1.7 (Classification of Lévy Processes) *For the Lévy process $X = \{X_t\}_{t \geqslant 0}$ with characteristic triplet (γ, σ, ν), we define the following three classes:*

1. *If $\sigma = 0$ and $\nu\big(\mathbb{R}\backslash\{0\}\big) < \infty$, then X is defined to be of **type A**: a purely non-Gaussian Lévy process with finite activity.*
2. *If $\sigma = 0$, $\nu\big(\mathbb{R}\backslash\{0\}\big) = \infty$, and $\int_{\{0<|x|\leqslant 1\}} |x|\nu(\mathrm{d}x) < \infty$, then X is defined to be of **type B**: a purely non-Gaussian Lévy process with infinite activity and finite variation.*
3. *If $\sigma > 0$ or $\int_{\{0<|x|\leqslant 1\}} |x|\nu(\mathrm{d}x) = \infty$, then X is defined to be of **type C**: a Lévy process with infinite variation.*

A Lévy process X is said to have **finite activity**, if P-almost all sample paths of X have only a finite number of jumps on any finite time interval $(0, t]$. Likewise, a Lévy process X is said to have **infinite activity**, if P-almost all sample paths have a (countably) infinite number of jumps on any finite time interval $(0, t]$. In order to relate these notions to properties of the Lévy measure ν, note that, due to the integrability condition in Theorem 1.3, any Lévy measure has finite mass in the tails, i.e.,

$$\int_{\{|x|>1\}} \nu(\mathrm{d}x) < \infty \,,$$

such that the number of 'large' jumps is bounded (P-a.s.). Put differently, there is only a finite number of 'large' jumps on any finite time interval $(0, t]$. Consequently,

$$\nu\big(\mathbb{R}\backslash\{0\}\big) = \int_{\mathbb{R}\backslash\{0\}} \nu(\mathrm{d}x) = \infty$$

is only possible, if there is an infinite number of 'small' jumps around the origin, i.e.,

$$\int_{\{0<|x|\leqslant 1\}} \nu(\mathsf{d}x) = \infty \ .$$

In sum, a Lévy process X has finite activity if $\nu\big(\mathbb{R}\backslash\{0\}\big) < \infty$, while it has infinite activity if $\nu\big(\mathbb{R}\backslash\{0\}\big) = \infty$. Recall that the **variation** of a stochastic process $X = \{X_t\}_{t\geqslant 0}$ is defined by

$$V_X[0,t] := \sup_{\Delta_n} \sum_{k=1}^{n} |X_{t_k} - X_{t_{k-1}}| \ ,$$

where the supremum is taken over all possible partitions $\Delta_n = \{t_0, t_1, \ldots, t_n : 0 = t_0 < t_1 < \cdots < t_n = t\}$ of $[0,t]$. The first sub-class of Lévy process, which are of type C, are models containing a continuous Brownian component. It is a well-known result that Brownian motions have infinite variation (Kallenberg, 2002, Corollary 13.10). What is even more interesting is that, for another subclass of type-C processes, **infinite variation** may be a result of the condition

$$\int_{\{0<|x|\leqslant 1\}} |x|\nu(\mathsf{d}x) = \int |x| \mathbb{1}_{\{0<|x|\leqslant 1\}} \nu(\mathsf{d}x) = \infty$$

implying that the sum of 'small' jumps,

$$\sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{0<|x|\leqslant 1\}} \ ,$$

does not converge for P-almost all sample paths of X. This technically delicate issue is indeed the reason why (1.1.8) is not the most general form of the characteristic exponent for Lévy processes, and it will be tackled in much more detail in the next section.

1.3 Poisson Random Measure & Lévy Density

In this section, we present a substantial refinement of the assertion of Theorem 1.6 and introduce the object of our statistical interest.

Recall that, according to Theorem 1.6, the sample paths of any Lévy process can be decomposed as

$$X_t = B_t + \sum_{0<s\leqslant t} \Delta X_s \ ,$$

where $B_t = \gamma t + \sigma W_t$ is a continuous Brownian motion with drift, while the second summand represents the discontinuous jump part of X_t. Without loss of generality, let

$B_t \equiv 0$ for all $t \geqslant 0$ such that we obtain a pure-jump Lévy process:

$$X_t = \sum_{0<s\leqslant t} \Delta X_s \ .$$

From Theorem 1.3, we also known that

$$\nu\big(\mathbb{R}\backslash\{0\}\big) = \int_{\mathbb{R}\backslash\{0\}} \nu(\mathsf{d}x) = \infty$$

may be possible, i.e., jumps may arrive at an infinitely high rate. When this happens, the jump part $\sum_{0<s\leqslant t} \Delta X_s$ has infinitely many summands which may diverge, in general. As we have already seen from the discussion of a Lévy process' variation in Subsection 1.2, it might be helpful to distinguish between the effects of 'large' and 'small' jumps in order to find a remedy to this divergence problem. Thus, we go on by separating 'large' and 'small' jumps which are defined in accordance with the general truncation scheme of Theorem 1.3:

$$X_t = \sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|>1\}} + \sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{0<|\Delta X_s|\leqslant 1\}} \ .$$

Here, the sum of 'large' jumps poses no problem because $\nu\big(\mathbb{R}\backslash[-1,1]\big) < \infty$ has already been derived such that there are only finitely many 'large' jumps. Another way to see this is by recalling the discussion of the characteristic exponent $\Psi^{③}(u)$ in Section 1.1, where the jump component involving 'large' jumps was associated with a compound Poisson process which has finite intensity $0 < \lambda < \infty$.

In contrast to the sum of 'large' jumps, we have already concluded from the integrability condition of the Lévy measure ν in Theorem 1.3 that it is the frequency of jumps located near the origin which may be infinite, i.e., $\int_{\{0<|x|\leqslant 1\}} \nu(\mathsf{d}x) = \infty$. Consequently, the sum of 'small' jumps is infinite and can diverge when there are too many 'small' jumps.[2]

The proposed solution for getting a handle on this problem begins with truncating all jumps smaller than some $\epsilon > 0$ and consider its limit as $\epsilon \searrow 0$:

$$X_t = \underbrace{\sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|>1\}}}_{=:X_t^L} + \lim_{\epsilon\searrow 0} \underbrace{\sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{\epsilon\leqslant|\Delta X_s|\leqslant 1\}}}_{=:X_t^\epsilon} \ . \tag{1.3.1}$$

Unfortunately, there is still no guarantee that the limit exists, and we have to resort to a downweighting scheme by subtracting the average change of X_t along $(0,t]$ due to the

[2] Actually, due to the càdlàg property of sample paths, there can only be countably infinite number of 'small' jumps and a finite number of 'large' jumps on any finite time interval (compare Appendix A.1).

jumps X_t^ϵ in the limit expression. Obviously, this is nothing else than compensation, where the corresponding average is deduced from the intensity with which these jumps arrive. Finally, it can be shown that the limit exists in probability as $\epsilon \searrow 0$:

$$X_t = \sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|>1\}} + \lim_{\epsilon \searrow 0} \left[\sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{\epsilon\leqslant|\Delta X_s|\leqslant 1\}} - t \int x \mathbb{1}_{\{\epsilon\leqslant|x|\leqslant 1\}} \nu(\mathsf{d}x) \right] .$$

It is important to note that this limit expression cannot, in general, be simplified by splitting it up into the difference of two separate limits because these individual limits may not exist.

Beyond these heuristic considerations, a more thorough derivation calls for the introduction of a devise allowing us to characterize the stochastic behavior of the jumps (or increments) of a Lévy process in a sensible way, and which is based on a *point-process* perspective. There arise two complications when trying to characterize the jumps of a Lévy process.

First, as we already know from the discussion of Condition C5 in Remark 1.2, the jump times of X_t are random. Thus, the occurrence of jumps $\{\Delta X_s : 0 < s \leqslant t\}$ is random and requires the introduction of a sequence of random times representing the jump times of a Poisson process $N_t(\omega)$ with intensity λ. From the basic properties of a Poisson process, we know that the inter-arrival times of its jumps are iid following an exponential law with parameter λ (see Appendix A.1). Moreover, the Poisson process has only jumps of unit size, by definition. An alternative approach to characterizing a Poisson process is via its **(Poisson) random measure**,

$$J\big(\omega; (0,t]\big) := \sharp\big\{n \in \mathbb{N} : 0 < \tau_n(\omega) \leqslant t\big\} ,$$

which counts the number of jumps of the Poisson process (for a given sample path ω) occurring up to time t such that

$$N_t(\omega) = \sum_{\substack{0<s\leqslant t \\ \Delta N_s(\omega)=1}} \Delta N_s(\omega) = \sharp\big\{n \in \mathbb{N} : 0 < \tau_n(\omega) \leqslant t\big\} = \sum_{n\in\mathbb{N}} \mathbb{1}_{\{\tau_n(\omega)\leqslant t\}} = \int_0^t J(\omega; \mathsf{d}s) .$$

Put differently, any Poisson process can be associated to and be completely characterized by its corresponding random measure.

This approach can easily be extended to tackle the second complication, i.e., the randomness of jump sizes. The idea is to define different Poisson processes with increments falling into different Borel subsets of $\mathbb{R}\backslash\{0\}$ and with corresponding intensities depending on these individual subsets. Then, the superposition of these Poisson processes is asso-

ciated to the extended random measure. The random jump times of a Poisson process with jump size B are given by

$$\begin{aligned}
\tau_1^B(\omega) &:= \inf\Big\{t > 0 : \Delta X_t(\omega) \subseteq B, B \in \mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)\Big\} \\
\tau_2^B(\omega) &:= \inf\Big\{t > \tau_1^B(\omega) : \Delta X_t(\omega) \subseteq B, B \in \mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)\Big\} \\
&\vdots \\
\tau_n^B(\omega) &:= \inf\Big\{t > \tau_{n-1}^B(\omega) : \Delta X_t(\omega) \subseteq B, B \in \mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)\Big\} \\
&\vdots
\end{aligned}$$

while the corresponding counting measure on $\mathscr{B}\big((0,t] \times B\big)$ is defined as

$$\begin{aligned}
J\big(\omega; (0,t] \times B\big) &:= \sharp\Big\{\big(s, \Delta X_s(\omega)\big) \subseteq A : s \in (0,t], \Delta X_s(\omega) \subseteq B, A = (0,t] \times B\Big\} \\
&= \sum_{0<s\leqslant t} \mathbb{1}_{\{\Delta X_s(\omega)\subseteq B\}} = \sum_{n\in\mathbb{N}} \mathbb{1}_{\{\tau_n^B(\omega)\leqslant t\}} \ ,
\end{aligned}$$

where we consider Borel subsets A of the product space $\mathbb{R}_+ \times \mathbb{R}\backslash\{0\}$. The last equality shows that it is indeed possible to observe a countably infinite number of jumps of magnitude falling into B during any finite time interval $[0,t]$.

For illuminating the roles played by ω, t, and B, we isolate their influences by a *ceteris paribus* analysis. On the one hand, the dependence on the sample path $\omega \in \Omega$ stresses the fact that $J\big(\omega; (0,t] \times B\big)$ is a random quantity. Hence, if we fix ω and t, the randomness and 'dynamics' of $J\big(\omega; (0,t] \times B\big)$ vanish, and $J\big(\boldsymbol{.}; (0,\boldsymbol{.}] \times B\big)$ becomes a σ-finite measure on the Borel sets B of $\mathbb{R}\backslash\{0\}$. On the other hand, if we fix B, the random process $J\big(\omega; (0,t] \times \boldsymbol{.}\big)$ counts its jumps occurring with fixed jump sizes falling into B.

For these extended random measures, Theorem 19.4 of Sato (1999) establishes their existence, while the following definition provides a complete characterization.

Definition 1.8 (Random Measure) *Let $\Big(\mathbb{R}\backslash\{0\}, \mathscr{B}\big(\mathbb{R}\backslash\{0\}\big), \nu\Big)$ be a σ-finite measure space. The counting process $J\big(\omega; (0,t] \times B\big)$ is a (Poisson) random measure, if it satisfies the following conditions:*

C1 For any $\omega \in \Omega$, $J\big(\,\boldsymbol{.}\,; (0,t] \times B\big)$ is a σ-finite measure on the product space $\mathscr{B}\big(\mathbb{R}_+ \times \mathbb{R}\backslash\{0\}\big)$.

C2 For any $B \in \mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)$, $J\big(\omega; (0,t] \times \boldsymbol{.}\big)$ follows a Poisson distribution with intensity $\nu(B)$.

C3 If $B_1, \ldots, B_n \in \mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)$ are disjoint, then $J\big(\omega; (0,t] \times B_1\big), \ldots, J\big(\omega; (0,t] \times B_n\big)$

are independent.

It is not hard to see that the random measure associated with the increments is indeed a Poisson random measure. First, note that the increments of $J := \Big\{J\big(\omega;(0,t]\times B\big)\Big\}_{t>0}$ are directly related to the increments of $X_t(\omega)$, i.e., the former counts the jumps occurring in the latter, or more formally,

$$J\big(\omega;(0,t]\times B\big) - J\big(\omega;(0,s]\times B\big) \in \mathscr{B}\Big(\big\{X_u(\omega) - X_v(\omega) : s \leqslant v < u \leqslant t\big\}\Big) \ .$$

Consequently, the stochastic properties of the increments of the Lévy process X, laid out in Definition 1.1, feed into those of the increments of J. In particular, $J\big(\omega;(0,t+s]\times B\big) - J\big(\omega;(0,t]\times B\big)$ is independent of $\mathscr{F}_t$, for all $t,s>0$, due to Condition C2 of Definition 1.1, while $J\big(\omega;(0,t+s]\times B\big) - J\big(\omega;(0,t]\times B\big) \stackrel{d}{=} J\big(\omega;(0,s]\times B\big)$, for all $t,s>0$, due to Condition C3 of Definition 1.1. Since the counting process J has iid increments, it is a Poisson process (Cont and Tankov, 2003), with intensity parameter depending on B, i.e., the expected number of jumps with size falling into B per unit of time.

Definition 1.9 (Lévy Measure) *Let $X = \big\{X_t(\omega)\big\}_{t\geqslant 0}$ be a Lévy process. The σ-finite measure ν on $\mathbb{R}\backslash\{0\}$ satisfying $\int_{\mathbb{R}\backslash\{0\}}(1\wedge x^2)\nu(\mathsf{d}x)<\infty$ and defined by*

$$\nu(B) := \frac{1}{t}\,\mathsf{E}\Big[\,\sharp\Big\{s\in(0,t] : \Delta X_s(\omega)\subseteq B, B\in\mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)\Big\}\,\Big]$$

is called the Lévy measure of X.

Remark 1.10 *From Definition 1.9, it follows immediately that the compensated random measure is obtained by subtracting its intensity measure, i.e.,*

$$\begin{aligned}\tilde{J}\big(\omega;(0,t]\times B\big) &:= J\big(\omega;(0,t]\times B\big) - \mathsf{E}\big[\,J\big(\omega;(0,t]\times B\big)\,\big]\\ &= J\big(\omega;(0,t]\times B\big) - t\nu(B)\ .\end{aligned}$$

It is noteworthy that although $J\big(\omega;(0,t]\times B\big)$ is a random measure on $\Omega\times\mathbb{R}_+\times\mathbb{R}\backslash\{0\}$, $t\nu(B)$ is a product measure on $\mathbb{R}_+\times\mathbb{R}\backslash\{0\}$. Moreover, $\nu(B)$ is the average number of jumps with sizes contained in B, per unit of time, such that

$$\nu(B) = \mathsf{E}\Big[\,\sharp\Big\{s\in(0,1] : \Delta X_s(\omega)\subseteq B, B\in\mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)\Big\}\,\Big]\ .$$

Let us now return to the jump decomposition (1.3.1) of a pure-jump Lévy process and suppress the dependence of $J\big(\omega;(0,t]\times B\big)$ upon ω for the sake of notational convenience.

From earlier discussions and Remark 1.10, we know that the component of 'large' jumps,

$$X_t^L = \sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{|\Delta X_s|>1\}} = \int_0^t \int_{\{|x|>1\}} xJ(\mathsf{d}s \times \mathsf{d}x) \ ,$$

is a compound Poisson process. Similarly, the component of 'small' jumps now reads

$$X_t^\epsilon = \sum_{0<s\leqslant t} \Delta X_s \mathbb{1}_{\{\epsilon\leqslant|\Delta X_s|\leqslant 1\}} = \int_0^t \int_{\{\epsilon\leqslant|x|\leqslant 1\}} xJ(\mathsf{d}s \times \mathsf{d}x) \ ,$$

which may not be convergent as $\epsilon \searrow 0$. But, as stated earlier, compensating X_t^ϵ guarantees that the integral expression converges as $\epsilon \searrow 0$:

$$\tilde{X}_t^\epsilon = \int_0^t \int_{\{\epsilon\leqslant|x|\leqslant 1\}} x\tilde{J}(\mathsf{d}s \times \mathsf{d}x) = \int_0^t \int_{\{\epsilon\leqslant|x|\leqslant 1\}} x\big[J(\mathsf{d}s \times \mathsf{d}x) - \mathsf{d}s\nu(\mathsf{d}x)\big] \ ,$$

which is a square-integrable martingale converging (in mean-square) to the pure-jump martingale[3]

$$\lim_{\epsilon\searrow 0} \tilde{X}_t^\epsilon = \int_0^t \int_{\{0<|x|\leqslant 1\}} x\tilde{J}(\mathsf{d}s \times \mathsf{d}x) \ .$$

One can also show that this convergence is uniform on $(0, t]$. Finally, this brings about the detailed version of Theorem 1.6,

$$X_t = \gamma t + \sigma W_t + X_t^L + \lim_{\epsilon\searrow 0} \tilde{X}_t^\epsilon \ ,$$

which should also be compared with the semimartingale decomposition in (1.1.10). For an analytically rigorous proof, see Theorem 19.2 of Sato (1999).

As it turns out in Section 2.5 integrals of bounded continuous functions (vanishing at the origin) with respect to the Poisson random measure of Definition 1.8, i.e.,

$$\int_0^t \int_B f(x)J(\omega; \mathsf{d}s \times \mathsf{d}x) = \sum_{0<s\leqslant t} f(\Delta X_s)\mathbb{1}_{\{\Delta X_s\subseteq B\}} \ , \tag{1.3.2}$$

[3]There are two instances where additional assumptions allow for some change of notation. First, when X has finite expectation, i.e., $\int_{\{|x|>1\}} |x|\nu(\mathsf{d}x) < \infty$, the integral expression $\int_0^t \int_{\{|x|>1\}} |x|\, \mathsf{d}s\nu(\mathsf{d}x) < \infty$ exists such that the two jump components can be rewritten as one compensated jump component:

$$\int_0^t \int_{\{|x|>1\}} x\big[J(\mathsf{d}s\times\mathsf{d}x)-\mathsf{d}s\nu(\mathsf{d}x)\big]+\int_0^t \int_{\{0<|x|\leqslant 1\}} x\big[J(\mathsf{d}s\times\mathsf{d}x)-\mathsf{d}s\nu(\mathsf{d}x)\big] = \int_0^t \int_{\mathbb{R}\setminus\{0\}} x\big[J(\mathsf{d}s\times\mathsf{d}x)-\mathsf{d}s\nu(\mathsf{d}x)\big] \ .$$

Second, if $\int_{\{0<|x|\leqslant 1\}} |x|\nu(\mathsf{d}x) < \infty$ holds, then there is obviously no divergence of the sum of 'small' jumps which renders compensation with passing to the limit redundant. For example, this holds for Lévy processes of type B.

will play a central role for constructing particular nonparametric estimators. Therefore, we summarize some results of Theorem 2.3.8 in Applebaum (2004), which will be relevant later on.

Theorem 1.11 (Applebaum (2004)) *Let $J(\omega; t, B)$ be a Poisson random measure and $f : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function that is finite on $B \in \mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)$.*

P1 For any $t > 0$, the integral $\int_0^t \int_B f(x) J(\omega; \mathsf{d}s \times \mathsf{d}x)$ is a compound Poisson process with characteristic function

$$\mathsf{E}\left[\exp\left\{ \mathrm{i}u \int_0^t \int_B f(x) J(\omega; \mathsf{d}s \times \mathsf{d}x) \right\} \right] = \exp\left\{ t \int_B (\mathrm{e}^{\mathrm{i}uf(x)} - 1) \nu(\mathsf{d}x) \right\} ,$$

for all $u \in \mathbb{R}$.

P2 If f is Lebesgue integrable on B, i.e., $f \in \mathcal{L}_1(B)$, then

$$\mathsf{E}\left[\int_0^t \int_B f(x) J(\omega; \mathsf{d}s \times \mathsf{d}x) \right] = t \int_B f(x) \nu(\mathsf{d}x) .$$

As is well-known from statistical analysis, a probability distribution function might be graphically less informative than its probability density function. Thus, we close this section by introducing the main object of our statistical interest: the density of a Lévy measure ν.

Definition 1.12 (Lévy Density) *If the Lévy measure ν is absolutely continuous with respect to (or dominated by) the Lebesgue measure $\mathsf{d}x$, i.e., $\mathsf{d}\nu \ll \mathsf{d}x$, then the Radon-Nikodým derivative*

$$p := \frac{\mathsf{d}\nu}{\mathsf{d}x}$$

exists and is called the Lévy density of ν.

Remark 1.13 *Alternatively, any Lévy density $p : \mathbb{R}\backslash\{0\} \to [0, \infty)$ can be implicitly defined by*

$$\nu(B) = \int_B p(x) \mathsf{d}x ,$$

for any $B \in \mathscr{B}\big(\mathbb{R}\backslash\{0\}\big)$. From the integrability condition of a Lévy measure ν in Theorem 1.3 and Definition 1.9, necessary and sufficient conditions,

$$\int_{\{|x| \geqslant 1\}} \nu(\mathsf{d}x) = \int_{\{|x| \geqslant 1\}} p(x)\, \mathsf{d}x < \infty \qquad \textit{and} \qquad \int_{\{|x|<1\}} x^2 \nu(\mathsf{d}x) = \int_{\{|x|<1\}} x^2\, p(x) \mathsf{d}x < \infty ,$$

for p to be a Lévy density can be deduced. Quantities like $p(x_0)$ provide information on the arrival rates or relative frequency of jumps with size 'close to' x_0.

Although p is a non-negative function, by definition, it does not necessarily integrate to one, i.e.,

$$\int_{\mathbb{R}\setminus\{0\}} p(x)\,\mathsf{d}x \neq 1\ .$$

Despite p not being defined at the origin, i.e., p has zero mass at the origin, we will sometimes replace $\mathbb{R}\setminus\{0\}$ by $\mathbb{R}$ for the sake of simplifying notations.

1.4 Why Pure-Jump Lévy Processes?

1.4.1 An Economic Point of View

In the next chapters, we will deal with estimating pure-jump Lévy processes only. At first sight, this might look as an unduly restrictive assumption prone to misspecification errors but as we will argue in the sequel, it might be quite reasonable to replace the diffusion component by an appropriately chosen Lévy measure.

Recall that the sample paths of purely Brownian driven models are continuous but nowhere differentiable (P-a.s.). In non-technical terms, such a process moves in very small steps and has an extremely vibrant activity. Financial economists developed a reasoning on the evolution of prices in security markets, which is indeed analogous to the heuristics of Robert Brown in the 1820s for describing the movement of grains of pollen in a fluid.

The dynamics of security prices are driven by trades which are the outcomes of decision making processes of traders (Detemple and Murthy, 1994). Assuming that traders make decisions when new information come in and that this information flow is continuous in time, i.e., behaves like a Brownian motion, then security prices should behave like Brownian motions, provided that continuous trading is possible (Duffie and Huang, 1985). One problem with this reasoning is that there may be other motives than new (superior) information on a security that might trigger trades. These motives may be completely unrelated to the security, like the personal preference for liquidity (Glosten and Milgrom, 1985). Because of this, and since the continuity of the information flow (with respect to time), is questionable from a practitioner's point of view, it appears that models for security prices should not be build on the postulate of an continuous information flow, but on (the arrival process of) trades per se.

As already mentioned, the marginal distribution of empirical returns is leptokurtic, i.e., the empirical distribution has more probability mass in a neighborhood of the origin and in the tails than a fitted normal distribution. Furthermore, a pure Brownian motion

(with drift), like the one depicted in Figure 1.1, is unable to reproduce this stylized fact. Although 'fat tails' can be generated by a jump-diffusion Lévy process, like the one depicted in Figure 1.4, this process may still have a hard time modeling the peakedness of the marginal distribution adequately (Geman, Madan, and Yor, 2001). Thus, for the sake of modeling peakedness, there does not seem to be much prospect in simply adding a compound Poisson process to a Brownian motion because, in jump-diffusion Lévy processes, the diffusion component captures frequent but small moves of asset prices, while the jump component captures rare but large changes. In sum, a more sophisticated device is needed.

If we dispense the diffusion component completely, then the returns correspond to (sums of) jumps of the postulated process. In this setup, peakedness can be interpreted as a high arrival rate of very small jumps. Indeed, one can argue that a pure-jump Lévy process with a 'high' rate of activity may render the diffusion component unnecessary. Roughly speaking, diffusions and jump processes with infinite activity from Section 1.2 can be considered as substitutes from a modeling perspective. To see this, recall that the Lévy measure of the compound Poisson process of Figure 1.2 is $\nu = \lambda F$, where F is a normal distribution, such that

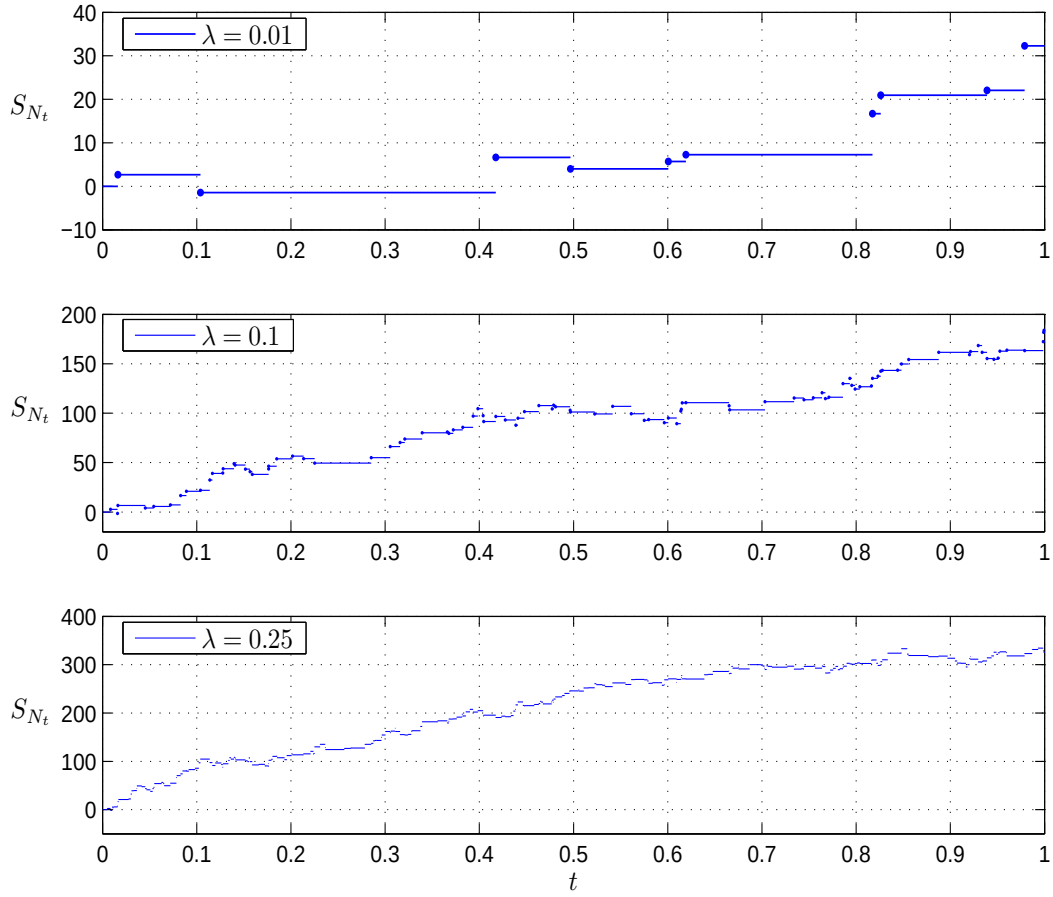
$$\nu\big(\mathbb{R}\backslash\{0\}\big) = \lambda F\big(\mathbb{R}\backslash\{0\}\big) = \lambda \, .$$

Then, $\nu\big(\mathbb{R}\backslash\{0\}\big) \to \infty$ as $\lambda \to \infty$. Figure 1.5 illustrates the effect of increasing the jump intensity of a compound Poisson process (1.1.4). Obviously, the sample paths are approaching those of a continuous Brownian motion. On the contrary, both peakedness and fat tails can easily be generated by choosing a general Lévy measure ν, which has a certain amount of flexibility.

Returning to the above practitioner's point of view, focussing of trades instead of information provides another reason which might speak in favor of pure-jump processes (with 'high' activity) and against a continuous component. An important assumption for deriving the Black-Merton-Scholes formula is the requirement of continuous trading in the underlying (Hull, 2000, p. 245). Even in age of high-frequency trading, this assumption might be questionable. Indeed, real securities are not only traded in discrete time points, but the inter-trade durations are stochastic which led to the development of a new branch in empirical market microstructure research (Engle and Russell, 1998).

Finally, Geman (2002) reckoned that processes with finite variation should be better models for real-world financial time series than processes with infinite variation. This looks reasonable when visually inspecting the observed trajectories of financial securities. If one is to accept this supposition, then we know from Definition 1.7 that the class of

Figure 1.5: *Effects of Increasing Intensity of Compound Poisson Process*
This figure depicts the effects of increasing the intensity λ of the compound Poisson process (1.1.4) in Figure 1.2. The sampling interval is scaled down to $[0, 1]$.
The first effect is that the activity increases, converging towards infinity activity. The second effect, due to $\mu_X \neq 0$, is that the drifting behavior of the sample path becomes more pronounced, converging towards the sample path of a Brownian motion with drift (1.1.3), although the sample path for high λ appears to be less irregular than in Figure 1.1. Hence, infinite variation may also be approached as $\lambda \to \infty$. In sum, this illustrates the potential of the Lévy measure ν to generate the infinite activity and infinite variation of Definition 1.7.



candidate models has to satisfy $\sigma = 0$ and $\int_{\{0<|x|\leqslant 1\}} |x|\nu(\mathrm{d}x) < \infty$. Thus, a diffusion component is ruled out, and we are considering pure-jump Lévy models of type B.

1.4.2 A Statistical Point of View

Besides reasons originating from economic theory and practice stated in Subsection 1.4.1, there is another very fundamental cause why it might be advantageous to refrain from including a continuous Brownian component. This is due to the problem of estimating continuous-time models based on discrete-time observations.

Let us first look at a classical case, and assume that we have a discrete-time stochastic process, say a pth-order autoregressive model, and a sample $\{X_t : t = 1, 2, \ldots, T\}$ of discrete-time observations which we want to use for estimating model parameters. In the next step, a parametric estimation method is selected based on statistical optimality properties. Usually, these properties can only be established by resorting to asymptotic theory. To be more precise, in order to derive consistency, central limit theorems, and asymptotic efficiency, we need so-called **long-span asymptotics**, i.e., $T \to \infty$. In other words, if the true data-generating process is a discrete-time process, then long-span asymptotics eventually renders the true process perfectly observable.

This situation changes dramatically when the true data-generating process is continuous-time. Let us assume that the true data-generating process is a jump-diffusion Lévy process with sample paths similar to the one depicted in Figure 1.4. Moreover, assume that it is only possible to sample the observations of X on a discrete skeleton

$$0 = t_0 < t_1 < \cdots < t_n = T$$

of $[0, T]$ where all n grid points are equally spaced, i.e.,

$$\Delta_n := t_i - t_{i-1} = T/n \ ,$$

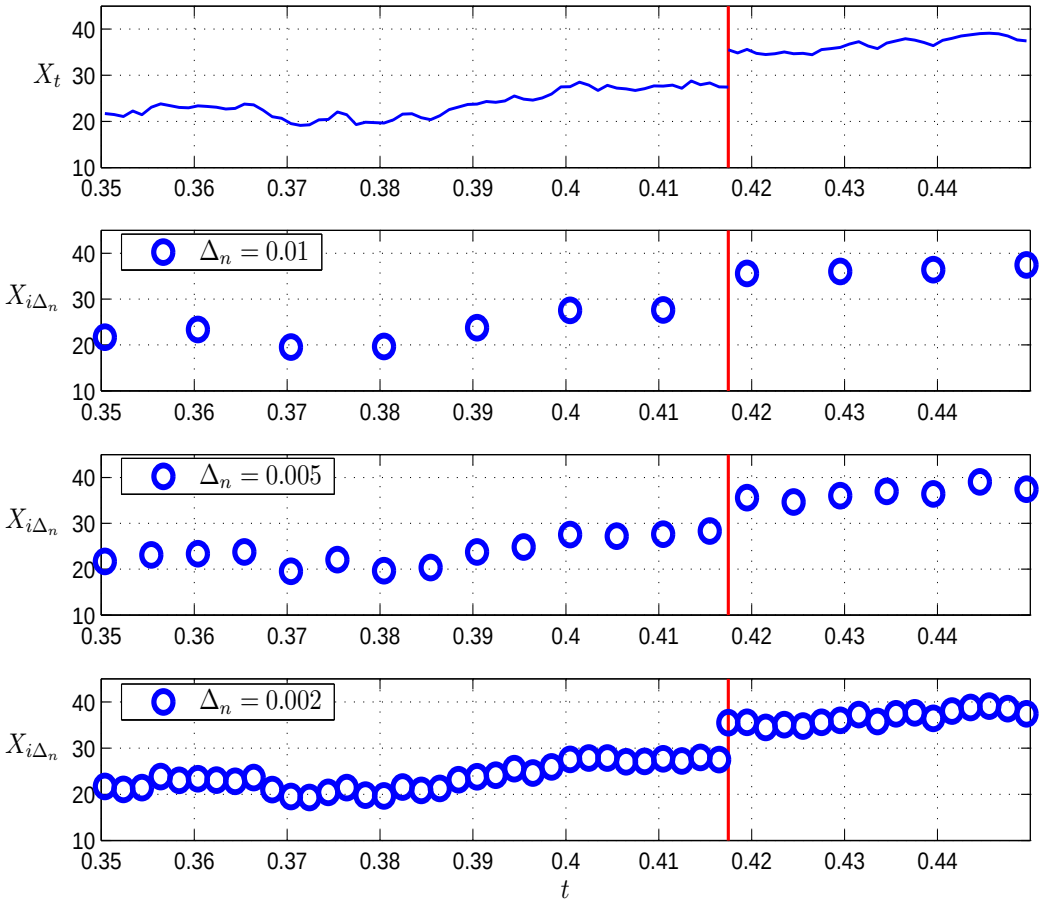
for all $i = 1, \ldots, n$. In contrast to the classical situation, long-span asymptotics is not sufficient for completely uncovering all the characteristics of the whole sample path. For example, it is impossible to statistically identify the jumps, which correspond to continuous-time increments ΔX_t of X, even as $T \to \infty$. Instead, we are merely able to observe the so-called **discrete-time increments**, i.e.,

$$\Delta X_i := X_{t_i} - X_{t_{i-1}} \ ,$$

for all $i = 1, \ldots, n$. For a jump-diffusion Lévy process, discrete-time sampling yields increments which are a mixture of changes due to both continuous and the jump component. For instance, Neumann and Reiß (2009) estimated the characteristic triplet of Definition 1.3 nonparametrically based on discrete-time observations by invoking the inverse Fourier transform of the empirical characteristic function, but they were not able to uniquely identify σ and ν. Of course, if there are good reasons to dispense σ, we can get rid of this identification problem between σ and ν. This is another reason for focussing the pure-jump Lévy processes. Figure 1.5 illustrates how identification of jump locations and sizes might fail for $\Delta_n > 0$.

More generally, disentangling the continuous component from the jump component is

Figure 1.6: *Effect of Increasing Sampling Frequency for Jump-Diffusion Process*
This figure depicts the effect of increased sampling of a discretely observed continuous-time process. To this end, the sample path of the jump-diffusion Lévy process of Figure 1.4 is zoomed in on the interval $[0.35, 0.45]$ which contains exactly one jump. The time and size of the jump can only be reliably identified when the sampling frequency $1/\Delta_n$ is sufficiently high.



only possible, if $\Delta_n \to 0$. This type of asymptotics is called **in-fill asymptotics** as it forces the sampling process to produce a skeleton which is dense on $[0, T]$, in the limit. If, at the same time, $T \to \infty$, then we are back in the classical asymptotic setting. Thus, if $\Delta_n \to 0$ and $T \to \infty$ holds simultaneously, then $n \to \infty$ at a faster rate than $T = t_n$, i.e., $t_n = o(n)$. Roughly speaking, while long-span asymptotics helps us to infer long-run behavior, like a trend, of an unknown continuous-time process, in-fill asymptotics helps us to infer its local dynamics, like diffusion and/or jump behavior (Florens-Zmirou, 1993).

Since in-fill asymptotics requires the sampling frequency n to approach infinity, it might appear to be an obscure, statistical notion to a practitioner. However, it is just as obscure as requiring an infinite sample of observations when $T \to \infty$. Hence, accepting the *asymptopia* of $T \to \infty$ is just as disturbing as accepting the *asymptopia* of $n \to \infty$.

Traditionally, analysis of estimators for continuous-time models were based on continuous-time sampling. See the monograph of Kutoyants (2004) for a historical review. Recently, econometricians became aware of the problem when sampling is in discrete time (Melino, 1994). For example, Lo (1988) showed that parametric maximum likelihood might not yield consistent estimators for a diffusion model when there is no in-fill asymptotics.

One might then ask whether the results obtained for continuous-time sampling do have any meaning nowadays? The answer is on the affirmative because the results obtained from continuous-time sampling are used as a benchmark to which the results obtained from discrete-time sampling should converge as $n \to \infty$. To be more specific, what is usually done is to derive, say, optimality results of a proposed estimation method for continuous-sampling. In a second step, it is shown that the discrete-time version of the proposed estimation method converges to its continuous-time counterpart. This two-step procedure is usually easier to handle than to establish long-span and in-fill asymptotics simultaneously.

Finally, it is noteworthy that we can relax the assumption of equally spaced observations. It poses no problem to consider irregularly spaced observations as long as we define

$$\Delta_n := \max_{1 \leqslant i \leqslant n} (t_i - t_{i-1}) \, .$$

Parametric estimation of continuous-time processes based on irregularly spaced observations was analyzed by Duffie and Glynn (2004) and Aït-Sahalia and Mykland (2004, 2008), for example.

1.4.3 Subordination & Random Time Change

As already mentioned in Subsection 1.4.1, the information flow is an incomplete proxy for explaining the driving force of asset prices. This led Clark (1973) to propose cumulated trading volume as a more adequate driving force, and which was shown by Ané and Geman (2000) to be an empirically reasonable choice. Moreover, we made the point in Subsection 1.4.1 that trading happens in discrete time rather than in continuous time. This implies that a mathematical representation of trading activity Y_t boils down to a (càdlàg) piecewise constant function with jumps located at times of incoming trades. For continuous trading, the corresponding mathematical representation is simply the identity mapping $Y_t = t$ defined on $\mathbb{R}_+$. Applying this idea to modeling some asset price X_t, we say that

$$X_t := Z_{Y_t} \tag{1.4.1}$$

is a **time-changed process**, where Z_t and Y_t are stochastic processes. Evidently, the driving process Y_t has to satisfy some extra condition in order to lend itself as a substitute for time t. A first minimal requirement is that Y_t should be positive and non-decreasing. Additionally, we require that Y_t increases with random step sizes in order to model varying trading intensity. Consequently, this allows us to get a handle on the fact that the trading process is not homogeneous in time by discriminating between calender time t and business time Y_t.

These considerations can be put into a concise notion called a **subordinator**. The following definition is due to Sato (1999, Definition 21.4), but the notion was first introduced by Bochner (1955). A rigorous exposition of subordination can be found in Bertoin (1996, Chapter III).

Definition 1.14 (Subordinator) *A Lévy process $Y = \{Y_t\}_{t\geqslant 0}$ is a subordinator, if the sample paths are non-decreasing (P-a.s.).*

Remark 1.15 *Definition 1.14 rules out subordinators with a diffusion component. For example, the Brownian motion with drift* (1.1.3) *of Figure 1.1 can only be non-decreasing for $\sigma = 0$. In this case, it would be a purely deterministic process not being able to model random market activity. Hence, the subordinators we consider as sensible are pure-jump Lévy processes.*

Sato (1999, Theorem 21.5) showed that Definition 1.14 implies the following restrictions on the characteristic triplet of a Lévy process:

$$\begin{aligned} \sigma &= 0 \\ \int_{\{-\infty<x<0\}} \nu(\mathsf{d}x) &= 0 \\ \int_{\{0<x\leqslant 1\}} x\nu(\mathsf{d}x) &< \infty \\ \gamma &\geqslant 0 \,. \end{aligned}$$

For $\mathbb{R}_+$-valued Lévy processes, it is more convenient to use the Laplace transform rather than the Fourier transform in order to derive its characteristic function (see Appendix A.1). In particular, if Y is a subordinator with characteristic triplet $(\gamma, 0, \nu)$, then the Laplace exponent $\Psi^+(z)$ of the Laplace transform $\mathsf{E}\big[\,\mathsf{e}^{-zY_t}\,\big] = \mathsf{e}^{-t\Psi_Y^+(z)}$, for all $z \geqslant 0$, has the general form

$$\Psi_Y^+(z) = z\gamma + \int_0^\infty (1 - \mathsf{e}^{-zx})\nu(\mathsf{d}x) \,,$$

see, for example, Kallenberg (2002, Corollary 15.8). Then, assuming analytical contin-

uation and using the relationship $\Psi_Y(u) = -\Psi_Y^+(-\mathsf{i}u)$ derived from (A.1.6), we end up with the characteristic exponent

$$\Psi_Y(u) = \mathsf{i}u\gamma + \int_0^\infty (\mathsf{e}^{\mathsf{i}ux} - 1)\nu(\mathsf{d}x) \ ,$$

where $\gamma \geqslant 0$, $\nu\big((-\infty, 0)\big) = 0$, and $\int_{(0,\infty)} x\nu(\mathsf{d}x) < \infty$.

The next theorem summarizes some important results of Section 1.3.2 of Applebaum (2004) and Theorem 30.1 of Sato (1999).

Theorem 1.16 (Properties of Random Time-Changed Processes) *Let $X_t = Z_{Y_t}$ be a random time-change process, where Z_t is a Lévy process and Y_t is a subordinator with characteristic triplet $(\gamma, 0, \nu)$ satisfying $\gamma \geqslant 0$, $\nu\big((-\infty, 0)\big) = 0$, and $\int_{(0,\infty)} x\nu(\mathsf{d}x) < \infty$. Then, X_t is a Lévy process with characteristic exponent*

$$\Psi_X = -\Psi_Y^+ \circ (-\Psi_Z) \ ,$$

where $\Psi_Y^+(z) = -\Psi_Y(\mathsf{i}z)$ is the Laplace exponent of Y and $\Psi_Z(u)$ is the characteristic exponent of Z.

In Section 4.1, a simple random time-changed process, called the **variance gamma process**, is presented which is of type B in Definition 1.7. Another popular random time-changed process is the **normal inverse Gaussian** process (Barndorff-Nielsen, 1997, 1998) which is of type C in Definition 1.7. The class of models, nesting—among others—the variance gamma and normal inverse Gaussian process, is the **generalized hyperbolic motion** (Eberlein, 2001) which is an extension of the **hyperbolic motion** (Barndorff-Nielsen, 1978). See also Eberlein and Keller (1995) and Eberlein, Keller, and Prause (1998). Finally, note that any semimartingale can be cast in the form (1.4.1), see Monroe (1978), which emphasizes the importance of subordination for arbitrage theory.

Chapter 2

Method of Sieves

2.1 Minimax Optimality & Adaptation

This section presents a very general introduction to nonparametric estimation theory which applies to both the current and the next chapter. Much of this content is covered in the monographs of Tsybakov (2009) and Györfi, Kohler, Krzyżak, and Walk (2002), and in particular, the seminal paper of Barron, Birgé, and Massart (1999).

The goal of this theory is not to construct nonparametric estimators in the first place, but to evaluate their performance. Put differently, the subject revolves around the question of optimality. In the nonparametric context, optimality of an estimator is related to minimax results that have been established for certain estimation problems. This program essentially consists of three basic ingredients which will sequentially be motivated and introduced.

Let us first assume that we have constructed an estimator $\hat{f}_n$ for a nonparametric estimation problem at hand. The next step is to analyze how well $\hat{f}_n$ performs relative to some other estimator (which, for example, is known to be the best one). Given a loss function $\ell(\,\cdot\,,\,\cdot\,)$, the **risk** of $\hat{f}_n$ at f is given by

$$R_n(\hat{f}, f) := \mathsf{E}_f\Big[\, \ell(\hat{f}_n, f) \,\Big] \;. \tag{2.1.1}$$

This is the first ingredient. Unfortunately, it is not admissible to consider any arbitrary f, since Farrell (1967) showed that, for any estimator $\hat{f}$, there exists some function f such that

$$\sup_{f \in \mathcal{F}} R_n(\hat{f}, f) \not\to 0 \qquad \text{as } n \to \infty,$$

where the supremum is taken over the class $\mathcal{F}$ of all functions. Consequently, we have to restrict the function class $\mathcal{F}$ to satisfy a certain smoothness condition. Usually, for kernel estimators, $\mathcal{F}$ is a subset (or ball) of a Hölder class, while for orthogonal projection estimators, $\mathcal{F}$ is a subset (or ball) of a Sobolev class. This is the second ingredient. Hence, the **maximal risk** of $\hat{f}_n$ over a function class $\mathcal{F}^s$, with smoothness parameter s, is given by

$$R_n(\hat{f}, \mathcal{F}^s) := \sup_{f \in \mathcal{F}^s} R_n(\hat{f}, f) = \sup_{f \in \mathcal{F}^s} \mathsf{E}_f\Big[\, \ell(\hat{f}_n, f) \,\Big] \;.$$

Finally, the **minimax risk** over $\mathcal{F}^s$ is given by

$$R_n(\mathcal{F}^s) := \inf_{\tilde{f}} R_n(\tilde{f}, \mathcal{F}^s) = \inf_{\tilde{f}} \sup_{f \in \mathcal{F}^s} R_n(\tilde{f}, f) = \inf_{\tilde{f}} \sup_{f \in \mathcal{F}^s} \mathsf{E}_f\Big[\, \ell(\tilde{f}_n, f) \,\Big] \;,$$

where the infimum is taken over the set of all estimators $\tilde{f}$.

What remains to be shown is whether the maximal risk of our estimator $\hat{f}$ corresponds to the minimax risk, i.e.,

$$R_n(\hat{f}, \mathcal{F}^s) = R_n(\mathcal{F}^s) \ ,$$

or is bounded by the minimax risk, i.e.,

$$R_n(\hat{f}, \mathcal{F}^s) \leqslant C(s) R_n(\mathcal{F}^s) \quad \Longleftrightarrow \quad \frac{R_n(\hat{f}, \mathcal{F}^s)}{R_n(\mathcal{F}^s)} \leqslant C(s) \ , \tag{2.1.2}$$

where the absolute constant should satisfy $C(s) \searrow 0$. If this holds, then $\hat{f}$ is said to be **minimax**. Sometimes minimax properties are hard or impossible to establish. Then, we have to resort to an asymptotic point of view and to show that $\hat{f}$ is **asymptotically minimax**, i.e.,

$$R_n(\hat{f}, \mathcal{F}^s) - R_n(\mathcal{F}^s) = o_P(1) \ ,$$

or that $\hat{f}$ attains the **optimal rate of convergence**, i.e.,

$$\frac{R_n(\hat{f}, \mathcal{F}^s)}{R_n(\mathcal{F}^s)} = O_P(1) \ .$$

This is the third ingredient.

Nowadays, many minimax results for various choices of loss functions and function classes can be found in the literature on nonparametric estimation theory. For example, Stone (1982) showed that, for the risk function based on the $\mathcal{L}_2$-loss,

$$R_n(\hat{f}, f) := \mathsf{E}_f\Big[\, \ell(\hat{f}_n, f) \,\Big] = \mathsf{E}_f\Big[\, ||\hat{f}_n - f||^2 \,\Big] = \mathsf{E}_f\Bigg[\int \big(\hat{f}_n(x) - f(x)\big)^2 \mathsf{d}x \Bigg] \ ,$$

and $f \in \mathcal{C}^m$, i.e., the space of all m-times continuously differentiable functions, with $m \in \mathbb{N}$, the best rate r attainable, such that

$$R_n(\hat{f}, \mathcal{C}^m) = \sup_{f \in \mathcal{C}^m} R_n(\hat{f}, f) \leqslant C(m) n^{-r}$$

holds, is

$$r = \frac{2m}{2m+1} \tag{2.1.3}$$

when considering the problem of nonparametric regression. For nonparametric regressions, the same rate for different settings were proved by Ibragimov and Hasminskii (1982), Nussbaum (1985), Speckman (1985), van de Geer (1990).

In this respect, we will shortly turn to the estimation problem of van de Geer (1990), which will be discussed in more detail in Section 2.2, in order to point out a shortcoming of

the minimax approach. Assume that we known the unknown target function $f : [0,1] \to \mathbb{R}$ belongs to some ball

$$\mathcal{F}^m(M) := \left\{ f \in \mathcal{L}_2[0,1] : \int \left| f^{(m)}(x) \right|^2 \mathsf{d}x \leqslant M^2, m \in \mathbb{N}, M > 0 \right\}$$

of the Sobolev space $\mathcal{W}_2^m[0,1]$. As will be explained later, the estimation method—like any another nonparametric procedure—depends on the choice of a tuning parameter, say λ. In order to attain the minimax rate in (2.1.3), it is necessary to optimally chose λ. Unfortunately, this optimal choice depends on the *a priori* knowledge of m and M. But if we know them *a priori*, then it can be shown that the risk can be uniformly bounded over the Sobolev ball $\mathcal{F}^m(M)$ by $\kappa M^{2/(2m+1)} n^{-2m/(2m+1)}$, where κ, which corresponds to the minimax rate (2.1.3). A similar situation arises, for example, in nonparametric kernel density estimation (Silverman, 1986).

This situation is a bit annoying since the estimator $\hat{f}$, though minimax, loses some of its nonparametric flavor, and brings up the issue known in the nonparametrics literature as **adaptation in the minimax sense**. Roughly speaking, adaptation in the minimax sense can be described as attaining the minimax rate without *a priori* knowledge of the particular function class (or ball) where f resides. In the above example, this boils down to the situation, where $\hat{f}$ attains the minimax rate without knowing the smoothness m and radius M of the ball.

The most common form of adaptation in the minimax sense assumes that we know that f belongs to one of the function spaces contained in the collection (or scale) $\{\mathcal{F}^s\}_{s \in S}$, and the goal is to find an estimator $\hat{f}$ which minimizes the ratio

$$\frac{R_n(\hat{f}, \mathcal{F}^s)}{R_n(\mathcal{F}^s)} = C_n(s) \geqslant 1 \ ,$$

over the whole scale of $\{\mathcal{F}^s\}_{s \in S}$. Efromovich and Pinsker (1984) considered the case of **exact asymptotic adaptation** where

$$\operatorname*{limsup}_{n \to \infty} C_n(s) = 1 \ .$$

Lepskii (1992) considered the case of **asymptotic adaptation** where

$$\operatorname*{limsup}_{n \to \infty} C_n(s) = C(s) \ .$$

Donoho and Johnstone (1995) considered the case of **asymptotic adaptation up to** (a

slowly varying function of n) $\boldsymbol{L_n}$ where

$$\operatorname*{\mathsf{limsup}}_{n\to\infty} \frac{C_n(s)}{L_n} = C(s) \ .$$

Finally, Barron (1994) and Birgé and Massart (1997) considered the cases of **nonasymptotic adaptation** where

$$C_n(s) \leqslant L_n C(s) \qquad \text{for } L_n \equiv 1,$$

and **nonasymptotic adaptation up to $\boldsymbol{L_n}$**, where L_n is a slowly varying function of n. They also showed that, in some cases, it is possible to obtain $C(s) \equiv C$.

Typically, L_n corresponds to a power of $\mathsf{ln}(n)$ and is the price we have to pay for adaptation in the minimax sense when $\{\mathcal{F}^s\}_{s\in S}$ is a scale of very flexible function classes. For instance, this turns out to be the case, when we move from Sobolev spaces to Besov spaces. However, this is a price one often willing to pay in applications. Moreover, it is noteworthy that nonasymptotic adaptation offers the tremendous advantage since it holds for any sample size, not necessarily for $n \to \infty$ only, such that it could be taken as an indication for better performance in small samples.

We close this section by mentioning a device which has turned out to be extremely useful for deriving minimax rates. In fact, much of the theoretical work on wavelets, to be discussed in Chapter 3, relies on this approach. It comes as a surprise that much of the minimax theory of nonparametric estimation can be worked out in the so-called **Gaussian white noise model** (with drift)

$$\mathsf{d}Y(t) = f(t)\,\mathsf{d}t + \epsilon\,\mathsf{d}W(t) \ ,$$

where $t \in [0,1]$, $f : [0,1] \to \mathbb{R}$, $0 < \epsilon < 1$, and W is a Wiener process on $[0,1]$. Then, the statistical task is to estimate the unknown function f, which is known to belong to a function class $\mathcal{F}$, from the noisy observations $Y(t)$. It can be shown that the Gaussian white noise model is asymptotically equivalent to the **Gaussian nonparametric regression model**,

$$Y_i = f(i/n) + \xi_i \ ,$$

and the **Gaussian sequence model**,

$$Y_i = \theta_i + \epsilon\xi_i \ ,$$

where $\xi_i \overset{\mathsf{iid}}{=} \mathsf{N}(\,0,1\,)$ for all $i = 1,\ldots,n$. Put differently, the estimation problems are

asymptotically identical. This result, which is closely related to Le Cam (1986)'s notion of equivalence of experiments, has important implications for nonparametric theory.

The major insight is that the risk functions of these models are asymptotically equivalent which renders it possible to transfer results in optimal convergence rates, minimax procedures, etc. from one model to another. Thus, we can take the simplest model, derive the desired results, and transfer them to the other models. From the perspective of nonparametric estimation theory, the simplest model is usually the Gaussian white noise model, whose first exact risk bound was derived by Pinsker (1980). Later on, Brown and Low (1996) established the first equivalence result for the Gaussian nonparametric regression model. Nussbaum (1996) showed asymptotic equivalence between the Gaussian white noise model and the problem of nonparametric density estimation. More recently, Brown, Carter, Low, and Zhang (2004) showed the asymptotic equivalence of the Gaussian white noise model and the nonparametric estimation of the intensity function of a compound Poisson process. The latter seems to be of some relevance for the nonparametric estimation of a Lévy density, which should be elaborated in future research.

We have mentioned these results as the Gaussian sequence model plays a central role in wavelet theory of Chapter 3. In particular, if $f \in \mathcal{L}_2[0,1]$ in the Gaussian white noise model, then the estimation of f is equivalent to the estimation of the Fourier coefficients θ_i in the Gaussian sequence model. Obviously, this is closely related to the isomorphism between the Lebesgue space $\mathcal{L}_2$ and the sequence space ℓ_2. This isomorphism paves the way for thresholding techniques which are variants of the James and Stein (1961) shrinking procedure giving wavelet-based estimation an edge over many standard nonparametric estimation methods. For more on these issues, we refer to the online manuscript of Johnstone (2011).

2.2 Nonparametric Estimation via Sieves

In applied econometrics, probably the most popular nonparametric estimators are **local estimators** based on kernel methods because of their intuitive appeal and ease of implementation. See, for example, the monographs of Härdle (1992), Li and Racine (2006), or Pagan and Ullah (1999). For the problem of estimating a probability density function nonparametrically, the corresponding kernel estimator was proposed by Rosenblatt (1956) and Parzen (1962). For the problem of estimating a regression function nonparametrically, the corresponding kernel estimator was proposed by Nadaraya (1964) and Watson (1964). As standard kernel estimators incur bias terms at the boundary of the support and the design points (Wasserman, 2006), a generalization, which nests standard

kernel estimators in a natural way, was put forward which is called the local polynomial estimator (Fan and Gijbels, 1996).

Generally speaking, all of these local estimators can be characterized as procedures which estimate an unknown function f around a fixed point x_0 in the support of f by using data contained in a local neighborhood of x_0. The crucial tuning (or smoothing) parameter in these local procedures, which has to be chosen optimally, determines the 'width' of these neighborhoods. A nonparametric estimator $\hat{f}$ for the whole of f is then obtained by repeating this procedure or a sufficiently large number of points in the support of f.

An alternative approach to estimating the whole shape of f is **global approximation**. The relationship between local and global approximation is similar to the relationship between Taylor series expansions and the Stone-Weierstrass Theorem. See the monograph of Christensen and Christensen (2004) for a comparison of these types of approximation. In the sequel, we will take on the local approach to nonparametric estimation, which offers two major advantages: First, since the global approach is based on the idea of f being an element in certain function spaces, it allows us to use the well-developed machinery of approximation theory (DeVore and Lorentz, 1993; Lorentz, Golitschek, and Makovoz, 1996) for deriving and characterizing the theoretical properties of the corresponding estimator $\hat{f}$. Second, the global approach presents a unifying framework which nests many optimization-based estimation procedures.

Let us now introduce this unifying framework which essentially corresponds to the classical M-estimator of Huber (1967). Let θ be an element of the parameter space Θ and $X_1, \ldots, , X_n$ iid random variables. A function $\gamma : \mathcal{X} \times \Theta \to \mathbb{R}$ is a **contrast function** if, for all $\theta_0 \in \Theta$,

$$\inf_{\theta \in \Theta} \mathsf{E}_{\theta_0}[\,\gamma(X;\theta)\,] = \mathsf{E}_{\theta_0}[\,\gamma(X;\theta_0)\,] =: P\gamma(X;\theta_0) \;,$$

θ_0 is the minimizer of the expectation (under the true model) of γ over Θ. However, as θ_0 is unknown, the expectation of γ cannot be computed and has to be substituted by its empirical counterpart P_n. The corresponding **empirical contrast** is defined by

$$\gamma_n(\theta) := P_n\gamma(X;\theta) := \frac{1}{n}\sum_{i=1}^{n} \gamma(X_i;\theta) \;,$$

where the X_i's are iid copies of X. Finally, the **minimum contrast estimator**,

$$\hat{\theta}_n := \operatorname*{arg\,min}_{\theta \in \Theta} \gamma_n(\theta) \;, \tag{2.2.1}$$

is defined as the minimizer of the empirical contrast. Of course, even if we assume that

$\hat{\theta}_n$ exists, it is not necessarily unique. If the existence of $\hat{\theta}_n$ cannot be guaranteed for some $n \in \mathbb{N}$, the so-called ϵ_n**-minimum contrast estimator** may be defined to satisfy

$$\gamma_n(\hat{\theta}_n) \leqslant \inf_{\theta \in \Theta} \gamma_n(\theta) + O_P(\epsilon_n) \,, \tag{2.2.2}$$

where $\epsilon_n \to 0$ as $n \to \infty$ (Birgé and Massart, 1993). If $\epsilon_n \equiv 0$ for all $n \in \mathbb{N}$, then $\hat{\theta}_n$ corresponds to the exact minimum contrast estimator (2.2.1).

Obviously, by choosing the appropriate (empirical) contrast function, the setup (2.2.1) collapses to a special parametric estimator such as the maximum likelihood estimator of Fisher (1912, 1921, 1922), the generalized method of moments estimator of Hansen (1982), the estimating equations approach of Godambe (1960), the quasi maximum likelihood estimator of Wedderburn (1974), the minimum distance estimator of Wolfowitz (1957), or the minimum Hellinger distance estimator of Beran (1977).

Although the analysis of these parametric estimators belongs to the statistical repertoire, severe problems may show up when the parameter of interest in (2.2.1) has infinite dimension, i.e., when θ is a function, which happens to be the case in a nonparametric context. For example, Neyman and Scott (1948) showed that maximum likelihood may fail to be consistent in infinite-dimensional parameter spaces. See also Bahadur (1958), Le Cam (1990), and Kiefer and Wolfowitz (1956). Less but still annoying are the results of Birgé and Massart (1993) or van de Geer (2000, Chapter 10) which showed that, even if consistency can be established, the estimators may exhibit slow (or suboptimal) rates of convergence, i.e., statistical inefficiency.

These pathological cases share a common ground. To further illuminate this issue, let us turn to the literature on maximum likelihood and least-squares estimation. One of the most fundamental results in mathematical statistics is due to Wald (1949) who showed that under suitable regularity conditions, including compactness of Θ and integrability conditions on $\gamma_n(\,\cdot\,;\theta)$, parametric maximum likelihood is consistent. Bahadur (1967) analyzed the consistency of maximum likelihood estimators in more general (non-Euclidean) compact parameter spaces. van de Geer (1990) analyzed the convergence rate of nonparametric least-squares while Wong and Severini (1991) considered nonparametric maximum likelihood. Both of these papers share a common reasoning that replaces explicit compactness of Θ by the notion of metric entropy of Θ. See Definitions 2.15 and 2.16 in the Sub-appendix 2.A for details. This is essentially because compactness in infinite-dimensional spaces is much harder to determine and characterize than in finite-dimensional, e.g. Euclidean, ones.

The basic motivation for this approach was not new and is essentially based on a

stochastic version of the Arzelà-Acoli Theorem (Dudley, 2002, Theorem 2.4.7) which provides a characterization of compactness in function spaces. A substantial improvement of this characterization was accomplished by the seminal work of Kolmogorov and Tihomirov (1961) who introduced the notion of metric entropy as a device to measure the complexity of general metric spaces, and computed metric entropies of many classical function spaces.

In mathematical statistics, the idea to relate the convergence rate of an estimator to the metric entropy of the underlying parameter was introduced by Le Cam (1973). See also Le Cam (1997). for example, much of modern statistical learning theory was founded on the notion of the VC-dimension of abstract sets introduced by Vapnik and Červonenkis (1971). Likewise, the extensive theory of nonparametric density estimation based on the $\mathcal{L}_1$-loss (Devroye and Lugosi, 2001) is rooted in the entropy considerations of Yatracos (1985).

Progress in these and many other fields were driven by new developments in the realm of empirical process theory. In particular, elaborating on the role of the order of magnitude of the increments of an empirical process indexed by a function class $\mathcal{F}$ led to new insights into stochastic limit theorems in function spaces. For more details, see the monographs of Dudley (1999), Kosorok (2008), van de Geer (2000), or van der Vaart and Wellner (1996).

Remark 2.1 *In order to illustrate these ideas, let us consider the nonparametric regression model of Section 2.1, where the aim is to estimate an unknown function* $f_0 : [0,1] \to \mathbb{R}$ *from noisy observations* Y_i *generated by the model*

$$Y_i = f_0(x_i) + \sigma\epsilon_i$$

with regular sampling at fixed points $x_i = i/n$ *and noise* $\epsilon_i \overset{\text{iid}}{=} \mathsf{N}(0,1)$ *for all* $i = 1, \ldots, n$. *The only a priori assumption that we use is that* f_0 *belongs to a class of all functions having a fixed number of derivatives* $f^{(m)}$, *i.e., the Sobolev space* $\mathcal{W}_2^m$ *with known* m*:*

$$\mathcal{F} := \left\{ f : [0,1] \to \mathbb{R} \,:\, \int \left[f^{(m)}(x)\right]^2 \mathsf{d}x \leqslant C, m \in \mathbb{N} \right\} .$$

Defining the empirical contrast function

$$\gamma_n(f) := \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - f(x_i)\right]^2$$

based on the (empirical) $\mathcal{L}_2$*-(semi-)norm*

$$||f|| := \left[\frac{1}{n}\sum_{i=1}^{n}\left[f(x_i)\right]^2\right]^{1/2} ,$$

yields the nonparametric least-squares estimator $\hat{f}_n$ *satisfying*

$$\gamma_n(\hat{f}_n) = \min_{f\in\mathcal{F}} \gamma_n(f) .$$

Assuming that $\gamma_n(\,\cdot\,)$ *satisfies an integrability condition, we can define the increments of an empirical process indexed by* $\mathcal{F}$*:*

$$\mathbb{G}_n(f - f_0) := \sqrt{n}\big\{\gamma_n(f_0) - \mathsf{E}[\,\gamma_n(f_0)\,]\,\big\} - \sqrt{n}\big\{\gamma_n(f) - \mathsf{E}[\,\gamma_n(f)\,]\,\big\} .$$

From the above property of the nonparametric least-squares estimator $\hat{f}_n$*, we can bound its rate of convergence by the increments of the empirical process indexed by* $\hat{f}_n$*:*

$$\sqrt{n}||\hat{f}_n - f_0||^2 \leqslant \mathbb{G}_n(\hat{f}_n - f_0) .$$

Using the metric entropy result in Lemma 6.1 of van de Geer (1990), the order of magnitude of $\left|\mathbb{G}_n(f - f_0)\right|$ *can be stochastically bounded, i.e.,*

$$\left|\mathbb{G}_n(f - f_0)\right| = O_P\left(||f - f_0||^{1-1/(2m)}\right) ,$$

uniformly for all $f \in \mathcal{F}$*. Since the last equality holds uniformly, it also holds when replacing* f *by* $\hat{f}_n$*. Then, by combining these results we obtain*

$$||\hat{f}_n - f_0||^2 = O_P\left(n^{-2m/(2m+1)}\right) ,$$

which corresponds to the minimax rate of convergence (2.1.3) *for nonparametric regression problems.*

These ideas were generalized by Birgé and Massart (1993) who extended this approach to general minimum contrast estimation and analyzed minimax adaptivity based upon the results of Birgé (1983). Of course, the price, they had to pay for increased generality, is a set of more stringent assumptions (for example, a known bound on f_0) than in van de Geer (1990).

In sum, the metric entropy determines the convergence rate of $\hat{f}_n$ via the oscillating behavior of $\mathbb{G}_n(\hat{f}_n - f_0)$. If the metric entropy is too large, then the sample paths of the

empirical process becomes too irregular such that no convergence to a (continuous) limit process may be guaranteed and consistency cannot be established. See Giné and Zinn (1984) and the monograph of Pollard (1984). In the econometrics literature, this type of convergence is usually called **stochastic equicontinuity** and was successfully applied to semi- and nonparametric estimation. See, for example, Andrews (1994) and Newey (1991).

Besides these theoretical caveats, there also arises a severe problem when putting an estimation method on infinite-dimensional parameter spaces to work. In practice, where the sample size is finite, estimating an infinite-dimensional object based upon a finite amount of information obviously appears to be a daunting task. This situation is known in statistics as an **ill-posed problem**. See Carrasco, Florens, and Ghysels (2007) for a review. Thus, a feasible nonparametric estimation method will be required to solve this ill-posed problem as well.

As a remedy to this 'large parameter space' problem, Grenander (1981) introduced what is nowadays known as the **method of sieves** and what boils down to replacing an infinite-dimensional (target) space Θ by a sequence $\{\Theta_n\}_{n\in\mathbb{N}}$ of finite-dimensional approximating spaces. Evidently, this turns an genuinely (infeasible) nonparametric estimation problem into a parametric one. However, in contrast to standard parametric estimation, the distinguishing feature of the method of sieves is that the dimension of the estimation problem increases with the sample size n, which adds a substantial amount of modeling flexibility.

Let (Θ, d) be a (semi-)metric space. A **sieve** is a sequence $\{\Theta_n\}_{n\in\mathbb{N}}$ of approximating spaces for Θ such that, for any $\theta \in \Theta$, there exists some $\theta_n \in \Theta_n$ satisfying

$$d(\theta_n, \theta) \to 0$$

as $n \to \infty$, i.e., the approximation error vanishes asymptotically. The approximate minimum contrast estimator $\hat{\theta}_n$ (2.2.2) on sieves is defined as the minimizer of $\gamma_n(\theta)$ over Θ_n, i.e.,

$$\gamma_n(\hat{\theta}_n) \leqslant \inf_{\theta\in\Theta_n} \gamma_n(\theta) + O_P(\epsilon_n) \ ,$$

where $\epsilon_n \to 0$ as $n \to \infty$.

Since we will deal with nonparametric estimation in function spaces only, the generic notion of a sieve is now replaced by a definition which is more suitable with respect to the basis functions used, in the sequel, to construct sieve subspaces.

Definition 2.2 (Sieve) *Let $\mathcal{F}$ be an infinite-dimensional (function) space and $\mathcal{M}_n$ be a collection (depending on the sample size n) of model labels $m := m_n$. A sieve is*

a sequence $\{F_m\}_{m\in\mathcal{M}_n}$ *of finite-dimensional, closed subspaces of* $\mathcal{F}$ *which satisfies the following conditions:*

C1 $F_m \subset F_{m+1}$

C2 $\{F_m\}_{m\in\mathcal{M}_n}$ *is dense in* $\mathcal{F}$ *as* $n \to \infty$.

Remark 2.3 *Basically, the collection* $\mathcal{M}_n$ *of models labeled by* m *can have two different forms, i.e.,* $\mathcal{M}_n$ *may be a collection of nested or non-nested models. For nested models, the collection* $\mathcal{M}_n$ *can be ordered in exactly the same way as the set of natural numbers such that* $m \in \mathbb{N}$. *Put differently, this structure allows us to totally order all models in* $\mathcal{M}_n$ *according to, say, their dimension* d_m. *This is the case which will be the relevant one in the sequel. For the sake of completeness, we mention that a collection* $\mathcal{M}_n$ *of non-nested models appears in the context of irregular histograms and wavelets, where we may have different models with the same dimension* d_m. *In this context, it is not possible to order all models in* $\mathcal{M}_n$ *according to their dimension* d_m, *but we rather need to resort to a lexicographical ordering.*

As already mentioned at the beginning of this section, the right choice of the bandwidth parameter is *the* crucial issue in kernel-based nonparametric estimation. From Definition 2.2, it seems to be obvious that the right choice of the model label m plays an analogous role in sieve-based nonparametric estimation. Moreover, this tuning parameter determines the degree of smoothing applied to the curve estimate. A non-optimal choice either leads to over- or undersmoothing. Oversmoothing means that significant details of the true function are blurred out, while undersmoothing means that the estimate is too rough or wiggly (relative to the true function) being pure artefacts of the sampling process. Since the sample size is finite in practice, it is important to relate this tuning parameter to the sample size n. In applications, popular procedures for such data-driven smoothing are cross-validation (Wahba, 1981) and the use of information criteria (Akaike, 1977).

Remark 2.4 *As an illustration of the problems of nonparametric estimation in (infinite-dimensional) function spaces, the usefulness of sieves in this context, and the role of the tuning parameter* m, *we look at a classical example from Grenander (1981) and Geman and Hwang (1982). Let* $X_1, \ldots, X_n$ *be an* iid *sample of random observations drawn from an absolutely continuous distribution function with an unknown probability density function* f_0 *belonging to the class* $\mathcal{F} := \{f : f \geqslant 0, \int f = 1\}$. *Moreover, define the contrast function* $\gamma(x; f) := \log f(x)$ *such that the nonparametric maximum likelihood estimator is*

given by

$$\hat{f}_n(X_1,\ldots,X_n) = \operatorname*{\mathsf{argmin}}_{f\in\mathcal{F}} \big[-\gamma_n(f)\big] = \operatorname*{\mathsf{argmin}}_{f\in\mathcal{F}} \left[-\sum_{i=1}^{n} \mathsf{log} f(X_i) \right]$$

See (2.2.1). *However, the maximizer of the empirical contrast function* $\gamma_n(f)$ *turns out to have a combe-type shape putting probability mass* $n^{-1}\delta_{x_i}$ *at the sample points* $x_1,\ldots,x_n$. *Hence,* $\hat{f}_n \in \mathcal{F}$ *is no sensible nonparametric estimator for an unknown probability density function, and it can be proved that, in this setup,* $\hat{f}_n$ *is not even consistent.*

The simplest sieve estimator for this problem is the histogram defined on the subspaces

$$F_m := \left\{ f : f \geqslant 0, \int f = 1, \textit{constant on } \left[\frac{k-1}{m}, \frac{k}{m} \right), m \in \mathbb{N}, k = 0, \pm 1, \pm 2, \ldots \right\} .$$

The maximizer (2.2.2) *for* $\gamma(x;f) := \mathsf{log} f(x)$ *is*

$$\hat{f}_m(x) = \frac{m}{n} \sharp \left\{ X_i : \frac{k-1}{m} \leqslant X_i < \frac{k}{m}, x \in \left[\frac{k-1}{m}, \frac{k}{m} \right) \right\} ,$$

which is just the histogram with bin width m^{-1}. *Although it can be proved that, for* $m \to \infty$, $\hat{f}_m$ *is strongly consistent, i.e.,*

$$P\left(\operatorname*{\mathsf{limsup}}_{n\to\infty} \int \big|\hat{f}_m(x) - f_0(x)\big| \mathsf{d}x = 0 \right) = 1 ,$$

it is necessary that $m = o(n^{-1})$. *The exact rate of* m *has to be tuned optimally such that it balances the effects of over- and undersmoothing. This example will be continued in Remark 2.7.*

We close this section with a short overview of the literature on sieve-based estimation. Following the initial impetus of Grenander (1981), Geman and Hwang (1982) proved and analyzed the conditions for the existence and consistency of sieve maximum likelihood estimation. Shen and Wong (1994) and Wong and Shen (1995) derived convergence rates of sieve maximum likelihood estimators. Convergence rates of sieve least-squares estimators were derived by van de Geer (1990). Moreover, Shen (1997) considered sieve maximum likelihood estimators, while van de Geer (1995b, 2002) and Birgé and Massart (1998) analyzed sieve minimum contrast estimators.

2.3 Orthogonal Projection Estimation on Fixed Sieve

The generality and flexibility of sieves allow us to use a unifying framework for nesting many popular approximating spaces derived from regular or irregular histograms, trigonometric polynomials, splines with fixed or variable knots, or wavelets (Barron, Birgé, and Massart, 1999). See also Chen (2007) for an extensive survey on sieves used in applied econometrics. As already mentioned in Section 2.2, although such a unifying framework looks appealing for comparing the theoretical properties of different sieves, the involved assumptions might be too general and stringent relative to direct derivations of these properties of a particular estimator such as, for example, least squares.

In what follows, we specialize the generic sieve of Definition 2.2 to approximating spaces for **orthogonal projections**. For orthogonal projection estimators it is natural to assume that the target function $f : D \to \mathbb{R}$ is an element of the infinite-dimensional space $\mathcal{F} = \mathcal{L}_2 := \mathcal{L}_2(D, \mathsf{d}x)$ and to equipped with the usual $\mathcal{L}_2$-(semi-)norm

$$||f|| := ||f||_{\mathcal{L}_2} = \langle f, f\rangle^{1/2} = \left(\int f^2\right)^{1/2} = \left(\int_I f^2(x)\,\mathsf{d}x\right)^{1/2},$$

where, for any $f, g \in \mathcal{F}$,

$$\langle f, g\rangle := \int fg = \int_D f(x)g(x)\,\mathsf{d}x$$

is the inner product of $\mathcal{L}_2$.[1] The **approximating function** f_m is an element of the d_m-dimensional linear space

$$F_m := \left\{\theta_1\varphi_1 + \cdots + \theta_{d_m}\varphi_{d_m} : \theta_1, \ldots, \theta_{d_m} \in \mathbb{R}, d_m \in \mathbb{N}\right\} \subset \mathcal{F},$$

where $\{\varphi_\lambda : 1 \leqslant \lambda \leqslant d_m\}$ is a set of orthonormal basis functions spanning F_m. Since F_m is a proper (closed) subspaces of $\mathcal{F}$, standard Hilbert space theory suggests that the **orthogonal projection**

$$\pi_m := \sum_{\lambda=1}^{d_m} \theta_\lambda\varphi_\lambda = \sum_{\lambda=1}^{d_m} \langle\varphi_\lambda, f_0\rangle\varphi_\lambda \tag{2.3.1}$$

[1] Figueroa-López and Houdré (2006) point out that it is possible to generalize Definition 1.12 by replacing the Lebesgue measure $\mathsf{d}x$ by a regularizing measure $\mathsf{d}\mu$, i.e., $\tilde{p} = \mathsf{d}\nu/\mathsf{d}\mu$. If, for example, $p = \mathsf{d}\nu/\mathsf{d}x$ blows off near the origin with rate x^{-1}, i.e., $p(x) = O(x^{-1})$ as $x \to 0$, which is the case for the gamma Lévy density in Section 4.1, then the regularizing measure $\mathsf{d}\mu = x^{-2}\mathsf{d}x$ guarantees that $\tilde{p}$ is still well-behaved near the origin and satisfies the condition $\int_{\mathbb{R}\setminus\{0\}} \tilde{p}^2\,\mathsf{d}\mu < \infty$ which has the advantage that it allows to extend $\mathcal{L}_2(D, \mathsf{d}x)$ to $\mathcal{L}_2(\mathbb{R}\setminus\{0\}, \mathsf{d}\mu)$. Additionally, this is expected to yield more accurate estimates of the Lévy density near the origin.

of f_0 onto F_m is the element $f_m \in F_m$ which is closest to $f_0 \in \mathcal{F}$ in terms of the distance $||f_m - f_0||^2$. Unfortunately, the orthogonal projection π_m is computationally infeasible from a statistical perspective since the θ_λ's depend upon the unknown target function f_0. Consequently, the next step is to find a computationally feasible estimator $\hat{f}_m$ of π_m. Given an iid sample of observations $X_1, \ldots, X_n$, a sensible estimator is based on a simple moment estimator of

$$\theta_\lambda = \langle \varphi_\lambda, f \rangle = P\varphi_\lambda = \int_D \varphi_\lambda(x) f_0(x) \, \mathrm{d}x \; , \tag{2.3.2}$$

i.e.,

$$\hat{\theta}_\lambda = P_n \varphi_\lambda = \frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda(X_i) \; , \tag{2.3.3}$$

for all $\lambda = 1, \ldots, d_m$, which turns out to be unbiased:

$$\mathsf{E}\Big[\, \hat{\theta}_\lambda \,\Big] = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}[\, \varphi_\lambda(X_1) \,] = \mathsf{E}[\, \varphi_\lambda(X_1) \,] = \int_D \varphi_\lambda(x) f_0(x) \, \mathrm{d}x = \theta_\lambda \; .$$

Thus,

$$\hat{f}_m := \sum_{\lambda=1}^{d_m} \hat{\theta}_\lambda \varphi_\lambda \tag{2.3.4}$$

is an unbiased estimator of π_m, i.e., $\mathsf{E}\Big[\, \hat{f}_m \,\Big] = \pi_m$. Moreover, since

$$\mathsf{Var}\Big[\, \hat{\theta}_\lambda \,\Big] = \mathsf{Var}\left[\sum_{i=1}^{n} \frac{\varphi_\lambda(X_i)}{n} \right] = \frac{1}{n} \mathsf{Var}[\, \varphi_\lambda(X_1) \,] \; , \tag{2.3.5}$$

the orthogonal projection estimator of π_m is mean-square consistent, i.e., $\hat{f}_m \xrightarrow{m.s.} \pi_m$.

In order to gauge the performance of the sieve estimator $\hat{f}_m$, we follow the common folklore in nonparametric estimation theory by assessing its risk derived from the $\mathcal{L}_2$-loss $||\hat{f}_m - f_0||^2$. The use of the $\mathcal{L}_2$-risk is a natural choice for function estimation and is usually justified by the fact that it allows for a neat decomposition of the global risk in a (squared) bias term and a variance term (more generally, the stochastic error)

$$\begin{aligned} \mathsf{E}\Big[\, ||\hat{f}_m - f_0||^2 \,\Big] &= \mathsf{E}\Big[\, ||\hat{f}_m - \pi_m + \pi_m - f_0||^2 \,\Big] \\ &= \mathsf{E}\big[\, ||f_0 - \pi_m||^2 \,\big] + \mathsf{E}\Big[\, ||\hat{f}_m - \pi_m||^2 \,\Big] \\ &= \underbrace{||f_0 - \pi_m||^2}_{\text{bias term}} + \underbrace{\mathsf{E}\Big[\, ||\hat{f}_m - \pi_m||^2 \,\Big]}_{\text{variance term}} , \end{aligned} \tag{2.3.6}$$

due to the unbiasedness of $\hat{f}_m$. This classical trade-off is optimally solved by minimizing the risk via balancing (the rates of) the bias and the variance term.

As a general loss function $\ell : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$ can be expressed in terms of a contrast function by

$$\ell(f_0, f) = \mathsf{E}_{f_0}[\,\gamma(X, f) - \gamma(X, f_0)\,] = \int \big[\gamma(X, f) - \gamma(X, f_0)\big]\,\mathsf{d}P\ ,$$

Birgé and Massart (1998) proposed the contrast function

$$\gamma(X; f) := ||f||^2 - 2f(X) \tag{2.3.7}$$

for the problem of nonparametric density estimation via orthogonal projections. The resulting loss function is indeed equivalent to the traditional $\mathcal{L}_2$-loss, i.e.,

$$\ell(f_0, f) = ||f - f_0||^2\ . \tag{2.3.8}$$

See Appendix 2.A for a derivation. Thus, the empirical contrast function of $\gamma(X; f)$,

$$\gamma_n(f) = \frac{1}{n}\sum_{i=1}^{n}\gamma(X_i; f) = \int f^2 - \frac{2}{n}\sum_{i=1}^{n} f(X_i)\ , \tag{2.3.9}$$

is minimized at $\hat{f}_m$ with minimum value

$$\begin{aligned}\gamma_n(\hat{f}_m) &= \int \hat{f}_m^2 - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_m(X_i) = \int \left(\sum_{\lambda=1}^{d_m}\hat{\theta}_\lambda\varphi_\lambda\right)^2 - \frac{2}{n}\sum_{i=1}^{n}\sum_{\lambda=1}^{d_m}\hat{\theta}_\lambda\varphi_\lambda(X_i)\\ &= \sum_{\lambda=1}^{d_m}\hat{\theta}_\lambda^2\int\varphi_\lambda^2 - 2\sum_{\lambda=1}^{d_m}\hat{\theta}_\lambda\frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda(X_i) = -\sum_{\lambda=1}^{d_m}\hat{\theta}_\lambda^2\ .\end{aligned} \tag{2.3.10}$$

Provided that the basis function satisfy a certain boundedness condition, the stochastic error in (2.3.6) can be bounded as well.

Proposition 2.5 *Let π_m be an orthogonal projection* (2.3.1) *of $f_0 \in \mathcal{F} = \mathcal{L}_2(D, \mathsf{d}x)$ on F_m and $\hat{f}_\lambda$ its corresponding estimator in* (2.3.4)*. If there exists a bound such that*

$$\left|\left|\sum_{\lambda=1}^{d_m}\varphi_\lambda^2\right|\right|_\infty = D_m\ ,$$

then the risk decomposition (2.3.6) *can be bounded as follows:*

$$\mathsf{E}\Big[\,||\hat{f}_m - f_0||^2\,\Big] \leqslant ||f_0 - \pi_m||^2 + \frac{D_m}{n}\;.$$

Remark 2.6 *Note that the upper bound D_m in Proposition 2.5 is proportional to the dimension d_m of the orthogonal projection. Thus, we can rewrite, with a little abuse of notation, the result of Proposition 2.5 as*

$$\mathsf{E}\Big[\,||\hat{f}_m - f_0||^2\,\Big] \leqslant ||f_0 - \pi_m|| + \frac{d_m}{n}\;,$$

which provides a better interpretation of the stochastic term.

On the one hand, the variance term increases linearly with the complexity of approximating space F_m, since the higher the dimension d_m, the more parameters θ_λ have to be estimated. Put differently, increasing the complexity of the model renders the estimation of its parameters less precise (for a fixed sample size). However, for $n \to \infty$, it follows from $\hat{f}_m \xrightarrow{m.s.} \pi_m$ that $\mathsf{E}\Big[\,||\hat{f}_m - \pi_m||^2\,\Big] \to 0$, for a fixed $m \in \mathcal{M}$ (or for a given approximating space F_m).

On the other hand, although the choice of the sieve, i.e., the sort of the underlying orthonormal basis functions $\{\varphi_\lambda : \lambda \in \mathbb{N}\}$, does not affect the properties of the estimation error, it can be a crucial issue for controlling the bias term in (2.3.6), since not all sieves may adapt equally well to important features of f. The bias term simply measures the discrepancy between the target function f and the best possible approximation from F_m. Clearly, the higher the dimension (or complexity) of F_m, the smaller the approximation error due to the denseness of the sieve in Condition C2 of Definition 2.2.

In sum, the complexity of $\hat{f}_m$ depends on the dimension d_m of a chosen approximating space F_m which should grow as the sample size n increases. But at the same time, approximating spaces with low dimension are preferable from the view point of estimation precision. Hence, we end up with two fundamental insights. First, the risk is determined by the classical trade-off between misspecification error and estimation error which typically occurs in nonparametric estimation theory, although the estimation is performed on a fixed parametric sieve. Second, this trade-off is 'tuned' by the choice of the model label $m \in \mathcal{M}_n$, where the collection of models depends on the sample size n.

Although the method of sieves allows us to transform an infeasible function estimation problem into a straightforward parametric estimation problem, it has been noted in the literature that it may be still hampered by (suboptimally) slow rates of convergence (Birgé and Massart, 1993). This happens to be the case in estimation procedures which

allow the dimension (or complexity) to grow with the sample size n such that they tend to opt for models with 'too' high dimensions. This has been known for quite some time in the nonparametrics literature, and one way of resolving it was by penalizing for model complexity.

2.4 Penalized Model Selection on Sieves

Before discussing penalization on sieves, we shortly return to the example of nonparametric maximum likelihood estimation of a probability density function introduced in Remark 2.4, since penalization is considered as an alternative to the method of sieves. At first sight, these methods seem to be rather different, but this is just ostensible as we will now show.

Remark 2.7 *Define the penalized contrast function*

$$\tilde{\gamma}(x; f) := \gamma(x; f) - \delta J_n(f) = \mathsf{log} f(x) - \delta J_n(f) \; ,$$

where δ is a Lagrange parameter and $J_n(f)$ is a non-negative penalty term, which leads to the empirical contrast function

$$\tilde{\gamma}_n(f) = \gamma_n(f) - \delta J_n(f) = \frac{1}{n} \sum_{i=1}^{n} \gamma(X_i; f) - \delta J_n(f) \; .$$

Then, the approximate penalized estimator is defined as the approximate minimizer $\hat{f}_n$ of $\tilde{\gamma}_n(f)$ over $\mathcal{F}$ such that

$$\tilde{\gamma}_n(\hat{f}_n) \leqslant \inf_{f \in \mathcal{F}} \big[- \tilde{\gamma}_n(f) \big] + O_P(\epsilon_n) \; , \tag{2.4.1}$$

where $\epsilon_n \to 0$ as $n \to \infty$. This corresponds to the penalized maximum likelihood estimator of a probability density function of Good and Gaskins (1971). Moreover, it can also be shown, by rewriting (2.4.1) *as*

$$\hat{f}_n(X_1, \ldots, X_n) = \operatorname*{arg\,min}_{f \in \mathcal{F}} \big[- \gamma_n(f) \big] \qquad s.t. \quad \mathsf{pen}(f) \leqslant m \; ,$$

if it exists, that it also corresponds to the minimum contrast estimation over the sieve subspace

$$F_m := \big\{ f \in \mathcal{F} : \mathsf{pen}(f) \leqslant m \big\} \; .$$

This shows the conceptual difference between the method of penalization and the method

of sieves. The former minimizes over the original parameter space $\mathcal{F}$, while the latter minimizes over F_m. However, the penalty term essentially 'transforms' the unconstrained, infinite-dimensional problem into a parametric one. Usually, the penalty term $J_n(\,\cdot\,)$ is a measure of smoothness of f and is chosen to penalize rough estimates heavier than smooth ones. This rules out combe-type estimates as in Remark 2.4 and forces the estimate to belong to a smoothness class, say a Sobolev space, of functions. More other examples, see Silverman (1982) and Wahba (1990).

In Section 2.3, we discussed the optimal estimation via orthogonal projections on a fixed sieve. Now, we want to look at how to gain additional modeling flexibility by allowing to choose a projection estimator $\hat{f}_m$ from the best model $\hat{m}$ in the given collection $\mathcal{M}_n$. Consequently, we have to solve two optimality problems where, fortunately, the first one has already been solved in Section 2.3. In practice, this boils down to the following two-step procedure:

1. **Projection step**: compute the orthogonal projection estimator $\hat{f}_m$ for all $m \in \mathcal{M}_n$

2. **Model selection**: select model label $\hat{m}$ indexing the best estimator $\hat{f}_m$ over all $m \in \mathcal{M}_n$

In order to avoid overfitting in the model-selection step, consider the penalized version of the empirical contrast function (2.3.9), i.e., $\gamma_n(f) + \mathsf{pen}(m)$. Then, the optimal estimator $\hat{f}_{\hat{m}}$ on the sieve $\{F_m\}_{m\in\mathcal{M}_n}$ is defined, if it exists, to satisfy

$$\gamma_n(\hat{f}_{\hat{m}}) + \mathsf{pen}(\hat{m}) = \inf_{m\in\mathcal{M}_n} \left[\inf_{f\in F_m} \gamma_n(f) + \mathsf{pen}(m) \right] . \tag{2.4.2}$$

The right-hand side of this equality exactly reflects the nested structure of the above two-step procedure. As we know from (2.3.10),

$$\gamma_n(\hat{f}_m) = -\sum_{\lambda=1}^{d_m} \hat{\theta}_\lambda^2$$

which can be use to eliminate the inner optimization by directly plugging in $\gamma_n(\hat{f}_m)$ in (2.4.2) such that we are left with the outer optimization which reduces to

$$\hat{m} = \operatorname*{arg\,min}_{m\in\mathcal{M}_n} \left\{ -\sum_{\lambda=1}^{d_m} \hat{\theta}_\lambda^2 + \mathsf{pen}(m) \right\} . \tag{2.4.3}$$

In Appendix 2.A, we show that the penalty term in (2.4.3) takes the form

$$\mathsf{pen}(m) = \frac{2}{n^2} \sum_{i=1}^{n} \sum_{\lambda=1}^{d_m} \varphi_\lambda^2(X_i) \; . \tag{2.4.4}$$

Moreover, a similar result as in Proposition 2.5 for model selection on sieves with nested models can be proved.

Theorem 2.8 (Birgé and Massart (1997)) *Assume that the boundedness condition*

$$\left|\left| \sum_{\lambda=1}^{d_m} \varphi_\lambda^2 \right|\right|_\infty < \Phi^2 d_m$$

and the nestedness condition

$$d_m < d_{m'} \quad \Longrightarrow \quad F_m \subset F_{m'}$$

hold true. Then, for $C_1 > 0$ and $C_2 \geqslant 1$ and the penalty term

$$\mathsf{pen}(m) = \frac{(C_1 + C_2^2 \Phi^2) d_m}{n} \; ,$$

the following risk inequality holds

$$\mathsf{E}\Big[\, ||\hat{f}_{\hat{m}} - f||^2 \, \Big] \leqslant C \inf_{m \in \mathcal{M}_n} \left[||f - \pi_m||^2 + \frac{d_m}{n} \right] \; .$$

Remark 2.9 *Nonasymptotic risk bounds such as in Theorem 2.8 are often called* ***oracle inequalities****. They are derived from concentration inequalities which are due to Talagrand (1994, 1996). For a more general treatment, see Ledoux and Talagrand (1991). Analogous to the example in Remark 2.1, the key idea of Theorem 2.8 is to use the penalty as a control on the oscillations of an empirical process based on the difference of empirical contrast functions. These increments are characterized by concentration inequalities. See Massart (2007). Massart (2000) provided a discussion of the constants of theses concentration inequalities.*

Oracle inequalities, like the one in Theorem 2.8, should not be confused with the minimax bound (2.1.2). They only describe how a proposed estimator behaves, for all $n \in \mathbb{N}$, relative to the so-called **oracle**. An additional step is required to show that the oracle is minimax (adaptive) which, in turn, translates back to $\hat{f}_{\hat{m}}$. Barron, Birgé, and Massart (1999) showed that many model-selection based estimators are indeed adaptive in the minimax sense. We refer to Section 2.6 for an illustration of the oracle approach.

We close this section by mentioning an interesting interpretation of the constants appearing in the penalty term of Theorem 2.8. The penalty term can be cast in the generic form

$$\mathsf{pen}(m) = \underbrace{\kappa}_{\substack{\text{depends on data} \\ \text{but not on } f}} \underbrace{L_m}_{\substack{\text{complexity} \\ \text{of } \mathcal{M}_n}} \underbrace{d_m}_{\substack{\text{complexity} \\ \text{within } F_m}} \big/ n \ ,$$

where the L_m's have to satisfy

$$\sum_{m \in \mathcal{M}_n} \mathsf{exp}(-L_m d_m) \leqslant \Sigma < \infty \ ,$$

with $\kappa, L_m > 0$. Put together, the weight L_m has a dual role. On the one hand, it should be small to keep the penalized risk at a low level. On the other hand, they it be large when

$$\sum_{m \in \mathcal{M}_n} \mathsf{exp}(-d_m) = \infty \ ,$$

which happens to be the case for collections $\mathcal{M}_n$ of non-nested models. More precisely, we shall usually choose $L_m = 1$ for nested models and $L_m = L\,\mathsf{ln}(n)$ for non-nested models. This turns out to be of great importance for the minimax adaptive rate of convergence. Note that, similar to the notion of minimum description length (Barron and Cover, 1991), $\mathsf{exp}(-L_m d_m)$ has a Bayesian flavor as they may be interpreted as a prior probability that we assign to a specific model m.

2.5 Lévy Density Estimation with Discretely Sampled Data

Recall from (2.3.4) that, given an iid sample $X_1, \ldots, X_n$, the orthogonal projection estimator on a fixed sieve was defined by

$$\hat{f}_m = \sum_{\lambda=1}^{d_m} \hat{\theta}_\lambda \varphi_\lambda$$

with estimated Fourier coefficients

$$\hat{\theta}_\lambda = P_n \varphi_\lambda = \frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda \ .$$

As our aim is to estimate a continuous-time process, we shall first assume that we have at our disposal a continuous record of observations $\{X_t\}_{t \in [0,T]}$ of the underlying pure-jump

Lévy process X. In this case, all jumps are perfectly identified by the (continuous-time) increments $\Delta X_t := X_t - X_{t-}$ of X. By a conjecture of Figueroa-López and Houdré (2006), a sensible estimator for the Fourier coefficients of (2.3.4) is defined by

$$\hat{\theta}_{\lambda,C} = \frac{1}{T}\int_0^T \int_D \varphi_\lambda J(\mathsf{d}t \times \mathsf{d}x) \ ,$$

where the domain of estimation is restricted to a compact subset of $\mathbb{R}$ excluding the origin, i.e.,

$$D = [a,b] \subset \mathbb{R}\backslash\{0\} \ ,$$

since this guarantees that $\nu(D) < \infty$, due to the σ-finiteness of any Lévy measure. In order to see that $\hat{\theta}_{\lambda,C}$ is indeed a reasonable estimator, note that, by P2 of Theorem 1.11,

$$\mathsf{E}\Big[\,\hat{\theta}_{\lambda,C}\,\Big] = \frac{1}{T}\mathsf{E}\bigg[\int_0^T \int_D \varphi_\lambda J(\mathsf{d}t \times \mathsf{d}x)\bigg] = \int_D \varphi_\lambda(x)\nu(\mathsf{d}x) = \int_D \varphi_\lambda(x)p(x)\,\mathsf{d}x \ ,$$

given that the conditions of Definition 1.12 and Theorem 1.11 are satisfied. Put differently, transferring the penalized model-selection approach of Section 2.4 simply amounts to replacing the empirical operator P_n by a properly scaled Poisson integral. Then, the empirical contrast function (2.3.9) reads as

$$\gamma_{n,C}(f) = \int_D f^2(x)\,\mathsf{d}x - \frac{2}{T}\int_{[0,T]}\int_D f(x)J(\mathsf{d}t \times \mathsf{d}x) \ ,$$

while the penalty term (2.4.4) can be computed by

$$\mathsf{pen}_C(m) = \frac{2}{T^2}\int_{[0,T]}\int_D \sum_{\lambda=1}^{d_m} \varphi^2(x)J(\mathsf{d}t \times \mathsf{d}x) \ .$$

As we know from (1.3.2), a Poisson integral can be represented as

$$\int_0^T \int_D f(x)J(\mathsf{d}s \times \mathsf{d}x) = \sum_{0<t\leqslant T} f(\Delta X_t) \ ,$$

which shows how to explicitly compute the estimator of the Fourier coefficients:

$$\hat{\theta}_{\lambda,C} = \frac{1}{T}\sum_{0<t\leqslant T} \varphi_\lambda(\Delta X_t) \ .$$

Unfortunately, observations of a continuous-time process cannot be sampled continuously. Instead, we are often confronted with the situation where our data set is discretely

sampled at, say n, equidistant time points; i.e.,

$$0 = t_0 < t_1 < \cdots < t_n = T \ .$$

Then, the sampling frequency is given by

$$\Delta_n = \frac{T}{n} \ .$$

Attached to these time points is a set of observations $\{X_{t_i}\}_{i=0}^n$. As an educated guess, one would naturally try to substitute the continuous-time increments by the corresponding discrete-time increments

$$\Delta X_i := X_{t_i} - X_{t_i - 1} \ ,$$

for all $i = 1, \ldots, n$, such that a feasible estimator of the Fourier coefficients is given by

$$\hat{\theta}_\lambda = \frac{1}{T} \sum_{i=1}^n \varphi_\lambda(\Delta X_i) \ . \tag{2.5.1}$$

Given the intuition in Figure 1.6, we would hope that $\hat{\theta}_\lambda$ converges to $\hat{\theta}_{\lambda,C}$ in some sense. This was indeed accomplished by Figueroa-López and Houdré (2006) who showed, by invoking the following result of Sato (1999, p. 45), that the discrete-time Poisson integral converges weakly (in distribution) to the continuous-time Poisson integral, i.e.,

$$\sum_{i=1}^n f(\Delta X_i) \rightsquigarrow \sum_{0<t\leqslant T} f(\Delta X_t) \ ,$$

as $n \to \infty$, for all f defined in Corollary 2.10.

Corollary 2.10 (Sato (1999)) *Let $\Delta_n \searrow 0$. If ν is the Lévy measure of an infinitely divisible distribution P, then for any $f \in \mathcal{C}_0^{\#}$ (the class of bounded continuous functions from $\mathbb{R}$ to $\mathbb{R}$ vanishing on a neighborhood of 0)*

$$\Delta_n^{-1} \int_{\mathbb{R}} f(x) P^{\Delta_n}(\mathsf{d}x) \to \int_{\mathbb{R}} f(x) \nu(\mathsf{d}x) \ .$$

Remark 2.11 *Note the close similarity of the relation stated in Corollary 2.10 and the form of infinite divisibility given in* (1.1.1)*. Actually, this is no coincidence since this result is a byproduct of the proof of Theorem 1.3. Moreover, Figueroa-López (2009) recently showed that $\Delta_n = o(T^{-1})$, implying that $\Delta_n \to 0$ at a faster rate than $T \to \infty$. This rate for high-frequency sampling is sufficient to guarantee that the Lévy measure can be identified.*

We close this section by pointing out that there seems to be alternative to Figueroa-López and Houdré (2006) by invoking the functional central limit theorem of Liese and Ziegler (1999) who established the weak convergence of a sequence of Poisson processes. This is indeed closely related to Corollary 8.8 of Sato (1999). Unfortunately, it is yet not clear how to relate this result in-fill asymptotics.

2.6 Histogram Estimation Based on Sieves

Let us now return to the penalized model selection of Section 2.4 and exemplify this approach by considering the problem of constructing an optimal histogram estimator for a Lévy density. Our approach follows Birgé and Rozenholc (2006) who considered the problem of nonparametric density estimation via model selection.

To this end, assume that we have computed the discrete-time increments $\{\Delta X_{t_i}\}_{i=1}^n$ from Section 2.5. For the ease of exposition, we drop the Δ-sign such that the sample of discrete-time increments $\{\Delta X_{t_i}\}_{i=1}^n$ is denoted by $\{X_{t_i}\}_{i=1}^n$ from now on. A **regular histogram estimator** of the Lévy density p with d_m bins is defined by

$$\hat{f}_m := \frac{d_m}{n} \sum_{\lambda=1}^{d_m} N_\lambda \mathbb{1}_{I_\lambda} , \tag{2.6.1}$$

where

$$N_\lambda = \sum_{i=1}^{n} \mathbb{1}_{\{X_{t_i} \in I_\lambda\}}$$

is the number of discrete-time increments whose value fall in the interval I_λ. To be more precise, this estimator implies a random partition $\mathcal{I}_m = \{I_1, \ldots, I_{d_m}\}$ of $[0,1]$ into d_m intervals of equal length $1/d_m$.

It is noteworthy that, without loss of generality, we have assumed $D = [0,1]$ such that each binwidth of the partition induced by d_m corresponds to $1/d_m$. In applications, we will stick to that convention by transforming the range of sampled observations, which is taken as a rough approximation of the unknown true support of p, to $[0,1]$ using an affine transformation of the data. Furthermore, recall that the Lévy density is not defined at the origin such that it would be more appropriate to restrict the support on $[\epsilon, 1]$. However, as this leads to unhandy expression, we keep $[0,1]$. Again, this poses no problem in applications since zero increments are excluded, if they appear.

Substituting N_λ in (2.6.1) shows that

$$\hat{f}_m := \sum_{\lambda=1}^{d_m} \left(\frac{d_m}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in I_\lambda\}} \right) \mathbb{1}_{I_\lambda} \tag{2.6.2}$$

is indeed a special orthogonal projection estimator from (2.3.4) based on the orthonormal basis $\{\mathbb{1}_{I_\lambda}\}_{\lambda \in \mathbb{N}}$. As already discussed in Remark 2.4, this type of estimator corresponds to the maximum likelihood estimator on a fixed, finite-dimensional sieve F_m spanned by $\{\mathbb{1}_{I_\lambda} : 1 \leqslant \lambda \leqslant d_m\}$, i.e., the space of all densities which are piecewise constant on the partition $\mathcal{I}_m$. More precisely,

$$\hat{f}_m = \operatorname*{\mathsf{arg\,min}}_{f \in F_m} \gamma_n(f) \ ,$$

with contrast function $\gamma(x; f) := -\mathsf{ln} f(x)$ which naturally leads to the the loss function

$$\ell(f, g) := \mathsf{E}_f[\, \gamma(X; g) - \gamma(X; f) \,] = \int_0^1 \mathsf{ln} \left(\frac{f(x)}{g(x)} \right) f(x) \, \mathsf{d}x =: \mathsf{K}(f, g) \ , \tag{2.6.3}$$

where K denotes the Kullback-Leibler divergence.

for going from an optimization problem on a fixed sieve to a model selection problem, we proceed as in Section 2.4 by adding a penalty term $\mathsf{pen}(m)$ which should guide us through the model class $\mathcal{M}_n$. Note that since we are not dealing with irregular histograms, $\mathcal{M}_n$ is nested. Then, the penalized maximum likelihood estimator $\hat{f}_{\hat{m}}$ is defined as *the* $\hat{f}_m$ which satisfies

$$\hat{m} = \operatorname*{\mathsf{arg\,min}}_{m \in \mathcal{M}_n} \Big\{ \gamma_n(\hat{f}_m) + \mathsf{pen}(m) \Big\} \ . \tag{2.6.4}$$

Unfortunately, we cannot use the penalty term (2.4.4) since it was explicitly motivated for least-squares problems based on the $\mathcal{L}_2$-equivalent contrast function (2.3.7). Moreover, van de Geer (1995b) argued against the $\mathcal{L}_2$-loss induced by (2.3.7) and in favor of the Hellinger loss, since the latter is expected to be a better measure of derivation for density estimation. See also Le Cam (1986).

Let us now show how the oracle approach may help us to find a sensible penalty term for the problem at hand. First, we need to compute the risk (2.1.1) of $\hat{f}_m$ at f. To this end, we need to define an appropriate risk function. As Birgé and Rozenholc (2006) argued, the Kullback-Leibler loss (2.6.3) is not a good choice since there exists the possibility that, for $d_m \geqslant 2$, a bin may contain no observations such that $\mathsf{K}(f, \hat{f}_m) = \infty$. Based on what has been argued in the preceding paragraph, we choose the Hellinger loss

$$\mathsf{h}^2(f, g) := \frac{1}{2} \int_0^1 \Big(\sqrt{f(x)} - \sqrt{g(x)} \Big)^2 \, \mathsf{d}x \ . \tag{2.6.5}$$

Consequently, the risk of $\hat{f}_m$ at f is defined by

$$R_n(f, \hat{f}_m) := \mathsf{E}_f\Big[\, \mathsf{h}^2(f, \hat{f}_m) \,\Big] \ , \tag{2.6.6}$$

It can be gauged by the following result, which is related to Proposition 2.5.

Theorem 2.12 *Let $X_1, \ldots, X_n$ be a random sample drawn from a Lévy density f, and let $\hat{f}_m$ be the histogram estimate* (2.6.1) *on the regular partition $\mathcal{I} = \{I_1, \ldots, I_{d_m}\}$ of $[0, 1]$. Define the orthogonal projection on the same partition by*

$$\pi_m = \sum_{\lambda=1}^{d_m} p_\lambda d_m \mathbb{1}_{I_\lambda} \ ,$$

where $p_\lambda = \int_{I_\lambda} f$. Then,

$$\begin{aligned} \mathsf{E}_f\Big[\, \mathsf{h}^2(f, \hat{f}_m) \,\Big] &\leqslant \mathsf{h}^2(f, \pi_m) + \frac{d_m - 1}{2n} \\ \mathsf{E}_f\Big[\, \mathsf{h}^2(f, \hat{f}_m) \,\Big] &= \mathsf{h}^2(f, \pi_m) + \frac{d_m - 1}{8n}\big[1 + o(1)\big] \ , \end{aligned}$$

as $n \to \infty$.

Next, assume that there exists an oracle telling us which of the model in $\mathcal{M}_n$ is best in the sense that it minimizes the risk over $\mathcal{M}_n$. More precisely, if we denote this best model by m^*, then it satisfies

$$m^* = \operatorname*{arg\,min}_{m \in \mathcal{M}_n} R_n(f, \hat{f}_m) \ .$$

Unfortunately, m^* cannot be taken as an estimator since it depends on f. However, it can be used as a benchmark for $R_n(f, \hat{f}_m)$ such that m is selected in such a way that it 'behaves' similar to $R_n(f, \hat{f}_{m^*})$. To be more precise, we seek to find a data-driven selection procedure for m such that the ratio of risks,

$$\frac{R_n(f, \hat{f}_m)}{R_n(f, \hat{f}_{m^*})} \leqslant C \qquad \text{with } C \geqslant 1,$$

is minimized. As this selection procedure is directly connected to the penalty term, we are able to derive an explicit expression (up to some constants) of $\mathsf{pen}(m)$. To this end, we now present a full-fledged result and oracle inequality for the problem at hand. It contains all the ingredients mentioned in Section 2.4.

Theorem 2.13 (Massart (2007)) *Assume that all conditions of Theorem 2.12 with*

$\hat{f}_{\hat{m}}$ *satisfying* (2.6.4). *Let* Σ *be some absolute constant and* $\{L_m\}_{m\in\mathcal{M}_n}$ *be a collection of nonnegative weights such that*

$$\sum_{m\in\mathcal{M}_n} \mathsf{e}^{-L_m(d_m-1)} \leqslant \Sigma$$

holds. Assume that there exists some penalty function $\mathsf{pen}(m)$ *such that, for all* $m \in \mathcal{M}_n$,

$$\mathsf{pen}(m) \geqslant c_1 \left(\sqrt{d_m - 1} + \sqrt{c_2 L_m (d_m - 1)}\right)^2$$

with $c_1 > 1/2$ *and* $c_2 = 2(1 + c_1^{-1})$. *If there exists some constant* $\rho > 0$, *such that* $f \geqslant \rho$ *(P-a.e.), and* $\int f(\mathsf{ln} f)^2 \leqslant L < \infty$, *then it holds, for some constant* $C(c_1, \rho, L, \Sigma)$,

$$\mathsf{E}_f\Big[\,\mathsf{h}^2(f, \hat{f}_{\hat{m}})\,\Big] \leqslant \frac{(2c_1)^{1/5}}{(2c_1)^{1/5} - 1} \inf_{m\in\mathcal{M}} \left[\mathsf{K}(f, \pi_m) + \frac{\mathsf{pen}(m)}{n}\right] + \frac{C(c_1, \rho, L, \Sigma)}{n}\,.$$

Remark 2.14 *The penalty term can be rewritten as*

$$\mathsf{pen}(m) = c_1(d_m - 1)\left(1 + \sqrt{c_2 L_m}\right)^2\,.$$

While it was shown, based on asymptotic considerations and simulations by Birgé and Rozenholc (2006), that $c_1 = 1$ *is optimal, the choice of the weights* L_m *is a delicate issue. On the one hand, the weights should be small to obtain a small penalty term. On the other hand, the weights should be large for decreasing the risk bounds via* Σ. *Massart (2007) provided some bounds on the weights which are still not sufficient to operationalize the penalty terms.*

Based on an extensive simulation study, including densities with spatial inhomogeneities like discontinuities etc., Birgé and Rozenholc (2006) found a robust calibration of the penalty term:

$$\mathsf{pen}(m) = d_m - 1 + \big[\mathsf{ln}(d_m)\big]^{2.5}\,,$$

for $1 \leqslant d_m \leqslant n/\mathsf{ln}(n)$. Thus, our two-step model selection procedure from Section 2.4 reads as follows:

1. **Projection step**: compute the orthogonal projection estimator $\hat{f}_m$ for all $1 \leqslant d_m \leqslant n/\mathsf{ln}(n)$
2. **Model selection**: select model satisfying label $\hat{m}$

$$\hat{m} = \operatorname*{arg\,min}_{1\leqslant d_m \leqslant n/\mathsf{ln}(n)} \Big\{\gamma_n(\hat{f}_m) + \mathsf{pen}(m)\Big\}$$

2.A Proofs & Auxiliary Results for Chapter 2

Proof of Equation 2.3.8

This equality follows from straightforward computations:

$$\begin{aligned}
\ell(f_0, f) &= \mathsf{E}_{f_0}[\,\gamma(X,f) - \gamma(X,f_0)\,] = \mathsf{E}_{f_0}\big[\,||f||^2 - 2f(X) - ||f_0||^2 + 2f_0(X)\,\big] \\
&= \int \big[||f||^2 - 2f - ||f_0||^2 + 2f_0\big]\, f_0 \\
&= \int ||f||^2 f_0 - 2\int f f_0 + 2\int f_0 f_0 - \int ||f_0||^2 f_0 \\
&= ||f||^2 \int f_0 - 2\langle f, f_0\rangle + 2\langle f_0, f_0\rangle - ||f_0||^2 \int f_0 \\
&= ||f||^2 - 2\langle f, f_0\rangle + 2||f_0||^2 - ||f_0||^2 = ||f||^2 - 2\langle f, f_0\rangle + ||f_0||^2 = ||f - f_0||^2 \,.
\end{aligned}$$

Proof of Proposition 2.5

This proof explicitly derives the assertions of Birgé and Massart (1998). Define the empirical process

$$\mathbb{G}_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int f f_0 \,.$$

Setting $f = \varphi_\lambda$ in $\mathbb{G}_n f$ leads to the following result

$$\begin{aligned}
\chi^2 := \sum_{\lambda=1}^{d_m} (\mathbb{G}_n\varphi_\lambda)^2 &= \sum_{\lambda=1}^{d_m}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda(X_i) - \int \varphi_\lambda f_0\right]^2 \\
&= \sum_{\lambda=1}^{d_m}\left[\theta_\lambda - \frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda(X_i)\right]^2 = \sum_{\lambda=1}^{d_m}(\theta_\lambda - \hat\theta_\lambda)^2 \\
&= \sum_{\lambda=1}^{d_m}(\theta_\lambda - \hat\theta_\lambda)^2 \int \varphi_\lambda^2 = \int\left[\sum_{\lambda=1}^{d_m}(\theta_\lambda - \hat\theta_\lambda)\varphi_\lambda\right]^2 \\
&= \int\left[\sum_{\lambda=1}^{d_m}\theta_\lambda\varphi_\lambda - \sum_{\lambda=1}^{d_m}\hat\theta_\lambda\varphi_\lambda\right]^2 = \int(\pi_m - \hat f_m)^2 \\
&= ||\hat f_m - \pi_m||^2 \,.
\end{aligned}$$

The expectation of χ^2 is given by

$$\mathsf{E}\big[\,\chi^2\,\big] = \mathsf{E}\left[\sum_{\lambda=1}^{d_m}(\mathbb{G}_n\varphi_\lambda)^2\right] = \sum_{\lambda=1}^{d_m}\mathsf{E}\big[\,(\mathbb{G}_n\varphi_\lambda)^2\,\big] = \sum_{\lambda=1}^{d_m}\mathsf{E}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda(X_i) - \int\varphi_\lambda f_0\right\}^2\right]$$

$$= \sum_{\lambda=1}^{d_m} \mathsf{E}\left[\left\{ \frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda(X_i) \right\}^2 - \frac{2}{n} \int \varphi_\lambda f_0 \sum_{i=1}^{n} \varphi_\lambda(X_i) + \left\{ \int \varphi_\lambda f_0 \right\}^2 \right]$$
$$= \sum_{\lambda=1}^{d_m} \mathsf{E}\Big[\hat{\theta}_\lambda^2 - 2\theta_\lambda \hat{\theta}_\lambda + \theta_\lambda^2 \Big] = \sum_{\lambda=1}^{d_m} \Big(\mathsf{E}\Big[\hat{\theta}_\lambda^2 \Big] - \theta_\lambda^2 \Big) \ ,$$

due to (2.3.3) and the unbiasedness of $\hat{\theta}_\lambda$. Moreover, since

$$\mathsf{Var}\Big[\hat{\theta}_\lambda \Big] = \mathsf{E}\Big[\hat{\theta}_\lambda^2 \Big] - \theta_\lambda^2 = \frac{1}{n} \mathsf{Var}[\, \varphi_\lambda(X_1) \,] \ ,$$

by (2.3.5), we obtain

$$\mathsf{E}[\, \chi^2 \,] = \frac{1}{n} \sum_{\lambda=1}^{d_m} \mathsf{Var}[\, \varphi_\lambda(X_1) \,] \ .$$

Thus, the risk decomposition (2.3.6) can be expressed as

$$\mathsf{E}\Big[\, ||\hat{f}_m - f_0||^2 \, \Big] = ||f_0 - \pi_m||^2 + \mathsf{E}[\, \chi^2 \,] \leqslant ||f_0 - \pi_m||^2 + \frac{1}{n} \mathsf{E}\left[\sum_{\lambda=1}^{d_m} \varphi_\lambda^2(X_1) \right] \ , \quad \text{(2.A.1)}$$

since

$$\frac{1}{n} \sum_{\lambda=1}^{d_m} \mathsf{Var}[\, \varphi_\lambda(X_1) \,] = \frac{1}{n} \sum_{\lambda=1}^{d_m} \mathsf{E}[\, \varphi_\lambda^2(X_1) \,] - \frac{1}{n} \sum_{\lambda=1}^{d_m} (\mathsf{E}[\, \varphi_\lambda(X_1) \,])^2 \ .$$

Additionally, if there exists a bound such that

$$\left|\left| \sum_{\lambda=1}^{d_m} \varphi_\lambda^2 \right|\right|_\infty = D_m \ ,$$

then it follows that

$$\mathsf{E}\Big[\, ||\hat{f}_m - f_0||^2 \, \Big] = ||f_0 - \pi_m||^2 + \mathsf{E}[\, \chi^2 \,] \leqslant ||f_0 - \pi_m||^2 + \frac{D_m}{n} \ .$$

This completes the proof. ■

Proof of Equation 2.4.4

We are not only going to derive 2.4.4, but we would also like to give another insight into (2.4.3). To be more precise, we will show that the optimal model $\hat{m}$ is the result of minimizing the corresponding risk in (2.3.6):

$$\hat{m} = \underset{m \in \mathcal{M}_n}{\mathsf{arg\,min}} \ \mathsf{E}\Big[\, ||\hat{f}_m - f||^2 \, \Big] \ .$$

According to (2.3.6), the risk can be decomposed into

$$\mathsf{E}\Big[\,||\hat{f}_m - f||^2\,\Big] = ||f - \pi_m||^2 + \mathsf{E}\Big[\,||\hat{f}_m - \pi_m||^2\,\Big]\ .$$

This can be further simplified by noting first that

$$\begin{aligned}||f - \pi_m||^2 &= \int (f - f_m)^2 = \int f^2 - 2\int f_m f + \int f_m^2 \\ &= ||f||^2 - 2\sum_{\lambda=1}^{d_m}\theta_\lambda \int \varphi_\lambda f + ||\pi_m||^2 = ||f||^2 - 2\sum_{\lambda=1}^{d_m}\theta_\lambda^2 + ||\pi_m||^2 \\ &= ||f||^2 - ||\pi_m||^2\ ,\end{aligned}$$

where we used

$$||\pi_m||^2 = \int \left(\sum_{\lambda=1}^{d_m}\theta_\lambda\varphi_\lambda\right)^2 = \sum_{\lambda=1}^{d_m}\theta_\lambda^2 \int \varphi_\lambda^2 = \sum_{\lambda=1}^{d_m}\theta_\lambda^2\ .$$

Second, note that

$$\begin{aligned}\mathsf{E}\Big[\,||\hat{f}_m - \pi_m||^2\,\Big] &= \mathsf{E}\left[\int (\hat{f}_m - \pi_m)^2\right] = \mathsf{E}\left[\int \hat{f}_m^2\right] - 2\int \mathsf{E}\Big[\,\hat{f}_m\,\Big]\,\pi_m + \int \pi_m^2 \\ &= \mathsf{E}\Big[\,||\hat{f}_m||^2\,\Big] - ||\pi_m||^2\ .\end{aligned}$$

such that

$$-||\pi_m||^2 = \mathsf{E}\Big[\,||\hat{f}_m - \pi_m||^2\,\Big] - \mathsf{E}\Big[\,||\hat{f}_m||^2\,\Big]\ .$$

Thus, the risk decomposition (2.3.6) reads as

$$\begin{aligned}\mathsf{E}\Big[\,||\hat{f}_m - f||^2\,\Big] &= ||f||^2 - ||\pi_m||^2 + \mathsf{E}\Big[\,||\hat{f}_m - \pi_m||^2\,\Big] \\ &= ||f||^2 - \mathsf{E}\Big[\,||\hat{f}_m||^2\,\Big] + 2\,\mathsf{E}\Big[\,||\hat{f}_m - \pi_m||^2\,\Big]\ .\end{aligned}$$

Finally, according to (2.A.1), we end up the risk decomposition

$$\mathsf{E}\Big[\,||\hat{f}_m - f||^2\,\Big] = ||f||^2 - \mathsf{E}\Big[\,||\hat{f}_m||^2\,\Big] + 2\,\mathsf{E}\big[\,\chi^2\,\big]\ ,$$

where

$$\mathsf{E}\big[\,\chi^2\,\big] = \frac{1}{n}\sum_{\lambda=1}^{d_m}\mathsf{Var}\big[\,\varphi_\lambda(X_1)\,\big]\ .$$

Since $||f||^2$ is irrelevant for the minimization (2.4.3), it follows that

$$\hat{m} = \operatorname*{\mathsf{arg\,min}}_{m\in\mathcal{M}_n}\left\{-\mathsf{E}\Big[\,||\hat{f}_m||^2\,\Big] + 2\,\mathsf{E}\big[\,\chi^2\,\big]\right\}\ .$$

Then, (2.4.3) and (2.4.4) are obtained by substituting the empirical counterparts

$$\hat{m} = \underset{m \in \mathcal{M}_n}{\mathsf{arg\,min}} \left\{ - \sum_{\lambda=1}^{d_m} \hat{\theta}_\lambda^2 + \frac{2}{n^2} \sum_{i=1}^{n} \sum_{\lambda=1}^{d_m} \varphi_\lambda^2(X_i) \right\} .$$

Proof of Theorem 2.12

This result was originally proposed by Birgé and Rozenholc (2006) for the problem of constructing a histogram estimator for an unknown density function. The crucial point in their proof is the use of a lemma which provides a bound and a limit on a moment of a binomial random variable.

Due to the σ-finiteness of the Lévy measure (or its Radon-property) and the compactness of its support, the estimation problem resembles the problem in Reynaud-Bouret (2003), where the intensity function of an inhomogeneous Poisson process was estimated via sieve based model selection. More precisely, in this case, the mean measure is finite and allows for a normalization of the Lévy density such that Lemma 1 of Birgé and Rozenholc (2006) applies.

A way to circumvent this lemma might be to follow van de Geer (1995a) who derived probability bounds for the Hellinger loss used in maximum likelihood estimation of general counting processes.

Miscellanea

Definition 2.15 (ϵ-Entropy) *Let $(\mathcal{X}, d)$ be a (semi-)metric space. For $\epsilon > 0$, the ϵ-covering number $N(\epsilon, \mathcal{X})$ is defined as the number of balls with radius ϵ necessary to cover $\mathcal{X}$, i.e., the cardinality of the smallest set, say X, such that, for all $x \in \mathcal{X}$,*

$$\min_{x_i \in X} d(x_i, x) \leqslant \epsilon .$$

A collection X satisfying the above condition is called an ϵ-covering set. The ϵ-entropy of $\mathcal{X}$ is

$$H(\epsilon, \mathcal{X}) := \log N(\epsilon, \mathcal{X}) .$$

Let $N(\epsilon, \mathcal{X}) = \infty$ if no such finite set X exists.

If $\mathcal{X}$ is not bounded, then we consider the entropy o a ball around some fixed $x_0 \in \mathcal{X}$.

Definition 2.16 (Local Entropy) *Let $B(x_0, \sigma) = \{x \in \mathcal{X} : d(x, x_0) \leqslant \sigma, \sigma > 0\}$ be a*

ball around x_0. The local entropy is defined by

$$H(\epsilon;\sigma) := H\big(\epsilon, B(x_0,\sigma)\big) \ .$$

Chapter 3

Nonparametric Estimation via Wavelets

3.1 Motivation & Definitions

This section introduces basic concepts and notions of wavelet analysis and motivates why it may be advantageous to use wavelets. The material is based on the monographs of Daubechies (1992), Meyer (1992), Ruch and van Fleet (2009), Walnut (2001), and Walter (1994).

Let us assume a target function $f \in \mathcal{F} = \mathcal{L}_2(\mathbb{R})$ and that there exists a nested sequence of closed subspaces $\{V_j\}_{j\in\mathbb{Z}}$ of $\mathcal{L}_2(\mathbb{R})$, i.e.,

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \subset \mathcal{L}_2(\mathbb{R}) \ ,$$

such that

$$\bigcap_{j\in\mathbb{Z}} V_j = \{0\} \qquad \text{and} \qquad \overline{\bigcup_{j\in\mathbb{Z}} V_j} = \mathcal{L}_2(\mathbb{R}) \ .$$

For the moment, let us focus on the **approximation space** with the so-called **resolution level** $j = 0$ which is defined as

$$V_0 := \left\{ v \in \mathcal{L}_2(\mathbb{R}) : v(x) = \sum_{k\in\mathbb{Z}} \alpha_{0,k}\phi_{0,k}(x) \right\} \ ,$$

where the set $\left\{\phi_{0,k}(x) := \phi(x-k) : k \in \mathbb{Z}\right\}$ forms an orthonormal basis of V_0, i.e., the integer translates of function ϕ span V_0. Under certain regularity conditions on ϕ to be laid out later on, this structure can be generalized to all approximation spaces in $\{V_j\}_{j\in\mathbb{Z}}$ by simple transformations on ϕ which also shows how the approximation spaces are interrelated: An orthonormal basis of the approximation space V_j with resolution level j is given by the set $\left\{\phi_{j,k}(x) := 2^{j/2}\phi(2^j x - k) : k \in \mathbb{Z}\right\}$. The orthogonal projection of $f \in \mathcal{L}_2(\mathbb{R})$ onto V_j is defined as

$$P_{V_j} f := \sum_{k\in\mathbb{Z}} \langle \phi_{j,k}, f\rangle \phi_{j,k} = \sum_{k\in\mathbb{Z}} \alpha_{j,k}\phi_{j,k} \ .$$

Up to now, this setup corresponds to an orthogonal projection on a finite-dimensional sieve of Section 2.3.

Besides the connection via ϕ, the actual framework may be shown to offer another relation between any pair of subspaces V_j and V_{j+1} in terms of a so-called 'residual space' or **detail space** W_j. To this end, let the subspace W_j be the orthogonal complement of V_j in V_{j+1}, i.e., $V_{j+1} = V_j \oplus W_j$ and $V_j \perp W_j$. This allows us to define a sequence $\{W_j\}_{j\in\mathbb{Z}}$ of detail spaces similar to $\{V_j\}_{j\in\mathbb{Z}}$ but with the important distinction that all

W_j's are mutually orthogonal. As before, we concentrate on a specific detail space with resolution level $j = 0$ which is defined as

$$W_0 := \left\{ w \in \mathcal{L}_2(\mathbb{R}) : w(x) = \sum_{k\in\mathbb{Z}} \beta_{0,k}\psi_{0,k}(x) \right\} ,$$

where the set $\{\psi_{0,k}(x) := \psi(x-k) : k \in \mathbb{Z}\}$ forms an orthonormal basis of W_0, i.e., the integer translates of function ψ span W_0. This can again be generalized to any W_j in $\{W_j\}_{j\in\mathbb{Z}}$ by noting that an orthonormal basis of the detail space W_j with resolution j is given by the set $\{\psi_{j,k}(x) := 2^{j/2}\psi(2^j x - k) : k \in \mathbb{Z}\}$. The orthogonal projection of $f \in \mathcal{L}_2(\mathbb{R})$ onto W_j is defined as

$$P_{W_j} f := \sum_{k\in\mathbb{Z}} \langle \psi_{j,k}, f\rangle \psi_{j,k} = \sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k} .$$

All of these considerations ultimately lead to what is known as the **wavelet decomposition** of a function $f \in \mathcal{L}_2(\mathbb{R})$, stated in terms of function spaces,

$$\begin{aligned} \mathcal{L}_2(\mathbb{R}) &= \cdots \oplus \underbrace{W_{j_0-1}}_{\substack{\text{detail}\\ \text{space}}} \oplus \underbrace{W_{j_0}}_{\substack{\text{detail}\\ \text{space}}} \oplus \underbrace{W_{j_0+1}}_{\substack{\text{detail}\\ \text{space}}} \oplus \cdots \\ &= \underbrace{V_{j_0}}_{\substack{\text{coarse}\\ \text{space}}} \oplus \underbrace{W_{j_0}}_{\substack{\text{detail}\\ \text{space}}} \oplus \underbrace{W_{j_0+1}}_{\substack{\text{detail}\\ \text{space}}} \oplus \cdots , \end{aligned} \tag{3.1.1}$$

or in terms of orthogonal projections,

$$\begin{aligned} f &= \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k} = \sum_{j\in\mathbb{Z}} P_{W_j} f \\ &= \sum_{k\in\mathbb{Z}} \alpha_{j_0,k}\phi_{j_0,k} + \sum_{j\geqslant j_0}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k} = P_{V_{j_0}} f + \sum_{j\geqslant j_0} P_{W_j} f . \end{aligned} \tag{3.1.2}$$

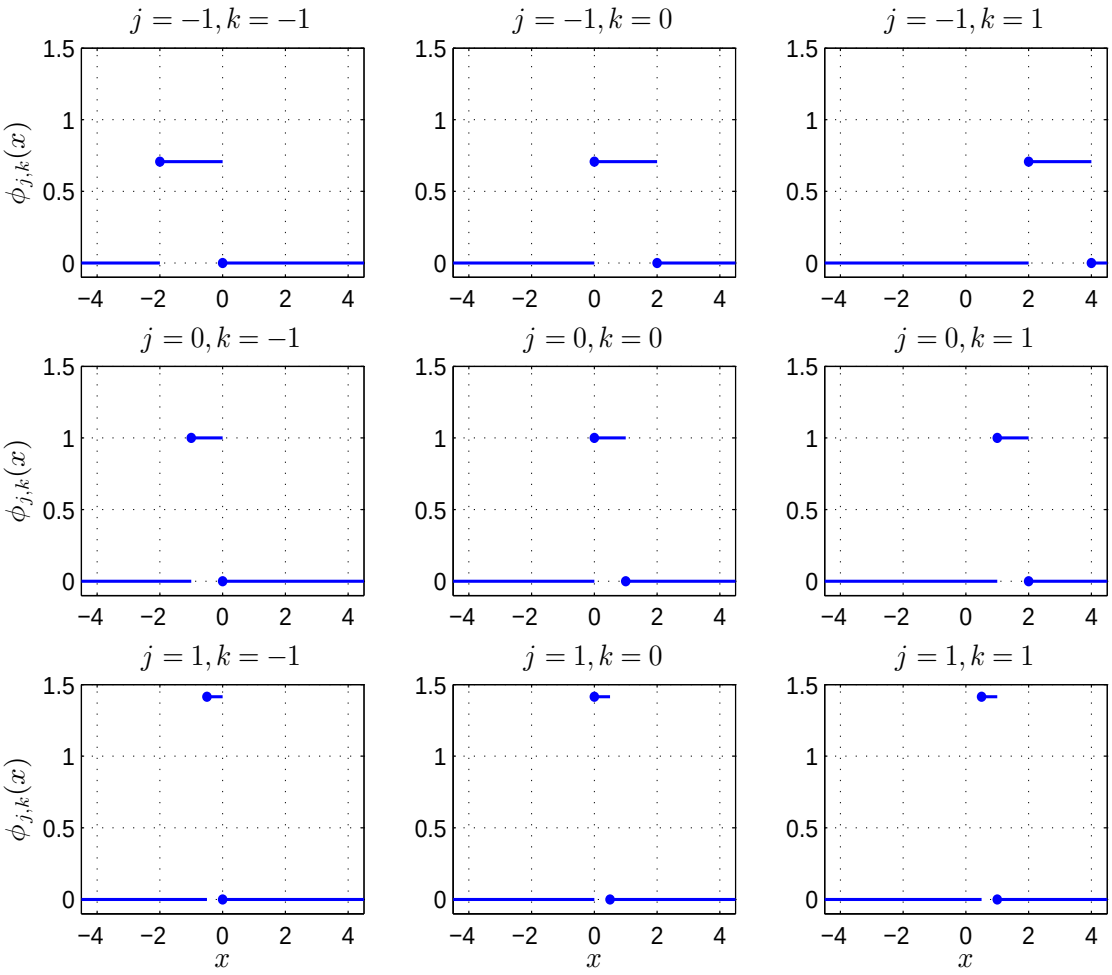
As already mentioned, applying $\{V_j\}_{j\in\mathbb{Z}}$ is essentially nothing more than a sieve approximation. Thus, one might ask what are the merits of wavelet-based approximation? For answering this question, it is convenient to illustrate the effects of transforming ϕ and ψ using the simplest basis functions. The **Haar scaling function** is defined as

$$\phi(x) := \begin{cases} 1 & \text{for } 0 \leqslant x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3.1 depicts what happens when the indices j and k are varied. Let us first look at

$\phi_{0,0}$ as the benchmark cases. Varying k while fixing $j = 0$, we see that the basis functions $\phi_{0,k}$ are translated, i.e., shifted, along the x-axis which, in the wavelet literature, is traditionally termed as "time." Next, fixing $k = 0$ and varying j, we recognize that the basis functions $\phi_{j,0}$ are locked-in at $k = 0$, while their support and amplitude are changing. In the wavelet literature, the ordinate is called the "scale" or "frequency."

Figure 3.1: *Effects of j and k on Shape of Haar Scaling Function*



Turning back to the wavelet decomposition in (3.1.2), it is obvious that the projection $P_{V_{j_0}} f$ fixes a baseline resolution level j_0 (or row) in Figure 3.1 such that it rules out any variation in the scale. This explains why ϕ is usually called the **scaling function** (or **father wavelet**). The rationale of using a baseline space V_{j_0} is that it provides a lower truncation of the infinite sum $\sum_{j \in \mathbb{Z}} P_{W_j}$ in (3.1.2).

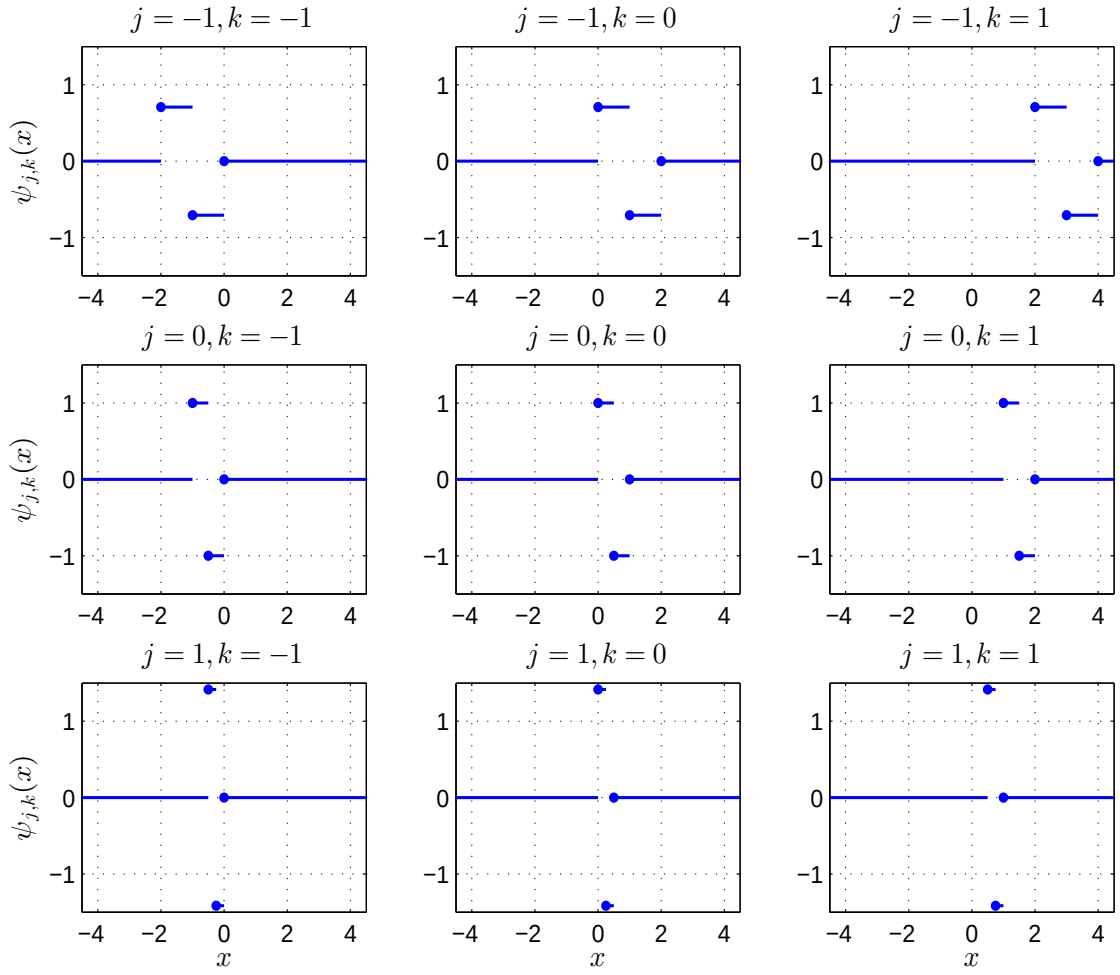
In contrast to standard orthogonal projections, the value-added of wavelet approxima-

tions is illustrated by introducing the **Haar wavelet function**:

$$\psi(x) := \begin{cases} 1 & \text{for } 0 \leqslant x < 0.5 \\ -1 & \text{for } 0.5 \leqslant x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3.2 depicts the effects of varying j and k, which turns out to be qualitatively analogous to the analysis of the scaling function.

Figure 3.2: *Effects of j and k on Shape of Haar Wavelet Function*



Turning back to the wavelet decomposition in (3.1.2), it is obvious that, starting from the baseline space V_{j_0}, the wavelet projection $P_{W_j} f$ with resolution levels $j \geqslant j_0$ add 'finer' function approximations to $P_{V_{j_0}} f$. This explains why V_{j_0} is called the **coarse space** while the W_j's are called the **detail spaces** in (3.1.1).

This insight gained from the analysis of the Haar scaling and wavelet functions can be

generalized to any pair of scaling and wavelet functions:

$$\phi_{j,k} := 2^{j/2}\phi(2^j x - k) \tag{3.1.3}$$

$$\psi_{j,k} := 2^{j/2}\psi(2^j x - k)\ . \tag{3.1.4}$$

The basis functions $\phi_{j,k}$ and $\psi_{j,k}$ are scaled and translated versions of ϕ and ψ, respectively. Increasing the translation index k has the effect of shifting ψ on the x-axis from the left to the right. Increasing the resolution level j has two different effects: First, the factor 2^j compresses the support of ψ. This is often described by saying that "ψ is well localized in time." Second, the factor $2^{j/2}$ dilates the amplitude of ψ.

Before starting with a more formal treatment of wavelet analysis, let us close this motivating section by pointing out the merits of using wavelets: Spatial adaptivity and sparse representation. To this end, note that since we can only deal with finitely many terms in computations, the infinite double sums have to be truncated in such a way that the wavelet representation provides the desired degree of approximation to the target function f, i.e.,

$$f \approx \sum_{|j|\leqslant J} \sum_{|k|\leqslant K} \beta_{j,k}\psi_{j,k}\ ,$$

for sufficiently, large $J, K \in \mathbb{N}$.

Spatial adaptivity of wavelets means that the superposition of different wavelet functions $\psi_{j,k}$ with varying degree of localization allows the wavelet decomposition to 'pick up' diverse spatial inhomogeneities of f such as discontinuities, high oscillations, wiggles, kinks, cusps, etc.

As an illustration, let us consider the following example, Assume that a function f is very smooth at the location $x_1 = k_1$ while it is 'non-smooth' at the location $x_2 = k_2$. In this case, spatial adaptivity works as follows: The smooth part of f is only picked up by the low resolution levels, say j_0, such that $|\beta_{j_0,k_1}| > 0$ but $|\beta_{j,k_1}| \approx 0$ for all $j > j_0$. On the contrary, the non-smooth part of f is picked up by the high resolution levels such that $|\beta_{j,k_2}| > 0$ for some $j > j_0$. Thus, at a fixed location k_* and for different resolution levels j, the absolute values of the wavelet coefficients convey information on the regularity of f.

This example leads us directly to the notion of **sparse representation**. Once we have obtained the coefficients $\beta_{j,k}$ via the wavelet decomposition, it is a remarkable feature of wavelet analysis that many of the $\beta_{j,k}$'s are close to or equal to zero. Recall that, for high resolution levels j, only wavelet coefficients $\beta_{j,k}$ near inhomogeneities of f are nonzero. It should be noted, however, that this does not mean, in general, that J is small. Indeed,

it usually turns out that J is large, but that the number of nonzero wavelet coefficients is small.

Let us return to the exact version of the above approximation, and note that wavelets allow for the decomposition

$$f(x) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k}(x) \tag{3.1.5}$$

of any function $f \in \mathcal{L}_2(\mathbb{R})$.[1] Turning $\mathcal{L}_2(\mathbb{R})$ into a Hilbert space by defining the usual inner product, (3.1.5) then implies that

$$||f||^2 = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \left|\langle f, \psi_{j,k}\rangle\right|^2 \qquad \text{for all } f \in \mathcal{L}_2(\mathbb{R})$$

$$\langle \psi_{j,k}, \psi_{j',k'}\rangle = \begin{cases} 1 & \text{for } k = k' \text{ and } j = j' \\ 0 & \text{otherwise.} \end{cases}$$

An important result in wavelet theory is that the convergence of the wavelet expansion to f in the $\mathcal{L}_2$-norm is unconditional, i.e., the ordering of basis functions is irrelevant. This is due to the fact that wavelets constitute a Riesz, and this was pointed out by Donoho (1993) to be the exceptional spatial adaptivity and compression properties of wavelet.

Definition 3.1 (Multiresolution Analysis) *A multiresolution analysis consists of a sequence $\{V_j\}_{j\in\mathbb{Z}}$ of closed subspaces of $\mathcal{L}_2(\mathbb{R})$ and a function $\phi \in V_0$ satisfying the following conditions:*

C1 $\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots$

C2 $\overline{\bigcup_j V_j} = \mathcal{L}_2(\mathbb{R})$ *and* $\bigcap_j V_j = \{0\}$

C3 $f \in V_j \iff \big\{x \mapsto f(2x)\big\} \in V_{j+1}$

C4 $f \in V_0 \Longrightarrow \big\{x \mapsto f(x-k)\big\} \in V_0$ *for all* $k \in \mathbb{Z}$

C5 $\big\{\phi(\,\cdot\, - k)\big\}_{k\in\mathbb{Z}}$ *is an orthonormal basis for* V_0.

[1] Note that by a simple change in notation, this expansion may be expressed as

$$f = \sum_{\lambda\in\Lambda} \theta_\lambda \varphi_\lambda$$

which looks similar to the orthogonal projection (2.3.1). Interestingly, if $\Lambda = \mathcal{M}_n$ were a finite (or countable) collection of models as in Section 2.4, then it would now be non-nested. Moreover, this implies that the bound on the risk of the oracle includes an additional $\mathsf{ln}(n)$ factor which slows down the rate of convergence. As it turns out, this is a common phenomenon for nonlinear wavelet estimators.

The existence of a multiresolution analysis allows us to relate the basis functions $\phi_{j,k}$ and $\psi_{j,k}$ such that they can recursively computed. This is important since most wavelets do not have closed forms. Another non-trivial, but for computations important, issue is the construction of compactly supported wavelets. This was accomplished by Daubechies (1988) by setting up the so-called db-family of wavelets which will be used in our applications.

3.2 Wavelet Estimators

For estimation based on wavelets, we refer to the monographs of Härdle, Kerkyacharian, Picard, and Tsybakov (1998), Ogden (1997), and Vidakovic (1999).

As already mentioned in the previous section, a multiresolution analysis allows us to rewrite the wavelet expansion (3.1.5) as

$$f = \sum_{k\in\mathbb{Z}} \alpha_{j_0,k}\phi_{j_0,k} + \sum_{j\geqslant j_0}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k} \ ,$$

defines a baseline, coarse space V_{j_0}. As $\sum_{k\in\mathbb{Z}} \alpha_{j_0,k}\phi_{j_0,k}$ and $\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k}$ are the orthogonal decompositions of V_{j_0} and W_j, respectively, we can apply the same reasoning as in Section 2.3 and define the estimators of the generalized Fourier coefficient as simple moment estimators: For a random sample of observations $\{X_i\}_{i=1}^n$, let

$$\begin{aligned} \hat{\alpha}_{j_0,k} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \phi_{j_0,k}(X_i) \\ \hat{\beta}_{j,k} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \psi_{j,k}(X_i) \end{aligned}$$

such that a wavelet estimator of f reads as

$$\hat{f} = \sum_{k\in\mathbb{Z}} \hat{\alpha}_{j_0,k}\phi_{j_0,k} + \sum_{j\geqslant j_0}\sum_{k\in\mathbb{Z}} \hat{\beta}_{j,k}\psi_{j,k} \ . \tag{3.2.1}$$

Unfortunately, this estimator is infeasible as it involved infinite sums. In order to resolve this issue, one first has to restrict the number of detail spaces involved by defining an upper truncation level j_n. Furthermore, the number of basis function for approximating

the spaces V_{j_0} and W_j is restricted by a finite dyadic decomposition:[2]

$$\hat{f} = \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_n}\sum_{k=0}^{2^j-1} \hat{\beta}_{j,k}\psi_{j,k} \; . \tag{3.2.2}$$

As this estimator depends linearly on the data, it is called a **linear wavelet estimator** of f.

The upper truncation parameter j_n in (3.2.2), which depends on the data, plays a similar role as the tuning (or smoothing) parameter, i.e., the bandwidth, in kernel density estimation. To see this, note that a large j_n includes high resolution detail spaces. If a smooth function is corrupted by noise, the basis functions of these high resolution spaces will pick up the oscillations due to noise. Thus, the estimate of the underlying function will be rough, i.e., it is undersmoothed. In order to get a smoother estimate, one is forced to decrease j_n.

For this problem, Donoho and Johnstone (1995) put forward a simple modification in order to work with high resolution levels, while optimally eliminating the noise component in the data. Since the noise is picked up by the corresponding $\hat{\beta}_{j,k}$'s, they derived two schemes to denoise the high resolution level coefficients. Both rely on the idea of thresholding. The first one is called **hard-thresholding**, which yields thresholded wavelet coefficients defined by $\eta_\lambda(\hat{\beta}_{j,k})$

$$\eta_\lambda^H(x) := \begin{cases} x & \text{if } |x| > \lambda \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda \geqslant 0$. The second one is called **soft-thresholding** and is based on the idea of Stein (1981)'s shrinkage procedure. There, the thresholded wavelet coefficients are defined by

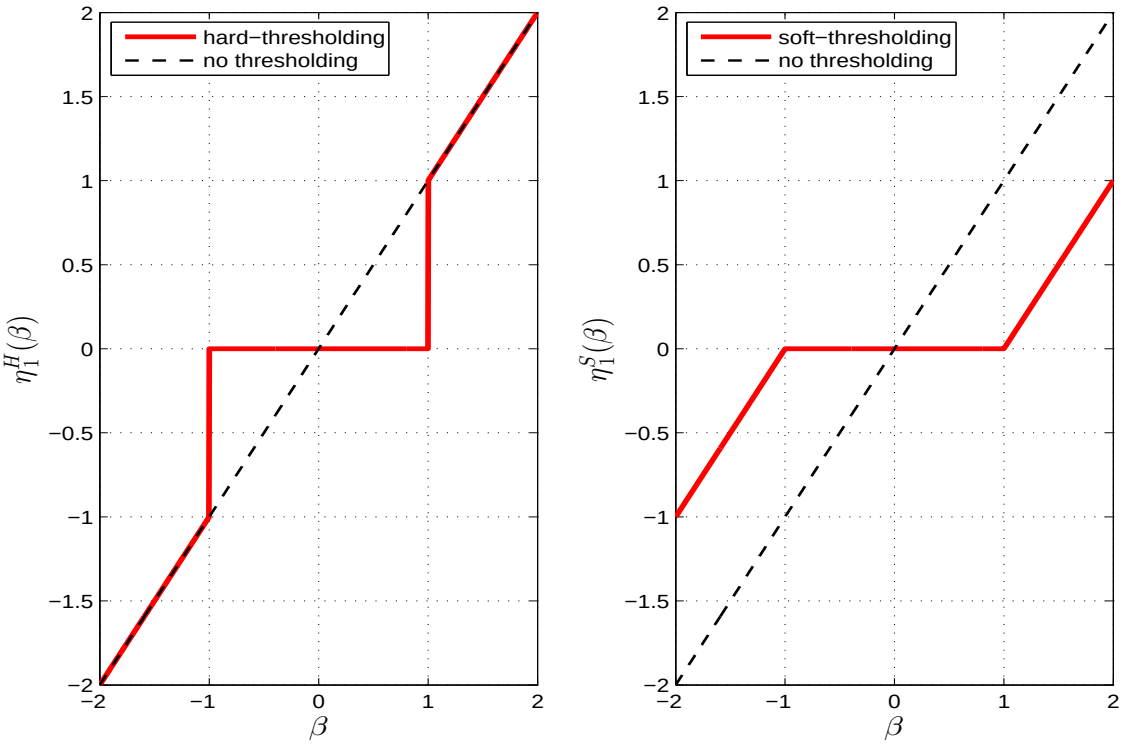
$$\eta_\lambda^S(x) := \begin{cases} x-\lambda & \text{if } x > \lambda \\ 0 & \text{if } |x| \leqslant \lambda \\ x+\lambda & \text{if } x < -\lambda. \end{cases}$$

The effects of these thresholding techniques are visualized in Figure 3.3

For the purpose of nonparametric density estimation via wavelets, Donoho, Johnstone, Kerkyacharian, and Picard (1996) derived a universal threshold level $\lambda = \sqrt{2\log n}$.

[2]On the one hand, dyadic decompositions derive from the so-called atomic decomposition of function space which are the precursor of the wavelet decomposition (Triebel, 1992, 2008). On the other hand, dyadic decompositions allow for the implementation of fast and efficient algorithms for computing of wavelet coefficients.

Figure 3.3: *Hard- vs. Soft-Thresholding Rules*



A problem of the above term-by-term thresholding schemes is that their rates of convergence is, in general, slowed by a $\ln(n)$ factor compared to the minimax rate. Thus, there were many attempts to alleviate this problem. For example, Kerkyacharian, Picard, and Tribouley (1996) proposed a soft-thresholding scheme, not term-by-term, but **levelwise**. To be more precise, if

$$\theta_j = \sum_{k \in \mathbb{Z}} |\beta_{j,k}|^2$$

denotes the 'energy' of the resolution level j, then the soft-thresholding is defined by:

$$\eta_j(\hat{\theta}_j) = \begin{cases} \dfrac{\hat{\theta}_j - 2^{j/n}}{\hat{\theta}_j} & \text{if } \hat{\theta}_j \geqslant 2^{j/n} \\ 0 & \text{otherwise.} \end{cases}$$

The rational why this scheme should provide more efficient estimates is that more 'information' is pooled for deciding whether to delete coefficients or to shrink them. The authors indeed showed that it is possible to get rid of the $\ln(n)$-factor.

The nonlinear wavelet estimator that we propose for nonparametrically estimating a Lévy density is a hybrid of term-by-term and levelwise thresholding and is called **block-**

thresholding. It was introduced by Hall, Kerkyacharian, and Picard (1998) and refined by Chicken and Cai (2005) for density estimation. Cai (1999) provided an oracle inequality. Here, the idea is to divide the wavelet coefficients in every resolution level into non-overlapping blocks of length $l = \mathsf{ln}(n)$. Then, hard-thresholding will be performed with respect to the estimated squared bias

$$\hat{B}_{i,k} = \frac{1}{l} \sum_{j \in B(k)} \hat{\beta}_{i,j}^2 \ ,$$

where $B(k)$ is the set of indices j contained in block k. To be more precise, the wavelet coefficients are kept, if $\hat{B}_{i,k}$ is larger than a threshold level, otherwise they are all deleted Hence, the block-thresholded estimator reads as

$$\hat{f}(x) = \sum_j \hat{\alpha}_j \phi_j(x) + \sum_{i=0}^{R} \sum_k \sum_{j \in B(k)} \hat{\beta}_{i,j} \psi_{i,j}(x) \mathbb{1}_{\{\hat{B}_{i,k} > cn^{-1}\}} \ ,$$

where $R = \left\lfloor \mathsf{log}_2(Dnl^{-1}) \right\rfloor$. Note that the introduction of blocks $B(k)$ led to a slight change of notation. For the exact calibration of D and c, we refer to Chicken and Cai (2005). Moreover, Theorem 1 of Chicken and Cai (2005) proves adaptation in the minimax sense of the block-thresholded estimator. We note that this theorem should also be valid for Lévy density estimation, due to the σ-finiteness of ν. Then, the only change necessary would be the replacement of the concentration inequality of Talagrand (1994) by an appropriate concentration for compensated Poisson processes of Reynaud-Bouret (2003) who considered nonparametric estimation of the intensity of inhomogeneous Poisson processes. However, the technical details of this proof are left for future research.

We close this chapter by noting that wavelet estimators have been developed within the nonparametric regression framework. Thus, they are not directly applicable to density estimation. Recall from Section 2.1 that we have noisy observations Y_i of f which we used in our estimator $\hat{f}$. However, observations Y_i drawn from a density function f do not correspond to noisy observations of f. Instead, we first have to construct these noisy observations from the sampled data. This is usually done by estimating a histogram based on the observations Y_i. The bin midpoints of this histogram are then interpreted as noisy observations of the unknown density function f. Michael Nussbaum was the first to advocate this approach in the discussion of Donoho, Johnstone, Kerkyacharian, and Picard (1995) by pointing out to an approximation results based on Haar functions of Koltchinskii (1994).

It is noteworthy that this approach is prone to a bias-variance trade-off. Clearly, more bins provide more 'noisy' observations to be used for the wavelet density estimator. These

will, however, be less efficient estimates (more noisy). This motivates the usage of the histogram estimator based on model selection from Section 2.6.

Chapter 4

Simulations & Applications

4.1 Variance Gamma Processes

In this section, we consider the variance gamma process as an example of a random time-changed Brownian motion. As it turns out, the variance gamma process has three different representations. Two of these involve gamma subordinators as building blocks. Thus,, we recap some basic facts about gamma random variables and processes at this point. A gamma random variable X has probability density function

$$f_X(x;\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)}\, x^{\alpha-1}\mathsf{e}^{-x/\beta}\mathbb{1}_{\{x>0\}}\ , \tag{4.1.1}$$

where $\alpha > 0$ and $\beta > 0$ are interpreted as shape parameter and scale parameter, respectively, and characteristic function

$$\Phi_X(u) = (1 - \mathsf{i}u\beta)^{-\alpha}\ ,$$

which is derived in (A.1.8).

The interpretations of α and β are partially due to the following important properties of gamma random variables:

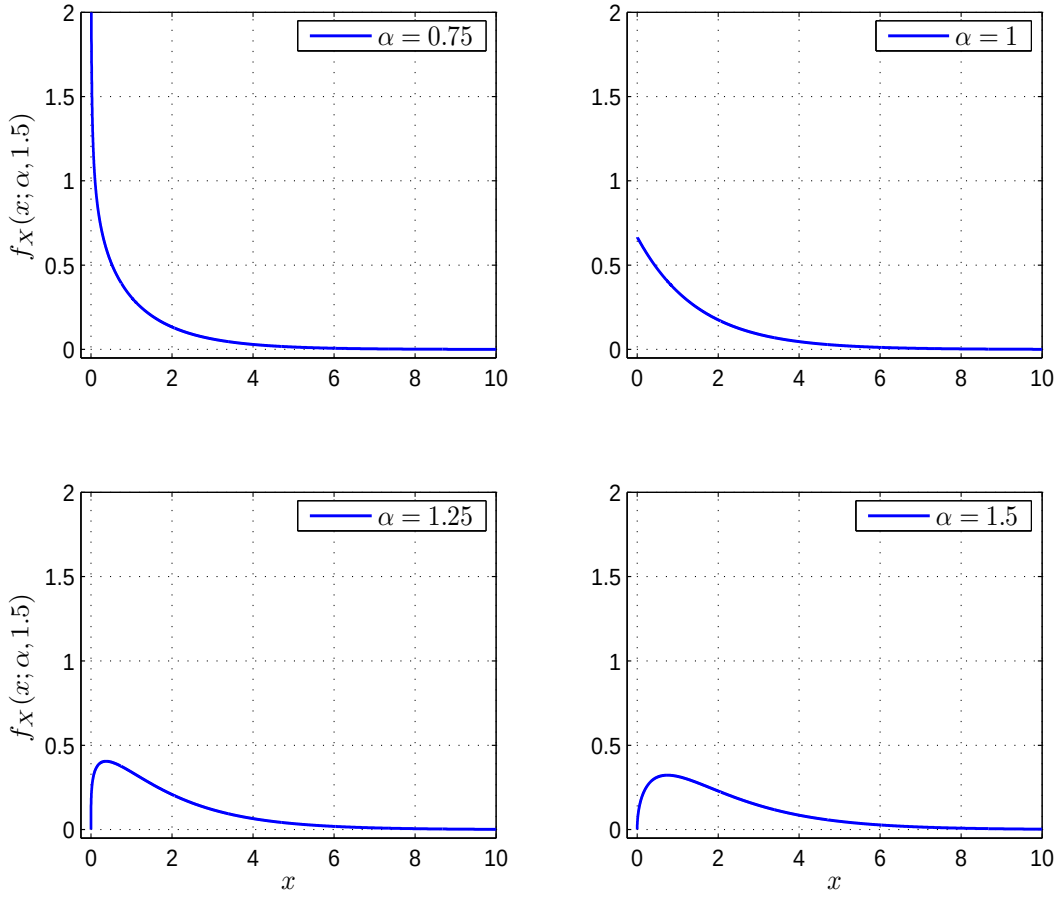
1. **Additivity of gamma random variables:** Let $X_1, \ldots, X_n$ be independent gamma random variables with respective probability density functions $f_{X_i}(x;\alpha_i,\beta)$ for $i = 1,\ldots,n$. Then, $Y = \sum_{i=1}^n X_i$ is a gamma random variable with probability density function
$$f_Y\left(y; \sum_{i=1}^n \alpha_i, \beta\right)\ . \tag{4.1.2}$$

2. **Scaling of gamma random variables:** Let X be a gamma random variable with probability density function $f_X(x;\alpha,\beta)$. If $c > 0$, then $Y = cX$ is a gamma random variable with probability density function
$$f_Y(y;\alpha,c\beta)\ . \tag{4.1.3}$$

From the moments of gamma random variable, derived in Appendix A.1, and Figures 4.1 and 4.2, we deduce that the its probability density is skewed to the right, is strictly decreasing for $0 < \alpha \leqslant 1$, and has a maximum at $x = (\alpha-1)\beta$ for $\alpha > 1$.

A **gamma process** is a Lévy process with gamma distributed increments. From the form (A.1.10) of the characteristic function of a gamma random variable X and the additivity property (4.1.2), we can immediately obtain the characteristic function of a

Figure 4.1: *Effect of Shape Parameter on Gamma Density Function*
This figure depicts the effect of increasing the shape parameter α on the shape of a gamma probability density function $f_X(x;\alpha,\beta)$, while the scale parameter $\beta = 1.5$ is kept constant. Due to the additivity property of gamma (4.1.2) and (4.1.6), this corresponds exactly to the time evolution of the marginal density of a gamma process $\{X_t\}_{t\geqslant 0}$.



gamma process $\{X_t\}_{t\geqslant 0}$

$$\Phi_{X_t}(u) = (1 - \mathsf{i}u\beta)^{-\alpha t} = \exp\left\{ t \int_0^\infty (\mathsf{e}^{\mathsf{i}ux} - 1)\, \frac{\alpha}{x}\, \mathsf{e}^{-x/\beta}\, \mathsf{d}x \right\} ,$$

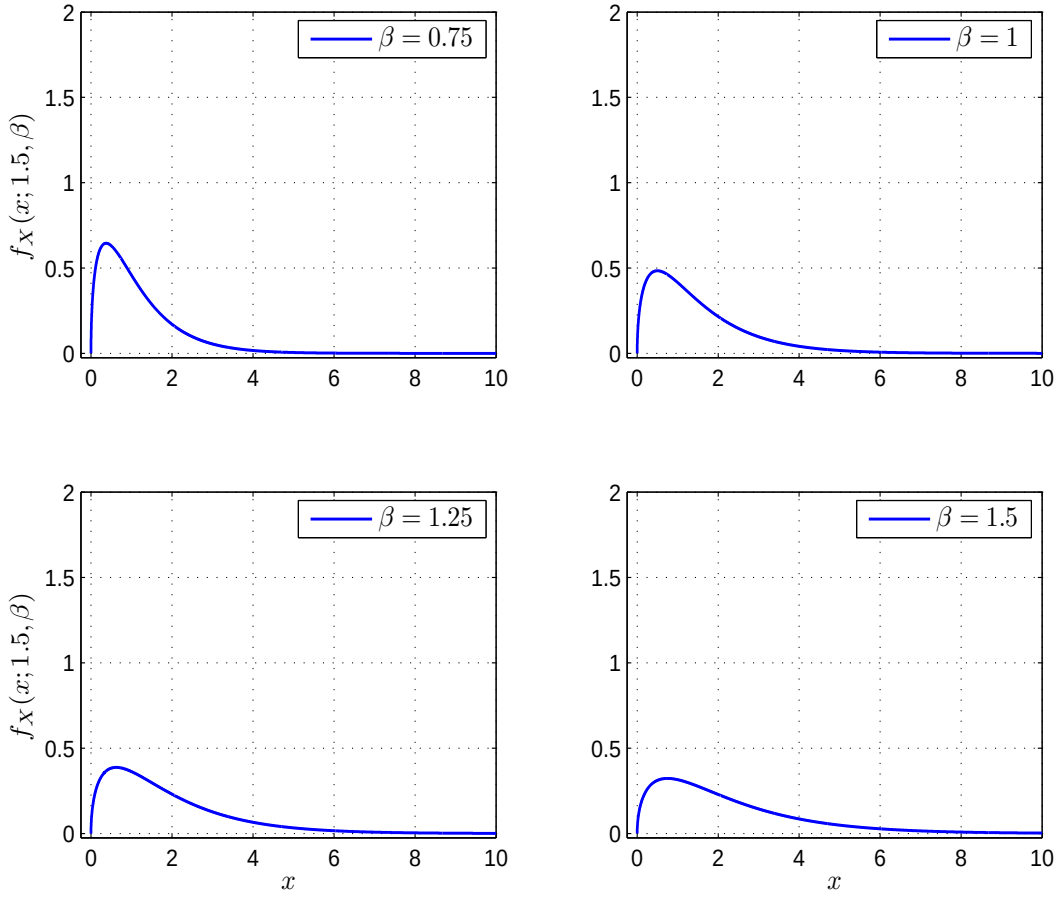
with characteristic triplet $(0, 0, \nu)$ and Lévy measure

$$\nu_\Gamma(\mathsf{d}x) = \frac{\alpha}{x}\, \mathsf{e}^{-x/\beta} \mathbb{1}_{\{x>0\}}\, \mathsf{d}x \, . \tag{4.1.4}$$

Put differently, the gamma process is a pure-jump Lévy process.

Following the literature on gamma processes and for reasons to be explained in Section 4.1, we set $\alpha = c$ and $\beta = 1/\lambda$. Then, it follows from Conditions C1 and C3 of

Figure 4.2: *Effect of Scale Parameter on Gamma Density Function*
This figure depicts the effect of increasing the scale parameter β on the shape of a gamma probability density function $f_X(x;\alpha,\beta)$, while the shape parameter $\alpha = 1.5$ is kept constant. Clearly, all subplots exhibit a mode.



Definition 1.1 that, for all $t > 0$,

$$\Delta X_t \stackrel{d}{=} \Delta X_1 = X_1 - X_0 = X_1 \stackrel{d}{=} \mathsf{Gam}(c, 1/\lambda)\ , \tag{4.1.5}$$

which allows us to easily simulate the increments of a gamma process from $\mathsf{Gam}(c, 1/\lambda)$. Moreover, due to the scaling property (4.1.3), it is sufficient to simulate

$$\Delta X_t \stackrel{d}{=} \mathsf{Gam}(c, 1)$$

in order to obtain $\mathsf{Gam}(c, 1/\lambda)$-increments because, then,

$$\Delta X_t/\lambda \stackrel{d}{=} \mathsf{Gam}(c, 1/\lambda)$$

for all $t > 0$. Figure 4.3 illustrates s some sample paths of gamma processes (4.1.5). Since

the increments of a gamma process are positive, its sample paths are non-decreasing such that, by Definition 1.14, a gamma process is indeed a subordinator.

Figure 4.3: *Sample Paths of Gamma Processes*
This figure depicts simulated paths of the gamma process (4.1.5) for varying shape parameter $\alpha = c$ and scale parameter $\beta = 1/\lambda$. The step size (or simulation frequency) is fixed at $\Delta t := t_i - t_{i-1} = 0.02$ for all $i = 1, \ldots, 50$ with $t_0 = 0$. The same seed for pseudo-random number generation is used along panel rows. The sampling interval is scaled down to $[0, 1]$. There are two important points to note: First, increasing c increases the overall jump activity of a sample path. Second, λ has only a scaling effect, leaving the overall jump activity unaffected, i.e., increasing λ solely scales down the sizes of jumps. The latter is perfectly in line with Figure 4.2, where increasing β leads to a heavier tail.

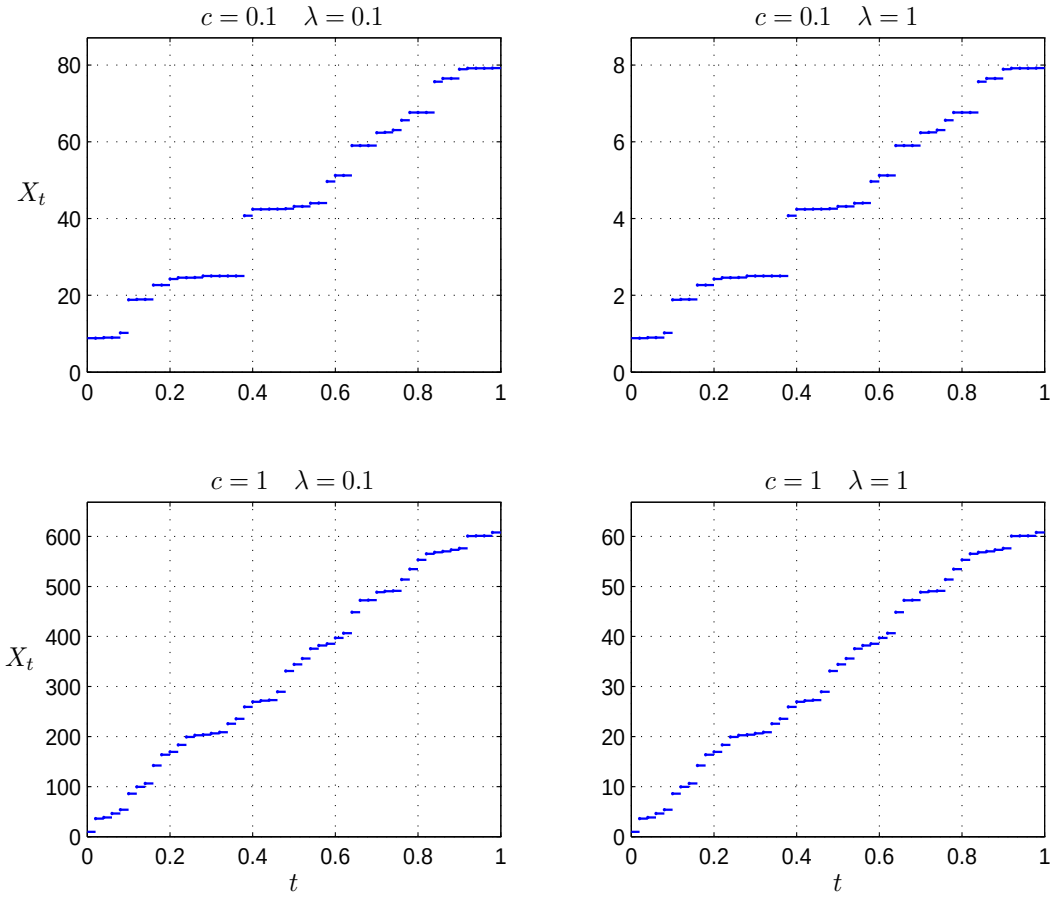


Figure 4.1 along with the additivity property (4.1.2) of gamma random variables allows us to infer the time evolution of the marginal density of a gamma process. Moreover, the moments of a gamma random variables, derived in Appendix A.1, yield the first four moments of a gamma process $\{X_t\}_{t \geqslant 0}$ with $\mathsf{Gam}(c, 1/\lambda)$-increments,

$$\mathsf{E}[\, X_t \,] = \frac{c}{\lambda}\, t$$
$$\mathsf{E}[\, X_t \,] = \frac{c}{\lambda^2}\, t$$

$$\mathsf{skewness}[X_t] = \frac{2}{\sqrt{ct}}$$
$$\mathsf{excess\ kurtosis}[X_t] = \frac{6}{ct} ,$$

and describe how they change in the course of time.

From the Lévy measure of a gamma process on (4.1.4), we obtain the Lévy density of a gamma process:

$$p_\Gamma(x) = \frac{\nu_\Gamma(\mathsf{d}x)}{\mathsf{d}x} = \frac{c}{x}\,\mathsf{e}^{-\lambda x}\mathbb{1}_{\{x>0\}} . \tag{4.1.6}$$

Clearly, any gamma process has infinity activity since

$$\int_{\{x>0\}} \nu_\Gamma(\mathsf{d}x) = \int_{\{x>0\}} p_\Gamma(x)\,\mathsf{d}x = \infty .$$

At the same time, any gamma process has finite variation since

$$\int_{\{0<x\leqslant 1\}} x\nu_\Gamma(\mathsf{d}x) = \int_{\{0<x\leqslant 1\}} x p_\Gamma(x)\,\mathsf{d}x < \infty .$$

Hence, according to Definition 1.7, any gamma process is a type-B Lévy process. Figure 4.4 depicts the Lévy densities underlying the simulated gamma processes in Figure 4.3.

Figures 4.3 and 4.4 lead to the following interpretations of parameters c and λ: c governs the overall arrival rate of jumps, while λ governs the arrival rate of large jumps.[1]

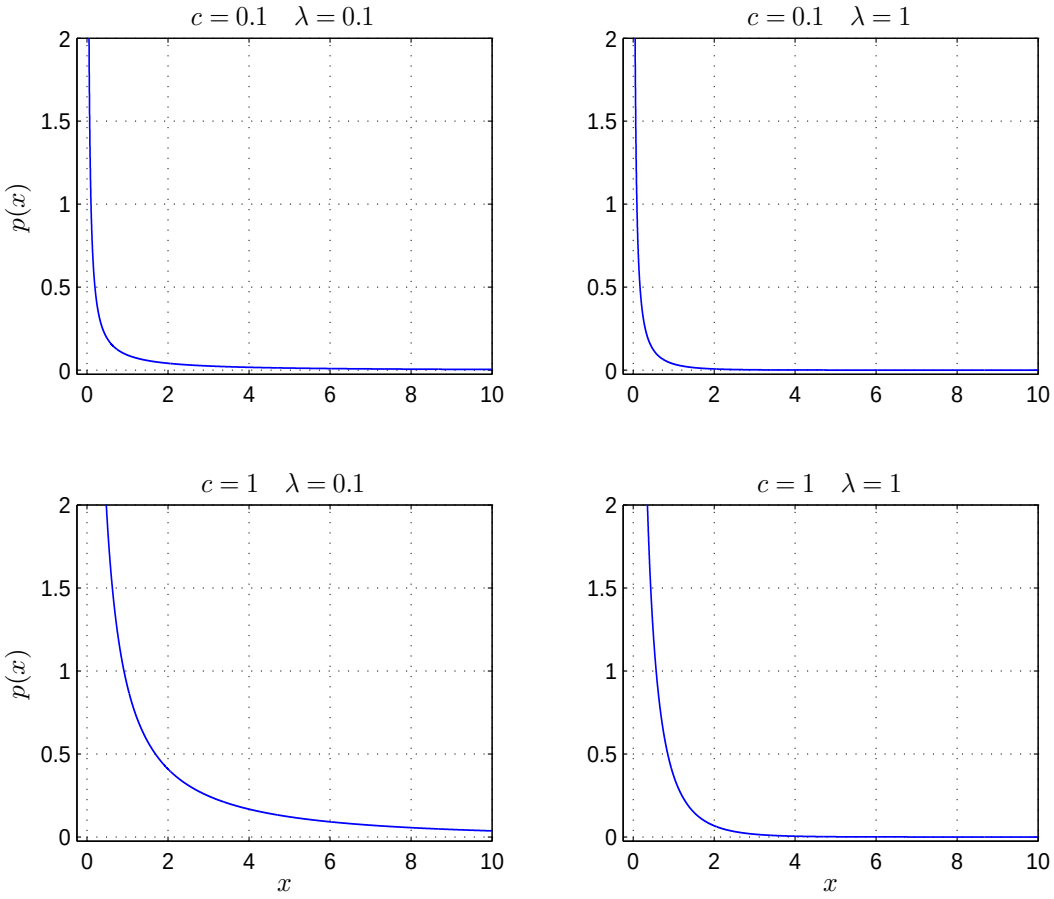
The **variance gamma process** was introduced for the first time by Madan and Seneta (1990). Further important properties and applications of the variance gamma process were analyzed by Madan, Carr, and Chang (1998) and Geman, Madan, and Yor (2001). As it turned out, the variance gamma process can be cast in three different representations, each of which emphasizes a distinctive feature.

The first representation of a variance gamma process X_t is to subordinate a Brownian motion to a random time change by a gamma process (4.1.5), i.e.,

$$X_t = B_{\Gamma_t} = \gamma\Gamma_t + \sigma W_{\Gamma_t} , \tag{4.1.7}$$

[1]Note that the gamma distribution is a popular choice when it comes to modeling inter-arrival times of jump processes. To see this, set $\alpha = 1$ and $\beta = \lambda^{-1}$ in (4.1.1): $f_X(x; 1, \lambda^{-1}) = \lambda\mathsf{e}^{-\lambda}\mathbb{1}_{\{x>0\}}$, i.e., $X \overset{d}{=} \mathsf{Exp}(\lambda)$. From Appendix A.1, the inter-arrival times between two consecutive jumps of a Poisson process $N_t \overset{d}{=} \mathsf{Poi}(\lambda t)$ is exponentially distributed with mean rate of occurrence per unit of time, i.e., intensity, $\lambda = \mathsf{E}[\,N_t\,]/t$. Consequently, the nth arrival time (or the arrival time of the nth jump) follows the law $\mathsf{Gam}(n, \lambda^{-1})$. However, the interpretations of the gama parameters differ from those laid out in Figure 4.4 since they model jump times and not jump sizes. Thus, these two notions should be not be mixed up when interpreting the model parameters.

Figure 4.4: *Levy Densities of Gamma Processes*
This figure depicts the Lévy density (4.1.6) of the gamma process (4.1.5) for different parameter values of c and λ.
The effects of changing c and λ are completely in line with the discussion of Figure 4.3. On the one hand, increasing c increases the overall jump activity of the gamma process as the area under the Lévy density p is scaled up. On the other hand, increasing λ decreases the intensity of large jumps as the tail of the Lévy density p decays at a faster rate.



where B_t is the Brownian motion with drift γ in (1.1.3) and $\{\Gamma_t\}_{t\geqslant 0}$ is the gamma process defined by (4.1.5) satisfying the important parameter restriction

$$\Gamma_t - \Gamma_{t-\Delta t} \stackrel{d}{=} \mathsf{Gam}(\Delta t/c, c) \ ,$$

or, by using Definition 1.1,

$$\Delta\Gamma_t := \Gamma_t - \Gamma_{t-1} = \Gamma_1 \stackrel{d}{=} \mathsf{Gam}(1/c, c) \ .$$

For the general form of a variance gamma process the parameters are defined as $\gamma \in \mathbb{R}$ and $\sigma, c > 0$.

The reason for reducing the number of parameters governing the gamma subordinator becomes immediately obvious when considering the mean and variance of its increments:

$$\begin{aligned}\mathsf{E}[\,\Gamma_t - \Gamma_{t-\Delta t}\,] &= \mathsf{E}[\,\mathsf{Gam}(\Delta t/c, c)\,] = \Delta t \\ \mathsf{Var}[\,\Gamma_t - \Gamma_{t-\Delta t}\,] &= \mathsf{Var}[\,\mathsf{Gam}(\Delta t/c, c)\,] = \Delta t \cdot c\ ,\end{aligned}$$

due to the moments of a gamma process in Subsection 1.4.3. Put together, the mean of the increments is equal to the time step on which the increments are computed, which conveniently lends itself to interpreting the subordinator as stochastic time, where its randomness is completely determined by its parameter c. Recall that, in Subsection 1.4.3, we interpreted the subordinator as a measure of business time which (randomly) deviates from calender time t.

Since we know from (4.1.5) how to simulate the increments of the gamma subordinator Γ_t, it is straightforward to simulate the sample paths of the variance gamma process by recognizing

$$\begin{aligned}X_t - X_{t-1} &= \gamma(\Gamma_t - \Gamma_{t-1}) + \sigma W_{\Gamma_t - \Gamma_{t-1}} \\ \Delta X_t &= \gamma \Delta\Gamma_t + \sigma W_{\Delta\Gamma_t}\ ,\end{aligned}$$

where $W_{\Delta\Gamma_t} \overset{d}{=} \mathsf{N}(\,0, \Delta\Gamma_t\,) = \sqrt{\Delta\Gamma_t}\,\mathsf{N}(\,0, 1\,)$. For simulation schemes for various Lévy processes, see Cont and Tankov (2003, Chapter 6) and Schoutens (2003). Figure 4.5 illustrates simulated sample path of the variance gamma process (4.1.7). In particular, this figure shows how the Brownian motion and the variance gamma process are related. Note that from the mean and variance of the increments of the subordinator Γ_t, we see that

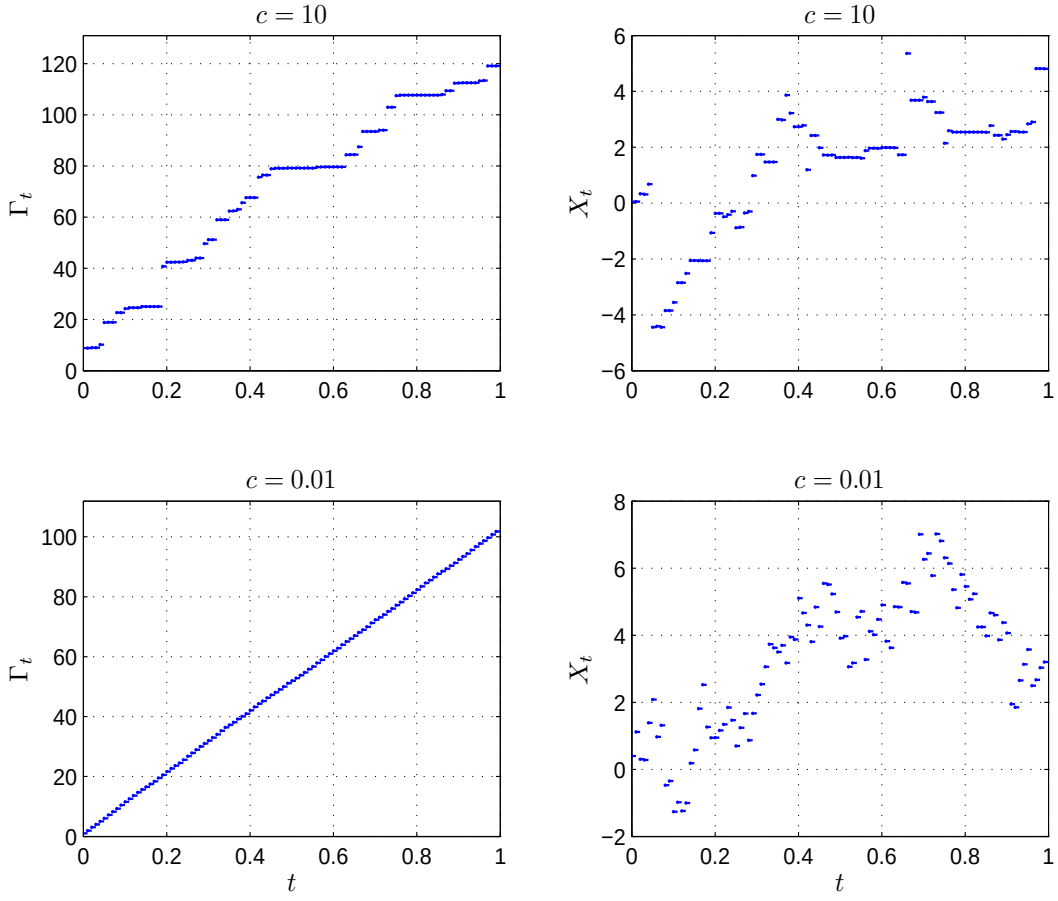
$$\begin{aligned}\mathsf{E}[\,\Gamma_t\,] &= t \\ \lim_{c \searrow 0} \mathsf{Var}[\,\Gamma_t\,] &= 0\ ,\end{aligned}$$

i.e., Γ_t corresponds to the deterministic business time $t \geqslant 0$. Thus, the Brownian motion is the limiting case of the variance gamma process as $c \searrow 0$. Moreover, the sample path behavior of a variance gamma process is lying in-between that of a compound Poisson process and a Brownian motion.

The random time-change representation of a variance gamma process,

$$X_t = B_{\Gamma_t}\ ,$$

Figure 4.5: *Sample Paths of Variance Gamma Processes & Subordinators*
This figure depicts simulated paths of the variance gamma process (4.1.7) and the respective gamma subordinators $\Gamma_t \stackrel{d}{=} \mathsf{Gam}(t/c, c)$ for differing parameter value c, while $\gamma = 0.02$ and $\sigma = 0.75$ are kept constant. The step size (or simulation frequency) is fixed at $\Delta t := t_i - t_{i-1} = 0.01$ for all $i = 1, \ldots, 100$ with $t_0 = 0$. The same seed for pseudo-random number generation is used along panel rows. The sampling interval is scaled down to $[0, 1]$.
From the discussion of the gamma process in Figures 4.3 and 4.4, we know that decreasing c increases the overall jump activity. Consequently, as subordinator Γ_t converges to t, the sample path of the corresponding variance gamma process resembles the sample path of the Brownian motion in Figure 1.1.



the characteristic exponent of a variance gamma process is easily derived by invoking Theorem 1.16:

$$\Psi_X = -\Psi_\Gamma^+ \circ (-\Psi_B) \,,$$

where the Laplace exponent

$$\Psi_\Gamma^+(z) = -\Psi_\Gamma(\mathsf{i}z) = \frac{1}{c}\,\mathsf{ln}(1 + zc)$$

follows from plugging in the characteristic function (A.1.8) of a gamma variable, i.e.,

$\Phi_{\Gamma_1} = (1 - \mathrm{i}uc)^{-1/c}$. Finally, since

$$\Psi_B(u) = \mathrm{i}u\gamma - \frac{u^2\sigma^2}{2} ,$$

it follows that

$$\Psi_X(u) = -\frac{1}{c} \ln\left(1 - \mathrm{i}u\gamma c + \frac{1}{2}\, u^2\sigma^2 c\right) . \tag{4.1.8}$$

This characteristic exponent, along with (A.1.2)–(A.1.5), allows us to derive, after some tedious calculations, the unit-time increments of a general variance gamma process with parameters γ, σ, and c:

$$\begin{aligned}
\mathsf{E}[\, X_1\,] &= \gamma \\
\mathsf{Var}[\, X_1\,] &= \sigma^2 + \gamma^2 c \\
\mathsf{skewness}[X_1] &= \frac{2\gamma^3\sigma^2 + 3\gamma\sigma^2 c}{(\sigma^2 + \gamma^2 c)^{3/2}} \\
\mathsf{excess\ kurtosis}[X_1] &= \frac{3\sigma^4 c + 6\gamma^4 c^3 + 12\gamma^2\sigma^2 c^2}{(\sigma^2 + \gamma^2 c)^2} .
\end{aligned}$$

Assuming $\sigma \neq 0$ in order to rule out deterministic variance gamma processes, these lower moments all exist and yield two interesting results: First, the skewness of the increments' law of a general variance gamma is determined by the drift parameter γ. A positive or negative drift γ implies that the increments' distribution is skewed to the right or left, respectively. Second, the increments' distribution of a variance gamma process is leptokurtic if its gamma subordinator is stochastic, i.e., $\mathsf{Var}[\,\Gamma_1\,] = c \neq 0$. This can be interpreted as excess kurtosis generated by stochastic volatility as c controls the jump activity. Moreover, the drift may also contribute to excess kurtosis.

Moreover, the characteristic exponent (4.1.8) yields after some cumbersome derivations the Lévy density of a general variance gamma process:

$$p_{VG}(x) = \frac{\nu_{VG}(\mathsf{d}x)}{\mathsf{d}x} = \frac{1}{c|x|} \exp\left(\frac{\gamma}{\sigma^2}\, x - \frac{1}{\sigma}\sqrt{\frac{2}{c} + \frac{\gamma^2}{\sigma^2}}|x|\right) . \tag{4.1.9}$$

Clearly, any variance gamma process has infinity activity since

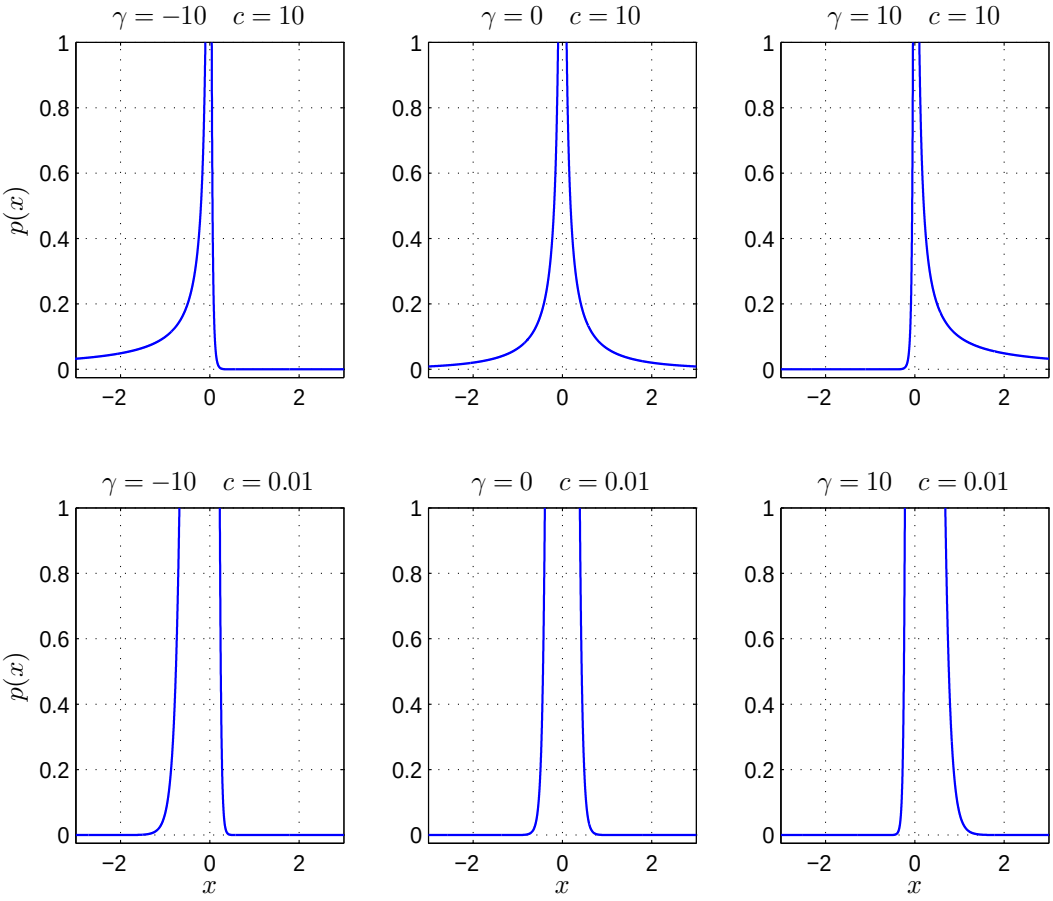
$$\int_{\{x>0\}} \nu_{VG}(\mathsf{d}x) = \int_{\{x>0\}} p_{VG}(x)\, \mathsf{d}x = \infty .$$

At the same time, any gamma process has finite variation since

$$\int_{\{0<x\leqslant 1\}} x\nu_{VG}(\mathsf{d}x) = \int_{\{0<x\leqslant 1\}} xp_{VG}(x)\,\mathsf{d}x < \infty\ .$$

Hence, according to Definition 1.7, any gamma process is a type-B Lévy process. Figure 4.6 depicts the Lévy densities of some variance gamma processes.

Figure 4.6: *Levy Densities of Variance Gamma Processes*
This figure depicts the Lévy density (4.1.9) of the variance gamma process (4.1.7) for different parameter values of γ and c, while $\sigma = 1$ is kept constant.
Note how skewness ($\gamma \neq 0$) shifts jump activity to the respective tail. Moreover, note how the overall jump activity increases as c decreases.



The last insight, e.g., the finite variation of any variance gamma process, leads to the second representation. Due to its finite variation, any variance gamma process can be represented as the difference two independent gamma processes, which are sometimes interpreted as gain and loss processes:

$$X_t = B_{\Gamma_t} = \gamma\Gamma_t + \sigma W_{\Gamma_t} = \Gamma_t^+ - \Gamma_t^-\ .$$

Due to the independence of Γ_t^+ and Γ_t^-, the characteristic function of the unit-time increments of X can be factorized as follows:

$$\begin{aligned} \Phi_{X_1}(u) &= \Phi_{\Gamma_1^+}(u)\Phi_{-\Gamma_1^-}(u) \\ \left(1 - \mathsf{i}u\gamma c + \frac{1}{2}\,u^2\sigma^2 c\right)^{-1/c} &= (1 - \mathsf{i}uM)^{-C}(1 - \mathsf{i}uG)^{-C}\,, \end{aligned}$$

where

$$\begin{aligned} C &:= 1/c > 0 \\ M &:= \sqrt{\frac{\gamma^2 c^2}{4} + \frac{\sigma^2 c}{2}} + \frac{\gamma c}{2} > 0 \\ G &:= \sqrt{\frac{\gamma^2 c^2}{4} + \frac{\sigma^2 c}{2}} - \frac{\gamma c}{2} > 0\,. \end{aligned}$$

This allows us to reconstruct the Lévy density (4.1.9) of a variance gamma process by merging the Lévy densities of the individual gamma processes. To this end, recall that, for $x > 0$,

$$\nu_{VG}(\mathsf{d}x) = \frac{C}{x}\,\mathsf{exp}\left(-\frac{x}{M}\right)\mathbb{1}_{\{x>0\}}\,\mathsf{d}x\,,$$

while we have

$$\nu_{VG}(\mathsf{d}x) = \frac{C}{|x|}\,\mathsf{exp}\left(-\frac{|x|}{G}\right)\mathbb{1}_{\{x<0\}}\,\mathsf{d}x\,,$$

for $x < 0$. Thus, we obtain

$$p_{VG}(x) = \begin{cases} \dfrac{C}{|x|}\,\mathsf{exp}\left(-\dfrac{|x|}{G}\right) & \text{for } x < 0 \\ \dfrac{C}{x}\,\mathsf{exp}\left(-\dfrac{x}{M}\right) & \text{for } x > 0. \end{cases} \tag{4.1.10}$$

Again, we note that increasing C increases the overall jump activity for both positive and negative jumps. Likewise, the parameters G and M measure the speed at which the arrival rate decays with the size of the jump.

A substantial extension of the variance gamma model was proposed by Carr, Geman, Madan, and Yor (2002, 2003), which is called the **CGMY model**. The CGMY model is able to generate both processes with finite or infinite activity and processes with finite and infinite variation. Surprisingly, this comes along with just a minor modification of

the variance gamma Lévy density (4.1.10):

$$p_{CGMY}(x) = \frac{\nu_{CGMY}(\mathsf{d}x)}{\mathsf{d}x} = \begin{cases} \dfrac{C}{|x|^{1+Y}} \exp\left(-\dfrac{|x|}{G}\right) & \text{for } x < 0 \\ \dfrac{C}{x^{1+Y}} \exp\left(-\dfrac{x}{M}\right) & \text{for } x > 0, \end{cases} \tag{4.1.11}$$

where $C, G, M > 0$, as for the variance gamma model, and $Y \in (-\infty, 1)$.

Similar to the discussion on the activity and variation of the variance gamma process following (4.1.9), it can be shown that the additional parameter Y allows us to classify CGMY processes according to Definition 1.7. A CGMY process has infinite activity for $-1 < Y < 1$, while it has infinite variation for $0 < Y < 1$. For $Y \leqslant -1$, the CGMY process has finite activity and finite variation such that it corresponds, for example, to a compound Poisson process or, more generally, to a Lévy process of type A. For $-1 < Y \leqslant 0$, the CGMY process has infinite activity and finite variation such that it corresponds, for example, to a variance gamma process ($Y = 0$) or, more generally, to a Lévy process of type B. For $0 < Y < 1$, the CGMY process has infinite activity and infinite variation such that it corresponds, for example, to a normal inverse Gaussian process or, more generally, to a Lévy process of type C. In order to satisfy integrability condition of the Theorem 1.3, $Y < 1$ is necessary. The CGMY process can also be represented as a random time change model (Madan and Yor, 2008).

Carr, Geman, Madan, and Yor (2002) performed goodness-of-fit testing on the S&P 500 index. Their favored model rejected any diffusion component, while a CGMY model of type B seemed to fit the data best.[2] Moreover, they showed that the CGMY model generates more realistic shapes of implied volatility.

The third representation of the variance gamma process essentially boils down to a very general scheme for simulating pure-jump Lévy processes, and which based upon the insight that any pure-jump Lévy process can be approximated by a sequence of compound Poisson processes. See Asmussen and Rosiński (2001) and Rosiński (2001).

Recall from Section 1.1 that the Lévy measure of a compound Poisson process is given by

$$\nu(\mathsf{d}x) = \lambda F(\mathsf{d}x) \, .$$

Since F is a probability distribution function with $F\big(\mathbb{R}\backslash\{0\}\big) = 1$, it follows that

$$\nu\big(\mathbb{R}\backslash\{0\}\big) = \lambda < \infty \, ,$$

[2]Note, however, that this result is a bit at odds with Aït-Sahalia and Jacod (2010) who found that there is a need for a continuous component for two equity stocks of the Nasdaq 100 composite index they considered.

due to the finite activity of any compound Poisson process. Next, define a sequence of compound Poisson processes with jump arrival rate

$$\lambda_n := \nu\big(\mathbb{R}\backslash[-1/n, 1/n]\big) = \int_{\mathbb{R}\backslash[-1/n,1/n]} \nu(\mathsf{d}x)$$

and jump size distribution

$$F(\mathsf{d}x) = \frac{\nu(\mathsf{d}x)}{\lambda_n} = \frac{\nu(\mathsf{d}x)}{\nu\big(\mathbb{R}\backslash[-1/n, 1/n]\big)} \ .$$

By passage to the limit, any pure-jump Lévy process with general Lévy measure ν can be arbitrarily well approximated by this sequence of compound Poisson processes (Sato, 1999, Corollary 8.8). As already discussed in Remark 2.11, this is the devise to handle the jump part in the proof of Theorem 1.6. In simulations, this procedure amounts to deleting jumps with absolute size smaller than $1/n$ and replacing them by a Brownian motion.

We now look at the results of a small simulation study where the simplest variance gamma model ($\gamma = 0$ and $\sigma = c = 1$),

$$p(x) = \frac{\nu(\mathsf{d}x)}{\mathsf{d}x} = \frac{1}{|x|}\, \mathsf{e}^{-\sqrt{2}|x|} \ ,$$

has been used to generate $M = 500$ trajectories, each with a sample size of $T = 5000$ returns. We then implement the block-thresholded wavelet estimator $\hat{p}$ of Section 3.2. This is done for various preliminary histogram estimators with a differing number of bins N and various wavelet bases (Daubechies, 1988) with differing degrees of smoothness. For each trajectory $m = 1, \ldots, M$ the integrated squared error

$$\mathsf{ISE}_m = \frac{1}{10^4} \sum_{i=1}^{10^4} \big[p(x_i) - \hat{p}(x_i)\big]^2 \ ,$$

via linear interpolation. That is, set up a fine grid of points $\{x_i\}_{i=1}^{10^4}$ and compute the value of $\hat{p}(x_i)$ by linearly interpolating between the adjacent $\hat{p}$ which has been estimate from the simulated data. This should guarantee a 'fairer' comparison to the 'true' Lévy density p. Finally, the overall integrated squared error is calculated as the mean value of the individual ISE_m's, i.e.,

$$\mathsf{ISE} = \frac{1}{M} \sum_{m=1}^{M} \mathsf{ISE}_m \ ,$$

The results are summarized in Table 4.1.

Table 4.1: *Simulation Results for Variance Gamma Model*
This table shows the integrated squared errors ISE. The number N denotes the number of bins used in the construction of the preliminary histogram estimators. N^* corresponds to the optimal value $\hat{m}$ (2.6.4). For the wavelet bases, the Daubechies (1988) class is used, where the corresponding number characterizes the degree of smoothness.

	N		
	25	N^*	1000
db4	0.602	0.532	0.663
db6	0.554	0.468	0.577
db8	0.479	0.398	0.526
db10	0.438	0.343	0.471

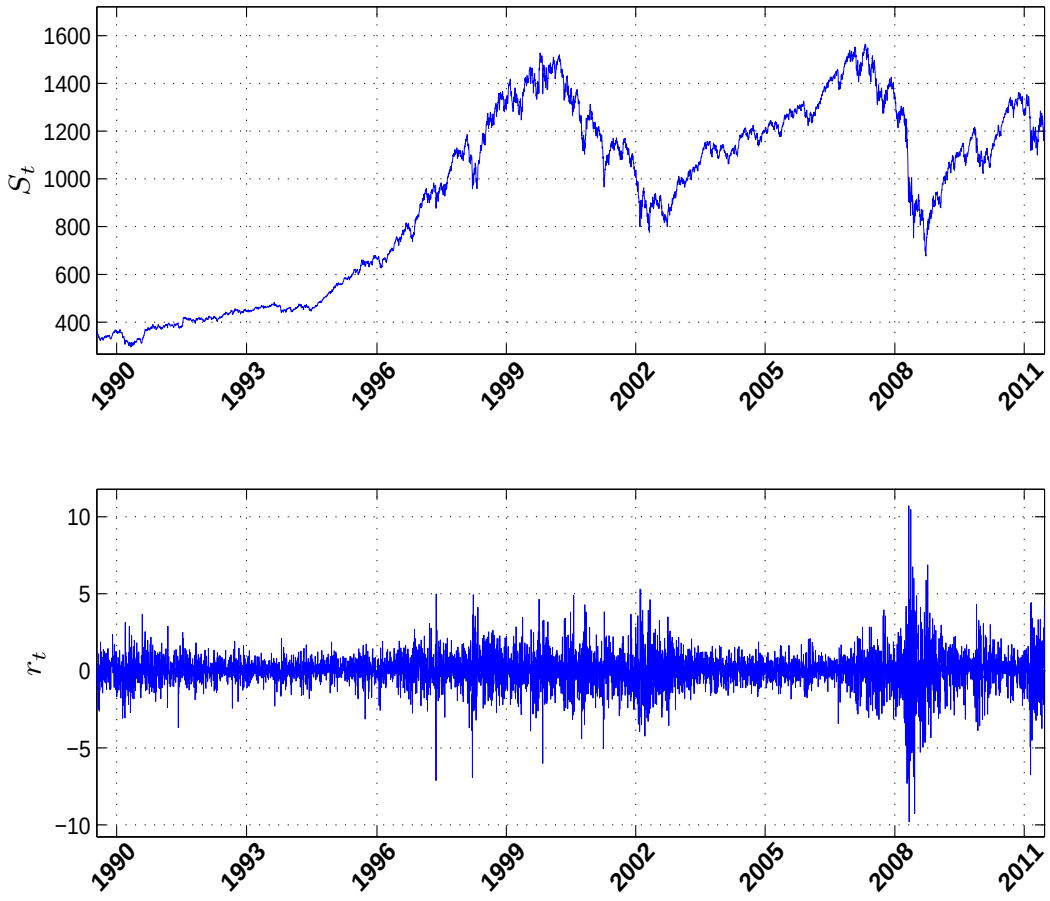
The results of Table 4.1 provide some indication that it may be advantageous to use an optimal bin selector. In particular, when estimating near the origin and using a binwidth too small, the estimated Lévy density 'falls off' as $|x| \to 0$. This may be explained by the fact that in simulations the bin nearest to the origin contains not enough observations, when there are too many bins. On the other hand, if the binwidth of the bin nearest to the origin is too large and might drag the blow-off effect at the origin into regions where this effect has actually vanished. The second result of Table 4.1, which is well known from the properties of wavelets, is that, for a smooth underlying density, smoother wavelets do perform better.

Figure 4.7 depicts the end-of-day price and returns series of the S&P 500 index from 01/02/1990 to 11/30/2011. Seneta (2004) provided an extensive study of the variance gamma process applied to financial data. Recently, Aït-Sahalia and Jacod (2011) developed a procedure to nonparametrically test whether a Lévy process has finite or infinite activity.

Panel (a) of Figure 4.8 depicts the block-thresholded wavelet density estimator from Section 3.2 along with the underlying histogram constructed from the model selection approach of Section 2.6 for nonparametrically estimating the Lévy density of the 15-seconds returns of the S&P 500 index. First, note that there is clear evidence for 'small' positive returns to have a higher arrival rate than their negative counterparts, which is an indication of asymmetry. However, there does not seem to be a significant difference between the wavelet estimator and the histogram, which might indicate that the former does not yield much of an improvement upon the latter, except that the wavelet estimator smooths the roughness of the histogram, as expected. This view is substantiated by Panel (b) of Figure 4.8 which zooms in at the area around the origin.

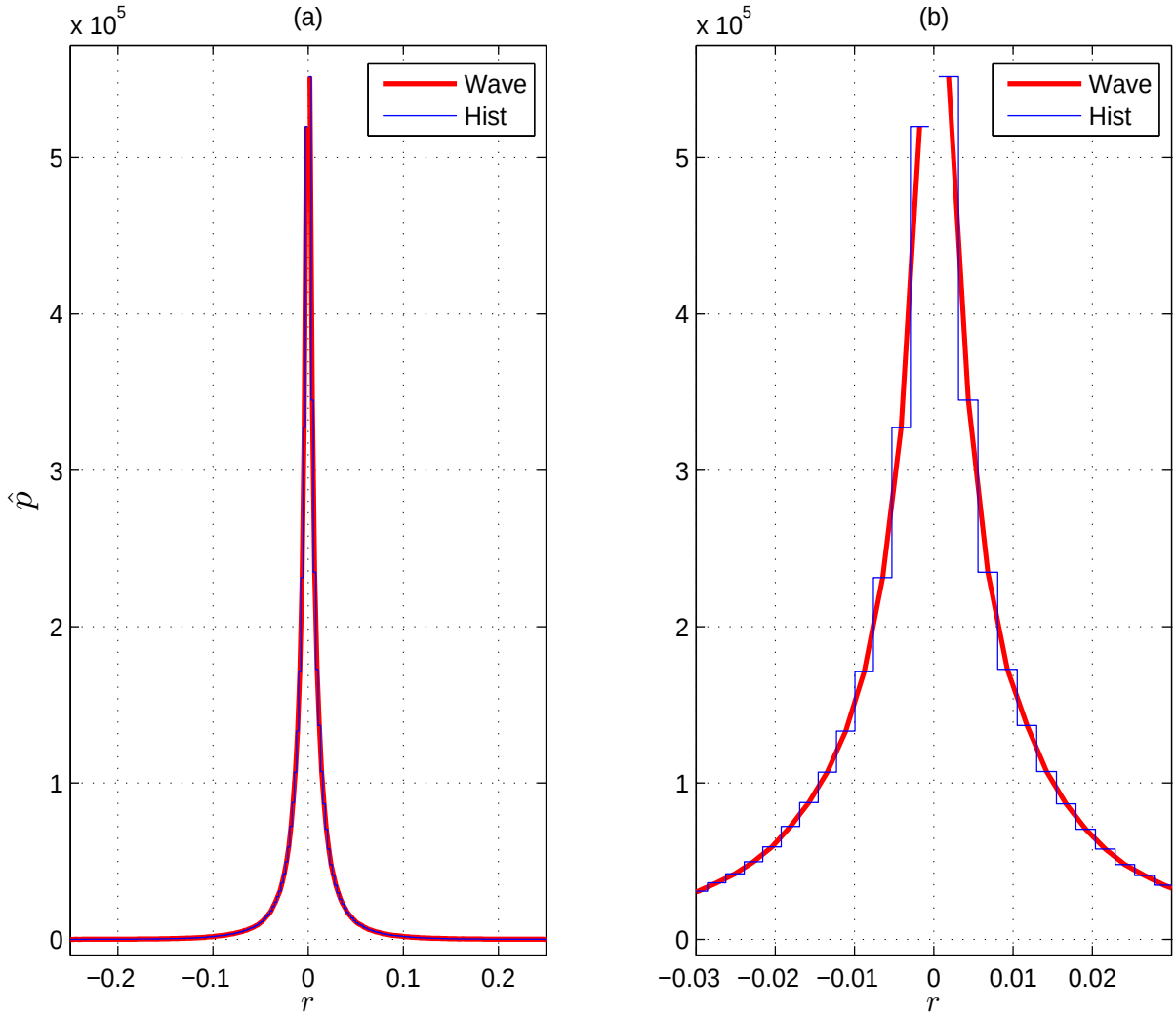
Figure 4.9 is analogous to Figure 4.8, except that the wavelet density estimator was replaced by a kernel density estimator which was scaled by the number of observations.

Figure 4.7: *S&P 500 Index and Returns Series*
This figure depicts daily series of the S&P 500 from Jan 02, 1990 to Nov 30, 2011. The upper panel displays the level series S_t. The lower panel displays the corresponding (continuously compounded) returns $r_t := \big[\ln(S_t) - \ln(S_{t-1})\big] \cdot 100\%$.
Note that although the data set consists of high-frequency observations (tick data), the panels show daily observations only. The reason for this is that at the beginning of the sample period, the S&P 500 was computed at a lower sampling frequency. Without daily aggregation, the time axes would be stretched out on the right-hand side.



The asymmetry of the arrival rate of 'small' jumps seems to be less pronounced than for the wavelet estimator. Moreover, the kernel estimator seems to have a harder time of gauging the overall arrival rate of 'small' jumps. This could be taken as an indication in favor of using the proposed approach for estimating Lévy densities with high arrival rates. Finally, note that another advantage of the wavelet estimator is that it is much faster to compute than the kernel density estimator. The 15-second returns series contained 4393337 observations, which increased the computational time of the kernel density estimator by a factor of 270 relative to the computational time of the wavelet estimator.

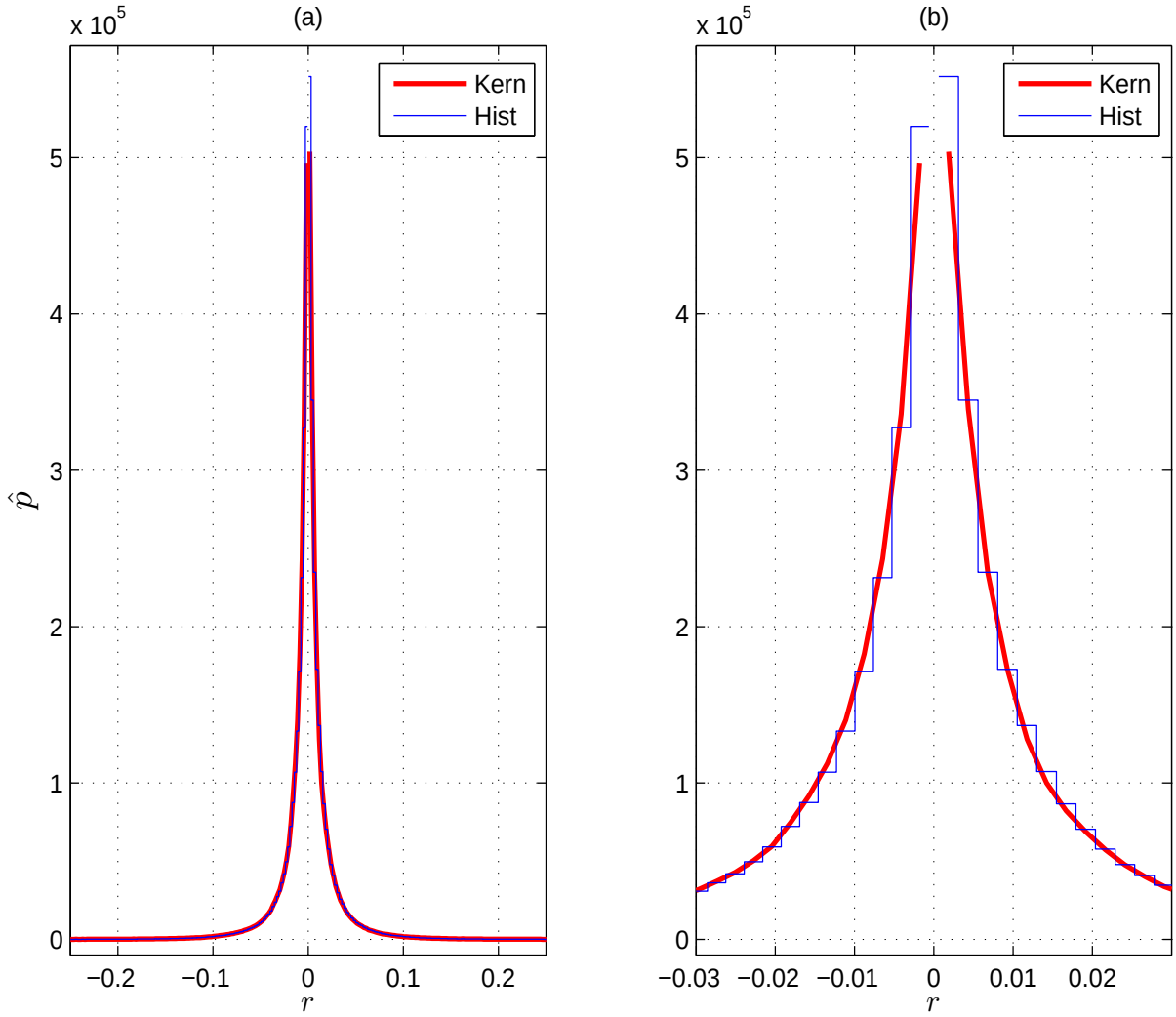
Figure 4.8: *Wavelet Density Estimator for S&P 500 Returns*
This figure illustrates the block-thresholded wavelet density estimator from Section 3.2 along with the underlying histogram constructed from the model selection approach of Section 2.6 for nonparametrically estimating the Lévy density of the 15-seconds returns of the S&P 500 index. Panel (a) shows the estimators on the whole support of the sampled returns, while Panel (b) zooms in at the areas around the origin.



4.2 Lévy-Driven Ornstein-Uhlenbeck Processes

Volatility modeling is an important ingredient for many branches of finance such as, for example, risk management and derivative pricing. In particular, dynamic and distributional aspects of volatility are the primary objects of interest. First attempts to tackle this problems were purely parametric such as the classical ARCH (Engle, 1982) and GARCH (Bollerslev, 1986) models or the stochastic volatility model (Hull and White, 1987; Heston, 1993).

Figure 4.9: *Kernel Density Estimator for S&P 500 Returns*
This figure compares the histogram constructed from the model selection approach of Section 2.6 to a kernel density, estimator which was scaled by the number of observations, for the problem of nonparametrically estimating the Lévy density of the 15-seconds returns of the S&P 500 index. Panel (a) shows the estimators on the whole support of the sampled returns, while Panel (b) zooms in at the areas around the origin.



Recently, the surge in computational power and the availability of high-frequency financial data paved the way for nonparametric estimation of (stochastic) volatility. See, for example, Andersen and Benzoni (2009) and Barndorff-Nielsen and Shephard (2002). In financial applications, interest often centers on daily volatility whose nonparametric counterpart is the integrated (or cumulative) volatility over time period $[0, T]$, where T equals one day. To be more precise, the **integrated volatility** (or variance) is defined by

$$\langle X, X \rangle_T := \int_0^T \sigma_t^2 \, \mathrm{d}t \ ,$$

where σ_t is the instantaneous (but latent) volatility, $X_t := \mathsf{ln}(S_t)$, and S_t is the price of a financial asset. However, note that the term "integrated volatility" is only accurate for models without jump component in the log-price equation, i.e.,

$$\mathsf{d}X_t = \gamma_t \, \mathsf{d}t + \sigma_t \, \mathsf{d}W_t \; .$$

In the presence of jumps, the variation of the increments due to jumps is also included in $\langle X, X \rangle_T$. In this case, it is more appropriate to call $\langle X, X \rangle_T$ the **quadratic variation** of X_t. Given a sample high-frequency observations of the log-price X_t on a discrete time grid over one day T, i.e.,

$$0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = T \; ,$$

with sampling frequency $\Delta_n := \mathsf{max}_{1 \leqslant i \leqslant n} \{t_i - t_{i-1}\}$, the sum of squared high-frequency returns constitutes a sensible estimator for $\langle X, X \rangle_T$. To be more precise, the **realized volatility** is defined by

$$[X, X]_T := \sum_{t_{i-1}, t_i \in [0,T]} (X_{t_i} - X_{t_{i-1}})^2 \; . \tag{4.2.1}$$

The upper panel of Figure 4.10 depicts the daily realized volatility computed from the tick data for the S&P 500 index from 01/02/1990 to 11/30/2011 shown in Figure 4.7.

The theoretical justification for (4.2.1) as an estimator of $\langle X, X \rangle_T$ is based upon Theorem I.4.47 of Jacod and Shiryaev (2003) which shows that

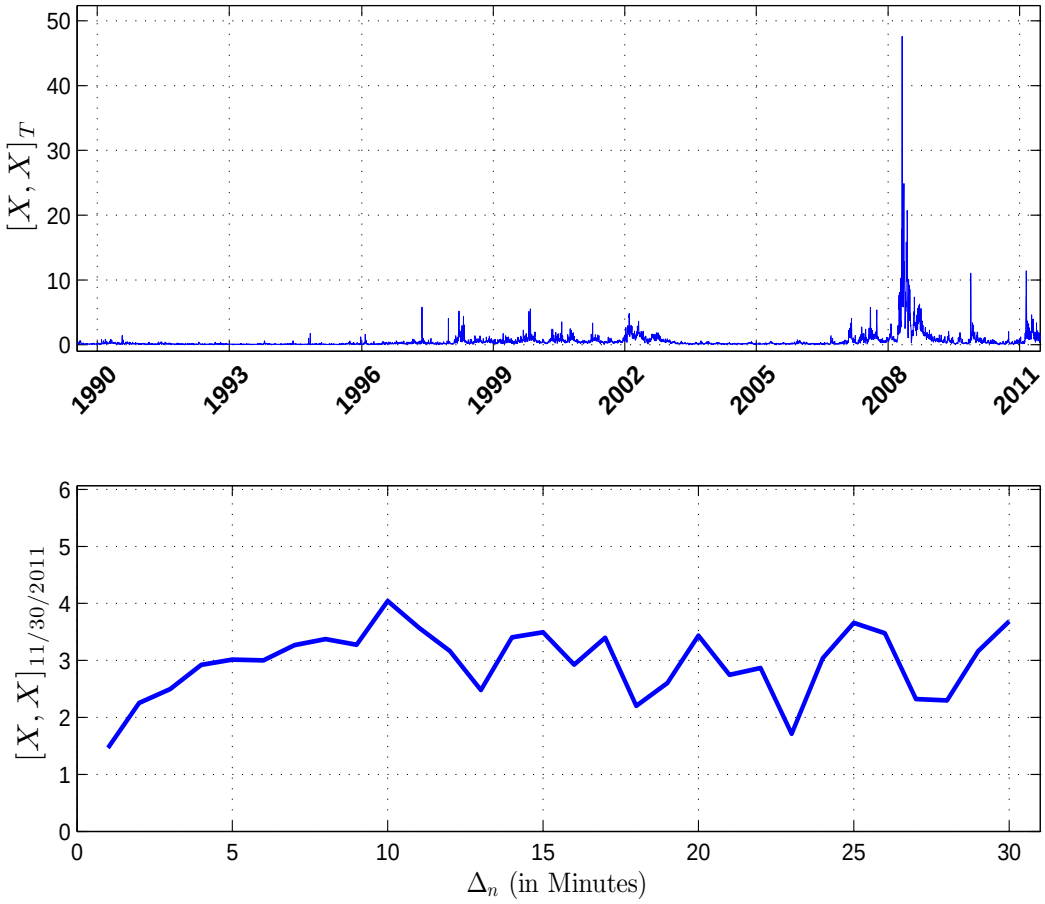
$$[X, X]_T \xrightarrow{P} \langle X, X \rangle_T$$

as $n \to \infty$ such that $\Delta_n \to 0$. Put differently, the estimation error of $[X, X]_T$ vanishes as the sampling frequency increases. Unfortunately, computing and plotting $[X, X]_T$ for increasing sample frequency leads to a picture (the so-called **signature plot**) where the realized volatility shows no tendency to converge or stabilize. Indeed, for $\Delta_n \searrow 0$, realized volatility seems to blow off for many financial assets.

The explanation for this phenomenon is that the efficient price process X_t is not directly observable at high sampling frequencies, but that we are only able to observe the transaction price process Y_t which is contaminated by some noise component ϵ_t:

$$\underbrace{Y_{t_i}}_{\substack{\text{transaction price} \\ \text{(observable)}}} = \underbrace{X_{t_i}}_{\substack{\text{efficient price} \\ \text{(unobservable)}}} + \underbrace{\epsilon_{t_i}}_{\substack{\text{market micro-} \\ \text{structure noise}}}$$

Figure 4.10: *Realized Volatility of S&P 500 Index*
This figure depicts ana analysis of the daily realized volatility of the S&P 500 from Jan 02, 1990 to Nov 30, 2011. The upper panel contained the time series of the realized volatility based upon the (continuously compounded) returns $r_t := \big[\ln(S_t) - \ln(S_{t-1})\big] \cdot 100\%$ displayed in the lower panel of Figure 4.7. The lower panel shows the signature plot of the realized volatility of Nov 30, 2011.



The noise component ϵ_t subsumes many factors known as **market microstructure effects** such as "bid-ask bounces, discreteness of price changes, differences in trade sizes or informational content of price changes, gradual response of prices to a block trade, the strategic component of the order flow, inventory control effects, etc." (Aït-Sahalia, 2007).

As a quick fix to the problem induced by market microstructure noise, Andersen, Bollerslev, Diebold, and Labys (2001) proposed to sample the data at a lower frequency, say 5 minutes, for which the signature plot shows that the realized volatility has settled on a stable level. However, as this sparse-sampling scheme implies that most the data will be discarded, more sophisticated procedures were suggested in order to correct for market microstructure noise. See, for example, Aït-Sahalia, Mykland, and Zhang (2005), Zhang,

Mykland, and Aït-Sahalia (2005), and Zhang (2006).

To check whether there is a need to apply one of these noise correction procedures, the lower panel of Figure 4.10 depicts the realized volatility of S&P 500 index on 11/30/2011 computed for varying sampling frequencies.[3] Clearly, the realized volatility does not exhibit any diverging behavior for increasing sample frequencies $1/\Delta_n$.

Next, let us take a look at what empirical evidence was found on the relationship between volatility and jumps. For example, Eraker, Johannes, and Polson (2003) estimated a parametric stochastic volatility model via maximum likelihood and found strong support for the hypothesis of jumps in the price and volatility series of the S&P 500 and the Nasdaq 100.

Wu (2011) used the realized volatility (4.2.1) in order to estimate the volatility of the S&P 500 index nonparametrically and found evidence for jumps in volatility. Moreover, the arrival rate of jumps is very high and proportional to the level of volatility.

In another study, Todorov and Tauchen (2011) analyzed the volatility index VIX computed by the *Chicago Board of Options Exchange* from close-to-maturity options written on the S&P 500 index. Consequently, the VIX is often interpreted as the market's risk-neutral expectations of future volatility extracted from implied volatilities. In their analysis, the authors used a generalized version of the (Blumenthal and Getoor, 1961) index (Aït-Sahalia and Jacod, 2009a; Todorov and Tauchen, 2010) which can be directly related to the stability index of α-stable processes. Their results suggest that the VIX is driven by 'small' and 'large' jumps. On the one hand, the existence of 'large' jumps require the inclusion of a jump component per se in any sensible model of the VIX. On the other hand, although 'small' jumps arrive at a high intensity that rate is not high enough to favor the inclusion of a continuous component.

As we have already argued in the Preface, simple Lévy processes are suitable models for unconditional (or marginal) phenomena of financial returns distributions, but they fail when it comes to reproducing or explaining conditional (or dynamic) phenomena of financial returns. Recall that the most important stylized facts about the dynamics of financial returns are volatility clustering, leverage, and long memory. Fortunately, there is still some scope for building more realistic models using pure-jump Lévy processes.

Barndorff-Nielsen and Shephard (2001) proposed a continuous-time stochastic volatility model for a security price $S_t = S_0\mathsf{exp}(X_t)$, where the volatility process σ_t^2 follows a weighted sum of independent Ornstein-Uhlenbeck processes $\sigma_{t,j}^2$, each of which has an

[3] We have picked a day from the end of the sample since they have the largest amount of data compared to days from the beginning of the sample period.

independent **background driving Lévy process** $Z_{t,j}$:

$$\begin{aligned} \mathsf{d}X_t &= (\mu + \beta\sigma_t^2)\,\mathsf{d}t + \sigma_t\,\mathsf{d}W_t + \sum_{j=1}^{m} \varrho_j\,\mathsf{d}\bar{Z}_{t,j} \\ \sigma_t^2 &= \sum_{j=1}^{m} w_j \sigma_{t,j}^2 \\ \mathsf{d}\sigma_{t,j}^2 &= -\lambda_j \sigma_{t,j}^2\,\mathsf{d}t + \mathsf{d}Z_{t,j}\ , \end{aligned} \tag{4.2.2}$$

where $\bar{Z}_{t,j} = Z_{t,j} - \mathsf{E}[\,Z_{t,j}\,]$, $\sum_{j=1}^{m} w_j = 1$, and $w_j, \lambda_j > 0$ for all $j = 1, \ldots, m$. The parameter $\mu > 0$ can be interpreted as the risk-free rate, while $\beta > 0$ can be interpreted as the risk premium. The parameters $\varrho_j < 0$ control the degree of leverage effect, whereas the parameters $\lambda_j > 0$ control the speed of mean-reversion. The weights w_j control the persistence of volatility. In particular, long-range dependence, i.e., long memory, is obtained for $m \to \infty$.

Besides the Wiener process W_t, the model dynamics is driven by Lévy processes $Z_{t,j}$, which are defined to have no drift and no diffusion component. In order to rule out negative volatilities, they are additionally required to be pure-jump Lévy processes with non-negative increments. Thus, the $Z_{t,j}$'s correspond to subordinators in the sense of Definition 1.14. Interestingly, it can be shown that the distribution of log-returns tends to the normal distribution when the time interval, on which returns are computed, increases. Thus, this model even provides an explanation of aggregational normality.

In the sequel, we work with a simplified model, since the superposition of factor volatilities in (4.2.2) is more involved and analytically less tractable. To this end, Barndorff-Nielsen and Shephard (2001) proposed a one-factor Ornstein-Uhlenbeck process for modeling stochastic volatility:

$$\mathsf{d}\sigma_t^2 = -\lambda\sigma_t^2 + \mathsf{d}Z_{\lambda t}\ , \tag{4.2.3}$$

whose strong solution is given by

$$\sigma_t^2 = \mathsf{e}^{-\lambda t}\sigma_0^2 + \int_0^t \mathsf{e}^{-\lambda(t-s)}\mathsf{d}Z_{\lambda s}\ . \tag{4.2.4}$$

Note that replacing the usual time index t by λt guarantees that the marginal (or stationary) law of σ_t^2 does not depend upon the parameter λ. Moreover, it can be shown that, for $\lambda > 0$, this process is indeed mean-reverting and strongly stationary.

Jongbloed and van der Meulen (2006) proposed a parametric estimator of discretely sampled subordinators and Ornstein-Uhlenbeck processes with background driving Lévy processes. Their estimator is based upon M-estimation of the cumulant function. The

same problem was considered by Jongbloed, van der Meulen, and van der Vaart (2005). However, their approach was nonparametric since the measure of the background driving Lévy process is related to the stationary law of Ornstein-Uhlenbeck process via its empirical characteristic function.

Simulations of Ornstein-Uhlenbeck process with background driving Lévy process can be implemented via its strong solution (4.2.4) or directly via an Euler-Maruyama discretization scheme of (4.2.3). The Lévy-driven Ornstein-Uhlenbeck processes of Barndorff-Nielsen and Shephard (2001) are constructed in such a way that their marginal (or stationary) laws are predetermined. Probably, the simplest of these non-Gaussian Ornstein-Uhlenbeck models is the $\Gamma(\alpha, \beta)$-Ornstein-Uhlenbeck model, which has a $\Gamma(\alpha, \beta)$ marginal law. Then, the subordinator is simulated as a compound Poisson process (1.1.4) with Lévy measure $\nu(\mathsf{d}x) = \alpha\lambda\mathsf{Exp}(\beta)\,\mathsf{d}x$. Figures 4.11 and 4.12 illustrates simulated sample paths of the $\Gamma(\alpha, \beta)$-Ornstein-Uhlenbeck process (4.2.3) for low and high λ, respectively.

The subordinators in the Barndorff-Nielsen and Shephard (2001) model may have infinite activity but only with finite variation. The latter property is somewhat at odds with one result of Todorov and Tauchen (2011) which favors subordinators with infinite variation.

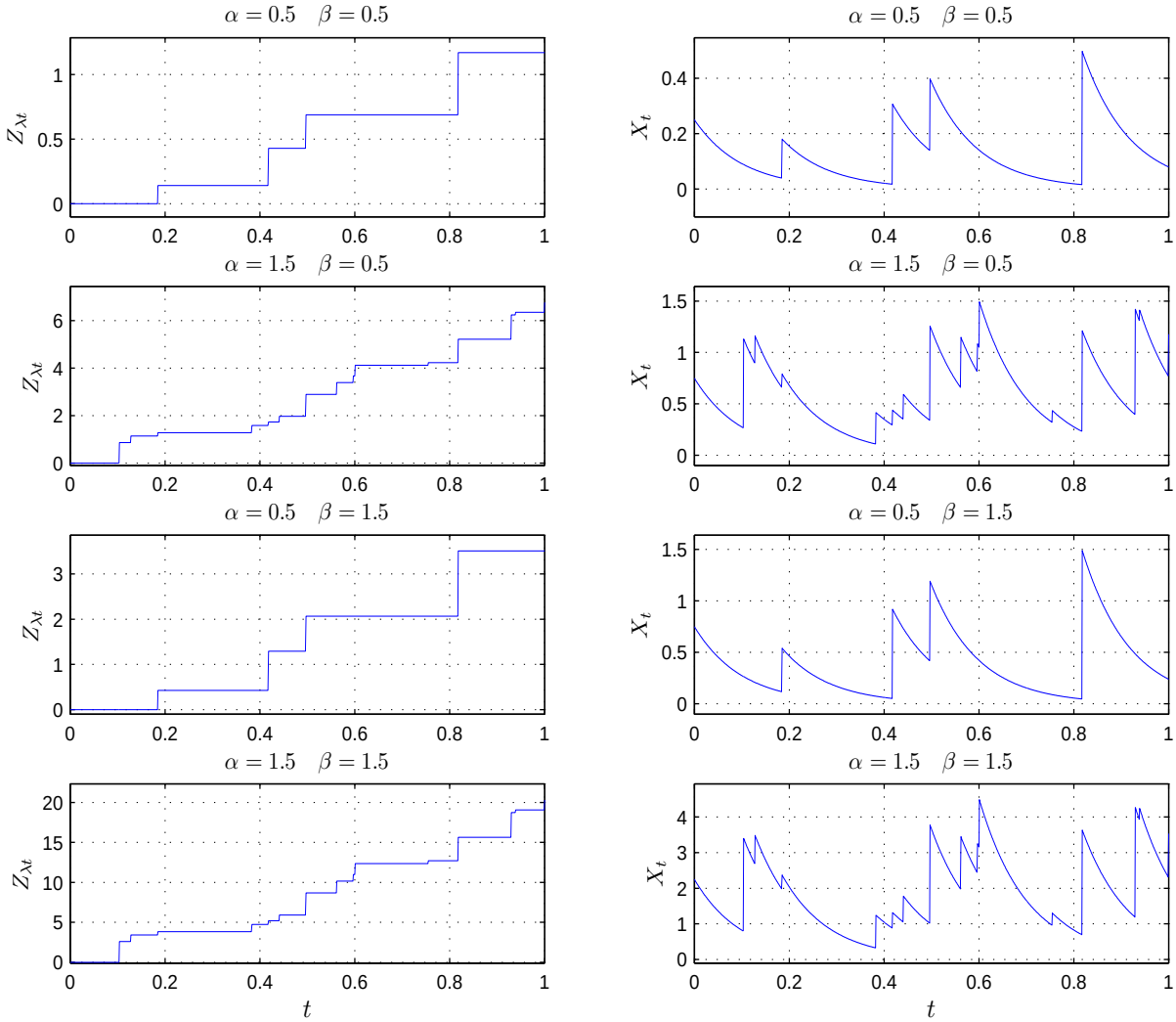
The dynamics of σ_t^2 are be summarized as follows: Upward movements of σ_t^2 are solely due to the jumps originating from the subordinator $Z_{\lambda t}$, while the effects of these jumps die out exponentially fast at the rate λ due to the mean-reversion of σ_t^2. This property motivates the estimation of the Lévy density of the driving subordinator by considering the positive increments of realized volatility. Of course, this correspondence is only accurate for continuous-time sampling. Thus, we need a consistency argument based upon in-fill asymptotics to warrant our estimation approach to be valid.

Although Eraker, Johannes, and Polson (2003) and Todorov and Tauchen (2011) suggested that there are jumps in the price series and, thus, $[X, X]_T$ contains a component induced by jump variation, it would nevertheless be possible to nonparametrically disentangle the integrated variance from the jump variation using the notion of **bipower variation** (Barndorff-Nielsen and Shephard, 2004, 2006). We leave this for future research.

We now look at the results of a small simulation study where used the $\Gamma(1.5, 0.5)$-Ornstein-Uhlenbeck model of Figure 4.12 with $\lambda = 0.1$ using the setup of Section 4.1. The results, summarized in Table 4.2, are similar to those of the simulation study in Section 4.1 yielding analogous conclusions and, therefore, omitted.

Figure 4.13 depicts the block-thresholded wavelet density estimator from Section 3.2

Figure 4.11: *Sample Paths of Levy-Driven Ornstein-Uhlenbeck Processes (a)*
This figure depicts simulated paths of the Lévy-driven Ornstein-Uhlenbeck process (4.2.3) and the respective subordinators for differing parameter values of α and β, while $\lambda = 0.01$ is kept constant. The step size (or simulation frequency) is fixed at $\Delta t := t_i - t_{i-1} = 10^{-3}$ for all $i = 1, \ldots, 1000$ with $t_0 = 0$. The same seed for pseudo-random number generation is used along panel rows. The initial value X_0 is is equated to its mean $\alpha\beta$. The sampling interval is scaled down to $[0, 1]$.
Clearly, increasing α increases the jump intensity, while increasing β increases the jump sizes. Obviously, these effects are to be expected from the structure of the $\Gamma(\alpha, \beta)$-Ornstein-Uhlenbeck model.



along with the underlying histogram constructed from the model selection approach of Section 2.6 for nonparametrically estimating the Lévy density of the subordinator of the daily realized volatility of the S&P 500 index. The analysis and the comparison to the corresponding kernel density estimator depicted in Figure 4.14 is analogous and, therefore, omitted.

Figure 4.12: *Sample Paths of Levy-Driven Ornstein-Uhlenbeck Processes (b)*
This figure depicts simulated paths of the Lévy-driven Ornstein-Uhlenbeck process (4.2.3) and the respective subordinators. The simulation setup is exactly the same as in Figure 4.11, expect that now $\lambda = 0.1$.
All effects are qualitatively analogous to Figure 4.11, when taking into account that increasing λ increases the jump intensity.

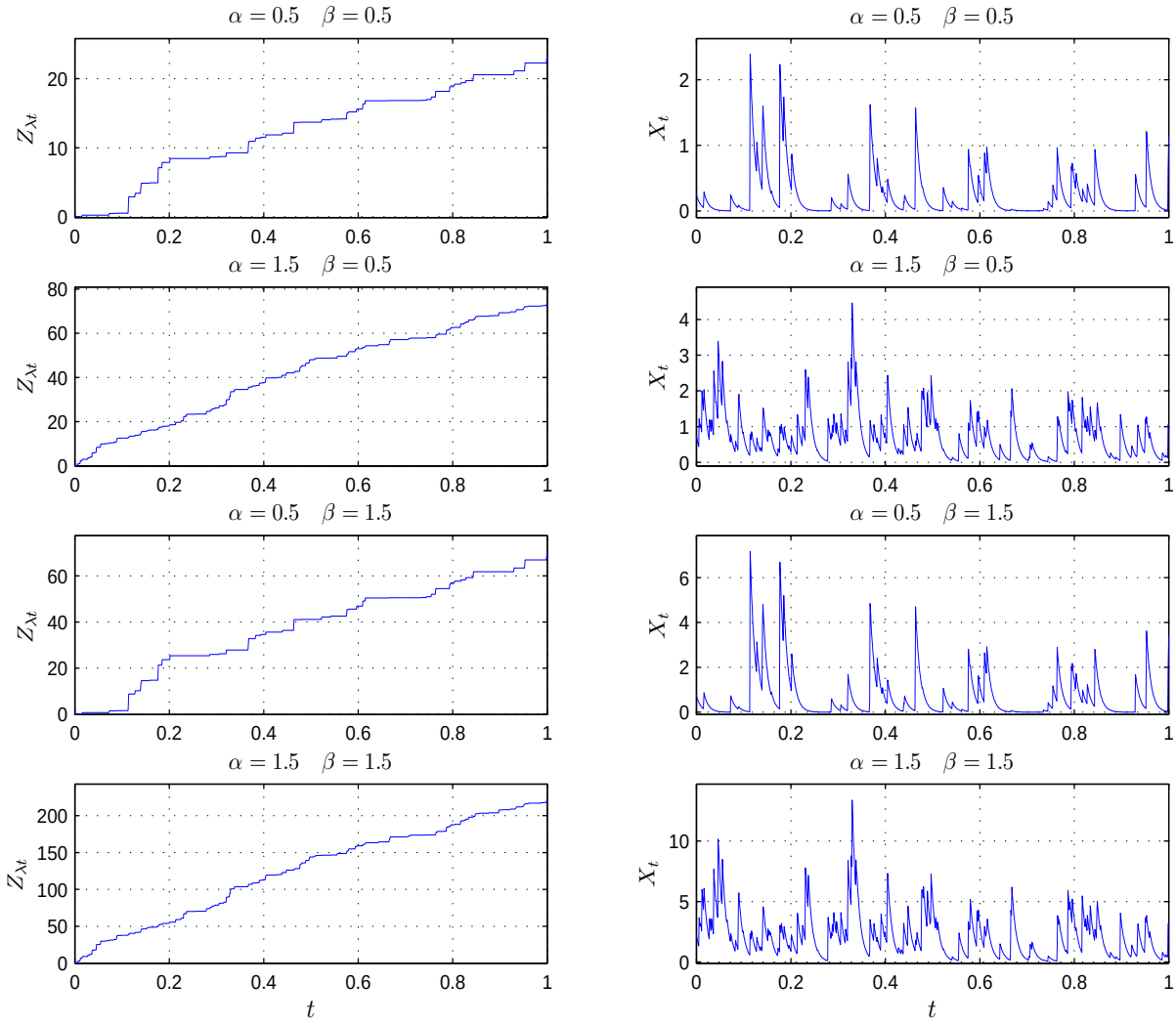


Table 4.2: *Simulation Results for Lévy-Driven Ornstein-Uhlenbeck Model*
This table shows the integrated squared errors ISE. The number N denotes the number of bins used in the construction of the preliminary histogram estimators. N^* corresponds to the optimal value $\hat{m}$ (2.6.4). For the wavelet bases, the Daubechies (1988) class is used, where the corresponding number characterizes the degree of smoothness.

	N		
	25	N^*	1000
db4	0.751	0.632	0.728
db6	0.704	0.609	0.693
db8	0.656	0.587	0.637
db10	0.615	0.510	0.598

Figure 4.13: *Wavelet Density Estimator for S&P 500 Realized Volatility*
This figure illustrates the block-thresholded wavelet density estimator from Section 3.2 along with the underlying histogram constructed from the model selection approach of Section 2.6 for nonparametrically estimating the Lévy density of the subordinator of the daily realized volatility of the S&P 500 index. Panel (a) shows the estimators on the whole support of the sampled returns, while Panel (b) zooms in at the areas around the origin.

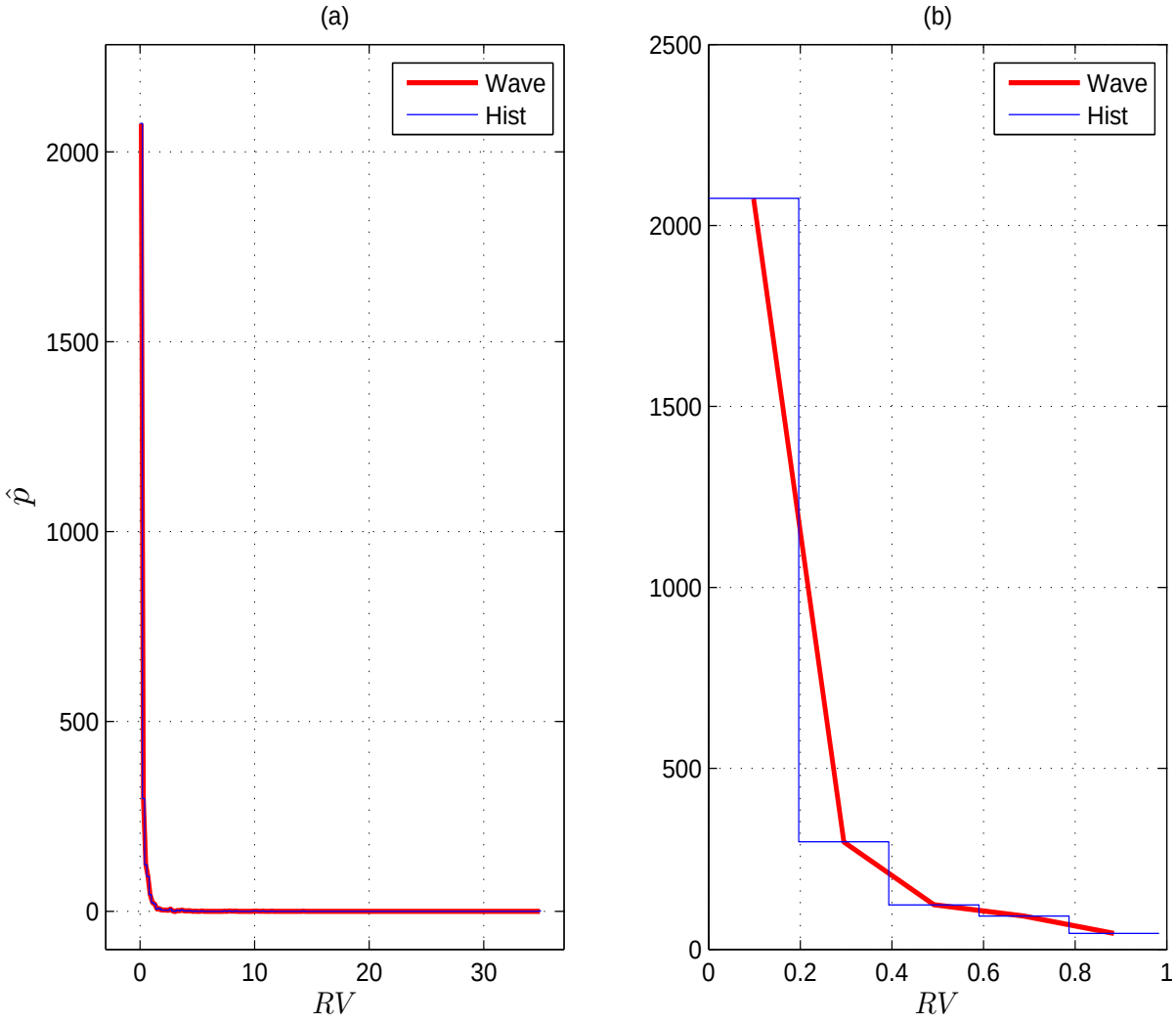
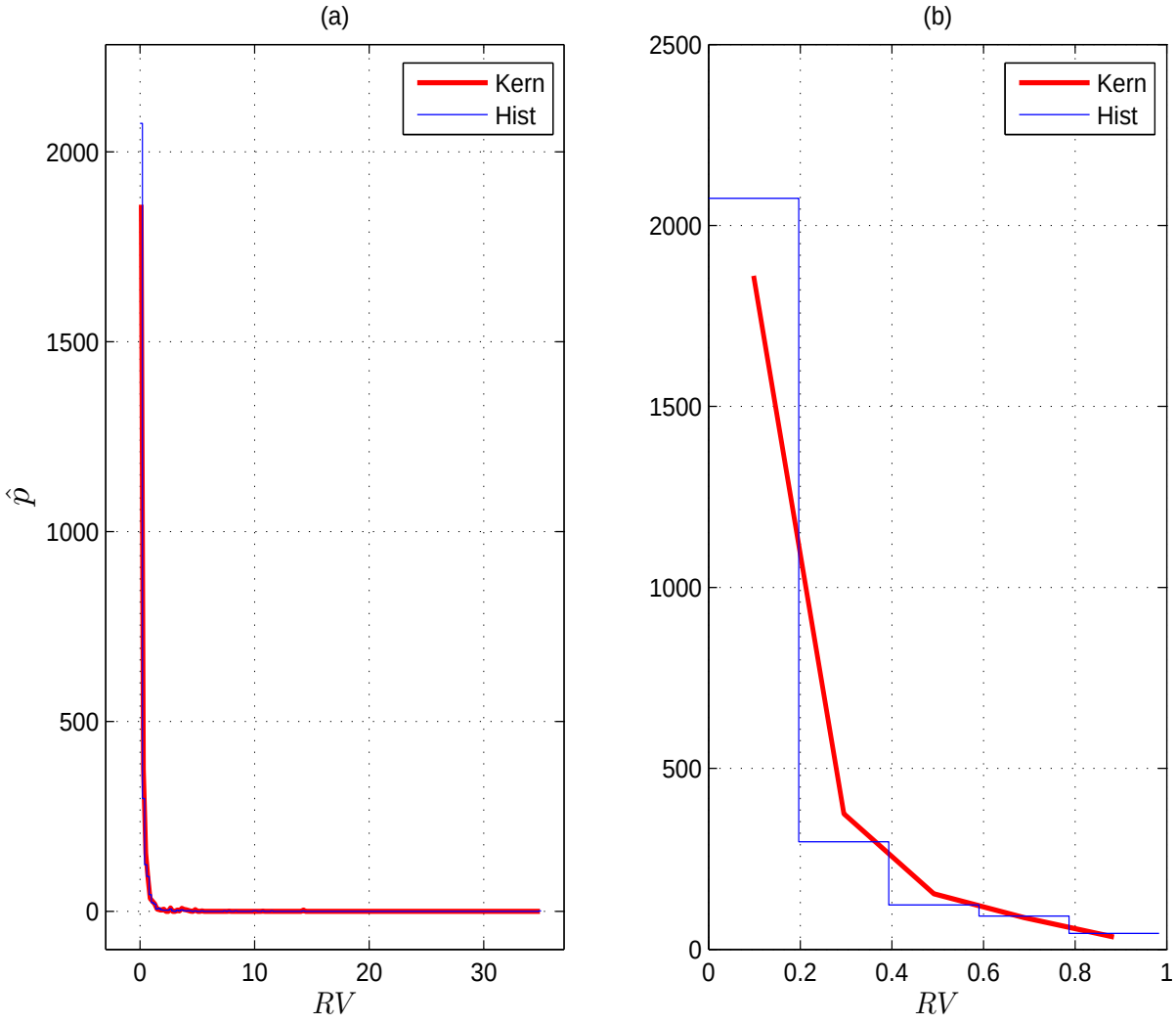


Figure 4.14: *Kernel Density Estimator for S&P 500 Realized Volatility*
This figure compares the histogram constructed from the model selection approach of Section 2.6 to a kernel density, estimator which was scaled by the number of observations, for the problem of nonparametrically estimating the Lévy density of the subordinator of the daily realized volatility of the S&P 500 index. Panel (a) shows the estimators on the whole support of the sampled returns, while Panel (b) zooms in at the areas around the origin.



Outlook

Finally, we point out some further directions which could be pursued in future work to complement or extend what has been done up to now.

One well-known problem of orthogonal projection estimators of density functions is that they may be negative and/or not integrate to one in small samples. Glad, Hjort, and Ushakov (2003) proposed some easy-to-implement post-processing remedies to these problems. Instead of ex-post corrections, there are essentially two approaches which correct for the shortcomings of orthogonal projection estimators at the outset.

The first one is based on the idea to estimate a transformation of the original density function. For example, Pinheiro and Vidakovic (1997) considered the estimation of the square root of the density function via wavelets. Another example is Song (2010) who estimated the log density function via wavelets. This approach is an extension of Kim and Koo (2002) who used information projection on an orthonormal wavelet basis with shrinkage to estimate the intensity function of an inhomogeneous Poisson process. Here, it would be interesting to derive analytical results for thresholded estimators and to compare their relative performance in a simulation study.

The second approach is based on the idea that a shape-preserving estimator of a density function can be implemented by using two different sets of wavelet basis functions. While the first set is used for decomposition and the second one for reconstruction, both sets are not mutually orthogonal. That is why these basis functions are called biorthogonal (Cohen, Daubechies, and Feauveau, 1992). An interesting point to note is that biorthogonal wavelets are essentially equivalent to B-splines and, therefore, particularly easy to implement as their 'basis functions' are explicitly given. Cosma, Scaillet, and von Sachs (2007) proposed a density estimator based on biorthogonal wavelets by using the approach of Dechevsky and Penev (1997, 1998). Again, it would be interesting to transfer this approach to the estimation of Lévy density using threshold rules. In this respect, the recent approach of Reynaud-Bouret and Rivoirard (2010) seems to be closely related and most straightforwardly to extend.

Another direction of future research could be the development of functional central limit theorems for the use in deriving confidence sets and goodness-of-fit procedures for wavelet-based Lévy density estimators. For example, Figueroa-López (2011) derived Sieve-based confidence bands for Lévy densities based on orthogonal projections, whereas Giné and Nickl (2009) derived Donsker theorems for wavelet-based density estimator.

In Section 2.1, we considered the risk based on the $\mathcal{L}_2$-loss as a global measure for estimation precision. For measuring of local accuracy of an estimator at a point x_0, the expected (pointwise) squared error loss at x_0 is used

$$R\left(\left(\hat{f}(x_0) - f(x_0)\right)^2\right) = \mathsf{E}\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right] .$$

Chicken and Cai (2005) proved that thresholding based on blocks of order $\log n$ are simultaneously adaptively rate-optimal (over the usual function spaces) in the global and local sense. This is in contrast to the original order $(\log n)^2$ of block lengths proposed by Hall, Kerkyacharian, and Picard (1999) which is too large to be locally adaptive. These results were derived for probability density estimation, and an extension to Lévy densities would complete the results of Section 3.2.

Finally, note that, in Section 4.2, we only considered a simplified version of the Lévy-driven Ornstein-Uhlenbeck model favored by Barndorff-Nielsen and Shephard (2001). As already mentioned in Section 4.2, in order to capture the stylized fact of long memory in volatility, a combination of Lévy-driven Ornstein-Uhlenbeck volatilities were considered by Barndorff-Nielsen and Shephard (2001). In this model, it would be interesting to have a method for nonparametric identification and estimation of a mixture of Lévy subordinators. Corsi (2009) provided a possible economic interpretation for factors driving volatility.

Appendix A

Mathematical Review

A.1 Review of Relevant Probability Theory

This section is based upon the monographs of Dudley (2002), Fristedt and Gray (1997), and Kallenberg (2002).

Definition A.1 (Radon measure) *A measure μ on a Borel σ-field $\mathscr{F}$ is a Radon measure, if $\mu(C) < \infty$, for every compact set $C \in \mathscr{F}$.*

Theorem A.2 (Radon-Nikodým) *Let μ and ν be σ-finite measures on a measurable space $(\Omega, \mathscr{F})$. $\nu \ll \mu$ if and only if there exists a non-negative measurable function f such that, for any $A \in \mathscr{F}$,*

$$\nu(A) = \int_A f \, \mathsf{d}\mu \; .$$

The function f is (μ-a.e.) unique. Let f be integrable with respect to μ, and define $\nu(A) = \int_A f \, \mathsf{d}\mu$. Then, ν is a signed measure.

Definition A.3 (Fourier Transform) *Let $f \in \mathcal{L}_1(\mathbb{R})$. The Fourier transform of f is defined by*

$$f^*(u) = \mathsf{F}\big[f(x)\big] := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathsf{e}^{-\mathsf{i}ux} f(x) \, \mathsf{d}x \; .$$

Sometimes the Fourier transform is stated without the factor $1/\sqrt{2\pi}$ and/or the minus sign in the exponent, i.e., $\int_{-\infty}^{\infty} \mathsf{e}^{\mathsf{i}ux} f(x) \, \mathsf{d}x$, which does not change anything in the theory of the Fourier transform.

The primary tool for characterizing and analyzing the laws of Lévy processes are the characteristic functions of their distributions, which are closely related to the Fourier transform of Definition A.3. For more details on the Fourier transform and its relation to the characteristic function, see the monographs of Kawata (1972) and Lukacs (1970).

Definition A.4 (Characteristic Function) *The characteristic function $\Phi_X(u)$ of an $\mathbb{R}$-valued random variable X (or of its probability law P_X) is defined by the $\mathbb{C}$-valued function*

$$\Phi_X(u) := \int_{\mathbb{R}} \mathsf{e}^{\mathsf{i}ux} P_X(\mathsf{d}x) = \mathsf{E}\big[\, \mathsf{e}^{\mathsf{i}uX} \,\big] = \mathsf{E}[\, \mathsf{cos}(uX) \,] - \mathsf{i}\, \mathsf{E}[\, \mathsf{sin}(uX) \,] \; ,$$

for all $u \in \mathbb{R}$, with complex number $\mathsf{i} = \sqrt{-1}$.

Characteristic functions always exist and are finite. The characteristic function of a random variable X allows for straightforward computation of the pth moment of X (if it

exists) by introducing the notion of the p**th cumulant** of X:

$$\kappa_p := \frac{1}{\mathsf{i}^p} \left. \frac{\mathsf{d}^p \big[\ln \Phi_X(u) \big]}{\mathsf{d}u^p} \right|_{u=0} . \tag{A.1.1}$$

For instance, the first four moments of X (if they exist) read as follows:

$$\mathsf{E}[\, X \,] = \kappa_1 \tag{A.1.2}$$

$$\mathsf{Var}[\, X \,] = \kappa_2 \tag{A.1.3}$$

$$\mathsf{skewness}[X] = \frac{\kappa_3}{\kappa_2^{3/2}} \tag{A.1.4}$$

$$\mathsf{excess\ kurtosis}[X] = \frac{\kappa_4}{\kappa_2^2} . \tag{A.1.5}$$

For positive random variables, there is an alternative to the characteristic function, which is often easier to work with, as we will see later on at the end of this section.

Definition A.5 (Laplace Transform) *The Laplace transform $\Phi_X^+(z)$ of an $\mathbb{R}_+$-valued random variable X (or of its probability law P_X) is defined by the $\mathbb{R}_+$-valued function*

$$\Phi_X^+(z) := \int_0^\infty \mathsf{e}^{-zx} P_X(\mathsf{d}x) = \mathsf{E}\big[\, \mathsf{e}^{-zX} \,\big] \ ,$$

for all $z \geqslant 0$.

There are two relationships which interrelate the Fourier transform and the Laplace transform:

$$\Phi_X(u) = \Phi_X^+(-\mathsf{i}u) \tag{A.1.6}$$

$$\Phi_X^+(z) = \Phi_X(\mathsf{i}z) \ . \tag{A.1.7}$$

However, although the Fourier transform always exists, the Laplace transform may not be extendable to an analytical function in the complex plane. Thus, the crucial condition for these relations to hold is analytical continuation. See also Sato (1999, p. 10) or Fristedt and Gray (1997, p. 219).

A **stochastic process** is a collection $X = \big\{ X_t(\omega) \in E : \omega \in \Omega, t \geqslant 0 \big\}$ is defined on a complete stochastic basis $\big(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t \geqslant 0}, P\big)$ where, as usual, Ω is a sample space, $\mathscr{F}$ is a σ-field of subsets of Ω, $\mathscr{F}_t$ is a filtration of $\mathscr{F}$, and P is a positive probability measure in $(\Omega, \mathscr{F})$. For any time point t, the corresponding random variable in $\{X_t\}_{t \geqslant 0}$ is a $(\mathscr{F}, \mathscr{E})$-measurable mapping $X_t : \Omega \to E$, where $(E, \mathscr{E})$ is a measurable space, where

we will be suppressing the dependence of X_t on ω when no ambiguity is involved. In the sequel, we will be considering state space $E = \mathbb{R}$ only.

Definition A.6 (Càdlàg Function) *A function $f : [0,1] \to \mathbb{R}$ is càdlàg, if*

C1 f is right-continuous, i.e., for all $0 \leqslant x_0 \leqslant 1$,

$$f(x_0+) := \lim_{x \searrow x_0} f(x) = f(x_0) < \infty$$

C2 and has left limits, i.e., for all $0 < x_0 \leqslant 1$,

$$f(x_0-) := \lim_{x \nearrow x_0} f(x) < \infty \, .$$

The space of all càdlàg function on $[0,1]$ is denoted by $\mathcal{D}[0,1]$.

We now turn to the most relevant sort of stochastic processes with respect to our purposes: point processes. A classical account on point processes is Kingman (1993).

A sequence $\{\tau_n\}_{n \in \mathbb{N}}$ of strictly increasing points, i.e., $0 < \tau_1 < \tau_2 < \cdots$ with $\tau_n \to \infty$ as $n \to \infty$, is a **simple point process**. When $\{\tau_n\}_{n \in \mathbb{N}}$ is interpreted as points in time, which indicate the occurrence of some pre-specified event, say B, it is usually assumed that $\tau_0 := 0$. Thus, the notion 'event' should not be confused with $\omega \in \Omega$, but rather deemed to be some primitive event like, for example, a stock order arriving at a market maker. The attribute 'simple' indicates that these events occur sequentially and not simultaneously. Consequently, when considering some fixed τ_n, we speak of it as the nth arrival time, i.e., the time when the nth event occurs.

If $\{\tau_n\}_{n \in \mathbb{N}}$ is a sequence of random variables, i.e., random times, then it is called a **random point process**. For all $n \in \mathbb{N}$, the nth inter-arrival time is defined as $T_n := \tau_n - \tau_{n-1}$ such that the nth arrival (or jump) time is simply the partial sum $\tau_n = T_1 + \cdots + T_n$ of the first n inter-arrival times.

If we define $N_t := \mathsf{max}\{n : \tau_n \leqslant t\}$ with $N_0 := 0$ as the number of points of $\{\tau_n\}_{n \in \mathbb{N}}$ falling into the time interval $(0,t]$, then $\{N_t\}_{t \geqslant 0}$ is the **counting process** for $\{\tau_n\}_{n \in \mathbb{N}}$. Obviously, the sample paths of a counting process are integer-valued, non-negative, and non-decreasing. Likewise, the increment $N_{\Delta t} := N_{t+\Delta t} - N_t$ of a counting process equals the number of events occurring in the corresponding time interval $(t, t+\Delta t]$.

A **Poisson process** $\{N_t\}_{t \geqslant 0}$ with rate (or intensity) $0 < \lambda < \infty$ is a counting process satisfying the following properties:

(P1) $N_0 = 0$.

(P2) $\{N_t\}_{t\geqslant 0}$ has stationary and independent increments $N_{\Delta t}$ for all $\Delta t > 0$.

(P3) For all $\Delta t \to 0$,

$$P(N_{\Delta t} = 1) = \lambda \Delta t + o(\Delta t) \quad \Longleftrightarrow \quad \lim_{\Delta t \to 0} \frac{P(N_{\Delta t} = 1)}{\Delta t} = \lambda$$
$$P(N_{\Delta t} \geqslant 2) = o(\Delta t) \quad \Longleftrightarrow \quad \lim_{\Delta t \to 0} \frac{P(N_{\Delta t} \geqslant 2)}{\Delta t} = 0 \, .$$

Property (P1) is the usual normalization. Property (P2) states that the distribution of $N_{\Delta t}$ is the same for all $\Delta t > 0$ and independent of t and, on the other hand, that the number of events in disjoint time intervals are independent. Note that the Poisson process is the only simple point process with stationary and independent increments. Finally, Property (P3) rules out simultaneous occurrences of event B. On the one hand, the probability that precisely one event B occurs in an arbitrarily small time interval of length Δt is approximately proportional to Δt with proportionality factor λ. On the other hand, the probability of two or more events B to occur in an arbitrarily small time interval of length Δt is negligible (i.e., of order Δt) relative to the probability of one event B to occur.

Interestingly, there are two alternative definitions of Poisson processes which are equivalent to the one given above, and which shed some light on further properties of Poisson processes. The first of these definitions reveals the reason why we speak of 'Poisson' processes:

(P1') $N_0 = 0$.

(P2') $\{N_t\}_{t\geqslant 0}$ has independent increments $N_{\Delta t}$ for all $\Delta t > 0$.

(P3') The increments $N_{\Delta t}$ follow a Poisson distribution, i.e., $N_{\Delta t} \stackrel{d}{=} \mathsf{Poi}(\lambda \Delta t)$, or for $n = 0, 1, 2, \ldots$,

$$P(N_{\Delta t} = n) = \frac{(\lambda \Delta t)^n}{n!} e^{-\lambda \Delta t} \, .$$

This definition relaxes Property (P2), and in turn, puts restriction (P3') on the increments which is more stringent than stationarity. Moreover, Property (P3') implies (P3). Property (P3') states that the number of events occurring in any finite time interval of length Δt is Poisson distributed with mean $\mathsf{E}[\, N_{\Delta t}\,] = \lambda \Delta t$ and $\mathsf{Var}[\, N_{\Delta t}\,] = \lambda \Delta t$. Consequently, the arrival rate equals the mean rate of occurrence of event B (per unit of time), i.e., $N_t \stackrel{d}{=} \mathsf{Poi}(\lambda t)$ implies

$$\lambda = \frac{\mathsf{E}[\, N_t\,]}{t} \, .$$

The second alternative definition of Poisson processes uses an equivalence relation between the distribution of the number of occurrences $P(N_t \geqslant n)$ and the distribution of the corresponding arrival times $P(\tau_n \leqslant t)$. To see this, note that Properties (P2') and (P3') imply

$$\begin{aligned} P(T_1 > t) &= P(\tau_1 - \tau_2 > t) = P(\tau_1 > t) = P(N_t < 1) = P(N_t = 0) = P(N_t - N_0 = 0) \\ &= \mathsf{e}^{-\lambda t} \, . \end{aligned}$$

Put differently, the probability of not observing the first occurrence of event B decays exponentially fast, i.e., the time interval T_1 until the first occurrence is $\mathsf{Exp}(\lambda)$–distributed. Due to Property (P2'), this result can be generalized: The inter-arrival times $\{T_n\}_{n\in\mathbb{N}}$ of a Poisson process are independent and identically distributed as

$$P(T_n > t) = \mathsf{e}^{-\lambda t} \iff P(T_n \leqslant t) = 1 - \mathsf{e}^{-\lambda t} \, ,$$

for all $t \geqslant 0$. As $\mathsf{E}[\, T_n \,] = 1/\lambda$ for $0 < \lambda < \infty$, we see that increasing the intensity λ of a Poisson process decreases the inter-arrival times, and thus, increases the activity of the process. As a very basic result in probability theory, the partial sum of n independent $\mathsf{Exp}(\lambda)$-distributed random variables is Gamma distributed, see (4.1.1), such that the nth arrival time τ_n follows the probability law $\mathsf{Gam}(n, \lambda^{-\lambda})$. Put together, the properties of the definition read as:

(P1") For any $\omega \in \Omega$, the sample paths of $\{N_t\}_{t\geqslant 0}$ are non-negative step functions with jump size 1.

(P2") The inter-arrival times $\{T_n\}_{n\in\mathbb{N}}$ are independent and $\mathsf{Exp}(\lambda)$–distributed, i.e.,

$$P(T_n > t) = \begin{cases} \mathsf{e}^{-\lambda t} & \text{for } t \geqslant 0 \\ 1 & \text{for } t < 0. \end{cases}$$

Property (P2") has two important implications: First, as is well known, the exponential distribution is the only continuous distribution which is memoryless, i.e., if $T \stackrel{d}{=} \mathsf{Exp}(\lambda)$, then

$$P\big(T > t + s \big| T > s\big) = P(T > s) \, .$$

This memoryless property along with the independence of inter-arrival times leads to the conclusion that the increments of such a process are stationary and independent. The second implication is that, given any finite time interval $(0, t]$, the location of jumps are uniformly distributed on $(0, t]$. To see this, assume that exactly one event has occurred in

$(0,t]$. Next, divide $(0,t]$ into a finite number of sub-intervals of equal length. Then, due to the stationarity and independence of the increments, the probability that the event has occurred in any of these sub-intervals is the same, i.e., uniform. This is indeed *the* fundamental insight for simulating a Poisson process: Conditional on the number N_t of jumps which have occurred with law $\mathsf{Poi}(\lambda t)$ on some given time interval $(0,t]$, the arrival times $\tau_1,\ldots,\tau_{N_t}$ have the same distribution as the order statistics of N_t independent random variables which follow a uniform distribution on $(0,t]$.

The **characteristic function of a Poisson process** $\{N_t\}_{t\geqslant 0}$ with intensity $0<\lambda<\infty$ follows directly from the knowledge of its marginal law $N_t \stackrel{d}{=} \mathsf{Poi}(\lambda t)$ and a simple Taylor series expansion:

$$\begin{aligned}
\Phi_t(u) &= \mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uN_t}\,\big] = \sum_{n=0}^{\infty}\mathsf{e}^{\mathsf{i}un}P(N_t=n)\\
&= \sum_{n=0}^{\infty}\mathsf{e}^{\mathsf{i}un}\,\frac{(\lambda t)^n}{n!}\,\mathsf{e}^{-\lambda t} = \mathsf{e}^{-\lambda t}\sum_{n=0}^{\infty}\frac{(\lambda t\mathsf{e}^{\mathsf{i}u})^n}{n!}\\
&= \mathsf{e}^{-\lambda t}\left(1+\frac{\lambda t\mathsf{e}^{\mathsf{i}u}}{1!}+\frac{(\lambda t\mathsf{e}^{\mathsf{i}u})^2}{2!}+\cdots\right)\\
&= \mathsf{e}^{-\lambda t}\mathsf{exp}(\lambda t\mathsf{e}^{\mathsf{i}u}) = \mathsf{exp}\big\{\lambda t(\mathsf{e}^{\mathsf{i}u}-1)\big\}\\
&= \mathsf{exp}\big\{t\Psi(u)\big\}\;,
\end{aligned}$$

where $\Psi(u)=\lambda(\mathsf{e}^{\mathsf{i}u}-1)$ is the characteristic exponent of a Poisson random variable. Obviously, N_t is infinitely divisible.

An important concept in the context of Poisson processes is the notion of **compensation**. Roughly speaking, compensation boils down to the operation of de-meaning. A compensated Poisson process $\tilde{N}_t$ is a Poisson process N_t adjusted to be a martingale with independent and stationary increments:

$$\tilde{N}_t \equiv N_t - \mathsf{E}[\,N_t\,] = N_t - \lambda t\;.$$

The **characteristic function of a compensated Poisson process** $\{\tilde{N}_t\}_{t\geqslant 0}$ is derived as follows:

$$\begin{aligned}
\Phi_t(u) &= \mathsf{E}\Big[\,\mathsf{e}^{\mathsf{i}u\tilde{N}_t}\,\Big] = \mathsf{E}\big[\,\mathsf{exp}\big\{\mathsf{i}u(N_t-\lambda t)\big\}\,\big]\\
&= \sum_{n=0}^{\infty}\mathsf{E}\Big[\,\mathsf{exp}\big\{\mathsf{i}u(N_t-\lambda t)\big\}\Big|N_t=n\,\Big]\,P(N_t=n)\\
&= \sum_{n=0}^{\infty}\mathsf{e}^{\mathsf{i}u(n-\lambda t)}P(N_t=n) = \sum_{n=0}^{\infty}\mathsf{e}^{\mathsf{i}u(n-\lambda t)}\,\frac{(\lambda t)^n}{n!}\,\mathsf{e}^{-\lambda t}
\end{aligned}$$

$$= \mathsf{e}^{-\lambda t}\mathsf{e}^{-\lambda t\mathsf{i}u}\sum_{n=0}^{\infty}\frac{(\lambda t\mathsf{e}^{\mathsf{i}u})^n}{n!} = \mathsf{e}^{-\lambda t(1+\mathsf{i}u)}\mathsf{exp}(\lambda t\mathsf{e}^{\mathsf{i}u}) = \mathsf{exp}\big\{t\lambda(\mathsf{e}^{\mathsf{i}u}-1-\mathsf{i}u)\big\}$$
$$= \mathsf{exp}\big\{t\Psi(u)\big\}\ ,$$

where $\Psi(u) = \lambda(\mathsf{e}^{\mathsf{i}u} - 1 - \mathsf{i}u)$ is the characteristic exponent of a compensated Poisson process $\tilde{N}_t$.

Based on the law of total probability and assuming $\Phi_X(u) = \mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX}\,\big]$ to be the characteristic function of X, we can easily derive the **characteristic function of a compound Poisson process** S_{N_t} in (1.1.4):

$$\begin{aligned}
\Phi_{S_{N_t}}(u) &= \mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uS_{N_t}}\,\big] = \sum_{n=0}^{\infty}\mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uS_{N_t}}\big|N_t=n\,\big]\,P(N_t=n)\\
&= \sum_{n=0}^{\infty}\mathsf{E}\left[\,\mathsf{exp}\left\{\mathsf{i}u\sum_{k=0}^{\infty}X_k\right\}\middle|\,N_t=n\right]P(N_t=n)\\
&= \sum_{n=0}^{\infty}\mathsf{E}\big[\,\mathsf{exp}\big\{\mathsf{i}u(X_1+\cdots+X_n)\big\}\,\big]\,P(N_t=n)\\
&= \sum_{n=0}^{\infty}\mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX_1}\cdots\mathsf{e}^{\mathsf{i}uX_n}\,\big]\,P(N_t=n) = \sum_{n=0}^{\infty}\mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX_1}\,\big]\cdots\mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX_n}\,\big]\,P(N_t=n)\\
&= \sum_{n=0}^{\infty}\underbrace{\Phi_X(u)\cdots\Phi_X(u)}_{n\text{-times}}P(N_t=n) = \sum_{n=0}^{\infty}\big[\Phi_X(u)\big]^n\frac{(\lambda t)^n}{n!}\,\mathsf{e}^{-\lambda t}\\
&= \mathsf{e}^{-\lambda t}\sum_{n=0}^{\infty}\frac{\big[\lambda t\Phi_X(u)\big]^n}{n!} = \mathsf{e}^{-\lambda t}\left(1+\frac{\lambda t\Phi_X(u)}{1!}+\frac{\big[\lambda t\Phi_X(u)\big]^2}{2!}+\cdots\right)\\
&= \mathsf{e}^{-\lambda t}\mathsf{exp}\big\{\lambda t\Phi_X(u)\big\} = \mathsf{exp}\Big\{\lambda t\big[\Phi_X(u)-1\big]\Big\} = \mathsf{exp}\Big\{\lambda t\mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX}-1\,\big]\,\Big\}\\
&= \mathsf{exp}\left\{\lambda t\int_{\mathbb{R}\setminus\{0\}}(\mathsf{e}^{\mathsf{i}ux}-1)F(\mathsf{d}x)\right\} = \mathsf{exp}\left\{t\int_{\mathbb{R}\setminus\{0\}}(\mathsf{e}^{\mathsf{i}ux}-1)\lambda F(\mathsf{d}x)\right\}\\
&= \mathsf{exp}\big\{t\Psi(u)\big\}\ ,
\end{aligned}$$

where $\Psi(u) = \int_{\mathbb{R}\setminus\{0\}}(\mathsf{e}^{\mathsf{i}ux}-1)\lambda F(\mathsf{d}x)$ is the characteristic exponent of a compound Poisson process.

Before deriving the **characteristic function of a compensated compound Poisson process**, we first derive the expectation and variance of a compound Poisson process. Assuming $\mu_X := \mathsf{E}[\,X\,] < \infty$ and using the law of iterated expectations, the expectation of a compound Poisson process S_{N_t} is

$$\mathsf{E}[\,S_{N_t}\,] = \mathsf{E}\big[\,\mathsf{E}\big[\,S_{N_t}\big|N_t\,\big]\,\big] = \mathsf{E}\left[\,\mathsf{E}\left[\sum_{k=1}^{N_t}X_k\middle|N_t\right]\right] = \mathsf{E}\left[\sum_{k=1}^{N_t}\mathsf{E}[\,X_k\,]\right]$$

$$= \mathsf{E}\left[\sum_{k=1}^{N_t} \mathsf{E}[\,X\,]\right] = \mathsf{E}[\,N_t \mathsf{E}[\,X\,]\,] = \mathsf{E}[\,N_t\,]\,\mathsf{E}[\,X\,] = \lambda t \mathsf{E}[\,X\,]\ ,$$

while we use the law of total variance in order to derive its variance

$$\begin{aligned}
\mathsf{Var}[\,S_{N_t}\,] &= \mathsf{E}\big[\,\mathsf{Var}\big[\,S_{N_t}\big|N_t\,\big]\,\big] + \mathsf{Var}\big[\,\mathsf{E}\big[\,S_{N_t}\big|N_t\,\big]\,\big] = \mathsf{E}\left[\,\mathsf{Var}\left[\sum_{k=1}^{N_t} X_k \Big| N_t\right]\right] + \mathsf{Var}[\,N_t \mathsf{E}[\,X\,]\,] \\
&= \mathsf{E}\left[\sum_{k=1}^{N_t} \mathsf{Var}[\,X\,]\right] + \mathsf{Var}[\,N_t\,]\,\mathsf{E}[\,X\,]^2 = \mathsf{E}[\,N_t \mathsf{Var}[\,X\,]\,] + \lambda t \mathsf{E}[\,X\,]^2 \\
&= \mathsf{E}[\,N_t\,]\,\mathsf{Var}[\,X\,] + \lambda t \mathsf{E}[\,X\,]^2 = \lambda t \left\{\mathsf{Var}[\,X\,] + \mathsf{E}[\,X\,]^2\right\} \\
&= \lambda t \mathsf{E}\big[\,X^2\,\big]\ .
\end{aligned}$$

The characteristic function of a compensated compound Poisson process $\tilde{S}_{N_t}$ in (1.1.6) is derived as follows:

$$\begin{aligned}
\Phi_{\tilde{S}_{N_t}}(u) &= \mathsf{E}\Big[\,\mathsf{e}^{\mathsf{i}u\tilde{S}_{N_t}}\,\Big] = \mathsf{E}\Big[\,\mathsf{exp}\Big\{\mathsf{i}u\big(S_{N_t} - \lambda t \mathsf{E}[\,X\,]\big)\Big\}\,\Big] \\
&= \sum_{n=0}^{\infty} \mathsf{E}\left[\,\mathsf{exp}\left\{\mathsf{i}u \sum_{k=0}^{N_t} X_k - \mathsf{i}u\lambda t \mathsf{E}[\,X\,]\right\} \Big| N_t = n\right] P(N_t = n) \\
&= \mathsf{exp}\big\{-\mathsf{i}u\lambda t \mathsf{E}[\,X\,]\,\big\} \sum_{n=0}^{\infty} \mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX_1} \cdots \mathsf{e}^{\mathsf{i}uX_n}\,\big]\, P(N_t = n) \\
&= \mathsf{exp}\big\{-\mathsf{i}u\lambda t \mathsf{E}[\,X\,]\,\big\} \sum_{n=0}^{\infty} \big[\Phi_X(u)\big]^n \frac{(\lambda t)^n}{n!}\, \mathsf{e}^{-\lambda t} \\
&= \mathsf{exp}\Big\{-\lambda t\big(1 + \mathsf{i}u\mathsf{E}[\,X\,]\big)\Big\} \sum_{n=0}^{\infty} \frac{\big[\lambda t \Phi_X(u)\big]^n}{n!} \\
&= \mathsf{exp}\Big\{-\lambda t\big(1 + \mathsf{i}u\mathsf{E}[\,X\,]\big)\Big\} \mathsf{exp}\big\{\lambda t \Phi_X(u)\big\} = \mathsf{exp}\Big\{\lambda t \mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX}\,\big] - 1 - \mathsf{i}u\mathsf{E}[\,X\,]\Big\} \\
&= \mathsf{exp}\Big\{\lambda t \mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX} - 1 - \mathsf{i}uX\,\big]\Big\} = \mathsf{exp}\left\{t \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}ux)\lambda F(\mathsf{d}x)\right\} \\
&= \mathsf{exp}\big\{t\Psi(u)\big\}\ ,
\end{aligned}$$

where $\Psi(u) = \int_{\mathbb{R}\setminus\{0\}} (\mathsf{e}^{\mathsf{i}ux} - 1 - \mathsf{i}ux)\lambda F(\mathsf{d}x)$ is the characteristic exponent of a compensated compound Poisson process $\tilde{S}_{N_t}$.

The **characteristic function of a gamma process** $\{X_t\}_{t \geqslant 0}$ follows directly from the knowledge of the characteristic function $\Phi_X(u)$ of a gamma random variable X via the gamma additivity property (4.1.2). Thus, it suffices to derive the characteristic function of a gamma random variable. Since a gamma random variable is positive, we can apply the Laplace transform of Definition A.5 for accomplishing this aim. From Definition A.5

and the gamma density function $f_X(x;\alpha,\beta)$ in (4.1.1), if follows that

$$\Phi_X^+(z) = \int_0^\infty \mathsf{e}^{-zx} f_X(x;\alpha,\beta)\,\mathsf{d}x = \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}\mathsf{e}^{-(z+1/\beta)x}\,\mathsf{d}x\ .$$

Using the change of variable $y=(z+1/\beta)x$, we conclude

$$\Phi_X^+(z) = \frac{(z+1/\beta)^{-\alpha}}{\beta^\alpha\Gamma(\alpha)}\underbrace{\int_0^\infty y^{\alpha-1}\mathsf{e}^{-y}\,\mathsf{d}y}_{=\Gamma(\alpha)} = (1+z\beta)^{-\alpha}$$

such that

$$\Phi_X(u) = (1-\mathsf{i}u\beta)^{-\alpha} \tag{A.1.8}$$

by setting $z=-\mathsf{i}u$ for $u<(\mathsf{i}\beta)^{-1}$. Finally, it follows from Conditions C2 and C3 of Definition 1.1 and (4.1.2) that the characteristic function of a gamma process X_t reads as

$$\Phi_{X_t}(u) = \mathsf{E}\big[\,\mathsf{e}^{\mathsf{i}uX_t}\,\big] = (1-\mathsf{i}u\beta)^{-\alpha t}\ . \tag{A.1.9}$$

An alternative, yet extremely useful, representation of the characteristic function (A.1.8) of a gamma random variable X is obtained by showing that

$$\mathsf{ln}(1+z\beta) = \int_0^\infty \mathsf{e}^{-x/\beta}\int_0^z \mathsf{e}^{-yx}\,\mathsf{d}y\,\mathsf{d}x\ .$$

To this end, note that

$$\int_0^z\int_0^\infty \mathsf{e}^{-(1/\beta+y)x}\,\mathsf{d}x\,\mathsf{d}y = \int_0^z\left[-\frac{1}{1/\beta+y}\,\mathsf{e}^{-(1/\beta+y)x}\right]_0^\infty \mathsf{d}y = \int_0^z \frac{1}{1/\beta+y}\,\mathsf{d}y\ .$$

Using the change of variable $x=1/\beta+y$, we arrive at

$$\int_0^\infty \mathsf{e}^{-x/\beta}\int_0^z \mathsf{e}^{-yx}\,\mathsf{d}y\,\mathsf{d}x = \big[\mathsf{ln}(1/\beta+y)\big]_0^z = \mathsf{ln}(1+z\beta)\ .$$

Thus, the Laplace transform of a gamma random variable X can be rewritten as

$$\begin{aligned}\Phi_X^+(z) &= (1+z\beta)^{-\alpha} = \mathsf{exp}\big\{-\alpha\mathsf{ln}(1+z\beta)\big\}\\ &= \mathsf{exp}\left\{-\alpha\int_0^\infty \mathsf{e}^{-x/\beta}\int_0^z \mathsf{e}^{-yx}\,\mathsf{d}y\,\mathsf{d}x\right\} = \mathsf{exp}\left\{-\alpha\int_0^\infty \mathsf{e}^{-x/\beta}\left[-\frac{1}{x}\,\mathsf{e}^{-yx}\right]_0^z \mathsf{d}x\right\}\\ &= \mathsf{exp}\left\{\int_0^\infty (\mathsf{e}^{-zx}-1)\,\frac{\alpha}{x}\,\mathsf{e}^{-x/\beta}\,\mathsf{d}x\right\}\ .\end{aligned}$$

Finally, by setting $z=-\mathsf{i}u$, the characteristic function of a gamma random variable X

reads

$$\Phi_X(u) = \exp\left\{\int_0^\infty (\mathrm{e}^{\mathrm{i}ux} - 1)\,\frac{\alpha}{x}\,\mathrm{e}^{-x/\beta}\,\mathrm{d}x\right\}\ . \tag{A.1.10}$$

Representation (A.1.10) of the characteristic function of a gamma random variable X allows us to explicitly derive, via the cumulant formula (A.1.1) and formulae (A.1.2)–(A.1.5), the first four **moments of a gamma random variable**. The mean

$$\mathsf{E}[\,X\,] = \kappa_1 = \alpha\beta$$

of a gamma random variable X follows from

$$\begin{aligned}
\Psi(u) &= \ln\Phi_X(u) = \alpha\int_0^\infty \frac{1}{x}\,\left(\mathrm{e}^{(\mathrm{i}u-1/\beta)x} - \mathrm{e}^{-x/\beta}\right)\,\mathrm{d}x\\
\Psi'(u) &= \mathrm{i}\alpha\int_0^\infty \mathrm{e}^{(\mathrm{i}u-1/\beta)x}\,\mathrm{d}x\\
\Psi'(0) &= \mathrm{i}\alpha\int_0^\infty \mathrm{e}^{-x/\beta}\,\mathrm{d}x\\
\kappa_1 &= \frac{1}{\mathrm{i}}\,\Psi'(0) = \alpha\beta\ .
\end{aligned}$$

The variance

$$\mathsf{Var}[\,X\,] = \kappa_2 = \alpha\beta^2$$

of a gamma random variable X follows from

$$\begin{aligned}
\Psi''(u) &= -\alpha\int_0^\infty x\mathrm{e}^{(\mathrm{i}u-1/\beta)x}\,\mathrm{d}x\\
\Psi''(0) &= -\alpha\int_0^\infty x\mathrm{e}^{-x/\beta}\,\mathrm{d}x = -\alpha\left\{\left[-\beta x\mathrm{e}^{-x/\beta}\right]_0^\infty + \int_0^\infty \beta\mathrm{e}^{-x/\beta}\,\mathrm{d}x\right\} = -\alpha\beta^2\\
\kappa_2 &= \frac{1}{\mathrm{i}^2}\,\Psi''(0) = \alpha\beta^2\ ,
\end{aligned}$$

via integration by parts. The skewness

$$\mathsf{skewness}[X] = \frac{\kappa_3}{\kappa_2^{3/2}} = \frac{2}{\sqrt{\alpha}}$$

of a gamma random variable X follows from

$$\begin{aligned}
\Psi^{(3)}(u) &= -\mathrm{i}\alpha\int_0^\infty x^2\mathrm{e}^{(\mathrm{i}u-1/\beta)x}\,\mathrm{d}x\\
\Psi^{(3)}(0) &= -\mathrm{i}\alpha\int_0^\infty x^2\mathrm{e}^{-x/\beta}\,\mathrm{d}x = -\mathrm{i}\alpha\left\{\left[-\beta x^2\mathrm{e}^{-x/\beta}\right]_0^\infty + \int_0^\infty 2x\beta\mathrm{e}^{-x/\beta}\,\mathrm{d}x\right\} = -2\mathrm{i}\alpha\beta^3
\end{aligned}$$

$$\kappa_3 = \frac{1}{\mathsf{i}^3}\,\Psi^{(3)}(0) = 2\alpha\beta^3\ ,$$

via integration by parts. The excess kurtosis

$$\mathsf{excess\ kurtosis}[X] = \frac{\kappa_4}{\kappa_2^2} = \frac{6}{\alpha}$$

of a gamma random variable X follows from

$$\begin{aligned}
\Psi^{(4)}(u) &= \alpha \int_0^\infty x^3 \mathsf{e}^{(\mathsf{i}u - 1/\beta)x}\,\mathsf{d}x \\
\Psi^{(4)}(0) &= \alpha \int_0^\infty x^3 \mathsf{e}^{-x/\beta}\,\mathsf{d}x = \alpha\left\{\left[-\beta x^3 \mathsf{e}^{-x/\beta}\right]_0^\infty + \int_0^\infty 3x^2\beta \mathsf{e}^{-x/\beta}\,\mathsf{d}x\right\} = 6\alpha\beta^4 \\
\kappa_4 &= \frac{1}{\mathsf{i}^4}\,\Psi^{(4)}(0) = 6\alpha\beta^4\ ,
\end{aligned}$$

via integration by parts.

Bibliography

Aït-Sahalia, Y. (2007): "Estimating Continuous-Time Models with Discretely Sampled Data," in *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress*, ed. by R. Blundell, T. Persson, and W. K. Newey. Econometrics Society Monograph, Cambridge University Press.

Aït-Sahalia, Y., and J. Jacod (2009a): "Estimating the Degress of Activity of Jumps in High Frequency Data," *The Annals of Statistics*, 37(5), 2202–2244.

——— (2009b): "Testing for Jumps in a Discretely Observed Process," *The Annals of Statistics*, 37(1), 184–222.

——— (2010): "Is Brownian Motion Necessary to Model High-Frequency Data?," *The Annals of Statistics*, 38(2), 3093–3128.

——— (2011): "Testing Whether Jumps Have Finite or Infinite Activity," *The Annals of Statistics*, 39(3), 1689–1719.

Aït-Sahalia, Y., and P. A. Mykland (2004): "Estimators of Diffusions with Randomly Spaced Discrete Observations: A General Theory," *The Annals of Statistics*, 32(5), 2186–2222.

——— (2008): "An Analysis of Hansen-Scheinkman Moment Estimators for Discretely and Randomly Sampled Diffusions," *Journal of Econometrics*, 144, 1–26.

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005): "How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise," *The Review of Financial Studies*, 18(2), 351–416.

Akaike, H. (1977): "On Entropy Maximization Principle," in *Applications of Statistics*, ed. by P. R. Krishnaiah, pp. 27–41. North-Holland.

Akgiray, V., and G. G. Booth (1988): "The Stable-Law Model of Stock Returns," *Journal of Business & Economic Statistics*, 6(1), 51–57.

Akritas, M. G. (1982): "Asymptotic Inference for Estimatng the Parameters of a Lévy Process," *Annals of the Institute of Statistical Mathematics*, 34, 259–280.

Akritas, M. G., and R. A. Johnson (1981): "Asymptotic Inference in Lévy Processes of the Discontinuous Type," *The Annals of Statistics*, 9(3), 604–614.

Andersen, T. G., and L. Benzoni (2009): "Operator Methods for Continuous-Time Markov Processes," in *Handbook of Financial Time Series*, ed. by T. G. Andersen, R. A. Davis, J.-P. Kreiß, and T. Mikosch, pp. 555–575. Springer.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001): "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*, 96(453), 42–55.

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62(1), 43–72.

Ané, T., and H. Geman (2000): "Order Flow, Transaction Clock, and Normality of Asset Returns," *The Journal of Finance*, 55(5), 2259–2284.

Applebaum, D. (2004): *Lévy Processes and Stochastic Calculus*. Cambridge University Press.

Asmussen, S., and J. Rosiński (2001): "Approximations of Small Jumps of Lévy Processes with a View Towards Simulation," *Journal of Applied Probability*, 38(2), 482–493.

Bachelier, L. (1900): "Théorie de la Spéculation," *Annales scientifiques de l'École Normale Supérieure*, 17(3), 21–86.

Bahadur, R. R. (1958): "Examples of Inconsistency of Maximum Likelihood Estimates," *Sankhyā*, 20(3/4), 207–210.

——— (1967): "Rates of Convergence of Estimates and Test Statistics," *The Annals of Mathematical Statistics*, 38(2), 303–324.

Ball, C. A., and W. N. Torous (1983): "A Simplified Jump Process for Common Stock Returns," *Journal of Financial and Quantitative Analysis*, 18(1), 53–65.

Barndorff-Nielsen, O. E. (1978): "Hyperbolic Distributions and Distributions on Hyperbolae," *Scandinavian Journal of Statistics*, 5(3), 151–157.

——— (1997): "Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling," *Scandinavian Journal of Statistics*, 24(1), 1–13.

——— (1998): "Processes of Normal Inverse Gaussian Type," *Finance and Stochastics*, 2(1), 41–68.

Barndorff-Nielsen, O. E., and N. Shephard (2001): "Non-Gaussian Ornstein-Uhlenbeck-Based Models and Some of Their Uses in Financial Economics (with discussion)," *Journal of the Royal Statistical Society: Series B*, 63(2), 167–241.

——— (2002): "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of the Royal Statistical Society: Series B*, 62(2), 253–280.

——— (2004): "Power and Bipower Variation with Stochastic Volatility and Jumps (with discussion)," *Journal of Financial Econometrics*, 2(1), 1–37.

——— (2006): "Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation," *Journal of Financial Econometrics*, 4(1), 1–30.

——— (2007): "Variation, Jumps, Market Frictions and High Frequency Data in Financial Econometrics," in *Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress*, ed. by R. Blundell, T. Persson, and W. K. Newey. Econometrics Society Monograph, Cambridge University Press.

Barron, A. R. (1994): "Approximation and Estimation Bounds for Artificial Neural Networks," *Machine Learning*, 14, 115–133.

Barron, A. R., L. Birgé, and P. Massart (1999): "Risk Bounds for Model Selection via Penalization," *Probability Theory and Related Fields*, 113(3), 301–413.

Barron, A. R., and T. M. Cover (1991): "Minimum Complexity Density Estimation," *IEEE Transactions on Automatic Control*, 37(4), 1034–1054.

Basawa, I. V., and P. J. Brockwell (1982): "Non-Parametric Estimation for Non-Decreasing Lévy Processes," *Journal of the Royal Statistical Society: Series B*, 44(2), 262–269.

Beran, R. (1977): "Minimum Hellinger Distance Estimates for Parametric Models," *The Annals of Statistics*, 5(3), 445–463.

Bertoin, J. (1996): *Lévy Processes*. Cambridge University Press.

Birgé, L. (1983): "Approximation dans les Espaces Métriques et Théorie de l'Estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2), 181–237.

Birgé, L., and P. Massart (1993): "Rates of Convergence of Minimum Contrast Estimators," *Probability Theory and Related Fields*, 97, 113–150.

——— (1997): "From Model Selection to Adaptive Estimation," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, ed. by D. Pollard, E. Torgersen, and G. L. Yang. Springer.

——— (1998): "Minimum Contrast Estimators On Sieves: Exponential Bounds and Rates of Convergence," *Bernoulli*, 4(3), 329–375.

Birgé, L., and Y. Rozenholc (2006): "How Many Bins Should Be Put in a Regular Histogram," *ESAIM: Probability and Statistics*, 10, 24–45.

Black, F., and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities," *The Journal of Political Economy*, 81(3), 637–654.

Blumenthal, R. M., and R. K. Getoor (1961): "Sample Functions of Stochastic Processes with Stationary Independent Increments," *Journal of Mathematics and Mechanics*, 10(3), 493–516.

Bochner, S. (1955): *Harmonic Analysis and the Theory of Probability*. University of California Press.

Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31(3), 307–327.

Brown, L. D., A. V. Carter, M. G. Low, and C.-H. Zhang (2004): "Equivalence Theory for Density Estimation, Poisson Processes and Gaussian White Noise with Drift," *The Annals of Statistics*, 32(5), 2074–2097.

Brown, L. D., and M. G. Low (1996): "Asymptotic Equivalence of Nonparametric Regression and White Noise," *The Annals of Statistics*, 24(6), 2384–2398.

Cai, T. T. (1999): "Adaptive Wavelet Estimation: A Block Thresholding and Oracle Inequality Approach," *The Annals of Statistics*, 27(3), 898–924.

Carr, P., H. Geman, D. B. Madan, and M. Yor (2002): "The Fine Structure of Asset Returns: An Empirical Investigation," *Journal of Business*, 75(2), 305–332.

——— (2003): "Stochastic Volatility for Lévy Processes," *Mathematical Finance*, 13(3), 345–382.

Carrasco, M., J.-P. Florens, and E. Ghysels (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6, pp. 5633–5751. North Holland.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6, pp. 5549–5632. North Holland.

Chicken, E., and T. T. Cai (2005): "Block Thresholding for Density Estimation: Local and Global Adaptivity," *Journal of Multivariate Analysis*, 95(1), 76–106.

Christensen, O., and K. L. Christensen (2004): *Approximation Theory: From Taylor Polynomials to Wavelets*. Birkhäuser, 3rd corrected edn.

Clark, P. K. (1973): "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices," *Econometrica*, 41(1), 135–155.

Cohen, A., I. Daubechies, and J.-C. Feauveau (1992): "Biorthogonal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, 45(5), 485–560.

Cont, R. (2001): "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues," *Quantitative Finance*, 1(2), 223–236.

Cont, R., and P. Tankov (2003): *Financial Modelling with Jump Processes*. Chapman & Hall/CRC.

Corsi, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7(2), 174–196.

Cosma, A., O. Scaillet, and R. von Sachs (2007): "Multivariate Wavelet-Based Shape Preserving Estimation for Dependent Observations," *Bernoulli*, 13(2), 301–329.

Daubechies, I. (1988): "Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, 41, 909–996.

——— (1992): *Ten Lectures on Wavelets*. SIAM: Society for Industrial and Applied Mathematics.

Dechevsky, L., and S. Penev (1997): "On Shape-Preserving Probabilistic Wavelet Approximators," *Stochastic Analysis and Applications*, 15(2), 187–215.

——— (1998): "On Shape-Preserving Wavelet Estimators of Cumulative Distribution Functions and Densities," *Stochastic Analysis and Applications*, 16(3), 423–462.

Detemple, J., and S. Murthy (1994): "Intertemporal Asset Pricing with Heterogeneous Beliefs," *Journal of Economic Theory*, 62(2), 294–320.

DeVore, R. A., and G. G. Lorentz (1993): *Constructive Approximation*. Springer.

Devroye, L., and G. Lugosi (2001): *Combinatorial Methods in Density Estimation*. Springer.

Donoho, D. L. (1993): "Unconditional Bases Are Optimal Bases for Data Compression and for Statistical Estimation," *Applied and Computational Harmonic Analysis*, 1(1), 100–115.

Donoho, D. L., and I. M. Johnstone (1995): "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90(432), 1200–1223.

Donoho, D. L., I. M. Johnstone, G. Kerkyacharian, and D. Picard (1995): "Wavelet Shrinkage: Asymptopia?," *Journal of the Royal Statistical Society: Series B*, 57(2), 301–369.

——— (1996): "Density Estimation by Wavelet Thresholding," *The Annals of Statistics*, 24(2), 508–539.

Dudley, R. M. (1999): *Uniform Central Limit Theorems*. Cambridge University Press.

——— (2002): *Real Analysis and Probability*. Cambridge University Press.

Duffie, D., and P. Glynn (2004): "Estimation of Continuous-Time Markov Processes Sampled at Random Time Intervals," *Econometrica*, 72(6), 1773–1808.

Duffie, D., and C. Huang (1985): "Implementing Arrow-Debreu Equilibria by Continuous Trading of Few Long-Lived Securities," *Econometrica*, 53(6), 1337–1356.

Duffie, D., J. Pan, and K. J. Singleton (2000): "Transform Analysis and Asset Pricing for Affine Jump-Diffusions," *Econometrica*, 68(6), 1343–1376.

Eberlein, E. (2001): "Application of Generalized Hyperbolic Lévy Motions to Finance," in *Lévy Processes: Theory and Applications*, ed. by O. E. Barndorff-Nielsen, T. Mikosch, and S. I. Resnick, pp. 319–336. Birkhäuser.

EBERLEIN, E., AND U. KELLER (1995): "Hyperbolic Distributions in Finance," *Bernoulli*, 1(3), 281–299.

EBERLEIN, E., U. KELLER, AND K. PRAUSE (1998): "New Insights into Smile, Mispricing, and Value at Risk: The Hyperbolic Model," *Journal of Business*, 71(3), 371–406.

EFROMOVICH, S., AND M. S. PINSKER (1984): "Learning Algorithm for Nonparametric Filtering," *Automation and Remote Control*, 11, 58–65.

ENGLE, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50(4), 987–1006.

ENGLE, R. F., AND J. R. RUSSELL (1998): "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66(5), 1127–1162.

ERAKER, B., M. JOHANNES, AND N. POLSON (2003): "The Impact of Jumps in Volatility and Returns," *The Journal of Finance*, 58(3), 1269–1300.

FAMA, E. F. (1970): "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, 25(2), 383–417.

FAN, J., AND I. GIJBELS (1996): *Local Polynomial Modelling and Its Applications*. Chapman & Hall.

FARRELL, R. H. (1967): "On the Lack of a Uniformly Consistent Sequence of Estimators of a Density Function in Certain Cases," *The Annals of Mathematical Statistics*, 38(2), 471–474.

FIGUEROA-LÓPEZ, J. E. (2009): "Nonparametric Estimation for Lévy Models Based on Discrete-Sampling," in *Optimality: The Third Erich L. Lehmann Symposium*, ed. by J. Rojo, vol. 57 of *Lecture Notes–Monograph Series*, pp. 117–146. Institute of Mathematical Statistics.

——— (2011): "Sieve-Based Confidence Intervals and Bands for Lévy Densities," *Bernoulli*, 17(2), 643–670.

FIGUEROA-LÓPEZ, J. E., AND C. HOUDRÉ (2006): "Risk Bounds for the Non-Parametric Estimation of Lévy Processes," in *High Dimensional Probability*, ed. by E. Giné, V. Koltchinskii, W. Li, and J. Zinn, vol. 51 of *Lecture Notes–Monograph Series*, pp. 96–116. Institute of Mathematical Statistics.

FISHER, R. A. (1912): "On an Absolute Criterion for Fitting Frequency Curves," *Messenger of Mathematics*, 41, 155–160.

——— (1921): "On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample," *Metron*, 1, 3–32.

——— (1922): "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London: Series A*, 222, 309–368.

Florens-Zmirou, D. (1993): "On Estimating the Diffusion Coefficient from Discrete Observations," *Journal of Applied Probability*, 30, 790–804.

Fristedt, B., and L. Gray (1997): *A Modern Approach to Probability Theory*. Birkhäuser.

Geman, H. (2002): "Pure Jump Lévy Processes for Asset Price Modelling," *Journal of Banking and Finance*, 26(7), 1297–1316.

Geman, H., D. B. Madan, and M. Yor (2001): "Time Changes for Lévy Processes," *Mathematical Finance*, 11(1), 79–96.

Geman, S., and C.-R. Hwang (1982): "Nonparametric Maximum Likelihood Estimation by the Methods of Sieves," *The Annals of Statistcs*, 10(2), 401–414.

Giné, E., and R. Nickl (2009): "Uniform Limit Theorems for Wavelet Density Estimators," *The Annals of Probability*, 37(4), 1605–1646.

Giné, E., and J. Zinn (1984): "Some Limit Theorems for Empirical Processes," *The Annals of Probability*, 12(4), 929–989.

Glad, I. K., N. L. Hjort, and N. G. Ushakov (2003): "Correction of Density Estimators That Are Not Densities," *Scandinavian Journal of Statistics*, 30(2), 415–427.

Glosten, L. R., and P. R. Milgrom (1985): "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 14(1), 71–100.

Godambe, V. P. (1960): "An Optimum Property of Regular Maximum Likelihood Estimation," *The Annals of Mathematical Statistics*, 31(4), 1208–1211.

Good, I. J., and R. A. Gaskins (1971): "Nonparametric Roughness Penalities for Probability Densities," *Biometrika*, 58(2), 255–277.

Grenander, U. (1981): *Abstract Inference*. Wiley.

Gugushvili, S. (2009): "Nonparametric Estimation of the Characteristic Triplet of a Discretely Observed Lévy Process," *Journal of Nonparametric Statistics*, 21(3), 321–343.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression*. Springer.

Hall, P., G. Kerkyacharian, and D. Picard (1998): "Block Threshold Rules for Curve Estimation Using Kernel and Wavelet Methods," *The Annals of Statistics*, 26(3), 922–942.

——— (1999): "On the Minimax Optimality of Block Threshold Wavelet Estimators," *Statistica Sinica*, 9(1), 33–49.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

Härdle, W. (1992): *Applied Nonparametric Regression*. Cambridge University Press.

Härdle, W., G. Kerkyacharian, D. Picard, and A. Tsybakov (1998): *Wavelets, Approximation, and Statistical Applications*. Springer.

Heston, S. L. (1993): "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bonds and Currency Options," *The Review of Financial Studies*, 6(2), 327–343.

Huber, P. J. (1967): "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed. by L. Le Cam, and J. Neyman, vol. 1, pp. 221–233. University of California Press.

Hull, J. (2000): *Options, Futures, & Other Derivatives*. Prentice Hall, 4th edn.

Hull, J., and A. White (1987): "The Pricing of Options on Assets with Stochastic Volatilities," *The Journal of Finance*, 42(2), 281–300.

Ibragimov, I. A., and R. Z. Hasminskii (1982): "Bounds for the Risks of Nonparametric Estimates of the Regression," *Theory of Probability and its Applications*, 27(1), 84–99.

Itō, K. (1951): "On Stochastic Differential Equations," *Memoirs of the American Mathematical Society*, 4, 1–51.

Jacod, J., and A. N. Shiryaev (2003): *Limit Theorems for Stochastic Processes*. Springer, 2nd edn.

James, W., and C. Stein (1961): "Estimation with Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, vol. 1, pp. 361–379. University of California Press.

Johnstone, I. M. (2011): *Gaussian Estimation: Sequence and Multiresolution Models*. `http://www-stat.stanford.edu/~imj/GE12-27-11.pdf`.

Jongbloed, G., and F. H. van der Meulen (2006): "Parametric Estimation for Subordinators and Induced OU Processes," *Scandinavian Journal of Statistics*, 33, 825–847.

Jongbloed, G., F. H. van der Meulen, and A. W. van der Vaart (2005): "Nonparametric Inference for Lévy-Driven Ornstein-Uhlenbeck Processes," *Bernoulli*, 11(5), 759–791.

Kallenberg, O. (2002): *Foundations of Modern Probability*. Springer, 2nd edn.

Kawata, T. (1972): *Fourier Analysis in Probability Theory*. Academic Press.

Kerkyacharian, G., D. Picard, and K. Tribouley (1996): "L^p Adaptive Density Estimation," *Bernoulli*, 2(3), 229–247.

Kiefer, J., and J. Wolfowitz (1956): "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27(4), 887–906.

Kim, W.-C., and J.-Y. Koo (2002): "Inhomogeneous Poisson Intensity Estimation via Information Projections onto Wavelet Subspaces," *Journal of the Korean Statistical Society*, 31(3), 343–357.

Kingman, J. F. C. (1993): *Poisson Processes*. Oxford University Press.

Kolmogorov, A. N., and V. M. Tihomirov (1961): "ϵ–Entropy and ϵ–Capacity of Sets in Functional Spaces," *American Mathematical Society Transactions*, 17(2), 277–364.

Koltchinskii, V. (1994): "Komlos-Major-Tusnady Approximation for the General Empirical Process and Haar Expansions of Classes of Functions," *Journal of Theoretical Probability*, 7(1), 73–118.

KOSOROK, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

KOU, S. G. (2002): "A Jump–Diffusion Model for Option Pricing," *Management Science*, 48(8), 1086–1101.

KUTOYANTS, Y. A. (2004): *Statistical Inference for Ergodic Diffusion Prozesses*. Springer.

LE CAM, L. (1973): "Convergence of Estimates under Dimensionality Restrictions," *The Annals of Statistics*, 1(1), 38–53.

——— (1986): *Asymptotic Methods in Statistical Decision Theory*. Springer.

——— (1990): "Maximum Likelihood: An Introduction," *International Statistical Review*, 58(2), 153–171.

——— (1997): "Metric Dimension and Statistical Estimation," in *Advances in Mathematical Sciences: CRM's 25 Years*, ed. by L. Vinet, vol. 11, pp. 303–311. The American Mathematical Society.

LEDOUX, M., AND M. TALAGRAND (1991): *Probability in Banach Spaces*. Springer.

LEE, S. L., AND J. HANNIG (2010): "Detecting Jumps from Lévy Jump Diffusion Processes," *Journal of Financial Economics*, 96(2), 271–290.

LEE, S. L., AND P. A. MYKLAND (2006): "Jumps in Real-Time Financial Markets: A New Nonparametric Test and Jump Clustering," in *Third International Workshop in Applied Probability*.

LEPSKII, O. V. (1992): "Asymptotically Minimax Adaptive Estimation I: Upper Bounds. Optimally Adaptive Estimates," *Theory of Probability and Its Applications*, 36(4), 682–697.

LI, Q., AND J. S. RACINE (2006): *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

LIESE, F., AND K. ZIEGLER (1999): "A Note on Empirical Process Methods in the Theory of Poisson Point Processes," *Scandinavian Journal of Statistics*, 26(4), 533–537.

LO, A. W. (1988): "Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data," *Econometric Theory*, 4, 231–247.

Lorentz, G. G., M. v. Golitschek, and Y. Makovoz (1996): *Constructive Approximation: Advanced Problems*. Springer.

Lukacs, E. (1970): *Characteristic Functions*. Griffin.

Madan, D. B., P. P. Carr, and E. C. Chang (1998): "The Variance Gamma Process and Option Pricing," *European Finance Review*, 2(1), 79–105.

Madan, D. B., and E. Seneta (1990): "The Variance Gamma Model for Share Market Returns," *Journal of Business*, 63(4), 511–524.

Madan, D. B., and M. Yor (2008): "Representing the CGMY and Meixner Lévy Processes as Time Changed Brownian Motions," *The Journal of Computational Finance*, 12(1), 27–47.

Mancini, C., and R. Renò (2011): "Threshold Estimation of Markov Models with Jumps and Interest Rate Modeling," *Journal of Econometrics*, 160(1), 77–92.

Mandelbrot, B. (1963): "New Methods in Statistical Economics," *The Journal of Political Economy*, 71(5), 421–440.

Massart, P. (2000): "About the Constants in Talagrand's Concentration Inequalities for Empirical Processes," *The Annals of Probability*, 28(2), 863–884.

——— (2007): *Concentration Inequalities and Model Selection*. Springer.

Masuda, H. (2009): "Notes on Estimating Inverse-Gaussian and Gamma Subordinators under High-Frequency Sampling," *Annals of the Institute of Statistical Mathematics*, 61, 181–195.

Melino, A. (1994): "Estimation of Continuous-Time Models in Finance," in *Advances in Econometrics: Sixth World Congress*, ed. by C. A. Sims, vol. II, pp. 313–351. Cambridge University Press.

Merton, R. C. (1973): "Theory of Rational Option Pricing," *The Bell Journal of Economics and Management Science*, 4(1), 141–183.

——— (1976): "Option Pricing when Underlying Stock Returns Are Discontinuous," *Journal of Financial Economics*, 3, 125–144.

Meyer, Y. (1992): *Wavelets and Operators*. Cambridge University Press.

Monroe, I. (1978): "Processes That Can Be Embedded in Brownian Motion," *The Annals of Probability*, 6(1), 42–56.

NADARAYA, E. A. (1964): "On Estimating Regression," *Theory of Probability and Its Applications*, 9(1), 141–142.

NEUMANN, M. H., AND M. REISS (2009): "Nonparametric Estimation for Lévy Processes from Low-Frequency Observations," *Bernoulli*, 15(1), 223–248.

NEWEY, W. K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59(4), 1161–1167.

NEYMAN, J., AND E. L. SCOTT (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16(1), 1–32.

NUSSBAUM, M. (1985): "Spline Smoothing in Regression Models and Asymptotic Efficiency in L_2," *The Annals of Statistics*, 13(3), 984–997.

——— (1996): "Asymptotic Equivalence of Density Estimation and Gaussian White Noise," *The Annals of Statistics*, 24(6), 2399–2430.

OGDEN, R. T. (1997): *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser.

PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics*. Cambridge University Press.

PARZEN, E. (1962): "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 33(3), 1065–1076.

PINHEIRO, A., AND B. VIDAKOVIC (1997): "Estimating the Square Root of a Density via Compactly Supported Wavelets," *Computational Statistics & Data Analysis*, 25(4), 399–415.

PINSKER, M. S. (1980): "Optimal Filtering of Square-Integrable Signals in Gaussian Noise," *Problems of Information Transmission*, 16(2), 120–133.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer.

PRAKASA RAO, B. L. S. (1999): *Semimartingales and their Statistical Inference*. Chapman & Hall/CRC.

PRESS, S. J. (1967): "A Compound Events Model for Security Prices," *Journal of Business*, 40(3), 317–335.

PROTTER, P. E. (2004): *Stochastic Integration and Differential Equations*. Springer, 2nd edn.

Rachev, S., and S. Mittnik (2000): *Stable Paretian Models in Finance*. Wiley.

Reynaud-Bouret, P. (2003): "Adaptive Estimation of the Intensity of Inhomogeneous Poisson Processes via Concentration Inequalities," *Probability Theory and Related Fields*, 126(1), 103–153.

Reynaud-Bouret, P., and V. Rivoirard (2010): "Near Optimal Thresholding Estimation of a Poisson Intensity on the Real Line," *Electronic Journal of Statistics*, 4, 172–238.

Rosenblatt, M. (1956): "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, 27(3), 832–837.

Rosiński, J. (2001): "Series Representations of Lévy Processes from the Perspective of Point Processes," in *Lévy Processes: Theory and Applications*, ed. by O. E. Barndorff-Nielsen, T. Mikosch, and S. I. Resnick, pp. 401–415. Birkhäuser.

Rubin, H., and H. G. Tucker (1959): "Estimating the Parameters of a Differential Process," *The Annals of Mathematical Statistics*, 30(3), 641–658.

Ruch, D. K., and P. J. van Fleet (2009): *Wavelet Theory: An Elementary Approach with Applications*. Wiley.

Samuelson, P. A. (1965): "Rational Theory of Warrant Pricing," *Industrial Management Review*, 6(2), 13–39.

Sato, K.-I. (1999): *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, 2nd edn.

Schoutens, W. (2003): *Lévy Processes in Finance: Pricing Financial Derivatives*. Wiley.

Seneta, E. (2004): "Fitting the Variance-Gamma Model to Financial Data," *Journal of Applied Probability*, 41A, 177–187.

Shen, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25(6), 2555–2591.

Shen, X., and W. H. Wong (1994): "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22(2), 580–615.

Shimizu, Y. (2006a): "Density Estimation of Lévy Measures for Discretely Observed Diffusion Processes with Jumps," *Journal of the Japan Statistical Society*, 36(1), 37–62.

——— (2006b): "M-Estimation for Discretely Observed Ergodic Diffusion Processes with Infinitely Many Jumps," *Statistical Inference for Stochastic Processes*, 9(2), 179–225.

——— (2009a): "Functional Estimation for Lévy Measures of Semimartingales with Poissonian Jumps," *Journal of Multivariate Analysis*, 100(6), 1073–1092.

——— (2009b): "Model Selection for Lévy Measures in Diffusion Processes with Jumps from Discrete Observations," *Journal of Statistical Planning and Inference*, 139(2), 516–532.

Shimizu, Y., and N. Yoshida (2006): "Estimation of Parameters for Diffusion Processes with Jumps from Discrete Observations," *Statistical Inference for Stochastic Processes*, 9(3), 227–277.

Shiryaev, A. N. (1999): *Essentials of Stochastic Finance: Facts, Models, Theory*. World Scientific Publishing Company.

Shreve, S. E. (2004): *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer.

Silverman, B. W. (1982): "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *The Annals of Statistics*, 10(3), 795–810.

——— (1986): *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.

Song, S. (2010): "Lévy Density Estimation via Information Projection onto Wavelet Subspaces," *Statistics & Probability Letters*, 80(21-22), 1623–1632.

Speckman, P. (1985): "Spline Smoothing and Optimal Rates of Convergence in Nonparametric Regression Models," *The Annals of Statistics*, 13(3), 970–983.

Stein, C. (1981): "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9(6), 1135–1151.

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10(4), 1040–1053.

Talagrand, M. (1994): "Sharper Bounds for Gaussian and Empirical Processes," *The Annals of Probability*, 22(1), 28–76.

——— (1996): "New Concentration Inequalities in Product Spaces," *Inventiones Mathematicae*, 126, 505–563.

TODOROV, V., AND G. TAUCHEN (2010): "Activity Signature Functions for High-Fequency Data Analysis," *Journal of Econometrics*, 154(2), 125–138.

——— (2011): "Volatility Jumps," *Journal of Business & Economic Statistics*, 29(3), 356–371.

TRIEBEL, H. (1992): *Theory of Function Spaces II*. Birkhäuser.

——— (2008): *Function Spaces and Wavelets on Domains*. European Mathematical Society.

TSYBAKOV, A. B. (2009): *Introduction to Nonparametric Estimation*. Springer.

VAN DE GEER, S. (1990): "Estimating a Regression Function," *The Annals of Statistics*, 18(2), 907–924.

——— (1995a): "Exponential Inequalities for Martingales, with Application to Maximum Likelihood Estimation for Counting Processes," *The Annals of Statistics*, 23(5), 1779–1801.

——— (1995b): "The Method of Sieves and Minimum Contrast Estimators," *Mathematical Methods of Statistics*, 4(1), 20–38.

——— (2000): *Empirical Processes in M-Estimation*. Cambridge University Press.

——— (2002): "M-Estimation Using Penalties or Sieves," *Journal of Statistical Planning and Inference*, 108, 55–69.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer.

VAPNIK, V., AND A. ČERVONENKIS (1971): "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Theory of Probability and Its Applications*, 16(2), 264–280.

VIDAKOVIC, B. (1999): *Statistical Modeling by Wavelets*. Wiley.

WAHBA, G. (1981): "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates," *The Annals of Statistics*, 9(1), 146–156.

——— (1990): *Spline Models for Observational Data*. SIAM: Society for Industrial and Applied Mathematics.

WALD, A. (1949): "Note on the Consistency of the Maximum Likelihood Estimate," *The Annals of Mathematical Statistics*, 20(4), 595–601.

Walnut, D. F. (2001): *An Introduction to Wavelet Analysis*. Birkhäuser.

Walter, G. G. (1994): *Wavelets and Other Orthogonal Systems With Applications*. CRC Press.

Wasserman, L. (2006): *All of Nonparametric Statistics*. Springer.

Watson, G. S. (1964): "Smooth Regression Analysis," *Sankhyā Series: A*, 26(4), 359–372.

Wedderburn, R. W. M. (1974): "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61(3), 439–447.

Wolfowitz, J. (1957): "The Minimum Distance Method," *The Annals of Mathematical Statistics*, 28(1), 75–88.

Wong, W. H., and T. A. Severini (1991): "On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces," *The Annals of Statistics*, 19(2), 603–632.

Wong, W. H., and X. Shen (1995): "Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLEs," *The Annals of Statistics*, 23(2), 339–362.

Wu, L. (2011): "Variance Dynamics: Joint Evidence from Options and High-Frequency Returns," *Journal of Econometrics*, 160(1), 280–287.

Yatracos, Y. G. (1985): "Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy," *The Annals of Statistics*, 13(2), 768–774.

Zhang, L. (2006): "Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach," *Bernoulli*, 12(6), 1019–1043.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2005): "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association*, 100(472), 1394–1411.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den ....

(Unterschrift)